# GEOMETRÍA E INFORMACIÓN

OPTATIVO

Mariela Adelina Portesi
Pedro Walter Lamberti
Steeve Zozor

Facultad de Ciencias Exactas

UNIVERSIDAD NACIONAL DE LA PLATA

edulp
Editorial de la Universidad 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 epģrafe con texto simulado.*
AUTOR DEL EPÍGRAFE, TÍTULO DE LA OBRA

# PRÓLOGO

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

*Los autores*

# 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

*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 Introducción

## 1.2 Probabilidades

El concepto de *probabilidad* es importante en situaciones donde el resultado 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 se puede dar 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 ocurrir o debe ser verdad. Si el proceso tiene $K$ resultados distinguibles, mutuamente excluyentes e igualmente probables (esto es, no se prefiere una posibilidad frente a otras), y si $k$ de esos $K$ resultados tienen un dado atributo, la probabilidad asociada a dicho atributo en un dado proceso es $\frac{k}{K}$. 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 $K$ muy grande de procesos independientes cierto atributo aparece $k$ veces, se identifica a la probabilidad asociada a un proceso o ensayo con la frecuencia relativa de ocurrencia $\frac{k}{K}$ del atributo (van Brakel, 1976; Hald, 1990; Shafer & Vovk, 2006, & Ref.) [1].

Los axiomas de Kolmogorov [2] proveen requisitos suficientes para determinar completamente las propiedades 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 de *muestras* (*outcomes*, en inglés) $\omega \in \Omega$. Se asocia $\mathcal{A}$ a una colección de sub-conjuntos de $\Omega$, donde los elementos de $\mathcal{A}$ son llamados *eventos*. Por ejemplo, para un dado de 6 caras, $\Omega$ es el conjunto de caras que se pueden etiquetar con los números naturales del 1 al 6 (o también con las letras *a, b, c, d, e, f*, u otro etiquetado), y $\mathcal{A}$ tiene los eventos $A$ "es un número natural par" y $B$ "es un número natural impar". En el caso de analizar el tiempo de vida de un aparato, $\Omega \equiv \mathbb{R}_+$. 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 y definiciones análogas a las de la teoría de conjuntos (ver entre otros (Spiegel, 1976; Brémaud, 1988; Mandel & Wolf, 1995; Sierpiński, 1975, 1976; Borel, 1898, 1909)):

- Combinación o unión de eventos: $A \cup B$ implica que se da $A$, ó $B$, o ambos (por ejemplo, para un dado de 6 caras, si $A$ son los eventos "cara par" y $B$ los eventos "cara menor o igual a 3", resulta $A \cup B = \{1\,,\,2\,,\,3\,,\,4\,,\,6\}$). Según la literatura, se denota a veces $A + B$ o $A \vee B$.

- Intersección de eventos: $A \cap B$ implica que se dan ambos $A$ y $B$ (en el ejemplo precedente, $A \cap B = \{2\}$). Se denota a veces $(A, B)$ o $A \wedge B$.

- Complemento de un evento: $\bar{A}$ indica que no se da $A$. Se denota a veces $-A$ o $A^{\complement}$ (en el ejemplo precedente, $\bar{A} = \{1\,,\,3\,,\,5\}$).

---

[1] A pesar de que las nociones de azar (que proviene del árabe *zahr* que significa dado, flor) o de aleatoriedad (del latín *alea* que es suerte, dado) son muy antiguas (Serrano Marugán, 2000), el matemático italiano y jugador de dados y cartas Gerolamo Cardano es "probablemente" uno de los primeros en tratar matemáticamente el concepto de probabilidad en el siglo XVI, en su libro sobre los juegos de azar escrito en 1564 pero publicado en 1663 (Cardano, 1663) (ver (Bellhouse, 2005) o (Hald, 1990, Cap. 4)). Entre los numerosos matemáticos que desarrollaron la teoría de las probabilidades, en particular los franceses Pierre de Fermat y Blaise Pascal (Hald, 1990, Cap. 5), hay que mencionar también al suizo Jacob Bernoulli (miembro de una dinastía de matemáticos) (Bernoulli, 1713, en latín) o ((E. D. Sylla, Translator), 1713) y al franco-inglés Abraham de Moivre (de Moivre, 1756). El francés Pierre Simon Laplace (de Laplace, 1820) fue quizás uno de los primeros en proveer un aporte importante al desarrollo de la teoría de las probabilidades en los siglos XVIII-XIX, a través del punto de vista "frecuentista" y combinatorial (ver también (Hald, 1990, Caps. 13, 15 & 22)).

[2] Un paso importante es debido a Kolmogorov en 1933 que se apoyó sobre trabajos de Richard von Mises (von Mises, 1932) y también sobre la teoría de la medida y de la integración, debidas entre otros a Émile Borel y Henri Lebesgue (Borel, 1898, 1909; Lebesgue, 1904, 1918; Halmos, 1950), para formalizar analíticamente la teoria de las probabilidades (Kolmogorov, 1956; Barone & Novikoff, 1978; Jacob & Protters, 2003).

- Eventos *disjuntos* o *mutuamente excluyentes* o *incompatibles*: son aquellos que no se superponen, se anota $A \cap B = \emptyset$ donde $\emptyset = \bar{\Omega}$ denota el *evento nulo* (evento que no puede ocurrir, es el complemento de $\Omega$). Por ejemplo los eventos "cara par" y "cara impar" son incompatibles.

- Denotaremos también $A \setminus B$ cuando el evento $A$ se realiza pero no $B$. Se lo denota también $A - B$, que es también $A \cap \bar{B}$ (en el ejemplo precedente, $A \setminus B = \{4\}$).

Esto es ilustrado en la Fig. 1-1 empleando lo que se conoce como diagramas de Venn [3]. La unión y la intersección de eventos satisfacen las mismas reglas que en la teoría de conjuntos, es decir cada una es conmutativa $A \cup B = B \cup A$, $A \cap B = B \cap A$, asociativa $(A \cup B) \cup C = A \cup (B \cup C)$, $(A \cap B) \cap C = A \cap (B \cap C)$, distributiva con respecto a la otra $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$, $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$ (ver por ejemplo (Jeffrey, 1948, 1973; Halmos, 1950; Feller, 1971; Brémaud, 1988; Mandel & Wolf, 1995; Ibarrola, Pardo & Quesada, 1997; Lehmann & Casella, 1998; Athreya & Lahiri, 2006; Cohn, 2013; Hogg, McKean & Craig, 2013)).

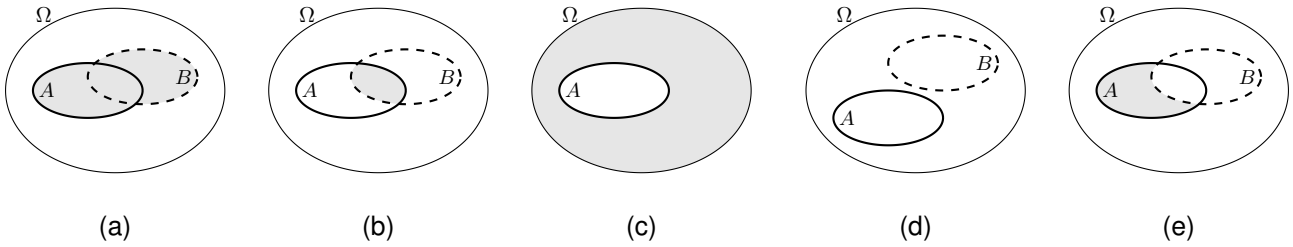



**Figura 1-1:** Ilustración de las operaciones entre eventos: (a) unión $A \cup B$, (b) intersección $A \cap B$, (c) complemento $\bar{A}$, (d) eventos excluyentes $A \cap B = \emptyset$, y (e) $A \setminus B$. $A$ es representado en línea llena, $B$ en línea discontinua; en (a)-(c) y (e), el resultado de la operación es la zona sombreada.

Formalmente, se define de manera abstracta un espacio medible $(\Omega, \mathcal{A})$ de la manera siguiente ( (Halmos, 1950; Feller, 1968, 1971; Brémaud, 1988; Ibarrola et al., 1997; Athreya & Lahiri, 2006; Bogachev, 2007a; Cohn, 2013); ver también (Barone & Novikoff, 1978; Borel, 1898; Sierpiński, 1918, 1975, 1976, & Ref.) para notas históricas):

**Definición 1-1** (Espacio medible)**.** $(\Omega, \mathcal{A})$, *formado por un espacio muestral* $\Omega$ *y una colección* $\mathcal{A}$ *de conjuntos de* $\Omega$, *es llamado espacio medible si satisface los requisitos*

*1.* $\emptyset \in \mathcal{A}$,

*2. si* $A \in \mathcal{A}$, *entonces* $\bar{A} \in \mathcal{A}$,

[3] Este tipo de diagramas fue popularizado por el inglés John Venn en 1880, pero en su trabajo (Venn, 1880) da la paternidad al matemático suizo Leonhard Euler, uno de los primeros en usar tal representación en el siglo XVIII en sus famosas "Cartas a una princesa alemana, acerca de diversas cuestiones de física y filosofía" (ver (Euler, 1768, L 102-105, pp. 95-126)), o antes a Christian Weise y Johan Christian Langius (Langius, 1712); apareció aún en trabajos de Gottfried Wilhelm Leibniz en el siglo anterior.

3. *la unión numerable de conjuntos de $\mathcal{A}$ queda en $\mathcal{A}$ ($\mathcal{A}$ es cerrado por la unión numerable).*

*Con estas propiedades, $\mathcal{A}$ es llamada una $\sigma$-álgebra. Los elementos de $\mathcal{A}$ son dichos medibles.*

Es sencillo mostrar que $\Omega$ también está en $\mathcal{A}$, y que $\mathcal{A}$ es cerrado por la intersección numerable. Un ejemplo de $\sigma$-álgebra sobre $\Omega = \{1\,,\,2\,,\,3\,,\,4\,,\,5\,,\,6\}$ puede ser $\mathcal{A} = \big\{\emptyset\,,\,\Omega\,,\,\{1\,,\,2\,,\,3\}\,,\,\{4\,,\,5\,,\,6\}\big\}$.

A partir de $(\Omega, \mathcal{A})$, se asocia una noción de probabilidad $P$ a un dado evento. Esta queda determinada por los siguientes requisitos llamados *Axiomas de Kolmogorov* (ver por ejemplo (Spiegel, 1976; Kolmogorov, 1956; Shafer & Vovk, 2006; von Plato, 2005)):

1. $P(A) \geq 0 \;\; \forall\, A \in \mathcal{A}$.
2. Si $A_1, \ldots, A_i, \ldots$ son eventos mutuamente excluyentes de $\mathcal{A}$, entonces $P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$.
3. $P(\Omega) = 1$.

Más formalmente, se define un *espacio de probabilidad* o *espacio probabilístico* de la manera siguiente (Halmos, 1950; Feller, 1968, 1971; Brémaud, 1988; Ibarrola et al., 1997; Athreya & Lahiri, 2006; Bogachev, 2007a; Jacob & Protters, 2003; Cohn, 2013):

**Definición 1-2** (Espacio de medida y espacio probabilístico)**.** *Sea $(\Omega, \mathcal{A})$ un espacio medible. Una función $\mu : \mathcal{A} \mapsto \mathbb{R}_+$ tal que*

1. *$\mu(\emptyset) = 0$, y*
2. *para cualquier conjunto numerable $\{A_i\}_{i\in I}$ ($I$ numerable) de elementos mutuamente excluyentes de $\mathcal{A}$ se tiene $\mu\left(\bigcup_i A_i\right) = \sum_i \mu(A_i)$,*

*es llamada función medida o medida $\sigma$-aditiva, y el espacio $(\Omega, \mathcal{A}, \mu)$ es llamado espacio de medida.*

- *Cuando $\mu$ es tal que existe un conjunto numerable $\{A_i\}_{i\in I}$ ($I$ numerable) de elementos de $\mathcal{A}$ tal que $\Omega = \cup_{i\in I} A_i$ y $\forall\, i \in I, \quad \mu(A_i) < +\infty$ finito, la medida se dice $\sigma$-finita y el espacio de medida se dice $\sigma$-finito.*
- *Cuando $\mu$ está acotada por arriba, $\mu(\Omega) < +\infty$, la medida se dice finita y el espacio de medida también se dice finito.*
- *Además, si $\mu(\Omega) = 1$, la medida es dicha medida de probabilidad. En general, se la denota $P$. En este caso, el espacio $(\Omega, \mathcal{A}, P)$ es llamado espacio probabilístico.*

(ver también (Kolmogorov & Fomin, 1961, Cap. 5 & 6)). Es importante notar que una combinación lineal positiva de medidas es una medida, pero el producto de dos medidas no es una medida más.

A partir de los axiomas de Kolmogorov 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: por ejemplo, $P(\emptyset) = 0$;

- el rango de las probabilidades está acotado: $0 \leq P(A) \leq 1 \ \ \forall\, A \in \mathcal{A}$;
- condición de normalización: si $\Omega = \bigcup_{i=1}^{n} A_i$, con $A_i$ mutuamente excluyentes, entonces $\sum_{i=1}^{n} P(A_i) = 1$; el conjunto $\{A_i\}_{i=1}^{n}$ se dice *conjunto completo de eventos posibles excluyentes entre sí* y es ilustrado en la Figura 1-2;
- si $A$ es subconjunto de $B$, lo que escribiremos $A \subset B$, es decir que si $B$ se realiza, $A$ se realiza también (pero no necesariamente al revés), entonces $P(A) \leq P(B)$; es ilustrado en la Figura 1-2.


(a) (b)

**Figura 1-2:** Ilustración de: (a) conjunto completo de eventos posibles excluyentes entre sí; (b) inclusión entre eventos, donde $A$ está en gris oscuro mientras que $B$ está en gris (claro y oscuro).

La probabilidad $P(A \cap B)$ del evento $A \cap B$ se llama también *probabilidad conjunta* de $A$ y $B$. Se demuestra que

- $P(A \cap B)$ está acotada: $0 \leq P(A \cap B) \leq \min\{P(A), P(B)\}$ (viene de $A \cap B \subset A$ y $A \cap B \subset B$);
- si $A$ y $B$ son mutuamente excluyentes, entonces $P(A \cap B) = 0$ (viene de $A \cap B = \emptyset$);
- si $\{B_j\}_{j=1}^{m}$ es un conjunto completo de eventos posibles excluyentes entre sí, entonces $\sum_{j=1}^{m} P(A \cap B_j) = P(A)$ (viene de $\{A \cap B_j\}_j$ mutuamente excluyentes y $\bigcup_j (A \cap B_j) = A \cap \left(\bigcup_j B_j\right) = A \cap \Omega = A$).

En el caso de eventos no necesariamente mutuamente excluyentes, se prueba que la *ley de composición* o *fórmula de inclusión-exclusión* es

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B),$$

y que para $n$ eventos resulta

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P\left(A_i\right).$$

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

Se prueba también que si $\{A_i\}_{i=1}^{+\infty}$ es una secuencia *creciente* de eventos, *i. e.,* $\forall\, i \geq 1, \quad A_i \subset A_{i+1}$, entonces

$$P\left(\bigcup_{i=1}^{+\infty} A_i\right) = \lim_{i \to +\infty} P(A_i).$$

Por otro lado, si $\{A_i\}_{i=1}^{+\infty}$ es una secuencia *decreciente* de eventos, *i. e.,* $\forall\, i \geq 1, \quad A_{i+1} \subset A_i$, entonces

$$P\left(\bigcap_{i=1}^{+\infty} A_i\right) = \lim_{i \to +\infty} P(A_i).$$

Podemos preguntarnos cuál es la probabilidad de un evento $A$, si sabemos que se da cierto evento $B$. Por ejemplo, para un dado de 6 caras equilibrado, cuál es la probabilidad de tener un número par sabiendo que tenemos un número menor o igual a 3. La respuesta se encuentra en la noción de *probabilidad condicional* (Hausdorff, 1901; Jeffrey, 1948, 1973; Brémaud, 1988; Mandel & Wolf, 1995; Jacob & Protters, 2003; Shafer & Vovk, 2006):

**Definición 1-3** (Probabilidad condicional)**.** *La probabilidad condicional de $A$ dado $B$, denotado $P(A|B)$, se define como la razón entre la probabilidad del evento conjunto y la probabilidad de que se dé $B$ (cuando éste es un evento de probabilidad no nula):*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

En el ejemplo precedente, la probabilidad condicional va a ser $P(A|B) = \frac{P(A \cup B)}{P(B)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$.

Claramente del hecho de que $P$ es una medida de probabilidad se tiene

$$P(A|B) \geq 0.$$

Luego, de $A \cap B \subseteq B$ resulta $P(A \cap B) \leq P(B)$; es decir,

$$P(A|B) \leq 1.$$

Además, $P(\Omega \cap B) = P(B)$ dando

$$P(\Omega|B) = 1.$$

Para cualquier conjunto $\{A_i\}$ de eventos mutuamente excluyentes, los $(A_i \cap B)$ son también mutuamente excluyentes, así que $P\left(\left(\bigcup_i A_i\right) \bigcap B\right) = P\left(\bigcup_i \left(A_i \bigcap B\right)\right) = \sum_i P\left(A_i \bigcap B\right)$ dando

$$P\left(\bigcup_i A_i \middle| B\right) = \sum_i P\left(A_i | B\right).$$

Dicho de otra manera, $P(A|B)$ es una medida de probabilidad [4]. Diversas situaciones de probabilidades condicionales son ilustradas en la Fig. 1-3.

Algunas propiedades interesantes son las siguientes:

- $P(A \cap B|B) = P(A|B)$ $\big($viene de $P(A \cap B \cap B) = P(A \cap B)\big)$;
- si $A$ y $B$ son mutuamente excluyentes, obviamente $P(A|B) = 0$;

[4] Se puede definir un espacio de probabilidad $(\Omega_B, \mathcal{A}_B, P_B)$ donde $P_B(A) \equiv P(A|B)$. La notación $P_B(A)$ suele utilizarse en la literatura pero no la usaremos en esta obra para no confundirla con la medida de probabilidad de una variable aleatoria que definiremos en la sección siguiente.

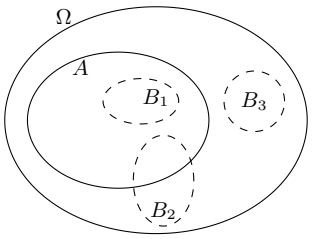



**Figura 1-3:** Ilustración de la probabilidad condicional con $A$ interior de la curva en línea llena y unos $B_i$ interiores de las curvas en líneas discontinuas. Se tiene $\omega \in B_1 \Rightarrow \omega \in A$ así que $P(A|B_1) = 1$; por otro lado, $\omega \in B_3 \Rightarrow \omega \notin A$ así que $P(A|B_3) = 0$. Entre estas situaciones extremas, si $P(\bar{A} \cap B_2) \neq 0$ y $P(A \cap B_2) \neq 0$ tenemos $0 < P(A|B_2) < 1$ (se puede tomar el ejemplo de probabilidad de un evento igual a su superficia sobre la de $\Omega$ para ver estas propiedades en este caso particular).

- si $B \subseteq C$, entonces $P(A|B \cap C) = P(A|B)$ $\big($viene de $P(A|B \cap C) = \frac{P(A \cap B \cap C)}{P(B \cap C)} = \frac{P(A \cap B)}{P(B)}$, pues $B \cap C = B\big)$;

- condición de normalización: si $\{A_i\}_{i=1}^n$ es un conjunto completo de resultados posibles mutuamente excluyentes, entonces $\sum_{i=1}^n P(A_i|B) = 1$;

- relación entre probabilidades condicionales inversas: $P(B|A) = \dfrac{P(B)}{P(A)} P(A|B)$, de donde $P(A|B)$ y $P(B|A)$ coinciden sólo cuando $A$ y $B$ tienen la misma probabilidad;

- *fórmula de probabilidad total*: si $\{B_j\}$ es un conjunto completo de eventos no nulos mutuamente excluyentes, entonces

$$P(A) = \sum_j P(A|B_j) P(B_j)$$

  $\big($viene de $A = A \cap \left(\bigcup_j B_j\right) = \bigcup_j (A \cap B_j)$ donde los $A \cap B_j$ son mutuamente excluyentes, y $P(A \cap B_j) = P(A|B_j)P(B_j)\big)$ ;

- *fórmula de Bayes*: si $\{B_j\}$ es un conjunto completo de eventos no nulos mutuamente excluyentes, entonces

$$P(B_i|A) = \frac{P(A \cap B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)};$$

  (ver (Brémaud, 1988; Jacob & Protters, 2003; Bayes, 1763; Barnard, 1958) [5] ).

Veamos ahora la noción de independencia entre dos eventos. Por ejemplo, si se tiran dos dados sobre sendas mesas, no hay ninguna razón para que la muestra de uno "influya" la del otro. Dicho de otra manera, dos eventos son independientes si el conocimiento de uno no lleva ninguna "información" sobre el otro (Brémaud, 1988; Mandel & Wolf, 1995; Hausdorff, 1901; Jacob & Protters, 2003; Borel, 1909):

---

[5] La obra del matemático y religioso inglés Thomas Bayes fue de hecho recopilada y publicada después de su muerte por Richard Price.

**Definición 1-4** (Independencia estadística)**.** *Dos eventos $A$ y $B$ se dicen estadísticamente independientes si la probabilidad condicional de $A$ dado $B$ es igual a la probabilidad incondicional de $A$:*

$$P(A|B) = P(A).$$

*Es equivalente al hecho de que la probabilidad conjunta se factoriza:*

$$P(A \cap B) = P(A)P(B).$$

Por inducción, la condición necesaria y suficiente para que $n$ eventos $A_1, \ldots, A_n$ sean mutuamente estadísticamente independientes es que la probabilidad conjunta se factorice como

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i).$$

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

Es importante notar que la independencia mutua no es equivalente a la independencia por pares de eventos, como lo ilustra el ejemplo siguiente.

**Ejemplo 1-1** (Independencia mutua vs por pares)**.** *Tiramos 2 dados independientemente y consideramos los eventos: $A_i, i = 1, 2$ "el dado $i$ es par" y $A_3$ "la suma de ambos dados es impar". Es claro que $A_1$ y $A_2$ son independientes y además para $i = 1$ o 2, $P(A_i \cap A_3) = \frac{1}{4} = P(A_i)P(A_3)$, mientras que $P(A_1 \cap A_2 \cap A_3) = 0 \neq \frac{1}{8}$: los eventos son independientes por pares, pero no son mutuamente independientes (Hogg et al., 2013).*

**Definición 1-5** (Independencia condicional)**.** *Dos eventos $A$ y $B$ se dicen estadísticamente independientes condicionalmente a un tercer evento $C$, si la probabilidad conjunta de $A$ y $B$ condicionalmente a $C$ es igual al producto de la probabilidad de $A$ condicionalmente a $C$ por la de $B$ condicionalmente a $C$:*

$$P(A \cap B|C) = P(A|C)P(B|C).$$

*Si $P(B|C) \neq 0$, es equivalente a $P(A|B \cap C) = P(A|C)$.*

Es importante notar que dos eventos pueden ser independientes, pero no serlo condicionalmente a un tercero, como lo ilustra el ejemplo siguiente.

**Ejemplo 1-2** (Independencia incondicional pero no condicional)**.** *Teniendo dos monedas bien equilibradas y tirándolas de manera independiente, consideramos los eventos $A$ "la primera faz es una cruz", $B$ "la segunda faz es una cara", $C$ "las faces son idénticas". Claramente $P(A \cap B) = \frac{1}{4} = P(A)P(B)$, mientras que $P(A \cap B|C) = 0 \neq P(A|C)P(B|C) = \frac{1}{16}$.*

Al revés, dos eventos pueden ser condicionalmente independientes a un tercero, pero ser dependientes.

**Ejemplo 1-3** (Independencia condicional pero no incondicional)**.** *Sea Alice tirando una moneda bien equilibrada y denotamos $A$ el evento "era una cruz". Claramente $P(A) = \frac{1}{2}$. Suponemos que Alice transmite el resultado a Bob a través de un intermediario Charlie con una probabilidad $\varepsilon$ de mentir a Charlie, y llamamos $C$ el evento "Alice dijo a Charlie que era una cruz". Tenemos que $P(C) = P(C|A)P(A) + P(C|\overline{A})P(\overline{A}) = (1-\varepsilon)\frac{1}{2} + \varepsilon\frac{1}{2} = \frac{1}{2}$.*

*Suponemos ahora que Charlie transmite a Bob lo que le dijo Alice, con una probabilidad $\vartheta$ de mentir (independientemente de Alice) y llamamos $B$ el evento "Charlie dijo a Bob que era una cruz". Es de nuevo sencillo ver que $P(B) = \frac{1}{2}$. Ahora, $P(A \cap B|C) = \frac{P(A \cap B \cap C)}{P(C)} = 2\,P(A \cap B \cap C)$. El evento $A \cap B \cap C$ es era una cruz y Alice no mintó y Charlie tampoco, es decir, por la independencia: $P(A \cap B|C) = (1-\varepsilon)(1-\vartheta)$. Inmediatamente $P(B|C) = 1 - \vartheta$ y $P(A|C) = 2\,P(A \cap C)$ siendo $A \cap C$ el evento "era una cruz y Alice no mintió", i.e. $P(A|C) = 1 - \varepsilon$. En conclusión, $P(A \cap B|C) = P(A|C)P(B|C)$: $A$ y $B$ son independientes condicionalmente a $C$. Ahora, $P(A \cap B) = P(A \cap B \cap C) + P(A \cap B \cap \overline{C}) = \frac{1}{2}(1-\varepsilon)(1-\vartheta) + \frac{1}{2}\varepsilon\vartheta \neq \frac{1}{4} = P(A)P(B)$ en general: $A$ y $B$ no resultan independientes. Este ejemplo es una instancia de lo que se llama un proceso de Markov, que vamos a ver un poco más en el cápitulo 2.*

## 1.3 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, continua o mixta.

Formalmente, la noción de variable aleatoria se apoya sobre la noción de función medible. Por esta formalización, vamos a necesitar definir la integración de manera general, más allá del enfoque de Riemann ("a la Lebesgue"), así como la noción de derivada de una medida con respecto a otra para definir densidades de probabilidad, en analogía a la densidad de masa en mecánica por ejemplo (Lebesgue, 1904, 1918; Kolmogorov & Fomin, 1961; Athreya & Lahiri, 2006; Bogachev, 2007a; Cohn, 2013).

### 1.3.1 Consideraciones preliminaries: Teorías de la medida y de la integración.

La primera noción que subyace a la definición formal de variable aleatoria es la de función medible:

**Definición 1-6** (Función medible)**.** *Sean $(\Omega, \mathcal{A})$ y $(\Upsilon, \mathcal{B})$ dos espacios medibles. Una función $f : \Omega \mapsto \Upsilon$ se dice $(\mathcal{A}, \mathcal{B})$-medible si*

$$\forall\, B \in \mathcal{B}, \quad A \equiv f^{-1}(B) = \{\omega \in A :\, f(\omega) \in B\} \,\in\, \mathcal{A}.$$

*Dicho de otra manera, la pre-imágen de un elemento dado de $\mathcal{B}$ (elemento medible) pertenece a $\mathcal{A}$ (elemento medible). Por abuso de escritura, se dice más simplemente que $f : (\Omega, \mathcal{A}) \mapsto (\Upsilon, \mathcal{B})$ es medible.*

Además, a partir de un espacio de medida y una función $f$ medible, se puede definir una medida imágen sobre el espacio de llegada (Athreya & Lahiri, 2006; Bogachev, 2007a; Cohn, 2013):

**Teorema 1-1** (Teorema de la medida imágen)**.** *Sean $(\Omega, \mathcal{A}, \mu)$ un espacio de medida, $(\Upsilon, \mathcal{B})$ un espacio medible y $f : (\Omega, \mathcal{A}) \mapsto (\Upsilon, \mathcal{B})$ una función medible. Sea $\mu_f$ tal que*

$$\forall\, B \in \mathcal{B}, \quad \mu_f(B) = \mu\left(f^{-1}(B)\right).$$

*Entonces, $\mu_f$ es una medida sobre el espacio medible $(\Upsilon, \mathcal{B})$, i. e., $(\Upsilon, \mathcal{B}, \mu_f)$ define un espacio de medida. Además, $\mu(\Omega) = \mu_f(\Upsilon)$ (posiblemente infinitas). Se dice que $\mu_f$ es la medida imágen de $\mu$ por $f$.*

*Demostración.* Por definición, claramente $\mu_f \geq 0$. Además, obviosamente $f^{-1}(\emptyset) = \emptyset$ dando $\mu_f(\emptyset) = \mu(\emptyset) = 0$. Luego, si para un conjunto numerable $\{B_j\}$ de elementos de $\mathcal{B}$ disjuntos entre sí, las pre-imágenes de los $B_j$ también son disjuntos entre sí (para $k \neq j$ no se puede tener $\omega \in f^{-1}(B_j) \cap f^{-1}(B_k)$ sino $\omega$ tendría dos imágenes distintas por $f$). Entonces $f^{-1}\left(\bigcup_j B_j\right) = \bigcup_j f^{-1}(B_j)$. Esto implica que $\mu_f\left(\bigcup_j B_j\right) = \mu\left(f^{-1}\left(\bigcup_j B_j\right)\right) = \mu\left(\bigcup_j f^{-1}(B_j)\right) = \sum_j \mu\left(f^{-1}(B_j)\right) = \sum_j \mu_f(B_j)$. Finalmente, necesariamente $f^{-1}(\Upsilon) = \Omega$ (obviamente $f(\Omega) \subseteq \Upsilon$) lo que cierra la prueba [6] □

A continuación, necesitaremos tratar de funciones medibles teniendo una propiedad (P) salvo sobre un conjunto de medida $\mu$ igual a cero. Más generalmente viene acá la noción de propiedad *casi siempre*:

**Definición 1-7** (Propiedad (e igualdad) $\mu$-casi siempre)**.** *Una funcion medible $f$ se dice tener una propiedad (P) dada $\mu$-casi siempre, si y solamente si la tiene excepto sobre un conjunto de medida nula,*

$$\mu\left(\{\omega : f(\omega) \text{ no satisface (P) }\}\right) = 0.$$

*Por ejemplo, dos funciones medibles $f_1$ y $f_2$ $(\Omega, \mathcal{A}, \mu) \to (\Upsilon, \mathcal{B})$ son iguales $\mu$-casi siempre,*

$$f_1 = f_2 \quad (\mu\text{-c.s.})$$

*si y solamente si son iguales excepto sobre un conjunto de medida nula,*

$$\mu\left(\{\omega : f_1(\omega) \neq f_2(\omega)\}\right) = 0.$$

Un espacio que juega un rol particular es $\mathbb{R}^d$, al cual se puede asociar una $\sigma$-álgebra particular conocida como *$\sigma$-álgebra de Borel* (Athreya & Lahiri, 2006; Bogachev, 2007a, 2007b; Cohn, 2013):

**Definición 1-8** ($\mathbb{R}^d$ y Borelianos)**.** *Para cualquier $d \geq 1$ entero, llamamos Borelianos $\mathcal{B}(\mathbb{R}^d)$ de $\mathbb{R}^d$ a la $\sigma$-álgebra más pequeña generada por los productos cartesianos $\overset{d}{\underset{i=1}{\times}}(-\infty\,;\, b_i]$ (similarmente, por los abiertos de $\mathbb{R}^d$, o también para los productos cartesianos de intervalos $\overset{d}{\underset{i=1}{\times}}(a_i\,;\, b_i]$), i. e., uniones numerables, intersecciones numerables, complementos de estos intervalos. $\mathcal{B}(\mathbb{R}^d)$ es también llamado $\sigma$-álgebra de Borel de $\mathbb{R}^d$.*

Se necesita ahora definir la noción de integración de una función medible con respecto a una medida:

---

[6] De hecho, se puede probar sencillamente que la pre-imágen de una unión numerable (disjuntos o no) es la unión de las pre-imágenes; lo mismo occure para la intersección y además la pre-imágen del complemento es el complemento de la pre-imágen. Esto se conoce como *leyes de de Morgan* (Athreya & Lahiri, 2006; Cohn, 2013; Hogg et al., 2013) (ver también (Kolmogorov & Fomin, 1957, Cap. 1) y (Kolmogorov & Fomin, 1961, Caps. 5 & 6)).

**Definición 1-9** (Medida e integración)**.** *Para una medida cualquiera, sobre un espacio de medida* $(\Omega, \mathcal{A}, \mu)$*, se define la integración a partir de*

$$\forall\, A \in \mathcal{A}, \quad \int_A d\mu(\omega) = \int_\Omega \mathbb{1}_A(\omega)\, d\mu(\omega) = \mu(A),$$

*donde*

$$\mathbb{1}_A(\omega) = \begin{cases} 1 & \textit{si} \quad \omega \in A \\ 0 & \textit{si} \quad \omega \notin A \end{cases}$$

*es la función indicadora del conjunto* $A$*.* $d\mu(\omega)$ *se escribe a veces también* $\mu(d\omega)$*, medida de un "infinitésimo". Claramente, por propiedades de una medida, para* $A_i, A_j$ *disjuntos* $\mathbb{1}_{A_i} + \mathbb{1}_{A_j} = \mathbb{1}_{A_i \cup A_j}$*, dando* $\int_\Omega \left(\mathbb{1}_{A_i} + \mathbb{1}_{A_j}\right) d\mu(\omega) = \mu(A_i \cup A_j) = \mu(A_i) + \mu(A_j) = \int_\Omega \mathbb{1}_{A_i} d\mu(\omega) + \int_\Omega \mathbb{1}_{A_j} d\mu(\omega)$ *y entonces, sin perdida de generalidad para un conjunto* $\{A_j\}$ *numerable y* $\{a_j\}$ *reales no negativos, la integral de la función escalonada* $\sum_j a_j \mathbb{1}_{A_j}$ *es dada por*

$$\int_\Omega \left( \sum_j a_j \mathbb{1}_{A_j}(\omega) \right) d\mu(\omega) = \sum_j a_j \int_\Omega \mathbb{1}_{A_j}(\omega)\, d\mu(\omega).$$

*Para los* $A_i$ *disjuntos es la consecuencia directa de la propiedad precedente, y si* $A_i, A_j$ *no son disjuntos. De hecho, sufice considerar* $A_i \setminus A_j,\ A_j \setminus A_i, A_i \cap A_j$ *con* $A \setminus B = \{\omega : \omega \in A \textit{ y } \omega \notin B\}$ *y respectivamente los coefficientes* $a_i,\ a_j,\ a_i + a_j$ *para volver al caso de conjuntos disjuntos.*

Antes de definir la integración de una función real, medible, cualquiera, el último paso que falta es el siguiente:

**Teorema 1-2** (Función medible como límite)**.** *Sea* $g : (\Omega, \mathcal{A}) \mapsto (\mathbb{R}, \mathcal{B}(\mathbb{R}))$*, no negativa y medible. Existe una sucesión* $\{g_n\}_{n\in\mathbb{N}}$ *creciente de funciones escalonadas que converge simplemente (punto a punto) hacia* $g$*.*

*Demostración.* La sucesión $g_n = \sum_{k=0}^{n2^n-1} \frac{k}{2^n}\, \mathbb{1}_{g^{-1}\left(\left[\frac{k}{2^n}\,;\,\frac{k+1}{2^n}\right)\right)} + n\, \mathbb{1}_{g^{-1}([n\,;+\infty))}$ es escalonada, creciente y converge hacia $g$ (notar que esta sucesión comparte la idea que subyace a la integración de Riemann). □

De este resultado, se puede generalizar la noción de integración de una función real:

**Definición 1-10** (Integración de una función real)**.** *Sea* $g : (\Omega, \mathcal{A}) \mapsto (\mathbb{R}, \mathcal{B}(\mathbb{R}))$*, no negativa y medible, y* $\{g_n\}_{n\in\mathbb{N}}$ *una sucesión creciente de funciones escalonadas que converge simplemente hacia* $g$*. Por definición,*

$$\int_\Omega g(\omega)\, d\mu(\omega) = \lim_{n\to\infty} \int_\Omega g_n(\omega)\, d\mu(\omega).$$

*Notar de que el límite puede ser infinita.*

*Sea ahora* $g : (\Omega, \mathcal{A}) \mapsto (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *medible cualquiera. Se verifica sencillamente que también* $|g|$ *(valor absoluto) es medible y, por definición,* $g$ *se dice* $\mu$*-integrable si la integral de* $|g|$ *es finita,*

$$g \textit{ es } \mu\textit{-integrable} \quad \Leftrightarrow \quad \int_\Omega |g(\omega)|\, d\mu(\omega) < +\infty.$$

*Además, se escribe* $g = g_+ + g_-$ *con* $g_+ = \max(g, 0)$ *y* $g_- = \min(g, 0)$*. Es sencillo ver de que si* $g$ *es medible,* $g_+$ *y* $g_-$ *son medibles. Si* $g$ *es* $\mu$*-integrable, necesariamente* $g_+$ *y* $g_-$ *son* $\mu$*-integrables, y, por definición*

$$\int_\Omega g(\omega)\, d\mu(\omega) = \int_\Omega g_+(\omega)\, d\mu(\omega) - \int_\Omega (-g_-(\omega))\ d\mu(\omega).$$

A continuación, damos unos teoremas que serán muy útiles más adelante, sin detallar las pruebas. Por esto, el lector se puede referir a (Lieb & Loss, 2001; Athreya & Lahiri, 2006; Bogachev, 2007a; Cohn, 2013).

**Teorema 1-3** (Teorema de convergencia monótona)**.** *Sea* $\{f_n\}_{n\in\mathbb{N}}$ *una sucesión creciente de funciones medibles sobre* $(\Omega, \mathcal{A}, \mu)$*, positivas, convergiendo simplemente hacia una función* $f$ *medible. Entonces*

$$\lim_{n\to+\infty} \int_\Omega f_n(\omega) d\mu(\omega) = \int_\Omega f(\omega) d\mu(\omega).$$

De hecho se prueba este teorema a partir de la definición de integración. Este teorema da una condición simple permitiendo intercambiar integración y límite.

**Corolario 1-1.** *Sea* $\{f_n\}_{n\in\mathbb{N}}$ *una sucesión de funciones medibles sobre* $(\Omega, \mathcal{A}, \mu)$*, positivas, tal que la serie* $\sum_n f_n$ *converge simplemente hacia una función* $f$*,* $\mu$*-integrable. Entonces*

$$\int_\Omega \sum_{n\in\mathbb{N}} f_n(\omega) d\mu(\omega) = \int_\Omega f(\omega) d\mu(\omega).$$

Es una consecuencia del teorema de convergencia monótona, considerando la sucesión creciente $\{\sum_{k=0}^n f_k\}_{n\in\mathbb{N}}$.

**Teorema 1-4** (Teorema de convergencia dominada)**.** *Sea* $\{f_n\}_{n\in\mathbb{N}}$ *una sucesión creciente de funciones medibles sobre* $(\Omega, \mathcal{A}, \mu)$ *convergiendo simplemente hacia una función* $f$*, medible. Suponemos que existe una función* $\mu$*-integrable* $g$ *que domina la sucesión, i. e.,* $\forall\, \omega \in \Omega, \quad |f_n(\omega)| \leq g(\omega)$*. Entonces*

$$\lim_{n\to+\infty} \int_\Omega f_n(\omega)\, d\mu(\omega) = \int_\Omega f(\omega)\, d\mu(\omega) \leq \int_\Omega g(\omega)\, d\mu(\omega).$$

Este teorema da una condición suficiente muy útil y muy usada para asegurarse de que se puede intercambiar límite e integración.

El último teorema que vamos a necesitar permite intercambiar dos integraciones. Antes, necesitamos definir la noción de espacio medible producto y medida producto.

**Definición 1-11** (Espacio medible producto, medida producto)**.** *Sean dos espacios de medida* $(\Omega_1, \mathcal{A}_1, \mu_1)$ *y* $(\Omega_2, \mathcal{A}_2, \mu_2)$*. Llamamos espacio medible producto* $(\Omega, \mathcal{A})$ *al espacio del producto cartesiano* $\Omega = \Omega_1 \times \Omega_2$ *con la* $\sigma$*-álgebra* $\mathcal{A}$ *generada por los productos cartesianos* $A_1 \times A_2$ *donde* $A_i \in \mathcal{A}_i, i = 1, 2$*. Además, llamamos medida producto* $\mu$ *definida sobre* $\mathcal{A}$ *a la medida tal que* $\forall\, (A_1, A_2) \in \mathcal{A}_1 \times \mathcal{A}_2, \quad \mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$*.*

**Teorema 1-5** (Teorema de Fubini)**.** *Sea* $(\Omega, \mathcal{A}, \mu)$ *espacio de medida producto de* $(\Omega_1, \mathcal{A}_1, \mu_1)$ *y* $(\Omega_2, \mathcal{A}_2, \mu_2)$ *donde* $\mu$ *es la medida producto. Sea* $f$ *una función integrable sobre* $(\Omega, \mathcal{A}, \mu)$ *entonces*

- $\omega_1 \mapsto f(\omega_1, \omega_2)$ *es* $\mu_1$*-integrable (*$\mu_2$*-c.s.) y* $\omega_2 \mapsto f(\omega_1, \omega_2)$ *es* $\mu_2$*-integrable (*$\mu_1$*-c.s.),*
- $\omega_1 \mapsto \int_{\Omega_2} f(\omega_1, \omega_2)\, d\mu_2(\omega_2)$ *es* $\mu_1$*-integrable y* $\omega_2 \mapsto \int_{\Omega_1} f(\omega_1, \omega_2)\, d\mu_1(\omega_1)$ *es* $\mu_2$*-integrable.*

*Además,*

$$\int_{\Omega_1\times\Omega_2} f(\omega)\, d\mu(\omega) = \int_{\Omega_1} \left( \int_{\Omega_2} f(\omega)\, d\mu_2(\omega_2) \right) d\mu_1(\omega_1) = \int_{\Omega_2} \left( \int_{\Omega_1} f(\omega)\, d\mu_1(\omega_1) \right) d\mu_2(\omega_2)$$

**Teorema 1-6** (Integral a parámetro: continuidad y diferenciabilidad)**.** *Sea $I$ un compacto de $\mathbb{R}^d$ y $\{f(\cdot,t)\}_{t\in I}$ una familia de funciones medibles sobre $(\Omega,\mathcal{A},\mu)$, tal que $t\mapsto f(\omega,t)$ sea continua sobre $I$ ($\mu$-c.s.). Si existe una función $\mu$-integrable $g$ tal que*

$$\forall\, t\in I,\quad \forall\,\omega\in\Omega,\quad |f(\omega,t)|\le g(\omega),$$

*entonces $\omega\to f(\omega,t)$ es $\mu$-integrable y la función $t\mapsto\displaystyle\int_\Omega f(\omega,t)\,d\mu(\omega)$ es continua sobre $I$. Además, si $f(\omega,\cdot)$ es diferenciable sobre $I$ y si existe una función $\mu$-integrable $h$ tal que*

$$\forall\, t\in I,\quad \forall\,\omega\in\Omega,\quad |\nabla_t f(\omega,t)|\le h(\omega),$$

*donde $\nabla_t$ indica el gradiente, i. e., el vector de componentes $\frac{\partial}{\partial t_1},\dots,\frac{\partial}{\partial t_d}$, entonces la función $t\mapsto\displaystyle\int_\Omega f(\omega,t)\,d\mu(\omega)$ es diferenciable sobre $I$, y*

$$\nabla_t\int_\Omega f(\omega,t)\,d\mu(\omega)=\int_\Omega \nabla_t f(\omega,t)\,d\mu(\omega).$$

Básicamente, este teorema es consecuencia del teorema de convergencia dominada.

Seguimos esta sección con la noción de derivada de una medida con respecto a otra, dando una definición muy general de densidad:

**Definición 1-12** (Densidad de una medida)**.** *Sean $\mu$ y $\nu$ dos medidas cualesquiera sobre un espacio medible $(\Omega,\mathcal{A})$. Si existe una función real no negativa $p:\Omega\mapsto\mathbb{R}_+$ medible tal que*

$$\forall\, A\in\mathcal{A},\quad \nu(A)=\int_A p(\omega)d\mu(\omega),$$

*$p$ es llamada densidad de $\nu$ con respecto a $\mu$, denotada*

$$p=\frac{d\nu}{d\mu},$$

*también llamada derivada de Radon-Nikodým.*

Notar que dos funciones pueden cumplir esta definición, por ejemplo si son iguales $\mu$-casi siempre. De hecho, si dos funciones $p_1=p_2$ ($\mu$-c.s.), y $C$ es el conjunto donde no son iguales, notando $A\setminus C=\{\omega:\omega\in A$ y $\omega\notin C\}$, de $\displaystyle\int_A p_1(\omega)d\mu(\omega)=\int_{A\setminus C}p_1(\omega)d\mu(\omega)=\int_{A\setminus C}p_2(\omega)d\mu(\omega)=\int_A p_2(\omega)d\mu(\omega)$ se ve que dos funciones iguales casi siempre pueden ser densidad de una medida con respecto a una otra.

Es sencillo ver que si $\mu(A)=0$, necesariamente $\nu(A)=0$. De eso viene la noción de absoluta continuidad:

**Definición 1-13** (Absoluta continuidad)**.** *Sean $\mu$ y $\nu$ dos medidas sobre un espacio medible $(\Omega,\mathcal{A})$. Se dice que $\nu$ es absolutamente continua con respecto a $\mu$, denotado*

$$\nu\ll\mu,$$

*si $\forall\, A\in\mathcal{A},\quad \mu(A)=0\ \Rightarrow\ \nu(A)=0$.*

De hecho, se muestra la recíproca de la definición Def. 1-12 a través de lo que se conoce como teorema de Radon-Nikodým (Nikodym, 1930; Athreya & Lahiri, 2006; Bogachev, 2007a; Cohn, 2013):

**Teorema 1-7** (Radon-Nikodým)**.** *Sean dos medidas $\mu$ y $\nu$, entonces*

$$\nu \ll \mu \quad \Longleftrightarrow \quad \nu \text{ admite una densidad con respecto a } \mu.$$

*Además, esta densidad $\frac{d\nu}{d\mu}$ es única en el sentido de que si dos funciones cumplen la definición, son iguales $\mu$-casi siempre.*

En todo lo que sigue, hablaremos de "la" densidad de una medida, salvo si se necesita explícitamente tener en cuenta esta sutileza.

A continuación, dos lemas van a ser muy útiles especialmente en el Capítulo 2, tratando con dos (o más) medidas y densidades.

**Lema 1-1.** *Sean $\nu$ y $\mu$ dos medidas sobre $(\Omega, \mathcal{A})$ tales que $\nu \ll \mu$. Entonces, para cualquier función medible $f$,*

$$\int_\Omega f(\omega)\, \frac{d\nu}{d\mu}(\omega)\, d\mu(\omega) = \int_\Omega f(\omega)\, d\nu(\omega)$$

*Demostración.* Tomando $f = \mathbb{1}_A$, de la definición Def. 1-12 se obtiene

$$\int_\Omega \mathbb{1}_A(\omega)\, \frac{d\nu}{d\mu}(\omega)\, d\mu(\omega) = \int_A \frac{d\nu}{d\mu}(\omega)\, d\mu(\omega) = \nu(A) = \int_A d\nu(\omega)$$

Se cierra la prueba usande el teorema 1-2 y la definición 1-10, tratando $f$ con su parte positiva y negativa separadamente. □

**Lema 1-2.** *Sean $\nu$, $\mu$ y $\lambda$ tres medidas sobre $(\Omega, \mathcal{A})$ y suponemos $\nu \ll \lambda$ y $\lambda \ll \mu$. Entonces*

- $\nu \ll \mu$;
- *equivalentemente, el soporte (ensemble de puntos que no anula la función) de $\dfrac{d\nu}{d\mu}$ está incluído ($\mu$-casi siempre) en el soporte de $\dfrac{d\lambda}{d\mu}$;*
- $\dfrac{d\nu}{d\lambda}\dfrac{d\lambda}{d\mu} = \dfrac{d\nu}{d\mu}$ *($\mu$-c.s.).*

*Demostración.* El primer resultado viene de la definición de la absoluta continuidad $\mu(A) = 0 \quad \Rightarrow \quad \lambda(A) = 0 \quad \Rightarrow \quad \nu(A) = 0$. El segunndo resultado se obtiene escribiendo la medida en su forma integral. Adémas, por definición de la densidad, $\forall\, A \in \mathcal{A}, \quad \nu(A) = \displaystyle\int_A \frac{d\nu}{d\mu}(\omega)\, d\mu(\omega)$. Luego, aplicando el lema anterior a $f = \mathbb{1}_A \frac{d\nu}{d\mu}$ se obtiene que, también, $\nu(A) = \int_A \frac{d\nu}{d\lambda}(\omega)\, d\lambda(\omega) = \int_A \frac{d\nu}{d\lambda}(\omega)\, \frac{d\lambda}{d\mu}(\omega)\, d\mu(\omega)$, lo que cierra la prueba. □

Unas medidas que juegan un rol particular son las medidas de Lebesgue o medidas discretas.

**Definición 1-14** (Medida de Lebesgue)**.** *La medida de Lebesgue $\mu_L$ sobre $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ se define tal que para cualquier producto cartesiano de intervalos,*

$$\mu_L\left(\overset{d}{\underset{i=1}{\times}} (a_i\,;\, b_i)\right) = \prod_{i=1}^{d} (b_i - a_i).$$

Acá, notamos dos hechos interesantes:

- $\mu_L$ es $\sigma$-finita. Viene de $\mathbb{R}^d = \cup_{i=1}^d \cup_{j_i \in \mathbb{Z}} \times_{i=1}^d (j_i \,;\, j_i+1]$ conjuntamente a $\mu_L(\times_{i=1}^d (j_i \,;\, j_i+1]) = 1 < +\infty$.
- $\forall\, A \in \mathcal{A}, \quad \mu_L(A) = |A|$ donde $|\cdot|$ denota el volumen de un conjunto (puede ser infinito).
- Para una función $g$ suficientemente "suave", la integración con respecto a la medida de Lebesgue coincide naturalmente con la integración de Riemann.

La medida de Lebesgue es así natural para la integración. Luego, en lo que sigue, al mencionar igualdad $\mu_L$-casi siempre, diremos simplemente "casi siempre" $(c.s.)$, entendiendo que es con respecto a la medida de Lebesgue. De la misma manera, hablando de densidad, sin precisiones, se entenderá que s con respecto a $\mu_L$.

Al "contrario" de la medida de Lebesgue, medidas discretas son también particulares. La más "elemental" es conocida como *medida de Dirac*, dando lugar a medidas discretas:

**Definición 1-15** (Medida de Dirac y medida discreta)**.** *La medida de Dirac al punto $x_0$, denotada $\delta_{x_0}$, es tal que*

$$\forall\, B \in \mathcal{B}(\mathbb{R}^d), \quad \delta_{x_0}(B) = \mathbb{1}_B(x_0) = \begin{cases} 1 & \textit{si} \quad x_0 \in B \\ 0 & \textit{si} \quad x_0 \notin B \end{cases}.$$

*Dado un conjunto $\mathcal{X} = \{x_i\}_i$ discreto (finito o infinito numerable), llamaremos medida discreta a la medida definida por*

$$\mu_{\mathcal{X}} = \sum_i \delta_{x_i}$$

*(en general, son definidas como combinaciones lineales positivas, siendo éste un caso particular).*

Notar que,

- $\mu_{\mathcal{X}}$ es $\sigma$-finita (se muestra con el mismo enfoque que para la medida de Lebesgue).
- $\forall\, A \in \mathcal{A}, \quad \mu_{\mathcal{X}}(A) = |\mathcal{X} \cap A|$ donde $|\cdot|$ denota el cardinal de un conjunto, equivalente discreto del volumen (puede ser infinito también).
- Para una función $g$ medible,

  $$\int_{\mathbb{R}^d} g(x)\, d\delta_{x_k}(x) = g(x_k) \qquad \text{y} \qquad \int_{\mathbb{R}^d} g(x)\, d\mu_{\mathcal{X}}(x) = \sum_{x \in \mathcal{X}} g(x),$$

  luego la integración se vuelve una suma. Se prueba saliendo de $g$ de la forma $g = \mathbb{1}_C$ y del Teorema 1-2 conjuntamente con las definiciones Def. 1-10 y Def. 1-15.

Con esta serie de definiciones, tenemos todo lo necesario para introducir la definición de variables/vectores aleatorios reales y sus caracterizaciones.

### 1.3.2 Variables aleatorias y vectores aleatorios. Distribución de probabilidad.

Empezamos con la noción de variable aleatoria real, que queremos ver como el resultado de un experimento o de un evento dado (Athreya & Lahiri, 2006; Cohn, 2013; Brémaud, 1988):

**Definición 1-16** (Variable aleatoria real)**.** *Una variable aleatoria real es una función medible*

$$X : (\Omega, \mathcal{A}, P) \mapsto (\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$$

*donde la medida $P_X$ sobre $\mathcal{B}(\mathbb{R})$ es la medida imágen de $P$. $P_X$ es frecuentemente llamada distribución de probabilidad o ley de la variable aleatoria $X$. En lo que sigue, escribiremos los eventos*

$$(X \in B) \equiv X^{-1}(B) = \{\omega \in \Omega : \ X(\omega) \in B\},$$

*así que, por definición,*

$$P_X(B) = P(X \in B)$$

Para ilustrar esta definición, tomando el ejemplo de un dado, $\Omega$ es discreto y representa las caras, mientras de que los numeros serán la imágen de $\Omega$ por $X$ (ej. $X(\omega_j) = j, \quad j = 1, \ldots, 6$).

Fijense de que, por las propiedades de una medida sobre una $\sigma$-álgebra, para caracterizar completamente la distribución $P_X$ es suficiente conocerla sobre los intervalos de la forma $(-\infty\,;\, b]$. Esto da lugar a la definición de la función de repartición (Athreya & Lahiri, 2006; Cohn, 2013; Brémaud, 1988; Hogg et al., 2013):

**Definición 1-17** (Función de repartición)**.** *Por definición, la función de repartición $F_X$ de una variable aleatoria es definida por*

$$F_X(x) = P_X((-\infty\,;\, x]) = P(X \leq x).$$

*A veces, por abuso de terminologia, se denomina $F_X$ como ley de la variable aleatoria. Se encuentra también en la literatura la terminologia de función cumulativa (cdf por cumulative density function en inglés).*

Naturalmente, de las propiedades de una medida de probabilidad,

- $0 \leq F_X(x) \leq 1$;
- $\lim\limits_{x \to -\infty} F_X(x) = 0$ y $\lim\limits_{x \to +\infty} F_X(x) = 1$ (viene de $P_X(\emptyset) = 0$ y $P_X(\mathbb{R}) = 1$);
- $F_X$ es creciente $\big($viene de que $x_1 \leq x_2 \ \Leftrightarrow \ (-\infty\,;\, x_1] \subseteq (-\infty\,;\, x_2]\big)$;
- $F_X$ no es necesariamente continua (lo vamos a ver más adelante), pero en cada punto $x$ es continua a su derecha (ver punto anterior).

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 $X = \begin{bmatrix} X_1 & \cdots & X_d \end{bmatrix}^t$ donde $\cdot^t$ denota la transpuesta, caracterizado por $d$-uplas de variables aleatorias reales. Como en el caso univariado, se define este vector de la manera siguiente (Athreya & Lahiri, 2006; Cohn, 2013; Brémaud, 1988):

**Definición 1-18** (Vector aleatorio real)**.** *Un vector aleatorio real es una función medible*

$$X : (\Omega, \mathcal{A}, P) \mapsto (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), P_X).$$

*donde* $\mathcal{B}(\mathbb{R}^d)$ *son los borelianos de* $\mathbb{R}^d$*,* $\sigma$*-álgebra generada por los productos cartesianos* $(-\infty\,;\, b_1] \times \cdots \times (-\infty\,;\, b_d]$ *y donde la medida* $P_X$ *sobre* $\mathcal{B}(\mathbb{R}^d)$ *es la medida imágen de* $P$ *llamada distribución de probabilidad de la variable aleatoria (o vector aleatorio)* $X$*. Como en el caso escalar,*

$$(X \in B) \equiv X^{-1}(B) = \{\omega \in \Omega :\ X(\omega) \in B\} \qquad \textit{y} \qquad P_X(B) = P(X \in B).$$

Nota: a veces, tenemos que considerar el caso de matrices aleatorias, o funciones medibles matriz-valuadas. Siendo de que se puede poner en biyección, una matriz con un vector (por ejemplo poniendo cada columna "debajo" de su columna antecedante), no desarollaremos más este caso, a pesar de que a veces sea más conveniente trabajar con matrices en lugar de su forma en vector.

De las propiedades de una medida sobre una $\sigma$-álgebra, para caracterizar completamente la distribución $P_X$ de nuevo es suficiente conocerla sobre los elementos de la forma $\overset{d}{\underset{i=1}{\times}}(-\infty\,;\, b_i]$, *i. e.,* la función de repartición multivariada (Athreya & Lahiri, 2006; Cohn, 2013; Brémaud, 1988; Hogg et al., 2013):

**Definición 1-19** (Función de repartición multivariada)**.** *Por definición, la función de repartición* $F_X$ *de un vector aleatorio es definida en* $x = (x_1, \ldots, x_d)$ *por*

$$F_X(x) = P_X\left(\overset{d}{\underset{i=1}{\times}}(-\infty\,;\, x_i]\right) = P\left(\bigcap_{i=1}^{d}(X_i \leq x_i)\right).$$

*Por abuso de escritura, escribiremos en lo que sigue*

$$F_x(x) = P(X \leq x),$$

*subentiendo de que* $(X \leq x)$ *es el evento* $\bigcap_{i=1}^{d}(X_i \leq x_i)$*.*

De nuevo, de las propiedades de una medida de probabilidad,

- $0 \leq F_X(x) \leq 1$;
- $\lim_{\forall i, x_i \to -\infty} F_X(x) = 0$ y $\lim_{\forall i, x_i \to +\infty} F_X(x) = 1$;
- $F_X$ es creciente con respecto a cada variable $x_i$.

Al final, para un subconjunto $I_k = (i_1, \ldots, i_k)$ de $1 \leq k \leq d$ elementos de $\{1, \ldots, d\}^k$, $X_{I_k} = \begin{bmatrix} X_{i_1} & \cdots & X_{i_k} \end{bmatrix}^t$ es obviamente un vector aleatorio $k$-dimensional. Es entonces sencillo ver de que

$$F_{X_{I_k}}(x_{I_k}) = \lim_{\forall i \notin I_k, x_i \to +\infty} F_X(x).$$

(viene de que $\bigcap_{j=1}^{k}(X_{i_j} \leq x_{i_j}) = \left(\bigcap_{j=1}^{k}(X_{i_j} \leq x_{i_j})\right) \bigcap \left(\bigcap_{i \notin I_k}(X_i \in \mathbb{R})\right)$). Esta función es dicha *función de repartición marginale* de $F_X$.

Cerramos estas generalidades con el caso de variables independientes:

**Definición 1-20** (Independencia)**.** *Sean $d$ variables aleatorias $X_i$ y $X = \begin{bmatrix} X_1 & \cdots & X_d \end{bmatrix}^t$. Los $X_i$ son mutuamente independientes si y solamente si, para cualquier ensemble de conjuntos $B_i$, los eventos $(X_i \in \mathcal{B}_i)$ son mutuamente independientes, i. e.,*

$$P_X\left(\overset{d}{\underset{i=1}{\times}} B_i\right) = \prod_{i=1}^{d} P_{X_i}(B_i).$$

*Es equivalente a*

$$F_X(x) = \prod_{i=1}^{d} F_{X_i}(x_i).$$

*La ley del vector aleatorio se factoriza. Necesariamente, $\mathcal{X} = X(\Omega)$ es de la forma $\mathcal{X} = \times_i \mathcal{X}_i$ con $\mathcal{X}_i = X_i(\Omega)$, producto cartesiano.*

Es importante notar de que no es equivalente a tener la independencia por pares, como ilustrado en el fin de la seción precediente.

