UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS
DEPARTAMENTO DE MATEMÁTICAS

GRANDES DESVIACIONES EN ESPACIOS DE MEDIDAS Y SUS REPRESENTACIONES

Tesis que presenta:
RODRIGO IÑIGO VARGAS

Para obtener el grado de:
LICENCIADO EN MATEMÁTICAS

Directora de la Tesis:
DRA. ANA MEDA GUARDIOLA

Ciudad Universitaria, Cd. Mx. Marzo 2020

1. Datos del alumno
Iñigo Vargas
Rodrigo
26 14 14 18
Universidad Nacional Autónoma de México
Facultad de Ciencias
Matemáticas
311095025

2. Datos del tutor
Dra.
Ana
Meda
Guardiola

3. Datos del sinodal 1
Dr.
Luis Antonio
Rincón
Solís

4. Datos del sinodal 2
Dr.
Francisco Javier
Torres
Ayala

5. Datos del sinodal 3
Dr.
Sergio Iván
López
Ortega

6. Datos del sinodal 4
Dr.
Fernando
Baltazar
Larios

7. Datos del trabajo escrito
Grandes desviaciones en espacios de medidas
y sus representaciones
193 p
2020

A mis padres, siempre. A la memoria de María de los Ángeles Canales Ruiz y a la memoria de Carlos Iñigo Martínez y Alfonso René Yñigo Martínez o más simplemente: a mi querida abuela y a mis queridos tíos Carlos y Poncho.

Agradecimientos

A la Dra. Ana Meda Guardiola: *Por confiar en mí para realizar este proyecto; por sus consejos, observaciones y la enorme paciencia requerida para completarlo.*

Al M. C. Arturo Govea Vargas: *Por sus sabios consejos acerca del lenguaje LaTeX y el haberme ayudado con la meticulosa tarea de dar una adecuada presentación a este trabajo.*

A los miembros del jurado: *Por sus valiosos comentarios y aportaciones que hicieron de este trabajo un mejor trabajo.*

A la planta docente del Departamento de Matemáticas: *Por compartir su conocimiento y experiencia pero sobre todo por mostrarme lo maravilloso de la actividad matemática y, por supuesto, de la actividad docente.*

A mis amigos: *Por compartir su inigualable compañía en los buenos y en los malos momentos.*

A mi estimada Familia: *Por ser mi primera escuela, y aunque este trabajo se realizó en la segunda, no hubiera sido posible sin las enseñanzas de la primera.*

Resumen

En este trabajo se desarrollan los métodos para encontrar, mediante la entropía relativa, una caracterización de la proyección de una medida de probabilidad en un conjunto. Asimismo se desarrollan las herramientas necesarias para saber bajo qué circunstancias un conjunto posee la propiedad de Sanov y una sucesión de variables aleatorias no independientes se comportan de manera independiente en el límite. Todo esto se definirá a lo largo del trabajo.

En el primer capítulo comenzamos por introducir los conceptos y teoremas básicos de la teoría de grandes desviaciones con especial énfasis en el teorema de Sanov para espacios de estado finito o numerable, así como el concepto de entropía relativa y sus respectivas propiedades. Después, en el segundo capítulo, presentamos el concepto de I-proyección e I-proyección generalizada de una medida de probabilidad Q (sobre un espacio medible $(S, \mathcal{B})$) en un subconjunto convexo Π. La I-proyección es la medida de probabilidad $P^* \in \Pi$ que minimiza la entropía relativa $D(P||Q)$ en Π, i.e.,

$$D(P^*||Q) = \inf_{P \in \Pi} D(P||Q),$$

y la I-proyección generalizada es la medida de probabilidad P^* que cumple que $P_n \to P^*$ para toda sucesión $\{P_n\}_{n \in \mathbb{N}}$ en Π con la propiedad de que

$$\lim_{n \to \infty} D(P_n||Q) = \inf_{P \in \Pi} D(P||Q).$$

Además se discuten algunas de las propiedades de las I-proyecciones con el objetivo de mostrar dos resultados que caracterizan a la I-proyección generalizada sobre dos conjuntos con ciertas características. Por último en el tercer capítulo se demuestra el teorema de Sanov en su versión más general; con ello introducimos la propiedad de Sanov junto con el concepto de cuasiindependencia asintótica y se presentan dos teoremas límite. El primero establece, bajo una condición de convexidad, una cota superior del tipo del teorema de Sanov pero para cada n natural, i.e.,

$$\frac{1}{n} \log \mathbb{P}(\{\hat{P}_X \in \Pi\}) \leq - \inf_{P \in \Pi} D(P||Q) \quad \forall n \in \mathbb{N}.$$

Este primer teorema también establece la posesión de la propiedad de Sanov por parte del conjunto Π y relaciona el comportamiento asintótico de una sucesión de variables aleatorias $X_1, ..., X_n, ...$ condicionadas con la I-proyección generalizada. En el último teorema se establece una caracterización de la I-proyección generalizada considerando el caso de un conjunto convexo $C \subset V$ en donde V es un espacio vectorial topológico localmente convexo y el evento $\{\hat{P}_X \in \Pi\}$ es el evento en el cual el promedio de la muestra $X_1, ..., X_n$ bajo un estadístico Ψ que toma valores en V cae en el conjunto C. Se analiza también el comportamiento límite de la sucesión de variables aleatorias condicionadas al evento anterior relacionando su distribución límite con la I-proyección generalizada.

El apéndice está dedicado a algunos conceptos y resultados de teoría de probabilidad, teoría de la medida, análisis, topología, etc. que fueron utilizados durante el desarrollo y construcción teórica de los capítulos anteriores.

Índice general

Introducción

El propósito principal de este trabajo es responder algunas preguntas: Las primeras dos preguntas son cuestiones de optimización y están relacionadas con el concepto de entropía relativa entre dos medidas de probabilidad que, *grosso modo*, es una función que nos permite comparar dos medidas de probabilidad y ver qué tan "parecidas" son entre ellas. Dado un conjunto de medidas de probabilidad Π nos preguntamos, en primer lugar, ¿Bajo qué circunstacias se puede minimizar la entropía relativa de las medidas de probabilidad en Π respecto a una medida de probabilidad fija? Y en caso de que se pueda minimizar, entonces ¿De qué manera se ve la medida de probabilidad que minimiza la entropía relativa? Es decir ¿Se puede dar una descripción analítica de esta medida de probabilidad? Para responder a estas dos preguntas se desarrollan los conceptos de entropía relativa, I-proyección generalizada de una medida de probabilidad sobre un conjunto de medidas de probabilidad y se prueban los resultados que nos permitirán abordar la cuestión. La respuesta final a esta pregunta se encuentra en los teoremas 2.3.1 y 2.3.2.

Ahora si se tiene una sucesión de variables aleatorias idénticamente dístribuidas, ¿Cuándo la probabilidad de que la medida empírica caiga en un conjunto de medidas de probabilidad Π es igual al ínfimo de la entropía relativa de las medidas de probabilidad en dicho conjunto respecto a su distribución común?, es decir, ¿Qué propiedades debe cumplir Π? Esta pregunta está intrínsecamente relacionada con la teoría de grandes desviaciones por lo cual es necesario desarrollar algunos resultados y conceptos de dicha teoría, especialmente lo que se conoce como propiedad de Sanov. Por último, nos preguntamos ¿Cuándo una sucesión de variables aleatorias condicionadas a cierto evento se comportan de manera "independiente" en el límite? Para responder a esta cuestión se desarrolla el concepto de cuasiindependencia asintótica.

Todo lo anterior queda sintetizado en un último teorema considerando una función que manda nuestras variables aleatorias a un espacio vectorial topológico localmente convexo. Ver el Teorema 3.3.1.

Por último comentamos que a lo largo de este trabajo utilizaremos las siguientes convenciones utilizadas dentro del marco de la teoría de grandes desviaciones. En lo sucesivo convendremos en que $\log(0) = -\infty$, $0 \cdot \infty = 0$, $0 \cdot -\infty = 0$, $\log(\frac{x}{0}) = \infty$ con x real positivo y $\log(\frac{0}{0}) = 0$.

Capítulo 1

Grandes desviaciones

El objetivo principal de este capítulo es introducir los conceptos y resultados básicos de la teoría de grandes desviaciones que nos serán de gran utilidad en los capítulos posteriores. Los principales resultados como el teorema de Cramér y el teorema de Sanov en ambas versiones son tomados de [22]. Otro material de consulta se tomó de [14], [19] y [9].

El teorema de Sanov nos permitirá definir la propiedad de Sanov que aparece en los resultados centrales de este trabajo. También se define la entropía relativa, que nos permitirá establecer la I-proyección generalizada en el siguiente capítulo.

1.1. ¿Qué es una desviación grande?

Cuando tenemos un sistema no determinista nos interesa saber si éste sigue un comportamiento predecible. En el caso de un experimento aleatorio nos interesa, por ejemplo, el resultado esperado al realizar dicho experimento. Dos de los teoremas centrales en la teoría de probabilidad nos dan información acerca de dicho comportamiento: el primero de ellos, conocido como la ley de los grandes números (LLN por sus siglas en inglés) nos da información acerca de la relación entre la esperanza de una variable aleatoria y la media muestral. Por otra parte, el segundo resultado conocido como teorema central del límite (CLT por sus siglas en inglés) nos indica cómo se da la convergencia en el caso anterior.

Comencemos con $\{X_i\}_{i\in\mathbb{N}}$ una sucesión de variables aleatorias independientes[1] e idénticamente distribuidas sobre un espacio de probabilidad $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P})$ con $\mathcal{B}(\mathbb{R})$ la σ-álgebra de Borel en $\mathbb{R}$, $\mathbb{E}(X_1) = \mu \in \mathbb{R}$ y $Var(X_1) = \sigma^2 \in (0, \infty)$. Ahora sea

[1] c.f. [16], pág 252.

$$S_n = \sum_{i=1}^{n} X_i.$$

Entonces enunciaremos los teoremas LLN y CLT que vamos a utilizar[2].

Ley de los Grandes Números (LLN).

$$\lim_{n\to\infty} \frac{1}{n} S_n = \mu \quad \text{casi seguramente,}$$

lo cual quiere decir que

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{1}{n} S_n = \mu \right) = 1.$$

Teorema del Límite Central.

$$\lim_{n\to\infty} \frac{1}{\sigma\sqrt{n}} (S_n - n\mu) \xrightarrow{\text{d}} N(0,1).$$

en donde $\xrightarrow{\text{d}}$ denota convergencia en distribución, cf. La subsección 4.1.1. Modos de convergencia del apéndice.

La cuestión es ¿qué sucede si $\frac{1}{n} S_n$ difiere de μ por una cantidad de orden mayor a $\sqrt{n}$? (Pues este es el orden de velocidad de convergencia que nos da el Teorema del Límite Central) Digamos, por ejemplo, n. Es decir ¿es posible que el promedio empírico se desvíe de la media cuando la muestra es "grande"? LLN nos dice que esto no puede suceder muy frecuentemente, por lo tanto tenemos que si $a > \mathbb{E}(X_1)$ entonces

$$\lim_{n\to\infty} \mathbb{P}(S_n \geq na) = 0.$$

$\{S_n \geq na\}_{n\in\mathbb{N}}$ es una sucesión de eventos raros y su probabilidad de ocurrencia converge a 0. Lo que a nosotros nos interesa es calcular con qué velocidad lo hace. Veremos que de hecho esta probabilidad decae exponencialmente con la velocidad dada por una función de tasa siempre que las variables tengan función generadora de momentos finita en alguna vecindad del origen. La velocidad buscada se da en el primer resultado de la teoría de grandes desviaciones que presentamos a continuación.

[2]No pretendemos dar sus versiones más generales.