Más allá de este enfoque general, dos casos particulares de variables aleatorias son de interés: las variables discretas y las continuas. En el primer caso $X(\Omega)$ es discreto, finito o no. La meta de las susecciones siguientes es estudiar las particularidades de cada caso.

Par fijar unas notaciones, en todo lo que sigue, escribiremos

$$\mathcal{X} = X(\Omega)$$

conjunto de llegada de $X$, o conjunto de valores que puede tomar la variable aleatoria. A veces, por razones de simplificaciones, se considera $\mathcal{X}$ como siendo el espacio muestral y se olvida de que $X$ sea una función medible entre espacios de probabilidades, *i. e.,* se trabaja en $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), P_X)$ como en el espacio pre-imágen.

### 1.3.3 Variable aleatoria discreta

**Definición 1-21** (Variable aleatoria discreta)**.** *Una variable aleatoria es dicha discreta cuando $\mathcal{X} = X(\Omega)$ es discreto, finito o infinito numerable. En lo que sigue, denotaremos por $|\mathcal{X}|$ el cardinal de $\mathcal{X}$, posiblemente infinito. En otras palabras, los posibles valores de una variable aleatoria discreta $X$ consisten en un conjunto contable (finito o infinito numerable) de números reales, $\mathcal{X} = \{x_j\}$ y se puede escribir $X$ como una variable escalonadas,*

$$X = \sum_j x_j \mathbb{1}_{A_j} \quad \textit{con} \quad A_j = X^{-1}(\{x_j\}).$$

(ver ej. (Athreya & Lahiri, 2006; Hogg et al., 2013)). Fijense de que $\Omega$ no es necesariamente discreto. Por ejemplo, si $\omega$ es la posición de un punto sobre una linea, y $X(\omega) = 0$ si $\omega$ es a la izquierda de un umbral, y $X(\omega) = 1$ si $\omega$ es a su derecha, $\mathcal{X} = \{0\,;\,1\}$ mientras de que $\Omega$ no es discreto.

En el caso de una variable aleatoria discreta $X$, las probabilidades $P_X(\{x_j\}) = P(X = x_j)$, $x_j \in \mathcal{X}$ caracterisan completamente esta variable aleatoria (Athreya & Lahiri, 2006; Hogg et al., 2013):

**Definición 1-22** (Función de masa de probabilidad)**.** *Por definición, la función de masa de probabilidad de $X$, variable aleatoria discreta tomando sus valores sobre $\mathcal{X}$ es dada por*

$$p_X(x) \equiv P(X = x) = P_X(\{x\}) \quad x \in \mathcal{X}.$$

*Por abuso de denominación, llamaremos en este libro $p_X$ distribución de probabilidad. Además, usaremos también la notación*

$$p_X = \begin{bmatrix} \cdots & p_X(x_j) & \cdots \end{bmatrix}^t$$

*dicho vector de probabilidad, de tamaño $|\mathcal{X}|$, posiblemente infinito.*

Fijense de que, $P_X$ siendo una medida de probabilidad, $p_X \geq 0$ y es obviamente normalizada en el sentido de que

$$\sum_{x_j \in \mathcal{X}} p_X(x_j) = 1.$$

En la Fig. 1-4-(a) se muestra una representación gráfica de una distribución de probabilidad discreta. En particular,

$$\forall\, B \in \mathcal{B}(\mathbb{R}), \quad P_X(B) = \sum_{x \in \mathcal{X} \cap B} p_X(x) = \int_B dP_X(x),$$

lo que da, tratando de la función de repartición,

$$F_X(x) = \sum_{x_j \leq x} p_X(x_j).$$

De esta forma, se justifica la denominación *cumulativa* para $F_X$. También, se puede ver inmediato de que $F_X$ es una función discontinua, con saltos finitos (en $x_j$, salto de altura $p_X(x_j)$). Esto es ilustrado figura Fig. 1-4-(b).

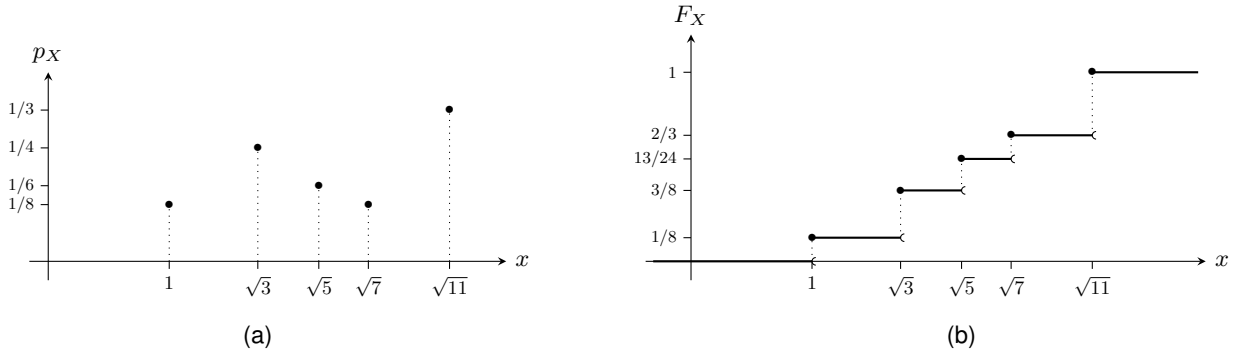



**Figura 1-4:** Ilustración de una distribución de probabilidad discreta (a), y la función de repartición asociada (b), con $\mathcal{X} = \{\, 1\,,\, \sqrt{3}\,,\, \sqrt{5}\,,\, \sqrt{7}\,,\, \sqrt{11}\,\}$ y $p_X = \begin{bmatrix} \frac{1}{8} & \frac{1}{4} & \frac{1}{6} & \frac{1}{8} & \frac{1}{3} \end{bmatrix}^t$.

Un caso especial se tiene cuando un valor $x_k$ es cierto o seguro, y no ocurre ninguno de los otros valores $x_j$ $(j \neq k)$. La forma de la distribución es: $p_X(x) = 1$ si $x = x_k$ y cero si no, o el vector de probabilidad se escribirá $p_X = \mathbb{1}_k$ donde $\mathbb{1}_k$ denotará es el vector (posiblemente de dimensión infinita) de componentes $k$-esima igual a 1, las otras siendo nulas, lo que se denota tambié con el *símbolo de Kronecker* $\delta_{jk} = 1$ si

$j = k$ y cero sino. En este libro, evitaremos usar este símbolo para no confundirlo con la medida de Dirac. Sin embargo, aparece de que $P_X$ es precisamente la medida de Dirac en $x_k$ (ver definición Def. 1-15).

Otra situación particular es la de *equiprobabilidad* o *distribución uniforme* cuando $|\mathcal{X}| = \alpha < +\infty$. La forma de la distribución es: $p_X(x_j) = \frac{1}{\alpha} \quad \forall j = 1, \ldots, \alpha$, *i. e.,* $p_X = \begin{bmatrix} \frac{1}{\alpha} & \cdots & \frac{1}{\alpha} \end{bmatrix}^t$ o, en termino de medida, $P_X = \frac{1}{\alpha} \sum_{j=1}^{\alpha} \delta_{x_j}$. La función de repartición resulta una función escalonada, con saltos de altura $\frac{1}{\alpha}$ en cada $x_j, \quad 1 \leq j \leq \alpha$.

De manera general, la medida de probabilidad de una variable discreta se escribe como combinación convexa de medidas de Dirac,

$$P_X = \sum_j p_j \delta_{x_j}, \qquad p_j = P(X = x_j) \geq 0, \quad \sum_j p_j = 1,$$

*i. e.,* como una medida. . . discreta.

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

**Definición 1-23** (Mayorización)**.** *Un vector de probabilidad (distribución) $p$ mayorizado por un vector de probabilidad (distribución) $q$, notado $p \prec q$, se define como:*

$$p \prec q \qquad \textit{ssi} \qquad \sum_{i=1}^{k} p_i^{\downarrow} \leq \sum_{i=1}^{k} q_i^{\downarrow}, \quad 1 \leq k < \alpha \qquad \textit{y} \qquad \sum_{i=1}^{\alpha} p_i^{\downarrow} = \sum_{i=1}^{\alpha} q_i^{\downarrow}$$

*(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 (sería equivalente a añadir estados fictivos de probabilidad nula).*

Por ejemplo, $\begin{bmatrix} 0{,}40 & 0{,}30 & 0{,}20 & 0{,}10 \end{bmatrix}^t \prec \begin{bmatrix} 0{,}50 & 0{,}30 & 0{,}15 & 0{,}05 \end{bmatrix}^t$ (ver figura Fig. 1-5-(a)).

Es importante resaltar que la mayorización provee un *orden parcial* (no total) entre distribuciones, existiendo pares de distribuciones tales que ninguna mayoriza a la otra. Por ejemplo, $\begin{bmatrix} 0{,}50 & 0{,}40 & 0{,}07 & 0{,}03 \end{bmatrix}^t$ y $\begin{bmatrix} 0{,}60 & 0{,}20 & 0{,}17 & ,03 \end{bmatrix}^t$ no se comparan por mayorización (ver figura Fig. 1-5-(b)).

Es interesante notar que la siguiente propiedad es válida para toda distribución $p$ de tamaño $\alpha$ (Marshall, Olkin & Arnold, 2011, p. 9, (6)-(8)):

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

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.

La relación de mayorización es ilustrada en la figura Fig. 1-5, donde se representa las sumas parciales

en función de $k$, llamadas *curvas de Lorentz* [7] (Marshall et al., 2011; Lorenz, 1905). Graficamente, $p \prec q$ es equivalente a tener la curva de Lorenz asociada a $p$ debajo de la asociada a $q$.

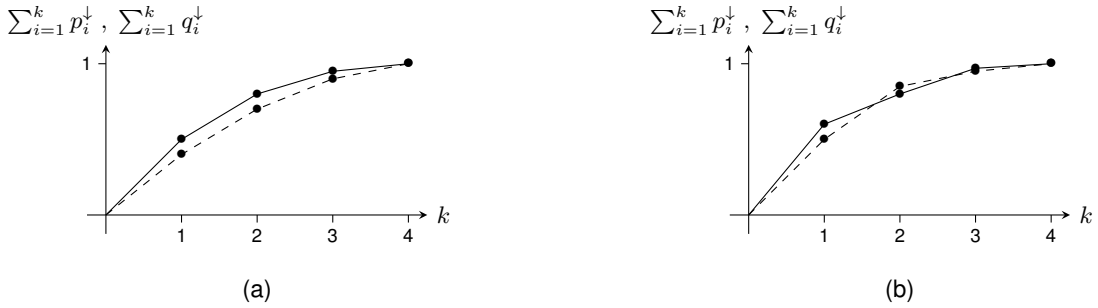



**Figura 1-5:** Orden parcial por mayorización: sumas parciales para $1 \leq k \leq \alpha = 4$ (a) para los vectores de probabilidades $p = [0{,}40 \quad 0{,}30 \quad 0{,}20 \quad 0{,}10]^t$ (linea punteada) y $q = [0{,}50 \quad 0{,}30 \quad 0{,}15 \quad 0{,}05]^t$ (linea llena) y (b) para los vectores de probabilidades $p = [0{,}50 \quad 0{,}40 \quad 0{,}06 \quad 0{,}04]^t$ (linea punteada) y $q = [0{,}70 \quad 0{,}15 \quad 0{,}13 \quad 0{,}02]^t$ (linea llena). En el caso (a), $p \prec q$ mientras que en el caso (b), $p \not\prec q$ y $q \not\prec p$ (no hay relaciones de mayorización).

### 1.3.4 Variable aleatoria continua

En varios contextos, puede tomar valores en un conjunto no numerable, por ejemplo cualesquiera de los números en un dado intervalo de la recta real. No son variables discretas más. En las variables que no son discretas, el caso particular de interés es el de variables continuas (Athreya & Lahiri, 2006; Hogg et al., 2013):

**Definición 1-24** (Variable aleatoria continua)**.** *Una variable aleatoria $X$ es dicha continua si su función de repartición $F_X$ es continua sobre $\mathbb{R}$.*

Cuando se puede, es conveniente asociar una *función densidad de probabilidad* (comúnmente anotada por su sigla en inglés: pdf por *probability density function*). Por esto, nos apoyamos sobre la definición 1-12 aplicada a la medida de probabilidad $P_X$:

**Definición 1-25** (Variable aleatoria admitiendo una densidad de probabilidad)**.** *Sea $X$ variable aleatoria continua y $P_X$ su medida de probabilidad. Por definición, se dice que $X$ admite una densidad de probabilidad con respecto a una medida $\mu$ sobre $\mathbb{R}$ si $P_X \ll \mu$ (teorema de Radon-Nikodým 1-7).*
*En general, nos enfocamos sobre la medida (dicha de referencia) $\mu = \mu_L$ de Lebesgue y denotando $d\mu_L(x) \equiv dx$, la definición se reduce a: Si existe una función no negativa $p_X$ medible sobre $\mathbb{R}$ tal que*

$$\forall\, B \in \mathcal{B}(\mathbb{R}), \quad P_X(B) = \int_B p_X(x)\, dx,$$

[7] Se proba sencillamente que estas curvas son crecientes y cóncavas.

*entonces $X$ es dicha admitiendo una densidad y $p_X$ es llamada densidad de probabilidad de $X$ (subentendido "con respecto a la medida de Lebesgue"). Notando de que $P_X(B) = P_X(B \cap \mathcal{X})$, el soporte de $p_X$ es necesariamente $\mathcal{X} = X(\Omega)$ (i. e., $p_X(\overline{\mathcal{X}}) = 0$ y $p_X(\mathcal{X}) \neq 0$), y*

$$\forall\, B \in \mathcal{B}(\mathbb{R}), \quad P_X(B) = \int_{B \cap \mathcal{X}} p_X(x)\, dx.$$

*Tratando de la función de repartición $F_X$, tenemos entonces*

$$F_X(x) = \int_{-\infty}^{x} p_X(u)\, du$$

*(queda valid con cualquier medida $\mu$, densidad con respecto a esta medida de referencia, e integración sobre $(-\infty; x]$ con la "diferencial" $d\mu(x)$). Dicho de otra manera, si $F_X$ es (continua y) derivable sobre $\mathbb{R}$, por lo menos por partes, $X$ admite una densidad de probabilidad (con respecto a la medida de Lebesgue) y* [8]

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

*Por abuso de terminologia, en lo que sigue llamaremos $p_X$ también distribución de probabilidad, a pesar de que no tiene el mismo sentido que la masa de probabilidad del caso discreto y denotaremos $|\mathcal{X}|$ el volumen (o medida de Lebesgue) de $\mathcal{X}$, posiblemente infinito.*

La escritura integral de $F_X$ justifica de nuevo la denominación *cumulativa* para $F_X$. Además, se puede ver por ejemplo que en este caso $P(a < X \leq b) = \int_a^b p_X(x)\, dx = F_X(b) - F_X(a)$ y que claramente

$$\forall x \in \mathbb{R}, \quad P_X(\{x\}) = P(X = x) = 0.$$

$\{x\}$ es de medida $P_X$ nula (es el caso de todos conjuntos numerable de $\mathbb{R}$).

Fijense de que si $0 \leq F_X \leq 1$, $p_X$ puede ser mayor que uno. Por ejemplo, para $F_X(x) = 2\,x\,\mathbb{1}_{[0\,;\,\frac{1}{2})}(x) + \mathbb{1}_{[\frac{1}{2}\,;\,+\infty)}(x)$, que define correctamente una función de repartición, $p_X(x) = 2\mathbb{1}_{[\frac{1}{2}\,;\,+\infty)}(x)$. No es contradictorio en el sentido de que $p_X$ no es una probabilidad, sino que $p_X(x)\, dx$ puede ser visto como la probabilidad de hallar a la variable con valores en el "intervalo infinitesimal entre $x$ y $x + dx$". Al final, la condición de normalización se escribe

$$\int_{\mathcal{X}} p_X(x)\, dx = \int_{\mathbb{R}} p_X(x)\, dx = 1.$$

En la figura Fig. 1-6-(a) se muestra una representación gráfica de una función densidad de probabilidad para una variable continua admitiendo una densidad, y en Fig. 1-6-(b) la función cumulativa correspondiente.

Fijense de que una variable aleatoria puede ser ni continua, ni discreta, como ilustrado en el ejemplo siguiente

**Ejemplo 1-4** (Ejemplo de variable mixta)**.** *Sean $U$ y $V$ variables continuas, independientes, de densidad de probabilidad $p_U = p_V = \mathbb{1}_{[0\,;\,1)}$ ($U$ y $V$ son dichas uniformes sobre $[0\,;\,1)$) y sea $X = V\mathbb{1}_{U<\frac{1}{2}} + \mathbb{1}_{U\geq\frac{1}{2}}$, es decir $X(\omega) = V(\omega)$ si $U(\omega) < \frac{1}{2}$ y $1$ si no. Entonces de la fórmula de probabilidades totales,*

[8] Recuendense que, rigurosamente, la igualdad debe ser entendido "casi siempre".

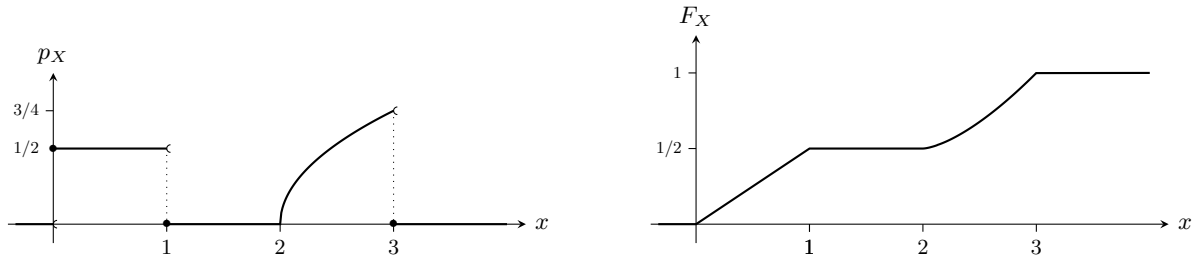



**Figura 1-6:** Ilustración de una distribución de probabilidad continua (a), y la función de repartición asociada (b), con $\mathcal{X} = [0\,;\,1) \cup [2\,;\,3)$ y $p_X(x) = \frac{1}{2}\mathbb{1}_{[0\,;\,1)}(x) + \frac{3\sqrt{x-2}}{4}\mathbb{1}_{[2\,;\,3)}(x)$, *i. e.,* $F_X(x) = \frac{x}{2}\,\mathbb{1}_{[0\,;\,1)}(x) + \frac{1}{2}\mathbb{1}_{[1\,;\,2)}(x) + \frac{(x-2)^{\frac{3}{2}}}{2}\mathbb{1}_{[2\,;\,3)}(x) + \mathbb{1}_{[3\,;\,+\infty)}(x)$.

$F_X(x) = P(X \leq x) = P\left((X \leq x)\,\middle|\,(U < \frac{1}{2})\right)\,P\left(U < \frac{1}{2}\right) + P\left((X \leq x)\,\middle|\,(U \geq \frac{1}{2})\right)\,P\left(U \geq \frac{1}{2}\right)$ *i. e.,* $F_X(x) = \frac{1}{2}P\left((V \leq x)\,\middle|\,(U < \frac{1}{2})\right) + P\left((1 \leq x)\,\middle|\,(U \geq \frac{1}{2})\right)$. *Ahora, de la independencia de* $U$ *y* $V$*, tenemos* $F_X(x) = \frac{1}{2}F_V(x) + \frac{1}{2}\mathbb{1}_{[1\,+\infty)}(x)$ *es decir*

$$F_X(x) = \frac{x}{2}\mathbb{1}_{[0\,;\,1)}(x) + \mathbb{1}_{[1\,;\,+\infty)}(x).$$

*Esta función de repartición es representada figura Fig. 1-7: es ni discreta, ni continua. Entonces, a pesar de que* $\mathcal{X} = [0\,;\,1]$ *sea un intervalo,* $X$ *no es continua (y tampoco no puede ser discreta).*

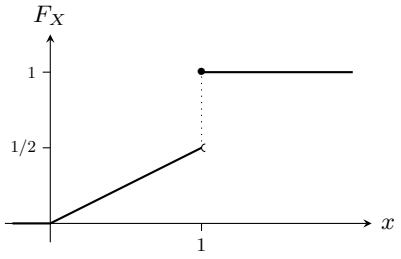



**Figura 1-7:** Función de repartición $F_X(x) = \frac{x}{2}\mathbb{1}_{[0\,;\,1)}(x) + \mathbb{1}_{[1\,;\,+\infty)}(x)$ asociada a $X = V\,\mathbb{1}_{U<\frac{1}{2}} + \mathbb{1}_{U\geq\frac{1}{2}}$ con $U$ y $V$ variables continuas uniformes sobre $\mathcal{X} = [0\,;\,1)$. No es tipo escalon, así que $X$ no es discreta. A pesar de que $\mathcal{X} = [0\,;\,1]$ sea un intervalo, de la presencia del salto en $x = 1$, tampoco $X$ no es continua.

Volvemos a las variables discretas $X$ sobre $\mathcal{X} = \{x_j\}_j$, de medida de probabilidad de la forma $P_X = \sum_j p_j \delta_{x_j}$. Considerando la medida discreta $\mu_{\mathcal{X}} = \sum_j \delta_{x_j}$, es claro de que $P_X \ll \mu_{\mathcal{X}}$. Entonces, formalmente, $P_X$ admite una densidad con respecto a la medida discreta $\mu_{\mathcal{X}}$ y esta densidad, definida sobre $\mathcal{X}$ es $p_X(x) = P(X = x)$. A pesar de que sea una tautología, este justifica de que usamos la escritura $p_X$ (minuscula) en el caso discreto como en el caso continuo, y de que hablamos (por abus de terminología) de distribución de probabilidad en ambos caso.

Recordamonos de que cualquier medida de probabilidad (caso continuo o no) se escribe también con una integral $P_X(B) = \int_B dP_X(x)$ y que en el caso discreto cierto $X = x_k$, la medida de probabilidad $P_X$ es la medida de Dirac. A veces, por abuso de escritura $dP_X(x)$ es denotado $\delta_{x_k}(x)\,dx$ o $\delta(x - x_k)\,dx$ donde ahora

$\delta$ es llamada *distribución (delta) de Dirac*. Se puede ver este Dirac como una densidad de probabilidad $p_X(x)$ con respecto a la medida de Lebesgue pero no es una función "ordinaria" dando de que $P_X$ no es diferenciable con respecto a la medida de Lebesgue. Se la llama *función generalizada* o *distribución de Schwartz* [9]. En particular, $F_X(x) = \mathbb{1}_{\mathbb{R}_+}(x - x_k)$ y en el sentido de las distribuciones, $\frac{dF_X}{dx} = \delta_{x_k}$. Además, se usan en general las propiedades, para cualquier function $f$ y real $x_0$,

$$f(x)\delta(x - x_0) = f(x_0)\delta(x - x_0) \qquad \text{y} \qquad \int_{\mathbb{R}} f(x)\delta(x - x_0)\, dx = f(x_0),$$

pero hay que entender la integración a través de la medida Dirac (insistamos en el hecho de que esta notación es un abuso de escritura, ej. (Gel'fand & Shilov, 1964)). Usando las distribuciones de Dirac, se puede unificar el tratamiento de las variables aleatorias discretas con las continuas en termino de densidad (con respecto a la medida de Lebesgue): si una variable aleatoria discreta toma los valores $x_j$ con probabilidades $p_j = P(X = x_j)$ respectivamente, entonces formalmente se puede describir mediante una variable aleatoria continua $X$ con "función densidad de probabilidad" $p_X(x) = \sum_j p_j\, \delta(x - x_j)$. Insistamos en el hecho de que rigurosamente debemos trabajar con medidas, como lo hemos formalizado al principio de este capítulo.

Terminamos mencionando el resultado siguiente, probado en (?, ?, Ec. (2.5) p. 47 & teorema 4.1.1) por ejemplo [10]:

**Teorema 1-8** (Descomposition de una medida de probabilidad)**.** *Cualquier función de repartición $F_X(x)$ se descompone como combinación convexa de una función de repartición $F_{\mathrm{d}}$ discreta y de una función de repartición $F_{\mathrm{c}}$ continua:*

$$\exists\, a \in [0\,;\,1] \quad \textit{tal que} \quad F(x) = aF_{\mathrm{d}}(x) + (1 - a)F_{\mathrm{c}}(x)$$

*En termino de medida, o como corrolario de (?, ?, teorema 4.1.1), cualquier medida de probabilidad $P_X$ se descompone como la combinación convexa de una medida discreta $P_{\mathrm{d}}$ y de una continua $P_{\mathrm{c}}$,*

$$\exists\, a \in [0\,;\,1],\ \widetilde{\mathcal{X}} \textit{ discreto} \quad \textit{tal que} \quad P = aP_{\mathrm{d}} + (1 - a)P_{\mathrm{c}} \qquad \textit{con} \qquad P_{\mathrm{d}} \ll \mu_{\widetilde{\mathcal{X}}} \quad \textit{y} \quad P_{\mathrm{c}} \ll \mu_L$$

*Entonces, $P_X \ll \mu_{\widetilde{\mathcal{X}}} + \mu_L$, i. e., admite una densidad con respecto a la medida $\sigma$-finita $\mu_{\widetilde{\mathcal{X}}} + \mu_L$.*

Dicho de otra manera, cualquier variable aleatoria es mixta, como en el ejemplo 1-4 pagina 37. Es discreta cuando $a = 1$, y continua cuando $a = 0$.

---

[9] La teoría de la distribuciones valió a Laurent Schwarz la medalla Field en 1950. Entre otros en el trabajo de Schwartz, se probó que el Dirac, visto como distribución de Schwartz, o función generalizada, tiene una "representación integral" $\delta(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{itx}dt$ o más rigurosamente transformada de Fourier de $x \mapsto 1$ en el sentido de las funciones generalizadas o distribuciones. Esto muestra claramente su caracter no ordinario (la integral siendo divergente en el sentido usual). Esto va más allá de la meta del capítulo y el lector se podrá referir a (Schwartz, 1966; Gel'fand & Shilov, 1964, 1968) por ejemplo.

[10] Basicamente, se muestra de que $\widetilde{X} = \{x \in \mathcal{X},\ \ p(x) = F_X(x) - \liminf_{u \to x} F(u) > 0\}$ es numerable. A continuación, $F(x) - \sum_{\widetilde{x} \in \widetilde{\mathcal{X}}} p(\widetilde{x})\mathbb{1}_{(-\infty\,;\,\widetilde{x}]}(x)$ es continua, y se recupera la descomposición con $a = \sum_{\widetilde{x} \in \widetilde{\mathcal{X}}} p(\widetilde{x})$.

### 1.3.5 Vector aleatorio discreto

Un ejemplo de vector aleatorio discreto puede verse a través de un conjunto de dados (que podrían ser dependientes si son ligados por un hilo por ejemplo).

**Definición 1-26** (Vector aleatorio discreto)**.** *Un vector aleatorio $d$-dimensional $X = \begin{bmatrix} X_1 & \cdots & X_d \end{bmatrix}^t$ y $\mathcal{X} = X(\Omega) \subset \overset{d}{\underset{i=1}{\times}} \mathcal{X}_i$ donde $\mathcal{X}_i = X_i(\Omega)$. $X$ es dicho discreto cuando $\mathcal{X} \subseteq \mathbb{N}^d$, es discreto, finito o infinito numerable. En lo que sigue, denotaremos también por $|\mathcal{X}|$ el cardinal de $\mathcal{X}$, posiblemente infinito.*

Obviamente, la medida de probabilidad en los $x = (x_1\,,\,\ldots\,,\,x_d) \in \overset{d}{\underset{i=1}{\times}} \mathcal{X}_i$ caracterisa completamente este vector aleatorio:

**Definición 1-27** (Función de masa de probabilidad conjunta)**.** *Por definición, la función de masa de probabilidad de $X$, vector aleatorio discreto tomando sus valores sobre $\mathcal{X} \subset \times_i \mathcal{X}_i$ es dada por*

$$p_X(x) \equiv P(X = x) = P\left(\bigcap_{i=1}^{d} (X_i = x_i)\right) \quad \forall\, x_i \in \mathcal{X}_i,\ 1 \leq i \leq d.$$

*Se la llama también función de masa de probabilidad conjunta de los $X_i$, o, por abuso de denominación, la llamaremos todavía $p_X$ distribución de probabilidad (conjunta). Fijense de que $\mathcal{X}$ no es necesariamente igual al producto cartesiano de los $\mathcal{X}_i$, siende de que una probabilidad conjunta en este producto puede ser nula. En el caso multivariado, la notación vectorial es más delicada a usar: $p_X$ sería un "tensor" $d$-dimensional (una matriz para $d = 2$,...). Pero queda posible usar una notación vectorial, recordandose de que $\mathbb{N}^d$ puede ser en biyección con $\mathbb{N}$ y una biyección elegida, usarla para etiquetar los componentes de $p_X$ puesto en vector. En el caso finito $\mathcal{X}_i = \{x_{j_i}\}_{j=1}^{\alpha_i}$ con $\alpha_i = |\mathcal{X}_i| < +\infty$, se puede organizar los componentes tales que $p_X(x_{j_1}\,,\,\ldots\,,\,x_{j_d})$ sea la $j = \sum_{i=1}^{d-1}(j_i - 1) \prod_{k=i+1}^{d} \alpha_k + j_d$-esima componente del vector $p_X$.*

De nuevo, $p_X$ puede ser vista como densidad con respecto a la medida discreta $\mu_{\mathcal{X}}$, Def. 1-15.

Como en el caso escalar, la función de repartición de un vector aleatorio discreto $d$-dimensional es echo de hyperplanos $d$-dimensionales constantes. Además, las componentes son mutuamente independientes si y solamente si la función de répartición se factoriza, o equivalentemente la función de masa se factoriza, *i. e.,*

$$X_i \text{ mutuamente independientes} \quad \Leftrightarrow \quad p_X = p_{X_1} \ldots p_{X_d}.$$

En notaciones tensoriales, $p_X = p_{X_1} \otimes \cdots \otimes p_{X_d}$ donde $\otimes$ denota el producto tensorial o externo entre los vectores de probabilidades [11]. Cuando los $\alpha_i$ son finito y la notación vectorial de la definición es adoptada,

[11] "Tensor" $d$-dimensional de componentes $(j_1, \ldots, j_d)$ el producto $\prod_i p_{X_i}(x_{j_i})$.

esta expresión queda valide donde $\otimes$ representa el producto de Kronecker [12].

Al final, de la fórmula de calculo de función de repartición marginales visto pagina 31, para un subconjunto $I_k = (i_1, \ldots, i_k)$ de $1 \leq k \leq d$ elementos de $\{1, \ldots, d\}^k$, $X_{I_k} = \begin{bmatrix} X_{i_1} & \cdots & X_{i_k} \end{bmatrix}^t$ la probabilidad marginal o distribución marginale de $X_{I_k}$ es dada por

$$p_{X_{I_k}}(x_{I_k}) = \sum_{\forall i \notin I_k, x_i \in \mathcal{X}_i} p_X(x).$$

### 1.3.6 Vector aleatorio continuo

Como para el caso de una variable, se puede considerar cualquiera medida de referencia $\mu$ sobre $\mathbb{R}^d$ para definir una noción de densidad ($d$-variada), pero en general nos enfocamos en el caso de la medida de Lebesgue.

**Definición 1-28** (Vector aleatorio continuo y densidad de probabilidad multivariada)**.** *Un vector aleatorio $X$ es dicho continuo si su función de repartición $F_X$ es continua sobre $\mathbb{R}^d$. Como en el caso escalar también, por definición 1-12 (y la reciproca evocaca al seguir), se dice que $X$ admite una densidad de probabilidad con respecto a una medida $\mu$ sobre $\mathbb{R}^d$ si $P_X \ll \mu$. De nuevo, nos enfocamos sobre la medida (dicha de referencia) $\mu = \mu_L$ de Lebesgue así que si existe una función no negativa y medible $p_X : \mathbb{R}^d \mapsto \mathbb{R}$ tal que*

$$\forall\, B \in \mathcal{B}(\mathbb{R}^d), \quad P_X(B) = \int_B p_X(x)\, dx = \int_{B \cap \mathcal{X}} p_X(x)\, dx$$

*con $\mathcal{X} = X(\Omega)$ soporte de $p_X$ y $d\mu_L(x) \equiv dx = dx_1 \cdots dx_d$, entonces $X$ es dicha admitiendo una densidad y $p_X$ es llamada densidad de probabilidad de $X$ ('subentendido "con respecto a la medida de Lebesgue"), o también densidad de probabilidad conjunta de los $X_i$. En particular,*

$$F_X(x) = \int_{\times_{i=1}^d (-\infty\,;\, x_i]} p_X(u)\, du$$

*o, equivalentemente, para $F_X$ (continua y) derivable sobre $\mathbb{R}^d$ (con respecto a la medida de Lebesgue), por lo menos por partes,*

$$p_X(x) = \frac{\partial^d F_X(x)}{\partial x_1 \cdots \partial x_d}.$$

*Usaremos todavía la terminología (por abuso) de distribución de probabilidad y denotaremos todavía $|\mathcal{X}|$ el volumen (o medida de Lebesgue) de $\mathcal{X}$, posiblemente infinito.*

Como en el caso escalar, $p_X \geq 0$ no es necesario menor que 1 y satisface la condición de normalización

$$\int_{\mathcal{X}} p_X(x)\, dx = \int_{\mathbb{R}^d} p_X(x)\; dx = 1.$$

---

[12] Para $p = \begin{bmatrix} p_1 & \cdots & p_n \end{bmatrix}^t$ y $q = \begin{bmatrix} q_1 & \cdots & q_m \end{bmatrix}^t$ el producto de Kronecker es dado por $p \otimes q$ vector de tamaño $nm$ de componente $(j-1)m + k$-esima el producto $p_j q_k, \quad 1 \leq j \leq n,\ 1 \leq k \leq m$. Fijense de que este producto es asociativo pero no es comutativo.

El teorema 1-8, pagina 38 se conserva en el caso $d$-dimensional: cualquier medida de probabilidad $P_X$ (resp. función de rpartición $F_X$) se descompone como combinación convexa de una medida de probabilidad (resp. función de repartición) discreta y una continua. En otro termino, existe un $\widetilde{X}$ discreto tal que $P_X \ll \mu_{\widetilde{\mathcal{X}}} + \mu_L$ ($\sigma$-finita).

Mencionamos de que las $d$ variables aleatorias $X_1, \ldots, X_d$, componentes de un vector aleatorio $X$ son independientes si y solamente si se factoriza la función de repartición, lo que da derivando esta,

$$X_i \text{ mutuamente independientes} \quad \Leftrightarrow \quad p_X(x) = p_{X_1}(x_1) \ldots p_{X_d}(x_d).$$

Seguimos esta sección mencionando que, de la fórmula de calculo de función de repartición marginales vista pagina 31, para un subconjunto $I_k = (i_1, \ldots, i_k)$ de $1 \leq k \leq d$ elementos de $\{1, \ldots, d\}^k$, $X_{I_k} = \begin{bmatrix} X_{i_1} & \cdots & X_{i_k} \end{bmatrix}^t$ la *densidad de probabilidad marginal* de $X_{I_k}$ es dada por

$$p_{X_{I_k}}(x_{I_k}) = \int_{\times_{i \notin I_k} \mathcal{X}_i} p_X(x) \prod_{i \notin I_k} dx_i = \int_{\mathbb{R}^{d-k}} p_X(x) \prod_{i \notin I_k} dx_i.$$

En particular, la función densidad de probabilidad marginal que caracteriza a la variable aleatoria $X_i$ es la ley que se obtiene integrando la densidad de probabilidad conjunta sobre todas las variables excepto la $i$-ésima.

Como en el caso discreto, se puede querrer comparar dos distribuciones de probabilidad, y por eso reordenar o rearreglar una densidad de probabilidad. Como en el caso discreto, se anota $p_X^{\downarrow}$ la densidad rearreglo simétrico. Primero, se necesita definir el rearreglo simétrico de un ensemble, y luego de una densidad de probabilidad:

**Definición 1-29** (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.,*

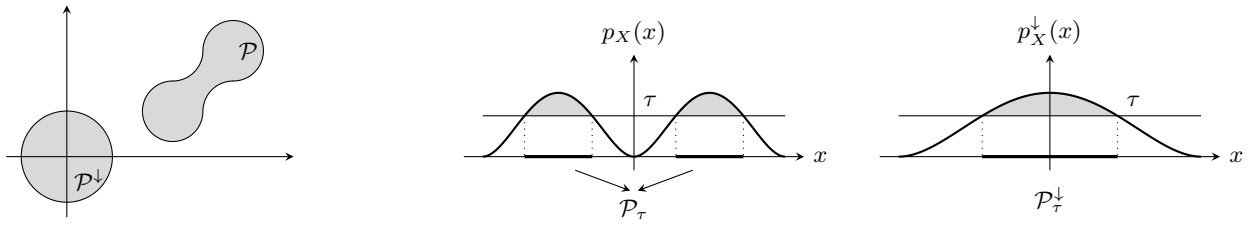
$$\mathcal{P}^{\downarrow} = \left\{ x \in \mathbb{R}^d \,:\, \frac{2\pi^{\frac{d}{2}} |x|^d}{\Gamma\left(\frac{d}{2}\right)} \leq |\mathcal{P}| \right\},$$

*donde $|\cdot|$ denota la norma euclideana. Eso es ilustrado figura Fig. 1-8-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 [13] rearreglada simétrico $p_X^{\downarrow}$ de $p_X$ es definida por*

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

*(recuerdense de que $\mathbb{1}_A$ es el indicator del conjunto $A$)*

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 $\mathcal{P}_{\tau+d\tau} \backslash \mathcal{P}_\tau$ la función $p_X$ "vale" $\tau$ y sobre $\mathcal{P}_{\tau+d\tau}^{\downarrow} \backslash \mathcal{P}_\tau^{\downarrow}$ la función $p_X^{\downarrow}$ "vale" tambien $\tau$, lo que da $\displaystyle\int_{\mathcal{P}_\tau^{\downarrow}} p_X^{\downarrow}(x)\, dx = \int_{\mathcal{P}_\tau} p_X(x)\, dx$ (ver (Lieb & Loss, 2001; Wang & Madiman, 2004) para une prueba más rigurosa). La representación de la definición es conocida como




(a) (b)

**Figura 1-8:** (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 $P_\tau^{\downarrow}$; dado un $x$, se busca el mayor $t$ tal que $x \in P_t^{\downarrow}$, este $t$ máximo siendo entonces $p_X^{\downarrow}(x)$; además, por construcción, las superficies en gris son iguales.

representación en capas de pastel ("layer cake" en ingles). Eso es ilustrado en la figura Fig. 1-8-b

A partir de esta definición del rearreglo, se puede ahora extender la noción de mayorización del caso discreto al caso continupo de la manera siguiente:

**Definición 1-30** (Mayorización en el contexto continuo)**.** *Una densidad de probabilidad $p$ es dicha mayorizada por una distribución $q$ si:*

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

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

**Equivalente de la curva de Lorentz???**

### 1.3.7 Transformación de variables y vectores aleatorios

En esta sección nos interesamos al effect de una variable o un vector aleatorio. Por ejemplo, en un juego con dos dados, nos podemos interesar a la ley de la suma que daría el número de casilla de que debemos adelantar en un juego de la oca.

**Teorema 1-9** (Transformación medible de un vector aleatorio)**.** *Sea $X : (\Omega, \mathcal{A}) \mapsto (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ una variable aleatoria, y $g : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \mapsto (\mathbb{R}^{d'}, \mathcal{B}(\mathbb{R}^{d'}))$ una función medible. Entonces, $Y = g(X)$ es una variable aleatoria $(\Omega, \mathcal{A}) \mapsto (\mathbb{R}^{d'}, \mathcal{B}(\mathbb{R}^{d'}))$. Además, la medida imágen $P_Y$ es vinculada a $P_X$ por*

$$\forall\, B \in \mathcal{B}(\mathbb{R}^{d'}), \quad P_Y(B) = P_X(g^{-1}(B)).$$

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

*Demostración.* Este resultado es obvio. $g$ siendo medible, para todo $B \in \mathcal{B}(\mathbb{R}^{d'})$, por definición $g^{-1}(B) \in \mathcal{B}(\mathbb{R}^d)$. Además, si $P_X$ es la medida (de probabilidad) asociado al espacio de salida de $g$, el resultado es consecuencia del teorema de la medida imágen 1-1, pagina 23. □

(Ver ej. (Jacob & Protters, 2003; Athreya & Lahiri, 2006; Bogachev, 2007b; Cohn, 2013)).

Es sencillo probar de que cualquier combinación de funciones medibles queda medible, cualquier producto (adecuado) de functiones medible queda medible, y que si $\{f_k\}_{k=1}^{d'}$ son $(\mathcal{B}(\mathbb{R}^d), \mathcal{B}(\mathbb{R}))$-medible, entonces $f = (f_1, \ldots, f_{d'})$ es $(\mathcal{B}(\mathbb{R}^d), \mathcal{B}(\mathbb{R}^{d'}))$-medible (Athreya & Lahiri, 2006).

Mencionamos de que si $\mathcal{X} = X(\Omega)$ es discreto, entonces $\mathcal{Y} = g(\mathcal{X}) = Y(\Omega)$ será discreto también, y:

**Teorema 1-10** (Función de masa por transformación medible)**.** *Sean $X$, vector aleatorio $d$-dimensional discreto, $g : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \mapsto (\mathbb{R}^{d'}, \mathcal{B}(\mathbb{R}^{d'}))$ una función medible, e $Y = g(X)$ necesariamente discreto $d'$-dimensional sobre $\mathcal{Y} = g(\mathcal{X})$. La distribución de $Y$ es relacionada a la de $X$ por la relación*

$$\forall\, y \in \mathcal{Y}, \quad p_Y(y) = \sum_{x \in g^{-1}(y)} p_X(x).$$

*Demostración.* El resultado es inmediato. □

En particular, si $g$ es inyectiva (necesariamente biyectiva de $\mathcal{X}$ en $\mathcal{Y}$), el vector de probabilidad queda invariante, $p_Y = p_X$; solamente cambian los estados.

Es importante mencionar de que con $\mathcal{Y}$ discreto, $\mathcal{X}$ no es necesariamente discreto (Athreya & Lahiri, 2006). Por ejemplo, $Y = \mathbb{1}_{X>0}$ es tal que $\mathcal{Y} = \{0\,;\,1\}$ a pesar de que $\mathcal{X}$ puede ser no discreto.

Tratar de las variables aleatorias continuas resuelta mas delicado. Vimos en el ejemplo precediente de que el caracter continuo puede perderse por transformación. De la misma manera, en un ejemplo de la sección precediente, vimos que $Y = X_1 \mathbb{1}_{X_2>0}$ con $X_i$ independientes uniformes es ni continua, ni discreta. En el enfoque de variables continuas, una clase importante de funciones en la cual no vamos a interesar son las funciones continuas (y diferenciables):

**Lema 1-3** (Continuidad y caracter medible)**.** *Sea $g : \mathbb{R}^d \mapsto \mathbb{R}^{d'}$ continua. Entonces, $g$ es $(\mathcal{B}(\mathbb{R}^d), \mathcal{B}(\mathbb{R}^{d'}))$-medible.*

*Demostración.* Por continuidad, la pre-imágen de un abierto de $\mathbb{R}^{d'}$ por $g$ es un abierto de $\mathbb{R}^d$ y entonces es en $\mathcal{B}(\mathbb{R}^d)$. La prueba se cierra recordandose de la definición de $\mathcal{B}(\mathbb{R}^{d'})$, $\sigma$-álgebra generada por los abiertos de $\mathbb{R}^{d'}$. □

En lo que sigue, nos interesamos más especialmente al caso de funciones $g : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \mapsto (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. De hecho, si $d' < d$, es sencillo llegar al caso considerado añandiendo $d - d'$ transformaciones. Por ejemplo, con $d = 2$ si nos interesamos a $X_1 + X_2$, se puede considerar $\begin{bmatrix} X_1 + X_2 & X_2 - X_1 \end{bmatrix}^t$ y llegar a la variable de interés por calculo de marginal. Si $d' > d$ la situación es más delicada, $g(Y)$ viviendo sobre una variedad $d$-dimensional de $\mathbb{R}^{d'}$.

En el caso de vectores aleatorios continuos $X$ admitiendo una densidad de probabilidad, una pregunta natural es entonces de saber si se conserva la continuidad y la existencia de una densidad, así que su forma.

La respuesta es dada por el teorema siguiente (Brémaud, 1988; Jacob & Protters, 2003; Athreya & Lahiri, 2006; Cohn, 2013; Hogg et al., 2013):

**Teorema 1-11** (Densidad de probabilidad por transformación continua inyectiva diferenciable)**.** *Sean $X$, vector aleatorio $d$-dimensional continuo y admitiendo una densidad de probabilidad $p_X$, $g : \mathbb{R}^d \mapsto \mathbb{R}^d$ una función continua inyectiva y diferenciable tal que $|\mathrm{J}_g| > 0$, donde $\mathrm{J}_g$ denota la matriz de componentes $\frac{\partial g_i}{\partial x_j}$, matriz Jacobiana de la transformación $g \equiv \begin{bmatrix} g_1(x_1, \ldots, x_d) & \cdots & g_d(x_1, \ldots, x_d) \end{bmatrix}^t$ y $|\cdot|$ representa el valor absoluto del determinante de la matriz. Sea $Y = g(X)$. Entonces $Y$ es continua admitiendo una densidad de probabilidad $p_Y$ de soporte $\mathcal{Y} = g(\mathcal{X}) = Y(\Omega)$ tal que*

$$\forall\, y \in \mathcal{Y}, \quad p_Y(y) = p_X(g^{-1}(y)) \left|\mathrm{J}_{g^{-1}}(y)\right|.$$

*Demostración.* Por definición, $X$ admitiendo una densidad y $g$ siendo medible,

$$\forall\, B \in \mathcal{B}(\mathbb{R}^d), \quad P_Y(B) = P_X(g^{-1}(B)) = \int_{g^{-1}(B)\cap\mathcal{X}} p_X(x)\, dx.$$

Por cambio de variable $x = g^{-1}(y)$ ($g$ siendo inyectiva, el antecedante es único por definicón) y notando de que $g\left(g^{-1}(B) \cap \mathcal{X}\right) = B \cap \mathcal{Y}$,

$$\forall\, B \in \mathcal{B}(\mathbb{R}^d), \quad P_Y(B) = \int_{B\cap\mathcal{Y}} p_X(g^{-1}(y)) \left|\mathrm{J}_{g^{-1}}(y)\right| dy$$

lo que cierra la prueba [14]. □

El caso escalar puede ser visto como caso particular, dando:

**Corolario 1-2.** *Sean $X$, variable aleatoria continua y admitiendo una densidad de probabilidad $p_X$, $g : \mathbb{R} \mapsto \mathbb{R}$ una función continua, inyectiva y diferenciable e $Y = g(X)$. Entonces $Y$ es continua admitiendo una densidad de probabilidad $p_Y$ tal que*

$$\forall\, y \in \mathcal{Y}, \quad p_Y(y) = p_X(g^{-1}(y)) \left|\frac{dg^{-1}(y)}{dy}\right|.$$

De hecho, se puede ver estos resultados esquematicamente como una "conservación" de probabilidad, $p_X(x)dx = p_Y(y)dy$, el volumen $dy$ siendo relacionado al $dx$ a traves de la Jacobiana (ver nota de pie **??**).

Una forma alternativa de derivar este corrolario consiste a salir de la función de repartición, notando de que $g$ es necesariamente monótona [15]: si $y \notin \mathcal{Y}$, necesariamente $p_Y = 0$ ($F_Y(y) = 1$ si $y > \sup \mathcal{Y}$ y $F_Y(y) = 0$

[14] La aparición de la Jacobiana viene del mismo enfoque que el cambio de variables en la integración de Rieman. De hecho, como lo hemos visto, $\mu_L(B) = |B|$ es el volumen y de la definición mismo del determinente, para cualquier matriz cuadrada el volumen se escribe $\mu_L(MB) = |MB| = |M||B| = |M|\mu_L(B)$ donde la misma escritura $|\cdot|$ representa el valor absoluto del determinente de una matriz. Esta notación se justifica precisamente por su significación de volumen, y el resultado es inmediato para $g(x) = Mx$. La forma general, para una transformación más general a partir de un desarollo de Taylor al orden 1 (Athreya & Lahiri, 2006; Cohn, 2013).

[15] Fijense de que $P(X \geq x) = 1 - P(X < x) = 1 - P(X \leq x) + P(X = x)$, pero $X$ siendo continua, $P(X = x) = 0$.

si $y < \inf \mathcal{Y}$) y para cualquier $y \in \mathcal{Y}$,

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = \begin{cases} P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)) & \text{si} \quad g \quad \text{es creciente} \\ P(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)) & \text{si} \quad g \quad \text{es decreciente} \end{cases}.$$

El resultado se obtiene calculando las derivadas del primer y último términos respecto de la variable transformada $y$.

Si $g$ no es inyectiva, $g^{-1}$ es multivaluada o multiforme. En este caso, se puede todavía tratar el problema, particionando $\mathbb{R}^d$ en conjuntos donde $g$ es inyectiva, dando

**Teorema 1-12.** *Sean $X$, vector aleatorio $d$-dimensional continuo y admitiendo una densidad de probabilidad $p_X$, $g : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \mapsto (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ una función continua y diferenciable. Denotamos $\left\{\mathcal{X}_{[k]}\right\}_{k=0}^m$ la partición de $\mathcal{X}$ tal que $|\mathrm{J}_g(y)| = 0$ sobre $\mathcal{X}_{[0]}$ y para todos $k \geq 1$, $g : \mathcal{X}_{[k]} \mapsto \mathcal{Y}$ sea inyectiva y tal que $|\mathrm{J}_g(y)| > 0$. Suponemos de que $\mathcal{X}_{[0]}$ sea de medida de Lebesgue nula, notamos $g_k^{-1}$ la función inversa de $g$ sobre $g(\mathcal{X}_{[k]})$ (rama $k$-ésima de la función multivaluada $g^{-1}$), $\mathrm{J}_{g_k^{-1}}$ su matriz Jacobiana y $I(y) = \{k, y \in g(\mathcal{X}_{[k]})\}$ los indices tales que $y$ tiene un inverso por $g_k$. Esto es ilustrado figura Fig. 1-9 para $d = 1$. Entonces $Y$ es continua admitiendo una densidad de probabilidad $p_Y$ tal que*

$$\forall\, y \in \mathcal{Y}, \quad p_Y(y) = \sum_{k \in I(y)} p_X(g_k^{-1}(y)) \left|\mathrm{J}_{g_k^{-1}}(y)\right|.$$

*En el caso escalar $d = 1$ esto se formula*

$$\forall\, y \in \mathcal{Y}, \quad p_Y(y) = \sum_{k \in I(y)} p_X(g_k^{-1}(y)) \left|\frac{dg_k^{-1}(y)}{dy}\right|.$$

*Demostración.* Sufice escribir $B = \bigcup_{k=0}^m (B \cap g(\mathcal{X}_k))$ unión de borelianos disjuntos, notar de que por consecuencia $g^{-1}(B) = \bigcup_{k=0}^m g^{-1}(B \cap g(\mathcal{X}_k))$ unión de borelianos disjuntos y por linearidad escribir la integración sobre $g^{-1}(B)$ como la suma de integrales sobre $g^{-1}(B \cap g(\mathcal{X}_k))$. Se cierra la prueba notando de que $g^{-1}(B \cap g(\mathcal{X}_0))$ es necesario de medida de Lebesgue nula, siendo la integral nula y de que $g^{-1}(B \cap g(\mathcal{X}_k)) = g_k^{-1}(B \cap g(\mathcal{X}_k))$. □

**Ejemplo 1-5** (Ejemplo de transformación no biyectiva)**.** *Sea $X$ definido sobre $\mathcal{X} = \mathbb{R}$ y la transformación de variables $Y = X^2$. Se tiene $y = g(x) = x^2$, continua diferenciable de derivada nula sobre $\mathcal{X}_{[0]} = \{0\}$, de medida nula, cuyas inversas son $g_1^{-1}(y) = \sqrt{y}$ sobre $\mathcal{X}_{[1]} = \mathbb{R}_-^*$ y $g_2^{-1}(y) = -\sqrt{y}$ sobre $\mathcal{X}_{[2]} = \mathbb{R}_+^*$; luego $p_Y(y) = \frac{p_X(\sqrt{y}) + p_X(-\sqrt{y})}{2\sqrt{y}}$, sobre $\mathcal{Y} = \mathbb{R}_+^*$.*

De nuevo, en el caso escalar, se puede salir de la función de repartición

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = \sum_{k=1}^m P\left(X \in \mathcal{X}_{[k]} \cap g_k^{-1}(-\infty\,;\, y]\right)$$

($\mathcal{X}_{[0]}$ siendo de medida nula, sobre este dominio la probabilidad es cero). Sea $\mathcal{Y}_{[k]} = g_k(\mathcal{X}_{[k]})$. Ahora, si $y \notin I(y)$,

$$P\left(X \in \mathcal{X}_k \cap g_k^{-1}(-\infty\,;\, y]\right) = \begin{cases} P(X \in \mathcal{X}_{[k]}) & \text{si} \quad y > \sup \mathcal{Y}_{[k]} \\ 0 & \text{si} \quad y < \inf \mathcal{Y}_{[k]} \end{cases}$$

dando una derivada nula. Si $y \in I(y)$,

$$P\left(X \in \mathcal{X}_k \cap g_k^{-1}(-\infty\,;\,y]\right) = \begin{cases} F_X(g_k^{-1}(y)) - F_X(\text{ínf}\,\mathcal{Y}_{[k]}) & \text{si} \quad g_k \quad \text{es creciente} \\ F_X(\sup \mathcal{Y}_{[k]}) - F_X(g_k^{-1}(y)) & \text{si} \quad g_k \quad \text{es decreciente} \end{cases}.$$

El resultado sigue diferenciando este resultado. Esto es ilustrado figura Fig. 1-9.

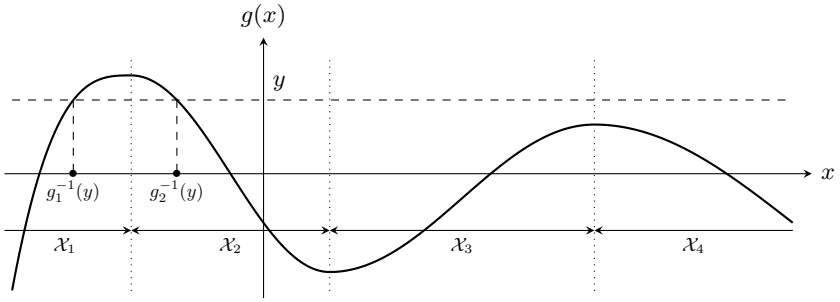



**Figura 1-9:** (a): Ilustración de una transformación $g$ no inyectiva, tal que $\mathcal{X}_{[0]} = \{x,\, g'(x) = 0\}$, representado por las lineas punteadas ($x$ correspondiente), es de medida de Lebesgue nula. Los $\mathcal{X}_{[k]}$ son descrito debajo de cada dominio. La linea discontinua da un nivel $y$ y los puntos en el eje $x$ representan $g_k^{-1}(y),\ k \in I(y)$; en el ejemplo, $I(y) = \{1\,;\,2\}$ y, suponiendo de que $\mathcal{X} = \mathbb{R},\ F_Y(y) = F_X(g_1^{-1}(y)) + 1 - F_X(g_2^{-1}(y))$.

Una tercera alternativa, a pesar que sea delicado, es de apoyarse sobre la teoria de las distribuciones y expresar como $p_Y(y) = \displaystyle\int_{\mathcal{X}} p_X(x)\,\delta(y - g(x))\,dx$, donde se usa la expansión de la función delta en términos de sus ceros: $\delta(y - g(x)\,) = \sum_{k \in I(y)} \frac{1}{\left|g_k'\left(g_k^{-1}(y)\right)\right|}\delta(x - g_k^{-1}(y))$ (Mandel & Wolf, 1995).

Es importante notar de que la condición $\mathcal{X}_{[0]}$ de medida nula es importante. El el caso contrario, $Y$ no queda continua como se lo puede ver en el ejemplo siguiente.

**Ejemplo 1-6** (Transformación con $\mu_L\left(\mathcal{X}_{[0]}\right) \neq 0$)**.** *Sea $X$ uniforme sobre $\mathcal{X} = (3\,;\,3)$ y $Y = g(X)$ con $g(x) = \left(1 + \cos\left((|x| - 1)\frac{\pi}{2}\right)\right)\mathbb{1}_{(1\,;\,3)}(|x|) + 2\mathbb{1}_{[0\,;1]}(|x|)$. Esta función es representado figura Fig. 1-10-(a). Claramente, $g$ es continua y diferenciable sobre $\mathcal{X}$, pero con $\mathcal{X}_{[0]} = [-1\,;\,1]$ que no es de medida nula. Saliendo de $F_Y(y) = P(g(X) \leq y)$ se calcula sencillamente $F_Y(y) = \frac{2}{3}\left(1 - \frac{1}{\pi}\arc\cos(y-1)\right)\mathbb{1}_{[0\,;\,2)} + \mathbb{1}_{[2\,;\,+\infty)}(y)$, ilustrada figura Fig. 1-10-(b). Claramente $F_Y$ es discontinua en $y = 2$: $Y$ no es continua.*

Un ejemplo de cambio de transformación puede sevir a calcular la densidad de probabilidad de una suma:

**Ejemplo 1-7** (Distribución de la suma de vectores aleatorios)**.** *Sean $X$ e $Y$ dos vectores aleatorios conjuntamente continuos, de densidad de probabilidad conjunta $p_{X,Y}$, y sea el vector*

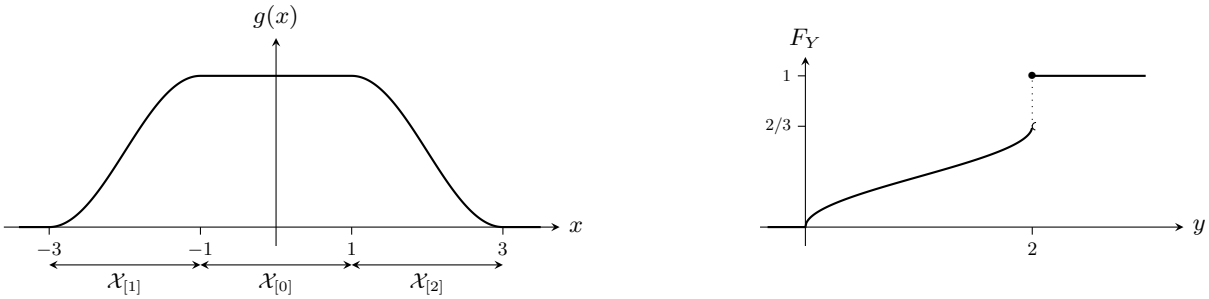
$$V = X + Y.$$

*Queremos calcular la a partir de la densidad de probabilidad de $V$. Por esto, se puede considerar la transformación biyectiva*

$$g : (x, y) \mapsto (u, v) = (x, x + y).$$

*Entonces*

$$g^{-1}(u, v) = (u, v - u)$$




(a) (b)

**Figura 1-10:** En (a) se dibuja $g(x) = \left(1 + \cos\left((1-|x|)\frac{\pi}{2}\right)\right) \mathbb{1}_{(1\,;\,3)}(|x|) + 2\mathbb{1}_{[0\,;1]}(|x|)$. Suponiendo de que $\mathcal{X} = (-3\,;\,3)$, claramente $\mathcal{X}_{[0]} = [-1\,;\,1]$ no es de medida nula, dando para $X$ uniforme sobre $\mathcal{X}$ la variable $Y = g(X)$ no continua de función de repartición representenda en (b).

*y la matriz Jacobiana es*

$$J_{g^{-1}} = \begin{bmatrix} I & -I \\ 0 & I \end{bmatrix}$$

*donde $I$ es la matriz identidad $d$-dimensional y $0$ la matriz nula de misma dimension. Claramente $\left|J_{g^{-1}}\right| = 1$ así que*

$$p_{U,V}(u,v) = p_{X,Y}(u, v-u)$$

*como lo pudimos intuir. Además, por marginalización, inmediatamente*

$$p_V(v) = \int_{\mathbb{R}^d} p_{X,Y}(u, v-u)\, du.$$

*Si $X$ e $Y$ son independientes, $p_{U,V}(u,v) = p_X(u)p_V(v-u)$ y la fórmula integral se escribe*

$$p_V(v) = \int_{\mathbb{R}^d} p_X(u)p_Y(v-u)\, du = \int_{\mathbb{R}^d} p_Y(u)p_X(v-u)\, du$$

*(por cambio de variable en la secunda expresión). Esta fórmula es conocida como producto de convolución entre las funciones*[16] *$p_X$ y $p_Y$.*

**No toque todavía. Se puede hacerlo con vectores. Caso circular. . .**

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

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

[16] Este producto no impone de que las funciones sean densidades de probabilidad. Una condición suficiente para que existe es que las funciones sean $L^1$ (ver desigualdad de Cauchy-Bunyakovsky-Schwarz).

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

### 1.3.8 Leyes condicionales

Tratando de un par de vectores aleatorios $X$ e $Y$, una pregunta natural puede ser de caracterizar el vector $Y$ si "observamos $X = x$". En palabras, la pregunta es de describir la ley de $Y$ "sabiendo de que $X = x$". En lo que sigue, para fijar las notaciones, consideramos $(X, Y) : (\Omega, \mathcal{A}) \mapsto (\mathbb{R}^{d_X} \times \mathbb{R}^{d_Y}\ , \mathcal{B}(\mathbb{R}^{d_X} \times \mathbb{R}^{d_Y}))$ tal que $X$ sea $d_X$-dimensional e $Y$ sea $d_Y$-dimensional (incluyendo los casos escalares).

**Caso $X$ discreto:** Un caso sencillo a estudiar es cuando $\mathcal{X} = X(\Omega)$ es discreto. En este caso, para cualquier $x \in \mathcal{X}$, tenemos $P_X(x) = P(X = x) \neq 0$ y de la definición de la probabilidad condicional Def. 1-3, $P(Y \in A|X = x) = \frac{P(Y \in A \cap X = x)}{P(X = x)}$ define una medida de probabilidad que llamamos medida de probabilidad condicional y que notaremos

$$P_{Y|X=x}(A) = P(Y \in A|X = x).$$

Siendo una medida de probabilidad, se puede referirse a la subsección anterior para definir una función de repartición tomando $A = \prod_{i=1}^{d}(-\infty\,;\, y_i]$, caracterisando completamento la medida de probabilidad:

**Definición 1-31** (Función de repartición condicional ($X$ discreto))**.** *Por definición, la función de repartición condicional es,*

$$\forall\, x \in \mathcal{X},\ y \in \mathcal{Y}, \quad F_{Y|X=x}(y) = P(Y \leq y|\, X = x) = \frac{P\,(Y \leq y \cap X = x)}{P\,(X = x)}.$$

Ahora, cuando $Y$ est discreta también, se puede definir la función de masa discreta de probabilidad, y si $Y$ es continua admitiendo una densidad de probabilidad, se puede definir una densidad de probabilidad condicional:

**Definición 1-32** (Función de masa o densidad de probabilidad condicional ($X$ discreto))**.** *Por definición, cuando $\mathcal{Y}$ est discreto, la función de masa de probabilidad condicional es,*

$$\forall\, x \in \mathcal{X},\ y \in \mathcal{Y}, \quad p_{Y|X=x}(y) = P(Y = y|\, X = x) = \frac{P\,(Y = y \cap X = x)}{P\,(X = x)}.$$

*Si $Y$ es continua, es sencillo ver de que $P_{Y|X=x} \ll P_Y$, i.e., $P_Y(B) = 0 \Rightarrow P_{Y|X=x}(B) = 0$. Si $Y$ adminte una densidad con respecto a la medida de Lebesgue, $P_Y \ll \mu_L$ medida de Lebesgue, es claro de que también $P_{Y|X=x} \ll \mu_L$, y por teorema de Radon-Nikodým 1-7, admite una densidad de probabilidad (con respecto a la medida de Lebesgue) que denotaremos $p_{Y|X=x}$,*

$$\forall\, B, \quad P_{Y|X=x}(B) = \int_B p_{Y|X=x}(y)\,dy.$$

*Saliendo de la función de repartición, obtenemos*

$$p_{Y|X=x}(y) = \frac{\partial^{d_Y} F_{Y|X=x}(y)}{\partial y_1 \ldots \partial y_{d_Y}}.$$

**Caso $X$ continuo admitiendo una densidad:** Cuando $X$ es continuo, el problema aparece más subtile porque $P(X = x) = 0$. Entonces, no se puede usar la definición de la probabilidad condicional, el evento $(X = x)$ siendo de probabilidad cero. Sin embargo, se puede seguir los pasos de Rényi (?, ?, Cap. 5) por ejemplo para resolver el problema, llegando a un resultado tan intuitivo que en el caso discreto.

Por esto, sea $B \in \mathcal{B}(\mathbb{R}^{d_Y})$ y definimos la medida $\nu_B(A) = P(X \in A \cap Y \in B)$ sobre $(\mathbb{R}^{d_Y}, \mathcal{B}(\mathbb{R}^{d_Y}))$. Es sencillo ver de que $B$ siendo dado, $\nu_B$ define una medida de probabilidad. Además, $\nu_B \ll P_X$, i.e., $P_X(A) = P(X \in A) = 0 \;\Rightarrow\; 0 = P(X \in A \cap Y \in B) = \nu_B(A)$. Por teorema de Radon-Nikodým 1-7, $\nu_B$ admite una densidad $g_B$ con respecto a $P_X$,

$$\forall\, A, \quad P\left((X \in A) \cap (Y \in B)\right) = \int_A g_B(x)\, dP_X(x).$$

Claramente $g_B \geq 0$, y de $P(X \in A) = P\left((X \in A) \cap (Y \in B)\right) + P\left((X \in A) \cap (Y \in B)\right)$, *i. e.,* $0 \leq P\left((X \in A) \cap (Y \in B)\right) = \int_A dP_X(x) - \int_A g_B(x)\, dP_X(x)$, se obtiene $0 \leq g_B \leq 1$. En realidad, tenemos $g_B \leq 1$ $P_X$-casi siempre, pero olvidando esta subtilesa, llamaremos la función $g_B$ medida de probabilidad condicional y, por continuación, la función de repartición condicional:

**Definición 1-33** (Medida de probabilidad y función de repartición condicional ($X$ continuo))**.** *La medida de probabilidad condicional de $P_{Y|X=x}$ es definida tal que*

$$\forall\, A \in \mathcal{X}, B \in \mathcal{Y}, \quad P\left((X \in A) \cap (Y \in B)\right) = \int_A P_{Y|X=x}(B)\, dP_X(x).$$

*Tomando $B = \times_i(-\infty\,;\, y_i]$ se obtiene la función de repartición condicional a partir de*

$$\forall\, A \in \mathcal{X}, y \in \mathcal{Y}, \quad P\left(X \in A) \cap (Y \leq y)\right) = \int_A F_{Y|X=x}(y)\, dP_X(x).$$

*Adémas, si $X$ admite una densidad de probabilidad $p_X$, $dP_X = p_X dx$ y tomando $A = \times_i(-\infty\,;\, x_i]$ se obtiene*

$$F_{X,Y}(x, y) = \int_{\times_i(-\infty\,;\,x_i]} F_{Y|X=x}(y)\, p_X(x)\, dx$$

*o, por diferenciación, para cualquier $y \in \mathcal{Y}$,*

$$F_{Y|X=x}(y) = \frac{\frac{\partial^{d_X}}{\partial x_1 \ldots \partial x_{d_X}} F_{X,Y}(x, y)}{p_X(x)}.$$

Nota: Tomando $A = \mathcal{X}$, de la primera fórmula definiendo la medida de probabilidad condicional se recupera el equivalente continuo de la fórmula de probabilidad total, que se escribe con la densidad $dP_X = p_X d\mu_L$ si $X$ admite una densidad.

Al final, si $(X, Y)$ admite una densidad, en sencillo ver de que $P_{Y|X=x} \ll \mu_L$, y entonces $P_{Y|X=x}$ admite una densidad que llamaremos *densidad de probabilidad condicional*. Sean $A \in \mathcal{B}(\mathcal{X})$ y $B$,

$$\begin{aligned} P\left((X \in A) \cap (Y \in B)\right) &= \int_{A \times B} p_{X,Y}(x, y)\, dx\, dy \\ &= \int_B \left( \int_A \frac{p_{X,Y}(x, y)}{p_X(x)}\, dy \right) p_X(x)\, dx \end{aligned}$$

Entonces, desde de que $p_X(x) \neq 0$, tenemos

$$P_{Y|X=x}(A) = \int_A \frac{p_{X,Y}(x, y)}{p_X(x)}\, dy.$$

**Teorema 1-13** (Densidad de probabilidad condicional)**.** *Si* $(X, Y)$ *admite una densidad de probabilidad, la medida de probabilidad condicional* $P_{Y|X=x}$ *admite una densidad, llamada densidad de probabilidad condicional definida por*

$$\forall\, x \in \mathcal{X}, \quad p_{Y|X=x}(y) = \frac{p_{X,Y}(x,y)}{p_X(x)}$$

*definida sobre* $\mathcal{Y}$*. Claramente, saliendo de la función de repartición condicional, aparece de que*

$$p_{Y|X=x} = \frac{\partial^{d_Y}}{\partial y_1 \dots \partial y_{d_Y}} F_{Y|X=x}.$$

De hecho, esta construcción rigurosa coincide con la intuición que podemos tener en este caso continuo. Por ejemplo, podemos pensar a $F_{Y|X=x}(y)$ como caso límite de $P(Y \leq y | \, x \leq X \leq x + \delta x) = \frac{P(Y \leq y \cap x \leq X \leq x+\delta x)}{P(x \leq X \leq x+\delta x)} = \frac{F_{X,Y}(x+\delta x, y) - F_{X,Y}(x,y)}{F_X(x+\delta x) - F_X(x)}$ cuando $\delta x$ tiende a 0. En el caso escalar, se calcula por ejemplo haciendo un desarollo de Taylor del numerador y del denominador al orden 1, o usando la regla de l'Hôpital [17] para re-obtener la función de repartición condicional de la definición Def. 1-33. En el caso multivariado, hace falta hacer los desarollos hasta el orden $d_X$ para concluir.

Fijense de que:

- si $X$ e $Y$ son independientes,

$$p_{Y|X=x} = p_Y;$$

- por la expresión $p_{Y|X=x}(y) = \frac{p_{X,Y}(x,y)}{p_X(x)}$, por integración con respecto a $y$ obtenemos la condición de normalización

$$\int_{\mathbb{R}^{d_Y}} p_{Y|X=x}(y)\, dy = 1;$$

- escribiendo $p_{X,Y}(x,y) = p_{Y|X=x}(y)\, p_X(x) = p_{X|Y=y}(x)\, p_Y(y)$, se obtiene

$$p_{Y|X=x}(y) = \frac{p_{X|Y=y}(x)\, p_Y(y)}{p_X(x)} = \frac{p_{X|Y=y}(x)\, p_Y(y)}{\displaystyle\int_{\mathbb{R}^{d_Y}} p_{X|Y=y}(x)\, p_Y(y)\, dy},$$

  equivalente continuo, con densidades, de la fórmula de Bayes;

- por la expresión $p_{X,Y}(x,y) = p_{Y|X=x}(y)\, p_X(x)$, por integración con respecto a $x$ obtenemos

$$p_Y(y) = \int_{\mathbb{R}^{d_X}} p_{Y|X=x}(y)\, p_X(x)\, dx,$$

  generalización de la fórmula de probabilidades totales al caso continuo con densidad de probabilidad.

Volvemos al ejemplo 1-7, pagina 46:

---

[17] De hecho, esta regla es debido al suizo J. Bernoulli que tuvo un acuerdo financiero con el Guillaume François Antoine, marqués de l'Hôpital, permitiendolo de publicar unos resultados de Bernoulli bajo su nombre.

**Ejemplo 1-8** (Distribución condicional de la suma de vectores aleatorios)**.** *Sea $V = X + Y$, con $X$ e $Y$ vecores $d$-dimensionales. Introduciendo $U = X$ obtuvimos $p_{U,V}(u,v) = p_{X,Y}(u, v-u)$ dando también $p_V(v) = \int_{\mathbb{R}^d} p_{X,Y}(u, v-u)\, du$. Entonces, recordandose de que $U = X$, se obtiene*

$$p_{V|X=x}(v) = \frac{p_{X,Y}(x, v-x)}{p_X(x)} = \frac{p_{X,Y}(x, v-x)}{\int_{\mathbb{R}^d} p_{X,Y}(x, v-x)\, dv},$$

*dando en el caso $X$ e $Y$ independientes*

$$p_{V|X=x}(v) = p_Y(v-x).$$

*Esto corresponde a la intución de que, con $V = X + Y$, fijando $X = x$ el vector aleatorio $V$ es nada mas que $Y$ desplazado de $x$. Pero hay que tomar muchas precauciones con este razonamiento, valide únicamente porque $X$ e $Y$ son independiente. En el caso contrario, fijando $X$ no coincide con un desplazamiento por la dependencia (esquematícamente, fijando $X$ no sólo mueve $Y$ pero "cambia" su estadística).*

## 1.4 Esperanza, momentos, identidades y desigualdades

*introducción...*

### 1.4.1 Media de un vector aleatorio

Una variable aleatoria $X$ tiene asociado un *promedio* o *media* (también llamado *valor esperado o de expectación o esperanza matemática*) que se obtiene pesando cada valor de $X$ con la medida de probabilidad asociada a ese valor,

**Definición 1-34** (Media o valor/vector medio)**.** *Formalmente, la media de una variable aleatoria $X$ integrable es definida por*

$$\mathrm{E}[X] = \int_\Omega X(\omega)\, dP(\omega).$$

*Por el teorema de la medida imagen 1-1, pagina 23, esta media se escribe también a partir de la medida de probabilidad $P_X$ como*

$$\mathrm{E}[X] = \int_{\mathbb{R}} x\, dP_X(x).$$

*En el caso vectorial $d$-dimensional, hay que entender la media, o vector medio, como un vector de componentes $i$-ésima la media $\mathrm{E}[X_i]$ de la componente $i$-ésima $X_i$ de $X$, dando*

$$\mathrm{E}[X] = \int_{\mathbb{R}^d} x\, dP_X(x).$$

*A veces, se encuentra también la notación $\langle x \rangle$ o $\langle x \rangle_{p_X}$ para el valor medio, especialmente en la literatura de física.*

La secunda formulación del valor medio se proba sencillamente, empezando por $X = \mathbb{1}_A$ para unos $A$. Entonces $P_X = (1-P(A))\delta_0 + P(A)\delta_1$. Luego $\int_\Omega \mathbb{1}_A(\omega) dP(\omega) = P(A) = (1-P(A))\times 0 + P(A)\times 1 = \int_{\mathbb{R}} x dP_X(x)$. Se cierra la prueba con el teorema 1-2 dando cualquier función medible como limite de funciones escalonadas, y por la definición 1-10 de la integral de cualquier función medible.

Luego, de la distribución marginal $P_{X_i}(B) = \int_{\mathbb{R}^{i-1}\times B\times\mathbb{R}^{d-i}} dP_X(x)$, se obtiene $\mathrm{E}[X_i] = \int_{\mathbb{R}^d} x_i \, dP_X(x)$, dando la última formulación en el caso vectorial.

Una variable aleatoria $X$ se dice integrable cuando $E[|X|] < \infty$. De la misma manera, un vector aleatorio admite una media si y solamente si cada componente es integrable. Veremos más adelante que existen variables aleatorias que no admiten una media.

Más allá de la formulación matemática de la media $\mathrm{E}[X]$ representa la posición alrededor de la cual se "distribuye las probabilidades de occurencia". Es el equivalente probabilístico de centro de gravedad o barycentro en mecánica.

En el caso de variables aleatorias discretas, de soporte $\mathcal{X} = \{x_i\}$, inmediatamente

$$\mathrm{E}[X] = \sum_i x_i P(X = x_i) = \sum_i x_i p_X(x_i).$$

Fijense de que $\mathrm{E}[X]$ no partenece necesariamente a $\mathcal{X}$:

**Ejemplo 1-9.** *Sea $X$ uniforme sobre $\mathcal{X} = \{1\,,\,3\,,\,7\}$, i. e., $\forall\, i \in \mathcal{X}, \quad P(X = i) = \frac{1}{3}$. Se calcula $\mathrm{E}[X] = 1\times\frac{1}{3} + 3\times\frac{1}{3} + 7\times\frac{1}{3} = \frac{11}{3} \notin \mathcal{X}$. Tampoco es el promedio de los valores extremos.*

Cuando $|\mathcal{X}| = +\infty$, $X$ no es necesariamente integrable:

**Ejemplo 1-10.** *Sea $\mathcal{X} = \mathbb{N}^*$ con $P(X = n) = \frac{6}{\pi^2 n^2}$. Claramente, $\sum_n \frac{6}{\pi^2 n}$ diverge, así que $X$ no tiene una media.*

En el caso de vectores aleatorios continuos, obtenemos la expresión siguiente de la media (o vector medio):

$$\mathrm{E}[X] = \int_{\mathbb{R}^d} x\, p_X(x)\, dx.$$

Las mismas observaciones que hicimos en el caso discreto se encuentra en el caso continuo:

**Ejemplo 1-11.** *Sea $X$ de densidad de probabilidad $p_X(x) = \frac{1}{2}\mathbb{1}_{[0\,;\,1)}(x) + \frac{3\sqrt{x-2}}{4}\mathbb{1}_{[2\,;\,3)}(x)$ como ilustrado figura Fig. 1-6, pagina 36. Se calcula $\mathrm{E}[X] = \frac{31}{20} \notin \mathcal{X} = [0\,;\,1] \cup [2\,;\,3]$.*

**Ejemplo 1-12.** *Un ejemplo de vector aleatorio no teniendo media es dado en el caso de una distribución de Cauchy-Lorentz (ver más adelante) $p_X(x) = \dfrac{\alpha}{(1 + x^t x)^{\frac{d+1}{2}}}$ donde $\alpha$ es un factor de normalización.*

En el caso general, para calcular la media, hay que pasar por la distribución $P_X$, como en el ejemplo 1-4 pagina 37:

**Ejemplo 1-13** (Continuación del ejemplo 1-4)**.** *Sea $X = V\,\mathbb{1}_{U<\frac{1}{2}} + \mathbb{1}_{U\geq\frac{1}{2}}$ con $U$ y $V$ variables aleatorias independientes de distribución uniformas sobre $[0\,;\,1)$, i. e., $p_U(x) = \mathbb{1}_{[0\,;\,1)}(x)$. De $X \in B \Leftrightarrow \big(\big(U < \frac{1}{2}\big) \cap (V \in B)\big) \cup \big(\big(U \geq \frac{1}{2}\big) \cap (1 \in B)\big)$, del hecho de que los eventos de la unión son incompatibles y*

*de la independencia de $U$ y $V$ (o saliendo de la función de repartición), se obtiene $P_X(B) = \frac{1}{2}P_V(B) + \frac{1}{2}\delta_1(x)$. A continuación,* $\mathrm{E}[X] = \dfrac{1}{2}\displaystyle\int_{\mathbb{R}} dP_V(x) + \frac{1}{2}\int_{\mathbb{R}} d\delta_1(x) = \frac{1}{2}\int_{\mathbb{R}} p_V(x)\,dx + \frac{1}{2}\times 1 = \frac{1}{2}\int_0^1 dx + \frac{1}{2} = \frac{3}{4}$.

Una nota intersante es de que, en el caso escalar, si $X \geq 0$ admitiendo una media, se obtiene

$$\mathrm{E}[X] = \int_{\mathbb{R}_+} P(X > t)\,dt = \int_{\mathbb{R}_+} (1 - F_X(t))\ dt.$$

Se proba saliendo de $x = \displaystyle\int_0^x dt = \int_{\mathbb{R}_+} \mathbb{1}_{(t\,;+\infty)}(x)\,dt$ dando $\mathrm{E}[X] = \displaystyle\int_{\mathbb{R}} \left(\int_{\mathbb{R}_+} \mathbb{1}_{(t\,;+\infty)}(x)\,dt\right) dP_X(x) = \int_{\mathbb{R}_+} \left(\int_{\mathbb{R}} \mathbb{1}_{(t\,;+\infty)}(x)\,dP_X(x)\right) dt$ por el teorema de Fubini Th. 1-5 pagina 26. Se cierra la prueba observando que la integral interior es nada más que $P(X > t)$. En el caso discreto con $\mathcal{X} = \mathbb{N}$, viene inmediatamente $\displaystyle\sum_{n\in\mathbb{N}} P(X > n)$ que podemos probar directamente saliendo de $P(X = n) = P(X > n) - P(X > n-1)$. En el caso de variable admitiendo una densidad, se obtiene también haciendo una integración por partes [18]

Esta fórmula se aplica al ejemplo 1-4 que tratamos:

**Ejemplo 1-14** (Continuación del ejemplo 1-4)**.** *Sea $X = V\,\mathbb{1}_{U<\frac{1}{2}} + \mathbb{1}_{U\geq\frac{1}{2}}$ con $U$ y $V$ variables aleatorias independientes de distribución uniformas sobre $[0\,;\,1)$. Obtuvimos pagina 37 $F_X(x) = \frac{x}{2}\mathbb{1}_{[0\,;\,1)}(x) + \mathbb{1}_{[1\,;+\infty)}(x)$. A continuación, reobtenemos* $\mathrm{E}[X] = \displaystyle\int_0^1 \left(1 - \frac{x}{2}\right) dx = \frac{3}{4}$.

.

Terminamos esta sección con la propiedad de linealidad de la esperanza matemática $\mathrm{E}$, como consecuencia de la linealidad de la integración y definición de la distribución marginal: para cualquier conjunto de vectores aleatorios $\{X_i\}$ integrables y cualquieras matrices $\{C_i\}$ dadas de dimensiones compatibles con las de $X$ (incluyendo el caso escalar),

$$\mathrm{E}\left[\sum_i C_i X_i\right] = \sum_i C_i\,\mathrm{E}\left[X_i\right]$$

(la integrabilidad de la suma se proba a partir de la desigualdad triangular).

### 1.4.2 Momentos de un vector aleatorio

Si $X$ es una variable aleatoria, para cualquier función medible $f$, $f(X)$ también lo es. Se puede entonces definir su valor medio, si existe. A pesar de necesitar evaluar la distribución de probabilidad de $Y = f(X)$, el valor medio se calcula a partir del de $X$:

---

[18] El el casi discreto, hay que tener precauciones separando la series de una diferencia de terminos. En el caso $X$ continuo admitiendo una densidad, hay que estudiar bien el coomportamiento de $t \mapsto t(1 - F_X(t))$ al infinito.

**Teorema 1-14** (Teorema de transferencia)**.** *Sea $X$ un vector aleatorio $d$-dimensional y $f : \mathbb{R}^d \mapsto \mathbb{R}^{d'}$ una función medible tal que $f(X)$ sea integrable. Entonces*

$$\mathrm{E}\left[f(X)\right] = \int_{\Omega} f(X(\omega))\, dP(\omega) = \int_{\mathbb{R}^{d'}} f(x)\, dP_X(x).$$

*En particular, en el caso $\mathcal{X} = X(\Omega)$ discreto se obtiene*

$$\mathrm{E}\left[f(X)\right] = \sum_i f(x_i) P(X = x_i)$$

*y para $X$ continuo admitiendo una densidad de probabilidad*

$$\mathrm{E}\left[f(X)\right] = \int_{\mathbb{R}^d} f(x)\, p_X(x)\, dx.$$

*Demostración.* Sea $B \in \mathcal{B}(\mathbb{R}^d)$ y consideramos $f(x) = \mathbb{1}_B(x)$. Entonces, $\mathcal{Y} = \{0\,,\,1\}$ y inmediatamente

$$P_Y = P_X(B)\, \delta_1 + (1 - P_X(B))\, \delta_0.$$

Entonces

$$\mathrm{E}[f(X)] = \int_{\mathbb{R}} P_X(B)\, d\delta_1 + \int_{\mathbb{R}} (1 - P_X(B))\, d\delta_0 = P_X(B) = \int_{\mathbb{R}^d} \mathbb{1}_B(x)\, dP_X(x).$$

En el caso $d' = 1$, para $f \geq 0$, se cierra entonces la prueba usando el teorema 1-2 pagina 25, escribiendo $f$ como límite creciente de una sucesión de funciones escalonadas, y la definición Def. 1-10 de la integración real. El caso $d' > 1$ es nada mas que $d' = 1$, componente a componente. □

De manera general, estas medias son llamadas *momentos* de la variable aleatoria $X$. Los momentos relevantes usuales son los siguientes:

- para el "monomio" $f(x) = x^{\otimes r}$ producto tensorial de $x$ $r$ veces [19] siendo $r \in \mathbb{N}^*$, se obtiene el tensor de los $r$*-ésimo momentos (ordinarios)* de $X$:

  $$m_r \equiv \mathrm{E}\left[X^{\otimes r}\right] = \int_{\mathbb{R}^d} x^{\otimes r}\, dP_X(x)$$

  que tiene unidades de $\prod_j X_{i_j}$ (de $X_i^r$ si los componentes de $X$ tienen la misma "unidad"). Se escribe también

  $$m_{r_1,\ldots,r_d} = \mathrm{E}\left[\prod_{i=1}^{d} X_i^{r_i}\right] \quad \text{con} \quad \sum_i r_i = r.$$

  Se puede incluir el caso $r = 0$ con la convención $x^{\otimes 0} = 1$, que corresponde a la condición de normalización: $m_0 = \int_{\mathbb{R}} dP_X(x) = 1$. La media es el primer momento: $m_1 = \mathrm{E}[X] = m_X$. Típicamente, los primeros momentos son más relevantes que los de órdenes mayores, para la caracterización de una distribución. Para $r = 2$, en el caso escalar, el momento de orden 2 es el análogo del momento de inercia de la mecánica.

[19] Recuerdense de que $x \otimes x$ es una matriz teniendo como componentes $x_i x_j$; entonces $x^{\otimes r}$ en un tensor $r$-dimensionale teniendo como componentes $\left[x^{\otimes r}\right]_{i_1,\ldots,i_r} = \prod_j x_{i_j}$.

Por ejemplo, para la distribución uniforme $p_X(x) = \frac{1}{b-a}$ en el intervalo $[a, b]$, resulta $m_r = \frac{b^{r+1}-a^{r+1}}{(r+1)(b-a)}$. En particular, $m_1 = \frac{a+b}{2}$, valor medio del intervalo.

Fijense de que $X^{\otimes r}$ no es siempre integrable, por ejemplo, el el caso con densidad, si $p_X(x)$ tiene soporte (semi)infinito, necesariamente la función $p_X$ debe tender a 0 cuando $|x| \to \infty$, donde $|\cdot|$ denota la norma euclideana. Si $p_X(x)$ 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 $p_X(x) = \frac{\alpha}{(1+(x-x_0)^t R^{-1}(x-x_0))^{\frac{d+1}{2}}}$ sobre $\mathbb{R}^d$, con la matriz cuadrada $R > 0$, $x_0 \in \mathbb{R}^d$ y $\alpha > 0$ coeficiente de normalización, no tiene momentos finitos de orden $r \geq 1$.

- En el caso de variables discretas $X$ sobre $\mathcal{X} = \mathbb{N}$, resulta útil introducir el $r$-ésimo *momento factorial* de $X$ ($r \in \mathbb{N}$) mediante

$$\mathrm{E}\left[X^{(r)}\right] \equiv \mathrm{E}\left[\prod_{k=0}^{r-1}(X-k)\right] = \sum_{n=r}^{\infty} \frac{n!}{(n-r)!}\, P(X=n).$$

(usamos la convención usual $\prod_0^{-1} = 1$).

- Los *momentos centrales* o *cumulantes* se definen alrededor de la media $\mathrm{E}[X]$, *i. e.,*, como el tensor de los $r$-ésimo momentos de la *desviación* $\Delta X \equiv X - \mathrm{E}[X]$:

$$\zeta_r \equiv \mathrm{E}\left[(X - \mathrm{E}[X])^{\otimes r}\right].$$

Se escribe también

$$\zeta_{r_1,\ldots,r_d} = \mathrm{E}\left[\prod_{i=1}^{d}(X_i - \mathrm{E}[X_i])^{r_i}\right] \quad \text{con} \quad \sum_i r_i = r.$$

Se deduce que si la distribución de probabilidad satisface a una simetría central con respecto a la media, *i. e.,* $X - m_X \overset{\mathrm{d}}{=} -(X - m_X)$ donde $\overset{\mathrm{d}}{=}$ significa que los vectores aleatorios tiene la misma distribución de probabilidad, entonces todos los momentos centrales impares son nulos. Los momentos (centrales) brindan medidas que caracterizan la distribución.

1. El primer momento, o media:

$$m_X = \mathrm{E}[X].$$

2. El segundo momento central se conoce como *matriz de covarianza*. En el caso escalar, hablamos de *varianza*, o *dispersión* o también *desviación cuadrática media*.

$$\Sigma_X \equiv \mathrm{Cov}[X] \equiv \zeta_2 = \mathrm{E}\left[(X - m_X)\,(X - m_X)^t\right].$$

En el caso escalar, la varianza se escribe en general

$$\mathrm{Var}[X] = \mathrm{E}\left[(X - m_X)^2\right]$$

y es una medida del cuadrado del ancho efectivo de una densidad de probabilidad (o vector de probabilidad). Para dos componentes $i \neq j$ hablamos de *covarianza entre variables*, y escribimos

$$\mathrm{Cov}[X_i, X_j] = \mathrm{E}\left[(X_i - m_{X_i})\left(X_j - m_{X_j}\right)\right].$$

La matriz de covarianza tiene las varianzas de los $X_i$ en su diagonal, y las covarianzas entre componentes en las componentes no diagonales. Es sencillo ver de que $\mathrm{Cov}[X]$ es simétrica, por construcción, y de que $\mathrm{Cov}[X] \geq 0$ donde $A \geq 0$ significa que la matriz es, definida no negativa (en el caso escalar la varianza es no negativa), con igualdad sólo cuando $P_X = \delta_{x_0}$ para un $x_0$ dado, esto es, cuando no hay incerteza sobre el resultado. De la desigualdad de Cauchy-Bunyakovsky-Schwarz (ver corolario **??**, pagina **??**) se proba sencillamente de que

$$|\mathrm{Cov}[X_i, X_j]|^2 \leq \sigma_{X_i}^2 \sigma_{X_j}^2,$$

así que se define también el *coeficiente de correlación* que es adimensional y toma valores entre $-1$ (variables completamente anticorrelacionadas) y 1 (variables completamente correlacionadas) como:

$$\rho_{ij} = \rho_{ji} \equiv \frac{\mathrm{Cov}[X_i, X_j]}{\sigma_{X_i} \sigma_{X_j}}.$$

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 $\rho_{12} = \frac{a}{|a|} = \pm 1$.

También, se puede ver de que

$$\mathrm{Var}[|X|] = \mathrm{Tr}\, \Sigma_X,$$

$\mathrm{Tr}$ siendo la traza y $|\cdot|$ la norma euclideana de un vector. La covarianza está bien definida si $|X|$ es una variable aleatoria de cuadrado integrable, esto es, cuando $E[|X|^2] < \infty$. Se proba sencillamente (desallorando el "cuadrado" y usando la linealidad de la esperanza) de que

$$\mathrm{Cov}[X] = \mathrm{E}\left[XX^t\right] - m_X m_X^t$$

conocido como *teorema de König-Huygens*. En el caso escalar, es el equivalente del teorema de Huygens de la mecánica relacionando el momento de inertia de un solido con respeto al origen en función del momento de inertia con respeto al centro de masa. Además, inmediatamente,

$$\forall\, A \in \mathbb{R}^{d' \times d},\, b \in \mathbb{R}^d, \quad \mathrm{Cov}[AX + b] = A\, \mathrm{Cov}[X] A^t.$$

En el caso escalar, $d = 1$, lo que es conocido también como el *ancho* de una distribución está dado por la *desviación estándar*

$$\sigma_X = \sqrt{\mathrm{Var}[X]}$$

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 $\frac{\sigma_X}{m_X} = \sqrt{\frac{\mathrm{E}[X^2]}{m_X^2} - 1}$ cuando $m_X \neq 0$.

Dado un vector aleatorio $X$, teniendo en cuenta que los dos primeros momentos dan las características más importantes de la distrubución de probabilidad, puede resultar conveniente hacer una transformación de variable aleatoria a la llamada *variable estándar*: $Y \equiv \Sigma_X^{-\frac{1}{2}} (X - m_X)$, donde $\Sigma^{-\frac{1}{2}}$ es la única matriz simétrica definida positiva tal que su cuadrado es igual a $\Sigma^{-1}$ (Horn & Johnson, 2013; Magnus & Neudecker, 1999) que entonces tiene media igual a 0 y una matriz de covarianza igual al identidad $I$ (en el caso escalar, desviación estándar igual a 1).

3. En el caso escalar, el tercer momento central permite definir el *coeficiente de asimetría* (o skewness en ingles) (Pearson, 1905):

$$\mathrm{Asim}[X] \equiv \gamma_X \equiv \mathrm{E}\left[\left(\frac{X - m_X}{\sigma_X}\right)^3\right] = \frac{\zeta_3}{\sigma_X^3},$$

momento de orden 3 de la variable estandar, que resulta adimensional y puede tener signo positivo o negativo, anulándose para distribuciones que son simétricas respecto del valor medio.

4. En el caso escalar, el cuarto momento central da lugar a la *curtosis* (Pearson, 1905; Westfall, 2014):

$$\mathrm{Curt}[X] \equiv \kappa_X \equiv \mathrm{E}\left[\left(\frac{X - m_X}{\sigma_X}\right)^4\right] = \frac{\zeta_4}{\sigma_X^4},$$

momento de orden 4 de la variable estandar, que posibilita diferenciar entre distribuciones altas y angostas. Veremos más adelante de que para la densidad Gausiana $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-m)^2}{2\,\sigma^2}\right)$, $m_X = m$, $\sigma_X = \sigma$, $\gamma_X = 0$, $\kappa_X = 3$. Se dice de que $p_X$ es alta y angosta, o sub-gausiana, o *con colas livianas* o también platicúrtica cuando $\kappa_X < 3$, y se dice bajas y anchas o sobre-gausiana, o *con colas pesadas* o también leptocúrtica cuando $\kappa_X > 3$ (para $\kappa_X = 3$ la distribución es a veces dicha mesocúrtica). A veces, se define entonces la *curtosis por exceso* $\mathrm{Curt}[X] - 3$. Más que el pico de distribución, la curtosis describe las colas de una distribución (pesadas o livianas) (Westfall, 2014).

Fijense de que, en el contexto escalar $d = 1$, se vinculan los cumulantes y los momentos ordinarios directamente de las definiciones:

$$\zeta_r = \sum_{s=0}^{r} \binom{r}{s} (-m_X)^{r-s}\, m_s$$

para cualquier $r \in \mathbb{N}$, siendo $m_0 = \zeta_0 = 1$. Por ejemplo, $\zeta_2 = m_2 - m_1^2$ que es nada más que la relación de König-Huyggens, mientras que $\zeta_3 = m_3 - 3m_1m_2 + 2m_1^3$. En el contexto multivariado, la relación momentos-cumulantes toma la expresión

$$\zeta_{r_1,\ldots,r_d} = \sum_{s_1=0}^{r_1} \cdots \sum_{s_d=0}^{r_d} \left(\prod_{i=1}^{d} \binom{r_i}{s_i} (-m_{X_i})^{r_i - s_i}\right) m_{s_1,\ldots,s_d}.$$

Tratando de covarianza, más generalmente, para dos vectores aleatorios $X$ e $Y$, se define la matriz de covarianza conjunta como

$$\Sigma_{X,Y} \equiv \mathrm{Cov}[X,Y] = \mathrm{E}\left[(X - m_X)\,(Y - m_Y)^t\right] = \mathrm{E}\left[XY^t\right] - m_X m_Y^t.$$

Esta matriz contiene las covarianzas $\mathrm{Cov}[X_i, Y_j]$.

### 1.4.3 Independencia, identidades y desigualdades

Una primera relación interesante concierna el caso de variables independientes y como se comporta la covarianza de estas:

**Propuesta 1-1.** *Sean $X$ e $Y$ dos vectores aleatorios integrables. Si son independientes, entonces*

$$\mathrm{E}[XY^t] = \mathrm{E}[X]\,\mathrm{E}[Y]^t \qquad \textit{i. e.,} \qquad \mathrm{Cov}[X,Y] = 0.$$

*En particular, para $X$ con componentes independientes, $\mathrm{Cov}[X]$ es una matriz diagonal.*

*Demostración.* Sean $X = \sum_j x_j \mathbb{1}_{A_j}$ e $Y = \sum_k y_k \mathbb{1}_{B_k}$ dos variables escalonadas. Entonces, $A_j = (X = x_j)$ y $B_k = (Y = y_k)$. Luego

$$\begin{aligned}
\mathrm{E}[XY] &= \sum_{j,k} x_j y_k \,\mathrm{E}[\mathbb{1}_{A_j}\mathbb{1}_{B_k}] \\
&= \sum_{j,k} x_j y_k \,\mathrm{E}[\mathbb{1}_{A_j \cap B_k}] \\
&= \sum_{j,k} x_j y_k P(A_j \cap B_k) \\
&= \sum_{j,k} x_j y_k P(X = x_j) P(Y = y_k) \quad \text{(de la independencia)}
\end{aligned}$$

dando el resultado para variables escalonadas. Se cierra la prueba para variables positivas como límite de crecientes de funciones escalonadas, y variables reales tratando las partes positivas y negativas aparte. El caso vectorial se deduce trabajando con pares de componentes. □

Fijense de que la reciproca es falsa en general:

**Ejemplo 1-15** (Uniforme sobre el disco unitario)**.** *Sea $X = (X_1, X_2)$ uniforme sobre el disco unitario, i.e., $p_X(x) = \frac{1}{\pi}\mathbb{1}_{\mathbb{S}^2}(x)$ con $\mathbb{S}^2 = \{(x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 \leq 1\}$. Claramente, los $X_i$ no pueden ser independientes del hecho de que $\mathcal{X}_i = [-1\,,\,1]$ y $\mathcal{X} \neq \mathcal{X}_1 \times \mathcal{X}_2$ (es estrictamente incluido en el producto cartesiano). Por simetría central de $p_X$, es sencillo ver de que $\mathrm{E}[X_1 X_2] = 0$ y similarmente $\mathrm{E}[X_i] = 0$: a pesar de que los $X_i$ no sean independientes, $\mathrm{Cov}[X_1, X_2] = 0$.*

Esta implicación facilita frecuentemente los calculos de media. Volviendo al ejemplo 1-4 de la pagina 37:

**Ejemplo 1-16** (Continuación del ejemplo 1-4)**.** *Tratando de la media de $X = V\,\mathbb{1}_{U<\frac{1}{2}} + \mathbb{1}_{U\geq\frac{1}{2}}$ con $U$ y $V$ variables independientes de distribución uniformas sobre $(0\,;\,1)$, se calcula gracia a la linealidad y a la independencia, $\mathrm{E}[X] = \mathrm{E}[V]\,\mathrm{E}[\mathbb{1}_{U<\frac{1}{2}}] + \mathrm{E}[\mathbb{1}_{U\geq\frac{1}{2}}] = \frac{1}{2}\times\frac{1}{2} + \frac{1}{2} = \frac{3}{4}$ como lo hemos obtenido usando $P_X$ en la pagina 52 o la positividad en la pagina 53.*

Una otra consecuencia de esta proposición trata de un conjunto de vectores aleatorios $\{X_i\}$ y un conjunto de matrices de dimensiones adecuadas,

$$\mathrm{Cov}\left[\sum_i A_i X_i + B\right] = \sum_i A_i \Sigma_{X_i} A_i^t + \sum_{j\neq i} A_i\,\mathrm{Cov}[X_i, X_j] A_j^t.$$

En particular, en el caso escalar,

$$\mathrm{Cov}\left[\sum_i A_i X_i + B\right] = \sum_i A_i^2\,\mathrm{Var}[X_i] + \sum_{j\neq i} A_i A_j\,\mathrm{Cov}[X_i, X_j].$$

Si los $X_i$ son independientes, entonces las covarianzas conjuntas son nulas así que, respectivamente,

$$\mathrm{Cov}\left[\sum_i A_i X_i + B\right] = \sum_i A_i \Sigma_{X_i} A_i^t \qquad \text{y} \qquad \mathrm{Cov}\left[\sum_i A_i X_i + B\right] = \sum_i A_i^2 \sigma_{X_i}^2.$$

Si el teorema da una implocación de la independencia, de hecho existe una reciproca que toma la forma siguiente:

**Teorema 1-15.** *Sean $X$ e $Y$ dos vectores aleatorios. Son independientes si y sólo si $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$ para todo par de funciones $f$ y $g$, medibles y acotadas de dimensiones adecuadas.*

*Demostración.* Se puede referirse a (Feller, 1971; Jacob & Protters, 2003) para unas pruebas rigurosa. En el caso escalar, el principio consiste a ver $f$ y $g$ como limites de funciones escalonadas. Para $f(x) = \sum_i a_i \mathbb{1}_{A_i}(x)$ y $g(y) = \sum_j b_j \mathbb{1}_{B_j}(y)$ se obtiene $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$ si y sólo si $\sum_{i,j} a_i b_j \left(P((X \in A_i) \cap (Y \in B_j)) - P(X \in A_i)P(Y \in B_j)\right) = 0$. Básicamente, eso debe valer para cualquieras $A_i, B_j$ y $a_i, b_j$, así que el término entre parentesis debe ser cero, lo que es nada más de la definición de la independencia de $X$ e $Y$. El caso vectorial se entiende por pares de componentes. □

Relaciones tambíen muy útiles son conocidas como *Desigualdades de Chebyshev* (Bienaymé, 1853; Tchébichev, 1867; Markov, 1884; Olkin & Pratt, 1958; Ferentinos, 1982; Navarro, 2013; Stellato, Van Parys & Goulart, 2017). 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.

**Teorema 1-16** (Desigualdades de Chebyshev)**.** *Sea un vector aleatorio $d$-dimensional $X$ y una función $g : \mathbb{R}^d \mapsto \mathbb{R}_+$ medible tal que $g(X)$ sea integrable. Entonces,*

$$\forall\, a > 0, \quad P(g(X) \geq a) \;\leq\; \frac{\mathrm{E}[g(X)]}{a}.$$

*Demostración.* Sea $\mathcal{D}_a = \{x \in \mathcal{X} :\, g(x) \geq a\} \subset \mathcal{X}$. Entonces, $g$ siendo non negativa,

$$\mathrm{E}[g(X)] = \int_{\mathcal{X}} g(x)\, dP_X(x) \geq \int_{\mathcal{D}_a} g(x)\, dP_X(x) \geq \int_{\mathcal{D}_a} a\, dP_X(x) = aP(X \in \mathcal{D}_a).$$

Se cierra la prueba notando de que $(X \in \mathcal{D}_a) = (g(X) \geq a)$. □

Existen varias formas similares, que son de hecho casos particulares de estas desigualdades.

**Corolario 1-3** (Bienaymé–Chebyshev)**.** *Sea $X$ un vector aleatorio $d$-dimensional admitiendo una esperanza $m_X$ y una covarianza $\Sigma_X$. Entonces,*

$$\forall\, \varepsilon > 0, \qquad P\left(\left|\Sigma_X^{-\frac{1}{2}}(X - m_X)\right|\right) > \varepsilon \;\leq\; \frac{d}{\varepsilon^2}.$$

Viene del teorema incial aplicado a $\Sigma_X^{-\frac{1}{2}}(X - m_X)$, $g(x) = |x|^2$ y $a = \varepsilon^2$.

**Corolario 1-4** (Markov)**.** *Sea* $X$ *un vector aleatorio y* $\varphi \geq 0$ *una función no decreciente tal que* $\varphi(|X|)$ *sea integrable. Entonces,*

$$\forall\, \varepsilon \geq 0, \quad \textit{tal que} \quad \varphi(\varepsilon) \neq 0, \qquad P(|X| > \varepsilon) \;\leq\; \frac{\mathrm{E}[\varphi(|X|)]}{\varphi(\varepsilon)}.$$

La versión inicial des esta desigualdad trataba de funciones $\varphi(u) = u^r,\ r > 0$. Viene del teorema incial aplicado a $g(x) = \varphi(|x|)$ y $a = \varphi(\varepsilon)$, notando de que $(\varphi(|X|) \geq \varphi(\varepsilon)) = (|X| \geq \varepsilon)$ por la no decrecencia de $\varphi$. El caso anterior (una vez la variable centrada) es nada más que un caso especial.

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, como se lo puede ver en el ejemplo siguiente

**Ejemplo 1-17.** *La desigualdad de Bienaymé–Chebyshev indica que la probabilidad de encontrar una fluctuación superior a* $\varepsilon = 3\sigma_X$*, tres desviaciones estándar alrededor de la media, está por debajo de* $1/9$*; el cálculo para una distribución típica como la Gaussiana,* $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(-\frac{x-m_X)^2}{2}\right)$ *ajusta dicha probabilidad por debajo de* $0{,}003$*.*

Una desigualdad muy importante que usaremos frecuentemente en el cápitulo siguiente, trata de funciones convexa, y del efecto sobre la media de una vector aleatorio.

**Definición 1-35** (Función convexa)**.** *Por definición, una función* $\phi : \mathcal{X} \subset \mathbb{R}^d \mapsto \mathbb{R}$ *con* $\mathcal{X}$ *un convexo, es convexa si para cualquier* $\pi_1 \in [0\,,\,1], \pi_2 = 1 - \pi_1$ *y* $x_1, x_2 \in \mathbb{R}^d$*,*

$$\phi(\pi_1 x_1 + \pi_2 x_2) \leq \pi_1 \phi(x_1) + \pi_2 \phi(x_2).$$

$\phi$ *es dicha estrictamente convexa si la desigualdad es estricta, salvo si* $x_2 = x_1$*.*
*Se puede ver de que si* $\phi$ *es dos veces diferenciable, su matriz Hessiana,* $\mathcal{H}\phi \geq 0$ *donde las componentes de la Hessiana son las derivadas partiales secundas de* $\phi$*,* $\frac{\partial^2 \phi}{\partial x_i \partial x_j}$*.*
*Por recurrencia, para cualquier conjunto* $\{x_i\}_i$ *numerable de elementos de* $\mathcal{X}$ *y reales positivos* $\{\pi_i\}_i$ *tales que* $\sum_i \pi_i = 1$*,*

$$\phi\left(\sum_i \pi_i x_i\right) \leq \sum_i \pi_i \phi(x_i).$$

*Dicho con palabras, la funcción del barycentro (combinación convexa) de los* $x_i$ *es debajo del barycentro de los* $\phi(x_i)$*. Eso es ilustrado el la figura Fig. 1-11.*

Intuitivamente, la media teniendo un sabor de barycentro, se intuye de que la media de $\phi(X)$ va a ser arriva de la función de la media de $X$. Es precisamente el teorema de Jensen [20] (Jensen, 1906; Feller, 1971; Brémaud, 1988; Athreya & Lahiri, 2006; Cohn, 2013):

[20] En (Jensen, 1906) se trata del en el caso discreto y integral; en (Hölder, 1889; Hadamard, 1893) se encuentran las primeras semillas de esta desigualdad, y entre otros (Jessen, 1931a, 1931b; Perlman, 1974; Rudin, 1991) para versiones más generales.

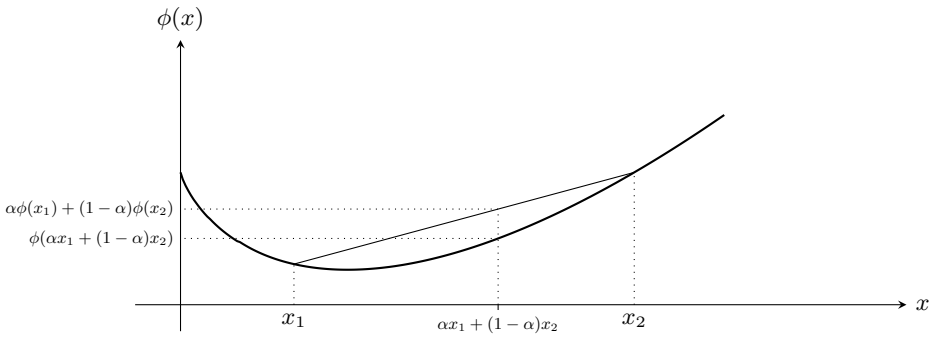



**Figura 1-11:** Ejemplo de función $\phi$ convexa: la cuerda, conteniendo los barycentros de $\{\phi(x_1), \phi(x_2)\}$, es siempre arriba de la curba, *i. e.,* función de los barycentros de $\{x_1, x_2\}$.

**Teorema 1-17** (Desigualdad de Jensen)**.** *Sea $X$ integrable y definida sobre $\mathcal{X} \subset \mathbb{R}^d$, convexo y $f : \mathcal{X} \mapsto \mathbb{R}$. Entonces*

$$\mathrm{E}[\phi(X)] \geq \phi\left(\mathrm{E}[X]\right).$$

*Si $\phi$ es estrictamente convexa, la igualdad se alcanza si y solamente si $X$ es determinista casi siempre.*

*Demostración.* Sea $X = \sum_i x_i \mathbb{1}_{A_i}$ variable escalonada. Entonces $\phi(X) = \sum_i \phi(x_i)\mathbb{1}_{A_i}$, dando

$$\mathrm{E}[\phi(X)] = \sum_i P(A_i)\phi(x_i) \geq \phi\left(\sum_i P(A_i)x_i\right) = \phi\left(\mathrm{E}[X]\right)$$

con igualdad (cuando la convexidad es estricta) si y solamente si todos los $x_i$ son iguales. Se cierra la prueba tomando $X \geq 0$ como limite de sucesión de funciones escalonadas (teorema 1-2, pagina 25), y cualquier $X$ tratando de la parte positiva y negativa (ver pagina 25). El caso vectorial se trata componente a componente para $X$ en termino de limite. Tomando el limite, la condición $x_i$ todos iguales vuelve "casi todos" los $x_i$ deben ser iguales, *i. e.,* $X$ debe ser constante casi siempre. □

Terminamos esta sección con una desigualdad también muy útil, y conocida en los espacios de Hilbert, conocida como *desigualdad de Hölder* (?, ?):

**Teorema 1-18** (Desigualdad de Hölder)**.** *Sean $X$ e $Y$ dos vectores aleatorios $d$-dimensionales y $r > 1$ real. $r^* > 1$ tal que $\frac{1}{r} + \frac{1}{r^*}$ es llamado conjugado de Hölder de $r$, y*

$$\left|\mathrm{E}\left[X^t Y\right]\right| \leq \mathrm{E}\left[\|X\|_r^r\right]^{\frac{1}{r}} \mathrm{E}\left[\|Y\|_{r^*}^{r^*}\right]^{\frac{1}{r^*}}$$

*donde $\|x\|_r = \left(\sum_i x_i^r\right)^{\frac{1}{r}}$ denota la norma $r$ de un vector ($\|\cdot\|_2 \equiv |\cdot|$). Se obtiene la igualda si y solamente si existe un $\lambda$ tal que $X = \lambda Y$ casi siempre.*

*Demostración.* Obviamente, $|\mathrm{E}[X^t Y]| \leq \mathrm{E}[|X^t Y|]$. Luego, de la convexidad de la función $-\log$ se obtiene la desigualdad $\log(|ab|) = \frac{1}{r}\log|a|^r + \frac{1}{r^*}\log|b|^{r^*} \leq \log\left(\frac{|a|^r}{r} + \frac{|b|^{r^*}}{r^*}\right)$ con igualdad si y solamente si $aesproporcionalab$. Aplicado a las componentes de dos vectores $a$ y $b$ se obtiene la desigualdad de Young

$|a^t b| \leq \frac{\|a\|_r}{r} + \frac{\|b\|_{r^*}^{r^*}}{r^*}$ con igualdad si y solamente si los vectores son proporcional. A continuación, denotando

$$\widetilde{X} = \frac{X}{\mathrm{E}\left[\|X\|_r^r\right]^{\frac{1}{r}}} \qquad \text{y} \qquad \widetilde{Y} = \frac{Y}{\mathrm{E}\left[\|Y\|_{r^*}^{r^*}\right]^{\frac{1}{r^*}}}$$

tenemos

$$\mathrm{E}\left[\left|X^t Y\right|\right] = \mathrm{E}\left[\|X\|_r^r\right]^{\frac{1}{r}} \; \mathrm{E}\left[\|Y\|_{r^*}^{r^*}\right]^{\frac{1}{r^*}} \; \mathrm{E}\left[\left|\widetilde{X}^t \widetilde{Y}\right|\right].$$

De la desigualdad de Young, se obtiene entonces

$$\mathrm{E}\left[\left|\widetilde{X}^t \widetilde{Y}\right|\right] \leq \frac{\mathrm{E}\left[\left\|\widetilde{X}\right\|_r^r\right]}{r} + \frac{\mathrm{E}\left[\left\|\widetilde{Y}\right\|_{r^*}^{r^*}\right]}{r^*} = \frac{1}{r} + \frac{1}{r^*} = 1,$$

lo que cierra la prueba. □

Un corolario es conocido como desigualdad de Cauchy-Bunyakovsky-Schwarz [21] para $p = \frac{1}{2}$:

**Corolario 1-5** (Desiguladad de Cauchy-Bunyakovsky-Schwarz)**.** *Sean $X$ e $Y$ dos vectores aleatorios $d$-dimensionales. Entonces*

$$\left|\mathrm{E}\left[X^t Y\right]\right|^2 \leq \mathrm{E}\left[|X|^2\right] \mathrm{E}\left[|Y|^2\right]$$

*(recuersense de que $|\cdot| \equiv \|\cdot\|_2$). Se obtiene la igualda si y solamente si existe un $\lambda$ tal que $X = \lambda Y$ casi siempre.*

Nota: se puede probar esta desigualdad considerando el polinomio $\mathrm{E}\left[|\lambda X + Y|^2\right] \geq 0$, del secundo orden en $\lambda$. Siendo no negativa para cualquier $\lambda$ el discriminente debe ser no positivo, conduciendo a la desigualdad.

De hecho, se puede ver $\mathrm{E}[X^t Y]$ como un producto escalar entre variables aleatorias. La sola subtileza es que $\mathrm{E}[|X|^2] = 0$ conduce a $X = 0$ casi siempre, *i. e.,* se puede tener $X \neq 0$ pero con medida de probabilidad igual a cero (ej. puntos $\omega$ "aislados" en el contexto continuo).

**Cambio de variable multivariado: citar Mukhopadhyay**

**Momento factorial: introducir el simbolo de Pockhammer $(x)_r = \prod_{k=0}^{r-1}, \quad r \geq 0$ con la convención $\prod_{k=0}^{-1} = 1$. Sencillamente, para $n \in \mathbb{N}$, si $n < r$ tenemos $(n)_r = 0$ y si $n \geq r$ tenemos $(n)_r = \frac{n!}{(n-r)!}$.**

**Momentos factoriales en el caso multivariado**

**poner el label "Th:MP:IndependenciaMomentos" al teorema de independencia con momentos**

**Hablar de esperanza condicional**

## 1.5 Funciones generadoras

[21] Esta desigualdad, fue probada por Cauchy para sumas en 1821 (Cauchy, 1821), para integrales por Bunyakovsky en 1859 (Bouniakowsky, 1859) y más elegamente por Schwarz en 1888 (Schwarz, 1888) en un enfoque más general. Ver también (Steele, 2004).

Como lo hemos visto, un vector aleatorio es completamente definida por su medida de probabilidad $P$, o equivalemente por la medida imagen $P_X$, o a través de la función de repartición $F_X$. Sin embargo, bajo el impulso de Laplace en el siglo XVII (entre otros), se introdujo caracterizaciones alternativas a través de transformaciones de la medida de probabilidad, conocidas como *funciones generadoras* o *funciones generatrices* [22] (de Laplace, 1820). Existen varias funciones, cuyas tienes propiedades particulares que vamos a ver en las subsecciones siguientes. Entre otros, estas funciones dadas como valores de expectación de funciones de la variable aleatoria (discreta o continua), con un parámetro real o complejo, permiten hallar fácilmente los distintos momentos de una distribución de probabilidad.

### 1.5.1 Función generadora de probabilidad

De manera general, siguiendo el enfoque de A. de Moivre (ver nota de pie 22) dada una sucesión $a_n, \quad n \in \mathbb{N}$, se define la función generadora dicha *ordinaria* de la sucesión como $G(\{a_n\}_{n\in\mathbb{N}}\,,\, z) = \sum_{n\in\mathbb{N}} z^n$. A veces, esta serie es conocida como transformada en $z$ de la sucesión $\{a_n\}_{n\in\mathbb{N}}$. Tratando de variables aleatorias discretas sobre $\mathbb{N}$, con $p_n = P_X(n) = P(X = n)$, se puede definir así la función generadora asociada a la sucesión $p_n$ y se puede ver que no es nada más que el momento $\mathrm{E}\left[z^X\right]$. De manera general, la función generadora de probabilidad se define de la manera siguiente (Feller, 1968; Johnson, Kotz & Balakrishnan, 1997; Mukhopadhyay, 2000; Athreya & Lahiri, 2006):

**Definición 1-36** (función generadora de probabilidad o de momentos factoriales)**.** *Sea* $X = \begin{bmatrix} X_1 & \cdots & X_d \end{bmatrix}^t$ *vector aleatorio* $d$ *dimensional definido sobre* $\mathcal{X} \subset \mathbb{R}^d$*. La función definida por*

$$G_X(z) = \mathrm{E}\left[\prod_{i=1}^{d} z_i^{X_i}\right] \quad \textit{con} \quad z = \begin{bmatrix} z_1 & \cdots & z_d \end{bmatrix}^t \in \mathbb{C}^d$$

*es conocida como* función generadora de probabilidad *o* función generadora de momentos factoriales *de* $X$*.*

Cuando los $X_i$ son todas variables aleatorias positivas, esta integral converge uniformamente en una bola $d$-dimensional [23] de rayo $r \geq 1$ $\mathbb{B}_d(r) = \left\{z \in \mathbb{C}^d : \|z\| \leq r\right\}$ donde $\|z\|^2 = \sum_i |z_i|^2$. En el caso contrario, puede resultar delicado en termino de existencia de esta función (convergencia de la integral); se puede por ejemplo que, para una variables, no existe ninguno dominio de existencia de esta función.

La denominación *generadora de probabilidad* (pgf para *probability generating function* en ingles) se entiende de sencillamente del hecho siguiente:

---

[22] De hecho, de manera general, se introdujeron tales funciones en un marco más general, asociado a sucesiones de números, bajo el impulso de A. de Moivre (de Moivre, 1730); ver también (Stirling, 1730; Euler, 1741, 1750; de Moivre, 1756) o (Knuth, 1997, Sec. 1.2.9).