1.2. Teorema de Cramér

El teorema de Cramér es el primer resultado que veremos de grandes desviaciones. Este resultado nos proporciona información acerca de la velocidad de decaimiento, es decir, información sobre la función $\mathcal{I}$ que definimos en la Definición 1.2.1. El resultado establece que la función de tasa es la transformada de Fenchel-Legendre[3] de la función generadora acumulada de X_1. Este teorema se puede probar de manera más general. La forma que aquí elegimos es para introducir de manera más amigable los problemas, técnicas y conceptos de la teoría de grandes desviaciones.

Definición 1.2.1. *Sean $\{X_i\}_{i\in\mathbb{N}}$ variables aleatorias reales independientes e idénticamentes distribuidas tales que su función generadora de momentos es finita en todo $\mathbb{R}$, es decir, $\varphi(t) = \mathbb{E}(e^{tX_1}) < \infty$ para toda $t \in \mathbb{R}$. Sea $z \in \mathbb{R}$ entonces definimos la función $\mathcal{I} : \mathbb{R} \to \mathbb{R}$ de la siguiente manera:*

$$\mathcal{I}(z) = \sup_{t\in\mathbb{R}}[zt - \log\varphi(t)].$$

A esta función se le conoce como la transformada de Fenchel-Legendre de φ.

Teorema 1.2.1. *(Cramér)*[4] *Sean $\{X_i\}_{i\in\mathbb{N}}$ variables aleatorias reales independientes e idénticamentes distribuidas tales que $\varphi(t)$ es finita para toda $t \in \mathbb{R}$. Sea $S_n = \sum_{i=1}^n X_i$. Entonces para cada $a > \mathbb{E}(X_1) = \mu$ se tiene que*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \geq an) = -\mathcal{I}(a), \tag{1.1}$$

con $\mathcal{I}$ la transformada de Fenchel-Legendre recién definida.

Demostración. Para demostrar el teorema primero observamos que podemos suponer sin pérdida de generalidad que $a = 0$ y $\mathbb{E}(X_1) = \mu < a$ (de otra forma podemos trabajar con $X_1 - \mu$). Con esta suposición debemos trabajar con $\mathcal{I}(0)$.

Consideremos el caso en el que X_1 es no degenerada ya que si, por el contrario, $X_1 \equiv \mu$, entonces $\mathbb{P}(S_n \geq an) = 0$ para toda n y por lo tanto

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \geq an) = -\infty.$$

Por otro lado

$$-\mathcal{I}(a) = -\sup_{t\in\mathbb{R}}[at - \log\varphi(t)] = -\sup_{t\in\mathbb{R}}[at - \log e^{t\mu}] = -\sup_{t\in\mathbb{R}}[t(a-\mu)] = -\infty.$$

[3] Función de gran relevancia en el análisis convexo, cf. [29].

[4] Enunciamos esta versión del teorema de Cramér aunque no es la más general.

Y así obtenemos que (1.1) es válido trivialmente para variables aleatorias degeneradas.

Para facilitar el trabajo haremos uso de la siguiente notación:

$$\rho = \inf_{t\in\mathbb{R}} \varphi(t).$$

Notemos que $\mathcal{I}(0) = \sup_{t\in\mathbb{R}}[-\log\varphi(t)] = -\log[\inf_{t\in\mathbb{R}}\varphi(t)] = -\log\rho$ donde en la penúltima igualdad hacemos uso del hecho de que $f(x) = \log(x)$ es continua y creciente, por lo cual, basta probar que

$$\lim_{n\to\infty} \frac{1}{n}\log\mathbb{P}(S_n \geq 0) = \log\rho. \tag{1.2}$$

Sea $F(x)$ la función de distribución de X_1. Como $\varphi(t)$ es finito en todo $\mathbb{R}$ tenemos que su n-ésima derivada es $\varphi^{(n)}(t) = \mathbb{E}(X^n e^{Xt})$ para toda $n \in \mathbb{N}$ (esto se prueba detalladamente en Lema 4.1.1 del apéndice). En particular

$$\varphi'(t) = \int_{\mathbb{R}} xe^{tx}dF(x)$$

y

$$\varphi''(t) = \int_{\mathbb{R}} x^2e^{tx}dF(x).$$

De lo anterior se sigue que $\varphi''(t) > 0$ para toda $t \in \mathbb{R}$ pues la función $f(x) = x^2e^{tx}$ es mayor que cero en todo $\mathbb{R}$ excepto en el cero; sin embargo, si la variable es continua este punto tiene medida de Lebesgue cero, si es discreta notamos que la masa de probabilidad no puede estar concentrada solamente en el cero y si es mixta de nuevo la masa de probabilidad no puede estar concentrada en el cero. Como la segunda derivada de φ es positiva entonces φ es estrictamente convexa, además $\varphi'(0) = \mathbb{E}(X_1)$. Consideraremos tres casos dependiendo de en donde se encuentre acumulada la masa de probabilidad de X_1.

Caso 1: $\mathbb{P}(X_1 < 0) = 1$.

En este caso tenemos que φ es decreciente. Esto se sigue de

$$\varphi'(t) = \int_{\mathbb{R}} xe^{tx}dF(x) = \int_{-\infty}^{0} xe^{tx}dF(x) + 0 = \int_{-\infty}^{0} xe^{tx}dF(x) < 0.$$

Lo anterior pues $xe^{tx} < 0$ para toda $x \in (-\infty, 0)$ y para toda $t \in \mathbb{R}$ y de nuevo observamos que la masa de probabilidad no puede estar concentrada sólo en el cero. Entonces φ es decreciente y observemos que $e^{tx} \leq 1$ para toda $x \in (0, -\infty)$ y para toda $t > 0$. Luego, por el teorema de convergencia

dominada de Lebesgue (CDL) pues ya vimos que la función constante 1 domina, cf. Teorema 4.4.4 del apéndice, se tiene que

$$0 = \int_{-\infty}^{0} \lim_{t\to\infty} e^{tx} dF(x) = \lim_{t\to\infty} \varphi(t) = \inf_{t\in\mathbb{R}} \varphi(t) = \rho.$$

Como $\mathbb{P}(X_1 < 0) = 1$ entonces $\mathbb{P}(S_n \geq 0) = 0$ para cada $n \in \mathbb{N}$. Por lo tanto se tiene que $\log \mathbb{P}(S_n \geq 0) = -\infty = \log \rho$, esto es, se obtiene (1.2).

Caso 2: $\mathbb{P}(X_1 \leq 0) = 1$ y $\mathbb{P}(X_1 = 0) > 0$.

Observemos que no puede suceder que $\mathbb{P}(X_1 = 0) = 1$ puesto que el caso en que X_1 es una constante ya fue probado. Análogamente al caso 1 se tiene que la función

$$\varphi(t) = \int_{-\infty}^{0} e^{tx} dF(x)$$

es decreciente pues de nuevo se observa que

$$\varphi'(t) = \int_{\mathbb{R}} xe^{tx} dF(x) = \int_{-\infty}^{0} xe^{tx} dF(x) + 0 = \int_{-\infty}^{0} xe^{tx} dF(x) < 0$$

y la masa de probabilidad no puede estar concentrada solo en el cero por hipótesis. De nuevo por el teorema CDL (con la constante 1 como la función que domina) y como en este caso $\mathbb{P}(X_1 = 0) > 0$, entonces

$$\lim_{t\to\infty} \varphi(t) = \int_{-\infty}^{0} \lim_{t\to\infty} e^{tx} dF(x) = 0 + \lim_{t\to\infty} e^{t\cdot 0}\mathbb{P}(X_1 = 0) = \mathbb{P}(X_1 = 0) > 0.$$

Así, como φ es decreciente y convexa, se tiene que

$$\rho = \inf_{t\in\mathbb{R}} \varphi(t) = \lim_{t\to\infty} \varphi(t) = \mathbb{P}(X_1 = 0) > 0.$$

Y observemos que

$$\mathbb{P}(S_n \geq 0) = \mathbb{P}(S_n > 0) + \mathbb{P}(S_n = 0) = 0 + \mathbb{P}(X_1 = 0, ..., X_n = 0)$$

$$= \prod_{i=1}^{n} \mathbb{P}(X_i = 0) = \rho^n.$$

Por lo tanto

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \geq 0) = \lim_{n\to\infty} \frac{1}{n} \log \rho^n = \log \rho.$$

Así hemos obtenido (1.2).

Caso 3: $\mathbb{P}(X_1 < 0) > 0$ y $\mathbb{P}(X_1 > 0) > 0$.

En este caso tenemos

$$\lim_{t\to\infty} \varphi(t) = \lim_{t\to\infty} \left[\int_{-\infty}^{0} e^{tx} dF(x) + \int_{0}^{\infty} e^{tx} dF(x) \right]$$

$$= \int_{-\infty}^{0} \lim_{t\to\infty} e^{tx} dF(x) + \lim_{t\to\infty} \int_{0}^{\infty} e^{tx} dF(x) = 0 + \lim_{t\to\infty} \int_{0}^{\infty} e^{tx} dF(x)$$

$$\geq \liminf_{n\to\infty} \int_{0}^{\infty} e^{tx} dF(x) \geq \int_{0}^{\infty} \liminf_{n\to\infty} e^{tx} dF(x) = \infty.$$

En donde en la segunda igualdad hacemos uso del teorema CDL y el la última desigualdad hacemos uso del lema de Fatou, cf. Lema 4.4.3 del apéndice. Por lo tanto

$$\lim_{t\to\infty} \varphi(t) = \infty.$$

De manera similar se obtiene que

$$\lim_{t\to-\infty} \varphi(t) = \infty.$$

Como φ es estrictamente convexa entonces tiene un único punto mínimo, i.e., existe un único $\tau \in \mathbb{R}$ tal que $\varphi(\tau) = \rho$ y $\varphi'(\tau) = 0$.

Ahora recordemos la desigualdad de Markov cf. Teorema 4.1.1 del apéndice:

$$\mathbb{P}(X \geq \epsilon) \leq \frac{\mathbb{E}(X)}{\epsilon}, \tag{1.3}$$

con X una variable aleatoria real no negativa y $\epsilon > 0$. Aplicando (1.3) a la variable aleatoria real y no negativa $e^{\tau S_n}$ obtenemos que

$$\mathbb{P}(S_n \geq 0) = \mathbb{P}(e^{\tau S_n} \geq 1) \leq \mathbb{E}(e^{\tau S_n}) = \prod_{i=1}^{n} \mathbb{E}(e^{\tau X_i}) = [\varphi(\tau)]^n = \rho^n.$$

Aplicando logaritmo, multiplicando por $\frac{1}{n}$ y tomando límite superior obtenemos

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \geq 0) \leq \log \rho. \tag{1.4}$$

Para obtener la cota inferior haremos uso de herramientas más refinadas. A saber, la transformada de Cramér y tres lemas adicionales.

Sean $\{\hat{X}_i\}_{i\in\mathbb{N}}$ variables aleatorias independientes e idénticamente distribuidas con distribución común dada por

$$\hat{F}(x) = \frac{1}{\rho}\int_{-\infty}^{x} e^{\tau y} dF(y). \tag{1.5}$$

A (1.5) se le conoce como la transformada de Cramér.

Observación.

$$\int_{\mathbb{R}} e^{\tau y} dF(y) = \mathbb{E}(e^{\tau X_1}) = \varphi(\tau) = \rho.$$

Lema 1.2.1. $\mathbb{E}(\hat{X}_1) = 0$ *y* $Var(\hat{X}_1) = \hat{\sigma}^2 \in (0,\infty)$.

Prueba. Sea $\hat{\varphi}(t) = \mathbb{E}(e^{t\hat{X}_1})$. Entonces

$$\hat{\varphi}(t) = \int_{\mathbb{R}} e^{tx} d\hat{F}(x) = \int_{\mathbb{R}} e^{tx}\frac{1}{\rho}e^{\tau x} f(x)\ dx = \frac{1}{\rho} \quad \varphi(t+\tau) < \infty \quad \forall t \in \mathbb{R}.$$

Así, por el Lema 4.1.1 del apéndice, existe la derivada de cualquier orden de $\hat{\varphi}$ y $\mathbb{E}(\hat{X}_1) = \hat{\varphi}'(0) = \frac{1}{\rho}\varphi'(\tau) = 0$. Y $Var(\hat{X}_1) = \hat{\varphi}''(0) = \frac{1}{\rho}\varphi''(\tau) \in (0,\infty)$ ya que como τ es mínimo se tiene que $\varphi''(\tau) > 0$.