[23] Claramente, en $\mathbb{B}_d \equiv \mathbb{B}_d(1)$ tenemos $\left|\int_{\mathbb{R}_+^d} \prod_{i=1}^{d} z_i^{x_i}\, dP_X(x)\right| \leq \int_{\mathbb{R}_+^d} \left|\prod_{i=1}^{d} z_i^{x_i}\right| dP_X(x) \leq \int_{\mathbb{R}_+^d} dP_X(x) = 1$. Además, si la integral converge uniformamente para un $z$ tal que $|z| = r$, por el mismo enfoque se prueba que converge en $\mathbb{B}_d(r)$.

**Lema 1-4.** *Cuando* $\mathcal{X} = \mathbb{N}^d$ *para cualquier* $k = \begin{bmatrix} k_1 & \dots & k_d \end{bmatrix}^t \in \mathbb{N}^d$, *con* $K = \sum_{i=1}^d k_i$

$$\frac{1}{\prod_{i=1}^d k_i!} \left. \frac{\partial^K G_X}{\partial z_1^{k_1} \dots \partial z_d^{k_d}} \right|_{z=0} = P_X(k) = P(X = k)$$

*Demostración.* Se puede escribir la función $G_X$ bajo su forma de generadora ordinaria $G_X(z) = \sum_{n \in \mathbb{N}^d} \left( \prod_{i=1}^d z_i^{n_i} \right) P(X = n)$ con $n = \begin{bmatrix} n_1 & \dots & n_d \end{bmatrix}^t$. A continuación, se nota que la serie converge uniformamente por lo menos en la bola $\mathbb{B}_d \equiv \mathbb{B}_d(1)$, probando que $G_X$ es diferenciable en $\mathbb{B}_d$, así que se puede ver esta series como el desarollo de Taylor de $G_X$ (o, equivalentemente, diferenciar bajo la suma y tomar la derivada en $z = 0$), lo que cierra la prueba. □

De este resultado, se puede notar que, en el caso discreto, hay una relación uno-a-uno entre la medida de probabilidad $P_X$ y la función generadora de probabilidad $G_X$. En el caso continuo **(y más general)**, veremos en la subsección **??** que para $z_j$ de la forma $z_j = e^{\imath u_j}$ con $u_j \in \mathbb{R}$ la transformación se inversa, de manera que se puede recuperar la densidad de probabilidad $p_X$ **(medida de probabilidad $P_X$)** a partir de $G_X$. Dicho de otra manera, como la medida $P_X$, la función $G_X$ caracteriza completamente el vector aleatorio $X$.

Aparece que la función generadora $G_X$ se vincula también con los momentos factoriales, justificando su secunda denominación, *generadora de momentos factoriales* (fmgf para *factorial moments generating function* en ingles):

**Lema 1-5.** *Para cualquier* $k = \begin{bmatrix} k_1 & \dots & k_d \end{bmatrix}^t \in \mathbb{N}^d$ *con* $K = \sum_{i=1}^d k_i$, *derivando* $G_X$ *se proba que, cuando existen* [24]

$$\left. \frac{\partial^K G_X}{\partial z_1^{k_1} \cdots \partial z_d^{k_d}} \right|_{z=1} = \mathrm{E}\left[ (X_1)_{k_1} \dots (X_d)_{k_d} \right]$$

*momento factorial* [25] *de* $X$.

De este resultado, se ve por ejemplo que, cuando existen, se recuperan los momentos de $X$ a través de las derivadas de $G_X$:

- $G_X(1) = 1$, condición de normalización.
- $\nabla_z G_X(1) = \mathrm{E}[X]$ donde $\nabla_z$ indica el gradiente, *i. e.,* el vector de componente $i$-esima $\frac{\partial}{\partial z_i}$.
- $\mathcal{H}_z G_X(1) + \mathrm{diag}\left(\nabla_z G_X(1)\right) = \mathrm{E}\left[X X^t\right]$ donde $\mathcal{H}_z$ es la matrice Hessiana, *i. e.,* la matriz de componente $(i,j)$-esima $\frac{\partial^2}{\partial z_i \partial z_j}$, y $\mathrm{diag}(a)$ es une matriz diagonal de componentes $(i,i)$-esima $a_i$ (vector $a$ sobre la diagonal). Entonces la matriz de covarianza es dada por $\mathrm{Cov}[X] = \mathcal{H}_z G_X(1) + \mathrm{diag}\left(\nabla_z G_X(1)\right) - \nabla_z G_X(1) \nabla_z^t G_X(1)$.

[24] En el caso extremo, el rayo de convergencia de la serie dando $G_X$ es igual a 1, así que no hay garantia que las derivadas en $z = 1$ existen.

[25] Recuerdense que $(x)_r = \prod_{k=0}^{r-1}(x-k), \quad r \geq 0$ símbolo de Pockhammer; ver pagina **??**

La función $G_X$ tiene unas propiedades permitiendo por ejemplo de manejar sencillamente distribuciones de probabilidades de combinaciones lineales de vectores aleatorios independientes, como lo vamos a ver a través del teorema siguiente.

**Teorema 1-19.** *Sean $X$ e $Y$ dos vectores aleatorios $d$-dimensionales independientes, $a = \begin{bmatrix} a_1 & \cdots & a_d \end{bmatrix}^t \in \mathbb{R}^d$ y $b = \begin{bmatrix} b_1 & \cdots & b_d \end{bmatrix}^t \in \mathbb{R}^d$. Entonces para cualquier $z = \begin{bmatrix} z_1 & \cdots & z_d \end{bmatrix} \in \mathbb{C}^d$ (donde existen las funciones):*

$$G_{\mathrm{diag}(a)X+b}(z) = \prod_{i=1}^{d} z_i^{b_i} G_X\left(z_1^{a_1}, \ldots, z_d^{a_d}\right),$$

$$G_{X+Y}(z) = G_X(z) G_Y(z)$$

*y para $z \in \mathbb{C}$*

$$G_X\left(z^{a_1}, \ldots, z^{a_d}\right) = G_{a^t X}(z)$$

*Demostración.* El primer resultado es inmediato, escribiendo $z_i^{a_i X_i + b_i} = z_i^{b_i} \left(z_i^{a_i}\right)^{X_i}$. El secundo viene de $z_i^{X_i+Y_i} = z_i^{X_i} z^{Y_i}$ conjuntamente con el teorema **??** con $f(X) = \prod_{i=1}^{d} z_i^{X_i}$ y $g(Y) = \prod_{i=1}^{d} z_i^{Y_i}$. El tercer resultado es consecuencia de $\prod_{i=1}^{d} \left(z^{a_i}\right)^{X_i} = z^{\sum_{i=1}^{d} a_i X_i}$. □

Estos resultados permiten manejar sencillamente la medida de probabilidad de combinaciones lineales de vectores aleatorios independientes y de marginales a través esta función generadora.

De la tercera identidad, se puede hacer un paso más tratando de sumas aleatorias de vectores aleatorios:

**Teorema 1-20.** *Sea $X_n, \quad n \in \mathbb{N}$ una sucesión de vectores aleatorios indepedientes de misma distribución (resp. generadora de probabilidad) $P_X$ (resp. $G_X$) y $N$ una variable definida sobre $\mathbb{N}$, independiente de los $X_n$. Sea el vector aleatorio $S_N = \sum_{n=0}^{N} X_n$. Entonces*

$$G_{S_N}(z) = G_N\big(G_X(z)\big),$$

*Demostración.* **Usando la formula de esperanza total ec. ??, se escribe**

$$\begin{aligned} G_{S_N}(z) &= \mathrm{E}\left[\sum_{n=0}^{N} X_n\right] \\ &= \mathrm{E}\left[\mathrm{E}\left[\sum_{n=0}^{N} X_n \,\middle|\, N\right]\right] \\ &= \mathrm{E}\left[G_X(z)^N\right] \end{aligned}$$

□

### 1.5.2 Función generadora de momentos

Como lo hemos visto, la función generadora de probabilidad permite recucperar los momentos de un vector aleatorio a través de combinaciones de sus derivadas. Con una pequaña modificación, se puede definir una función generada permitiendo recuperar más directamente los momentos, de manera siguiente (Feller, 1968; Johnson et al., 1997; Mukhopadhyay, 2000; Athreya & Lahiri, 2006):

**Definición 1-37** (función generadora de momentos)**.** *La* función generadora de momentos *(mgf para* moment generating function *en ingles) de un vector aleatorio* $d$*-dimensional se define como*

$$M_X(u) = \mathrm{E}\left[e^{z^t X}\right]$$

*para* $u \in \mathbb{C}^d$*.*

De esta definición se nota inmediatamente que

$$M_X(u) = G_X\left(e^u\right) \qquad \text{donde} \qquad e^u = \begin{bmatrix} e^{u_1} & \cdots & e^{u_d} \end{bmatrix}^t$$

Entonces, como $G_X$, la generadora de los momentos caracteriza completamente el vector aleatorio $X$. Además, resuelta de esta relación que cuando los $X_i$ son todas variables aleatorias positivas, $M_X$ es definida en un dominio de $\mathbb{C}^d$ tal que $\Re e\{u_i\} \leq v_i$ donde $\Re e\{\cdot\}$ denota la parte real de un número complejo y los $v_i$ son positivos llamados indices de convergencia. En el caso de variables escalares admitiendo una densidad de probabildad $p_X$, denotando $s = -u$, esta función se interpreta como la transformada (bilateral) de Laplace de $p_X$.

La generadora de los momentos permite recuperar directamente los momentos a través de derivadas, sin hacer combinaciones:

**Lema 1-6.** *Para cualquier* $k = \begin{bmatrix} k_1 & \cdots & k_d \end{bmatrix}^t \in \mathbb{N}^d$ *con* $K = \sum_{i=1}^d k_i$*, derivando* $M_X$ *se proba que, cuando existen*

$$\left.\frac{\partial^K M_X}{\partial u_1^{k_1} \cdots \partial u_d^{k_d}}\right|_{u=0} = \mathrm{E}\left[\prod_{i=1}^d X_i^{k_i}\right] = m_{k_1,\ldots,k_d}$$

*momento de orden* $k$ *de* $X$*.*

En particular, se recuperan

- $M_X(0) = 1$, condición de normalización.
- $\nabla_u M_X(0) = \mathrm{E}[X]$ promedio,
- $\mathcal{H}_u M_X(0) = \mathrm{E}\left[XX^t\right]$, *i. e.,* $\mathrm{Cov}[X] = \mathcal{H}_u M_X(0) - \nabla_u M_X(0) \nabla_u^t M_X(0)$ matriz de covarianza.

Como la función $G_X$, la generadora de los momentos tiene unas propiedades similares a las del teorema 1-19:

**Teorema 1-21.** *Sean* $X$ *e* $Y$ *dos vectores aleatorios* $d$*-dimensionales independientes,* $A$ *una matriz de* $\mathbb{R}^{d' \times d}$ *y* $b = \begin{bmatrix} b_1 & \cdots & b_{d'} \end{bmatrix}^t \in \mathbb{R}^{d'}$*. Entonces para cualquier* $u = \begin{bmatrix} u_1 & \cdots & u_{d'} \end{bmatrix}^t \in \mathbb{C}^{d'}$ *(donde la función existe):*

$$M_{AX+b}(u) = e^{u^t b} M_X\left(A^t u\right),$$

*y para cualquier* $u = \begin{bmatrix} u_1 & \cdots & u_d \end{bmatrix}^t \in \mathbb{C}^d$ *(donde la función existe):*

$$M_{X+Y}(u) = M_X(u) M_Y(u)$$

*Demostración.* Las pruebas siguen punto a punto los mismos pasos que las del teorema 1-19. □

De nuevo, se puede hacer un paso más tratando de sumas aleatorias de vectores aleatorios como en el teorema 1-20:

**Teorema 1-22.** *Sea* $X_n, \quad n \in \mathbb{N}$ *una sucesión de vectores aleatorios indepedientes de misma distribución (resp. generadora de probabilidad)* $P_X$ *(resp.* $M_X$*) e* $N$ *una variable aleatoria definida sobre* $\mathbb{N}$*, independiente de los* $X_n$*. Sea el vector aleatorio* $S_N = \sum_{n=0}^{N} X_n$*. Entonces*

$$M_{S_N}(u) = G_N\big(M_X(u)\big),$$

*Demostración.* El resultado es consecuencia directa del teorema 1-20. □

### 1.5.3 Función característica

Si la función generadora de momentos permite recuperar los momentos de un vector aleatorio, no es definida sobre todo $\mathbb{C}^d$. Sin embargo, cuando $\Re e\{u_i\} = 0$, esta función es siempre definida ($e^u \in \mathbb{S}_d \subset \mathbb{B}_d$ donde $\mathbb{S}_d$ es la hiperespfera $d$-dimensional unitaria). Entonces, una función generadora muy útil que se usa frecuentemente es la de momentos para este tipo de argumentos, lo que es conocida como función característica y que es al final definida sobre $\mathbb{R}^d$ de manera siguiente (Lukacs, 1961; Golberg, 1961; Feller, 1968; Johnson et al., 1997; Mukhopadhyay, 2000; Athreya & Lahiri, 2006):

**Definición 1-38** (función característica)**.** *La* función característica *(cf para* characteristic function *en ingles) de un vector aleatorio* $d$*-dimensional se define como*

$$\Phi_X(\omega) = \mathrm{E}\left[e^{\imath \omega^t X}\right]$$

*para* $\omega \in \mathbb{R}^d$*.*

De esta definición se nota inmediatamente que

$$\Phi_X(\omega) = M_X(\imath\omega) = G_X\left(e^{\imath\omega}\right) \qquad \text{donde} \qquad e^{\imath\omega} = \begin{bmatrix} e^{\imath u_1} & \cdots & e^{\imath u_d} \end{bmatrix}^t$$

De hecho, se puede definir esta función para un argumento complejo, pero es equivalente a volver a la definición de la generadora de momentos.

En su forma general, la función característica se escribe

$$\Phi_X(\omega) = \int_{\mathbb{R}^d} e^{\imath\omega^t x}\, dP_X(x)$$

y es relacionada a la transformada de Fourier-Stieltjes de la medida $P_X$ (Pinsky, 2009, Chap. 5). Cuando $P_X$ admite una densidad $p_X$, la función es una transformada de Fourier usual de la densidad $p_X$, introducida bajo el impulso de Fourier en 1822 para estudiar la difusión del calor (?, ?).

Insistamos sobre el hecho que la importancia de esta función reside en que siempre existe y está bien definida, dado que $\int_{\mathbb{R}^d} \left|e^{\imath\omega^t x}\right| \, dP_X(x) = \int_{\mathbb{R}^d} dP_X(x) = 1$.

Como para las generadoras ya introducidas, la función característica permite recuperar directamente los momentos a través de derivadas:

**Lema 1-7.** *Para cualquier* $k = \begin{bmatrix} k_1 & \cdots & k_d \end{bmatrix}^t \in \mathbb{N}^d$ *con* $K = \sum_{i=1}^d k_i$*, derivando* $\Phi_X$ *se proba que, cuando existen*

$$(-\imath)^K \left.\frac{\partial^K \Phi_X}{\partial\omega_1^{k_1}\cdots\partial\omega_d^{k_d}}\right|_{\omega=0} = \mathrm{E}\left[\prod_{i=1}^d X_i^{k_i}\right] = m_{k_1,\ldots,k_d}$$

*momento de orden* $k$ *de* $X$*.*

En particular, se recuperan

- $\Phi_X(0) = 1$, condición de normalización.
- $-\imath\nabla_\omega M_X(0) = \mathrm{E}[X]$ promedio,
- $-\mathcal{H}_\omega M_X(0) = \mathrm{E}\left[XX^t\right]$, *i. e.,* $\mathrm{Cov}[X] = -\mathcal{H}_\omega M_X(0) + \nabla_\omega M_X(0)\nabla_\omega^t M_X(0)$ matriz de covarianza.

Fijense de que $\Phi_X$ no es siempre diferencial en $\omega = 0$; Por ejemplo, en el caso de la distribución de Cauchy–Lorentz univariada [26] $p_X(x) = \frac{\gamma}{\pi(\gamma^2+(x-x_0)^2)}$ con $\gamma > 0$, resulta $\Phi_X(\omega) = e^{-\imath x_0\omega-\gamma|\omega}$. Esta función está definida para todo $\omega$, pero no es derivable en $\omega = 0$, lo que coincide con el hecho de que no están definidos los momentos para esta densidad de probabilidad.

Como las funciónes $G_X$ y $M_X$, la función característica tiene entre otros propiedades similares a las del teorema 1-23:

**Teorema 1-23.** *Sean* $X$ *e* $Y$ *dos vectores aleatorios* $d$*-dimensionales independientes,* $A$ *una matriz de* $\mathbb{R}^{d'\times d}$ *y* $b = \begin{bmatrix} b_1 & \cdots & b_{d'} \end{bmatrix}^t \in \mathbb{R}^{d'}$*. Entonces para cualquier* $\omega = \begin{bmatrix} \omega_1 & \cdots & \omega_{d'} \end{bmatrix}^t \in \mathbb{R}^{d'}$*:*

$$\Phi_{AX+b}(\omega) = e^{\imath\omega^t b}\Phi_X\left(A^t\omega\right),$$

*y para cualquier* $\omega = \begin{bmatrix} \omega_1 & \cdots & \omega_d \end{bmatrix}^t \in \mathbb{R}^d$*:*

$$\Phi_{X+Y}(\omega) = \Phi_X(\omega)\Phi_Y(\omega)$$

[26] Lo mismo occure en la extensión multivariada (?, ?).

Resumimos algunas otras propiedades importantes de la función característica:

1. $\Phi_X$ es una función continua en $\mathbb{R}^d$ (Pinsky, 2009, Prop. 5.2.1). Eso es una consecuencia del teorema de convergencia dominada (ver teorema **??** pagina **??**).
2. $|\Phi_X(\omega)| \leq 1\Phi_X(0)$: $|\Phi_X(\omega)|$ es máxima en $\omega = 0$. Eso viene directamente de $\left|e^{\imath\omega^t x}\right| = 1$.
3. $\Phi_X(-\omega) = \Phi_X^*(\omega)$: $\Phi_X$ tiene una símetria hermitica.
4. $\Phi_X$ es una función no negativa definida, *i. e.,* para un conjunto arbitrario de $k \geq 1$ números complejos $a_1, \ldots, a_k$ y $k$ vectores de $\mathbb{R}^d$ $w_1, \ldots, w_k$, se cumple

$$\sum_{i,j=1}^{k} a_i^* a_j \Phi_X(w_j - w_i) \geq 0$$

   Dicho de otra manera, la matriz de componente $(i,j)$-esima $\Phi_X(w_j - w_i)$ es a hermitica (símetria hermítica dada por la propiedad anterior, y no negativa definida).

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

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

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:

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

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

**Teorema 1-24.** *(Bochner, Goldberg)....*

Cumulant generating function ....

**hablar de la cota de Chernoff con la pgf?**
**hablar del CLT y prueba**

## 1.6 Algunos ejemplos de distribuciones de probabilidad

*introducción...*

### 1.6.1 Distribuciones de variable discreta

#### 1.6.1.1. Variable con certeza

...

### 1.6.1.2. Ley de Bernoulli

Se denota $X \sim \mathcal{B}(p)$ con $p \in [0\,;\,1]$ y sus características son las siguientes:

| | |
|---|---|
| Dominio de definición | $\mathcal{X} = \{0\,,\,1\}$ |
| Parametro | $p \in [0\,;\,1]$ |
| Distribución de probabilidad | $p_X(1) = 1 - p_X(0) = p$ |
| Promedio | $m_X = p$ |
| Varianza | $\sigma_X^2 = p(1-p)$ |
| Asimetría | $\gamma_X = 0$ |
| Curtosis | $\kappa_X = 0$ |
| Generadora de probabilidad | $G_X(z) = 1 - p + pz$ sobre $\mathbb{C}$ |
| Generadora de momentos | $M_X(u) = 1 - p + pe^u$ sobre $\mathbb{C}$ |
| Función característica | $\Phi_X(\omega) = 1 - p + pe^{\imath\omega}$ |

Su masa de probabilidad y función de repartición son representadas en la figura Fig. 1-12.

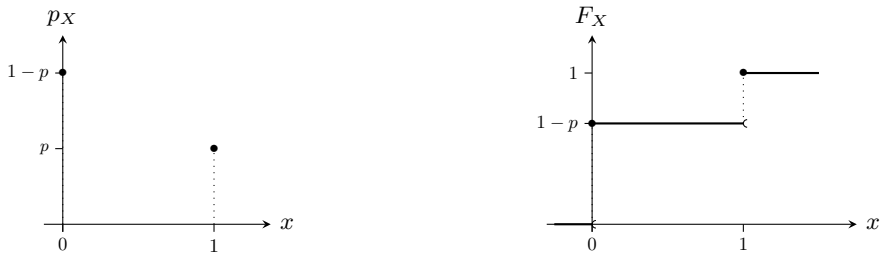



**Figura 1-12:** Ilustración de una distribución de probabilidad de Bernoulli (a), y la función de repartición asociada (b), con $p = \frac{1}{3}$.

Nota que cuando $p = 0$ (resp. $p = 1$) la variable es cierta $X = 0$ (resp. $X = 1$).

### 1.6.1.3. Ley Binomial

Se denota $X \sim \mathcal{B}(n, p)$ con $n \in \mathbb{N} \setminus \{0\,;\,1\}, \quad p \in [0\,;\,1]$ y sus características son las siguientes:

| | |
|---|---|
| Dominio de definición | $\mathcal{X} = \{0\,,\,\ldots\,,\,n\}$ |
| Parametros | $n \in \mathbb{N} \setminus \{0\,,\,1\}, \quad p \in [0\,;\,1]$ |
| Distribución de probabilidad | $p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$ |
| Promedio | $m_X = np$ |
| Varianza | $\sigma_X^2 = np(1-p)$ |
| Asimetría | $\gamma_X = \dfrac{1-2p}{\sqrt{np(1-p)}}$ |
| Curtosis | $\kappa_X = \dfrac{1-6p}{np(1-p)}$ |
| Generadora de probabilidad | $G_X(z) = (1-p+pz)^n$ sobre $\mathbb{C}$ |
| Generadora de momentos | $M_X(u) = (1-p+pe^u)^n$ sobre $\mathbb{C}$ |
| Función característica | $\Phi_X(\omega) = (1-p+pe^{\imath\omega})^n$ |

con $\binom{n}{k} = \dfrac{n!}{k!(n-k)!}$ coefficiente binomial.

Su masa de probabilidad y función de repartición son representadas en la figura Fig. 1-13.

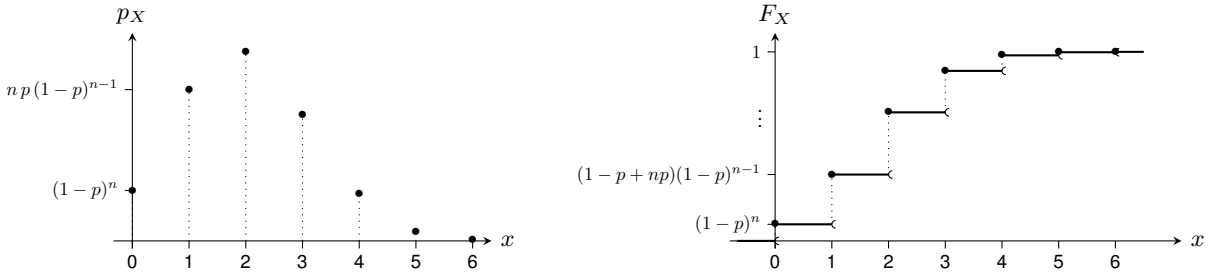



**Figura 1-13:** Ilustración de una distribución de probabilidad Binomial (a), y la función de repartición asociada (b), con $n = 6, \quad p = \frac{1}{3}$.

Cuando $n = 2$, se recupera la lei de Bernoulli $\mathcal{B}(p) \equiv \mathcal{B}(2, p)$. Adémas, se muestra sencillamente usando la generadora de probabilidad que si $X_i \sim \mathcal{B}(p), \quad i = 1, \ldots, n$ independientes, $X = \sum_{i=1}^n X_i \sim \mathcal{B}(n, p)$: esta distribución aparece en el conteo de eventos independientes de misma probabilidad entre $n$.

Nota que cuando $p = 0$ (resp. $p = 1$) la variable es cierta $X = 0$ (resp. $X = n$). Adémas, s

...

Ley uniforme

...

### 1.6.1.4. Ley Geometrica

Se denota $X \sim \mathcal{G}(p)$ con $p \in (0\,;\,1]$ y sus características son las siguientes:

| | |
|---|---|
| Dominio de definición | $\mathcal{X} = \mathbb{N}^*$ |
| Parametro | $p \in (0\,;\,1]$ |
| Distribución de probabilidad | $p_X(k) = (1-p)^{k-1}p$ (convención $0^0 = 1$) |
| Promedio | $m_X = \frac{1}{p}$ |
| Varianza | $\sigma_X^2 = \dfrac{1-p}{p^2}$ |
| Asimetría | $\gamma_X = \dfrac{2-p}{\sqrt{1-p}}$ |
| Curtosis | $\kappa_X = 6 + \dfrac{p^2}{1-p}$ |
| Generadora de probabilidad | $G_X(z) = \dfrac{pz}{1-(1-p)z}$ para $\lvert z\rvert < \frac{1}{1-p}$ |
| Generadora de momentos | $M_X(u) = \dfrac{pe^u}{1-(1-p)e^u}$ para $\Re e\,\{u\} < -\ln(1-p)$ |
| Función característica | $\Phi_X(\omega) = \dfrac{pe^{\imath\omega}}{1-(1-p)e^{\imath u}}$ |

Su masa de probabilidad y función de repartición son representadas en la figura Fig. 1-14.

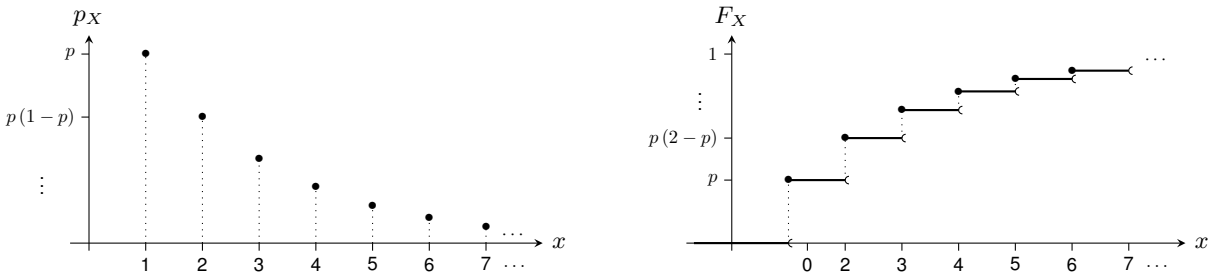



**Figura 1-14:** Ilustración de una distribución de probabilidad Geometrica (a), y la función de repartición asociada (b), con $p = \frac{1}{3}$.

Esta distribución aparece en el conteo de conteo de une repetición de una experiencia de maneja independiente hasta que occure un evento de probabilidad $p$; por ejemplo el número de tiro de un dado equilibriado hasta que occurre un "6" sigue una ley geometrica de parametro $p = \frac{1}{6}$.

Nota que cuando $p = 1$ la variable es cierta $X = 1$.

### 1.6.1.5. Ley de Poisson

Se denota $X \sim \mathcal{P}(\lambda)$ con $\lambda \in \mathbb{R}_+^*$ llamada *taza*, y sus características son las siguientes:

| | |
|---|---|
| Dominio de definición | $\mathcal{X} = \mathbb{N}$ |
| Parametro | $\lambda \in \mathbb{R}_+^*$ |
| Distribución de probabilidad | $p_X(k) = \dfrac{\lambda^k e^{-\lambda}}{k!}$ |
| Promedio | $m_X = \lambda$ |
| Varianza | $\sigma_X^2 = \lambda$ |
| Asimetría | $\gamma_X = \lambda^{-\frac{1}{2}}$ |
| Curtosis | $\kappa_X = \lambda^{-1}$ |
| Generadora de probabilidad | $G_X(z) = e^{\lambda(z-1)}$ para $z \in \mathbb{C}$ |
| Generadora de momentos | $M_X(u) = e^{\lambda(e^u-1)}$ para $u \in \mathbb{C}$ |
| Función característica | $\Phi_X(\omega) = e^{\lambda(e^{i\omega}-1)}$ |

Su masa de probabilidad y función de repartición son representadas en la figura Fig. 1-15.

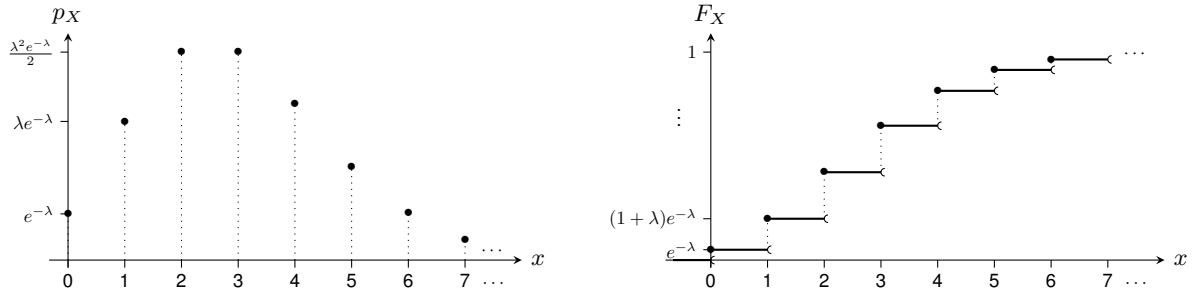



**Figura 1-15:** Ilustración de una distribución de probabilidad de Poisson (a), y la función de repartición asociada (b), con $\lambda = 3$.

Cuando $\lambda = 0$ la variable es cierta $X = 0$ (usando la convención $0^0 = 1$). **Esta distribución aparece...**

...

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

...

Ley multinomial

### 1.6.2 Distribuciones de variable continua

$\sigma \to 0$ **caso cierto**

### 1.6.2.1. Distribución uniforme sobre un producto cartesiano de intervalos

Se denota $X \sim \mathcal{U}(\mathcal{D})$ con, en este caso, $\mathcal{D} = \times_{i=1}^{d}[a_i\,;\,b_i] \subset \mathbb{R}^d$. Se puede escribir $X \overset{\mathrm{d}}{=} a + \mathrm{diag}(b-a)U \overset{\mathrm{d}}{=}$ significando que la equalidad es en distribución (las variables tienen la misma distribución de probabilidda), con $a = \begin{bmatrix} a_1 & \cdots & a_d \end{bmatrix}^d$, $b = \begin{bmatrix} b_1 & \cdots & b_d \end{bmatrix}^d$ y $U \sim \mathcal{U}\left([0\,;\,1]^d\right)$ llamadas *uniforme estandar*. Las caracteristicas de $X \sim \mathcal{U}\left([0\,;\,1]^d\right)$ son las siguientes (se deducen para cualquier uniforme sobre $\mathcal{D}$ por transformación lineal; ver secciones anteriores):

| | |
|---|---|
| Dominio de definición | $\mathcal{D} = [0\,;\,1]^d$ |
| Densidad de probabilidad | $p_X(x) = 1$ |
| Promedio | $m_X = \frac{1}{2}\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}^t$ |
| Covarianza | $\Sigma_X = \frac{1}{12} I$ |
| Asimetría (caso escalar) | $\gamma_X = 0$ |
| Curtosis (caso escalar) | $\kappa_X = -\frac{6}{5}$ |
| Generadora de momentos | $M_X(u) = \prod_{k=1}^{d} \frac{e^{u_k} - 1}{u_k}$ para [27] $u \in \mathbb{C}^d$ |
| Función característica | $\Phi_X(\omega) = (-\imath)^d \prod_{k=1}^{d} \frac{e^{\imath\omega_k} - 1}{\omega_k}$ |

Su densidad de probabilidad y función de repartición son representadas en la figura Fig. 1-16 en el caso escalar.

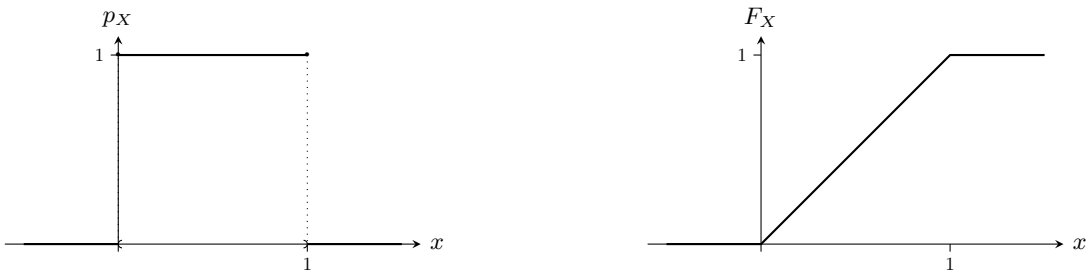



**Figura 1-16:** Ilustración de una densidad de probabilidad uniforme (a), y la función de repartición asociada (b).

**Hacer el caso** $d = 2$

**Esta distribución aparece...**

### 1.6.2.2. Distribución exponencial

Se denota $X \sim \mathcal{E}(\lambda)$ con $\lambda \in \mathbb{R}_+^*$ llamada *taza* (inversa de *escala*), y sus características son las siguientes:

| | |
|---|---|
| Dominio de definición | $\mathcal{X} = \mathbb{R}_+$ |
| Parametro | $\lambda \in \mathbb{R}_+^*$ |
| Densidad de probabilidad | $p_X(x) = \lambda e^{-\lambda x}$ |
| Promedio | $m_X = \dfrac{1}{\lambda}$ |
| Varianza | $\sigma_X^2 = \dfrac{1}{\lambda^2}$ |
| Asimetría | $\gamma_X = 2$ |
| Curtosis | $\kappa_X = 6$ |
| Generadora de momentos | $M_X(u) = \dfrac{\lambda}{\lambda - u}$ para $\Re e\{u\} < \lambda$ |
| Función característica | $\Phi_X(\omega) = \dfrac{\lambda}{\lambda - \imath\omega}$ |

Su densidad de probabilidad y función de repartición son representadas en la figura Fig. 1-17.

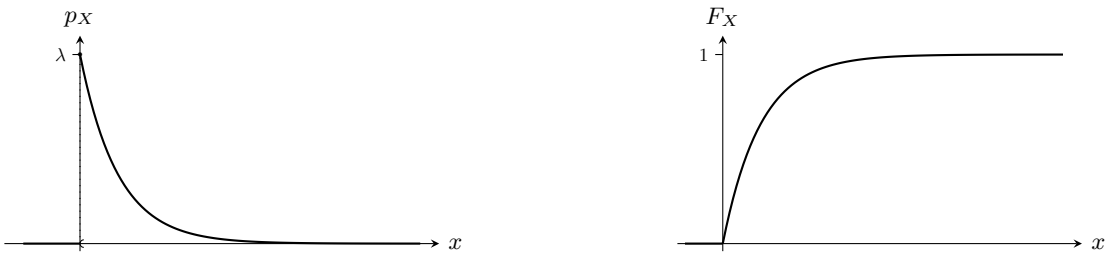



**Figura 1-17:** Ilustración de una densidad de probabilidad exponencial (a), y la función de repartición asociada (b), con $\lambda = 1{,}5$.

Cuando $\lambda \to +\infty$ la variable tiende a una variable cierta $X = 0$. **Esta distribución aparece...**

### 1.6.2.3. Distribución normal o Gaussiana multivariada

Se denota $X \sim \mathcal{N}(m, \Sigma)$ con $m \in \mathbb{R}^d$ y $\sigma$ matriz $d \times d$ símetrica definida positiva. Se puede escribir $X \overset{\mathrm{d}}{=} \Sigma^{\frac{1}{2}} N + m$ con $N \sim \mathcal{N}(0, I)$ donde $I$ es la identidad y $N$ es dicha Gausiana estandar o centrada-normalizada. Las caracteristicas de $X \sim \mathcal{N}(0, I)$ son las siguientes (se deducen para cualquier Gaussiana por transformación lineal; ver secciones anteriores):

| | |
|---|---|
| Dominio de definición | $\mathcal{X} = \mathbb{R}^d$ |
| Densidad de probabilidad | $p_X(x) = \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2}x^t x}$ |
| Promedio | $m_X = 0$ |
| Covarianza | $\Sigma_X = I$ |
| Asimetría (caso escalar) | $\gamma_X = 0$ |
| Curtosis (caso escalar) | $\kappa_X = 0$ |
| Generadora de momentos | $M_X(u) = e^{u^t u}$ para $u \in \mathbb{C}^d$ |
| Función característica | $\Phi_X(\omega) = e^{-\frac{1}{2}\omega^t \omega}$ |

Su densidad de probabilidad y función de repartición en el caso escalar son representadas en la figura Fig. 1-18.

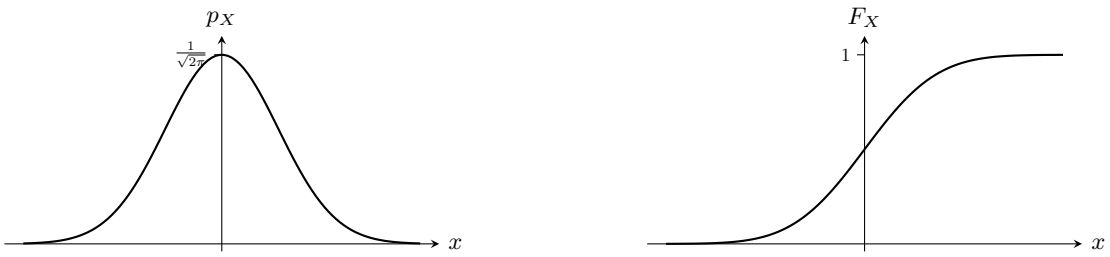



**Figura 1-18:** Ilustración de una densidad de probabilidad gaussiana escalar estandar (a), y la función de repartición asociada (b).

**Esta distribución aparece...**

### 1.6.2.4. Distribución Gamma

Se denota $X \sim \mathcal{G}(\alpha, \beta)$ con $\alpha \in \mathbb{R}_+^*$ llamado *parametro de forma* y $\beta \in \mathbb{R}_+^*$ llamada *taza* (inversa de *escala*). Las caracteristicas son:

| | |
|---|---|
| Dominio de definición | $\mathcal{X} = \mathbb{R}_+$ |
| Parametros | $\alpha \in \mathbb{R}_+^*$ (forma), $\beta \in \mathbb{R}_+^*$ (taza) |
| Densidad de probabilidad | $p_X(x) = \dfrac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$ |
| Promedio | $m_X = \dfrac{\alpha}{\beta}$ |
| Varianza | $\sigma_X^2 = \dfrac{\alpha}{\beta^2}$ |
| Asimetría | $\gamma_X = \dfrac{2}{\sqrt{\alpha}}$ |
| Curtosis | $\kappa_X = \dfrac{6}{\alpha}$ |
| Generadora de momentos | $M_X(u) = \left(1 - \dfrac{u}{\beta}\right)^{-\alpha}$ para $\Re e\{u\} < \beta$ |
| Función característica | $\Phi_X(\omega) = \left(1 - \dfrac{\imath\omega}{\beta}\right)^{-\alpha}$ |

$\Gamma(\alpha) = \int_{\mathbb{R}_+} x^{\alpha-1}\, e^{-x}\, dx$ es la función Gamma (?, ?, ?, ?).

Su densidad de probabilidad y función de repartición son representadas en la figura Fig. 1-19 para varios $\alpha$ y $\beta = 1$.

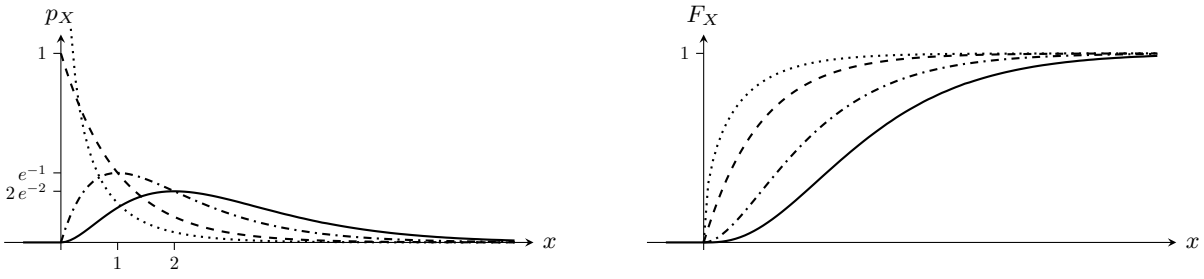



**Figura 1-19:** Ilustración de una densidad de probabilidad gamma (a), y la función de repartición asociada (b). $\beta = 1$ y $\alpha = 0{,}5$ (linea punteada), $1$ (guiones), $2$ (linea mixta) y $3$ (linea llena).

Nota que para $\mathcal{G}(1,\beta) \overset{\mathrm{d}}{=} \mathcal{E}(\beta)$, ley exponencial. Se muestra también sencillamente con las funciones características que para variables independientes

$$\mathcal{G}(a,\beta) + \mathcal{G}(b,\beta) \overset{\mathrm{d}}{=} \mathcal{G}(a+b,\beta)$$

Además, se muestra sencillamente por cambio de variables que

$$\mathcal{N}\left(0,\sigma^2\right)^2 \overset{\mathrm{d}}{=} \mathcal{G}\left(\frac{1}{2}, \frac{1}{2\,\sigma^2}\right)$$

así que que para variables aleatorias $X_i \sim \mathcal{N}\left(0,\sigma^2\right),\ i = 1,\ldots,n,$ independientes, $\sum_{i=1}^n X_i^2 \sim \mathcal{G}\left(\frac{n}{2}, \frac{1}{2\,\sigma^2}\right)$.

**Esta distribución aparece...**

### 1.6.2.5. Distribución Beta

Se denota $X \sim \mathcal{B}e(\alpha, \beta)$ con $(\alpha, \beta) \in \mathbb{R}_+^{*\,2}$ llamados *parametros de forma*. Las caracteristicas son:

| | |
|---|---|
| Dominio de definición | $\mathcal{X} = [0\,;\,1]$ |
| Parametros | $(\alpha, \beta) \in \mathbb{R}_+^{*\,2}$ (forma) |
| Densidad de probabilidad | $p_X(x) = \dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta))}$ |
| Promedio | $m_X = \dfrac{\alpha}{\alpha+\beta}$ |
| Varianza | $\sigma_X^2 = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ |
| Asimetría | $\gamma_X = \dfrac{2\,(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}}$ |
| Curtosis | $\kappa_X = \dfrac{6\,\left((\alpha-\beta)^2(\alpha+\beta+1) - \alpha\beta(\alpha+\beta+2)\right)}{\alpha\,\beta\,(\alpha+\beta+2)\,(\alpha+\beta+3)}$ |
| Generadora de momentos | $M_X(u) = {}_1F_1\,(\alpha, \alpha+\beta; u)$ para $u \in \mathbb{C}$ |
| Función característica | $\Phi_X(\omega) = {}_1F_1\,(\alpha, \alpha+\beta; \imath\omega)$ |

$B(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ es la función beta y ${}_pF_q$ es la función confluent hipergeometrica (?, ?, ?, ?).

Su densidad de probabilidad y función de repartición son representadas en la figura Fig. 1-20 para varios $\alpha$ y $\beta$.

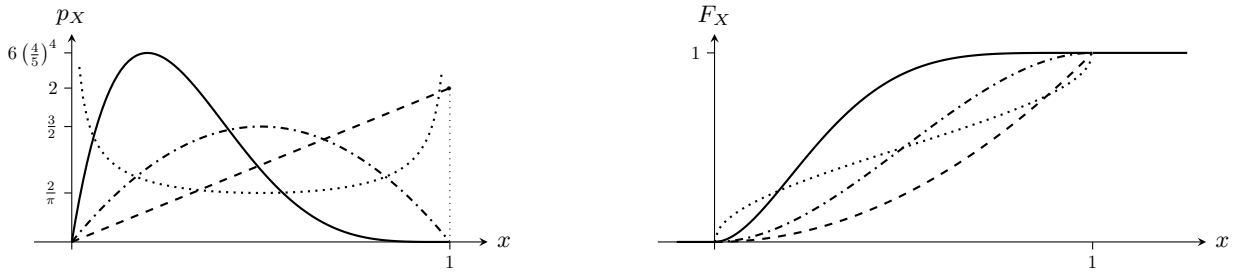



**Figura 1-20:** Ilustración de una densidad de probabilidad beta (a), y la función de repartición asociada (b). $(\alpha, \beta) = (0{,}5\,,\,0{,}5)$ (linea punteada), $(2\,,\,1)$ (guiones), $(2\,,\,2)$ (linea mixta) y $(2\,,\,5)$ (linea llena).

**Arcsin, unif...**

**Nota que para $\mathcal{G}(1,\beta) \overset{\mathrm{d}}{=} \mathcal{E}(\beta)$, ley exponencial. Se muestra también sencillamente con las funciones características que para variables independientes**

$$\mathcal{G}(a,\beta) + \mathcal{G}(b,\beta) \overset{\mathrm{d}}{=} \mathcal{G}(a+b,\beta)$$

**Además, se muestra sencillamente por cambio de variables que**

$$\mathcal{N}\left(0,\sigma^2\right)^2 \overset{\mathrm{d}}{=} \mathcal{G}\left(\frac{1}{2}, \frac{1}{2\,\sigma^2}\right)$$

**así que que para variables aleatorias** $X_i \sim \mathcal{N}\left(0, \sigma^2\right), i = 1, \ldots, n$**, independientes,** $\sum_{i=1}^{n} X_i^2 \sim \mathcal{G}\left(\frac{n}{2}, \frac{1}{2\sigma^2}\right)$**. Esta distribución aparece...**

Teorema del límite central **Ref relaxando la independencia, y versiones con leyes diferentes pero uniformamente acotadas.**

...

**Familia exponencial (Darmois, 1935; Koopman, 1936; Andersen, 1970; Kay, 1993; Lehmann & Casella, 1998; Robert, 2007).; Wishart como ejemplo de 'matrix variate'**

**Familia invariante por rotacion**

**hablar de simulación? Metoto inverso, mezcla, rejeccion, a traves de la condicional para el caso vectorial?**

**hablar del CLT y prueba**

# CAPÍTULO 2
# Nociones de teoría de la información

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

## 2.1 Introducción

La noción de información encuentra su origen con el desarrollo de la comunicación moderna, por ejemplo a través del telégrafo siguiendo la patente de Morse 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 Paris), 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) o (Arndt, 2001, Cap. 3). 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 la manera 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 como 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, íntimamente vinculada a la noción de información, nació bajo el impulso de Claude 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” (Una) por un “The” (La). Desde estos años, las herramientas de dicha teoría de la información dieron lugar a muchas aplicaciones especialmente en comunicación (Cover & Thomas, 2006; Verdu, 1998; Gallager, 2001, y ref.), pero también en otros campos muy diversos tal como la estimación o la discriminación (Cover & Thomas, 2006; Kay, 1993; van den Bos, 2007; Lehmann & Casella, 1998, y ref.), la inferencia estadística (Robert, 2007; Pardo, 2006), el procesamiento de señal o de datos (Phillips & Rousseau, 1992; Ebeling, Molgedey, Kurths & Schwarz, 2000; Basseville, 2013, y Ref.), en ciencias de la ingeneria (Arndt, 2001; Kapur, 1989; Kapur & Kesavan, 1992; Phillips & Rousseau, 1992), física (Arndt, 2001; Ohya & Petz, 1993; Merhav, 2018, y Ref.) entre muchas otras (ver por ejemplo el esquema pagina 2 de (Cover & Thomas, 2006)).

La meta de este capítulo es 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 subyace 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 entropía condicional, lo que va a dar lugar a la noción de información compartida entre dos sistemas o variables aleatorias, concepto fundamental en el marco de la transmisión de información o de mensajes. A continuación, se presentará la noción de entropía relativa a una distribución de probabilidad de referencia, así que el concepto de distancia estadística o divergencia de una distribución con respecto a una referencia. En este capítulo veremos como estas medidas informacionales son entrelazadas a través varias identidades y desigualdades, así que varias relaciones con medidas del mundo de la estimación. Al final, se darán ejemplos y aplicaciones, así que varias generalizaciones de las medidades informacionales.

## 2.2 Entropía como medida de incerteza

### 2.2.1 Entropía de Shannon, propiedades

Uno 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 del 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 Boltzmann de la mecánica estadística, la famosa fórmula $S = k_B \log W$ (Boltzmann, 1896, 1898; Jaynes, 1965; Merhav, 2010, 2018).

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 Boltzmann, misma probabilidad de cada configuración), siendo la información menos el logaritmo de estas probabilidades. Al 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 estar íntimamente vinculada a la probabilidad de aparición de estos símbolos $x$. Siguiendo la idea de Hartley, la información elemental asociada 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) [28].

**Definición 2-39** (Entropía de Shannon)**.** *Sea $X$ una variable aleatoria definida sobre un alfabeto discreto $X(\Omega) = \mathcal{X} = \{x_1, \ldots, x_\alpha\}$ de cardinal $\alpha = |\mathcal{X}| < +\infty$ finito. Sea $p_X$ la distribución de probabilidad de $X$, i. e., $\forall\, x \in \mathcal{X}, \quad p_X(x) = P(X = x)$. La entropía de Shannon de la variable $X$ está definida por*

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

*con la convención $0 \log 0 = 0$ ($\lim_{t \to 0} t \log t = 0$).*

La base del logaritmo es arbitraria; si es $\log_2$ el logaritmo de base 2, $H$ está en unidades binarias o bits (se encuentra tambíen la denominación Shannons), si se usa el logaritmo natural $\ln$, $H$ está en unidades naturales o nats, si es el de base 10, $H$ se da en dígitos decimales o dits (se encuentra tambíen la denominación bans o Hartleys). *En este capítulo, se usará $H$ sin especificar la base del logaritmo. Si es necesario que tenga una base $a$ dada, se denotará la entropía correspondiente $H_a$ y se especificará la base del logaritmo $\log_a$.* Notar que $\log_a x = \frac{\log x}{\log a}$, dando

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

[28] En la misma época que Shannon, independientemente, medidades informacionales aparecieron en cálculos de capacidad de canal 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, y Ref.)).

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 indistintamente tanto la notación $H(p_X)$ como $H(X)$ según lo más conveniente. Además, $p_X$ podrá denotar indistintamente la distribución de probabilidad, o el vector de probabilidad $p_X \equiv \begin{bmatrix} p_X(x_1) & \cdots & p_X(x_\alpha) \end{bmatrix}^t$. En lo que sigue, de vez a cuando, usaremos $p_i \equiv p_X(x_i)$ por simplificación de escritura.

$H$ es el equivalente de la entropía de Gibbs en mecánica estadística. La letra $H$ viene del teorema-H debido a... Ludwig Boltzmann (Jaynes, 1965; Merhav, 2010, 2018).

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

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

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

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

lo que se escribe también $H(\sigma(X)) = H(X)$. En particular, denotando $p_X^\downarrow$ el vector de probabilidades obtenido a partir de $p_X$, clasificando las probabilidades en orden decreciente, $p_1^\downarrow \geq p_2^\downarrow \geq \cdots \geq p_\alpha^\downarrow$ donde $p_i^\downarrow$ es la $i$-ésima componente de $p_X^\downarrow$,

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

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

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

A través tal transformación los estados cambian, pero no cambia la distribución de probabilidad vinculada al alfabeto transformado. Tomando el ejemplo de un dado, la incerteza vinculada al dado no debe depender de los 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 $x_j \in \mathcal{X}$ tal que $p_X(x_j) = 1$ y $p_X(x) = 0$ para $x \neq x_j$, *i. e.,* $p_X = \mathbb{1}_j$,

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

En otras palabras, cuando $X$ no es aleatoria, *i. e.,* $X = x_j$, 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 términos positivos vale cero si y solamente si cada término 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.,*

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

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 técnica del Lagrangiano para tomar en cuenta el vínculo (Miller, 2000; Cambini & Martein, 2009), notando $p_i = p_X(x_i)$, hay que minimizar $\sum_i (-p_i \log p_i + \eta\, p_i)$ donde el factor de Lagrange $\eta$ se determinará para satisfacer el vínculo. Se obtiene sencillamente que $\log p_i = -\eta$, dando la distribución uniforma.

La figura Fig. 2-21 representa la entropía de un sistema a dos estados, de probabilidades $p_X = \begin{bmatrix} r & 1-r \end{bmatrix}^t$ (ley de Bernoulli de parametro $r$), entropía a veces dicha *entropía binaria*, en función de $r$. Esta figura ilustra ambas cotas ($r = 1$ o 1, $r = \frac{1}{2}$) así que la invariancia bajo una permutación ($h(r) = H(r, 1-r) = H(1-r, r) = h(1-r)$).

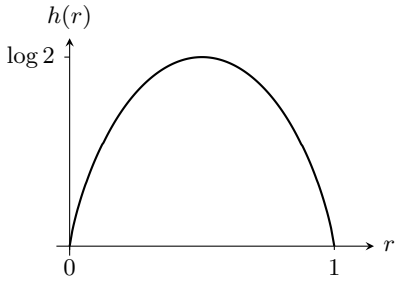



**Figura 2-21:** Entropía binaria (de una variable de Bernoulli) $h(r) = H(r, 1-r)$ en función de $r \in [0, 1]$.

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

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

Esta propiedad es obvia, consecuencia de $\lim\limits_{t \to 0} t \log t = 0$.

[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. De la invarianza de la entropía por permutación, sin perdida de generalidad se puede considerar que los estados que se juntan son los dos últimos, *i. e.,* sean $X$ definido sobre $\mathcal{X}$ y $\breve{X}$ sobre $\breve{\mathcal{X}}$ tales que,

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

entonces,

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

lo que se escribe también

$$H(p_1, \ldots, p_\alpha) = H(p_1, \ldots, p_{\alpha-2}, p_{\alpha-1} + p_\alpha) + (p_{\alpha-1} + p_\alpha)\, H\left(\frac{p_{\alpha-1}}{p_{\alpha-1} + p_\alpha}, \frac{p_\alpha}{p_{\alpha-1} + p_\alpha}\right).$$

Esta relación viene de $a \log a + b \log b = (a+b)\left(\frac{a}{a+b} \log\left(\frac{a}{a+b}\right) + \frac{b}{a+b} \log\left(\frac{a}{a+b}\right) - \log(a+b)\right)$ es ilustrada en la figura Fig. 2-22.

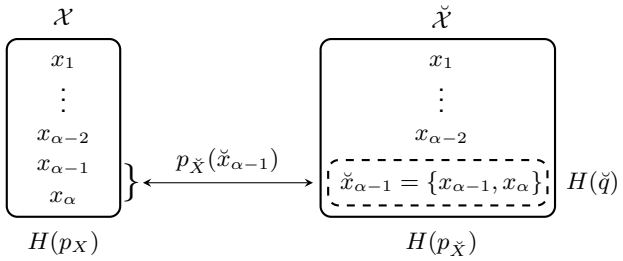



**Figura 2-22:** Ilustración de la propiedad de recursividad, que cuantifica como decrece la entropía en un conjunto cuando se juntan dos estados, relacionando la entropía total, la entropía después del la agrupación y la entropía interna a los dos estados juntados.

[P8] *Concavidad:* la entropía es cóncava ($-H$ es convexa, ver Def. 1-35 pagina 59), 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:

$$\forall\, \{\pi_i\}_{i=1}^n, \quad 0 \leq \pi_i \leq 1, \quad \sum_{i=1}^n \pi_i = 1 \quad \text{and cualquier conjunto de distribuciones} \quad \{p_{(i)}\}_{i=1}^n,$$

$$H\left(\sum_{i=1}^n \pi_i\, p_{(i)}\right) \geq \sum_{i=1}^n \pi_i H(p_{(i)}).$$

Esta relación es conocida también como desigualdad de Jensen (Jensen, 1906). Es una consecuencia directa de la convexidad de la función $\phi : t \mapsto t \log t$, como ilustrado en la figura Fig. 2-23-(a). La figura Fig. 2-23-(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 $\pi_1 = 1 - \pi_2$ de elegir el dado izquierda).

[P9] *Schur-concavidad:* como se lo puede querrer, 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 resuma a partir de la noción de mayorización vista en la definición Def. 1-23, pagina 34 (recuerdense que si los vectores no tienen el mismo tamaño, el más pequeño es completado por ceros; es equivalente

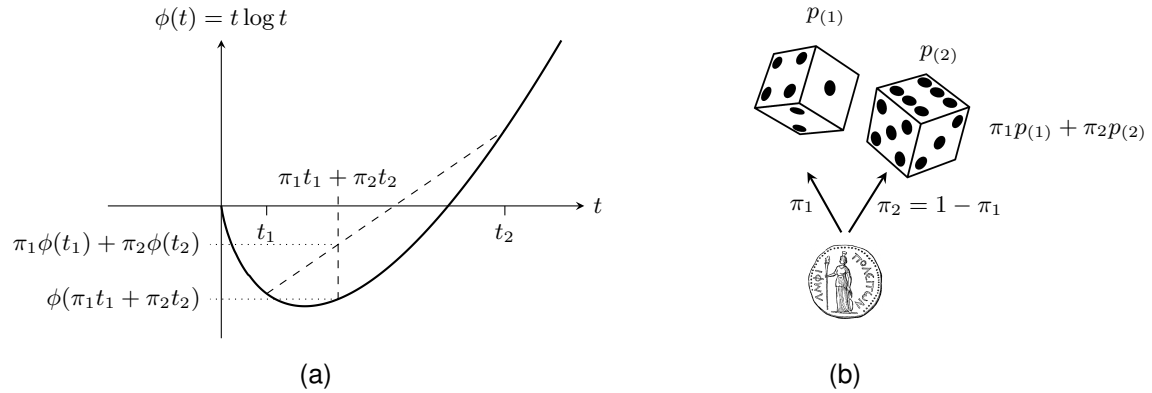



**Figura 2-23:** (a) $\phi(t) = t \log t$ es convexa: la curva es siempre debajo de sus cuerdas; entonces, cada promedio de $\phi(t_1)$ y $\phi(t_2)$ estando en la cuerda juntando estos puntos, queda arriba de la función tomada en el promedio de $t_1$ y $t_2$. Escribiendo eso para (más de dos puntos) sobre los $\sum_i \pi_i \, p_{(i)}(x)$ 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).

a añadir estados fictivos de probabilidad nula, lo que no cambia 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 cualquiera distribución y cualquiera distribución mayoriza la distribución uniforme (Marshall et al., 2011, p. 9, (6)-(8)). Además, de la Schur-concavidad se obtiene que

$$H\left(\begin{bmatrix} \frac{1}{\alpha} & \dots & \frac{1}{\alpha} \end{bmatrix}^t\right) \quad \text{es una función creciente de } \alpha.$$

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 cualquiera función $\phi$ convexa. Basta 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 define naturalmente la entropía para esta variable; tal entropía es llamada *entropía conjunta* de $X$ y $Y$:

**Definición 2-40** (Entropía conjunta)**.** *Sean $X$ e $Y$ dos variables 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., $\forall\, (x, y) \in \mathcal{X} \times \mathcal{Y}, \quad p_{X,Y}(x, y) = P((X = x) \cap (Y = y))$. 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 "⇒" 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

$$H\left(p_X \otimes p_Y\right) = H\left(p_X\right) + H\left(p_Y\right),$$

donde $\otimes$ es el producto de Kronecker [29]. Se generaliza sencillamente a un conjunto de variables aleatorias $\{X_i\}_{i=1}^n$ (o, equivalentemente a un producto de Kronecker 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:

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

Dicho de otra manera, 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 cualesquiera de las entropías individuales

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

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 expansabilidad [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).

---

[29] Recuerdense de que $p_X \otimes p_Y$ es un vector de tamaño $\alpha\beta$ de componente $(i-1)\alpha+j$-ésima $p_X(x_i)\, p_Y(y_j), \quad 1 \leq i \leq \alpha, 1 \leq j \leq \beta$. Se lo puedo ver también como un producto tensorial o externo de $p_X$ definido sobre $\mathcal{X}$ y $p_Y$ definido sobre $\mathcal{Y}$, el producto tensorial siendo definido sobre $\mathcal{X} \times Y$. Ver nota de pie 12 pagina 39.

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

**Definición 2-41** (Tasa de entropía)**.** *Sea $X \equiv \{X_i\}_{i\in\mathbb{N}^*}$ una serie de variables aleatorias, o proceso estocástico. La tasa de entropía del proceso es definida por*

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

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

Se termina esta subsección 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. 2-39. 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 $p(x) \propto \frac{1}{(x+2)\left(\log(x+2)\right)^2}$, $x \in \mathbb{N}$, correctamente normalizada ($\propto$ significa "proporcional a"): $\frac{\log\log(x+2)}{(x+2)\left(\log(x+2)\right)^2} \geq 0$ y la serie $\sum_x \frac{1}{(x+2)\log(x+2)}$ es divergente, así que la serie $-\sum_x p(x)\log p(x)$ diverge.

## 2.2.2 Entropía diferencial

Volviendo a la definición Def. 2-39 de la entropía de Shannon, usando el operador $\mathrm{E}$ promedio estadístico o esperanza matemática, se puede reescribir la entropía de Shannon como $H(X) = \mathrm{E}\left[-\log p_X(X)\right]$. 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 2-42** (Entropía diferencial)**.** *Sea $X$ una variable aleatoria continua admitiendo una densidad de probabilidad $p_X$, definida sobre $\mathbb{R}^d$ y $X(\Omega) = \mathcal{X} = \{x \in \mathbb{R}^d : p_X(x) > 0\} \subseteq \mathbb{R}^d$ el soporte de $p_X(x)$. 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 $X = (X_1,\ldots,X_d)$, 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. 2-24. Se define la variable aleatoria discreta $X^\Delta = \sum_k x_k \mathbb{1}_{(X\in[k\Delta,(k+1)\Delta))}$ sobre $\mathcal{X}^\Delta$ tal que $P(X^\Delta = x_k) = p_{X^\Delta}(x_k) = p_X(x_k)\Delta$. 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

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

lo que se escribe también

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

Entonces, de la integración de Riemann sale que

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

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 término "diferencial" $\log\Delta$.

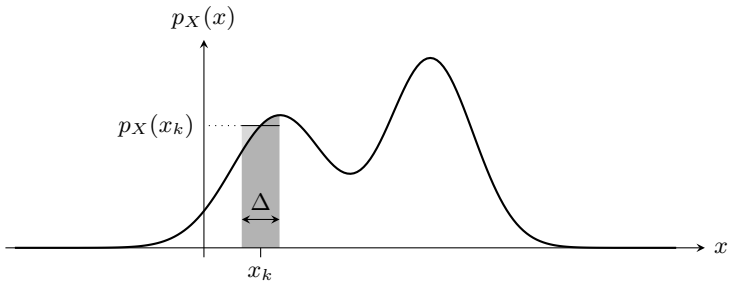



**Figura 2-24:** Densidad de probabilidad $p_X$ de $X$, construcción del alfabeto $\mathcal{X}^\Delta$ donde se define la versión cuantificada $X^\Delta$ de $X$ con su distribución discreta de probabilidad $p_{X^\Delta}$. La superficie en gris oscuro es igual a la superficie definida por el rectángulo 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:

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

donde $\mathrm{J}_g$ es la matriz Jacobiana de la transformación $g:\mathbb{R}^d\mapsto\mathbb{R}^d$ y $|\cdot|$ representa el valor absoluto del determinante de la matriz. Recordando que $p_X(x) = p_Y(g(x))|\,\mathrm{J}_g(x)|$ (ver subsección 1.3.7, pagina 42), se obtiene la propiedad siguiente:

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

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

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

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

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 2-43** (Potencia entrópica)**.** *Sea $X$ una variable aleatoria $d$-dimensional. La potencia entrópica de $X$ es definida por*

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

Por construcción, $N(X)\geq 0$. Además, en el caso continuo, $N(aX+b) = |a|^2N(X)$ (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, sorprendentemente, se conserva la entropía bajo el equivalente continuo del rearreglo.

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

$$H\left(p_X^{\downarrow}\right) = H(p_X).$$

donde $p_X^{\downarrow}$ es el rearreglo símetrico de $p_X$ definido Def. 1-29 pagina 41.

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) [30], y entonces para el caso particular $\phi(t) = t \log t$.

Una pregunta natural es de saber lo que pasa en término de mayorización en el contexto continuo $d$-dimensional. Aparece que la Schur-concavidad [P9] se conserva en el caso continuo, *i. e.,*

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

con la relación de mayorización continua vista Def. 1-30 pagina 41. La desigualdad inversa es probada para cualquier 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 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 ponen vínculos:

[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 $\Sigma_X = \mathrm{E}\left[XX^t\right] - m_X m_X^t$ ($m_X = \mathrm{E}[X]$), la entropía es también acotada por arriba,

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

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 subsección Sec. 2.4.1.

Al final, se conservan obviamente las propiedades de concavidad [P8], de aditividad [P10] y de sub-aditividad [P11]. Es interesante 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).

Como lo hemos visto, la entropía y su versión diferencial no tienen ni las mismas propiedades, ni completamente la misma interpretación. Sin embargo, varias propiedades se comparten y se proban de la misma

[30] 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.

manera. De las escrituras, con una suma o una integral, a veces se encuentra en la literatura la escritura única $\int\!\!\!\!\!\!\sum$ para significar que se usa la suma en el caso discreto, y la integración en el caso continuo con densidad (Rioul, 2007). Sin embargo, volviendo al fin de la subsección 1.3.4, pagina 37 y a la definición Def. 1-15 de una medida discreta sobre $\mathcal{X} = \{x_j\}_j$ dada por $\mu_{\mathcal{X}} = \sum_j \delta_{x_j}$, vimos de que en el caso discreto $p_X$ es la densidad de la medida de probabilidad con respeto a $\mu_{\mathcal{X}}$. Además, de la propiedad $\int_{\mathbb{R}} f(x)\, d\delta_{x_j} = f(x_j)$ se puede ver una suma como una integral con respeto a une medida discreta. De estas observacione, se puede escribir de la misma forma la entropía discreta y diferencial:

**Definición 2-44** (Escritura única de la entropía)**.** *Sea $X$ variable aleatoria definida sobre $\mathcal{X} \subseteq \mathbb{R}^d$, admitiendo una densidad de probabilidad $p_X$ con respeto a una medida $\mu$ (ej. $\mu_{\mathcal{X}}$ en el caso discreto $\mu = \mu_L$ en el caso diferencial). La entropía de $X$ con respeto a $\mu$ se escribe como*

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

*Insistamos en el hecho de que se puede entender esta definición para cualquier $\mu$ y densidad con respeto a $\mu$, que sea discreta, de Lebesgue, o cualquiera.*

## 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 $Y = y$, ¿que información lleva sobre $X$? La respuesta a esta interogación se encuentra en la noción de entropía condicional. Si uno mide $Y = y$, la descripción estadística de $X$ conociendo este $Y = y$ se resuma a la distribución condicional de probabilidad $p_{X|Y=y} = \frac{p_{X,Y}(\cdot,y)}{p_Y(y)}$. Con esta restricción, se puede evaluar una incerteza sobre $X$, sabiendo que $Y = y$,

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

Entonces, condicionalmente a la variable aleatoria $Y$, la incerteza va a ser el promedio estadístico sobre todos los estados $Y$ es decir $H(X|Y) = \int_{\mathbb{R}^d} p_Y(y) H(X|Y = y)\, d\mu(y)$ (con la medida $\mu$ adecuada).

**Definición 2-45** (Entropía condicional)**.** *Sean $X$ e $Y$ dos variables aleatorias, respectivamente $d_X$ y $d_Y$-dimensionales. La entropía condicional de $X$, con respeto a $Y$, es definida por*

$$H(X|Y) = -\int_{\mathbb{R}^{d_X} \times \mathbb{R}^{d_Y}} p_{X,Y}(x,y) \log p_{X|Y=y}(x)\, d\mu(x,y),$$

*con $\mu$ medida adecuada (discreta en el caso discreto o de Lebesgue en el caso diferencial).*

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

[P13]

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

Esta propiedad 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 (o entropía diferencial). Además, de $p_{X,Y}(\cdot, y) = p_{X|Y=y}p_Y(y)$ se deduce la propiedad siguiente

[P14] *Regla de cadena*

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

Esta regla se generaliza sencillamente a

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

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

Siendo $H(X|Y = y)$ una 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 2-46** (Información mutua)**.** *Sean $X$ e $Y$ dos variables aleatorias, la información mutua entre $X$ e $Y$ es la cantidad simétrica*

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

*Se expresa*

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

*con $\mu$ medida adecuada (discreta en el caso discreto o de Lebesgue en el caso diferencial).*

Las diferentes cantidades pueden ser vistas a través de una visión ensemblista, como descrita en la figura Fig. 2-25, diagrama de Venn o de Euler (ver nota de pie 3 pagina 17.

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. En particular, de $I(X; X) = H(X)$ se denoma a veces $H(X)$ *auto información* de $X$.

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 2-47** (Entropía relativa)**.** *La entropía relativa, o divergencia de una medida de probabilidad $Q$, con respeto a una medida de probabilidad de referencia $P$ tal que $Q \ll P$, ambas definidas sobre $\mathbb{R}^d$, es definida como*

$$D_{\mathrm{kl}}(Q \,\|P) = \int_{\mathbb{R}^d} \frac{dQ}{dP}(x) \, \log \left( \frac{dQ}{dP}(x) \right) dP(x) = \int_{\mathbb{R}^d} \log \left( \frac{dQ}{dP}(x) \right) dQ(x).$$

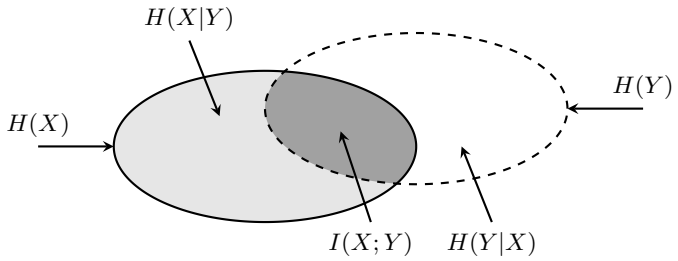



**Figura 2-25:** Diagrama de Venn: Ilustración de la definición de la entropía condicional, de la información mutua, y de las relaciones entre cada medida. La superficia del elipse en linea llena (parte grise) representa $H(X)$ y el interior de la en linea discontinua representa $H(Y)$. La parte grise clara representa $H(X|Y)$ superficia del "conjunto $H(X)$" quitando la parte que partenece a $H(Y)$. La parte blanca representa $H(Y|X)$ superficia del "conjunto $H(Y)$" quitando la parte que partenece a $H(X)$. La parte en grise oscuro es entonces lo que $X$ e $Y$ comparten, es decir $I(X;Y)$.

*Si $P$ y $Q$ admiten una densidad con respeto a una medida $\mu$ (basicamente nos interesamos a $\mu_L$ y $\mu_{\mathcal{X}}$), se escribe a través de las densidades como* [31]

$$D_{\mathrm{kl}}(q\,\|p) \equiv D_{\mathrm{kl}}(Q\,\|P) = \int_{\mathbb{R}^d} \log\left(\frac{q(x)}{p(x)}\right) q(x)\, d\mu(x).$$

Inicialmente, esta medida fue introducida por Kullback y Leibler en la misma linea que Shannon, interpretando $\log\left(\frac{dQ}{dP}(x)\right)$ como una información de discriminación entre dos hipótesis de distribuciones $Q$ y $P$ a partir de la observación $x$, la divergencia siendo la información de discriminación promedia. Introdujeron también una versión simétrica, que veremos más adelante. Se notara de que, $D_{\mathrm{kl}}(Q\,\|P) = -\int_{\mathbb{R}^d} q(x)\,\log(q(x))\,d\mu(x) + \int_{\mathbb{R}^d} p(x)\,\log(q(x))\,d\mu(x)$. El primer termino es nada mas que la entropía de $q$, que se puede ver como distribución "presupuesta". Menos el secundo termino se interpreta como el promedio de la incerteza elemental $\log(q)$ con respeto a la distribución de referencia ("verdadera"), a veces llamado *entropía cruzada*. Por eso, esta divergencia es una entropía relativamente a la distribución $p$. En la misma linea, se puede inmediatamente ver de la definición general Def. 2-44, pagina 90, que $D_{\mathrm{kl}}(Q\,\|P) \equiv -H\left(\frac{dQ}{dP}\right)$ con respeto a la medida $\mu = P$. Por ejemplo, en el caso discreto finito, si $p$ es la distribución uniforme sobre un alfabeto de cardinal $\alpha$, $D_{\mathrm{kl}}(q\,\|p) = \log\alpha - H(q)$, 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 $q$ teniendo la misma matriz de covarianza). *Como para la entropía, cuando se necesitará un logaritmo especificamente de base $a$, se notará la divergencia $D_{\mathrm{kl},a}$.*

[31] 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 $i$-esima 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 perdida de generalidad. En el caso continuo, esta razonamiento no vale más, esta cantidad dependiendo de los estados...

**Lema 2-8** (Positividad de la entropía relativa)**.**

$$D_{\mathrm{kl}}\left(Q\,\|P\right) \geq 0 \quad \textit{con igualdad ssi} \quad P = Q.$$

*Demostración.* Existen varias pruebas, pero la más linda puede ser la usando la desigualdad de Jensen [32], teorema 1-17, pagina 60: para $\phi$ convexa e $Y$ variable aleatoria escalar, $\mathrm{E}[\phi(Y)] \geq \phi(\mathrm{E}[Y])$ con igualdad ssi $Y$ es determinista (casi siempre) si $\phi$ es estrictamente convexa. Sea $X$ de medida de probabilidad $P$. Se escribe la entropía relativa $D_{\mathrm{kl}}\left(Q\,\|P\right) = \mathrm{E}\left[\frac{dQ}{dP}(X)\log\left(\frac{dQ}{dP}(X)\right)\right]$. Sea $Y = \frac{dQ}{dP}(X)$ y $\phi(u) = u\log u$, función estrictamente convexa. Entonces $D_{\mathrm{kl}}\left(Q\,\|P\right) = \mathrm{E}[\phi(Y)] \geq \phi(\mathrm{E}[Y])$. Pero $\mathrm{E}[Y] = \mathrm{E}\left[\frac{dQ}{dP}(X)\right] = \int_{\mathbb{R}^d}\frac{dQ}{dP}(x)\,dP(x) = 1$ según el lema 1-1, pagina 28. Se cierra la prueba con el hecho de que $\phi(1) = 0$ El caso de igualdad apareciendo si y solamente si $Y$ es determinista, es decir $\frac{dP}{dQ}(X)$ determinista, es equivalente a $\frac{dP}{dQ} = 1$ ($P$-c.s.) (constante, la constante siendo igual a 1 del lema 1-1 y $P$ y $Q$ siendo medidas de probabilidad). □

Esta propiedad tiene consecuencias fijandose de que

$$I(X;Y) = D_{\mathrm{kl}}\left(P_{X,Y}\,\|P_X P_Y\right),$$

*i. e.,* la información mutua es la divergencia de Kullback-Leibler de la distribución conjunta relativa al producto de las marginales (obviamente $P_{X,Y} \ll P_X P_Y$). Con este enfoque, no se necesita de que $(X,Y)$ sea discreta o continua admitiendo una densidad.

[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)).

Fijense que si $D_{\mathrm{kl}}$ 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, se mencionará las propiedades adicionales siguientes:

1. $D_{\mathrm{kl}}\left(q\,\|p\right)$ queda invariante bajo una misma transformación biyectiva sobre ambos $p$ y $q$. Es trivial en el caso discreto y si no se prueba sencillamente por un cambio de variables en la forma integral.

[32] En el caso discreto, 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) (Gibbs, 1902; Cover & Thomas, 2006; Rioul, 2007; Merhav, 2010, 2018).

2. $D_{\mathrm{kl}}$ es convexa con respecto al par $(P, Q)$ en el sentido de que, para $\pi_i \geq 0$, $\sum_i \pi_i = 1$, y dos conjuntos $\{P_{(i)}\}_i$, $\{Q_{(i)}\}_i$ de medidas de probabilidades tales que $Q_{(i)} \ll P_{(i)}$,

$$D_{\mathrm{kl}}\left(\sum_i \pi_i\, Q_{(i)} \middle\| \sum_i \pi_i\, P_{(i)}\right) \geq \sum_i \pi_i D_{\mathrm{kl}}\left(Q_{(i)} \middle\| P_{(i)}\right).$$

   La prueba de esta desigualdad es dada subsección Sec. **??**, pagina **??** en un contexto más general.

3. Para $Q$ fijo, $D_{\mathrm{kl}}\left(Q \middle\| P\right)$ es convexa con respecto a $P$ en el sentido de que, para $\pi_i \geq 0$, $\sum_i \pi_i = 1$, y un conjunto $\{P_{(i)}\}_i$ de medidas de probabilidades tales que $Q \ll P_{(i)}$,

$$D_{\mathrm{kl}}\left(Q \middle\| \sum_i \pi_i\, P_{(i)}\right) \geq \sum_i \pi_i\, D_{\mathrm{kl}}\left(Q \middle\| P_{(i)}\right).$$

   Eso es la consecuencia obvia de la concavidad de $u \mapsto \log u$ (escribiendo la divergencia como densidad con respeto a una medida dada). Es sencillo ver de que si $D_{\mathrm{kl}}$ siendo convexa con respecto a $P$ ($Q$ dada) y al par $(P, Q)$, no puede ser convexa con respecto a $Q$ (con $P$ dada).

## 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 desarolló 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; Merhav, 2010, 2018, y ref.)), 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 une infinidad de copias del sistema en el mismo estado macroscópico), es decir su probabilidad de occurencia. 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 [33] (claramente, identificando las

[33] Ver epígrafe del capítulo...

frecuencias a probabilidades de ocurrencia). 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 interactuan (aparte chocandose), 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 subsección Sec. 2.5.4 se va a desarollar 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; Merhav, 2010, 2018)), encontró un echo en varios campos: en inferencia bayesiana para elegir distribuciones del a priori [34] conociendo unos momentos de la ley (Robert, 2007; Jaynes, 1968, 1982; Csiszàr, 1991), hacer estimación espectral o de procesos estocasticos autoregresivos (Burg, 1967, 1975; Jaynes, 1982) o (Cover & Thomas, 2006, cap. 12), entre otros (Arndt, 2001; Kapur, 1989; Kapur & Kesavan, 1992, & ref.).

Sea $X$ variable aleatoria viviendo sobre (de distribución de probabilidad de soporte) $X(\Omega) = \mathcal{X} \subseteq \mathbb{R}^d$ con $K \geq 0$ momentos $\mathrm{E}\left[M_k(X)\right] = m_k$ fijos, con $M_x : \mathcal{X} \to \mathbb{R}$. Suponemos de que $X$ tiene una densidad $p$ con respeto a una medida $\mu$, basicamente $\mu_L$ en el contexto continuo y $\mu_{\mathcal{X}}$ en el caso discreto. El problema de entropía máxima se formula de la manera siguiente: sean $M(x) = \begin{bmatrix} 1 & M_1(x) & \cdots & M_K(x) \end{bmatrix}^t$ y $m = \begin{bmatrix} 1 & m_1 & \cdots & m_K \end{bmatrix}^t$ (si $K = 0$, $M = m = 1$), se busca,

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

donde los dos primeros vínculos (positividad, normalización) aseguran de que $p^*$ sea una distribución de probabilidad. En el ejemplo del gas perfecto, $K = 1$, $M_1(x) = \sum_i x_i^2$ (los $x_i$ son las velocidades). Introduciendo factores de Lagrange $\eta = \begin{bmatrix} \eta_0 & \eta_1 & \cdots & \eta_K \end{bmatrix}^t$ 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)

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

donde $\eta$ será determinado para satisfacer a los vínculos. En el caso continuo $\mu = \mu_L$ se usa la ecuación de Euler-Lagrange (Gelfand & Fomin, 1963; van Brunt, 2004), esquematicamente anulando la "derivada" del integrande con respeto a $p$; en el caso discreto, se anula realmente un gradiente con respeto a los componentes de $p$. Reparametrizando los factores de Lagrange, se obtiene así

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

[34] A partir de una distribución parametrizada por un parámetro $\theta$. El enfoque bayesiano consiste a modelizar $\theta$ aleatorio, digamos $\Theta$, tal que la distribución de observaciones se escribe entonces $p_{X|\Theta=\theta}(x)$. Inferir $\theta$ a partir de observaciones $x$ consiste a determinar la distribución dicha *a posteriori* $p_{\Theta|X=x}(\theta)$. Por eso, hace falta darse una distribución dicha *a priori* $p_\Theta(\theta)$. 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.

con $\eta$ tal que se satisfacen los vínculos de normalización y momentos. Esta distribución cae en la familia exponencial que **hemos visto sección ??**, donde los $M_k$ son las estadísticas suficientes y los $\eta_k$ los parámetros naturales.

Un problema que puede aparecer es que no se puede determinar $\eta$ 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 [35]. En otros terminos, en este caso, el problema no tiene solución [36].

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

**Lema 2-9.** *Sea* $\mathcal{P}_m = \left\{ p \geq 0 : \int_{\mathcal{X}} M(x)\, p(x)\, d\mu(x) = m \right\}$ *conjunto de densidades con respeto a* $\mu$*, con los mismos momentos* $M$*, y sea* $p^* \in \mathcal{P}_m$ *que sea de la forma* $p^*(x) = \mathrm{e}^{\eta^t M(x)}$*. Entonces*

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

*donde* $H$ *es de la definición general Def. 2-44, pagina 90, de la entropía con respeto a* $\mu$*.*

*Demostración.*

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

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

$$\begin{aligned} H(p) &= -D_{\mathrm{kl}}\left(p \,\|p^*\right) - \int_{\mathcal{X}} \eta^t M(x) p(x)\, d\mu(x) \\ &= -D_{\mathrm{kl}}\left(p \,\|p^*\right) - \int_{\mathcal{X}} \eta^t M(x)\, p^*(x)\, d\mu(x) \\ &= -D_{\mathrm{kl}}\left(p \,\|p^*\right) - \int_{\mathcal{X}} p^*(x) \log p^*(x)\, d\mu(x) \\ &= -D_{\mathrm{kl}}\left(p \,\|p^*\right) + H(p^*) \end{aligned}$$

porque $p, p^* \in \mathcal{P}_m$ (segunda linea) y $\eta^t M = \log p^*$ (tercera linea). La prueba se cierra notando que $D_{\mathrm{kl}}\left(p \,\|p^*\right) \geq 0$ con igualdad si y solamente si $p = p^*$ ($\mu$-c.s.). □

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,

---

[35] 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.

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

- 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 subsección 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 como $K = d^2$ momentos), la distribución de entropía máxima es la distribución gausiana de la propiedad [P'5]*b* de la sección Sec. 2.2.2.

Con el enfoque del lema 2-9, se necesita solamente que $p$ sea una densidad con respeto a una medida $\mu$ fija, cualquiera. En particular, si es una medida de probabilidad (de referencia) $\widetilde{P}$, el problema de entropía máxima vuelve ser un problema de minimización de la divergencia de Kullback-Leibler entre $P$ y la medida de referencia, siendo $p$ la densidad con respeto a esta medida (ver definición Def. 2-47). Es decir, tomando $\mu = \widetilde{P}$ medida de probabilidad aparece inmediatamente

$$P^* = \operatorname*{argmáx}_P D_{\mathrm{kl}}\left(P \middle\| \widetilde{P}\right) \quad \text{sujeto a} \quad \int_{\mathcal{X}} M(x)\, dP(x) = m \qquad \Leftrightarrow \qquad \frac{dP^*}{d\widetilde{P}}(x) = \mathrm{e}^{\eta^t x}\,.$$

## 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 concierna la relación que podría tener $X + Y$ con cada variable en término de entropía. La respuesta no es trivial, y el resultado general concierna el caso de variables continuas sobre $\mathbb{R}^d$. Es conocido como desigualdad de la potencia entrópica (EPI para entropy power inequality en inglés). No vincula las entropías, sino que las potencias entrópicas.

**Teorema 2-25** (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 desiguladad (Shannon, 1948; Lieb, 1978; Cover & Thomas, 2006; Dembo et al., 1991; Rioul, 2007):

1. Sean $\widetilde{X}$ y $\widetilde{Y}$ gaussianas independientes de matrices de covarianza proporcionales y tal que $H(\widetilde{X}) = H(X)$ y $H(\widetilde{Y}) = H(Y)$. Entonces
$$N(X+Y) \geq N\left(\widetilde{X} + \widetilde{Y}\right),$$
con igualdad si y solamente si $X$ y $Y$ son gaussianas. De hecho, la primera formulación es equivalente a $N(X+Y) \geq N\left(\widetilde{X}\right) + N\left(\widetilde{Y}\right) = \frac{1}{2\pi\,\mathrm{e}}\left(\left|\Sigma_{\widetilde{X}}\right|^{\frac{1}{d}} + \left|\Sigma_{\widetilde{Y}}\right|^{\frac{1}{d}}\right) \geq \frac{1}{2\pi\,\mathrm{e}}\left|\Sigma_{\widetilde{X}} + \Sigma_{\widetilde{Y}}\right|^{\frac{1}{d}} = N\left(\widetilde{X} + \widetilde{Y}\right)$ (la última desigualdad viniendo de la desigualdad matricial de Minkowski (Hardy et al., 1952; Minkowski, 1910)). Se notará que, de la relación uno-uno entre $H$ y $N$ la desigualdad se escribe también
$$H(X+Y) \geq H\left(\widetilde{X} + \widetilde{Y}\right).$$

2. *Desigualdad de preservación de covarianza:*

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

con igualdad si y solmente si $X$ e $Y$ son gaussianas con matrices de covarianza proporcionales. Claramente, se cumple la igualdad para $\widetilde{X}$ e $\widetilde{X}$, entonces $H\left(\sqrt{a}\, X + \sqrt{1-a}\, Y\right) \geq a\, H(X) + (1-a)\, H(Y) \quad \Leftrightarrow \quad H\left(\sqrt{a}\, X + \sqrt{1-a}\, Y\right) \geq H\left(\sqrt{a}\, \widetilde{X} + \sqrt{1-a}\, \widetilde{Y}\right)$ lo que es nada más que la desigualdad anterior reemplazando $X$ por $\sqrt{a}\, X$ e $Y$ por $\sqrt{1-a}\, Y$ (y vice-versa).

La prueba de esta(s) desigualdad(es) no es trivial. Númeras 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.

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 cualquieras 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).

Se mencionará que existe una versión de la desigualdad de la potencia entrópica con rearreglo (Wang & Madiman, 2004):

**Teorema 2-26** (Desigualdad de la potencia entrópica con rearreglo)**.** *Sean $X$ e $Y$ dos variables $d$-dimensionales continuas independientes de densidades de probabilidades $p_X$ y $p_Y$ respectivamente. Sean $p_X^{\downarrow}$ y $p_Y^{\downarrow}$ los rearreglos de $p_X$ y $p_Y$ respectivamente y denotamos $X^{\downarrow}$ y $Y^{\downarrow}$ vectores independientes de distribución de probabilidad $p_X^{\downarrow}$ y $p_Y^{\downarrow}$ respectivamente. Enconces, Entonces*

$$N(X+Y) \geq N(X^{\downarrow} + Y^{\downarrow}).$$

Se referirá a (Madiman & Barron, 2007, y Ref.) por ejemplo para varias generalizaciones de la desigualdad de la potencia entrópica.

Para cerrar esta secciíon, se mencionará de que en el caso discreto, no hay un resultado general y aún existen contra-ejemplos (Johnson & Yu, 2010, Sec. IV). Existen solamente resultados para variables particulares como para variables binarias (Shamai & Wyner, 1990), leyes binomiales (Harremoës & Vignat, 2003; Sharma, Das & Muthukrishnan, 2011) (ver también (Johnson & Yu, 2010; Haghighatshoar, Abbe & Telatar, 2014)).

### 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 2-48** (Proceso de Markov)**.** *Una secuencia $X_1 \mapsto X_2 \mapsto \ldots \mapsto X_n$ es dicha de Markov si para cualquier $i > 1$,*

$$\forall\, x_i, \quad P_{X_{i-1},X_{i+1}|X_i=x_i} = P_{X_{i-1}|X_i=x_i}\, P_{x_{i+1}|X_i=x_i}.$$

*Dicho de otra manera, condicionalmente a $(X_i = x)$, las variables $X_{i-1}$ y $X_{i+1}$ son independientes. Eso es equivalente a*

$$P_{X_{i+1}|(X_i,X_{i-1},\ldots)=(x_i,x_{i-1},\ldots)} = P_{X_{i+1}|X_i=x_i}.$$

*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 $P_{X_{i+1}|X_i=x_i}$ y la distribución inical $P_{X_1}$ caracterizan completamente el proceso). Es sencillo fijarse de que $X_n \mapsto X_{n-1} \mapsto \ldots \mapsto X_1$ es tambien un proceso de Markov.*

**Teorema 2-27** (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 de que $Z \mapsto Y \mapsto X$ sea también un proceso de Markov, es sencillo probar que $H(X|Y,Z) = H(X|Y)$ (conciendo $Y$ sufice para 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 $\frac{p_{X|(Y,Z)=(y,z)}(x)}{p_{X|Z=z}(x)} = \frac{p_{X|(Y,Z)=(y,z)}(x)\, p_{Y|Z=z}(y)}{p_{X|Z=z}(x) p_{Y|Z=z}(y)} = \frac{p_{X,Y|Z=z}(x,y)}{p_{X|Z=z}(x) p_{Y|Z=z}(y)}$ se nota de que $H(X|Z) - H(X|Y,Z)$ es la divergencia de Kullback-Leibler de $p_{X,Y|Z=z}$ relativamente a $p_{X|Z=z} p_{Y|Z=z}$, 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. □

### 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; Merhav, 2010, 2018, y ref.)).

**Lema 2-10** (Versión informacional de la segunda ley de la termodinámica)**.** *Sea $X_1 \mapsto X_2 \mapsto \cdots \mapsto X_n \mapsto \cdots$ un proceso de Markov, con probabilidades de transición $r_{X_{n+1}|X_n=x_n}$ dadas (independientemente de la condición inicial). Estas últimas modelizan el sistema, independiente de las condiciones iniciales. Sean dos distribuciones (condiciones) iniciales diferentes $p_{X_1}$ y $q_{X_1}$, conduciendo a las distribuciones $p_{X_n}$ y $q_{X_n}$ (con respeto a una medida $\mu$ dada) para $X_n$. Entonces:*

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

  $$D_{\mathrm{kl}}\left(p_{X_{n+1}} \middle\| q_{X_{n+1}}\right) \leq D_{\mathrm{kl}}\left(p_{X_n} \| q_{X_n}\right);$$

  *las distribuciones $p_{X_n}$ y $q_{X_n}$ no se "alejan" (tiende a acercarse);*

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

  $$D_{\mathrm{kl}}\left(p_{X_{n+1}} \| p^*\right) \leq D_{\mathrm{kl}}\left(p_{X_n} \| p^*\right);$$

  *la distribución no se aleja de la distribución estacionaria.*

- *Además, si los $X_n$ tienen $K$ momentos fijos $m = \mathrm{E}[M(X_n)] \quad \forall\, n$ y si $p^*$ es la densidad (con respeto a la medida $\mu$ dada) de entropía máxima tiendo los mismos momentos como descrito subsección Sec. 2.4.1, (ej. $K = 0$, $\mathcal{X}$ de cardinal o volumen finito y ley uniforme, $K = 2$, $M_k(x) = x^k$ y ley gausiana),*

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

  *el sistema tiende a desorganizarse (dando los vinculos/momentos).*

*Demostración.* Se muestra sencillamente que $D_{\mathrm{kl}}\left(p_{X_{n+1},X_n} \middle\| q_{X_{n+1},X_n}\right) = D_{\mathrm{kl}}\left(p_{X_{n+1}} \middle\| q_{X_{n+1}}\right) + \int_{\mathcal{X}_{n+1}} D_{\mathrm{kl}}\left(p_{X_n|X_{n+1}=x_{n+1}} \middle\| q_{X_n|X_{n+1}=x_{n+1}}\right) d\mu(x_{n+1}) = D_{\mathrm{kl}}\left(p_{X_n} \| q_{X_n}\right) +$

$\int_{\mathcal{X}_n} D_{\mathrm{kl}}\left(p_{X_{n+1}|X_n=x_n}\middle\|q_{X_{n+1}|X_n=x_n}\right)\,d\mu(x_n)$. Además, $p_{X_{n+1}|X_n=x_n}=r_{X_{n+1}|X_n=x_n}=q_{X_{n+1}|X_n=x_n}$ (transición independiente de la condición inicial), conduciendo a $D_{\mathrm{kl}}\left(p_{X_{n+1}|X_n=x_n}\middle\|q_{X_{n+1}|X_n=x_n}\right)=0$ con consecuencia de que $D_{\mathrm{kl}}\left(p_{X_n}\|q_{X_n}\right)=D_{\mathrm{kl}}\left(p_{X_{n+1}}\middle\|q_{X_{n+1}}\right)+\int_{\mathcal{X}_{n+1}} D_{\mathrm{kl}}\left(p_{X_n|X_{n+1}=x_{n+1}}\middle\|q_{X_n|X_{n+1}=x_{n+1}}\right)\,d\mu(x_{n+1})$. $p_{X_n|X_{n+1}=x_{n+1}}$ no es necesariamente igual a $q_{X_n|X_{n+1}=x_{n+1}}$, pero la divergencia siendo no negativa, se obtiene la primera desigualdad. La segunda desigualdad se obtiene tomando $q_{X_n}=p^*$. Además, si $p^*$ es le entropía máxima de mismos momentos $m=\mathrm{E}[M(X_n)]$ que los $X_n$, hemos visto de que $p^*(x)=\mathrm{e}^{\eta^t M(x)}$ ley de la familia exponencial, dando $D_{\mathrm{kl}}\left(p_{X_n}\|p^*\right)=-H(X_n)-\eta^t m$, conduciendo a la última desigualdad. □

### 2.4.5 Principio de incerteza entrópico

**Bourret 58, Leipnik 59, Stam59, entre otros que ya citamos un par de veces**

### 2.4.6 Un foco sobre la información de Fisher

Si la entropía y las heramientas 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).

**Definición 2-49** (Matriz información de Fisher parámetrica)**.** *Sea $X$ una variable aleatoria parametrizada por un parámetro $m$-dimensional, $\theta\in\Theta\subseteq\mathbb{R}^m$, de distribución de probabilidad $p_X(\cdot;\theta)$ (con respeto a una medida $\mu$ dada) sobre $\mathcal{X}\subseteq\mathbb{R}^d$ su soporte. Suponga que $p_X$ sea diferenciable con respeto a $\theta$ sobre $\Theta$. La matriz de Fisher, de tamaño $m\times m$, es definida por*

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

*donde $\nabla_\theta=\left[\cdots\ \frac{\partial}{\partial\theta_i}\ \cdots\right]^t$ es el gradiente en $\theta$ y $\log$ el logaritmo natural. Es la matriz de covarianza del score paramétrico $S(X)=\nabla_\theta\log p_X(X;\theta)$ notando que su media es igual a cero (escribiando el promedio y intercambiando integral y gradiente), siendo $\log p_X$ la log-verosimilitud. Bajo condiciones de regularidad, se puede mostrar*[37] *que $J_\theta(X)=-\mathrm{E}[\mathcal{H}_\theta\log p_X(X;\theta)]$ con $\mathcal{H}_\theta$ la Hessiana*[38] *$\mathcal{H}_\theta$ de $\log p_X(X;\theta)$. Nota: a veces se define la información de Fisher como $\mathrm{Tr}(J)$, traza de la matriz información de Fisher.*

[37] 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.

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

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.

En el caso continuo, $\mu = \mu_L$, con una densidad diferenciable, tomando el gradiente en $x$ en lugar de $\theta$ da la matriz de información de Fisher no paramétrica,

**Definición 2-50** (Matriz información de Fisher no paramétrica)**.** *Sea $X$ una variable aleatoria continua admitiendo una densidad de probabilidad $p_X$ definida sobre $\mathcal{X} \subseteq \mathbb{R}^d$ su soporte. Suponga que $p_X$ sea diferenciable con respeto a $x$. La matriz de Fisher no paramétrica, $d \times d$, es definida por*

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

*Es la matriz de covarianza de la función score $\nabla_x \log p_X(X)$ (escribiendo la media y $p_X$ siendo cero el los bordes de $\mathcal{X}$, se ve que el promedio de $\nabla_x \log p_X(X)$ también vale cero) o, bajo condiciones de regularidad, $J(X) = -\mathrm{E}[\mathcal{H}_x \log p_X(X;\theta)]$.*

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 $J(X) > 0$ (matrices definidas positivas). Además,

  $$\forall\ A \text{ matrix no singular}, \quad J(AX) = A^{-t} J(X) A^{-1},$$

  con $A^{-t} = \left(A^{-1}\right)^t = (A^t)^{-1}$ (Cover & Thomas, 2006; Dembo et al., 1991; Barron, 1986). Esta relación da a $J(X)$ un sabor de información en el sentido de que, cuando $A$ es real y 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 $\pi_k \geq 0$, $\sum_{k=1}^K \pi_k = 1$ y cualquier conjunto de distribuciones $p_{(k)}$, $k = 1, \ldots, K$ (Cohen, 1968; Frieden, 2004),

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

  donde $A < B$ significa que $B - A > 0$. 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 $\big(\nabla \log p\big)\big(\nabla \log p\big)^t p = \frac{\big(\nabla p\big)\big(\nabla p\big)^t}{p}$, considerando el gradiente con respeto a $\theta$ (resp. a $x$) tratando de $J_\theta$ (resp. $J$), se obtiene $\sum_k \pi_k \frac{\big(\nabla p_{(k)}\big)\big(\nabla p_{(k)}\big)^t}{p_{(k)}} - \frac{\big(\nabla \sum_k \pi_k p_{(k)}\big)\big(\nabla \sum_k \pi_k p_{(k)}\big)^t}{\sum_k \pi_k p_{(k)}} = \frac{1}{\sum_k \pi_k p_{(k)}} \sum_{k,l} \pi_k \pi_l \Big(\frac{p_l}{p_{(k)}}\big(\nabla p_{(k)}\big)\big(\nabla p_{(k)}\big)^t - \big(\nabla p_{(k)}\big)\big(\nabla p_l\big)^t\Big)$, lo que vale, tratando del caso $K = 2$, $\frac{\pi_1 \pi_2}{p_{(2)} p_{(2)} (\pi_1 p_{(1)} + \pi_2 p_{(2)})}\big(p_{(2)} \nabla p_{(1)} - p_{(1)} \nabla p_{(2)}\big)\big(p_{(2)} \nabla p_{(1)} - p_{(1)} \nabla p_{(2)}\big)^t \geq 0$. No puede

ser idénticamente cero (salvo si $\pi_1\pi_2 = 0$ o $p_{(1)} = p_{(2)}\ldots$) así que se obtiene la desigualdad sobre la matriz de Fisher integrando esta última desigualdad.

### 2.4.6.1. Desigualdad de Cramér-Rao

Una otra interpretación de $J$ como información es debido a la desigualdad de Cramér-Rao que la relaciona a la covarianza de estimación [39] (Rao, 1945, 1992; Rao & Wishart, 1947; Cramér, 1946; Rioul, 2007; Cover & Thomas, 2006; Frieden, 2004; Kay, 1993; van den Bos, 2007). Sea $X$ parametrizada por $\theta$. La meta es estimar $\theta$ a partir de $X$. Tal estimador va a ser una función únicamente de $X$, lo que se escribe usualmente [40] $\widehat{\theta}(X)$ (la función no depende explícitamente de $\theta$). Las características de la calidad de un estimator es naturalmente su sesgo $b(\theta) = \mathrm{E}\left[\widehat{\theta}(X)\right] - \theta$ y su matriz de covarianza $\Sigma_{\widehat{\theta}}$ (la varianza da la dispersión alrededor de su promedio). La desigualdad de Cramér-Rao acota por debajo esta covarianza.

**Teorema 2-28** (Desigualdad de Cramér-Rao)**.** *Sea $X$ parametrizada por $\theta$, de densidad de soporte $\mathcal{X} \subseteq \mathbb{R}^d$ independiente de $\theta$, y $\widehat{\theta}(X)$ un estimador de $\theta$. Sea $b(\theta)$ su sesgo y $\Sigma_{\widehat{\theta}}$ su matriz de covarianza. Sea $\mathrm{J}_b(\theta)$ la matriz Jacobiana del sesgo $b$. Entonces,*

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

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

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

*donde $b'$ es la derivada de $b$.*
*Tomando $\theta$ parámetro de posición y $\widehat{\theta} = X$, 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,*

$$\sigma^2_X \geq \frac{1}{J(X)}.$$

*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 $\theta$.

---

[39] 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).

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

*Demostración.* Sea $S = \nabla_\theta \log p_X$ y $\theta_0 = \mathrm{E}\left[\widehat{\theta}(X)\right] = \theta + b(\theta)$. Fijandose que $p_X \nabla_\theta \log p_X = \nabla_\theta p_X$, que $\widehat{\theta}$ no es función de $\theta$, y que el soporte $\mathcal{X}$ no depende de $\theta$, se obtiene [41]

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

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

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

La prueba se termina tomando $u = J_\theta(X)^{-1}\left(I + \mathrm{J}_b(\theta)\right)^t v$ (recordandose que $J_\theta$ es simétrica).
Para $\theta$ parametro de posición, $\widehat{\Theta} = X$, 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 $x$ (c.s.), est decir $\nabla_x p_X(x) \propto J(X)(x - \theta)p_X(x)$, lo que es la ecuación diferencial que satisface (solamente) la gaussiana: en este caso, se verifica a posteriori que $J(X) = \Sigma_X^{-1}$, y entonces que se alcanza la cota de la desigualdad de Cramér-Rao no paramétrica. □

En el caso paramétrico, no se puede estudiar el caso de igualdad del hecho de que $\widehat{\Theta}$ 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; van den Bos, 2007).

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

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

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

[41] Se supone que los integrandes sean $\theta$-localmente integrables, tal que se puede invertir derivada en $\theta$ e integración; ver también teorema 1-6, pagina 26.

$I$ (Magnus & Neudecker, 1999, cap. 11, sec. 4), dando otras versiones escalares de la desigualdad de Cramér-Rao, por ejemplo, de la versión no paramétrica,

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

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 determinente-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) [42].

### 2.4.6.2. Fisher como curvatura de la entropía relativa

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

**Teorema 2-29** (Fisher como curvatura de la entropía relativa)**.** *Sea $X$ parametrizado por $\theta_0 \in \mathring{\Theta}$ interior de $\Theta$ ($\Theta$ contiene un vecinaje de $\theta_0$). Siendo $D_{\mathrm{kl}}\left(p_X(\cdot;\theta)\,\|p_X(\cdot;\theta_0)\right)$ función de $\theta \in \Theta$, aparece que*

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

*donde $o(\cdot)$ es un resto pequeño con respecto a su argumento. En otros 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 $D_{\mathrm{kl}}\left(p_X(\cdot;\theta)\,\|p_X(\cdot;\theta_0)\right)$ de $\theta$, tomada en $\theta=\theta_0$. Por propiedad de $D_{\mathrm{kl}}$, la divergencia es positiva y se cancela cuando $\theta=\theta_0$. Entonces, el primer término del desarrollo vale cero y el segundo también, $D_{\mathrm{kl}}$ siendo mínima en $\theta=\theta_0$. Además,

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

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

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

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

---

[42] Ver nota de pie 34 pagina 95. 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 $p_\Theta$.

Este teorema, ilustrado en la figura Fig. 2-26, relaciona claramente dos objectos 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)).

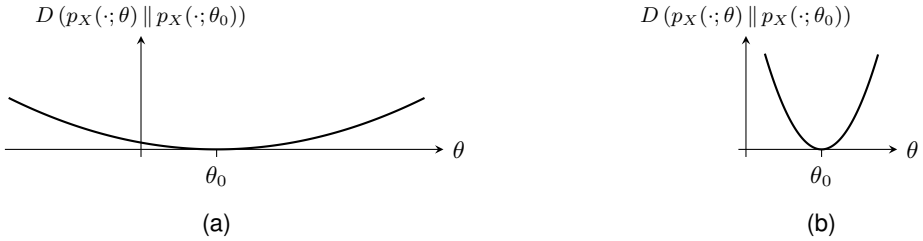



**Figura 2-26:** Ilustración del comportamiento local de $D_{\mathrm{kl}}\left(p_X(\cdot;\theta)\,\|p_X(\cdot;\theta_0)\right)$ en función de $\theta$ en $\theta_0$ en el contexto escalar $\Theta \subseteq \mathbb{R}$. (a) Caso con $J_{\theta_0}(X)$ "pequeño" y (b) caso con $J_{\theta_0}(X)$ "grande". En el caso (b), la determinación de $\theta$ usando $D_{\mathrm{kl}}$ va a ser más "sencillo" que en el caso (a) porque el mínimo es más "picado".

### 2.4.6.3. Identidad de de Bruijn

Un otro vínculo entre el mundo de la información y el de la estimación aparece a través de la identidad de de Bruijn [43] (Stam, 1959; Cover & Thomas, 2006; Johnson, 2004; Barron, 1984, 1986; Palomar & Verdú, 2006; Toranzo, Zozor & Brossier, 2018). Esta identidad caracteriza lo que es conocido como canal gaussiano de la figura Fig 2-27-(a), *i. e.,* la salida $Y$ es una versión ruidosa de la entrada. La identidad vincula las variaciones de entropía de salida con respeto al nivel de ruido, y la información de Fisher.

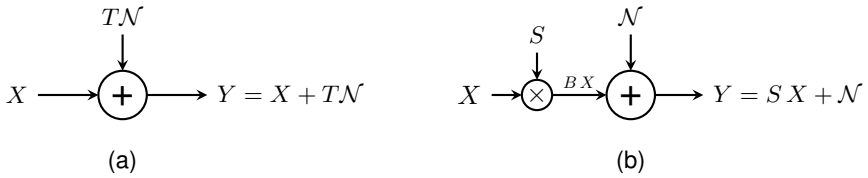



**Figura 2-27:** Canal de comunicación gaussiano de entrada $X$. (a) Canal gaussiano usual, donde $T$ maneja los parámetros (nivel) del ruido. (b) canal gaussiano con un preprocesamiento $S$ de la entrada.

**Teorema 2-30** (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. 2-27-(a)). Entonces, la entropía de Shannon y la información de Fisher de $Y$ satisfacen*

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

[43] 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.

*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*[44] $\theta$,

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

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

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

Además, de $\left(T\Sigma_{\mathcal{N}}T^t\right)\left(T\Sigma_{\mathcal{N}}T^t\right)^{-1} = I$ viene $\frac{\partial \left(T\Sigma_{\mathcal{N}}T^t\right)^{-1}}{\partial T_{i,j}} = -\left(T\Sigma_{\mathcal{N}}T^t\right)^{-1} \frac{\partial \left(T\Sigma_{\mathcal{N}}T^t\right)}{\partial T_{i,j}} \left(T\Sigma_{\mathcal{N}}T^t\right)^{-1}$. Usando le vector $\mathbb{1}_i$ con 1 en su $i$-ésima componente, y cero si no, se obtiene

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

usando la relación $x^t A \mathbb{1}_k \mathbb{1}_l^t B x = \mathbb{1}_l^t B x x^t A \mathbb{1}_k = \mathbb{1}_k^t A^t x x^t B^t \mathbb{1}_l$ (escalares comutan y un escalar es igual a su transpuesta) y usando la simetría de $T\Sigma_{\mathcal{N}}T^t$. Eso significa que

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

dando

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

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

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

Suponiende que se puede intervertir derivadas y integrales (ver (Barron, 1984, 1986) donde se dan condiciones rigurosas, y el teorema 1-6, pagina 26), $p_Y(y) = \int_{\mathbb{R}^d} p_X(x)\, p(y-x)\, dx$ (ver ejemplo 1-7, pagina 46)

[44] Si el parámetro es multivariado, hace falta entender la desigualdad a través de derivas parciales con respeto a los componentes de $\theta$.

satisface también esta ecuación diferencial, y además

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

usando la ecuación diferencial en la 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 (intergración por partes) aplicada respectivamente a los componentes de $\nabla_y p_Y \log p_Y$ y $\nabla_y p_Y$, suponiendo que estos gradientes se cancelan en el borde del dominio de integración, los dos términos integrales valen cero, lo que cierra la prueba de la desigualdad general. Además, si $T = T(\theta)$, la segunda desigualdad sigue de $\frac{\partial \cdot}{\partial \theta} = \sum_{i,j} \frac{\partial \cdot}{\partial T_{i,j}} \frac{\partial T_{i,j}}{\partial \theta} = \mathrm{Tr}\left(\nabla_T \cdot \frac{\partial T^t}{\partial \theta}\right)$. □

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

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

se recupera en el caso particular $T = \sqrt{\theta} I$. En este caso, la ecuacíon diferencial 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 respeto al nivel de ruido gaussiano (la gaussiana juega de nuevo un rol central acá).

Existe una otra forma muy similar de esta desigualdad debido a Guo, Shamai, Verdú, Palomar (Guo, Shamai & Verdú, 2005; Palomar & Verdú, 2006; Toranzo et al., 2018). 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 $S$ coresponde a un pre-tratamiento antes de la salida. Eso es ilustrado en la figura Fig. 2-27-(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 $\mathrm{E}\left[\left|\widehat{X}(Y) - X\right|^2\right]$ es la esperanza condicional $\widehat{X}(Y) = \mathrm{E}[X|Y]$ (Kay, 1993; Robert, 2007; Lehmann & Casella, 1998). Una característica de un estimador siendo su matriz de covarianza, se notará $\mathcal{E}(X|Y) = \mathrm{E}\left[(X - \mathrm{E}[X|Y])\,(X - \mathrm{E}[X|Y])^t\right]$ esta matriz. Sorpredentemente, existe también una identidad entre $I(X;Y)$ y $\mathcal{E}(X|Y)$:

**Teorema 2-31** (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. 2-27-(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* $S = S(\mathfrak{s})$ *depende de un parámetro escalar* $\mathfrak{s}$*,*

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

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

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

(multiplicando ambos lados por $p_X$). Ahora, $I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(\mathcal{N})$ (de la independencia, cuando $X = x$, $Y = Sx + \mathcal{N}$ gaussiana de misma convarianza que $\mathcal{N}$ y de promedio $Sx$ (ver ejemplo 1-8, pagina 50), así que

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

La segunda línea viene de la escritura de $H(Y)$ usando $p_Y$ como marginale de $p_{X,Y}$ en $x$ y intercambiando gradiente e integral (ver pasos de la prueba de la desigualdad de de Bruijn); la tercera de $\frac{p_{X,Y}(x,y)}{p_Y(y)} = p_{X|Y=y}(x)$; en la cuarta se usa la ecuación diferencial satisfecha por $p_{X,Y}$ en la primera integral y 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 el la segunda ($p_Y$ sumando a 1, el término se cancela). Además,

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

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

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

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 $\mathrm{E}\left[\mathrm{E}[X|Y]\right] = \mathrm{E}[X]$ y por la formula de König-Huyggens (ver capítulo **??**, subsección 1.4.2, pagina 53).

La segunda identidad viene de $\frac{\partial \cdot}{\partial \mathfrak{s}} = \mathrm{Tr}\left(\nabla_S \frac{\partial S^t}{\partial \mathfrak{s}}\right)$ (ver prueba de la identidad de de Bruijn). □

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

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

### 2.4.6.4. Desigualdad de Stam

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 [45] (Cover & Thomas, 2006; Rioul, 2007; Stam, 1959), o a veces "desigualdad isoperimetrica para la entropía" (Wang & Madiman, 2004).

**Teorema 2-32** (Desigualdad de Stam)**.** *Sea $X$ una variable aleatoria continua sobre $\mathcal{X} \subseteq \mathbb{R}^d$. Entonces,*

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

*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 $\theta \to 0$, aparece que $\frac{d}{d\theta} N(X + \sqrt{\theta}\mathcal{N})\Big|_{\theta=0} \geq 1$. La prueba se cierra con $\frac{d}{d\theta} N(X + \sqrt{\theta}\mathcal{N}) = \frac{1}{2\pi\,\mathrm{e}} \frac{d}{d\theta} \exp\left(\frac{2}{d} H(X + \sqrt{\theta}\mathcal{N})\right) = \frac{2}{d} N(X + \sqrt{\theta}\mathcal{N}) \frac{d}{d\theta} H(X + \sqrt{\theta}\mathcal{N}) = dN(X + \sqrt{\theta}\mathcal{N})\,\mathrm{Tr}\left(J(X + \sqrt{\theta}\mathcal{N})\right)$ (por la 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. □

[45] 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.

Se puede ver de nuevo el rol central que juega la gaussiana en esta desigualdad. Además, de la desigualdad de Stam se puede deducir tambíen 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 desiguldad de Stam, $|\Sigma_X|^{\frac{1}{d}} \operatorname{Tr}(J(X)) \geq d$ (y las otras versiones escalares de la relación determinente-traza). Como se lo puede esperar, se obtiene la igualdad si y solamente $X$ es gaussiano (potencia entrópica alcanzando su cota superior) y de matriz la identidad (desiguladad de Stam se saturando).

Varias otras pruebas de la desigualdad de Stam pueden provenir de generalizaciones (Bercher, 2012, 2013; Lutwak, Yang & Zhang, 2005; Lutwak, Lv, Yang & Zhang, 2012; Zozor, Puertas-Centeno & Dehesa, 2017). **La sección ZZZ lo va a rapidamente evocar. Ver caso discreto Kagan (Kagan, 2001).**

### 2.4.6.5. Fisher aditividad, procesamiento de datos y convolución

Además del grán número de relaciones entre la información de Fisher y otras medidas informacionales, la información de Fisher satisface también desigualdades en si mismo, muy parecidas a las satisfechas por la entropía o información mutua.

Primero, al imagen de la entropía condicional, se puede definir una información condicional al imagen de la definición Def. 2-45,

**Definición 2-51** (Matriz información de Fisher parámetrica condicional)**.** *Sean $X$ e $Y$ dos variables aleatoria parametrizada por el mismo parámetro $m$-dimensional, $\theta \in \Theta \subseteq \mathbb{R}^m$, de distribución de probabilidad conjunta $p_{X,Y}(\cdot,\cdot;\theta)$ continua sobre $\mathcal{X} \times \mathcal{Y}$ su soporte, $p_{X|Y=y}(\cdot;\theta)$ la distribución condicional de $X$ connociendo $Y = y$ y $p_Y$ la distribución marginal. Suponga que estas distribuciones sean diferenciable en $\theta$ sobre $\Theta$. La matriz de Fisher de $X$ condicionalmente a $Y$ es el promedio estadístico sobre $p_Y$ de la matriz de Fisher de $p_{X|Y}(\cdot;\theta)$, es decir*

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

*donde $p_{X|Y}(\cdot;\theta) = \frac{p_{X,Y}(\cdot,Y;\theta)}{p_Y(Y)}$ es acá una variable aleatoria.*

De esta definición, es sencillo probar de que la matriz de Fisher paramétrica sigue una regla de cadena al imagen de la propiedad [P14],

$$J_\theta(X,Y) = J_\theta(X|Y) + J_\theta(Y).$$

Además, si $X$ e $Y$ son independientes, la información de Fisher es aditiva de la misma manera que $H$ satisface las propiedades [P10] y [P13], *i. e.,*

$$J_\theta(X|Y) = J_\theta(X) \quad \Leftrightarrow \quad J_\theta(X,Y) = J_\theta(X) + J_\theta(Y) \quad \Leftrightarrow \quad X \,\&\, Y \text{ son independientes.}$$

En particular, tratando de una secuencia $X = \{X_i\}_{i=1}^n$ de vectores aleatorias independientes parametrizados por $\theta$, $J_\theta(X) = nJ_\theta(X_i)$, lo que significa que estimando $\theta$ a partir de la secuencia se baja a la taza $1/n$ la cota de Cramér-Rao. Se referirá a (Fisher, 1925; Stam, 1959; Kay, 1993; Kagan & Smith, 1999; Johnson, 2004; Cover & Thomas, 2006; Rioul, 2007) entre otros para estas propiedades.

De la regla de cadena, viene obviamente la desigualdad siguiente, parecida a la propiedad de super-aditividad [P12],

$$J_\theta(X_1, \ldots, X_n) \geq J_\theta(X_i) \quad \forall\, 1 \leq i \leq n,$$

y una desigualdad de procesamiento de datos via la información de Fisher (Zamir, 1998; Rioul, 2007; Cover & Thomas, 2006; Frieden, 2004; Kagan & Smith, 1999):

**Teorema 2-33** (Desigualdad de procesamiento de datos tipo Fisher)**.** *Sea $\theta \mapsto X \mapsto Y$ un proceso de Markov con $\theta$ determinista y $p_{X,Y}$ parametrizado por $\theta$, es decir en este contexto que, $p_{Y|X=x}$ no es parametrizado por $\theta$. Entonces*

$$J_\theta(X) \geq J_\theta(Y),$$

*con igualdad si y solamente si $\theta \mapsto Y \mapsto X$ es también de Markov. En particular,*

$$\forall\, g, \quad J_\theta(X) \geq J_\theta(g(X)).$$

*Demostración.* De la regla de cadena tenemos

$$J_\theta(Y|X) + J_\theta(Y) = J_\theta(X|Y) + J_\theta(X).$$

Del hecho de que $p_{Y|X=x}$ no es parametrizado por $\theta$ es sencillo ver que $J_\theta(Y|X) = 0$, la prueba se cerrando de $J_\theta(X|Y) \geq 0$. Además se obtiene la igualdad si y solamente si $J_\theta(X|Y) = 0$, es decir de la "positividad" del integrante dando la matrix de Fisher, si y solamente si $p_{X|Y=y}$ no es parametrizado por $\theta$. □

Mencionamos de que existe también una desigualdad parecida a la de la potencia entrópica, teorema 2-25, dada en el caso escalar en (Johnson, 2004; Blachman, 1965; Zamir, 1998; Dembo et al., 1991; Kagan & Yu, 2008):

**Teorema 2-34** (Desigualdad convolucional de Fisher)**.** *Sean $X$ e $Y$ dos variables $d$-dimensionales continuas independientes parametrizadas. Entonces*

$$\forall\, a \in [0\,,\,1], \quad J\left(\sqrt{a}\,X + \sqrt{1-a}\,Y\right) \leq a\,J(X) + (1-a)\,J(Y),$$

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

*Demostración.* $X$ e $Y$ siendo independientes, tenemos para $W = X + Y$, $p_W(w) = \int_{\mathcal{X}} p_X(x) p_Y(w-x)\,dx$ convolución de las distribuciones de $X$ y de $Y$ (ver ejemplo 1-7, pagina 46). Escribiendo $S_X = \nabla_x \log p_X$ el

score de $X$ y lo mismo para $Y$ y $W$,

$$\begin{aligned} S_W(w) &= \int_{\mathcal{X}} \frac{p_X(x)}{p_W(w)} \nabla_w p_Y(w-x)\, dx \\ &= -\int_{\mathcal{X}} \frac{p_X(x)}{p_W(w)} \nabla_x p_Y(w-x)\, dx \\ &= \int_{\mathcal{X}} \frac{p_Y(w-x)}{p_W(w)} \nabla_x p_X(x)\, dx \\ &= \int_{\mathcal{X}} \frac{p_Y(w-x)p_X(x)}{p_W(w)} \nabla_x \log p_X(x)\, dx \\ &= \int_{\mathcal{X}} p_{X|W=w}(w)\, \nabla_x \log p_X(x)\, dx \\ &= \mathrm{E}\left[S_X(X) \middle| W=w\right] \end{aligned}$$

Intercambiando los roles de $X$ e $Y$, tenemos también $S_W(w) = \mathrm{E}\left[S_Y(Y) \middle| W=w\right]$, así que, para cualquier $0 \le a \le 1$,

$$S_W(w) = \mathrm{E}\left[aS_X(X) + (1-a)S_Y(Y) \middle| W=w\right].$$

A continuación, de la formula de König-Huyggens (ver capítulo **??**, subsección 1.4.2, pagina 53),

$$S_W(w)S_W(w)^t \le \mathrm{E}\left[\left(aS_X(X) + (1-a)S_Y(Y)\right)\left(aS_X(X) + (1-a)S_Y(Y)\right)^t \middle| W=w\right],$$

es decir, tomando el promedio en $W$,

$$J(X+Y) \le a^2 J(X) + (1-a)^2 J(Y) + a(1-a)\left(\mathrm{E}\left[S_X(X)S_Y(Y)^t\right] + \mathrm{E}\left[S_Y(Y)S_X(X)^t\right]\right).$$

Luego, $X$ e $Y$ siendo independientes, $S_X(X)$ y $S_Y(Y)$ son también independientes. Además son centradas, probando que el término en $a(1-a)$ vale cero, dando una versión equivalente del teorema; La versión dada se recupera re-emplazando $X$ por $\sqrt{a}\,X$ e $Y$ por $\sqrt{1-a}\,Y$.