□

Lema 1.2.2. *Sea* $\hat{S}_n = \sum_{i=1}^{n} \hat{X}_i$ *entonces* $\mathbb{P}(S_n \geq 0) = \rho^n \mathbb{E}\Big(e^{-\tau\hat{S}_n} 1_{\{\hat{S}_n \geq 0\}}\Big)$.

Prueba.

$$\mathbb{P}(S_n \geq 0) = \int_{\{(x_1,\ldots,x_n)|x_1+\ldots+x_n\geq 0\}} dF(x_1)...dF(x_n) =$$

$$\int_{\{(x_1,\ldots,x_n)|x_1+\ldots+x_n\geq 0\}} (\rho e^{-\tau x_1} d\hat{F}(x_1))...(\rho e^{-\tau x_n} d\hat{F}(x_n)) =$$

$$\rho^n \int_{\{(x_1,\ldots,x_n)|x_1+\ldots+x_n\geq 0\}} e^{-\tau\sum_{i=1}^{n} x_i} d\hat{F}(x_1)...d\hat{F}(x_n) = \rho^n \mathbb{E}(e^{-\tau\hat{S}_n} 1_{\{\hat{S}_n\geq 0\}}).$$

□

Lema 1.2.3. $\liminf_{n\to\infty} \frac{1}{n}\log \mathbb{E}\Big(e^{-\tau\hat{S}_n} 1_{\{\hat{S}_n\geq 0\}}\Big) \geq 0.$

Prueba. Sea $C > 0$ tal que

$$\frac{1}{4} < \frac{1}{\sqrt{2\pi}} \int_0^C e^{\frac{x^2}{2}}\ dx < 1.$$

Ahora, por la desigualdad de Markov cf. (1.3) tenemos que

$$e^{-\tau C\hat{\sigma}\sqrt{n}}\mathbb{P}\Big(e^{-\tau \hat{S}_n}1_{\{\hat{S}_n\geq 0\}} \geq e^{-\tau C\hat{\sigma}\sqrt{n}}\Big) \leq \mathbb{E}\Big(e^{-\tau \hat{S}_n}1_{\{\hat{S}_n\geq 0\}}\Big). \tag{1.6}$$

Por otro lado notemos que

$$\mathbb{P}\Big(e^{-\tau \hat{S}_n}1_{\{\hat{S}_n\geq 0\}} \geq e^{-\tau C\hat{\sigma}\sqrt{n}}\Big) = \mathbb{P}\Big(-\tau \hat{S}_n 1_{\{\hat{S}_n\geq 0\}} \geq -\tau C\hat{\sigma}\sqrt{n}\Big).$$

Lo anterior ya que si $1_{\{\hat{S}_n\geq 0\}} = 1$ casi seguramente entonces

$$\Big\{e^{-\tau \hat{S}_n}1_{\{\hat{S}_n\geq 0\}} \geq e^{-\tau C\hat{\sigma}\sqrt{n}}\Big\} = \Big\{-\tau \hat{S}_n 1_{\{\hat{S}_n\geq 0\}} \geq -\tau C\hat{\sigma}\sqrt{n}\Big\},$$

simplemente aplicando logaritmo. Ahora, si $1_{\{\hat{S}_n\geq 0\}} = 0$ casi seguramente entonces los eventos

$$\Big\{e^{-\tau \hat{S}_n}1_{\{\hat{S}_n\geq 0\}} \geq e^{-\tau C\hat{\sigma}\sqrt{n}}\Big\} \quad \text{y} \quad \Big\{-\tau \hat{S}_n 1_{\{\hat{S}_n\geq 0\}} \geq -\tau C\hat{\sigma}\sqrt{n}\Big\}$$

tienen la misma probabilidad de ocurrencia pues los eventos

$$\big\{0 \geq e^{-\tau C\hat{\sigma}\sqrt{n}}\big\} \quad \text{y} \quad \{-\infty \geq -\tau C\hat{\sigma}\sqrt{n}\}$$

tienen probabilidad igual a cero para toda n.

Ahora, ocurre que

$$\mathbb{P}\Big(-\tau \hat{S}_n 1_{\{\hat{S}_n\geq 0\}} \geq -\tau C\hat{\sigma}\sqrt{n}\Big) = \mathbb{P}\Big(\hat{S}_n 1_{\{\hat{S}_n\geq 0\}} \leq C\hat{\sigma}\sqrt{n}\Big)$$

$$= \mathbb{P}\left(\frac{\hat{S}_n 1_{\{\hat{S}_n\geq 0\}}}{\hat{\sigma}\sqrt{n}} \leq C\right) = \mathbb{P}\left(\frac{\hat{S}_n}{\hat{\sigma}\sqrt{n}} \in [0, C]\right)$$

si $\tau \geq 0$ y

$$\mathbb{P}\Big(-\tau \hat{S}_n 1_{\{\hat{S}_n\geq 0\}} \geq -\tau C\hat{\sigma}\sqrt{n}\Big) = 1 - \mathbb{P}\left(\frac{\hat{S}_n}{\hat{\sigma}\sqrt{n}} \in [0, C]\right)$$

si $\tau < 0$. Entonces de (1.6) se sigue que

$$e^{\tau C\hat{\sigma}\sqrt{n}}\mathbb{P}\left(\frac{\hat{S}_n}{\hat{\sigma}\sqrt{n}} \in [0, C]\right) \leq \mathbb{E}\Big(e^{-\tau \hat{S}_n}1_{\{\hat{S}_n\geq 0\}}\Big)$$

ó bien

$$e^{\tau C\hat{\sigma}\sqrt{n}}\left[1-\mathbb{P}\left(\frac{\hat{S}_n}{\hat{\sigma}\sqrt{n}}\in[0,C]\right)\right]\leq\mathbb{E}\Big(e^{-\tau\hat{S}_n}1_{\{\hat{S}_n\geq 0\}}\Big).$$

Observemos que el Lema 1.2.1 nos permite aplicar el Teorema del Límite Central a la variable aleatoria $\frac{\hat{S}_n}{\hat{\sigma}\sqrt{n}}$. Así, para n suficientemente grande se tiene que

$$\frac{1}{4}<\frac{1}{\sqrt{2\pi}}\int_0^C e^{\frac{x^2}{2}}\ dx=\mathbb{P}\left(\frac{\hat{S}_n}{\hat{\sigma}\sqrt{n}}\in[0,C]\right)<1.$$

luego

$$e^{\tau C\hat{\sigma}\sqrt{n}}\frac{1}{4}\leq\mathbb{E}\Big(e^{-\tau\hat{S}_n}1_{\{\hat{S}_n\geq 0\}}\Big),$$

ó bien, como $0<1-\mathbb{P}\Big(\frac{\hat{S}_n}{\hat{\sigma}\sqrt{n}}\in[0,C]\Big)$ entonces existe $\epsilon>0$ tal que $1-\mathbb{P}\Big(\frac{\hat{S}_n}{\hat{\sigma}\sqrt{n}}\in[0,C]\Big)=\epsilon$, luego

$$e^{\tau C\hat{\sigma}\sqrt{n}}\epsilon\leq\mathbb{E}\Big(e^{-\tau\hat{S}_n}1_{\{\hat{S}_n\geq 0\}}\Big).$$

Así, se obtiene que

$$\frac{1}{\sqrt{n}}\tau C\hat{\sigma}+\frac{1}{n}\log\Big(\frac{1}{4}\Big)=\frac{1}{n}\tau C\hat{\sigma}\sqrt{n}+\frac{1}{n}\log\Big(\frac{1}{4}\Big)=\frac{1}{n}\log\Big[e^{\tau C\hat{\sigma}\sqrt{n}}\frac{1}{4}\Big]$$

$$\leq\frac{1}{n}\log\mathbb{E}\Big(e^{-\tau\hat{S}_n}1_{\{\hat{S}_n\geq 0\}}\Big),$$

ó bien

$$\frac{1}{\sqrt{n}}\tau C\hat{\sigma}+\frac{1}{n}\log\epsilon\leq\frac{1}{n}\log\mathbb{E}\Big(e^{-\tau\hat{S}_n}1_{\{\hat{S}_n\geq 0\}}\Big).$$

Así, al tomar el límite inferior tenemos que

$$0=\liminf_{n\to\infty}\left[\frac{1}{\sqrt{n}}\tau C\hat{\sigma}+\frac{1}{n}\log\Big(\frac{1}{4}\Big)\right]\leq\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{E}\Big(e^{-\tau\hat{S}_n}1_{\{\hat{S}_n\geq 0\}}\Big).$$

que es lo que se quería demostrar.

□

Observación. *Podemos notar que la elección de $\frac{1}{4}$ como cota es arbitraria, de hecho podemos utilizar cualquier número mayor que cero.*

Para concluir la demostración del teorema basta observar que del Lema 1.2.2 se sigue que

$$\frac{1}{n}\log\big(\mathbb{P}(S_n \geq 0)\big) = \frac{1}{n}\log\left[\rho^n \mathbb{E}\Big(e^{-\tau \hat{S}_n} 1_{\{\hat{S}_n \geq 0\}}\Big)\right]$$

$$= \log\rho + \frac{1}{n}\log\mathbb{E}\Big(e^{-\tau \hat{S}_n} 1_{\{\hat{S}_n \geq 0\}}\Big).$$

Al tomar límite inferior y por el Lema 1.2.3 obtenemos

$$\liminf_{n\to\infty}\left[\frac{1}{n}\log\big(\mathbb{P}(S_n \geq 0)\big) - \log\rho\right] = \liminf_{n\to\infty}\frac{1}{n}\log\mathbb{E}\Big(e^{-\tau \hat{S}_n} 1_{\{\hat{S}_n \geq 0\}}\Big) \geq 0.$$

Es decir

$$\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}(S_n \geq 0) \geq \log\rho. \tag{1.7}$$

Por lo tanto, de (1.4) y (1.7) se sigue (1.2). □

Existen varias generalizaciones de este teorema cf. [14]. Lo que ahora nos interesa son generalizaciones en espacios de medidas que se conocen como teoremas de Sanov.

1.3. Teorema de Sanov (Versión finita)

En la sección anterior nos familiarizamos con los conceptos de la teoría de grandes desviaciones y demostramos el primer teorema concerniente a esta. El teorema de Cramér nos da información sobre la velocidad de decaimiento de la probabilidad de que el promedio se desvíe de la esperanza de la variable aleatoria que representa el experimento aleatorio. Ahora es tiempo de subir el nivel de abstracción y obtener información para la distribución empírica que definimos a continuación.

Consideremos por ahora un experimento aleatorio con un número finito de resultados. Si se registra el número de ocurrencias de un resultado determinado y lo dividimos entre el número de realizaciones se obtiene la proporción de veces que se obtuvo dicho resultado. Ahora la Ley de los Grandes Números nos dice que conforme más experimentos se realicen dicha proporción se acercará cada vez más a la probabilidad de que si se realiza una vez el experimento se obtenga ese resultado. Lo que a nosotros nos interesa es obtener información acerca del decaimiento de la probabilidad de que las distribuciones empíricas se encuentren lejos de la distribución en común de las variables aleatorias. Dicha información nos la proporciona el teorema de Sanov que, por el momento, analizaremos en su versión más simple que

es cuando el experimento tiene un número finito de resultados. Dicho de otra forma, consideremos a un conjunto $\Gamma = \{1, ..., r\} \subset \mathbb{N}$ y $\{X_i\}_{i\in\mathbb{N}}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas en el espacio $(\Omega, \mathcal{B}, \mathbb{P})$, $X_i : \Omega \to \Gamma$ y con densidad de probabilidad común $\rho = (\rho_1, ..., \rho_r)$, i.e., $\mathbb{P}(X_i = s) = \rho_s$ con $s \in \Gamma$.