Escribiendo la desigualdad viniendo de la formula de König-Huyggens, se nota de que la igualdad es satisfecha si y solamente si $aS_X(w-x)+(1-a)S_Y(x) = S_W(w)$ para cualquier $x, w$. Integrando en $x$ se obtiene $-a\log p_X(w-x) + (1-a)\log p_Y(x) = xS_W(w) + g(w)$. Derivando en $w$ obtenemos $-a\nabla_w \log p_X(w-x) = S_W(w) + \nabla_w g(w)$, es decir, en $w=0$, notando de que $\nabla_w \log p_X(w-x) = -\nabla_x \log p_X(w-x)$, se nota de que $\nabla_x \log p_X(x)$ es constante, *i. e.,* $X$ es necesariamente gausiana. Similarmente, $Y$ es necesariamente gausiana también. Además, calculando las informaciones de Fisher en el caso gausiano, obtenemos $(\Sigma_X + \Sigma_Y)^{-1} = \Sigma_X^{-1} + \Sigma_Y^{-1}$ lo que es posible si y solamente si $\Sigma_X$ y $\Sigma_Y$ son proporcionales. □

Este teorema tiene varias consecuencias. En particular, interviene en la prueba de la desigualdad de la potencia entrópica.

**(2) ver MinFisher Frieden p. 235, Berchet Vignat 2009, Ernst 2017; cf. travaux rederivant MQ de Frieden-Plastino-Soffer (1999, 2002), Reginato 98, Bickel 81**

## 2.5 Unos ejemplos y aplicaciones

### 2.5.1 Canal de transmisión y su capacidad

Siguiendo el esquema de comunicación de Shannon, un mensaje que se modeliza como un vector aleatorio [46] $X$ 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 2-52** (Capacidad de canal)**.** *Sea un canal de transmisión, $X$ su entrada e $Y$ su salida, como ilustrado figura Fig. 2-28. Sea $p_X$ la distribución de probabilidad de $X$. La capacidad $C$ del canal es definida por*

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

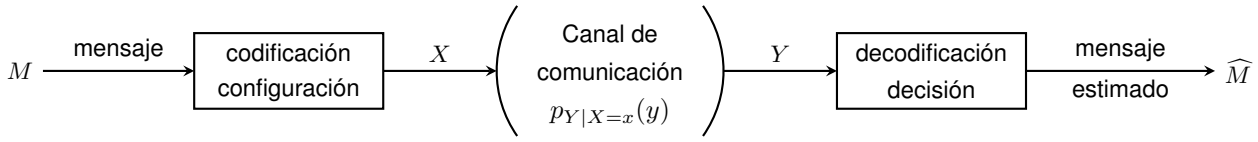



**Figura 2-28:** Esquema de comunicación de Shannon. En una primera etapa, un mensaje $M$ a transmitir es códificado (ej. código binario) o puesto en forma (ej. símbolos modulando una función para que sea analógica y en una banda frecuencial dada). Sea $X$ este mensaje codificado o puesto en forma. A la recepción, se mide $Y$ (ej. versión ruidosa de $X$), antes de ser decodificado o usado para tomar una decisión, $\widehat{M}$ 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 respeto a su entrada.

#### 2.5.1.1. Canal binario

[46] 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.

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 probabilidaded de transición $p_{Y|X=x}(y)$, *i. e.,* las probabilidades que un 0 (resp. un 1) se transmite correctamente o cambia en un 1 (resp. 0), *i. e.,*

$$\varepsilon = PY = 1|X = 0) = 1 - P(Y = 0|X = 0) \qquad \text{y} \qquad \vartheta = P(Y = 0|X = 1) = 1 - P(Y = 1|X = 1).$$

$\varepsilon$ y $\vartheta$ representan errores de comunicación. Tal canal es descrito figura Fig. 2-29-(a). La figura Fig. 2-29-(b) da un esquema "práctico" que podría ser al origen de un tal canal. Cuando $\varepsilon = \vartheta$, el canal es conocido como *canal binario simétrico*. Cuando $\varepsilon = 0$ y $\vartheta \in (0\,;\,1)$, el canal es conocido como *canal binario en Z*.

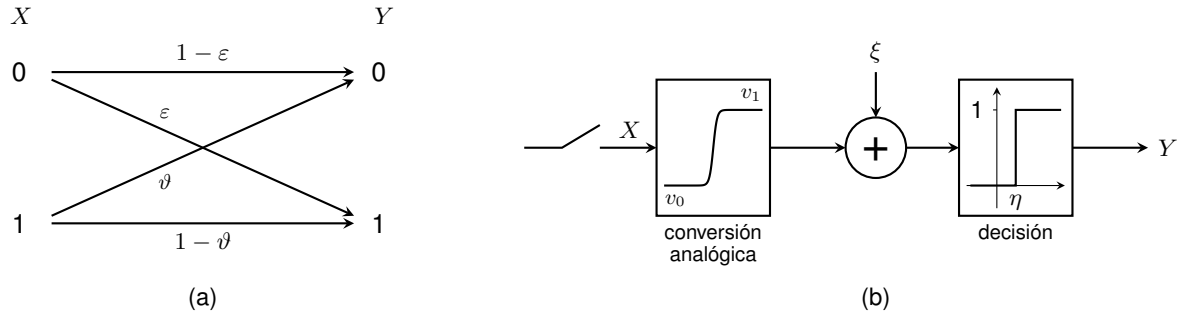



**Figura 2-29:** (a): Canal binario. La entrada $X$ definida sobre $\mathcal{X} = \{0, 1\}$ pasa por este canal e $Y$ definida sobre $\mathcal{Y} = \mathcal{X}$ es recibido. Este canal es caracterizado por las probabilidades de transición $p_{Y|X=x}(y)$. (b): Esquema que puede conducir al canal binario; 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 añandiendo un ruido $\xi$. En la recepción, se toma una decisión, por ejemplo 0 (resp. 1) si la medida es mayor (resp. menor) que $\eta = \frac{v_0+v_1}{2} + \mathrm{E}[\xi]$. En este ejemplo, $\varepsilon$ y $\vartheta$ 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, sean

$$r = P(X = 0),$$

dando la distribución de la entrada. La distribución de la salida va a ser dada a partir de $s = P(Y = 0) = P(Y = 0|X = 0)P(X = 0) + P(Y = 0|X = 1)P(X = 1)$ es decir

$$s = P(Y = 0) = \vartheta + r(1 - \varepsilon - \vartheta).$$

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

$$I_2(X;Y) = h_2(s) - rh_2(\varepsilon) - (1 - r)h_2(\vartheta),$$

donde $h_2(u) = -u\log_2 u - (1-u)\log_2(1-u)$ es la entropía binaria en bits. Para calcular la capacidad $C_2$ en bits, hace falta maximizar $I_2$ con respeto a $r$. Diferenciando $I_2$ en $r$, *i. e.,* $\frac{\partial I_2(X;Y)}{\partial r} = \frac{\partial h_2(s)}{\partial s}\frac{\partial s}{\partial r} - h_2(\varepsilon) + h_2(\vartheta)$, es decir

$$\frac{\partial I_2(X;Y)}{\partial r} = (1-\varepsilon-\vartheta)\log_2\left(\frac{1-s}{s}\right) - h_2(\varepsilon) + h_2(\vartheta).$$

- Claramente,

$$\vartheta = 1-\varepsilon \quad \Rightarrow \quad C_2 = 0.$$

  Viene del hecho de que para $\vartheta = 1-\varepsilon$, de $h_2(\varepsilon) = h_2(1-\varepsilon)$ 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 $s = \vartheta$, dando $p_{Y|X=x} = p_Y$, dando una información mutua nula, y entonces una capacidad nula.

- Si $\vartheta \neq 1-\varepsilon$, la derivada de $I_2$ con respeto a $r$ se anula para $s = s^{\rm opt}$ ($r = r^{\rm opt}$),

$$s^{\rm opt} = \frac{1}{1+2^{\frac{h(\varepsilon)-h(\vartheta)}{1-\varepsilon-\vartheta}}} \qquad \text{siendo} \qquad r^{\rm opt} = \frac{s^{\rm opt}-\vartheta}{1-\varepsilon-\vartheta},$$

  y dando un extremo para $I_2$. A continuación, $\frac{\partial^2 I_2}{\partial r^2} = \frac{(1-\varepsilon-\vartheta)^2}{s(1-s)} > 0$ (en particular para el $s$ "óptimo"), probando de que el extremo es un máximo. Poniendo la expresión de $r^{\rm opt}$ en la formula de $I_2(X;Y)$, luego de muchos cálculos (básicos), se obtiene

$$C_2 = \log_2\left(1+2^{\frac{h_2(\varepsilon)-h_2(\vartheta)}{1-\varepsilon-\vartheta}}\right) - \frac{(1-\vartheta)\,h_2(\varepsilon) - \varepsilon\,h_2(\vartheta)}{1-\varepsilon-\vartheta}.$$

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

De $I_2(X;Y) = H_2(Y) - H_2(Y|X) \leq H_2(Y) \leq 1$ bit ($Y$ es binario, de 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 [47]: 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 $rh_2(\varepsilon) + (1-r)h_2(\vartheta) = 0$ y además (ii) $h_2(s) = 1$. Estudiando cada caso (ej. con $r = 0$ y $\vartheta = 0$ se satisface (i) pero no (ii) porque $s = 0$), se obtiene que

$$C_2 = 1 \qquad \Leftrightarrow \qquad r = \frac{1}{2} \quad \text{y} \quad \varepsilon = \vartheta = \frac{1\pm 1}{2}.$$

---

[47] 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, $p_{Y|X=x}$ depende solo del canal y no de la entrada, así que para $p_X = \pi_1 p_{(1)} + \pi_2 p_{(2)}$ ($\pi_2 = 1-\pi_1$) se obtiene $p_Y = \pi_1 q_{(1)} + \pi_2 q_{(2)}$ con $q_{(i)}$ distribución de la salida corespondiente a una entrada de distribución $p_{(i)}$. 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 respeto a $p_X$. A continuación, $p_X$ parteneciendo a un convexo, $I$ tiene un máximo que es único.

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

La figura Fig. 2-30 representa la información mutua $I(X;Y)$ para unos canales ($\varepsilon$ y $\vartheta$ dados) en función de $r$. Se nota que la curva es cóncava y tiene un máximo, capacidad del canal. La figura Fig. 2-31 representa la capacidad del canal en función de $\varepsilon$ y $\vartheta$ así que unos casos particulares/cortes.

En el caso particular $\varepsilon = \vartheta$, conocido como *canal simético*, la capacidad es

$$C_2 = 1 - h_2(\varepsilon)$$

(alcanzada con una entrada uniforme). Como visto en el caso general, la capacidad vale 1 bit si y solamente si $h_2(\varepsilon) = 0$, es decir $\varepsilon = 0$ o $\varepsilon = 1$. Al revés, la capacidad es mínima cuando $H_2$ est máximo, es decir para $\varepsilon = \vartheta = \frac{1}{2}$, y $C_2 = 0$ (instancia particular de $\vartheta = 1 - \varepsilon$). $h_2(\varepsilon)$ es la perdida en bit para cada bit transmitido. La capacidad $C_2$ en función de $\varepsilon$ es dada figura Fig. 2-31-(b).

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

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

Se nota en este caso también que la capacidad alcanza 1, su máximo, si y solamente si $\vartheta = 0$ (canal perfecto). Al revés, cuando $\vartheta \to 1$, $C \to 0$, instancia particular de $\vartheta = 1 - \varepsilon$. La capacidad $C_2$ en función de $\vartheta$ es dada figura Fig. 2-31-(c).

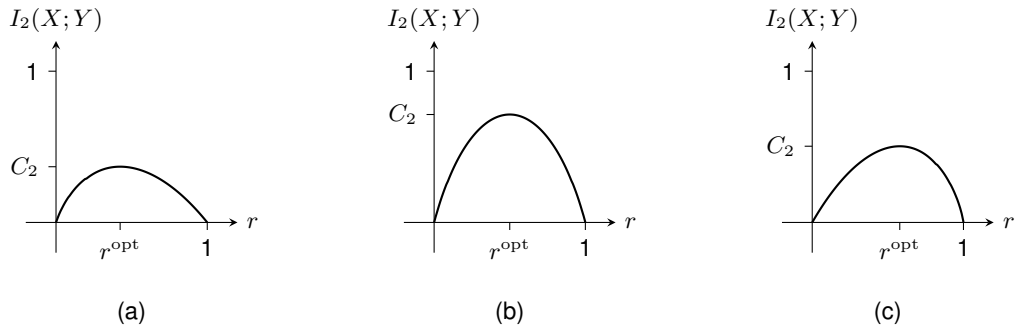



**Figura 2-30:** Información mutua (en bits) entrada-salida $I_2(X;Y)$ del canal binario en función de $r = P(X = 0)$. (a): $\varepsilon = 0{,}4$ y $\vartheta = 0{,}01$; (b): $\varepsilon = \vartheta = 0{,}05$ (canal simético); (c): $\varepsilon = 0$ y $\vartheta = 0{,}05$ (canal en Z).

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. 2-32 (ver también (Shannon, 1948; Elias, 1957) o (Arimoto, 1972) para el cálculo numérico de la capacidad en el caso general).

[48] Se mencionará de que en toda esta subsección, no se necesita de que $X$ y/o $Y$ sean variables aleatorias reales, *i. e.,* pueden tomar sus valores sobre cualquier espacio discreto (por ejemplo de letras).

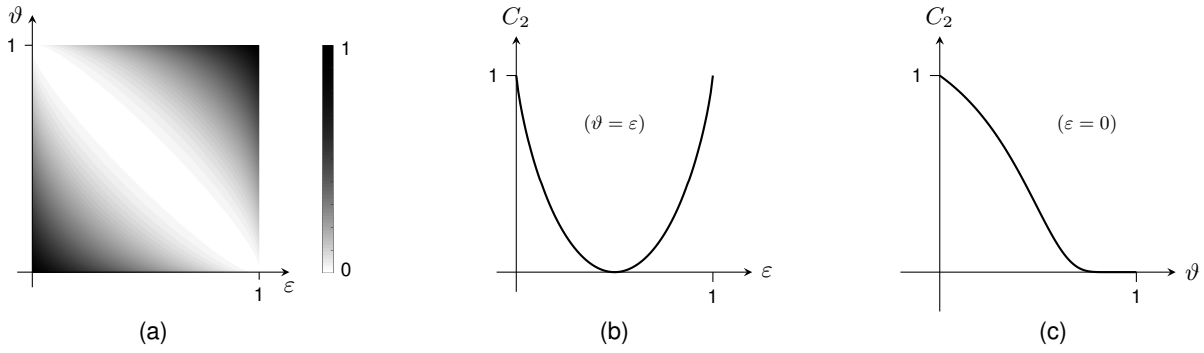



**Figura 2-31:** Capacidad $C_2$ del canal binario. (a): en función de $\varepsilon$ y $\vartheta$. (b): en función de $\varepsilon$ para el canal simético ($\varepsilon = \vartheta$); (c): en función de $\vartheta$ para $\varepsilon = 0$ (canal en Z).

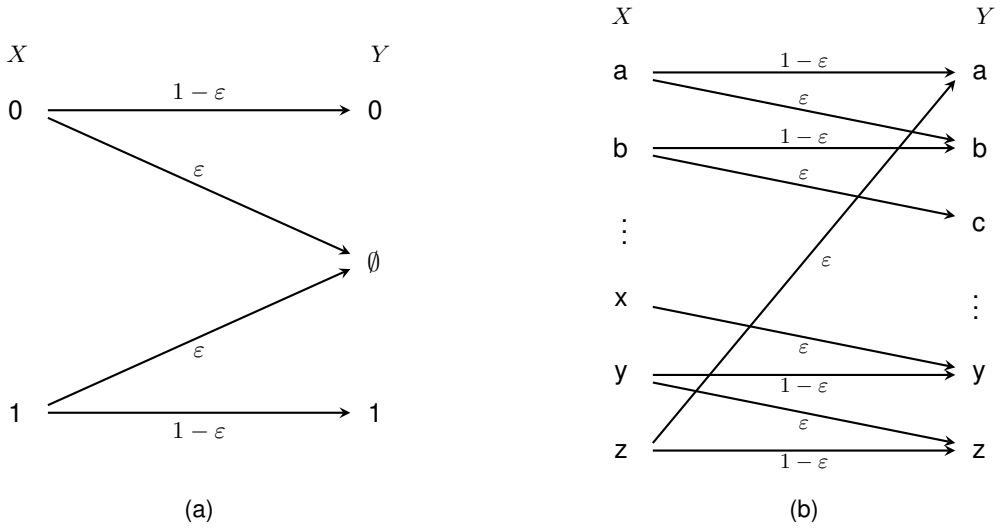



**Figura 2-32:** Ejemplos de canales discretos usuales. (a): canal borrador, donde un 0 (de probabilidad de ocurrencia $r$) o 1 (de probabilidad de occurrencia $1-r$) puede transmitirse correctamente or ser borado/perdido (estado $\emptyset$) con una probabilidad $\varepsilon$. Se calcula $I_2(X;Y) = (1-\varepsilon)h_2(r)$, dando la capacidad $C_2 = 1-\varepsilon$, alcanzada para una entrada uniforme. (b): canal tipo machina de escribir [48], 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 uniforma, dando $C_n = 1 - h_n(\varepsilon)$.

## 2.5.2 Canal de transmisión continuo gaussiano y su capacidad

Un canal de comunicación continuo relativamente simple es conocido como *canal gaussiano* (Shannon, 1948, Sec. 25), (Cover & Thomas, 2006; Rioul, 2007): $X$ es una variable continua definida sobre $\mathcal{X} \subseteq \mathbb{R}^d$ y la 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" $p_{Y|X=x}(y)$, *i. e.,* por la distribución del ruido. Tal canal es descrito figura Fig. 2-33. 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

$$\mathrm{Tr}\,(\Sigma_X) \leq \mathcal{P},$$

potencia límite permitida por componente (sampleo).

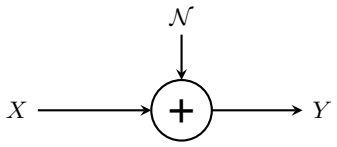



**Figura 2-33:** Canal gaussiano. La entrada $X$, modelizada por un vector aleatorio, es corrupta aditivamente por un ruido gaussiano $\mathcal{N}$ independiente de $X$. La salida es entonces $Y = X + \mathcal{N}$ y el canal es completamente descrito por $p_{Y|X=x}(y) = p_{\mathcal{N}}(y - x)$ (obviamente independiente de la distribución de la entrada).

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 $\mathrm{Tr}\,(\Sigma_X) \leq \mathcal{P}$. 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 $X$ est gaussiana, dando $I(X;Y) = \frac{1}{2}\log|\Sigma_X + \Sigma_{\mathcal{N}}| - \frac{1}{2}\log|\Sigma_{\mathcal{N}}|$. Tomando en cuenta el límite de potencia, hace falta maximizar $|\Sigma_X + \Sigma_{\mathcal{N}}|$ sujeto a $\mathrm{Tr}\,\Sigma_X \leq \mathcal{P}$ y $\Sigma_X \geq 0$ simética lo que no es trivial. Se encuentra el enfoque permitiendo solucionar el problema en (Cover & Thomas, 2006, Sec. 9.4). Sea $U$, matriz ortogonal ($UU^t = U^tU = I$) de los autovectores de la matriz $\Sigma_{\mathcal{N}} \geq 0$ simética [49], de columnas $u_i$ ordenadas tal que los autovalores corespondientes $\lambda_i^{\mathcal{N}}$ sean en orden creciente, *i. e.,*

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

donde $\mathrm{diag}$ 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 $\mathrm{Tr}\,\Sigma_X = \mathrm{Tr}\,R_X$ (de $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$). Entonces, el problema se reduce a maximizar $|R_X + \Lambda_{\mathcal{N}}|$ sujeto a $\mathrm{Tr}\,R_X \leq \mathcal{P}$ y $R_X \geq 0$ simética. La desigualdad de Hadamard ya evocada da $|R_X + \Lambda_{\mathcal{N}}| \leq \prod_i (R_X + \Lambda_{\mathcal{N}})_{i,i} = \prod_i \left((R_X)_{i,i} + \lambda_i^{\mathcal{N}}\right)$ donde $(\cdot)_{i,i}$ denota la componente $i,i$ de la matriz, 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 $\prod_i \left(\lambda_i^X + \lambda_i^{\mathcal{N}}\right)$ sujeto a $\sum_i \lambda_i^X \leq \mathcal{P}$ y $\lambda_i^X \geq 0$. Este problema de optimización sujeto a una desigualdad se resuelva con el enfoque de Karush-Kuhn-Tucker [50] (KKT) (Miller, 2000; Cambini & Martein,

[49] Se recordará de que $A \geq 0$ significa que $A$ es definida no negativa.

[50] Se introduce el factor de Lagrange y se maximiza $\prod_i \left(\lambda_i^X + \lambda_i^{\mathcal{N}}\right) + \eta \sum_i \lambda_i^X$. 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, $\lambda_i^X = \left(\lambda - \lambda_i^{\mathcal{N}}\right)_+$ con $\lambda$ el factor de Lagrande 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 $\sum_i \lambda_i^X \leq \mathcal{P}$, *i. e.,* alcanzando la igualdad.

2009), dando $\lambda_i^X = \left(\lambda - \lambda_i^{\mathcal{N}}\right)_+$ con $(\cdot)_+ = \max(\cdot, 0)$ y $\lambda$ tal que $\sum_i \left(\lambda - \lambda_i^{\mathcal{N}}\right)_+ = \mathcal{P}$. En conclusión, la capacidad es dada por

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

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), illustrado figura Fig. 2-34. 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 $\mathcal{P}$; en cada $\lambda_i^{\mathcal{N}}$ se ha añadido el $\lambda_i^X$ (Cover & Thomas, 2006, Sec. 9.4).

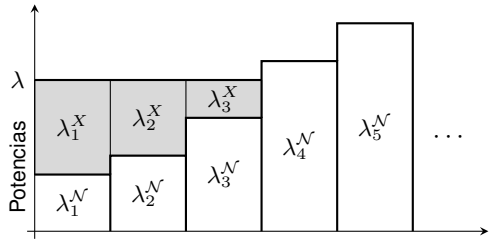



**Figura 2-34:** Principio del "water-filling" para obtener los $\lambda_i^X$ satisfaciendo el vínculo de potencia límite y permitiendo de construir $\Sigma_X$ a partir de la matriz diagonal de los $\lambda_i^X$ y la base que diagonaliza la covarianza $\Sigma_{\mathcal{N}}$ del ruido. La zona en grise representa esquemáticamente $\mathcal{P}$.

En el caso escalar, se obtiene

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

donde $\frac{\mathcal{P}}{\sigma_{\mathcal{N}}^2}$ es conocido como relación señale-ruido [51]

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 retroacción (o feedback), etc.).

[51] 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 $\frac{\mathcal{P}}{N_0 B}$. Además, se calcula en general la capacidad por unidad de tiempo es decir la capacidad por símbolo divido por $T = \frac{1}{2B}$, *i. e.,* $C = B\log\left(1 + \frac{\mathcal{P}}{N_0 B}\right)$ por segundos, lo que es precisamente la capacidad calculdada por Shannon. Esta es a veces conocida como formula de Shannon-Hartley.

### 2.5.3 Codificación entrópica sin perdida

El problema de codificación de fuente puede presentarse de la manera siguiente (Cover & Thomas, 2006, cap. 5) o (Rioul, 2007, cap. 13). Sea un proceso aleatorio $\{X_t\}_{t\in\mathbb{Z}}$, supuesto estacionario, llamado *fuente*, donde los $X_t$ toman sus valores sobre un alfabeto discreto finito no necesariamente real ($X$ puede tomar cualquier etiqueta)

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

de distribución $p_X$. A cada posible secuencia [52] $s_1 \cdots s_n \in \mathcal{X}^n$ de letras de $\mathcal{X}$, se quiere asignar un código $c(s_1 \cdots s_n)$ de letras de un alfabeto discreto finito,

$$\mathcal{C} = \{\mathfrak{c}_1, \ldots, \mathfrak{c}_d\} \qquad \text{alfabeto código.}$$

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

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

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

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

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

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

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 $1 \leq i \leq \alpha$ se puede escribir de manera única $i - 1 = (i_0 - 1) + (i_1 - 1)d + \cdots + (i_K - 1)d^K$ donde $K = \left\lceil \log_d |\mathcal{X}| \right\rceil$ y $1 \leq i_k \leq \alpha$. Entonces, se puede asignar la palabra código $c(x_i) = \mathfrak{c}_{i_0} \cdots \mathfrak{c}_{i_k}$ al símbolo $x_i$. Haciendo eso, cada palabra código tieno el mismo largo. Pero, es más económico hacer una codificación dicha de largos variables, teniendo en cuenta las probabilidades de aparición de cada $x_i$. Implicitamente, es la idea del código de Moorse, 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

[52] Por abuso de escritura una cadena de $n$ símbolos puede ser vista como un $n$-uplet.

que son raras (ej. bara-bara-punto-bara para el 'q' o cinco baras 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 est dicho *descifrable* o *a decodificación única* (o sin perdida) si todas las extensiones son no singulares.

**Ejemplo 2-18** (Código non singular, pero no decifrable)**.** *Sean, sean* $\mathcal{X} = \{א, ב, ג, ד\}$, $\mathcal{C} = \{0, 1\}$ *y* $c(א) = 0$, $c(ב) = 00$, $c(ג) = 1$, $c(ד) = 01$ *(*$m = 1$*). El código es no singular, pero no descifrable. La secuencia* $0010$ *puede provenir de* אגאא*, de* אדא *o de* אגב.

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 instantaneo*.

**Ejemplo 2-19** (Código decifrable, pero no instantaneo)**.** *Sea el código* $c(א) = 00$, $c(ב) = 10$, $c(ג) = 11$, $c(ד) = 110$. *Este código es descifrable, pero no instantaneo. Considera la secuencia* $0011011$ *y marcha sobre ella.* $0$ *no es una palabra código;* $00$ *es y sin ambigüedad proviene de un* א *(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* ג *o de un* ד*; la letra siguiente siendo un* $0$*, todavía no se puede concluir si* $110$ *vino de* ג *y algó o* ד*. Al final, con* $1101$*, se sabe que se tuvimos un* ד *porque ninguna palabra código empieza por* $01$*. Al final, sin ambigüedad el antecedente de la secuencia binaria era* גדא*. Pero se necesitó marchar sobre toda la secuencia antes de decodificar.*

Obviamente, un código instantaneo es tal que ninguna palabra código es prefijo de una otra, *i. e.,* si $c_m(x)$ es una palabra código, las otras palabras código no pueden empezar con $c_m(x)$; el código es también dicho *libre de prefijo*. Estas distinciones estan ilustradas en la figura Fig. 2-35 (ver (Cover & Thomas, 2006, cap. 5)).

Además de la decodificación sin ambigüedad, una caracterización importante del código es la taza de codificación [53]

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

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,

[53] En (Rioul, 2007) por ejemplo, se define esta taza suponiendo que cada secuencia fuente es codificada por el mismo número de bits. La taza es entonces el número de bits por símbolo.

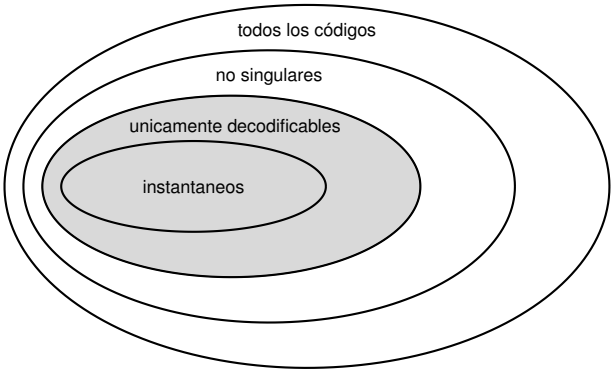



**Figura 2-35:** Clases de códigos. Los códigos contienen la clase de los no singulares. La misma contiene la clase de los códigos descifrables. Ella contiene los códigos instantaneos. En grise se representan las clases de códigos sin perdida a lo cuales se dedica esta sección.

nos concentramos en el problema de compresión, sin tener en cuenta el paso de transmisión de mensajes codificados en un canal. Minimizar la taza es equivalente a minimizar el largo promedio. Además, se puede focalisarse 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

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

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) [54]

**Teorema 2-35** (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 $l_{c_k}(s) = \sum_{i=1}^{k} l_c(s_i)$. 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 códicos 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

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

inyectiva, imponiendo $\#(m) \leq d^m$ (no hay más palabras de largo $m$ que el cardinal de $\mathcal{C}^m$), dando inmediatemente que necesariamente

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

Un estudio rápido de $u \mapsto \left(u\, l_c^{\max}\right)^{\frac{1}{u}}$ 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 $\ell = 1, \ldots, \ell^{\max} \leq \alpha$. Consideramos ahora un arbol 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 $\mathfrak{c}_1, \ldots, \mathfrak{c}_d$. Esto nudos se dividen cada uno en $d$ otras ramas, y los nudos de "padre" $\mathfrak{c}_i$ se va a asociar las palabras códigos $\mathfrak{c}_i\mathfrak{c}_1, \ldots, \mathfrak{c}_i\mathfrak{c}_\alpha$, etc. Este arbol, conocido como arbol de Kraft, es ilustrado en la figura Fig. 2-36 para $d = 2$ y $\mathcal{C} = \{0\,,\, 1\}$. Claramente, $n_1 \leq d$ si no $n_1\, d^{-1} > 1$ y los largos no podrían satisfacer la desigualdad de Kraft-McMillan. El principio es entonce de asociar a los $n_1$ (posiblemente igual a 0) largos iguales a $1$ unos nudos con las palabras código asociadas de largo $1$ (primera profundez de ramas) y de prohibir todos las ramas de padre los nudos seleccionados (lineas punteadas en la figura Fig. 2-36). Estos nudos son llamados *hojas* (no hay ramas). En la capa de "hijos" de profundez/largos $2$, quedan $d^2 - n_1\, d$ nudos (accessibles) que se pueden 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 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 profundez), y de prohibir que salen de estos nudos nuevas ramas (son entonces hojas en la segunda profundez), etc. Haciendo así, se asocia un código $c$ de largos $l_c(x) = \ell_x$ que aparece libre de prefijo, es decir instantaneo. Entonces, este código es también descifrable. □

A este punto, se menciona 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 instantaneo. Claramente, se puede buscar un código de largo promedio mínimo en los códigos instantaneo, sin perdida de optimalidad (buscar en la clase más amplia de los descifrable no permite bajar el largo promedio).
- En los códigos libres de prefijo, si se fija el número de hojas (última profundez) 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 instantaneos (Kraft Jr, 1949; Karush, 1961).
- El teorema se generaliza obviamente para codificar una fuente (discreta) con un número infinito de estados, tomando el límite $\alpha \to \infty$.
- Si se conocen los largos óptimos, es suficiente para poder construir un código libre de prefijo.

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

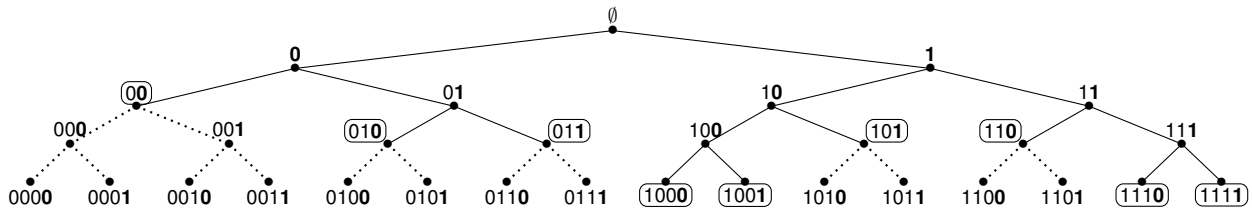



**Figura 2-36:** Arbol de Kraft en el caso binario ($d = 2$). La raíz, de código $\emptyset$ de largo 0, se divide en dos ramas, de códigos respectivamente $0$ y $1$ (profundez $1$). Cada nudo de esta profundez se divide en dos ramas (profundez dos), dando cuatros nuevos nudos con los códigos $00$ y $01$ de padre $0$, y $10$ y $11$ de padre $1$. Etc. En cada nudo 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 boran las ramas saliendo de un nudo-palabra código (ramas punteadas).

**Teorema 2-36** (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

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

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

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

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

Este resultado significa que la taza de compresión sin perdida 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 2-37** (Cota superior del código descifrable óptimo)**.** *El largo promedio $L^{\mathrm{opt}}$ del código $c^{\mathrm{opt}}$ descifrable, de largo promedio mínimo es acotado por arriba por la entropía de Shannon de base $d$ de $X$ más un dit (1 símbolo de $\mathcal{C}$),*

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

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

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

Escribiendo $l_c(x) = \log_d d^{-l_c(x)}$, se puede expresar el largo promedio de la forma

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

Olvidando que los $l_i \equiv l_c(x_i)$ 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 [55], optimización con vínculos tipo desigualdades (Miller, 2000; Cambini & Martein, 2009), conduciendo a los "largos"

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

$\widetilde{l}(x)$ no es necesariamente entero, así que una posibilidad para volver a largos enteros puede ser de tomar la parte entera superior de $\widetilde{l}(x)$,

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

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

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

□

De

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

se revela el rol fundamental de la entropía en la codificación de fuente sin perdida. La codificación es a veces dicha *codificación entrópica* y da un rol operacional a la entropía de Shannon. Se notará también que de la demostración precediente de que aparece un código particular a través de los $\left\lceil -\log_d p_X(x) \right\rceil$:

**Definición 2-53** (Código de Shannon)**.** *Un código $c^{\mathrm{sh}}$ de una fuente $X$, de largos $l^{\mathrm{sh}}(x) = \left\lceil -\log_d p_X(x) \right\rceil$, libre de prefijo (construido sobre el arbol de Kraft) es llamado código de Shannon.*

Obviamente, también

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

Al lo contrario de primer vista, un código de Shannon no es óptimo, como se lo puede ver con el ejemplo siguiente.

**Ejemplo 2-20.** *Sea $\mathcal{X} = \mathcal{C} = \{0\,,\,1\}$ y una fuente $X$ tal que $p_X(0) = 0{,}999 = 1 - p_X(1)$. Los largos de Shannon van a ser $l^{\mathrm{sh}}(0) = 1$ y $l^{\mathrm{sh}}(1) = 10$, y el largo promedio vale $L\left(c^{\mathrm{sh}}\right) = 1{,}009$. Obviamente, un código óptimo es $c(x) = x$ de largos $l_c(x) = 1$ dando $L^{\mathrm{opt}} = 1$ bit.*

---

[55] Ver nota de pie 50 pagina 119.

De hecho, volviendo al problema con largos virtualmente no enteros, el mínimo se alcanza para $\widetilde{l}(x) = -\log_d p_X(x)$, es decir que, los largos siendo enteros, se alcanza la cota mínima del código óptimo si y solamente si $-\log_d p_X(x)$. Una distribución satifaciendo esta condición es dicha $d$-ádica. Sin embargo, el código de Shannon es "competitivo" en el sentido de que:

**Teorema 2-38** (Competitividad del código de Shannon)**.** *Sea $X$ fuente sobre $\mathcal{X}$, de distribución $p_X$ y $c^{\mathrm{sh}}$ el código de Shannon asociado sobre el alfabeto código $\mathcal{C} = \{\mathfrak{c}_1, \ldots, \mathfrak{c}_d\}$, de largos $l^{\mathrm{sh}}(x) = \Big\lceil -\log_d p_X(x) \Big\rceil$. Para cualquier código $c$ descifrable y cualquier $k \geq 1$,*

$$P\Big(l^{\mathrm{sh}}(X) \geq l_c(X) + k\Big) \leq \frac{1}{d^{k-1}}.$$

*Demostración.* Por definición de la parte entera superior, $a + 1 > \lceil a \rceil$, así que $\lceil a \rceil \geq b \;\Rightarrow\; a > b - 1$. A continuación, de la implicación de eventos $(Y \geq a) \subset (Y \geq b - 1)$ dando $P(A \geq a) \leq P(Y \geq b - 1)$ y de la definición de un código de Shannon se obtiene

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

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

$$P\Big(l^{\mathrm{sh}}(X) \geq l_c(X) + k\Big) \leq \frac{1}{d^{k-1}} \sum_{x \in \mathcal{X} : p_X(x) \leq d^{-l_c(x)-k+1}} d^{-l_c(x)} \leq \frac{1}{d^{k-1}} \sum_{x \in \mathcal{X}} d^{-l_c(x)}$$

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

Este teorema traduce el hecho de que a pesar de que $c^{\mathrm{sh}}$ no sea óptimo, tomando cualquier otro código, incluyendo el óptimo, la probabilidad que $c^{\mathrm{sh}}(X)$ tenga un largo más grande que $c(X) + k$ decrece exponencialmente con $k$.

**Ejemplo 2-21.** *De manera general, con $d = 2$ y para $k = 9$, $P\Big(l^{\mathrm{sh}}(X) \geq l_c(X) + 9\Big) \leq 0{,}391\,\%$. En particular, en el ejemplo 2-20, notando que sólo en la codificación de $x = 1$ se puede tener $l^{\mathrm{sh}}(x) \geq l_c(x) + 9$, si no se usa 1 bit, este resultado significa que la probabilidad de usar más de 1 bit con el código de Shannon es menor que $0{,}391\,\%$. De hecho, una palabra código de largo $10$ aparece con una probabilidad $0{,}1\,\%$...*

En el problema de minimización, el hecho de que los largos deben ser enteros no permite solucionar explicitamente el problema de busqueda 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* [56] $c^{\mathrm{fa}}$ se basa

[56] 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).

sobre el hecho de que se alcanza la cota mínima para una distribución $d$-ádica.

**Definición 2-54** (Código de Fano)**.** *El principio es de clasificar los estados de $\mathcal{X}$ para obtener las probabilidades clasificadas en orden decreciences ($p_X^{\downarrow}$). 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 $\mathfrak{c}_i$ 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 $\mathfrak{c}_i\mathfrak{c}_j$, etc. Eso es ilustrado en la figura Fig. 2-37-(a).*

**Probar/mencionar que también**

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

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 $\mathfrak{c}_i$ da los mismos largos y el mismo largo promedio sin cambiar el aspecto libre de prefijo. De la misma manera, en el arbol de Kraft, en cada profundez se puede permutar los símbolos asociados a las hojas de esta profundez 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) [57]

**Definición 2-55** (Código de Huffman)**.** *Suponemos que existe un $\beta \in \mathbb{N}$ tal que* [58] $\alpha = |\mathcal{X}| = d + \beta(d-1)$*. 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: por cambio de escritura, llamamos $p_i$ las probabilidades rearregladas y $x_i$ los símbolo fuente correspondientes.*
2. *A cada $x_i$, $\alpha - d + 1 \leq i \leq \alpha$, associar un nudo y la letra "hijo" $\mathfrak{c}_i$.*
3. *Crear $d$ ramas saliendo de un nudo padre hasta los $d$ nudos $x_i$, $\alpha - d + 1 \leq i \leq \alpha$.*
4. *Crear un nuevo conjunto de símbolos fuente $\widetilde{x}_i = x_i$, $1 \leq i \leq \alpha - d$ de probabilidades respectivas $\widetilde{p}_i = p_i$ y $\widetilde{x}_{\alpha-d+1} = \{x_j,\ \alpha - d + 1 \leq j \leq \alpha$ de probabilidad $\widetilde{p}_{\alpha-d+1} = p_{\alpha-d+1} + \ldots + p_\alpha$. El útimo "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 $p \equiv \widetilde{p}$ y $x \equiv \widetilde{x}$.*

---

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

[58] Si no, se puede elegir $\beta = \left\lceil \frac{\alpha - d}{d-1} \right\rceil$, y completar $\mathcal{X}$ con $d + \beta(d-1) - \alpha$ símbolos fuente fictivos de probabilidades ceros, lo que no va a cambiar ni la entropía, ni el largo promedio del códico aferente.

*Como descrito tratando del código usando el arbol de Kraft,* $c^{\text{huf}}(x_i)$ *se construye saliendo de la raíz del arbol así construido, agregando las letras del camino que llega hasta la hoja* $x_i$*. Eso es ilustrado en la figura Fig. 2-37-(b) en el caso binario.*

Se mencionará que a cada etapa, el nuevo conjunto de super-símbolos fuente contiene exactamente $d-1$ símbolos menos que a la etapa precediente. Así, con $\alpha = d + \beta(d-1)$ el algoritmo tiene exactamente $\beta + 1$ bucles y en cada profundez no hay ningún nudo vacio 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 exemplo, con $d = 3$ si tuvieramos $\alpha = 4$, en la segunda etapa tendriamos $2$ estados a juntar, dando un código de largos $2, 2, 2, 1$. Empezando la primera etapa con la asociaciación de $2$ estados, es decir $3$ teniendo en cuenta un estado fictivo ($\alpha = 5$, $\beta = 1$) 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 2-39** (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 descrifrables y los libre de prefijo (se recordará que con los largos de códigos descifrables, siempre se puede construir un código libre de prefijo), es decir* $L^{\text{opt}} = L\left(c^{\text{huf}}\right)$*.*

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

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

- Sea $\eta$ el número de símbolos fuente con un código de largo máximo $l_{\text{máx}}$ y $\eta' = \min(\eta, d)$. Del punto anterior, los $\eta$ símbolos con palabra código de largo máximo son los de probabilidades más pequeñas.

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

- Suponemos que $\eta' = \eta < d$. Sea una "super-fuente" $\mathcal{X}^{(2)} = \left\{x_i^{(2)}\right\}_{i=1}^{\alpha-\eta'+1}$ con $x_i^{(2)} = x_i, \quad 1 \leq i \leq \alpha - \eta'$ de probabilidades respectivas $p(x_i)$ y $x_{\alpha-\eta'+1}^{(2)} \equiv \{x_i\}_{i=\alpha-\eta'+1}^{\alpha}$ de probabilidad $p_{\alpha-\eta'+1} + \cdots + p_\alpha$ (se "plegan" las $\eta'$ hojas en un super-símbolo). La códificación óptima es entonces una codificación libre de prefijo de $\mathcal{X}^{(2)}$, "arbol raíz" del código óptimo, a la cual se añade una letra código $\mathfrak{c}_j$ diferente a cada símbolo del super-símbolo $x_{\alpha-\eta'+1}^{(2)}$. La profundez máxima del código arbol es $l_{\text{máx}} - 1$ y debe ser llena, en el sentido de que no debe tener un nudo que sea ni una hoja, ni un prefijo. En el caso contrario, se podría desplazar un símbolo de $x_{\alpha-\eta'+1}^{(2)}$ al nudo "vacio" de la profundez $l_{\text{máx}} - 1$, 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 códificar $\mathcal{X}^{(2)}$, se necesita por lo menos $\lceil \log_d(\alpha - \eta' + 1) \rceil$ profundez en el arbol raíz. En esta profundez (máxima en el caso optimista), hay $d^{\lceil \log_d(\alpha-\eta'+1) \rceil} \geq \alpha - \eta' + 1$ nudos. En la última profundez pueden ser todos ocupados si y solamente si $d^{\lceil \log_d(\alpha-\eta'+1) \rceil} = \alpha - \eta' + 1$. En otras palabras, es posible si y solamente si existe un entero $k$ tal que $\alpha - \eta' + 1 = d^k$, es decir, con $\alpha = d + \beta(d-1)$, que teniamos el entero $\beta = \frac{d^k - d}{d-1} + \frac{\eta' - 1}{d-1}$. La primera fracción $\frac{d^k-d}{d-1} = d^{k-1} + \cdots + 1$ siendo entera, $\beta$ no puede ser entero con $\eta' < d$. En otros términos, necesariamente $\eta' = d$, *i. e.,* los $d$ símbolos de probabilidad más debiles son el la última profundez y se puede elegir que compartent el mismo nudo padre.

- Sea $c^{\mathrm{opt},(1)}$ el código óptimo correspondiente a $\mathcal{X}$ y $c^{(2)}$ el código "padre" sobre $\mathcal{X}^{(2)}$ ($c^{\mathrm{opt},(1)}$ quitando la último letra código de los símbolos juntados, *i. e.,* con la raíz común de estos). De la misma manera, sea $c^{\mathrm{opt},(2)}$ un código óptimo sobre $\mathcal{X}^{(2)}$ y $c^{(1)}$ el que se obtiene deplegando el super-símbolo $x^{(2)}_{\alpha-d+1}$ en $d$ hojas. De $L^{\mathrm{opt},(1)} = L\left(c^{(2)}\right) + p_{\alpha-d+1} + \cdots + p_\alpha$ (pasar de $\mathcal{X}^{(2)}$ a $\mathcal{X}$ se añade solo una letra palabra a los símbolos del super-símbolo) y $L\left(c^{(1)}\right) = L^{\mathrm{opt},(2)} + p_{\alpha-d+1} + \cdots + p_\alpha$ se obiene $\left(L^{\mathrm{opt},(1)} - L\left(c^{(1)}\right)\right) + \left(L^{\mathrm{opt},(2)} - L\left(c^{(2)}\right)\right) = 0$. Cada 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 $c^{\mathrm{opt},(1)}$ queda óptimo, $c^{(2)} \equiv c^{\mathrm{opt},(2)}$ (y $c^{(1)} \equiv c^{\mathrm{opt},(1)}$).

- Notando que $|\mathcal{X}^{(2)}| = \alpha - (\beta - 1)(d-1)$, el razonamiento se propaga por inducción, pasando de $c^{\mathrm{opt},(k)}$ a $c^{\mathrm{opt},(k+1)}$ juntando los $d$ super-símbolos de probabilidades más debiles, hasta tener un super-símbolo tendiendo todos los símbolos, $\left|\mathcal{X}^{(K)}\right| = 1$, raíz del arbol.

□

De esta prueba, se puede ver que

- Cada profundez siendo llena, los largos obtenidos van a saturar la desigualdad de Kraft-McMillan.

- Si $\frac{\alpha-d}{d-1}$ no es entero, en lugar de completar $\mathcal{X}$ con símbolos fictivos se puede empezar el algoritmo de Huffman juntando los $\alpha - d - \left\lfloor \frac{\alpha-d}{d-1} \right\rfloor (d-1) + 1$ símbolos fuentes de probabilidades más debiles 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, pero no es contradictorio con el punto anterior que contaba los largos de estados fictivos(que no codificamos en realidad).

- Obviamente, en el caso binario $d = 2$, no es necesario completar $\mathcal{X}$ por estados fuentes, o empezar con menos de $d$ símbolos juntados ($\alpha$ es necesariamente de la forma $\alpha = d + \beta(d-1) = 2 + \beta$).

- 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 el código). 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 $H_d(X) = \log_d |\mathcal{X}|$, precisamente la incerteza del enfoque de Hartley que corresponde a los números de dits necesarios para codificar (ingenuosamente) la fuente.

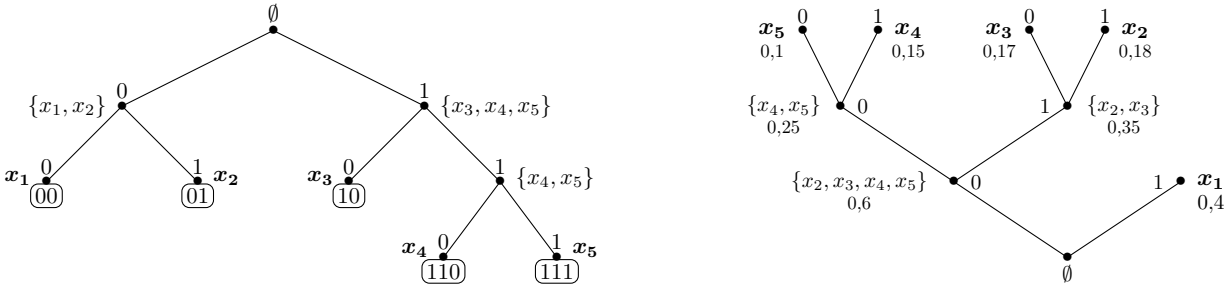



**Figura 2-37:** Construcción de un código binario sobre $\mathcal{C} = \{0, 1\}$ asociado al vector de probabilidad $p_X = \begin{bmatrix}0{,}4 & 0{,}18 & 0{,}17 & 0{,}15 & 0{,}1\end{bmatrix}^t$ sobre el arbol de Kraft. En este caso, $H_2(X) \approx 2{,}1514$ (a): Enfoque de Fano, saliendo de la raíz. 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 profundez a la otra dividiendo los conjunto 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 $L\left(c^{\text{fa}}\right) = 2{,}25$. (b): Enfoque de Huffman, saliendo de las hojas. En cada nudo, se menciona el correspondiente (i) conjunto de símbolos, (ii) $\mathfrak{c}_i$ de esta profundez/posición, (iii) la probabilidad asociada al conjunto. Se pasa de una profundez a la otra juntando los conjuntos menos probables en sobre-conjuntos. En negro son indicados los 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 $\begin{bmatrix}X_1 & \cdots & X_n\end{bmatrix}^t$ viviendo sobre $\mathcal{X}^n$, llamado *extensión de orden $n$ de la fuente*, con un código descifrable (o libre de prefijo) de esta extensión; es llamado *codificación de la extensión de la fuente* pero no es necesariamente una extension de $c$. Así, $H_d(X_1, \ldots, X_n) \leq L^{\text{opt},n} < H_d(X_1, \ldots, X_n) + 1$, 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 $\lim_{n\to\infty} \dfrac{L^{\text{opt},n}}{n} \to H(X)$, $\frac{L^{\text{opt},n}}{n}$ no es necesariamente decreciente con respeto a $n$. Eso es descrito figura Fig. 2-38. Lo mismo puede ocurir 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.

Para codificar una fuente, que se haga el código óptimo de Fano, 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 códifica con una distribución que no es la distribución verdadera de la fuenta. 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:

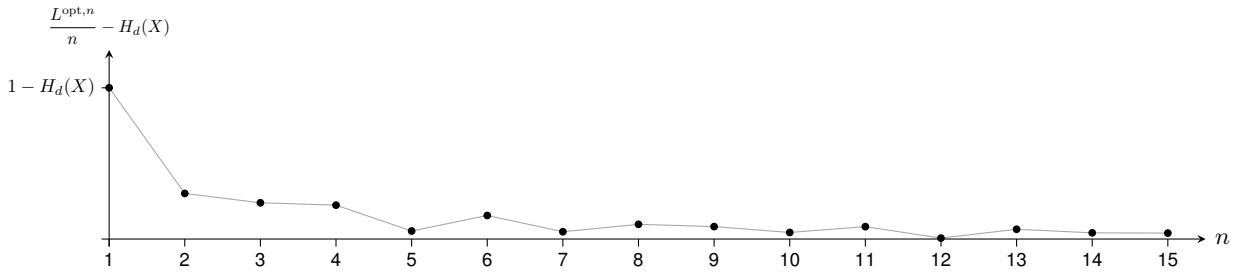



**Figura 2-38:** $\frac{L^{\mathrm{opt},n}}{n} - H_d(X)$ (puntos), diferencia entre el largo promedio óptimo por símbolo de las extensiones $\mathcal{X}^n$ de orden $n$ de la fuente $\mathcal{X}$ y la cota inferior en función de $n$. La linea llena en grise sirve como guía. En esta ilustración se usa el ejemplo lo más simple con $d = 2$ y $p = [0{,}33 \quad 0{,}67]^t$.

**Teorema 2-40** (Código falso de Shannon)**.** *Sea $c^{\mathrm{sh}}(p)$ el código de Shannon sobre el alfabeto código $\mathcal{C} = \{\mathfrak{c}_1, \ldots, \mathfrak{c}_d\}$ asociado a la distribución $p$. Sea $X$ fuente sobre $\mathcal{X}$, de distribución $p_X$ y $q$ una distribución cualquiera (ej. estimada de $p_X$ presupuesta. . . ). Entonces el largo promedio $L_{c^{\mathrm{sh}}(q)}$ del código $c^{\mathrm{sh}}(q)$ aplicado a la fuente $X$ satisface las desigualdades siguientes*

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

*Demostración.* Por definición,

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

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

Olvidando el posible extra dit (pensar a la códification por bloques), este teorema da una interpretación operacional a la entropía relativa, o divergencia de Kullback-Leibler. Esta cantidad cuantifica la perdida 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 códifica la fuente $X$, pero suponiendo la distribucíon $q$, se piedre lo que representa la información relativa de $p_X$ con respeto 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 drasticamente el mensaje. Existen por ejemplo codificaciones que permiten una corrección de error en la recepción. Pueden tomar en cuenta las caracteristicas 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 generalises Gok75 etc**

## 2.6 Entropías y divergencias generalizadas

**Partout, voir convexite stricte, en 1, idem pour h monotone, etc. vis a vis des cas d'egalite avec les divergences.**

A pesar de que la entropía de Shannon y sus cantidades asociadas demostraron sus potencias tan de un punto de vista descriptivo que en término de aplicaciones en la transmisión de la información y la compresión, varias nociones informacionales, tipo entropías o divergencias, aparecieron luego. En esta sección no se desarollará todos los enfoques ni todas las aplicaciones tan la literatura es importante. La meta es dar los caminos conduciendo a las generalizaciones de la 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 deseable que a cada entropía sean asociados nociones de entropías condicionales y relativas.

### 2.6.1 Entropías y propiedades

Si la entropía de Shannon fue el punto de salida fundamental en todo el desarollo de la teoría de la información, un poco más de una decada después 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) para probabilidades incompletas [59] $p = \begin{bmatrix} p_1 & \cdots & p_n \end{bmatrix}^t$, $p_i \geq 0$, $w_p = \sum_i p_i \leq 1$: (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) $H\left(\frac{1}{2}\right) = 1$, (iv) la aditividad $H(p \otimes q) = H(p) + H(q)$ donde $p \otimes q$ es el producto de Kronecker o tensorial [60], *i. e.,* probabilidad conjunta de dos variables independientes, y consideró en lugar de la recursividad un axioma dicho de valor promedio, axioma muy parecido a la recursividad. Para $p$ y $q$ probabilidades incompletas tales que $p \cup q = \begin{bmatrix} p_1 & \cdots & p_n & q_1 & \cdots & q_m \end{bmatrix}^t$ sea incompleta ($w_p + w_q \leq 1$), el axioma (v) es $H(p \cup q) = \frac{w_p\, H(p) + w_q\, H(q)}{w_p + w_q}$. Demostró que con (v) en lugar de la recursividad, el conjunto de axiomas conduce de nuevo a la entropía de Shannon. La generalización propuesta por Rényi era de generalizar el axioma (v) reemplazando la media

[59] En esta sección, los $p_i$ son componentes del vector $p$ no necesariamente asociado a una variable aleatoria; Hay que entender de que $p_i = p_X(x_i)$ si son asociados a una variable aleatoria $X$.

[60] Ver nota de pie 12 pagina 39.

aritmética por una media generalizada (v') $H^{\mathrm{r}}(p \cup q) = g^{-1}\left(\frac{w_p\, g\left(H^{\mathrm{r}}(p)\right)+w_q\, g\left(H^{\mathrm{r}}(q)\right)}{w_p+w_q}\right)$ con $g$ estrictamente monótona y continua, llamado media *cuasi-aritmética, o cuasi-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 $H^{\mathrm{r}}(p_i)$ y reemplazar la media aritmética $\sum_i p_i H^{\mathrm{r}}(p_i)$ por una media de Kolmogorov-Nagumo, $g^{-1}\left(\sum_i p_i g\big(H^{\mathrm{r}}(p_i)\big)\right)$. Rényi propusó la función de Kolmogorov-Nagumo $g_\lambda(x) = 2^{(\lambda-1)x}, \quad \lambda > 0, \quad \lambda \neq 1$, probando de que los axiomas (i)-(ii)-(iii)-(iv)-(v') se cumplen, conduciendo a la entropía de Rényi de un vector de probabilidad $p$,

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

Relaxando el axioma (iii), se puede elegir $g_\lambda(x) = a^{(\lambda-1)x}, \quad a > 0, \quad a \neq 1$; 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* $\lambda$. Notablemente,

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

En otros términos, la clase de Rényi contiene como caso particular 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] (additividad) 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 entopia de Shannon, *i. e.,* considerando (i) la invarianza por permutación, (ii) la continuidad, (iii) la expansividad, (iv) $H^{\mathrm{hc}}(1) = 0$ y $H^{\mathrm{hc}}\left(\frac{1}{2},\frac{1}{2}\right) = 1$, pero generalizando la recursividad por (v) $H^{\mathrm{hc}}(p_1,\ldots,p_n) = H^{\mathrm{hc}}(p_1,\ldots,p_{n-2},p_{n-1}+p_n) + \lambda(p_{n-1}+p_n)^\lambda H^{\mathrm{hc}}\left(\frac{p_{n-1}}{p_{n-1}+p_p},\frac{p_n}{p_{n-1}+p_p}\right), \quad \lambda > 0$ [61]. Con $\lambda = 1$ se recupera la recursividad estandar, pero con $\lambda \neq 1$ eso permite dar un peso diferente a la incerteza del estado interno, *i. e.,* a las probabilidades que se juntan (la describen como clasificación refinada). Estos axiomas conducen necesariamente a la entropía (teorema 1 del papel)

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

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

$$H^{\mathrm{hc}}_1(p) \equiv \lim_{\lambda\to 1} H^{\mathrm{hc}}_\lambda(p) = H(p) \quad \text{entropía de Shannon.}$$

[61] En sus papel, lo imponen para cualquier par $(p_i, p_j)$ sin imponer la invarianza por permutación, pero es equivalente a la exposición de este párrafo.