Definición 1.3.1. *Consideremos la siguiente función* $L_n : \Gamma^n \times \Gamma \to \mathbb{R}^+ \cup \{0\}$

$$L_n(x, \cdot) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(\cdot) \quad x = (x_1, ..., x_n) \in \Gamma^n.$$

En donde $\delta_{x_i}(s) = \begin{cases} 1 & si \quad x_i = s \\ 0 & si \quad x_i \neq s. \end{cases}$

Se define a la medida empírica o distribución empírica del vector aleatorio $X^n = (X_1, ..., X_n)$ *como* $L_n(X^n, \cdot)$. *Es decir, a la composición en la primera entrada de* $L_n(\cdot, \cdot)$ *con el vector aleatorio* X^n.

En algunas ocasiones haremos uso de la siguiente notación: $L_n(X^n, \cdot) = L_n(\cdot)$.

Observación. $L_n(x, s)$ *con* $x \in \mathbb{R}^n$ *registra la proporción de veces que resultó* $s \in \Gamma$ *en la muestra* $x = (x_1, ..., x_n)$ *y no es una cantidad aleatoria.*

Consideremos ahora el conjunto de todas las medidas de probabilidad sobre $(\Gamma, \mathcal{B}(\Gamma))$. Como Γ es finito (con r elementos) entonces dicho conjunto puede ser identificado con el siguiente conjunto (cf. Lema 4.1.2 del apéndice):

$$\mathcal{M}_1(\Gamma) = \left\{ \nu = (\nu_1, ..., \nu_r) \in [0, 1]^r \,\middle|\, \sum_{s=1}^{r} \nu_s = 1 \right\}.$$

A este conjunto también se le suele identificar con el r-simplejo en $\mathbb{R}^r$. En este espacio definimos la distancia de variación total entre dos medidas de probabilidad $d : \mathcal{M}_1(\Gamma) \times \mathcal{M}_1(\Gamma) \to \mathbb{R}^+ \cup \{0\}$ y que además tiene la siguiente identificación (cf. Lema 4.4.4 del apéndice):

$$d(\mu, \nu) = \sup_{A\subset\Gamma} \mid \mu(A) - \nu(A) \mid \; = \; \frac{1}{2}\sum_{s=1}^{r} \mid \mu_s - \nu_s \mid .$$

Se tiene que d es una métrica en $\mathcal{M}_1(\Gamma)$, de hecho también es cierto que $(\mathcal{M}_1(\Gamma), d)$ es un espacio métrico completo (cf. Teorema 4.4.6 del apéndice). Más aún, observamos que para cualesquiera dos medidas $\nu, \mu \in \mathcal{M}_1(\Gamma)$ $d(\nu, \mu) \leq 1$ (pues $\frac{1}{2}\sum_{s=1}^{r} |\mu_s - \nu_s| \leq \frac{1}{2}\sum_{s=1}^{r} |\mu_s| + \sum_{s=1}^{r} |\nu_s| = \frac{1}{2}(1+1) = 1$); además como $\mathbb{Q}_1 = \{q = (q_1, ..., q_r) \in [0, 1]^r | \quad q_s \in \mathbb{Q} \quad \sum_{s=1}^{r} q_s = 1\}$ es un subconjunto denso y numerable de $\mathcal{M}_1(\Gamma)$ entonces $(\mathcal{M}_1(\Gamma), d)$ es un espacio Polaco[5].

[5]Espacio métrico completo y separable.

Ahora, notamos que $L_n(X^n, \cdot)$, la distribución empírica del vector aleatorio $X^n = (X_1, ..., X_n)$, es un elemento aleatorio de $\mathcal{M}_1(\Gamma)$ pues δ_{X_i} es una función que depende de la variable aleatoria X_i $i = 1, ..., n$.

El teorema de Glivenko-Cantelli (cf. Teorema 4.1.2 del apéndice) nos dice que $d(L_n, \rho)$ tiende a cero casi seguramente respecto a la medida de probabilidad $\mathbb{P}$ cuando n tiende a infinito.

La pregunta natural es ¿qué sucede con las grandes desviaciones? Es decir ¿qué sucede con la velocidad con la cual L_n se desvía de ρ? El siguiente teorema contesta a esta pregunta.

Teorema 1.3.1. *(Sanov) Sean $\{X_i\}_{i\in\mathbb{R}}$ variables aleatorias independientes e idénticamente distribuidas tales que X_i toma valores en $\Gamma = \{1, ..., r\} \subset \mathbb{N}$ y con densidad de probabilidad común $\rho = (\rho_1, ..., \rho_r)$, $\rho_s > 0$ para toda $s \in \Gamma$. Sea $a > 0$ y definimos $B_a(\rho) = \{\nu \in \mathcal{M}_1(\Gamma) | \ \ d(\nu, \rho) \leq a\}$ y $\mathcal{I}_\rho(\nu) = \sum_{s=1}^r \nu_s \log(\frac{\nu_s}{\rho_s})$. Entonces se tiene que*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(L_n \in B_a^c(\rho)) = - \inf_{\nu\in B_a^c(\rho)} \mathcal{I}_\rho(\nu). \tag{1.8}$$

Demostración. Consideremos el siguiente conjunto

$$K_n = \left\{ k = (k_1, ..., k_r) \in \mathbb{N}^r \Big| \ \sum_{s=1}^r k_s = n \right\}.$$

Convendremos que $0 \in \mathbb{N}$.

Antes de comenzar la demostración del teorema procederemos a enunciar una serie de lemas técnicos que nos serán de gran utilidad. Sus correspondientes pruebas se realizarán en el apéndice.

Lema 1.3.1. *Sea $k \in K_n$ entonces $\sum_{s=1}^r \left(\frac{-O(\log k_s)}{n}\right) + \frac{O(\log n)}{n} = O\left(\frac{\log n}{n}\right)$.*

Prueba. cf. Lema 4.1.3 del apéndice. □

Utilizaremos la siguiente notación: Si A es un conjunto entonces $|A|$ denota el número de elementos del conjunto A.

Lema 1.3.2. $|K_n| = \binom{n+r-1}{r-1}$ *y* $\binom{n+r-1}{r-1} = O(n^{r-1})$.

Prueba. cf. Lema 4.1.4 del apéndice. □

Lema 1.3.3. *Sean $r > 1$ y $M \in \mathbb{R}$ fijo entonces $\frac{1}{n} \log M + \frac{r-1}{n} \log n = O(\frac{\log n}{n})$.*

Prueba. cf. Lema 4.1.6 del apéndice. □

Lema 1.3.4. $O(\frac{\log n}{n}) + O(\frac{\log n}{n}) = O(\frac{\log n}{n})$.

Prueba. cf. Lema 4.1.5 del apéndice aplicado a $f(n) = \frac{\log n}{n}$. □

Ahora procedemos con la demostración del teorema. Observemos lo siguiente:

(i) $\frac{1}{n}K_n \subset \mathcal{M}_1(\Gamma)$, $\forall n \in \mathbb{N}$.

(ii) $\sum_{i=1}^n \delta_{X_i}(s)$ cuenta las veces en las cuales el experimento tuvo como resultado s en n realizaciones; dicho de otra forma, $\sum_{i=1}^n \delta_{X_i}(s)$ cuenta las veces que la variable aleatoria pertenece a la categoría $s \in \Gamma$ en n realizaciones; i.e. $\sum_{i=1}^n \delta_{X_i}(s)$ sigue una distribución multinomial con parámetros n y $\rho_1, ..., \rho_r$.

Por (ii) tenemos que:

$$\mathbb{P}\bigg(L_n(s) = \frac{k_s}{n} \quad \forall s \in \Gamma\bigg) = \mathbb{P}\left(\sum_{i=1}^n \delta_{X_i}(s) = k_s \quad \forall s \in \Gamma\right)$$

$$= \mathbb{P}\left(\sum_{i=1}^n \delta_{X_i}(1) = k_1, ..., \sum_{i=1}^n \delta_{X_i}(r) = k_r\right) = n! \prod_{s=1}^r \frac{\rho_s^{k_s}}{k_s!} \quad k = (k_1, ..., k_r) \in K_n.$$

Para $k \in K_n$ sea $\nu_n(k) = \frac{1}{n}k \in \mathcal{M}_1(\Gamma)$, y definimos

$$Q_n(a) = \max_{k \in K_n | \nu_n(k) \in B_a^c(\rho)} n! \prod_{s=1}^r \frac{\rho_s^{k_s}}{k_s!}.$$

Si A es un conjunto finito y $A \subset \mathbb{R}$ entonces

$$\max A \leq \sum_{a \in A} a \leq |A| \max A.$$

Además tenemos que

$$\mathbb{P}(L_n \in B_a^c(\rho)) = \mathbb{P}\bigg(\bigcup \Big\{ L_n(s) = \frac{k_s}{n} \quad \forall s \in \Gamma \Big| \quad k \in K_n \quad \nu_n(k) \in B_a^c(\rho) \Big\}\bigg)$$

$$= \sum_{k \in K_n | \nu_n(k) \in B_a^c(\rho)} \mathbb{P}\bigg(L_n(s) = \frac{k_s}{n} \quad \forall s \in \Gamma\bigg),$$

pues los eventos

$$\left\{ L_n(s) = \frac{k_s}{n} \quad \forall s \in \Gamma \Big| \quad k \in K_n \quad \nu_n(k) \in B_a^c(\rho) \right\}$$

son ajenos dos a dos. De lo anterior se sigue que:

$$Q_n(a) \quad \leq \quad \mathbb{P}(L_n \in B_a^c(\rho)) \quad \leq \quad |K_n| Q_n(a). \tag{1.9}$$

Ahora recordemos la fórmula de Stirling: $n! \approx \sqrt{2\pi n}(\frac{n}{e})^n$ lo que implica que $\log n! = n \log n - n + O(\log n)$. Tenemos que:

$$\mathbb{P}\bigg(L_n = \frac{k_s}{n} \quad \forall s \in \Gamma\bigg)$$

$$= \frac{1}{n} \log \left(n! \prod_{s=1}^{r} \frac{\rho_s^{k_s}}{k_s!} \right) = \frac{1}{n} \left(\log n! + \sum_{s=1}^{r} k_s \log \rho_s - \log k_s! \right)$$

$$= \frac{1}{n} \bigg[n \log n - n + O(\log n) + \sum_{s=1}^{r} (k_s \log \rho_s - k_s \log k_s + k_s - O(\log k_s)) \bigg]$$

$$= \log n - 1 + \frac{O(\log n)}{n} + \sum_{s=1}^{r} \bigg(\frac{k_s}{n} \log \rho_s - \frac{k_s}{n} \log k_s + \frac{k_s}{n} - \frac{O(\log k_s)}{n} \bigg)$$

$$= \sum_{s=1}^{r} \bigg(\frac{k_s}{n} \log n \bigg) + \sum_{s=1}^{r} \bigg(\frac{k_s}{n} \log \rho_s - \frac{k_s}{n} \log k_s + \frac{k_s}{n} - \frac{O(\log k_s)}{n} \bigg) + \frac{O(\log n)}{n} - 1$$

$$= \sum_{s=1}^{r} \bigg(\frac{k_s}{n} (\log \rho_s + \log n - \log k_s) \bigg) + \sum_{s=1}^{r} \bigg(\frac{k_s}{n} \bigg) + \sum_{s=1}^{r} \bigg(\frac{-O(\log k_s)}{n} \bigg) + \frac{O(\log n)}{n} - 1$$

$$= \sum_{s=1}^{r} \left[\frac{k_s}{n} \left(\log \rho_s - \log \frac{k_s}{n} \right) \right] + \sum_{s=1}^{r} \left(\frac{-O(\log k_s)}{n} \right) + \frac{O(\log n)}{n}$$

$$= \sum_{s=1}^{r} \left[\frac{k_s}{n} \left(\log \rho_s - \log \frac{k_s}{n} \right) \right] + O\left(\frac{\log n}{n} \right).$$

En donde en la tercera igualdad aplicamos la fórmula de Stirling. En la penúltima igualdad hacemos uso del hecho de que $\sum_{s=1}^{r} k_s = n$. Y la última igualdad utilizamos el Lema 1.3.1.

Por otro lado observamos que

$$-\mathcal{I}_\rho(\nu_n(k)) = -\sum_{s=1}^{r} \nu_{n_s}(k) \log \left(\frac{\nu_{n_s}(k)}{\rho_s} \right) = \sum_{s=1}^{r} \frac{k_s}{n} \left(\log \rho_s - \log \frac{k_s}{n} \right).$$

Es decir,

$$\frac{1}{n}\log\left(n!\prod_{s=1}^{r}\frac{\rho_s^{k_s}}{k_s!}\right) = -\mathcal{I}_\rho(\nu_n(k)) + O\left(\frac{\log n}{n}\right).$$

Por lo tanto tenemos

$$\frac{1}{n}\log Q_n(a) = \max_{k\in K_n|\nu_n(k)\in B_a^c(\rho)} -\mathcal{I}_\rho(\nu_n(k)) + O\left(\frac{\log n}{n}\right).$$

Además observemos que los elementos en K_n son soluciones naturales a la ecuación $\sum_{s=1}^{r} k_s = n$. Por el Lema 1.3.2, $|K_n| = O(n^{r-1})$. Ahora, de (1.9) tenemos que

$$\frac{1}{n}\log Q_n(a) \leq \frac{1}{n}\log \mathbb{P}(L_n \in B_a^c(\rho)) \leq \frac{1}{n}(\log|K_n| + \log Q_n(a)).$$

Se sigue entonces que $\frac{1}{n}\log\mathbb{P}(L_n \in B_a^c(\rho)) = \frac{1}{n}\log Q_n(a) + g(n)$ con $g(n) > 0$ y $g(n) \leq \frac{1}{n}\log|K_n|$. Ahora, por el Lema 1.3.2 se tiene que

$$\frac{1}{n}\log|K_n| = \frac{1}{n}\log O(n^{r-1})$$

y además

$$\frac{1}{n}\log O(n^{r-1}) = O\Big(\frac{\log n}{n}\Big)$$

pues

$$\left|\frac{1}{n}\log O(n^{r-1})\right| = \frac{1}{n}\log O(n^{r-1}) \leq \frac{1}{n}\log(Mn^{r-1}) = \frac{1}{n}(\log M + \log(n^{r-1}))$$

$$= \frac{1}{n}\log M + \frac{r-1}{n}\log n = O\Big(\frac{\log n}{n}\Big) + O\Big(\frac{\log n}{n}\Big) = O\Big(\frac{\log n}{n}\Big).$$

En donde la primera igualdad se sigue de que $1 \leq |K_n| = \frac{1}{n}\log O(n^{r-1})$, la penúltima y última igualdad de los lemas 1.3.3 y 1.3.4 respectivamente. Así $|g(n)| = g(n) \leq O(\frac{\log n}{n})$ y entonces $g(n) = O(\frac{\log n}{n})$ y obtenemos que

$$\frac{1}{n}\log\mathbb{P}(L_n \in B_a^c(\rho)) = O\left(\frac{\log n}{n}\right) + O\left(\frac{\log n}{n}\right) - \min_{k\in K_n|\nu_n(k)\in B_a^c(\rho)} \mathcal{I}_\rho(\nu_n(k)).$$

Por el Lema 1.3.4 se tiene que

$$\frac{1}{n}\log\mathbb{P}(L_n \in B_a^c(\rho)) = O\left(\frac{\log n}{n}\right) - \min_{k\in K_n|\nu_n(k)\in B_a^c(\rho)} \mathcal{I}_\rho(\nu_n(k)) \qquad (1.10)$$

Para concluir la prueba basta observar dos cosas:

(i) $\bigcup_{n\in\mathbb{N}}\{\nu_n(k)|k\in K_n\}$ es denso en $\mathcal{M}_1(\Gamma)$.
(ii) La función $\nu\to\mathcal{I}_\rho(\nu)$ es continua en $\mathcal{M}_1(\Gamma)$.

Lo primero se sigue del hecho de que nuestro espacio $\mathcal{M}_1(\Gamma)$ es de dimensión finita y de la densidad de $\mathbb{Q}$ en $\mathbb{R}$; lo segundo se sigue ya que $\mathcal{I}_\rho$ es una suma y composición de funciones continuas de $\mathcal{M}_1(\Gamma)$ en $\mathbb{R}$. Ahora (i) y (ii) nos garantizan que para toda $\nu\in\mathcal{M}_1(\Gamma)$ existe una sucesión $(k_n)_{n\in\mathbb{N}}$ con $k_n\in K_n$ para toda $n\in\mathbb{N}$ tal que

$$\lim_{n\to\infty} d(\nu_n(k_n),\nu)=0 \quad \text{y} \quad \lim_{n\to\infty}\mathcal{I}_\rho(\nu_n(k_n))=\mathcal{I}_\rho(\nu).$$

En particular para $\nu\in B_a^c(\rho)\subset\mathcal{M}_1(\Gamma)$ existe dicha sucesión $(k_n)_{n\in\mathbb{N}}$ y además se puede construir tal que $\nu_n(k_n)\in B_a^c(\rho)$ $\forall n\in\mathbb{N}$. Entonces

$$\min_{k\in K_n|\nu_n(k)\in B_a^c(\rho)}\mathcal{I}_\rho(\nu_n(k))\le\mathcal{I}_\rho(\nu_n(k_n)).$$

Observamos que en el lado derecho de la desigualdad cuando tomamos el mínimo obtenemos un número que no depende de la elección inicial de ν ya que estamos corriendo sobre todos los $k\in K_n$ tales que $\nu_n(k_n)\in B_a^c(\rho)$ y por ende tampoco depende de la sucesión $(k_n)_{n\in\mathbb{N}}$, luego

$$\limsup_{n\to\infty}\min_{k\in K_n|\nu_n(k_n)\in B_a^c(\rho)}\mathcal{I}_\rho(\nu_n(k))\le\mathcal{I}_\rho(\nu)\quad\forall\nu\in B_a^c(\rho).$$

En particular

$$\limsup_{n\to\infty}\min_{k\in K_n|\nu_n(k_n)\in B_a^c(\rho)}\mathcal{I}_\rho(\nu_n(k))\le\inf_{\nu\in B_a^c(\rho)}\mathcal{I}_\rho(\nu),$$

y es claro que

$$\inf_{\nu\in B_a^c(\rho)}\mathcal{I}_\rho(\nu)\le\liminf_{n\to\infty}\min_{k\in K_n|\nu_n(k)\in B_a^c(\rho)}\mathcal{I}_\rho(\nu_n(k))$$

ya que

$$\{\nu_n(k)\in B_a^c(\rho)\quad k\in K_n\}\subset B_a^c(\rho).$$

Por lo que tomando límite en (1.10) obtenemos

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}(L_n\in B_a^c(\rho))=\lim_{n\to\infty}O\Big(\frac{\log n}{n}\Big)-\min_{k\in K_n|\nu_n(k)\in B_a^c(\rho)}\mathcal{I}_\rho(\nu_n(k))$$

$$=-\inf_{\nu\in B_a^c(\rho)}\mathcal{I}_\rho(\nu).$$

□

Antes de extender el Teorema 1.3.1 al caso en que $|\Gamma| = \aleph_0$ necesitamos hacer unas observaciones. En primer lugar es importante hacer notar que la elección de la métrica d es flexible siempre y cuando se cumplan las propiedades (i) y (ii) que se presentan en la demostración del teorema. En segundo lugar tenemos que el teorema también es válido reemplazando a B_a^c por conjuntos más generales, sin embargo, esto no es nuestro cometido por el momento: el caso general (Γ y los conjuntos generales) se demostrará en el capítulo 3. A continuación discutimos la generalización del teorema de Sanov a conjuntos infinitos numerables.

1.4. Teorema de Sanov (Versión infinito numerable)

Consideremos una medida de probabilidad sobre $\mathbb{N}$

$$\rho = (\rho_s)_{s\in\mathbb{N}} \quad \rho_s > 0 \quad \forall s \in \mathbb{N}. \tag{1.11}$$

Sean $\mathcal{M}_1(\mathbb{N}) = \{\nu | \nu \quad \text{es medida de probabilidad sobre} \quad (\mathbb{N}, \mathcal{B})\}$ y $\{X_i\}_{i\in\mathbb{N}}$ variables aleatorias independientes e idénticamente distribuidas con distribución común ρ. Sea L_n la medida empírica generada por la sucesión, que es un elemento aleatorio de $(\mathcal{M}_1(\mathbb{N}), d)$ con la métrica $d(\mu,\nu) = \frac{1}{2}\sum_{s\in\mathbb{N}} |\mu_s - \nu_s|$. De nuevo observamos que la desigualdad del triángulo implica que para cualesquiera dos medidas $\nu, \mu \in \mathcal{M}_1(\mathbb{N})$ se tiene que $d(\nu,\mu) \leq 1$).

Observación. *Utilizaremos la siguiente notación: Sea f una función real de variable real. Entonces $f(a^-)$ denota el límite por la izquierda de $f(a)$, i.e., $f(a^-) = \lim_{\epsilon\to 0} f(a+\epsilon)$ con $\epsilon < 0$.*

Observación. *En el Teorema 1.3.1 utilizamos la función $\mathcal{I}_\rho$ definida para el caso finito. En el Teorema 1.4.1 utilizaremos la misma función extendida al caso infinito numerable; esta función es conocida como entropía relativa y se definirá de manera general más adelante, cf. Definición 1.5.1.*

Teorema 1.4.1. *Sean $a > 0$, $B_a(\rho) = \{\nu \in \mathcal{M}_1(\mathbb{N}) | d(\nu,\rho) \leqslant a\}$ ρ como en (1.11) y $\nu = (\nu_s)_{s\in\mathbb{N}} \in \mathcal{M}_1(\mathbb{N})$ una probabilidad sobre $\mathbb{N}$. Definimos*

$$I_\rho(\nu) = \sum_{s\in\mathbb{N}} \nu_s \log\Big(\frac{\nu_s}{\rho_s}\Big) \quad y \quad J(a) = \inf_{\nu\in B_a^c(\rho)} \mathcal{I}_\rho(\nu). \tag{1.12}$$

Entonces se verifican las siguientes cotas (inferior y superior respectivamente) para L_n definida en la Definición 1.3.1.

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(L_n \in B_a^c(\rho)) \geqslant -J(a) \quad \forall a > 0. \tag{1.13}$$

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(L_n \in B_a^c(\rho)) \leqslant -J(a^-) \quad \forall a > 0. \tag{1.14}$$

Demostración. Definimos $\pi_N : \mathbb{N} \to \{1, ..., N\}$ como $\pi_N(s) = \text{mín}\{s, N\}$ la identidad truncada. Asimismo denotamos $\pi_N\nu = \nu \circ \pi_N^{-1}$, en donde

$$\nu \circ \pi_N^{-1}(i) = \begin{cases} \nu_i & si \quad i < N \\ \sum_{s \geqslant N} \nu_s & si \quad i = N. \end{cases}$$

Observamos que $\pi_N\nu$ es una medida de probabilidad sobre el conjunto $\{1, ...N\}$, lo que estamos haciendo es relativamente intuitivo: pensando en la medida de probabilidad como una eneada (una sucesión en el caso infinito numerable) truncamos hasta la N-1 entrada y concentramos el resto de la masa de probabilidad en la N-ésima entrada, lo que nos da como resultado "casi" la misma medida (difiere en la entrada N) pero sobre un conjunto finito. Todo lo anterior nos será de mucha utilidad ya que nos permitirá utilizar el Teorema (1.3.1) con la medida de probabilidad truncada. Para ello necesitaremos unas propiedades enunciadas en el siguiente lema.