Por axioma, se conservan las propiedades [P1] (continuidad) y [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 uniforma (teorema 2). Aun que no aparece así en el papel, satisface la propiedad de Schur-concavidad [P9] (teorema 3). A pesar de que mencionan que $H_\lambda^{\mathrm{hc}}$ sea diferente de $H_\lambda^{\mathrm{r}}$, es sencillo ver que hay un mapa uno-uno entre las dos entropías. Se mencionarán en un marco más general otras propiedades de esta entropía.

Independiente de Havrda & Charvát, de la escuela húgara de la teoría de la información, Z. Daróczy en (Daróczy, 1970) definó la entropía $H^f$ a partir de una *función información* $f$ satifaciendo (i) $f(0) = f(1)$, (ii) $f\left(\frac{1}{2}\right) = 1$ y la ecuación funcional (ii) $f(x) + (1-x)f\left(\frac{y}{1-x}\right) = f(y) + (1-y)f\left(\frac{x}{1-y}\right)$ sobre $\{(x,y) \in [0;1)^2, \quad x+y \leq 1\}$, siendo $H^f(p) = \sum_{i=2}^n s_i f\left(\frac{p_i}{s_i}\right), \quad s_i = \sum_{j=1}^{i-1} p_j$. Daróczy mostró que si $f$ es medible, o continua en $0$, o no 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 términos, su axioma (v) es alternativa a la recursividad. Para extender la entropía de Shannon, propuso extender este axioma (v) por la ecuación funcional $f_\lambda(x)+(1-x)^\lambda f_\lambda\left(\frac{y}{1-x}\right) = f_\lambda(y)+(1-y)^\lambda f_\lambda\left(\frac{x}{1-y}\right)$, lo que condujo necesariamente a la entropía (teoremas 2 y 3)

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

es decir nada más que la entropía introducida por Havdra & Charvát. En lo que sigue, se la denotará $H_\lambda^{\mathrm{hcd}}$. 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 probó que se conserva la propiedad [P2] (invarianza por permutación, que no era un axioma en su enfoque), $H_\lambda^{\mathrm{hcd}}\left(\frac{1}{2},\frac{1}{2}\right) = 1$ (lo llama normalización), la expansividad [P6], una aditividad 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 $H_\lambda^{\mathrm{hcd}}\left(\frac{1}{\alpha},\ldots,\frac{1}{\alpha}\right)$ crece con el cardinal $|\mathcal{X}| = \alpha$. Muy interesante también es que se puede definir una entropía condicional en el mismo modelo que en el caso de Shannon, $H_\lambda^{\mathrm{hcd}}(X|Y) = \sum_y \left[p_{X|Y=y}(x)\right]^\lambda H_\lambda^{\mathrm{hcd}}(p_{X|Y=y})$, que existe una regla de cadena [P14], $H_\lambda^{\mathrm{hcd}}(X,Y) = H_\lambda^{\mathrm{hcd}}(Y) + H_\lambda^{\mathrm{hcd}}(X|Y)$ y que condicionar reduce la entropía $H_\lambda^{\mathrm{hcd}}(X|Y) \leq H_\lambda^{\mathrm{hcd}}(X)$ (teorema 8) [P16]. Mostró también que si se pierde la aditividad, se obiene para $X$ e $Y$ independientes $H_\lambda^{\mathrm{hcd}}(X,Y) = H_\lambda^{\mathrm{hcd}}(X) + H_\lambda^{\mathrm{hcd}}(Y) + \left(2^{1-\lambda}-1\right)H_\lambda^{\mathrm{hcd}}(X)H_\lambda^{\mathrm{hcd}}(Y)$. La propiedades de regla de cadena le permitió revisitar la caracterisación de un canal de transmisión y redefinir una capacidad canal extendidas (capacidad tipo $\lambda$; basicamente se usa el mismo enfoque que Shannon, pero usando $H_\lambda^{\mathrm{hcd}}$ en lugar de $H$, ver sección 6 del papel).

Las entropías tipo Havdra-Charvát-Daróczy fueron (re)descubiertos varias otras veces y/o estudiadas 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; Boekee & 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 decada 1970 (Arimoto, 1971). Fue rediscubierto y estudiado con más detalles una decada después por J. Burbea y C. R. Rao (Burbea & Rao, 1982) y luego por M. Salicrú (Salicrú, 1987) o M. Teboulle (Teboulle, 1992) entre otros. La medida propuesta, llamada $\phi$-*entropía*, es definida por

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

Burbea y Rao asociaron una medida de divegencia a esta entropía. Las $\phi$-entropías contienen Shannon como caso particular ($\phi(x) = x\log x$), así que la clase de Havdra-Charvát-Daróczy ($\phi(x) = \frac{x-x^\lambda}{2^{1-\lambda}-1}$) 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 acà substancialmente su escritura de la literatura por razones de homogeneidad con la $\phi$-entropía (y las divergencias que se introducirán luego) [62]

**Definición 2-56** ($(h,\phi)$-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*

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

*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 $\phi(0) = 0$ (la incerteza elemental asociada a un estado de probabilidad nula vale cero) y, sin perdida de generalidad, que $h(-\phi(1)) = 0$.*

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

Como en el caso de Shannon, para $X = (X_1, \ldots, X_d)$, 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 las propiedades [P1] (continuidad), [P2] (invariaza por permutación), [P3] (invarianza por transformación biyectiva de $X$), [P6] (expansabilidad, debido a $\phi(0) = 0$).

Además se conserva la Schur-concavidad con una reciproca:

---

[62] En la literatura, no hay el signo $-$, y hay que invertir cóncava y convexa.

[P$_\phi$9] Schur-convavidad:

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

En otros términos, se obtiene la relación de mayorisación si se cumple la relación de ordre entrópicas para todos los pares 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, y reciprocamente.

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

$$H_{(h,\phi)}\left(\begin{bmatrix}\frac{1}{\alpha} & \cdots & \frac{1}{\alpha}\end{bmatrix}^t\right) \quad \text{función creciente de } \alpha.$$

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

[P$_\phi$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 $\lambda < 1$, la entropía de Rényi es cóncava, pero se proba que existe un $\lambda^*(\alpha) > 1$ tal que para cualquier $\lambda \leq \lambda^*(\alpha)$ 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$_\phi$7] Sean $X$ definido sobre $\mathcal{X}$ y $\breve{X}$ sobre $\breve{\mathcal{X}}$,

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

$$H_{(h,\phi)}(p_X) \geq H_{(h,\phi)}\left(p_{\breve{X}}\right).$$

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). Aparte en el caso de Shannon y el de Havdra-Charvát-Daróczy, no hay ningún vínculo explicito general entre $H_{(h,\phi)}(p_X)$ y $H_{(h,\phi)}\left(p_{\breve{X}}\right)$.

Se conserva la super-aditividad [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, $\phi\left(p_{X,Y}(x_i, y_j)\right) =$

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

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

**Teorema 2-41.** *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).*

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

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

*Demostración.* La reciproca de este teorema es nada más que la propidad [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 cardenal 3 cada unos se proba de que la desigualdad no se puede cumplir, salvo si la derivada $\phi'$ de la función entrópica satisface a una ecuación funcional.
- la sola solución admisible de esta ecuación se reduce a $\phi(x) = -x\ln x$.

**Etapa 1**: Sea el vector de probabildad

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

donde $(a,\tilde{a},b,\tilde{b}) \in D$,

$$D = \{(u,\tilde{u},v,\tilde{v}) \in [0\,;\,1]^4 : \quad 0 < \tilde{u} \leq 1-u \quad \& \quad 0 < \tilde{v} \leq 1-v\},$$

y $c \in C_{a,\tilde{a},b,\tilde{b}}$,

$$C_{a,\tilde{a},b,\tilde{b}} = \left[-1 + \max\left\{ab, \tilde{a}\tilde{b}, 1-a\tilde{b}, 1-\tilde{a}b\right\},\, \min\left\{ab, \tilde{a}\tilde{b}, 1-a\tilde{b}, 1-\tilde{a}b\right\}\right].$$

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 $u$, como ilustrado figura Fig. 2-39 (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 $C_{a,\tilde{a},b,\tilde{b}}$ no sea reducido a $\{0\}$), y $c \in \mathring{C}_{a,\tilde{a},b,\tilde{b}}$ donde $\mathring{\cdot}$ denota el interior de un conjunto, se obtiene para $\phi$ convexa,

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

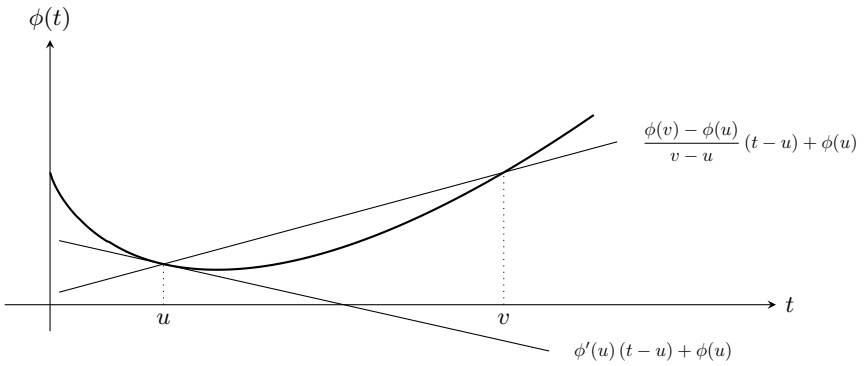



**Figura 2-39:** $\phi$ estrictamente convexa: la variación (cuerda) $\frac{\phi(v)-\phi(u)}{v-u}$ es mayor que la derivada $\phi'(u)$. Aplicado a dos distribuciones $p$ y $q$, de componentes $p_i$ y $q_i$, con $u = p_i$ y $v = q_i$ y sumando, se obtiene $H_\phi\,(q) - H_\phi\,(p) \geq \sum_i (p_i - q_i)\phi'(p_i)$ con $H_\phi \equiv H_{(\mathrm{id},\phi)}$, id siendo la identidad.

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

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

Supongamos que existe un $(s,\tilde{s},t,\tilde{t}) \in D$, con $(s,t) \in (0\,;\,1)^2$, tal que $g(s,\tilde{s},t,\tilde{t},0) \neq 0$. De la continuidad de $\phi'$, la función $g$ es continua, entonces existe un vecinaje $V_0 \subset \mathring{C}_{s,\tilde{s},t,\tilde{t}}$ de $0$ tal que la función $c \mapsto g(s,\tilde{s},t,\tilde{t},c)$ tiene un signo constante sobre $V_0$. Eso permite concluir que $c \mapsto c\,g(s,\tilde{s},t,\tilde{t},c)$ no tiene un signo constante sobre $V_0$, y entonces de concluire que, de la desigualdad dedibo a la concavidad de $\phi$ (resp. convexidad), $H_\phi\,(p_{X,Y})$ puede ser major (resp. menor) que $H_\phi\,(p_X \otimes p_Y)$, y entonces, con la crecencia (resp. decrecencia) de $h$ que si $g(a,\tilde{a},b,\tilde{b},0)$ no es identicamente cero sobre $D$, $H_{(h,\phi)}$ no puede ser sub-aditiva (distribución conjunta vs producto de las marginales).

**Etapa 2**. De la etapa 1, se sabe que la sub-aditividad es potencialmente posible solamente si $g(a,v,s,t,0) = 0$ sobre $D_{a,\tilde{a},b,\tilde{b}} \cap (0\,;\,1)^4$. Eso significa que $\phi'$ debe necesariamente satisfacer la ecuación funcional

$$\phi'\big(ab\big) + \phi'\big(\tilde{a}\tilde{b}\big) - \phi'\big(a\tilde{b}\big) - \phi'\big(\tilde{a}b\big) = 0,$$

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

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

Eso significa de que $x\,\phi''(x)$ es constante sobre $x \in (0\,,\,(1-a)\,\mathrm{máx}\{b,1-b\})$, y para cualquier par $(a,b) \in (0\,;\,1)^2$. Entonces, $x\,\phi''(x)$ es constante sobre $x \in (0\,;\,1)$, es decir que $\phi$ es necesariamente de la forma $\phi(x) = \eta\,x\ln x + \theta x + \vartheta$. 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 $\theta = 0$ (poniendo $\theta$ adentro de $\vartheta$ sin cambiar el valor de entropía obtenida). Además, del requisito $\phi(0) = 0$ tenemos $\vartheta = 0$. Para que $\phi$ sea

convexa (resp. cóncava) hace falta tener $\eta > 0$ (resp. $\eta < 0$) así que, sin perdida de generalidad, $\eta$ 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. En conclusión, $g = 0$ sobre $D \cap (0\,;\,1)^2$ es decir, por continuidad, sobre $D$ se reduce a necesitar tener $H_\phi = H$. Esta entropía siendo sub-aditiva (propidad [P11]), cualquiera función creciente de $H$ va obviamente quedar sub-aditiva, lo que cierra la prueba. □

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) $H_{(h,\phi)}(p_{XY}) = h\left(-2\,\phi\left(\frac{1}{2}\right)\right)$, $H_{(h,\phi)}(p_X \otimes p_Y) = h\left(-4\,\phi\left(\frac{1}{4}\right)\right)$ y $H_{(h,\phi)}(p_X) + H_{(h,\phi)}(p_Y) = 2\,h\left(-2\,\phi\left(\frac{1}{2}\right)\right)$, así que, en este ejemplo $H_{(h,\phi)}(p_{XY}) > H_{(h,\phi)}(p_X \otimes p_Y)$ (consecuencia de la Schur-concavidad) y $H_{(h,\phi)}(p_{XY}) > H_{(h,\phi)}(p_X) + H_{(h,\phi)}(p_Y)$: tampoco las $(h,\phi)$-entropía son super-aditivas.

La definición de entropías generalizadas condicionales aparece mucho más problemático. Por ejemplo, si se define a la Shannon, es decir definiendo $H_{(h,\phi)}(X|Y)$ tomando $\sum_{y\in\mathcal{Y}} p_Y(y) H_{(h,\phi)}\left(p_{X|Y=y}\right)$ 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 $p_Y^\lambda$. 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 problemático también. Por ejemple 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 super-aditividad garantiza la positividad de la entropía condicional, se pierde la propiedad [P13] por perdida de la aditividad, 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 divergencias.

Como en el caso de Shannon, se puede extender la generalización de la entropía al caso de vectores aleatorios discretos sobre de cardenal infinito, con las mismas debilidades que en el caso de Shannon. A continuación, se puede también extenderla a vectores aleatorios admitiendo una densidad de probabilidad, reemplazando la suma por una integración.

**Definición 2-57** ($(h,\phi)$-entropía diferencial)**.** *Sea $X$ una variable aleatoria continua sobre $\mathbb{R}^d$ y sea $p_X(x)$ la densidad (distribución) de probabilidad de $X$ de soporte $\mathcal{X}$. La $(h,\phi)$-entropía diferencial de la variable $X$ es definida por*

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

*con $h$ y $\phi$ cumpliendo los requisitos de la definición discreta 2-56 (de $\phi(0)$, se puede escribir la integración sobre $\mathbb{R}^d$).*

De nuevo para $X = (X_1, \ldots, X_d)$, 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 con respecto a 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 propidedades, y entre otros si $h$ es cóncava, la $(h,\phi)$-entropía diferencia es cóncava [P$_\phi$8]. Más sorprendentemente a primer vista, se conserva la $(h,\phi)$-entropía diferencial bajo un rearreglo [P'2],

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

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) [63].

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.,*

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

**Quide de la reciproca? Quid sub-aditividad ssi fct creciente de Shannon?**

Terminamos esta subsección notando de que, como para la entropía de Shannon, el enfoque discreto y diferencial son contenido en la forma general usando densidades con respeto a una medida (respectivamente discreta y de Lebesgue en estos casos).

**Definición 2-58** (Escritura única de las $(h,\phi)$-entropías)**.** *Sea $X$ variable aleatoria definida sobre $\mathcal{X} \subseteq \mathbb{R}^d$, admitiendo una densidad de probabilidad $p_X$ con respeto a una medida $\mu$ (ej. $\mu_{\mathcal{X}}$ en el caso discreto $\mu = \mu_L$ en el caso diferencial). La $(h,\phi)$-entropía de $X$ con respeto a $\mu$ se escribe como*

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

*con $h$ y $\phi$ cumpliendo los requisitos de la definición discreta 2-56 (de $\phi(0)$, se puede escribir la integración sobre $\mathbb{R}^d$).*
*Insistamos de nuevo en el hecho de que se puede entender esta definición para cualquier $\mu$ y densidad con respeto a $\mu$, que sea discreta, de Lebesgue, o mixta.*

## 2.6.2 Divergencias y propiedades

[63] Recuerdense que en (Lieb & Loss, 2001, Sec. 3.3) lo muestran para $\phi$ diferencia de dos funciones monótonas, siendo una función convexa un caso particular.

En esta sub-sección vamos a ver que la literatura trató casi conjuntamente de tres enfoques dando lugar a generalizaciones de la divergencia de Kullback-Leibler. Lamentablemente, ninguna generalización contiene las otras, a pesar de que divergencias conocidas pueden partenecer a varias clases distinctas. Practicamente, cada clase tiene sus ventajas y justificación en termino de aplicaciones.

### 2.6.2.1. Clase de Jensen

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) [64],

$$D_J(Q\|P) = D_{\mathrm{kl}}\left(P\,\|Q\right) + D_{\mathrm{kl}}\left(Q\,\|P\right).$$

Esta versión simetrizada de la divergencia queda naturalmente positiva, pero sufre todavía de unas debilidades de $D_{\mathrm{kl}}$. Esta bien definida siempre que $P \ll Q$ conjuntamente a $Q \ll P$ (las medidas son dichas *medidas equivalentes* en este caso). 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 solamente si cada uno vale cero).

Unas decadas después, Lin introdujo lo que llamó $K$-divergencia directada, $K(P,Q) = D_{\mathrm{kl}}\left(P\left\|\frac{P+Q}{2}\right.\right)$, su versión simetrizada, antes de generalizarla bajo la terminologia de *divergencia de Jensen* (Lin, 1991) [65].

$$\begin{aligned} D^{\pi}_{\mathrm{js}}(P_{(1)},P_{(2)}) &= \pi_1 D_{\mathrm{kl}}\left(P_{(1)}\left\|\pi_1 P_{(1)}+\pi_2 P_{(2)}\right.\right) + \pi_2 D_{\mathrm{kl}}\left(P_{(2)}\left\|\pi_1 P_{(1)}+\pi_2 P_{(2)}\right.\right) \\ &= H(\pi_1 p_{(1)}+\pi_2 p_{(2)}) - \pi_1 H(p_{(1)}) - \pi_2 H(p_{(2)}) \qquad \pi=[\pi_1 \quad \pi_2],\quad 0\le\pi_1=1-\pi_2\le 1 \end{aligned}$$

con $p_{(i)}$ densidades de $P_{(i)}$ con respeto a una misma medida $\mu$ (puede ser discreta, de Lebesgue, o mixta).

$D^{\pi}_{\mathrm{js}}$ heride obviamente de $D_{\mathrm{kl}}$ su positividad con igualdad si y solamente si $P_{(1)} = P_{(2)}$. La misma propiedad puede ser vista a través de la desigualdad de Jensen, dando este nombre a la medida. Además, se quita el problema de definición, siendo de que $P_{(i)} \ll \pi_1 P_{(1)} + \pi_2 P_{(2)}$. No es simétrica en general, pero se obtiene esta propiedad cuando $\pi = \pi_{\mathrm{u}} \equiv [\frac{1}{2} \quad \frac{1}{2}]^t$. Además, en este caso, a pesar de que la divergencia no cumpla la desigualdad triangular, aparece que $\left(J^{\pi_{\mathrm{u}}}_{\mathrm{js}}(P_{(1)},P_{(2)})\right)^s$, $\quad 0 < s \le \frac{1}{2}$ es una metrica [66] (Österreicher & Vajda, 2003, Teorema 1 & Nota 2) o (Endres & Schindelin, 2003; Kafka, Österreicher & Vincze, 1991; Osán, Bussandri

---

[64] Esta expresión apareció en (Jeffrey, 1946, Ec. (1)) o en (Jeffrey, 1948), antes de la introduccíon de la divergencia de Kullback-Leibler en el campo de la estimación Bayesiana, Jeffrey siendo citado por Kullback y Leibler.

[65] De hecho, apareció implicitamente en varios trabajos anteriores, por ejemplo en mecanica cuántica (Holevo, 1973, 2011) o en reconocimiento de patrones (Wong & You, 1985)

[66] Se necesita soló de que los $P_{(i)}$ admiten una densidad con respeto a una medida $\sigma$-finita; nos referiremos al resultado 1-8, pagina 38.

& Lamberti, 2018). Si puede parecer más lógico definir tal divergencia con a priori/proporciones $\pi_i$ iguales, de hecho la versión no simétrica, con pesos $\pi_i$ se vuelve natural en el marco de la discriminación donde apareció implicitamente esta cantidad. En particular, cuando estamos frente a dos hypotesis $i = 1, 2$ 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 hypotesis). El enfoque Bayesiano más natural consiste maximizar la probabilidad a posteriori (probabilidad de estar en hypotesis $i$ condicionalmente a la observación), y se prueba que la probabilidad de error es dada por $P_e = \int_{\mathcal{X}} \min(\pi_1 p_{(1)}(x)\,,\, \pi_2 p_{(2)}(x))\, d\mu(x)$ con $p_{(i)}$ densidad con respeto a $\mu$ (Kay, 1993). Probó Lin de que

$$\frac{1}{4}\left(H(\pi) - D^{\pi}_{\mathrm{js}}(P_{(1)}, P_{(2)})\right)^2 \leq P_e \leq \frac{1}{2}\left(H(\pi) - D^{\pi}_{\mathrm{js}}(P_{(1)}, P_{(2)})\right),$$

lo que da naturalmente un rol operacional a esta divergencia. Incidentalmente, de esta desigualdad es inmediato ver de que $D^{\pi}_{\mathrm{js}}(P_{(1)}, P_{(2)}) \leq H(\pi) - 2P_e$. $P_e$ siendo positivo, da

$$0 \leq D^{\pi}_{\mathrm{js}}(P_{(1)}, P_{(2)}) \leq H(\pi) \leq \log(2)$$

Usando el logaritmo de base 2, adaptado a este caso de dos distribuciones, la cota vale 1: $D^{\pi}_{\mathrm{js}}$ es dicha *normalizada*.

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. 2-23-(b), pagina 85. Sea $Z$ la variable aleatoria binaria sobre $\{1, 2\}$ tal que $P(Z = i) = \pi_i$, 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 $D^{\pi}_{\mathrm{js}}(P_{(1)}, P_{(2)}) = H(p) - \sum_i \pi_i H(p_{(i)})$ viene $D^{\pi}_{\mathrm{js}}(P_{(1)}, P_{(2)}) = H(X) - H(X|Z)$, es decir

$$D^{\pi}_{\mathrm{js}}(P_{(1)}, P_{(2)}) = I(X; Z).$$

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 las cotas mayor de $D^{\pi}_{\mathrm{js}}$.

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

Más allá, en el campo de la clasificación, se puedre 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,

$$\begin{aligned} D^{\pi}_{\mathrm{js}}(P_{(1)}, \ldots, P_{(n)}) &= H\left(\sum_i \pi_i p_{(i)}\right) - \sum_i \pi_i H(p_{(i)}) \\ &= \sum_i \pi_i D_{\mathrm{kl}}\left(P_{(i)} \middle\| \sum_j \pi_j P_{(j)}\right). \end{aligned}$$

De la desigualdad de Jensen, esta cantitad queda positiva con igualdad si y solamente si todos los $P_{(i)}$ son iguales. Se conserva una cota superior

$$D^{\pi}_{\mathrm{js}}(P_{(1)},\ldots,P_{(n)}) \leq H(\pi) \leq \log(n),$$

así que $D^{\pi}_{\mathrm{js}}(P_{(1)},P_{(2)}) = I(X;Z)$ con $X$ de distribución la mezcla $\sum_i \pi_i P_{(i)}$ y $Z$ definida sobre $\{1,\ldots,n\}$ variable de selección de distribución $\pi$.

**convexidad?**

Un punto clave que dio lugar a la definición de la divergencia de Jensen-Shannon es la concavidad de la entropía de Shannon. Naturalmente, el mismo enfoque se generaliza a cualquier entropía cóncava de un vector de probabilidad. Tal generalización fue propuesta de maner formal por Burbea-Rao e (Burbea & Rao, 1982), y luego generalizado y estudiado más detenidamente por Nielsen et al. (Nielsen & Boltz, 2011; Nielsen & Nock, 2017). A pesar que que apareció ya en el papel de Burbea & Rao, Nielsen llamó tal generalización "divergencia de Burbea-Rao asimetrizada". Más formalemente, se puede definir una divergencia de Jensen de la manera siguiente:

**Definición 2-59** (Divergencias de Jensen)**.** *Sea* $f : \mathcal{U} \subset \mathbb{R}^m \mapsto \mathbb{R}$ *convexa y de clase* $C^1$ *sobre* $\mathcal{U}$*, un cerrado convexo de* $\mathbb{R}^m$ *y* $\pi = \begin{bmatrix}\pi_1 & \pi_2\end{bmatrix}^t$ *con* $0 \leq \pi_1 = 1 - \pi_2 \leq 1$*. Las divergencias de Jensen entre dos puntos* $u_1, u_2 \in \mathcal{U}$ *son definidas por*

$$J^{\pi}_f(u_1,u_2) = \pi_1 f(u_1) + \pi_2 f(u_2) - f(\pi_1 u_1 + \pi_2 u_2).$$

*Se ilustra a que corresponde esta cantidad con respecto a* $f$ *en la figura Fig. 2-40 más adelante.*

Esta definición se generalizada a densidad de probabilidad, donde $f$ es a valor reales, actuando sobre el convexo de las medidas de probabilidades (o tomando densidades en un $x$ e integrando sobre $\mathcal{X}$) (Nielsen & Boltz, 2011; Nielsen & Nock, 2017).

**Definición 2-60** (Divergencia $(h,\phi)$-Jensen entrópica)**.** *Para* $(h,\phi)$*-entropías cóncavas (ej. con* $h$ *cóncava), siendo* $-H_{(h,\phi)}$ *convexa, se puede entonces asociar una divergencia de Jensen*

$$D^{\mathrm{j},\pi}_{(h,\phi)}\left(p_{(1)},p_{(2)}\right) \equiv J^{\pi}_{-H_{(h,\phi)}}(p_{(1)},p_{(2)}) = H_{(h,\phi)}\left(\pi_1 p_{(1)} + \pi_2 p_{(2)}\right) - \pi_1 H_{(h,\phi)}\left(p_{(1)}\right) - \pi_2 H_{(h,\phi)}\left(p_{(2)}\right).$$

*con* $p_{(i)}$ *densidad con respeto a una medida* $\mu$*. Cuando* $h \equiv \mathrm{id}$*, se notará* $D^{\mathrm{j},\pi}_{\phi}$*.*

*La definición se generaliza a cualquier conjunto* $\{p_{(i)}\}_{i=1}^n$ *de distribuciones de probabilidades y* $\pi$ *vector de probabilidad* $n$*-dimensional,*

$$D^{\mathrm{j},\pi}_{(h,\phi)}(\{p_{(i)}\}) = H_{(h,\phi)}\left(\sum_i \pi_i p_{(i)}\right) - \sum_i \pi_i H_{(h,\phi)}\left(p_{(i)}\right).$$

Por analogía a la información mutua, Burbea y Rao llamarón esta medida "información mutua generalizada". Eso viene de que si se define una información condicional en el mismo esquema que el de Shannon, *i. e.,* para $Y$ discreta, $H_{(h,\phi)}(X|Y) = \sum_y p_Y(y) H_{(h,\phi)}\left(p_{X|Y=y}\right)$, entonces, con $\pi \equiv p_Y$ y $\{p_{(i)}\}_i \equiv \{p_{X|Y=y}\}_y$ aparece

de que $D^{\mathrm{j},p_Y}_{(h,\phi)}(\{p_{X|Y=y}\}_y) = H_{(h,\phi)}(X) - H_{(h,\phi)}(X|Y)$. Esta expresión es parecida a una de las formas de la información mutua de Shannon, justificando la terminilogía de Burbea-Rao. Sin embargo, hay que tener conciencia de que no todo se translata obviamente del mundo Shannon al mundo generalizado. Por ejemplo, con tal definicón de la entropía condicional, se pierde la regla de cadena, y por consecuencia la simetría de tal información mutua generalizada o la forma usando la entropía conjunta y las marginales.

Se notará de que Nielsen propusó generalizaciones mas avanzadas, usando generalizaciones de la noción de convexidad. Estas generalizaciones van más allá de la meta del capítulo y el lector so puede referir a (Nielsen & Nock, 2017).

Las divergencias de Jensen tiene las propiedades siguientes

1. Positividad:

   $$J^{\pi}_f(P,Q) \geq 0 \quad \text{con igualdad si y solamente si} \quad P = Q.$$

   Esta propiedad es la consecuencia directa de la convexidad estricta de $f$, como ilustrado figura Fig. 2-40.

2. Pensando a $J^{\pi}_f$ con respecto a $f$, es lineal en el sentido de que $J^{\pi}_{a_1 f_1 + a_2 f_2} = a_1 J^{\pi}_{f_1} + a_2 J^{\pi}_{f_2}$ (con $f_i$ convexas y $a_i \geq 0$).

Desgraciamente, las divergencias de Jensen no cumplen la desigualdad triangular en general, y entonces no son métricas entre distribuciones de probabilidad. Se refierá a (Burbea & Rao, 1982; Nielsen & Boltz, 2011; Nielsen & Nock, 2017) para tener más propiedades.

Se notará que la clase de las divergencias de Jensen contiene el cuadrado de la distancia de Mahalanobis (por un factor), *i. e.,* con $f(u) = u^t K u$ con $K > 0$ simétrica se obtiene $J_f(u,v) = \pi_1\pi_2(v-u)^t K(v-u)$ (siendo la distancia $L^2$ un caso particular). Se generaliza al caso continuo y distancias $L^2$ con un nucleo.

### 2.6.2.2. Clase de Bregman

Estas divergencias fueron intoducidos en el campo de la programación lineal convexa, para resolver problemas de minimización convexa [67] (Bregman, 1967), pero con aplicaciones en varios campos (Basseville, 1989, 2013, y ref.):

**Definición 2-61** (Divergencias de Bregman)**.** *Sea $f : \mathcal{U} \subset \mathbb{R}^m \mapsto \mathbb{R}$ convexa y de clase $C^1$ sobre $\mathcal{U}$, un cerrado convexo de $\mathbb{R}^m$. Las divergencias de Bregman de un punto $v \in \mathcal{U}$ relativamente a un punto $u \in \mathcal{U}$ son definidas por*

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

*Dicho de otra manera, $B_f$ corresponde al desarollo 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. 2-40 más adelante.*

[67] Aún que aparece en una revista de matemática y física matemática, una gracia del papel de Bregman es que toma el ejemplo de maximización de la entropía de Shannon sujeto a momentos...

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 capitulo, tratando posiblemente de densidad de probabilidades, nos interesamos a funciones de funciones (Frigyik, Srivastava & Gupta, 2008; Nielsen & Nock, 2017):

**Definición 2-62** (Divergencias de Bregman funcional)**.** *Sea $f : \mathcal{U} \mapsto \mathbb{R}$ convexa y de clase $C^1$ sobre $\mathcal{U}$, un cerrado convexo de un espacio de Banach. Las divergencias de Bregman de un "punto" (una función) $v \in \mathcal{U}$ relativamente a un "punto" $u \in \mathcal{U}$ son definidas por*

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

*El último término de esta formula es connocida como derivada de Gâteau (o derivada direccional) de $f$ en $u$ en la dirección $v - u$ (siendo $u$ una función)* [68]*.*

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

**Definición 2-63** (Divergencia $(h, \phi)$-Bregman entrópica)**.** *Para $(h, \phi)$-entropías cóncavas (ej. con $h$ cóncava), se puede entonces asociar una divergencia de Bregman*

$$D^{\mathrm{b}}_{(h,\phi)}(q\|p) = H_{(h,\phi)}(p) - H_{(h,\phi)}(q) - h'\big(H_\phi(p)\big) \int_{\mathcal{X}} (q(x) - p(x))\phi'(p(x))\, d\mu(x).$$

*Cuando $h \equiv \mathrm{id}$, se notará $D^{\mathrm{b}}_\phi$ y es equivalente a salir de la definición inicial $u = p(x)$, $v = q(x)$ y sumar la divergencia obtenida sobre $\mathcal{X}$ con respeto a la medida $\mu$. En el caso particular discreto toma la expresión*

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

Aparece de que las divergencias de Jensen se escriben a partir de divergencias de Bregman, y vice-versa:

**Lema 2-11.** *De la definiciones 2-59, 2-61 y 2-62, se muestra sencillamente de que las divergencias de Jensen se escriben como combinaciones convexas de divergencias de Bregman,*

$$J^\pi_f(P_{(1)}, P_{(2)}) = \pi_1 B_f(P_{(1)}\|\pi_1 P_{(1)} + \pi_2 P_{(2)}) + \pi_2 B_f(P_{(1)}\|\pi_1 P_{(1)} + \pi_2 P_{(2)}),$$

*Al revés, las divergencias de Bregman se escriben como limites de divergencias de Jensen,*

$$B_f(P_{(2)}\|P_{(1)}) = \lim_{\pi_2 \to 0} \frac{J^\pi_f(P_{(1)}, P_{(2)})}{\pi_1 \pi_2}$$

*(o similarmente con densidad $p_{(i)}$ con respeto a una medida $\mu$).*

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

La figura Fig. 2-40 ilustra a que corresponden $D_f$ y $J_f$ con respecto a la función convexa $f$.

La divergencia de Bregman tiene las propiedades siguientes

[68] 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 identica independientemente de la dirección. Entonces, si una función tiene una derivada de Fréchet, tiene necesariamente derivadas de Gâteau, pero no es reciproca. Esta subtileza va más allá de la meta de esta sección.

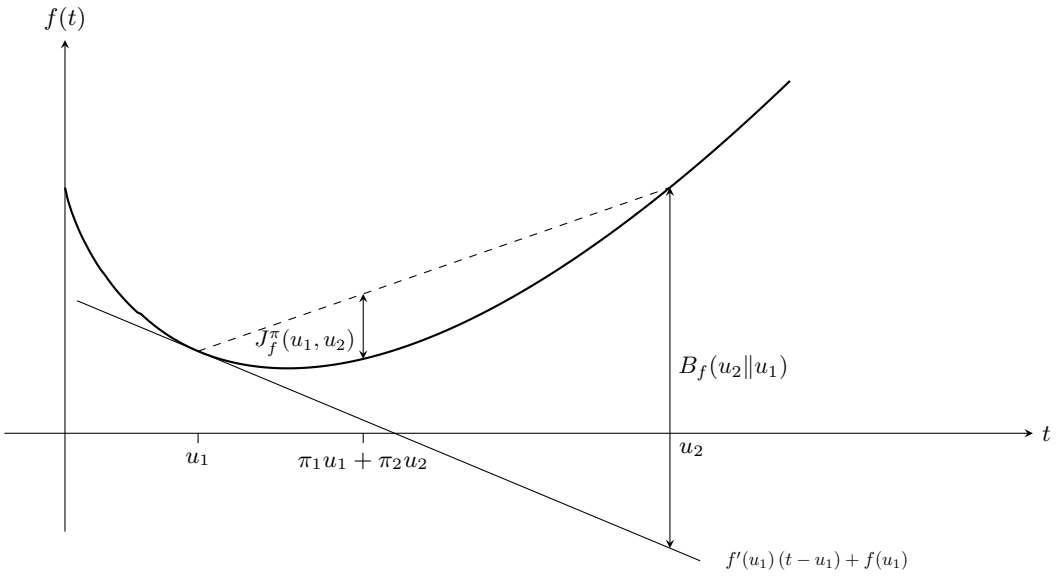



**Figura 2-40:** $f$ estrictamente convexa. Las cantidad positiva marcada por la dupla-flecha representan respectivamente la divergencia de $f$-Jensen $J_f^{\pi}(u_1, u_2)$, diferencia entre la combinaci¨on convexa de los $f(u_i)$ y $f$ de la combinación convexa de los $u_i$, y la divergencia de Bregman $B_f(u_2 \| u_1)$ 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.

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. 2-40.

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 $B_{a_1 f_1 + a_2 f_2} = a_1 B_{f_1} + a_2 B_{f_2}$ (con $f_i$ convexas y $a_i \geq 0$).

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 de Mahalanobis con $f(u) = u^t K u$ con $K > 0$ simétrica (siendo la distancia $L^2$ un caso particular), el cuadrado de la distancia $L^1$ con $f(u) = \left( \sum_i u_i \right)^2$, la distancia de Itakura-Saito cuando $f(u) = -\log u$ (asociado a la entropía de Burg), entre otros. Unas se exteinden sencillamente al caso continuo (Frigyik et al., 2008).

Como en el caso de divergencias de Jensen, Nielsen propusó generalizaciones más avanzadas, usando las generalizaciones de la noción de convexidad usada para generalizar las divergencias de Jensen. Estas generalizaciones también van más allá de la meta del capítulo y el lector so puede referir a (Nielsen & Nock, 2017).

También, varias aplicaciones se encuentran en la literatura (Basseville, 1989; Csiszár, 1995; Csiszàr & Matúš, 2012; Basseville, 2013, y ref.) en adición de las referencias de esta sección, para dar unas.

### 2.6.2.3. Clase de Csiszár o Ali-Silvey

Un primer paso generalizando la noción de entropía relativa o divergencia, siguiendo el enfoque de Kullback y Leibler y sus versiones tipo $J$-divergencia o divergencia de Jensen-Shannon fue debido a Rényi. En su papel (Rényi, 1961), A. Rényi introdujo una noción de ganancia o perdida de información de una distribucíon (incompleta) de probabilidad $q$ relativa a una referencia $p$, $I^{\mathrm{r}}(q\|p)$, teniendo un enfoque axiomático similar al que uso para definir su entropía: (i) la medida sea invariante a una misma permutación de los componentes de $p$ y de $q$, (ii) si $\forall\, i,\ p_i \leq q_i$ entonces $I^{\mathrm{r}}(q\|p) \geq 0$ y vice versa $I(p\|q) \leq 0$, (iii) $I([1]\|[1/2]) = 1$, (iv) $I(q_{(1)} \otimes q_{(2)}\|p_{(1)} \otimes p_{(2)}) = I(q_{(1)}\|p_{(1)}) + I(q_{(2)}\|p_{(2)})$ y (v) una propiedad de media generalizada $I(q_{(1)} \cup q_{(2)}\|p_{(1)} \cup p_{(2)}) = g^{-1}\left(\frac{w_{q_1} I^{\mathrm{r}}(q_{(1)}\|p_{(1)}) + w_{q_2} I^{\mathrm{r}}(q_{(2)}\|p_{(2)})}{w_{q_1} + w_{q_2}}\right)$ conduciendo a

$$I^{\mathrm{r}}_{\lambda}(q\|p) = \frac{1}{\lambda - 1} \log_2 \left( \sum_i p_i \left( \frac{q_i}{p_i} \right)^{\lambda} \right).$$

Unos años después, se introdujo una clase más general debido a I. Csiszár (Csiszár, 1963; Csiszàr, 1967; Csiszár & Shields, 2004), T. Morimoto (Morimoto, 1963) o S. M. Ali & S. D. Silvey (Ali & Silvey, 1966), clase que llamaremos *$\phi$-divergencias de Csiszàr*. De manera general, con $Q \ll P$, toma la forma

$$D^{\mathrm{c}}_{\phi}(Q\|P) = \int_{\mathcal{X}} \phi\left(\frac{dQ}{dP}(x)\right) dP(x),$$

donde $\phi$ es una función convexa Estas divergencias o casos particulares fueron muy estudiadas las decadas que siguieron, dando también lugar a varias aplicaciones (Gupta & Sharma, 1976; Burbea & Rao, 1982; Cressie & Read, 1984; Ben-Tal, Charnes & Teboulle, 1989; Teboulle, 1992; Ben-Tal, Bornwein & Teboulle, 1992; Salicrú et al., 1993; Salicrú, 1994; Csiszár, 1995; Cressie & Pardo, 2000; Liese & Vajda, 2006).

Como para el caso de las $\phi$-entropías, esta clase se enmarca dentro de una clase un poco más general (Ali & Silvey, 1966, Secs. 4.5 & 5) (ver también (Orsak & Paris, 1995, Sec. I)):

**Definición 2-64** ($(h,\phi)$-divergencia)**.** *La $(h,\phi)$-divergencia de una distribución de probabilidad $Q$ con respeto a una distribución de referencia $P$ tal que $Q \ll P$*

$$D^{\mathrm{c}}_{(h,\phi)}(Q\|P) = h\left(\int_{\mathcal{X}} \phi\left(\frac{dQ}{dP}(x)\right) dP(x)\right)$$

*Si $P$ y $Q$ admiten una densidad con respeto a una medida $\mu$,*

$$D^{\mathrm{c}}_{(h,\phi)}(q\|p) = h\left(\int_{\mathcal{X}} p(x)\phi\left(\frac{q(x)}{p(x)}\right) d\mu(x)\right)$$

*en su versión diferencial, donde o*

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

*y* $\mathcal{X} = X(\Omega)$ *para* $X$ *de distribución* [69] $P$*. Frecuentemente, se supone adicionalmente que* $\phi$ *y* $h$ *son de clase* $C^2$ *y sin perdida de generalidad, que* $h(\phi(1)) = 0$*.*

Se notará de que, obviamente, $D^{\mathrm{c}}_{\phi} = D^{\mathrm{c}}_{(\mathrm{id},\phi)}$.

Notablemente, cuando $\phi(u) = u \log u$ y $h = \mathrm{id}$ se recupera de nuevo la divergencia de Kullback-Leibler: esta última partenece simultaneamente a la clase de Csiszàr y a la de Bregman y es la sola en este caso (Csiszàr, 1991). Cuando $\phi(u) = \pi_2 u \log u - (\pi_1 + \pi_2 u) \log(\pi_1 + \pi_2 u)$ y $h = \mathrm{id}$ se recupera la divergencia de Jensen-Shannon **sola de la clase de Jensen en este caso?**. Además de $D_{\mathrm{kl}}$, la clase de Csiszár contiene la ganancia de información de Rényi para $\phi(u) = u^{\lambda}$ y $h(u) = \frac{\log u}{\lambda - 1}$ apareciendo tambien en una forma muy parecida en (Hellinger, 1909; Chernoff, 1952; Cressie & Read, 1984; Liese & Vajda, 2006) y conocida como divergencia de Chernoff o de Hellinger. Contiene varias otra como la $J$-divergencia para $\phi(u) = u \log u - \log u$ y $h = \mathrm{id}$, la distancia de Bhattacharyya (Bhattacharyya, 1943, 1946) $-\log \int_{\mathcal{X}} \sqrt{\frac{dQ}{dP}(x)}\, dP(x)$ para $\phi(u) = \sqrt{u}$ y $h(u) = -\log u$, instancia particular de la de Rényi ($\lambda = \frac{1}{2}$), la divergencia de variación total (o $L^1$ distancia) para $\phi(u) = |u - 1|$ y $h = \mathrm{id}$, la divergencia de Pearson o divergencia $\chi^2$ para $\phi(u) = (u-1)^2$ o $u^2 - 1$ y $h = \mathrm{id}$, para mencionar unas.

Las divergencias de Csiszár tienen las propiedades siguientes

1. Positividad:

   $$D^{\mathrm{c}}_{(h,\phi)}(Q \| P) \geq 0 \quad \text{con igualdad si y solamente si} \quad P = Q.$$

   Esta propiedad es la consecuencia directa de la convexidad estricta de $\phi$ conjuntamente a $h(\phi(1)) = 0$. De hecho, de la desigualdad de Jensen con $X$ de distribución $P$ tenemos en el caso $\phi$ convexa y $h$ creciente $D^{\mathrm{c}}_{(h,\phi)}(Q \| P) = h\left(\mathrm{E}\left[\phi\left(\frac{dQ}{dP}(X)\right)\right]\right) \geq h\left(\phi\left(\mathrm{E}\left[\frac{dQ}{dP}(X)\right]\right)\right) = h(\phi(1))$ (y similarmente en el caso $\phi$ cóncava y $h$ decreciente). Fijense de que la positividad no es en contradicción con el enfoque de Rényi en su caso, porque conideró el caso discreto finito con probabilidades incompletas, *i. e.,* su axioma (ii) se cumpla potencialemente solamente para los vectores de probabilidades incompletos.

2. $D^{\mathrm{c}}_{(h,\phi)}$ satisface un teorema de procesamiento de datos (o secunda ley de la termodinámica) en el sentido de que si dos distribuciones son consecuencias de la misma probabilidad de transición (condicional) $p_{X_{n+1}|X_n = x_n} = q_{X_{n+1}|X_n = x_n}$ (densidades con respeto a una medida $\mu$), entonces

   $$D^{\mathrm{c}}_{(h,\phi)}\left(p_{X_{n+1}} \middle\| q_{X_{n+1}}\right) \leq D^{\mathrm{c}}_{(h,\phi)}\left(p_{X_n} \| q_{X_n}\right).$$

   **Probar**

3. $D^{\mathrm{c}}_{(h,\phi)}$ es convexa con respecto al par $(P, Q)$, pero no necesariamente con respecto a $p$ solamente y/o $q$. En el caso $\phi$ convexa, es consecuencia directa de la convexidad [70] (resp. cóncavidad) de $(u, v) \mapsto u\, \phi\left(\frac{v}{u}\right)$ sobre $\mathbb{R}^2_+$ conjuntamente a la crecencia (resp. decrecencia) de $h$.

---

[69] En general, por convención, $0\, \phi\left(\frac{0}{0}\right) = 0$. Además, se requiere de que $0\, \phi\left(\frac{a}{0}\right) = \lim_{\varepsilon \to 0^+} \varepsilon\, \phi\left(\frac{a}{\varepsilon}\right) = a \lim_{u \to +\infty} \frac{\phi(u)}{u}$.

[70] Con la hipotesis de que $\phi$ sea de clase $C^2$, es sencillo ver de que la Hessiana de la función $(u, v) \mapsto u\, \phi\left(\frac{v}{u}\right)$ con respeto a $(u, v)$ es no negativa, implicando la convexidad de esta función bi-variada (Cambini & Martein, 2009).

4. Pensando a $D^{\mathrm{c}}_{\phi}$ con respecto a $\phi$, es lineal en el sentido de que $D^{\mathrm{c}}_{a_1\phi_1+a_2\phi_2} = a_1 D^{\mathrm{c}}_{\phi_1} + a_2 D^{\mathrm{c}}_{\phi_2}$ (con $\phi_i$ convexas y $a_i \geq 0$).

5. Sea $\phi^*(u) = u\phi\left(\frac{1}{u}\right)$. Es sencillo ver de que si $\phi$ es convexa (resp. cóncava), $\phi^*$ es también convexa (resp. cóncava). $\phi^*$ es llamada $^*$*-conjugada convexa (resp. cóncava)* de $\phi$. Luego,

$$D^{\mathrm{c}}_{(h,\phi)}\left(Q\|\,P\right) = D^{\mathrm{c}}_{(h,\phi^*)}\left(P\|\,Q\right).$$

6. $D^{\mathrm{c}}_{(h,\phi)}$ es simétrica si y solamente si $\phi = \phi^* + c(\mathrm{id} - 1)$; sin perdida de generalidad, consideramos $c = 0$. Sin embargo, en el caso general, se puede definir una versión simetrizada al imagen de la $J$-divergencia, considerando $D^{\mathrm{c}}_{(h,\phi)} + D^{\mathrm{c}}_{(h,\phi^*)}$. En particular, cuando $h = \mathrm{id}$, tenemos

$$D^{\mathrm{c}}_{\phi} + D^{\mathrm{c}}_{\phi^*} = D^{\mathrm{c}}_{\phi+\phi^*}$$

   que es simétrica ($(\phi^*)^* = \phi$).

7. Cota superior:

$$D^{\mathrm{c}}_{(h,\phi)} \leq h\left(\phi(0) + \phi^*(0)\right)$$

   posiblemente infinita [71].

Estas propiedades con varias otras se encuentran por ejemplo en (Vajda, 1972; Csiszár, 1974; ?, ?; Kafka et al., 1991; Österreicher, 1996; Österreicher & Vajda, 2003; Vajda, 2009; Kumar & Chhina, 2005; Liese & Vajda, 2006). Como en el caso de las divegencias de Jensen, en general las divergencias simétricas ($\phi = \phi^*$) no satisfacen en general a la desigualdad triangular, y entonces no dan lugar a una distancia entre distribuciones de probabilidad, aparte en casos particulares (ej. divergencia de la variación total, divergencia de Hellinger o Rényi con $\lambda = \frac{1}{2}$). Sin embargo, se probó en (Kafka et al., 1991, Teoremas 1 & 2, Remark 6) y (Österreicher, 1996; Österreicher & Vajda, 2003; Vajda, 2009) el lema siguiente, condición suficiente para que una potencia de la divergencia satisfaga a la desigualdad triangular:

**Lema 2-12.** *Sea $\phi$ una función convexa tal que $\phi^* = \phi$, $\phi(0) \neq 0$ y $D^{\mathrm{c}}_{\phi}$ su divergencia de Csiszár asociada ($h = \mathrm{id}$). Adicionalmente de supone de que $\phi(1) = 0$ (ver definición Def. 2-64) y de que $\phi$ es estrictamente convexa en $1$. Si existe $\kappa \in \mathbb{R}^*_+$ tal que*

$$h(u) = \frac{\left(1 - u^{\kappa}\right)^{\frac{1}{\kappa}}}{\phi(u)}, \quad u \in [0\,,\,1) \quad \textit{es no decreciente,}$$

*entonces*

$$\left(D^{\mathrm{c}}_{\phi}\left(\cdot\|\,\cdot\right)\right)^{s} \quad s \in (0\,,\,\kappa] \quad \textit{satisface la desigualdad triangular}$$

*y entonces es una métrica entre dos discribuciones de probabilidades.*

---

[71] Por ejemplo, para la divergencia de Kullback-Leibler, $\phi(u) = u\log u$, dando $\phi^*(u) = -\frac{\log u}{u}$, tales que $\phi(0) = 0$ y $\phi^*(0) = +\infty$: no es acotada por arriba.

Además, en (Kafka et al., 1991, Sec. 3) se dan condiciones necesarias que debe cumplir $\kappa$ cuando $\phi$ tiene un comportamiento particular en $u \to 0$ y/o $u \to 0$; eso va más allá de la meta de esta sección y el lector se podrá referir a (Kafka et al., 1991; Österreicher, 1996; Österreicher & Vajda, 2003) para tener más detalles.

Este lema se usó para probar el caracter métrico de $\left(J_{\mathrm{js}}^{\pi_{\mathrm{u}}}(p_{(1)}, p_{(2)})\right)^s, \quad 0 < s \leq \frac{1}{2}$ (Österreicher & Vajda, 2003; Osán et al., 2018, y ref.), siendo $J_{\mathrm{js}}$ una divergencia de Csiszár particular. Se usó también para probar de que $\left(D_\phi^{\mathrm{c}}\right)^s$ con $\phi(u) = \frac{\lambda}{\lambda-1}\left(\left(1+u^\lambda\right)^{\frac{1}{\lambda}} - 2^{\frac{1}{\lambda}-1}(1+u)\right)$ y $\kappa = \min\left(\lambda, \frac{1}{2}\right)$ es una métrica (Österreicher & Vajda, 2003).

Para cerrar esta sección, se mencionará de que varias aplicaciones se encuentran en la literatura que sea en estimación, discriminación, reconocimiento de patrones, pruebas de adecuación o inferencia estadística, entre otros (Kailath, 1967; Boekee & van der Lubbe, 1979; Poor, 1988; Basseville, 1989; Csiszár, 1995; Orsak & Paris, 1995; Menéndez, Morales, Pardo & Vajda, 1977; Pardo, 1999; Liese & Vajda, 2006; Pardo, 2006; Nielsen & Boltz, 2011; Csiszàr & Matúš, 2012; Basseville, 2013, y ref.) en adición de las referencias de esta sección, para dar unas.

**cf Bha 43 egalement**

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

**Principio de entropía máxima** Si este principio nació en el marco de la termodynamica o física, con la entropía de Shannon (Boltzman), tratando de las nociones generalizadas de incertas, vuelve natural preguntarse sobre la extensión de este problema en el marco general. **Tal estudio fue hecho en varios trabajos (?, ?; Ben-Tal et al., 1992) 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 $\mathrm{E}\left[M_k(X)\right] = m_k$ fijos, con $M_x : \mathcal{X} \to \mathbb{R}$, el problema de $(h,\phi)$-entropía máxima se formula de la manera siguiente: sean $M(x) = \begin{bmatrix} 1 & M_1(x) & \cdots & M_K(x) \end{bmatrix}^t$ y $m = \begin{bmatrix} 1 & m_1 & \cdots & m_K \end{bmatrix}^t$, se busca,

$$p^* = \operatorname*{argmáx}_p H_{(h,\phi)}(p) \qquad \text{sujeto a} \qquad p \geq 0, \quad \int_{\mathcal{X}} M(x)\, p(x)\, d\mu(x) = m.$$

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 perdida de generalidad, se puede considerar la situación $\phi$ convexa. Como en el caso de Shannon, introduciendo factores de Lagrange $\eta = \begin{bmatrix} \eta_0 & \eta_1 & \cdots & \eta_K \end{bmatrix}^t$ 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)

$$p^* = \operatorname*{argmáx}_p \int_{\mathcal{X}} \left(-\phi\left(p(x)\right) + \eta^t M(x)\, p(x)\right) d\mu(x)$$

donde $\eta$ será determinado para satisfacer los vínculos. De nuevo, de la ecuación de Euler-Lagrange (Gelfand & Fomin, 1963; van Brunt, 2004) en el caso continuo, o derivando con respeto a las probabilidades en el caso discreto, se obtiene la ecuación $-\phi'(p(x)) + \eta^t M(x) = 0$. La función entrópica $\phi$ es cóncava y de clase $C^2$, así que $\phi'$ es continua decreciente, y de la monotonicidad es invertible. Entonces,

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

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

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

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

**Lema 2-13.** *Sea* $\mathcal{P}_m = \left\{p \geq 0 : \int_{\mathcal{X}} M_k(x)\, p(x)\, d\mu(x) = m\right\}$ *y* $p^* \in \mathcal{P}_m$ *que satisfaga* $\phi'(p^*(x)) = \eta^t M(x)$. *Entonces*

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

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

$$\begin{aligned} D_\phi^{\mathrm{b}}(p \| p^*) &= H_\phi(p^*) - H_\phi(p) - \int_{\mathcal{X}} \left(p(x) - p^*(x)\right) \phi'(p^*(x))\, d\mu(x) \\ &= H_\phi(p^*) - H_\phi(p) - \eta^t \int_{\mathcal{X}} \left(p(x) - p^*(x)\right) M(x)\, d\mu(x) \\ &= H_\phi(p^*) - H_\phi(p) \end{aligned}$$

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

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 $\phi(u) = u^\lambda$ (Rényi o Havrda-Charvát-Daróczy), la distribución de entropía máxima es Student; Costa and son on. Gausiana se recupera caso límite.**

Como en el caso de Shannon, si $\mu = Q$ es una medida de referencia, el problema vuelev ser un problema de minimización de la $(h,\phi)$-divergencia de $P \ll Q$ con respeto a $Q$. La densidad obtenida es $p^* = \frac{dP^*}{dQ}$.

**reapparition Fisher comme courbure, cf Varma, Jizba, MenMor97...**

**EPI variaciones Madiman Barron MadBar07**

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

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

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

## 2.7 Entropias cuanticas discretas

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

# CAPÍTULO 3
# Elementos de geometría diferencial

*Pedro Walter Lamberti*

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

## 3.1 Estructuras

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

- Espacio topológico (continuidad)
- Espacio métrico (distancia)
- Variedad topológica (coordenadas)
- Variedad diferenciable (diferenciabilidad)
- Estructura afin (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 espacio vectorial 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.

## 3.2 Espacio Topológico

Un conjunto arbitrario $X$ está desprovisto de toda estructura que permita definir nociones tales como la *convergencia* de una sucesión de elementos de $X$, la *proximidad* de dos elementos de $X$, etc. En principio se dispone sólo de las operaciones elementales de *unión* $\bigcup$ e *intersección* $\bigcap$ de subconjuntos. Estas operaciones también pueden realizarse entre distintos conjuntos. Denotaremos con $\emptyset$ al conjunto vacío. Surge entonces el desafío 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 3-65** (Topología)**.** *Una topología $\Upsilon$ sobre el conjunto $X$ es una familia de subconjuntos de $X$ que cumple con las siguientes condiciones:*

1. *$X$ y $\emptyset$ están en $\Upsilon$: $X, \emptyset \in \Upsilon$*

2. *La intersección de cualquier colección finita de elementos de $\Upsilon$ está en $\Upsilon$:*

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

3. *La unión de una colección arbitraria –finita o no– de elementos de $\Upsilon$, pertenece a $\Upsilon$:*

$$A_i \in \Upsilon \quad \Rightarrow \quad \bigcup_{i} A_i \in \Upsilon$$

**Definición 3-66** (Espacio topológico y abiertos)**.** *Al par $(X, \Upsilon)$ lo llamaremos espacio topológico. Los conjuntos que están en $\Upsilon$ se llaman abiertos.*

Ejemplos:

- *Topología trivial*. Es la que consta de sólo dos elementos, el conjunto vacío y el conjunto total $X$ : $\Upsilon = \{\emptyset, X\}$.

- *Topología discreta*. Es la que en todo subconjunto de $X$ está en $\Upsilon$, es decir $\Upsilon = \mathcal{P}(X)$ 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 $p = (p_1, ..., p_n) \in \mathbb{R}^n$ y de radio $r > 0$ es el conjunto

$$\mathcal{B}_{r,p} = \left\{ (x_1, \ldots, x_n) \in \mathbb{R}^n : 0 \leq \sqrt{\sum_i (x_i - p_i)^2} < r \right\}$$

  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 $\mathbb{R}^n$.

  Obsérvese que un subconjunto $A$ de $\mathbb{R}^n$ es abierto (en el sentido usual), cuando para todo $x \in A$, existe un $\varepsilon > 0$ tal que $\mathcal{B}_{\varepsilon,x} \subset A$.

**Definición 3-67** (Entorno)**.** *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 $V \in \Upsilon$.*

**Definición 3-68** (Función continua)**.** *Sea $f : X \to Y$ una función entre dos espacios topológicos $(X, \Upsilon)$ 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 $f(V) \subset U$. Equivalentemente se puede definir una función continua de la siguiente manera: $f$ es una función continua sii la imagen inversa de cada conjunto abierto es un abierto.*

Es fácil demostrar la equivalencia entre ambas definiciones, y hacerlo queda como ejercicio para el lector.

**Definición 3-69** (Homomorfismo)**.** *Un homomorfismo $\Psi$ entre dos espacios topológicos $(X, \Upsilon)$ e $(Y, \omega)$ es una función $\Psi : X \to V \subseteq Y$ biyectiva, continua y con inversa continua.*

**Definición 3-70** (Sucesión)**.** *Una sucesión en un conjunto $X$ es una aplicación $s : \mathbb{N} \to X$ donde $\mathbb{N}$ es el conjunto de los números naturales. Denotaremos a la sucesión por $\{x_n\}_{n \in \mathbb{N}}$.*

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 3-71** (Límite)**.** *Sea $(X, \Upsilon)$ 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 $(X, \Upsilon)$ 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 cumplen con esta condición se los llama espacios de Hausdorff o espacios $T_2$.

## 3.3 Espacios métricos

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

**Definición 3-72** (Espacio métrico)**.** *Un espacio métrico en un conjunto $X$ munido de una función $d : X \times X \to \mathbb{R}_+$ tal que se cumplen las condiciones:*

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

La última condición se conoce como *desigualdad triangular*. Mas adelante en este libro veremos funciones $d : X \times X \to \mathbb{R}_+$ 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* definida sobre $X$.

## 3.4 Variedad Topológica

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

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

**Definición 3-73** (Variedad topológica $n$-dimensional)**.** *Una Variedad topológica $n$-dimensional es un espacio topológico $\mathcal{M}$ tal que es localmente euclídeo, es decir que para cada $x \in \mathcal{M}$ existe un entorno abierto $U$ de $x$, homeomorfo a un abierto $V$ de $\mathbb{R}^n$: $\phi : U \subseteq \mathcal{M} \to \mathbb{R}^n$ tal que $\phi : U \to V$ y $\phi$ es un homeomorfismo. También pediremos que $\mathcal{M}$, como espacio topológico, sea un espacio Hausdorff.*

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

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

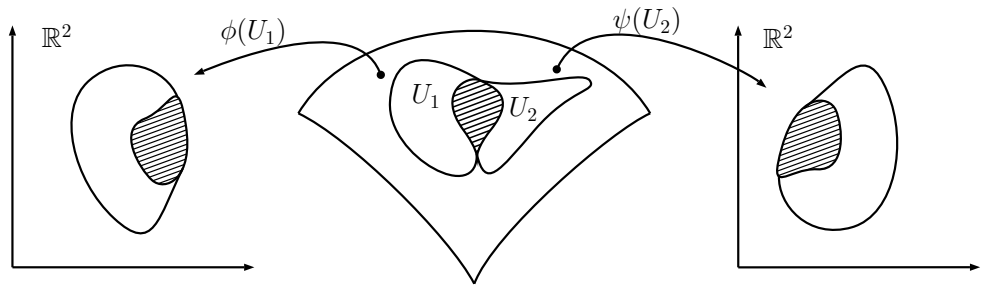



**Figura 3-41:** Cartas coordenadas usadas en la definición de una variedad topológica.

A la colección de números reales $(p_1, \ldots, p_n)$ se llaman las coordenadas de $p$ de acuerdo a la carta $(U, \phi)$. La existencia de coordenadas, es el aspecto fundamental por el que el concepto de variedad es tan útil en física.

Podría suceder que un mismo punto $p$ pertenezca a más de una carta, digamos $(U_1, \phi_1)$ y $(U_2, \psi_2)$. En ese caso hablaremos de un cambio de coordenadas:

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

Si denotamos por $(p_1, \ldots, p_n)$ a las coordenadas correspondientes a la carta $(U_1, \phi_1)$ y por $(\tilde{p}_1, \ldots, \tilde{p}_n)$ a las correspondientes a la carta $(U_2, \psi_2)$, entonces las funciones $\tilde{p}_i = \tilde{p}_i(p_1, \ldots, p_n)$ son funciones continuas, y dan el cambio de coordenadas. Estas funciones son invertibles con inversa continua.

Ejemplos de variedades topológicas son:

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

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

Se debe observar que al definir $\mathbb{S}^n$ no estamos pensando que está inmersa en $\mathbb{R}^n$. En este caso podemos usar las siguientes cartas: $(U_N, \phi_N)$ y $(U_S, \phi_S)$, donde $U_N = \mathbb{S}^n - \{(0, 0, \ldots, 1)\}, \quad U_S = \mathbb{S}^n - (1, 0, \ldots, 0)$ y los mapas

$$\phi_N : U_N \to \mathbb{R}^n / (\phi_N(x_1, \ldots, n_{n+1}))_i = \frac{x_i}{1 - x_{n+1}}$$

y

$$\phi_S : U_S \to \mathbb{R}^n / (\phi_N(x_1, \ldots, n_{n+1}))_i = \frac{x_i}{1 + x_{n+1}}$$

Ambos mapas son homeomorfismos. Observemos que $\phi_N(x_1, \ldots, x_{n+1}) = (tx_1, \ldots, tx_n)$ y $\phi_S(x_1, \ldots, x_{n+1}) = (ux_1, \ldots, ux_n)$ con $t = \frac{1}{1 - x_{n+1}}$ y $u = \frac{1}{1 + x_{n+1}}$, respectivamente. Es directo verificar la inyectividad pues si $(tx_1, \ldots, tx_n) = (ty_1, \ldots, ty_n) \Rightarrow x_i = y_i \quad \forall i$. Entonces los puntos $x$ e $y$ son idénticos. Para ver la suryectividad consideremos el punto $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$. Si tomamos $x = \left(t^{-1}y_1, \ldots, t^{-1}y_n, y_{n+1}\right)$ con $t \neq 0$ e $y_{n+1} = t\sqrt{1 - (t^{-1}y_1)^2 - \cdots - (t^{-1}y_n)^2}$ vemos que para cada $y \in \mathbb{R}^n$ existe un $x \in \mathbb{S}^n$ tal que $\phi(x) = y$. Usando las expresiones explćitas de $\phi_N$ y $\phi_S$ es directo verificar que se trata de funciones continuas.

**Nota**: Hay propiedades de las variedades topológicas que no tienen que ver con sus características locales, las que hemos dicho son similares a las de $\mathbb{R}^n$, sino con sus propiedades globales. Por ejemplo una esfera $2$-dimensional es homeomorfa a la superficie de una pelota de futbol, aún cuando pensemos en una pelota de futbol verdadera, la cual es una colección de parches hexagonales o pentagonales, unidos unos con otros. Ambos objetos, la esfera y la pelota de futbol, son objetos compactos, cerrados y simplemente conexos. Sin embargo un toro y una esfera no comparten todas estas características: un toro es cerrado, compacto pero no simplemente conexo, es decir no todo lazo sobre él puede contraerse continuamente a un punto. Por ello diremos que un toro y una esfera son localmente homeomorfos, pero no lo son globalmente. Este tipo de situaciones ha llevado a introducir cantidades que de alguna manera caractericen a las propiedades globales de una variedad topológicas. Un ejemplo muy conocido es la característica de Euler. Para un poliedro de tres dimensiones la caracteristica de Euler $\Xi$ está definida por

$$\Xi = V - A + C$$

donde $V, A$ y $C$ son el número de vertices, de aristas y de caras del poliedro, respectivamente. Para un cubo, por ejemplo, $\Xi = 2$. Supongamos que el cubo está hecho en un material elástico, apoyado sobre un armazón (las aristas) de metal. Si inflamos ese cubo, obtenemos una esfera. Matemáticamente eso significa que el cubo y la efera son globalmente homeomorfos entre si, y por lo tanto topológicamente equivalentes. Es posible extender el concepto de característica de Euler a la superficie de una esfera, a través de la triangularización de la superficie esférica, es decir cubriendo la esfera por triángulos. En ese caso la característica de Euler se calcula como el número de triángulos menos el número de aristas más el número de vértices. Haciéndo ese cálculo para la esfera resulta el valor $2$. Lo mismo sucede con cualquier otro poliedro que se pueda deformarse continuamente a una esfera. Hay maneras de definir la característica de Euler para una variedad topológica arbitraria y esa cantidad es un invariante topológico, es decir una cantidad que no cambia entre variedades homeomórficos. Para un toro la característica de Euler vale $0$.

## 3.5 Variedad Diferenciable

Sobre una variedad topológica se puede "montar" una nueva estructura. Es posible hacer eso imponiendo condiciones de diferenciabilidad a los mapas coordenados de la definición de una variedad topológica. Sin embargo, no tenemos definida la noción de diferenciablidad sobre una variedad cualquiera. Por ello, para definir una estructura diferenciable sobre una variedad topológica arbitraria, recurrimos a $\mathbb{R}^n$ donde si está definida la noción de diferenciabilidad. Por ello hacemos la siguiente:

**Definición 3-74** ($C^r$-compatibilidad)**.** *Diremos que dos cartas coordenadas $(U, \phi)$ y $(V, \psi)$ sobre una variedad $\mathcal{M}$ son $C^r$-compatibles si cuando $U \bigcap V \neq \emptyset$ entonces $\phi \circ \psi^{-1}$ y $\psi \circ \phi^{-1}$ son de clase $C^r$ sobre los subconjuntos $\phi(U \bigcap V)$ y $\psi(U \bigcap V)$ de $\mathbb{R}^n$, respectivamente.*

Con esto podemos avanzar en la siguiente:

**Definición 3-75** (Variedad diferenciable)**.** *Una Variedad diferenciable $n$-dimensional de clase $C^r$, $\mathcal{M}$, es una variedad topológica y una familia de cartas coordenadas $\mathcal{B} = (U_\alpha, \phi_\alpha)$, tales que:*

1. *los $U_\alpha$ cubren $\mathcal{M}$,*
2. *para cualquier par $\alpha, \beta$, los entornos $(U_\alpha, \phi_\alpha)$ y $(U_\beta, \phi_\beta)$ son $C^r$ - compatibles,*
3. *Cualquier entorno coordenado $(V, \psi)$ $C^r$-compatible con cualquiera de los $(U_\alpha, \phi_\alpha) \in \mathcal{B}$ está en $\mathcal{B}$.*

Cualquier superficie "suave" en $\mathbb{R}^3$ es un ejemplo de (sub) variedad diferenciable. Este ejemplo no debe conducir a la confusión de pensar que una variedad debe estar inmersa en $\mathbb{R}^n$. Otro ejemplo de variedad diferenciable de dimensión $n$ es la esfera $\mathbb{S}^n$, definida previamente.

**Definición 3-76** (Diferenciabilidad de clase $C^k$)**.** *Dadas dos variedades $\mathcal{M}$ y $\mathcal{M}'$ de clase $C^r$, una aplicación $f : \mathcal{M} \to \mathcal{M}'$, se dice diferenciable de clase $C^k$, $\quad k \leq r$ si para toda carta $(U_\alpha, \phi_\alpha)$ de $\mathcal{M}$ y toda carta de*

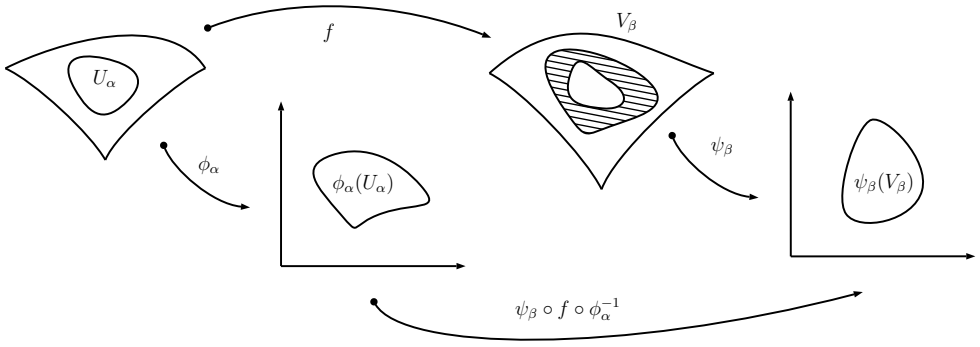



**Figura 3-42:** Cartas coordenadas usadas en la definición de una función diferenciable.

$(V_\beta, \psi_\beta)$ *de* $\mathcal{M}'$ *tal que* $f\,(U_\alpha) \subset V_\beta$*, la aplicación* $\psi_\beta \circ f \circ \phi_\alpha^{-1}$ *de* $\phi_\alpha\,(U_\alpha)$ *en* $\psi_\beta\,(V_\beta)$*, es diferenciable de clase* $C^k$*.*

El disponer de la noción de función diferenciable, permite asignar a cada punto de una variedad diferenciable, un espacio vectorial. Éste estará dado por operadores lineales que actúan sobre funciones diferenciables y dan por resultado un número. Antes de ir a la definición de ese espacio vectorial, introducimos el concepto de curva suave sobre una variedad.

**Definición 3-77** (Curva de clase $C^k$ sobre una variedad)**.** *Sea* $\mathcal{M}$ *una variedad de clase* $C^r$*. Una curva* $\lambda$ *en* $\mathcal{M}$ *de clase* $C^k$*,* $\quad k \leq r$ *es una función del intervalo real* $[a\,,\,b]$ *en* $\mathcal{M}$ *tal que para toda carta* $(U_\alpha, \phi_\alpha)$ *en* $\mathcal{M}$ *la composición*

$$\phi_\alpha \circ \gamma : [a\,,\,b] \to \phi_\alpha\,(U_\alpha)$$

*es de clase* $C^k$*. En coordenadas*

$$\phi_\alpha \circ \gamma(t) = \{x^1(t), \ldots, x^n(t)\}$$

Con esto podemos ahora dar la noción de vector tangente a una variedad:

**Definición 3-78** (Tangente a una variedad)**.** *Sea* $\mathcal{F}(p)$ *el conjunto de funciones diferenciables de clase* $C^1$ *definidas en un entorno del punto* $p$*. Sea* $\gamma(t)$ *una curva de clase* $C^1$*,* $a \leq t \leq b$ *tal que* $\gamma(t_0) = p$*. El vector tangente a la curva* $\gamma(t)$ *en el punto* $p$ *es una aplicación* $\mathbb{X}_p : \mathcal{F}(p) \to \mathbb{R}$ *cuyo efecto es*

$$\mathbb{X}_p f = \frac{df(\gamma(t))}{dt}|_{t_0}$$

El vector $\mathbb{X}_p$ satisface las siguientes propiedades

- $\mathbb{X}_p$ es una aplicación lineal de $\mathcal{F}(p)$ en $\mathbb{R}$,
- $\mathbb{X}_p(fg) = (\mathbb{X}_p f)\, g(p) + f(p)\, (\mathbb{X}_p g)$ para $f, g \in \mathcal{F}(p)$.

Dejamos para el lector demostrar estas propiedades.

Sean $(u^1, \ldots, u^n)$ coordenadas locales en un entorno $U$ de $p$. Para cada $j$, $\left(\frac{\partial}{\partial u^j}\right)|_p$ es una aplicación de $\mathcal{F}(p)$ en $\mathbb{R}$ la cual satisface las propiedades (i) e (ii). Veremos a continuación que el conjunto de todas las

aplicaciones $\mathbb{X}$ de $\mathcal{F}(p)$ en $\mathbb{R}$ es un espacio vectorial $n$-dimensional, siendo $n$ la dimensión de la variedad diferenciable $\mathcal{M}$.

Dada una curva $\gamma(t)$ con $\gamma(t_0) = p$, sean $u^j(t) = \gamma^j(t), \quad j = 1, \ldots, n$ las coordenadas locales de esa curva. Entonces $\frac{df(\gamma(t))}{dt}|_{t_0} = \sum_j \left(\frac{\partial f}{\partial u^j}|_p\right)\left(\frac{d\gamma^j(t)}{dt}\right)|_{t_0}$. Extra expresión indica que todo vector en $p$ es una combinación lineal de los vectores (operadores).

$$\left(\frac{\partial}{\partial u^1}|_p\right), \ldots, \left(\frac{\partial}{\partial u^n}|_p\right) \tag{3}$$

Sea la combinación lineal $\sum_j \xi^j \frac{\partial}{\partial u^j}|_p$ y sea la curva definida por

$$u^j(t) = u^j(p) + \xi^j t \quad j = 1, \ldots, n$$

El vector tangente a esta curva en $t = 0$ es $\sum \xi^j \frac{\partial}{\partial u^j}|_p$. Además si

$$\sum \xi^j \frac{\partial}{\partial u^j}|_p = 0,$$

entonces

$$0 = \sum \xi^j \left(\frac{\partial u^k}{\partial u^j}\right)|_p = \xi^k \quad k = 1, \ldots, n$$

Esto demuestra la independencia lineal de los vectores (3).

**Definición 3-79** (Espacio tangente)**.** *El conjunto de vectores tangentes en $p \in \mathcal{M}$, es llamado el espacio tangente de $\mathcal{M}$ en $p$, y lo denotaremos por $T_p(\mathcal{M})$.*

La colección de todos los espacios tangentes, $\bigcup_{p\in\mathcal{M}} T_p(\mathcal{M})$ se llama *fibrado tangente*.

Al fibrado tangente se le puede dar la estructura de un álgebra (álgebra de Lie). Esta surge de calcular el conmutador $[\mathbb{X}, \mathbb{Y}]$ entre dos campos vectoriales $\mathbb{X}$ e $\mathbb{Y}$:

$$[\mathbb{X}, \mathbb{Y}]f \equiv (\mathbb{X}\mathbb{Y} - \mathbb{Y}\mathbb{X})\, f$$

Si los vectores se escriben en término de los vectores de la base coordenada $\left(\frac{\partial}{\partial x^a}\right)$, el conmutador entre ellos resulta ser el vector:

$$\sum_{ab} X^a \frac{\partial Y^b}{\partial x^a}\frac{\partial}{\partial x^b} - \sum_{ab} Y^a \frac{\partial X^b}{\partial x^a}\frac{\partial}{\partial x^b}$$

A cada espacio tangente $T_p(\mathcal{M})$ podemos asignar su dual, $T_p^*(\mathcal{M})$, es decir el conjunto de todos los operadores lineales y homogéneos que actúan sobre $T_p(\mathcal{M})$. A un elemento del espacio dual lo llamaremos $1$-*forma*. Denotaremos a la acción de un elemento de $T_p^*(\mathcal{M})$, digamos $\omega_p$, por:

$$\omega_p(\mathbb{X}_p) = \langle \omega_p, \mathbb{X}_p \rangle .$$

Para cada función $f \in \mathcal{F}(p)$, el *diferencial de* $f$, denotado por $(df)_p$, es el elemento de $T_p^*(\mathcal{M})$ que tiene por acción:

$$\langle (df)_p, \mathbb{X}_p \rangle = \mathbb{X}_p f, \quad \mathbb{X}_p \in T_p(\mathcal{M})$$

Cada función coordenada $u^j$ es una función de $\mathcal{M}$ sobre $\mathbb{R}$. Entonces podemos calcular el diferencial de $u^j$, cuya acción sobre un vector $\mathbb{X}_p \in T_p(\mathcal{M})$ está dada por

$$\left\langle (du^j)_p, \mathbb{X}_p \right\rangle = \mathbb{X}_p^j$$

En particular, si $\mathbb{X}_p = \left(\frac{\partial}{\partial u^k}\right)$ resulta

$$\left\langle (du^j)_p, \left(\frac{\partial}{\partial u^k}\right) \right\rangle = \delta_k^j;$$

es decir $\left\{(du^j)_p\right\}_{j=1}^n$ es la base dual de $\left\{\left(\frac{\partial}{\partial u^j}\right)_p\right\}_{j=1}^n$. Toda $1$-forma $\omega$ se puede escribir en término de esta base:

$$\omega = \sum_a \omega_a dx^a$$

Con los espacios $T_p(\mathcal{M})$ y $T_p^*(\mathcal{M})$ podemos construir el espacio producto cartesiano

$$(T_p(\mathcal{M}))_s^r = T_p(\mathcal{M}) \times T_p(\mathcal{M}) \ldots T_p(\mathcal{M}) \times T_p^*(\mathcal{M}) \times T_p^*(\mathcal{M}) \ldots \times T_p^*(\mathcal{M})$$

con $r$ factores de $T_p(\mathcal{M})$ y $s$ factores de $T_p^*(\mathcal{M})$.

**Definición 3-80** (Tensor de tipo $(r,s)$)**.** *Un tensor de tipo $(r,s)$ es un operador $S$,*

$$S : (T_p(\mathcal{M}))_s^r \to \mathbb{R}$$

*que es lineal y homogéneo en cada uno de sus argumentos.*

**Definición 3-81** (Campo tensorial)**.** *Un campo tensorial $S$ de clase $C^k$ de tipo $(r,s)$ sobre $V \subseteq \mathcal{M}$ es un mapa $C^k$ que asigna un tensor de tipo $(r,s)$ a cada punto $p \in V$.*

En término de las bases $\left\{\left(\frac{\partial}{\partial u^j}|_p\right)\right\}_{j=1}^n$ y $\left\{(du^j)_p\right\}_{j=1}^n$, el campo tensorial $S$ se puede escribir:

$$S(p) = S^{a_1 \ldots a_r}_{b_1 \ldots b_s}(p) \frac{\partial}{\partial x^{a_1}} \bigotimes \ldots \bigotimes \frac{\partial}{\partial x^{a_r}} \bigotimes dx^{b_1} \bigotimes \ldots \bigotimes dx^{b_s}$$

donde las funciones $S^{a_1 \ldots a_r}_{b_1 \ldots b_s}$ son de clase $C^k$ y $\bigotimes$ es el producto tensorial.

Entre los campos tensoriales que se pueden definir sobre una variedad $\mathcal{M}$, hay uno particularmente importante, y es conocido como el tensor métrico. Éste se define por medio de un producto escalar:

**Definición 3-82** (Producto escalar)**.** *Un* **producto escalar** *sobre $T_p(\mathcal{M})$ es una función*

$$g : T_p(\mathcal{M}) \times T_p(\mathcal{M}) \to \mathbb{R}$$

*que satisface*

1. $g(\mathbb{X}, \mathbb{Y}) = g(\mathbb{Y}, \mathbb{X}), \quad$ *para* $\quad \mathbb{X}, \mathbb{Y} \in T_p(\mathcal{M})$

2. $g(\mathbb{X}, a\mathbb{Y} + b\mathbb{Z}) = ag(\mathbb{X}, \mathbb{Y}) + bg(\mathbb{X}, \mathbb{Z})$

El producto escalar se dice *no degenerado* si $g(\mathbb{X},\mathbb{Y})=0 \quad \forall\, \mathbb{Y}\in T_p(\mathcal{M})$ implica $\mathbb{X}=0$. Obviamente el producto escalar es un tensor de tipo $(0,2)$. Como campo tensorial $g(\ ,\ )$ se puede expresar en términos de la base $\left(\frac{\partial}{\partial u^1}|_p\right),\ldots,\left(\frac{\partial}{\partial u^n}|_p\right)$

$$g(\mathbb{X},\mathbb{Y})=\sum_{ab}X^aY^bg\left(\frac{\partial}{\partial u^a},\frac{\partial}{\partial u^b}\right)$$

o, de manera equivalente

$$g(\mathbb{X},\mathbb{Y})=\sum_{ab}g_{ab}X^aY^b$$

donde $g_{ab}=g\left(\frac{\partial}{\partial u^a},\frac{\partial}{\partial u^b}\right)$. Si el producto escalar es no degenerado, entonces existe la matriz inversa de la matriz $g_{ab}$, a cuyos elementos los denotaremos por $g^{ab}$, de modo que

$$\sum_c g_{ac}g^{cb}=\delta^b_a$$

La existencia de un campo tensorial métrico (o producto escalar definido localmente), permite introducir la idea de *longitud de una curva*. En efecto, sea $\gamma(t),\quad t\in[a\,,\,b]$ una curva de clase $C^1$ sobre $\mathcal{M}$, que une los puntos $p$ y $q$: $\gamma(a)=p,\quad \gamma(b)=q$. En el punto $\gamma(t)$ tenemos el vector tangente a la curva $\gamma$ dado por

$$\left(\frac{\partial}{\partial t}\right)_\gamma=\sum_j\frac{d\gamma^j}{dt}\frac{\partial}{\partial x^j}$$

**Definición 3-83** (Longitud de una curva)**.** *La longitud de la curva $\gamma$ entre los puntos $p$ y $q$ está dada por la cantidad*

$$L=\int_a^b|g\left(\frac{\partial}{\partial t},\frac{\partial}{\partial t}\right)|^{\frac{1}{2}}dt \tag{4}$$

O, equivalentemente

$$L=\int_a^b\left|\sum_{ij}g_{ij}(x)\frac{d\gamma^i}{dt}\frac{d\gamma^j}{dt}\right|^{\frac{1}{2}}dt \tag{5}$$

## 3.6 Estructura Afin

En el espacio euclídeo $n$-dimensional (pensado aquí como una variedad diferenciable), cuando usamos coordenadas cartesianas, caracterizamos a dos vectors paralelos como aquellos que tienen iguales componentes. Si reemplazamos las coordenadas cartesianas por las polares, por ejemplo, esta caracterización deja de ser válida. Veamos cómo podemos introducir la noción de paralelismo de vectores, usando cualquier sistema de coordenadas. Sea $\{x^a\}$ el sistema de coordenadas cartesiano del espacio. En este sistema, hemos dicho que dos vectores paralelos, por ejemplo $\mathbb{V}$ y $\tilde{\mathbb{V}}$ tienen iguales componentes:

$$V^a=\tilde{V}^a$$

Si el vector $\mathbb{V}$ es tangente al espacio en el punto $p$ con coordenadas $\{x^a\}$ y el vector paralelo $\tilde{\mathbb{V}}$ es tangente al punto $q$ con coordenadas $x^a+\delta x^a$, vale

$$\tilde{V}^a(q)-V^a(p)=0$$

Dado un vector $\mathbb{V}$ en $p$, podemos definir un campo de vectores paralelos a $\mathbb{V}$ en un entorno de $p$. Denotemos a este campo por $\tilde{\mathbb{V}}$. Este campo cumple que en el punto $p$ coincide con $\mathbb{V}$ y con la condición:

$$\tilde{V}^a(x+\delta x) - V^a(x) = \frac{\partial \tilde{V}^a}{\partial x^b}(p)\delta x^b$$

Sea $\xi^a$ otro sistema de coordenadas para el espacio euclídeo, vinculado con $x^a$ mediante las relaciones

$$\xi^a = \xi^a(x^b), \quad x^b = x^b(\xi^a) \tag{6}$$

A partir de ellas, resulta

$$\delta\xi^a = \frac{\partial \xi^a}{\partial x^b}\delta x^b, \qquad \delta x^b = \frac{\partial x^b}{\partial \xi^a}\delta\xi^a \tag{7}$$

Las componentes de $\tilde{\mathbb{V}}$ se transforman de acuerdo con

$$\tilde{V}^a = \frac{\partial x^a}{\partial \xi^b}\tilde{V'}^b$$

donde $\tilde{V'}^a$ son las componentes de $\tilde{\mathbb{V}}$ en las coordenadas $\{\xi^a\}$. Entonces, podemos escribir

$$\begin{aligned} \frac{\partial \tilde{V}^a}{\partial x^b} &= \frac{\partial}{\partial \xi^c}\left(\frac{\partial x^a}{\partial \xi^d}\tilde{V'}^d\right)\frac{\partial \xi^c}{\partial x^b} \\ &= \frac{\partial^2 x^a}{\partial \xi^c \partial \xi^d}\tilde{V'}^d\frac{\partial \xi^c}{\partial x^a} + \frac{\partial x^a}{\partial \xi^d}\frac{\partial \tilde{V'}^d}{\partial \xi^c}\frac{\partial \xi^c}{\partial x^b} \end{aligned} \tag{8}$$

Si definimos la cantidad $\delta\tilde{V'}^d = \frac{\partial \tilde{V'}^d}{\partial x^e}\delta\xi^e$ y después de un poco de álgebra, llegamos a la relación

$$\delta\tilde{V'}^n = -\frac{\partial^2 x^a}{\partial \xi^e \partial \xi^d}\frac{\partial \xi^n}{\partial x^a}\tilde{V'}^d\delta\xi^e \tag{9}$$

Esta expresión puede reescribirse de la siguiente forma:

$$\delta\tilde{V'}^n = -\Gamma'^n_{ed}\tilde{V'}^d\delta\xi^e \tag{10}$$

en donde los coeficientes $\Gamma'$ están definidos en la expresión (9). De su definición resulta que las cantidades $\Gamma'$ se anulan para cambios *lineales* de coordenadas (6).

Obsérvese que al haber arribado a la definición de los coeficientes $\Gamma$ no hemos hecho uso de ninguna propiedad especial del espacio euclídeo. Es por ello que la expresión (9) es válida para cualquier variedad n-dimensional. Es fácil ver que frente a un cambio de coordenadas

$$x^a \to x'^a = x'^a\left(x^b\right) \tag{11}$$

las cantidades $\Gamma$ cambian según la expresión

$$\Gamma^f_{de} = \Gamma'^a_{mn}\frac{\partial x^f}{\partial x'^a}\frac{\partial x'^m}{\partial x^d}\frac{\partial x'^n}{\partial x^e} + \frac{\partial x^f}{\partial x'^a}\frac{\partial^2 x'^a}{\partial x^e \partial x^d} \tag{12}$$

Debemos remarcar que esta ley de transformación es lineal y homogénea (tensorial) sólo cuando el cambio de coordenadas (6) es lineal. Esta propiedad de los coeficientes $\Gamma$ nos permite generalizar la idea de paralelismo en una variedad arbitraria:

**Definición 3-84** (Conexión afín)**.** *Cuando en una variedad n-dimensional arbitraria $\mathcal{M}$ se introducen $n^3$ coeficientes $\Gamma$ que se transforman de acuerdo con la ley* (12)*, diremos que sobre esa variedad se ha definido una conexión afín*

A partir de los coeficientes $\Gamma$ es posible definir una nueva derivada para un campo vectorial arbitrario, digamos $V^a(x)$:

**Definición 3-85** (Derivada covariante de campo)**.** *Sea un campo vectorial $V$ definido en un entorno del punto $x$. La derivada covariante del campo $V$ está dado por las componentes de un tensor de tipo $(1,1)$*

$$V^a_{;c} = V^a_{,c} + \Gamma^a_{bc} V^b$$

**Definición 3-86** (Derivada covariante en una dirección)**.** *Dados dos campos vectoriales $U(x)$ y $V(x)$, la derivada covariante de $V$ en la dirección de $U$ es el campo vectorial definido por*

$$U(x) \cdot \nabla V(x) \equiv \sum_{ab} V^a_{;b}(x) U^b(x) \mathbb{E}_a \equiv \nabla_U V$$

donde $\mathbb{E}^a$ es el campo de vectores coordenados asociados con las coordenadas $x^a$. Esta última definición permite trasladar paralelamente a un vector a lo largo de una curva. Basta con tomar como $\mathbb{U}$ al campo tangente a la curva.

## 3.7 Variedad Riemmanniana

Sea $\mathcal{M}$ una variedad diferenciable $n$-dimensional. Si $\mathcal{M}$ tiene definida una métrica no singular sobre ella, recibe el nombre de *variedad Riemanniana*. La existencia de una métrica sobre $\mathcal{M}$ permite introducir una conexión afín particular, conocida como la conexión de Levi-Civita. Sean $g_{ab}$ y $g^{ab}$ los coeficientes de la métrica $g$ y su inversa, en las coordenadas $\{x^a\}$, respectivamente.

Para dos puntos próximos, la separación entre ellos viene dada por la expresión:

$$ds^2 = \sum_{ab} g_{ab} dx^a dx^b \tag{13}$$

Además de definir una distancia entre puntos próximos, la existencia de una métrica permite definir una conexión particular sobre una variedad riemanniana:

**Definición 3-87** (Conexión de Levi-Civita)**.** *La conexión de Levi-Civita en las coordenadas $x^a$ está dada por:*

$$\Gamma^a_{bc} = \frac{1}{2} \sum_d g^{ad} \left( g_{bd,c} + g_{cd.b} - g_{bc.d} \right) \tag{14}$$

La existencia de esta particular conexión no imposibilita la existencia de otras conexiones definidas sobre $\mathcal{M}$.

Como hemos visto más arriba, el tener definida una métrica permite definir la longitud de una curva. Bajo ciertas condiciones, que supondremos que se satisfacen, podemos plantearnos el problema de determinar

la curva que minimiza (en realidad extremiza) su longitud al unir dos puntos fijos sobre la variedad. Esto se puede tratar resolviendo el problema variacional asociado con el funcional (5). La ecuación de Euler-Lagrange conduce en este caso a:

$$\frac{d^2x^d}{dt^2} + \Gamma^d_{ca}\frac{dx^c}{dt}\frac{dx^a}{dt} = 0 \tag{15}$$

donde $x^a(t)$ son las coordenadas de la curva y $t$ es un parámetro adecuadamente elegido. Una curva que satisface (15), se llama una *curva geodésica*. Es posible caracterizar a una curva geodésica de otro modo. Sea $\mathbb{U}(t)$ el vector tangente a una curva $\gamma(t)$ definida sobre $\mathcal{M}$. La curva $\gamma$ se dice una geodésica si su vector tangente es trasladado paralelamente a lo largo de ella:

$$\mathbb{U}\cdot\nabla\mathbb{U} = f(t)\mathbb{U}$$

Siempre es posible elegir al parámetro $t$ de forma tal que $f(t) = 0$, con lo cual reobtenemos la ecuación (15).

El disponer de geodésicas, permite dar a una variedad riemanniana el carćter de espacio métrico. En efecto, podemos definir la distancia entre dos puntos $p$ y $q$ de la variedad $\mathcal{M}$ a través de la expresión:

$$d(p,q) = \min_{\gamma} L(\gamma) \tag{16}$$

donde el mínimo se evalúa entre todas las curvas que unen los puntos $p$ y $q$, y $L$ es la longitud (5). Como siempre, todo esto es posible ser realizado localmente. Las geodésicas son las curvas que localmente minimizan la distancia entre dos puntos. La distancia definida por (16) verifica la desigualdad triangular, y por eso es una métrica.

Dada una conexión $\nabla$ se define un tensor de tipo $(1,3)$, llamado *tensor de curvatura* asociado a la conexión $\nabla$, cuya espresión es:

$$\mathcal{R}(\mathbb{X},\mathbb{Y})\,\mathbb{Z} = \nabla_{\mathbb{X}}(\nabla_{\mathbb{Y}}\mathbb{Z}) - \nabla_{\mathbb{Y}}(\nabla_{\mathbb{X}}\mathbb{Z}) - \nabla_{[\mathbb{X},\mathbb{Y}]}\mathbb{Z}$$

Si los vectores $\mathbb{X}$, $\mathbb{Y}$ y $\mathbb{Z}$ son reemplazados por los vectors coordenados $\frac{\partial}{\partial x^a}$, $\frac{\partial}{\partial x^b}$ y $\frac{\partial}{\partial x^c}$, respectivamente, resulta

$$\mathcal{R}\left(\frac{\partial}{\partial x^a},\frac{\partial}{\partial x^b}\right)\frac{\partial}{\partial x^c} = R^d_{cab}\frac{\partial}{\partial x^d}$$

con

$$R^d_{cba} \equiv \left(\frac{\partial\Gamma^d_{cb}}{\partial x^a} + \Gamma^d_{ra}\Gamma^r_{cb} - \frac{\partial\Gamma^d_{ca}}{\partial x^b} - \Gamma^d_{rb}\Gamma^r_{ca}\right)\frac{\partial}{\partial x^d}$$

**Nota:** Si bien existe una motivación geométrica para introducir el tensor de curvatura, aquí no la hemos dado. Ella tiene que ver con la idea de cuánto cambia un vector al desplazarlo paralelamente a lo largo de una curva cerrada. En general diremos que una variedad es plana, si todas las componentes de su tensor de curvatura, se anulan.

Concluimos este capítulo con una breve nota histórica. En sus trabajos originales sobre geometría, B. Riemann introdujo el elemento de línea entre dos puntos vecinos $p$ y $q$ por medio de la expresión

$$ds = F(p,\mathbb{X})dt \tag{17}$$

con $F(p,\mathbb{X})$ una función homogénea de grado 2 en la segunda variable. Aquí estamos suponiendo que los puntos $p$ y $q$ tienen coordenadas $x^a$ y $x^a + X^a dt$, respectivamente. La geometría basada sobre el elemento

de línea se conoce como geometría de Finsler. Obsérvese que el elemento (13) (de Riemann) es un caso particular de la geometría de Finsler.

**CF libro de Bullet et al. (Bullet, Fearn & Smith, 2017), Cencov (Cencov, 1982), Amari (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.

Ali, S. M. & Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society B*, *28*(1), 131–142.

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

Arndt (2001). *Information Measures: Information and its Description in Sciences and Ingineering*. Berlin: Springer Verlag.

Athreya, K. B. & Lahiri, S. N. (2006). *Measure Theory and Probability Theory*. New-York: Springer.

Barnard, G. A. (1958). Studies in the history of probability and statistics: IX. Tomas Bayes's essay towards solving a problem in the doctrine of chances. *Biometrika*, *45*(3-4), 293–295.

Barone, J. & Novikoff, A. (1978). A history of the axiomatic formulation of probability from Borel to Kolmogorov: Part I. *Archive for History of Exact Sciences*, *18*(2), 123–190.

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.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418.

Beck, C. (2009). Generalised information and entropy measures in physics. *Contemporary Physics*, *50*(4),

495–510.

Bellhouse, D. (2005). Decoding cardano's liber de ludo aleae. *Historia Mathematica*, *32*(2), 180–202.

Ben-Tal, A., Bornwein, J. M., & Teboulle, M. (1992). Spectral estimation via convex programming. In F. Y. Phillips & J. J. Rousseau (Eds.), *Systems and Management Science by Extremal Methods* chapter 18, (pp. 275–290). Springer.

Ben-Tal, A., Charnes, A., & Teboulle, M. (1989). Entropic means. *Journal of Mathematical Analysis and Applications*, *139*(2), 537–551.

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

Bercher, J.-F. (2012). On a $(\beta, q)$-generalized Fisher information and inequalities invoving $q$-Gaussian distributions. *Journal of Mathematical Physics*, *53*(6), 063303.

Bercher, J.-F. (2013). On multidimensional generalized Cramér-Rao inequalities, uncertainty relations and characterizations of generalized $q$-Gaussian distributions. *Journal of Physics A*, *46*(9), 095303.

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

Bernoulli, J. (1713). *Ars conjectandi, opus posthumum. Accedit Tractatus de seriebus infinitis, et epistola gallicé scripta de ludo pilæ reticularis*. Basel, Switzeland: Thurneysen Brothers.

Bhatia, R. (1997). *Matrix Analysis*. New-York: Springer Verlag.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, *35*, 99–109.

Bhattacharyya, A. (1946). On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics*, *7*(4), 401–406.

Bienaymé, I.-J. (1853). Considérations à l'appui de la découverte de Laplace sur la loi de probabilité dans la méthode des moindres carrées. *Comptes Rendus de l'Académie des Sciences.*, *37*, 158–176.

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. (1979). Some aspects of error bounds in feature selection. *Pattern Recognition*, *11*(5-6), 353–360.

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

Bogachev, V. I. (2007a). *Measure Theory*, volume I. Berlin: Springer.

Bogachev, V. I. (2007b). *Measure Theory*, volume II. Berlin: Springer.

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.

Borel, E. (1898). *Le0.98628cons sur la théorie des fonctions*. Paris: Gauthier-Villars et fils.

Borel, E. (1909). *Éléments de la théorie des probabilités*. Paris: A. Hermann & fils.

Bouniakowsky, V. (1859). Sur quelques inégalités concernant les intégrales ordinaires et les intégrales aux différences finies. *Mémoires de l'Académie Impériale des Sciences de Saint-Petersbourg*, *I*(9).

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.

Brémaud, P. (1988). *An Introduction to Probabilitic Modeling*. New-York: Springer.

Bullet, S., Fearn, T., & Smith, F. (2017). *Analysis and Mathematical Physics*. London: World Scientific.

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 the 37th Meeting of the Society of Exploration Geophysicists*, 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.

Cardano, J. (1663). *Liber de ludo aleae, en "Opera Omnia"*, volume 1, (pp. 262–276). Lyon: cura Caroli Sponii.

Cauchy, A.-L. (1821). *Cours d'analyse de l'école royale polytechnique*, volume 1: analyse algébrique. Paris: Imprimerie royale (digital version, Cambrige, 2009).

Cencov, N. N. (1982). *Statistical Decision Rules and Optimal Inference*. Providence, Rhode Island, USA: American Mathematical Society.

Chenciner, A. (2017). La force d'une idée simple. *Gazette de la Société de Mathématiques Fran0.98628caise*, *152*, 16–22.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, *23*(4), 493–507.

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.

Cohn, D. L. (2013). *Measure Theory* (2nd ed.). New-York: Springer.

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.

Cressie, N. & Pardo, L. (2000). Minimum $\phi$-divergence estimator and hierarchical testing in loglinear models. *Statistica Sinica*, *10*(3), 867–884.

Cressie, N. & Read, L. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society B*, *46*(3), 440–464.

Csiszár, I. (1963). Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, *8*(1-2), 85–108.

Csiszàr, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, *2*, 299–318.

Csiszár, I. (1974). Information measures: A critical survey. In *Transactions of the Seventh Prague Conference on Information Theory*, volume B, (pp. 73–86)., Prague, 18-23 august.

Csiszàr, 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.

Csiszár, I. (1995). Generalized projections for non-negative functions. *Acta Mathematica Hungarica*, *68*(1-2), 161–186.

Csiszàr, I. & Matúš, F. (2012). Generalized minimizers of convex integral functionals, Bregman distance, Pythgorean identities. *Kybernetica*, *48*(4), 637–689.

Csiszár, I. & Shields, P. C. (2004). Information theory and statistics: A tutorial. *Foundations and Trends$^{TM}$ in Communications and Information Theory*, *1*(4), 417–528.

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

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

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

Daróczy, Z. & Járai, A. (1979). On the measurable solution of a functional equation arising in information theory. *Acta Mathematica Academiae Scientiarum Hungaricae*, *34*(1-2), 105–116.

de Laplace, P. S. (1820). *Théorie analytique des Probabilités* (3ème ed.). Paris: Gauthier-Villars.

de Moivre, A. (1730). *Miscellanea analytica de seriebus et quadraturis*. London: Londini: J. Tonson & J. Watts.

de Moivre, A. (1756). *The Doctrine of Chances : or, a method for calculating the probabilities of events in play* (3rd ed.). London: AMS Chelsea Publishing.

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.

(E. D. Sylla, Translator), J. B. (1713). *The Art of Conjecturing - Together with a "Letter to a Friend on Set in Court Tennis"*. Johns Hopkins University Press.

Ebeling, W., Molgedey, L., Kurths, J., & Schwarz, U. (2000). Entropy, complexity, predictability and data analysis of time series and letter sequences. In *Theory of Disaster* (A. Bundle and H.-J. Schellnhuber ed.). Berlin:

Springer Verlag.

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.

Euler, L. (1741). Observationes analyticae varias de combinationibus. *Commentarii academiae scientiarum Petropolitanae*, *13*, 64–93.

Euler, L. (1750). De partitione numerorum. *Novi Commentarii academiae scientiarum Petropolitanae*, *3*, 125–169.

Euler, L. (1768). *Lettres à une princesse d'Allemagne sur divers sujets de physique & de philosophie*, volume 2. Saint Petersbourg, Rusia: Académie Impériale des Sciences de Saint Petersbourg.

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

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

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

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications* (3 ed.)., volume 1. New-York: John Wiley & Sons, Inc.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, volume 2. New-York: John Wiley & Sons, Inc.

Ferentinos, K. (1982). On Tchebycheff's type inequalities. *Trabajos de Estadística e Investigación Operativa*, *33*(1), 125–132.

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

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

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

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

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

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

Frigyik, B. A., Srivastava, S., & Gupta, M. R. (2008). Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Transactions on Infomation 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.

Gel'fand, I. M. & Shilov, G. E. (1964). *Generalized Functions*, volume 1: Properties and Operations. New-York: Academic Press.

Gel'fand, I. M. & Shilov, G. E. (1968). *Generalized Functions*, volume 2: Spaces of Fundamental and Generalized Functions. New-York: Academic Press.

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.

Gupta, H. C. & Sharma, B. D. (1976). On non-additive measures of inaccuracy. *Czechoslovak Mathematical Journal*, *26*(4), 584–595.

Hadamard, J. (1893). Etude sur les propriétés des fonctions entières et en particulier d'une function considérée par Riemann. *Journal de Mathématiques Pures et Appliquées*, *58*(9), 171–215.

Haghighatshoar, S., Abbe, E., & Telatar, I. E. (2014). A new entropy power inequality for integer-valued random variables. *IEEE Transactions on Information Theory*, *60*(7), 3787–3796.

Hald, A. (1990). *History of Probability and Statistics and Their Applications before 1750*. John Wiley & Sons, Inc.

Halmos, P. R. (1950). *Measure Theory*. New-York: Springer.

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.

Harremoës, P. & Vignat, C. (2003). An entropy power inequality for the binomial family. *Journal of Inequalities in Pure and Applied Mathematics*, *4*(5), 93.

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

Hausdorff, F. (1901). Beiträge zur wahrscheinlichkeitsrechnung. *Berichte über die Verhandlungen der Köeniglich Sächsische Akademie der Wissenschaften zu Leipzig*, *53*(1), 152–178.

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

Hellinger, E. (1909). Neue bergründung der theorie quadratischer formen von unendlichvielen veränderlichen.

*Journal für die riene und angewandte Mathematik*, 210–271.

Hogg, R. V., McKean, J. W., & Craig, A. (2013). *Introduction to Mathematical Statistics* (7th ed.). Boston: Pearson.

Hölder, O. (1889). Ueber einen mittelwerthabsatz. *Nachrichten von der Königl. Gesellschaft der Wissenschaften und der Georg-Augusts-Universität zu Göttingen*, *2*, 38–47.

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.

Horn, R. A. & Johnson, C. R. (2013). *Matrix Analysis* (2nd ed.). Cambridge University Press.

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

Ibarrola, P., Pardo, L., & Quesada, V. (1997). *Teoría de la Probabilidad*. Madrid: Síntesis.

Jacob, J. & Protters, P. (2003). *Probability Essentials* (2nd ed.). Berlin: Springer.

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.

Jeffrey, H. (1973). *Scientific Inference* (3rd ed.). Cambridge: Cambridge University Press.

Jensen, J. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, *30*(1), 175–193.

Jessen, B. (1931a). Bemærkninger om konvekse funktioner og uligheder imellem middelærdier. I. *Matematisk Tidsskrift. B*, 17–28.

Jessen, B. (1931b). Bemærkninger om konvekse funktioner og uligheder imellem middelærdier. II. *Matematisk Tidsskrift. B*, 84–95.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. New-York: John Wiley & Sons.

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

Johnson, O. & Yu, Y. (2010). Monotonicity, thinning, and discrete versions of the entropy power inequality. *IEEE Transactions on Information Theory*, *56*(11), 5387–5395.

Kafka, P., Österreicher, F., & Vincze, I. (1991). On powers of $f$-divergences defining a distance. *Studia Scientiarum Mathematicarum Hungarica*, *24*(4), 415–422.

Kagan, A. (2001). A discrete version of the stam inequality and a characterization of the Poisson distributions.

*Journal of Statistical Planning and Inference*, *92*(1-2), 7–12.

Kagan, A. & Smith, P. J. (1999). A stronger version of matrix convexity as applied to functions of Hermitian matrices. *Journal of Inequalities and Applications*, *3*(2), 143–152.

Kagan, A. & Yu, T. (2008). Some inequalities related to the Stam inequality. *Applications of Mathematics*, *53*(3), 195–205.

Kailath, T. (1967). The divergence and Bhattacharyya distance measures in signal selection. *IEEE Transactions on Communications*, *15*(1), 52–60.

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. (1989). *Maximum Entropy Model in Sciences and Engineering*. New-Dehli: Wiley Eastern Limited.

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

Karamata, J. (1932). Sur une inegalité 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.

Knuth, D. E. (1997). *The Art of Computer Programming* (3rd ed.)., volume 1 / fundamental algorithms. Reading: Addison Wesley Longman.

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

Kolmogorov, A. N. (1956). *Foundations of the Theory of Probability* (2nd ed.). New-York: Chelsea Publishing Company.

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.

Kolmogorov, A. N. & Fomin, S. V. (1957). *Elements of the Theory of Function and Functional Analysis*, volume 1: Metric and Normed Spaces. Rochester, NY, USA: Graylock Press.

Kolmogorov, A. N. & Fomin, S. V. (1961). *Elements of the Theory of Function and Functional Analysis*, volume 2: Measure. The Lebesgue Integral. Hilbert Space. Rochester, NY, USA: Graylock Press.

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

Kraft Jr, L. G. (1949). A device for quantizing, grouping, and coding amplitude-modulated pulses. Master's

thesis, Department of Electrical Engineering, MIT, Massachusetts Institute of Technology.

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

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

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.

Kumar, P. & Chhina, S. (2005). A symmetric information divergence measure of the Csiszár's $f$-divergence class and its bounds. *Computers and Mathematics with Applications*, *49*(4), 575–588.

Langius, J. C. (1712). *Nvclevs Logicae Weisianae*. Giessen: Henningius Müllerus.

Laplume, J. (1948). Sur le nombre de signaux discernables en présence de bruit erratique dans un système de transmission à bande passance limitéee. *Comptes Rendus de l'Academie des Sciences*, *226*, 1348–1349. Séance du 26 avril.

Lebesgue, H. (1904). *Le0.98628cons sur l'Intégration et la recherche des Fonctions Primitives*. Paris: Gauthier-Villars et fils.

Lebesgue, H. (1918). Remarques sur les théories de la mesure et de l'intégration. *Annales Scientifiques de l'Ecole Normale Supérieure*, *35*, 191–250.

Lee, P. M. (1964). On the axioms of information theory. *The Annals of Mathematical Statistics*, *35*(1), 415–418.

Lehmann, E. L. & Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). New-York: Springer-Verlag.

Lieb, E. H. (1975). Some convexity and subadditvity properties of entropy. *Bulletin of the American Mathematical Society*, *81*(1), 1–13.

Lieb, E. H. (1978). Proof of an entropy conjecture of Wehrl. *Communications in Mathematical Physics*, *62*(1), 35–41.

Lieb, E. H. & Loss, M. (2001). *Analysis* (2nd ed.). Providence, Rhode Island: American Mathematical Society.

Liese, F. & Vajda, I. (2006). On divergence and informations in statistics and information theory. *IEEE Transactions on Information Theory*, *52*(10), 4394–4412.

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.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, *9*(70), 209–219.

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". *Telektronikk*, *98*(1), 20–29.

Lutwak, E., Lv, S., Yang, D., & Zhang, G. (2012). Extension of Fisher information and Stam's inequality. *IEEE Transactions on Information Theory*, *58*(3), 1319–1327.

Lutwak, E., Yang, D., & Zhang, G. (2005). Cramér-Rao and moment-entropy inequalities for Rényi entropy and generalized Fisher information. *IEEE Transactions on Information Theory*, *51*(2), 473–478.

Madiman, M. & Barron, A. (2007). Generalized entropy power inequalities and monotonicity properties of information. *IEEE Transactions on Information Theory*, *53*(7), 2317–2329.

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.

Markov, A. (1884). *On certain applications of algebraic continued fractions*. PhD thesis, University of Saint Petersburg, St. Petersburg, Russia.

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.

Menéndez, M. L., Morales, D., Pardo, L., & Vajda, I. (1977). Testing in stationary models based on divergences of observed and theoretical frequencies. *Kybernetica*, *33*(5), 465–475.

Merhav, N. (2010). Statistical physics and information theory. *Foundations and Trends® in Communications and Information Theory*, *6*(1-2), 1–212.

Merhav, N. (2018). *Statistical Physics for Electrical Engineering*. Springer.

Miller, R. E. (2000). *Optimization: Foundations and Applications*. New-York: John Wiley & Sons, inc.

Minkowski, H. (1910). *Geometrie der Zahlen*. Leipzig, Germany: Teubner.

Mittal, D. P. (1975). On additive and non-additive entropies. *Kybernetica*, *11*(4), 271–276.

Montagné, J.-C. B. (2008). *Transmissions. L'histoire des moyens de communication à distance depuis l'Antiquité jusqu'au milieu du xxe siècle*. Bagneux, JCB Montagné.

Morimoto, T. (1963). Markov processes and the H-theorem. *Journal of the Physical Society of Japan*, *18*(3), 328–331.

Mukhopadhyay, N. (2000). *Probability and Statistical Inference* (5th ed.)., volume 162 of *"Statistics: textbooks and monographs"*. New-York: Marcel Dekker.

Nagumo, M. (1930). Über eine klasse der mittelwerte. *Japanese journal of mathematics: transactions and abstracts*, *7*, 71–79.

Navarro, J. (2013). A very simple proof of the multivariate Chebyshev's inequality. *Communications in Statistics - Theory and Methods*, *45*(12), 3458–3463.

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.

Nikodym, O. (1930). Sur une généralisation des intégrales de M. J. Radon. *Fundamenta Mathematicae*, *15*(1), 131–179.

Ohya, M. & Petz, D. (1993). *Quantum Entropy and Its Use*. Berlin: Springer Verlag.

Olkin, I. & Pratt, J. W. (1958). A multivariate tchebycheff inequality. *The Annals of Mathematical Statistics*, *29*(1), 226–234.

Onicescu, O. (1966). Energie informationnelle. *Comptes rendus de l'académie des Sciences. série 1, mathématiques*, *263*(3), 841–842.

Orsak, G. C. & Paris, B.-P. (1995). On the relationship between measures of discrimination and the performance of suboptimal detectors. *IEEE Transactions on Information Theory*, *41*(1), 188–203.

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. (1996). On a class of perimeter-type distances of probability distributions. *Kybernetika*, *32*(4), 389–393.

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.

Pardo, L. (2006). *Statistical Inference Based on Divergence Measures*. Boca Raton, FL, USA: Chapman & Hall.

Pardo, M. C. (1999). On Burbea-Rao divergence based goodness-of-fit tests for multinomial models. *Journal of Multivariate Analysis*, *69*(1), 65–87.

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. (1905). "das fehlergesetz und seine verallgemeinerungen durch fechner und pearson.". A rejoinder. *Biometrika*, *4*(1/2), 169–212.

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.

Perlman, M. D. (1974). Jensen's inequality for a convex vector-valued function on an infinite-dimensional space. *Journal of Multivariate Analysis*, *4*(1), 52–65.

Petz, D. (2007). Bregman divergence as relative operator entropy. *Acta Mathematica Hungarica*, *116*(1-2), 127–131.

Phillips, F. Y. & Rousseau, J. J. (Eds.). (1992). *Systems and Management Science by Extremal Methods*. Springer.

Pigeon, S. (2003). Huffman coding. In K. Sayood (Ed.), *Lossless Compression Handbook* chapter 4, (pp. 79–99). San Diego, CA: Academic Press.

Pinsky, M. A. (2009). *Introduction to Fourier Analysis and Wavelets*, volume 102. Providence, Rhode Island, USA: American Mathematical Society.

Planck, M. (2015). *Eight Lectures on Theoretical Physics*. New-York: Columbia University Press.

Poor, H. V. (1988). Fine quantization in signal detection and estimation. *IEEE Transactions on Information Theory*, *34*(5), 960–972.

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.

Rudin, W. (1991). *Functional Analysis* (2nd ed.). New-York: McGraw-Hill.

Salicrú, M. (1987). Funciones de entropía asociada a medidas de Csiszár. *Qüestiió*, *11*(3), 3–12.

Salicrú, M. (1994). Measures of information associated with Csiszár's divergences. *Kybernetica*, *30*(5), 563–573.

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.

Schwartz, L. (1966). *Théorie des distributions*. Paris: Hermann.

Schwarz, H. A. (1888). Ueber ein die flächen kleinsten flächeninhalts betreffendes problem der variationsrechnung. *Acta societatis scientiarum Fennicæ*, *15*, 315–362.

Serrano Marugán, E. (2000). Etimología de algunos términos matemáticos. *Suma*, *35*, 87–96.

Shafer, G. & Vovk, V. (2006). The sources of Kolmogorov's grundbegriffe. *Statistical Science*, *21*(1), 70–98.

Shamai, S. & Wyner, A. (1990). A binary analog to the entropy-power inequality. *IEEE Transactions on Information Theory*, *36*(6), 1428–1430.

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.

Sharma, N., Das, S., & Muthukrishnan, S. (2011). Entropy power inequality for a family of discrete random variables. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, (pp. 1945–1949)., Saint Petersburg, Russia.

Sierpiński, W. (1918). Sur les définitions axiomatiques des ensembles mesurables. *Bulletin international de l'Académie des sciences de Cracovie: Série A. Classe des sciences mathématiques et naturelles – Sciences mathématiques*, 29–34.

Sierpiński, W. (1975). *Oeuvres choisies, Tome II: Théorie des ensembles et ses applications*. Warszawa, Poland: PWM Éditions scientifiques de Pologne.

Sierpiński, W. (1976). *Oeuvres choisies, Tome III: Théorie des ensembles et ses applications*. Warszawa, Poland: PWM Éditions scientifiques de Pologne.

Spiegel, M. (1976). *Probabilidad y Estadística*. México: McGrew Hill.

Stam, A. J. (1959). Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, *2*(2), 101–112.

Steele, J. M. (2004). *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*. Cambridge: Cambridge University Press.

Stellato, B., Van Parys, B. P. G., & Goulart, P. J. (2017). Multivariate Chebyshev inequality with estimated mean and variance. *The American Statistician*, *71*(2), 123–127.

Stirling, J. (1730). *Methodus Differentialissme Tractus de Summatione et Interpolatione Serierum Infinitarum*. London: Londini: Typis Gul. Bowyer; impensis G. Strahan.

Stix, G. (1991). Profile: Davis a. Huffman. *Scientific American*, *265*(3), 54–58.

Tchébichev, P. (1867). Des valeurs moyennes. *Journal de Mathématiques Pures et Appliquées*, *12*, 177–184.

Teboulle, M. (1992). On $\Phi$-divrgence and its applications. In F. Y. Phillips & J. J. Rousseau (Eds.), *Systems and Management Science by Extremal Methods* chapter 17, (pp. 255–273). Springer.

Toranzo, I. V., Zozor, S., & Brossier, J.-M. (2018). Generalization of the de Bruijn identity to general $\phi$-entropies and $\phi$-fisher informations. *IEEE Transactions on Information Theory*, on press.

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 $\alpha$-entropy of a generalized probability scheme. *Kybernetika*, *4*(2), 105–112.

Vajda, I. (1972). On the $f$-divergence and singularity of probability measures. *Periodica Mathematica Hungarica*, *2*(1-4), 223–234.

Vajda, I. (2009). On metric divergences of probability measures. *Kybernetika*, *45*(6), 885–900.

van Brakel, J. (1976). Some remarks on the prehistory of the concept of statistical probability. *Archive for History of Exact Sciences*, *16*(2), 119–136.

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.

Venn, J. M. A. (1880). I. On the diagrammatic and mechanical representation of propositions and reasonings. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *10*(59), 1–18.

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.

von Mises, R. (1932). Théorie des probabilités. fondements et applications. *Annales de l'institut Henri Poincaré*, *3*(2), 137–190.

von Plato, J. (2005). A.N. Kolmogorov, Grundbegriffe der wahrscheinlichkeitsrechnung (1933). In *Landmark Writings in Western Mathematics 1640-1940* chapter 75, (pp. 960–969). Elsevier.

Wang, L. & Madiman, M. (2004). Beyond the entropy power inequality via rearrangements. *IEEE Transactions on Information Theory*, *60*(9), 5116–5137.

Westfall, P. H. (2014). Kurtosis as peakedness, 1905-2014 R.I.P. *The American Statistician*, *68*(3), 191–195.

Wiener, N. (1948). *Cybernetics: or Control and Communication in the Animal end 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.

Zamir, R. (1998). A proof of the Fisher information inequality via a data processing argument. *IEEE Transactions on Information Theory*, *44*(3), 1246–1250.

Zhang, J. (2004). Divergence function, duality, and convex analysis. *Neural Computation*, *16*(1), 159–195.

Zozor, S., Puertas-Centeno, D., & Dehesa, J. S. (2017). On generalized Stam inequalities and Fisher–Rényi complexity measures. *Entropy*, *19*(9), 493.

# 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 Politénica 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 National de la Investigación Cientifica (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 Imagenes, 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 téoria de la información

(medidades informacionales generalizadas clásicas y cúanticas) con aplicaciones en procesamientos de datos, mecanica cúantica 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.