Lema 1.4.1. *Sea $\rho \in \mathcal{M}_1(\mathbb{N})$ como en (1.11). Entonces*

(a) $0 \leqslant d(\nu, \rho) - d(\pi_N\nu, \pi_N\rho) \leqslant \sum_{s \geqslant N} \rho_s$ para cada $\nu \in \mathcal{M}_1(\mathbb{N})$.

(b) La función $a \mapsto J(a)$ es no decreciente y continua por la derecha, J como en (1.12).

(c) $\mathcal{I}_\rho(\nu)$ definida en (1.12) converge en $[0, \infty]$ (puede tomar el valor infinito). La función $N \mapsto \mathcal{I}_{\pi_N\rho}(\pi_N\nu)$ es no decreciente y tiene límite $\mathcal{I}_\rho(\nu)$ cuando N tiende a infinito, para toda $\nu \in \mathcal{M}_1(\mathbb{N})$.

Prueba. (a) Como $\pi_N\nu = \nu \circ \pi_N^{-1}$ entonces

$$d(\pi_N\nu, \pi_N\rho) = \frac{1}{2}\left(\sum_{s=1}^{N-1} |\nu_s - \rho_s| + |\sum_{s \geq N} \nu_s - \rho_s|\right).$$

Así

$$d(\nu, \rho) - d(\pi_N\nu, \pi_N\rho) = \frac{1}{2}\sum_{s=1}^{\infty} |\nu_s - \rho_s| - \frac{1}{2}\left(\sum_{s=1}^{N-1} |\nu_s - \rho_s| + \left|\sum_{s \geqslant N} \nu_s - \rho_s\right|\right)$$

$$= \frac{1}{2}\left(\sum_{s \geqslant N} |\nu_s - \rho_s| - \left|\sum_{s \geqslant N} \nu_s - \rho_s\right|\right) \geq 0 \quad \text{pues} \quad \sum_{s \geqslant N} |\nu_s - \rho_s| \geq \left|\sum_{s \geqslant N} \nu_s - \rho_s\right|.$$

donde la última desigualdad se da por la desigualdad del triángulo.

Con esto tenemos la primera afirmación en (a). Para mostrar la segunda afirmación en (a) definimos $W_n = \{s \geq N | \nu_s \geq \rho_s\}$, $A_n = \frac{1}{2}\sum_{s \in W_n}(\nu_s - \rho_s)$ y $B_n = -\frac{1}{2}\sum_{s \notin W_n, s \geq N}(\nu_s - \rho_s)$. Así, tenemos que

$$A_n + B_n - |A_n - B_n| = \frac{1}{2}\sum_{s\in W_n}(\nu_s - \rho_s) - \frac{1}{2}\sum_{s\notin W_n, s\geq N}(\nu_s - \rho_s)$$

$$-\left|\frac{1}{2}\sum_{s\in W_n}(\nu_s - \rho_s) + \frac{1}{2}\sum_{s\notin W_n, s\geq N}(\nu_s - \rho_s)\right|$$

$$= \frac{1}{2}\left(\sum_{s\geq N}|\nu_s - \rho_s| - |\sum_{s\geq N}(\nu_s - \rho_s)|\right) = d(\nu,\rho) - d(\pi_N\nu, \pi_N\rho).$$

Ahora bien, $A_n + B_n - |A_n - B_n| \leq 2B_n$ ya que si $A_n - B_n \geq 0$ entonces $A_n + B_n - |A_n - B_n| = 2B_n$ y si $A_n - B_n \leq 0$ entonces $A_n + B_n - |A_n - B_n|$ $= 2A_n \leq 2B_n$ pues $A_n \leq B_n$. Así

$$0 \leq d(\nu,\rho) - d(\pi_N\nu, \pi_N\rho) \leq 2B_n = \sum_{s\notin W_n, s\geq N}(\rho_s - \nu_s) \leq \sum_{s\geq N}(\rho_s - \nu_s) \leq \sum_{s\geq N}\rho_s.$$

(b) J es no decreciente pues si $b < d$ entonces $B_d^c(\rho) \subset B_b^c(\rho)$ y así $J(d)$ $= \text{ínf}_{\nu\in B_d^c(\rho)}\mathcal{I}_\rho(\nu) \leq \text{ínf}_{\nu\in B_b^c(\rho)}\mathcal{I}_\rho(\nu) = J(b)$. Para probar la continuidad por la derecha tenemos por la definición de J que para toda $\epsilon > 0$ existe $\nu_\epsilon \in B_a^c(\rho)$ tal que $\mathcal{I}_\rho(\nu_\epsilon) \leq J(a) + \epsilon$. Ahora, para toda $\delta > 0$ suficientemete pequeño (basta tomar $\delta < d(\nu_\epsilon, \rho) - a$) tenemos que $\nu_\epsilon \in B_{a+\delta}^c(\rho)$ por lo que $J(a+\delta) \leq \mathcal{I}_\rho(\nu_\epsilon) \leq J(a) + \epsilon$ para toda $\epsilon > 0$. Por lo tanto

$$\lim_{\delta\to 0} J(a+\delta) \leq J(a).$$

Como J es no decreciente tenemos $J(a) \leq J(a+\delta)$ para toda $\delta > 0$. Así

$$\lim_{\delta\to 0} J(a+\delta) \geq J(a),$$

por lo tanto J es continua por la derecha.

(c) Definimos

$$h(x) = x\log(x), \tag{1.15}$$

de esta forma observemos que $\mathcal{I}_\rho(\nu) = \sum_{s\in\mathbb{N}}\rho_s h(\frac{\nu_s}{\rho_s})$. Consideraremos varios casos:

Caso 1. Si $\frac{\nu_s}{\rho_s} \in [0,1)$ para toda $s \in \mathbb{N}$.

En este caso h está acotada inferiormente por $-\frac{1}{e}$ entonces $|h(x)| < \frac{1}{e}$ en $[0,1)$. Tenemos entonces que $\sum_{s\in\mathbb{N}}|\rho_s h(x)| \leq \sum_{s\in\mathbb{N}}\rho_s\frac{1}{e} = \frac{1}{e}$ en $[0,1)$, i.e., $\mathcal{I}_\rho(\nu)$ converge absolutamente.

Caso 2. Si $\frac{\nu_s}{\rho_s} \in [1, \infty)$ para toda $s \in \mathbb{N}$.

En este caso tenemos que $h(\frac{\nu_s}{\rho_s}) > 0$ para toda $s \in \mathbb{N}$; así, si definimos $S_n = \sum_{i=1}^{n} \left|\rho_i h(\frac{\nu_i}{\rho_i})\right| = \sum_{i=1}^{n} \rho_i h(\frac{\nu_i}{\rho_i}) \geq 0$ entonces $(S_n)_{n\in\mathbb{N}}$ es no decreciente por lo que $\sum_{s=1}^{\infty} \left|\rho_s h(\frac{\nu_s}{\rho_s})\right|$ converge (posiblemente a ∞), por lo tanto $\mathcal{I}_\rho(\nu)$ converge absolutamente.

Caso 3. Si $\frac{\nu_s}{\rho_s} \in [0, \infty)$ para toda $s \in \mathbb{N}$.

Sea $A = \{s \in \mathbb{N} | \frac{\nu_s}{\rho_s} \in [0, 1)\}$. Consideraremos dos subcasos.

Subcaso 3.1. $\sum_{s\notin A} \left|\rho_s h(\frac{\nu_s}{\rho_s})\right| = \infty$.

$\sum_{s\in\mathbb{N}} \left|\rho_s h(\frac{\nu_s}{\rho_s})\right| = \sum_{s\in A} \left|\rho_s h(\frac{\nu_s}{\rho_s})\right| + \sum_{s\notin A} \left|\rho_s h(\frac{\nu_s}{\rho_s})\right| = \infty$. Por lo tanto $\mathcal{I}_\rho(\nu)$ converge absolutamente a ∞.

Subcaso 3.2. $\sum_{s\notin A} \left|\rho_s h(\frac{\nu_s}{\rho_s})\right| = \alpha$ para algún $\alpha \in \mathbb{R}$.

$\sum_{s\in\mathbb{N}} \left|\rho_s h(\frac{\nu_s}{\rho_s})\right| = \sum_{s\in A} \left|\rho_s h(\frac{\nu_s}{\rho_s})\right| + \sum_{s\notin A} \left|\rho_s h(\frac{\nu_s}{\rho_s})\right| \leq \frac{1}{e} + \alpha$. Por lo tanto $\mathcal{I}_\rho(\nu)$ converge absolutamente.

Concluimos entonces que $\mathcal{I}_\rho(\nu)$ converge absolutamente.

Ahora veremos que la serie converge a algún valor en $[0, \infty]$ para lo cual es suficiente mostrar que es no negativa: definimos una variable aleatoria Z que toma valores en $\mathbb{N}$ con densidad ν y definimos $Y = \frac{\rho(Z)}{\nu(Z)}$. Sea $\phi(x) = -\log(x)$ (que es convexa). Gracias a la desigualdad de Jensen (cf. Teorema 4.4.1 el apéndice) se tiene que

$$0 = \phi(\mathbb{E}(Y)) \leq \mathbb{E}(\phi(Y)) = \sum_{s\in\mathbb{N}} \nu_s \left[-\log(\frac{\rho_s}{\nu_s})\right] = \mathcal{I}_\rho(\nu),$$

por lo que $\mathcal{I}_\rho(\nu) \geq 0$.

Por último veamos que la función $N \mapsto \mathcal{I}_{\pi_N\rho}(\pi_N\nu)$ es no decreciente. Para ello sean

$$\nu_{[N]} = \sum_{s>N} \nu_s \quad \text{y} \quad \rho_{[N]} = \sum_{s>N} \rho_s. \tag{1.16}$$

Sea

$$\Delta = \mathcal{I}_{\pi_{N+1}\rho}(\pi_{N+1}\nu) - \mathcal{I}_{\pi_N\rho}(\pi_N\nu)$$

un incremento. Debemos mostrar que $\Delta \geq 0$. Primero observemos que

$$\Delta = \sum_{i=1}^{N} \nu_i \log\Big(\frac{\nu_i}{\rho_i}\Big) + \left[\sum_{s>N} \nu_s\right] \log\Big(\frac{\sum_{s>N} \nu_s}{\sum_{s>N} \rho_s}\Big) - \sum_{i=1}^{N-1} \nu_i \log\Big(\frac{\nu_i}{\rho_i}\Big)$$

$$-\left[\sum_{s\geq N}\nu_s\right]\log\Big(\frac{\sum_{s\geq N}\nu_s}{\sum_{s\geq N}\rho_s}\Big)=\nu_N\log\Big(\frac{\nu_N}{\rho_N}\Big)+\nu_{[N]}\log\Big(\frac{\nu_{[N]}}{\rho_{[N]}}\Big)$$

$$-(\nu_N+\nu_{[N]})\log\Big(\frac{\nu_N+\nu_{[N]}}{\rho_N+\rho_{[N]}}\Big)$$

$$=\rho_N h\Big(\frac{\nu_N}{\rho_N}\Big)+\rho_{[N]}h\Big(\frac{\nu_{[N]}}{\rho_{[N]}}\Big)-(\rho_N+\rho_{[N]})h\Big(\frac{\nu_N+\nu_{[N]}}{\rho_N+\rho_{[N]}}\Big).$$

La primera igualdad es la definición de Δ, la segunda introduce la nueva notación y la tercera también ocupa la notación h(x) establecida en (1.15). Ahora, como h es convexa, se sigue que

$$\rho_N h\Big(\frac{\nu_N}{\rho_N}\Big)+\rho_{[N]}h\Big(\frac{\nu_{[N]}}{\rho_{[N]}}\Big)-(\rho_N+\rho_{[N]})h\Big(\frac{\nu_N+\nu_{[N]}}{\rho_N+\rho_{[N]}}\Big)$$

$$\geq h\bigg(\rho_N\Big(\frac{\nu_N}{\rho_N}\Big)+\rho_{[N]}\Big(\frac{\nu_{[N]}}{\rho_{[N]}}\Big)-(\rho_N+\rho_{[N]})\Big(\frac{\nu_N+\nu_{[N]}}{\rho_N+\rho_{[N]}}\Big)\bigg)$$

$$=h(\nu_N+\nu_{[N]}-(\nu_N+\nu_{[N]}))=0\log 0=0.$$

Así, la función es no decreciente.

Por otro lado

$$\lim_{N\to\infty}\mathcal{I}_{\pi_N\rho}(\pi_N\nu)=\lim_{N\to\infty}\left[\sum_{s=1}^{N-1}\bigg(\nu_s\log\Big(\frac{\nu_s}{\rho_s}\Big)\bigg)+\bigg(\sum_{s=N}^{\infty}\nu_s\bigg)\log\bigg(\frac{\sum_{s=N}^{\infty}\nu_s}{\sum_{s=N}^{\infty}\rho_s}\bigg)\right].$$

Entonces para ver que la función $N\mapsto\mathcal{I}_{\pi_N\rho}(\pi_N\nu)$ tiene límite $\mathcal{I}_\rho(\nu)$ cuando N tiende a infinito para toda $\nu\in\mathcal{M}_1(\mathbb{N})$ es suficiente mostrar que

$$\lim_{N\to\infty}\bigg(\sum_{s=N}^{\infty}\nu_s\bigg)\log\bigg(\frac{\sum_{s=N}^{\infty}\nu_s}{\sum_{s=N}^{\infty}\rho_s}\bigg)=0$$

pues

$$\lim_{N\to\infty}\sum_{s=1}^{N-1}\bigg(\nu_s\log\Big(\frac{\nu_s}{\rho_s}\Big)\bigg)=\mathcal{I}_\rho(\nu).$$

Efectivamente,

$$\lim_{N\to\infty}\bigg(\sum_{s=N}^{\infty}\nu_s\bigg)\log\bigg(\frac{\sum_{s=N}^{\infty}\nu_s}{\sum_{s=N}^{\infty}\rho_s}\bigg)=\bigg(\lim_{N\to\infty}\sum_{s=N}^{\infty}\nu_s\bigg)\log\bigg(\frac{\lim_{N\to\infty}\sum_{s=N}^{\infty}\nu_s}{\lim_{N\to\infty}\sum_{s=N}^{\infty}\rho_s}\bigg)$$

$$=0\log\Big(\frac{0}{0}\Big)=0.$$

□

Procedemos ahora con la demostración del Teorema (1.4.1).

(1) Cota inferior (1.13):

En primer lugar afirmamos que

$$\{x \in \mathbb{N}^n | d(L_n(x), \rho) > a\} \supset \{x \in \mathbb{N}^n | d(\pi_N L_n(x), \pi_N \rho) > a\}$$

para cada $N > 0$. Esto ya que si $d(\pi_N L_n(x), \pi_N \rho) > a$ entonces por (a) del Lema 1.4.1, con $L_n(x)$ como ν, obtenemos que $d(L_n(x), \rho) > d(\pi_N L_n(x), \pi_N \rho)$ $> a$, es decir, $d(L_n(x), \rho) > a$. Así

$$\mathbb{P}(d(L_n, \rho) > a) \geq \mathbb{P}(d(\pi_N L_n, \pi_N \rho) > a) \quad \forall N > 0. \tag{1.17}$$

Además, observamos que para toda $\mu \in \mathcal{M}_1(\Gamma)$ con $|\Gamma| = N$ se tiene que $\mu = \pi_N \mu'$ para alguna $\mu' \in \mathcal{M}_1(\mathbb{N})$ ya que si definimos μ' de la siguiente manera:

$$\mu'(i) = \begin{cases} \mu_i & si \quad i \leq N, \\ 0 & si \quad s > N, \end{cases}$$

entonces $\mu' \in \mathcal{M}_1(\mathbb{N})$ y además $\mu = \pi_N \mu'$. Por lo tanto,

$$\nu \in \mathcal{M}_1(\mathbb{N}) | d(\pi_N \nu, \pi_N \rho) > a = \{\nu \in \mathcal{M}_1(\Gamma) | d(\nu, \rho) > a\}$$

y como $\pi_N L_n = \frac{1}{n} \sum_{i=1}^n \delta_{\pi_N X_i}$ podemos aplicar el Teorema (1.3.1) y obtener que para toda $N > 0$ se tiene que

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}(d(\pi_N L_n, \pi_N \rho) > a) = - \inf_{\nu \in \mathcal{M}_1(\mathbb{N}) | d(\pi_N \nu, \pi_N \rho) > a} \mathcal{I}_{\pi_N \rho}(\pi_N \nu). \tag{1.18}$$

Por otro lado afirmamos que para toda $\delta > 0$ existe $n_0(\delta)$ tal que

$$\{\nu \in \mathcal{M}_1(\mathbb{N}) | d(\pi_N \nu, \pi_N \rho) > a\} \supset \{\nu \in \mathcal{M}_1(\mathbb{N}) | d(\nu, \rho) > a + \delta\}$$

con $N \geqslant n_0$. Lo cual es sencillo de ver: tomamos $\delta > 0$ y
$\nu \in \{\mu \in \mathcal{M}_1(\mathbb{N}) | d(\mu, \rho) > a + \delta\}$, i.e.,

$$d(\nu, \rho) > a + \delta, \tag{1.19}$$

y supongamos lo contrario, es decir, que

$$d(\pi_N \nu, \pi_N \rho) < a. \tag{1.20}$$

Como $\sum_{s=N}^{\infty} \rho_s$ tiende a cero cuando N tiende a infinito entonces para toda δ positiva existe $n_0(\delta)$ tal que si $N > n_0$ $\sum_{s=N}^{\infty} \rho_s < \delta$. Así, por (a) del Lema 1.4.1 tenemos que

$$d(\nu,\rho) - d(\pi_N\nu, \pi_N\rho) < \delta.$$

Sumando (1.19) y (1.20) obtenemos $d(\nu,\rho) < a + \delta$, lo cual contradice (1.19). Por lo tanto tenemos que $d(\pi_N\nu, \pi_N\rho) > a$, i.e.,

$$\nu \in \{\nu \in \mathcal{M}_1(\mathbb{N}) | d(\pi_N\nu, \pi_N\rho) > a\}.$$

Utilizando el inciso (c) del Lema 1.4.1 se tiene que $\mathcal{I}_{\pi_N\rho}(\pi_N\nu) \leq \mathcal{I}_\rho(\nu)$ para toda $N > 0$, así tenemos que

$$-\inf_{\nu\in\mathcal{M}_1(\mathbb{N})|d(\pi_N\nu,\pi_N\rho)>a} \mathcal{I}_{\pi_N\rho}(\pi_N\nu) \geq -\inf_{\nu\in\mathcal{M}_1(\mathbb{N})|d(\nu,\rho)>a+\delta} \mathcal{I}_\rho(\nu) = J(a+\delta). \tag{1.21}$$

De (1.17), (1.18) y (1.21) se sigue que

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(L_n \in B_a^c(\rho)) \geq -J(a+\delta) \quad \forall \delta > 0.$$

Haciendo δ tender a cero y por (b) del Lema 1.4.1 obtenemos (1.13).

(2) Cota superior:

Sea $\delta > 0$ fijo. Por (a) del Lema 1.4.1 existe $n_0'(\delta)$ tal que

$$\mathbb{P}(d(L_n,\rho) > a) \leq \mathbb{P}(d(\pi_N L_n, \pi_N\rho) > a - \delta) \quad \forall N \geq n_0'. \tag{1.22}$$

Lo anterior ya que si suponemos que $d(L_n,\rho) > a$ y de nuevo como $\sum_{s=N}^{\infty} \rho_s \to 0$ cuando N tiende a infinito entonces para $d(L_n,\rho) - a + \delta > 0$ existe $n_0'(\delta)$ tal que $d(L_n,\rho) - d(\pi_N Ln, \pi_N\rho) < \sum_{s=N}^{\infty} < d(L_n,\rho) - a + \delta$, i.e., $d(\pi_N Ln, \pi_N\rho) > a - \delta$ y así $\{d(L_n,\rho) > a\} \subset \{d(\pi_N L_n, \pi_N\rho) > a - \delta\}$ para toda $N \geq n_0'$.

De nuevo usando el Teorema (1.3.1) obtenemos

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(d(\pi_N L_n, \pi_N\rho) > a - \delta) = -\inf_{\nu\in\mathcal{M}_1(\mathbb{N})|d(\pi_N\nu,\pi_N\rho)>a-\delta} \mathcal{I}_{\pi_N\rho}(\pi_N\nu)$$

para toda $N \geq 1$. Mostraremos que

$$\inf_{\nu\in\mathcal{M}_1(\mathbb{N})|d(\pi_N\nu,\pi_N\rho)>a-\delta} \mathcal{I}_{\pi_N\rho}(\pi_N\nu) \geq J(a-\delta) \quad \forall N \geq 1.$$

Sea $N \geq 1$ fijo. Sea $\nu \in \mathcal{M}_1(\mathbb{N})$ tal que $d(\pi_N\nu, \pi_N\rho) > a - \delta$. Consideremos $\nu_{[N]}$ y $\rho_{[N]}$ cf. (1.16), definimos la siguiente medida de probabilidad $\tilde{\nu} \in \mathcal{M}_1(\mathbb{N})$:

$$\tilde{\nu}(s) = \begin{cases} \nu_s & si \quad s < N \\ \frac{\nu_{[N]}}{\rho_{[N]}}\rho_s & si \quad s \geq N. \end{cases}$$

Efectivamente $\tilde{\nu} \in \mathcal{M}_1(\mathbb{N})$ pues tiene valores no negativos con:

$$\sum_{s\in\mathbb{N}} \tilde{\nu} = \sum_{s=1}^{N-1} \nu_s + \sum_{s=N}^{\infty} \frac{\nu_{[N]}}{\rho_{[N]}}\rho_s = 1 - \sum_{s=N}^{\infty} \nu_s + \frac{\nu_{[N]}}{\rho_{[N]}} \sum_{s=N}^{\infty} \rho_s = 1 - \nu_{[N]} + \frac{\nu_{[N]}}{\rho_{[N]}}\rho_{[N]} = 1.$$

Ahora, observamos que $\pi_N\tilde{\nu} = \pi_N\nu$ pues $\tilde{\nu}(s) = \nu_s$ cuando $s < N$ y así $\pi_N\tilde{\nu}(s) = \nu_s = \pi_N\nu(s)$; si $s \geq N$ entonces $\pi_N\tilde{\nu}(s) = \sum_{s\geq N} \frac{\nu_{[N]}}{\rho_{[N]}}\rho_s = \frac{\nu_{[N]}}{\rho_{[N]}} \sum_{s\geq N} \rho_s = \frac{\nu_{[N]}}{\rho_{[N]}}\rho_{[N]} = \nu_{[N]} = \sum_{s\geq N} \nu_s = \pi_N\nu(s)$. Además $\mathcal{I}(\tilde{\nu}) = \mathcal{I}_{\pi_N\rho}(\pi_N\nu)$ ya que

$$\begin{aligned} \mathcal{I}(\tilde{\nu}) &= \sum_{s=1}^{N-1} \left(\nu_s \log\Big(\frac{\nu_s}{\rho_s}\Big) \right) + \sum_{s\geq N} \frac{\nu_{[N]}}{\rho_{[N]}}\rho_s \log\left(\frac{\frac{\nu_{[N]}}{\rho_{[N]}}\rho_s}{\rho_s} \right) \\ &= \sum_{s=1}^{N-1} \nu_s \log\Big(\frac{\nu_s}{\rho_s}\Big) + \frac{\nu_{[N]}}{\rho_{[N]}} \log\Big(\frac{\nu_{[N]}}{\rho_{[N]}}\Big) \sum_{s\geq N} \rho_s \\ &= \sum_{s=1}^{N-1} \nu_s \log\Big(\frac{\nu_s}{\rho_s}\Big) + \frac{\nu_{[N]}}{\rho_{[N]}}\rho_{[N]} \log\Big(\frac{\nu_{[N]}}{\rho_{[N]}}\Big) \\ &= \sum_{s=1}^{N-1} \nu_s + \nu_{[N]} \log\Big(\frac{\nu_{[N]}}{\rho_{[N]}}\Big) = \mathcal{I}_{\pi_N\rho}(\pi_N\nu). \end{aligned}$$

Más aún, $d(\tilde{\nu}, \rho) = d(\pi_N\nu, \pi_N\rho)$ pues

$$2d(\tilde{\nu}, \rho) = \sum_{s=1}^{N-1} |\nu_s - \rho_s| + \sum_{s\geq N} \Big| \frac{\nu_{[N]}}{\rho_{[N]}}\rho_s - \rho_s \Big| = \sum_{s=1}^{N-1} |\nu_s - \rho_s| + \Big| \frac{\nu_{[N]}}{\rho_{[N]}} - 1 \Big| \sum_{s\geq N} \rho_s$$

$$= \sum_{s=1}^{N-1} |\nu_s - \rho_s| + \Big| \frac{\nu_{[N]}}{\rho_{[N]}}\rho_{[N]} - \rho_{[N]} \Big| = \sum_{s=1}^{N-1} |\nu_s - \rho_s| + |\nu_{[N]} - \rho_{[N]}| = 2d(\pi_N\nu, \pi_N\rho).$$

Por lo anterior tenemos que $d(\tilde{\nu}, \rho) = d(\pi_N\tilde{\nu}, \pi_N\rho) = d(\pi_N\nu, \pi_N\rho) > a - \delta$ por lo que $\tilde{\nu} \in B^c_{a-\delta}(\rho)$ y así

$$\mathcal{I}_{\pi_N\rho}(\pi_N\nu) = \mathcal{I}_\rho(\tilde{\nu}) \geq \inf_{\nu\in B^c_{a-\delta}(\rho)} \mathcal{I}_\rho(\nu) = J(a - \delta).$$

Como ν era arbitrario entonces

$$-\inf_{\nu\in\mathcal{M}_1(\mathbb{N})|d(\pi_N\nu,\pi_N\rho)>a-\delta}\mathcal{I}_{\pi_N\rho}(\pi_N\nu)\leq -J(a-\delta).$$

Por (1.22) tenemos que

$$\limsup_{n\to\infty}\frac{1}{n}\log P(d(L_n,\rho)>a)\leq -J(a-\delta).$$

Haciendo δ tender a cero y de nuevo por (b) del Lema 1.4.1 obtenemos (1.14).

□

Antes de pasar al siguiente capítulo es necesario hacer un par de observaciones: primero es importante notar que el argumento principal en la demostración del teorema anterior es el teorema de Sanov en el caso finito y la δ que utilizamos es la que nos permite relacionar la medida truncada (que nos permite usar el teorema de Sanov en su versión finita) y las cotas a las cuales queremos llegar.

En segundo lugar tenemos que la velocidad de decaimiento, es decir, la función de tasa $\mathcal{I}$, es una función altamente utilizada en teoría de la información, grandes desviaciones, etc., en este caso es conocida como entropía relativa. En el Teorema 1.3.1 no hubo ambigüedades a la hora de utilizar la función $\mathcal{I}$ ahí mismo definida pues trabajamos sólo con medidas sobre un conjunto finito (es decir, la suma es finita y $\log(\frac{\nu_s}{\rho_s})$ está bien definida pues $\nu_s>0$ y $\rho_s>0$ para toda s); también es el caso del Teorema 1.4.1 y la función $\mathcal{I}$ definida ahí mismo, ya que en el Lema 1.4.1 probamos que la serie converge en $[0,\infty]$. Sin embargo, es necesario definir la entropía relativa de manera general pues los dos casos anteriores son casos muy particulares. A continuación damos una pequeña introducción a dicha función y demostramos algunas propiedades que nos serán útiles en el desarrollo de los próximos dos capítulos. Para profundizar en este concepto se puede consultar los libros [19] y [14]. Por último, la versión general del teorema de Sanov se mostrará en el capítulo 3 ya que para formularlo de una manera ligeramente distinta se requiere definir algunos conceptos que serán utilizados en ese mismo capítulo.

1.5. Entropía relativa

La entropía relativa o divergencia de Kullback-Leibler de una medida de probabilidad ν respecto a otra medida de probabilidad ρ, denotada por $D_{KL}(\nu||\rho)$, es un caso particular de las f-divergencias[6]; una divergencia es una función que cuantifica lo que difiere una medida de probabilidad respecto

[6]cf. [13], pág. 447.

a otra; sin embargo, las divergencias no son métricas (en general) pues pueden no cumplir ni con la simetría ni con la desigualdad del triángulo.

La entropía relativa cuantifica la variación logarítmica esperada de ν respecto a ρ, es decir, registra qué tanto esperamos que difiera una medida de probabilidad respecto a otra en proporción logarítmica. También suele interpretarse como la cantidad de información que se espera perder al aproximar ρ con ν. Cuando log se toma en base 2 la entropía relativa se mide en bits y cuando log es el logaritmo natural (el cual es nuestro caso) se mide en nats. A continuación damos una definición formal.

Definición 1.5.1. *Sea $(\Omega, \mathcal{B})$ un espacio medible y sea*

$$\Theta = \{A = (A_1, ..., A_k) | A \text{ es partición de } \Omega, \ \ A_i \in \mathcal{B} \ \forall i, \ \ k \in \mathbb{N}\}.$$

Sean ν y μ dos medidas de probabilidad sobre $(\Omega, \mathcal{B})$ y A una partición medible y finita de Ω (como en Θ). Sean

$$\nu^A = (\nu(A_1), ..., \nu(A_k)), \quad \mu^A = (\mu(A_1), ..., \mu(A_k))$$

y

$$D(\nu^A || \mu^A) = \sum_{i=1}^{k} \nu(A_i) \log\left(\frac{\nu(A_i)}{\mu(A_i)}\right).$$

Definimos la entropía relativa de ν respecto a μ como:

$$D_{KL}(\nu||\mu) = \sup_{A \in \Theta} D(\nu^A||\mu^A).$$

Observemos que $\log\left(\frac{\nu(A_i)}{\mu(A_i)}\right)$ sólo tiene sentido si siempre que se tiene $\mu(A_i) = 0$ sucede que $\nu(A_i) = 0$ con $A_i \in \mathcal{B}$, de la partición A. De otro modo definimos $D(\nu^A||\mu^A) = \infty$.

Recordemos que una medida de probabilidad ν es absolutamente continua respecto a otra medida de probabilidad μ si $\nu(B) = 0$ siempre que $\mu(B) = 0$ con $B \in \mathcal{B}$. Esta propiedad se denota como $\nu \ll \mu$[7].

Observación. *Si ν es absolutamente continua respecto a μ no significa que $D_{KL}(\nu||\mu) < \infty$ pues el supremo puede no existir, es decir, el conjunto $\{D(\nu^A||\mu^A)|A \in \Theta\}$ puede no estar acotado y así el supremo quedará definido como ∞. Claramente $\{D(\nu^A||\mu^A)|A \in \Theta\}$ es no vacío pues siempre se tiene la partición trivial $\{\Omega, \Omega^c = \emptyset\}$. Por otro lado, si ν no es absolutamente continua respecto a μ entonces $D_{KL}(\nu||\mu) = \infty$. En efecto, como ν no es absolutamente continua respecto a μ entonces existe $B \in \mathcal{B}$ tal que $\mu(B) = 0$ y $\nu(B) > 0$ luego para la partición $A = \{B, B^c\}$ se tiene que $D(\nu^A||\mu^A) = \infty$ entonces $D_{KL}(\nu||\mu) = \infty$. Lo anterior nos da que si $D_{KL}(\nu||\mu) < \infty$ entonces $\nu \ll \mu$.*

[7] cf. Ver sección 4.4 del apéndice.

En lo que sigue denotaremos a la entropía relativa simplemente como $D_{KL}(\nu||\mu) = D(\nu||\mu)$.

Lema 1.5.1. *Sean ν, μ dos medidas de probabilidad sobre $(\Omega, \mathcal{B})$. Entonces:*

(1) $D(\nu||\mu) \geq 0$. $D(\nu||\mu) = 0$ si y sólo si $\nu = \mu$.

(2) $D(\cdot||\cdot)$ es semi-continua inferiormente en ambas entradas.

(3) $D(\cdot||\cdot)$ es convexa.

Prueba. (1): Observemos que basta probar la afirmación para $D(\nu^A||\mu^A)$ con A una partición medible y finita pues si $D(\nu^A||\mu^A) \geq 0$ para toda A entonces $\sup_{A\in\Theta} D(\nu^A||\mu^A) \geq 0$. Más aún, se tiene entonces que $\sup_{A\in\Theta} D(\nu^A||\mu^A) = 0$ si y sólo si $D(\nu^A||\mu^A) = 0$ para toda partición medible y finita A.

Tomemos una partición $A = (A_1, ..., A_k)$ con $k \in \mathbb{N}$. Recordemos la desigualdad elemental: $\log x \leq x - 1$ para todo $x > 0$; la desigualdad es equivalente a la desigualdad $-\log x \geq 1 - x$ y la igualdad se da si y sólo si $x = 1$. Aplicando esta última desigualdad con $x = \frac{\mu(A_i)}{\nu(A_i)}$ se obtiene que

$$D(\nu^A||\mu^A) = \sum_{i=1}^{k} \nu(A_i) \log\left(\frac{\nu(A_i)}{\mu(A_i)}\right) = \sum_{i=1}^{k} \nu(A_i) \log\left(\frac{\mu(A_i)}{\nu(A_i)}\right)^{-1}$$

$$= \sum_{i=1}^{k} \nu(A_i) \left[-\log\left(\frac{\mu(A_i)}{\nu(A_i)}\right)\right] \geq \sum_{i=1}^{k} \nu(A_i) \left[1 - \frac{\mu(A_i)}{\nu(A_i)}\right] = 1 - 1 = 0.$$

La igualdad se da si y sólo si $\frac{\mu(A_i)}{\nu(A_i)} = 1$, i.e., $\mu(A_i) = \nu(A_i)$, $i = 1, ..., k$.

(2): En este caso también se reduce a probar que $D(\cdot||\mu^A)$ y $D(\nu^A||\cdot)$ son continuas para cualquier partición finita A, pues el supremo de funciones semi-continuas inferiormente es semi-continua inferiormente, en particular el supremo de funciones continuas lo es (cf. Lema 4.1.7 del apéndice). En efecto, sea $A = (A_1, ..., A_k)$ una partición finita. Se sabe que $\nu \mapsto \nu(B)$, $B \in \mathcal{B}$ es continua en la topología de la convergencia puntual o la topología producto[8] en $\Lambda(\Omega)$[9] (cf. Definición 4.3.5 del apéndice y el desarrollo ulterior); como composición de funciones continuas es continua, suma de funciones continuas es continua y multiplicar una funcion continua por algún escalar resulta ser una función continua tenemos que $D(\cdot||\mu^A)$ y $D(\nu^A||\cdot)$ son continuas.

[8] Además notemos que el caso en que Ω es finito o infinito numerable entonces $\Lambda(\Omega) = \mathcal{M}_1(\Omega) \subset \mathbb{R}^n$ o bien $\Lambda(\Omega) = \mathcal{M}_1(\Omega) \subset \mathbb{R}^\infty$ y la topología producto coincide con la topología euclideana.

[9] cf. Definición 2.0.1 del Capítulo 2.

(3): De nuevo observemos que todo se reduce a probar que $D(\cdot||\cdot)^A$ (con $D(\nu||\mu)^A = D(\nu^A||\mu^A)$) es convexa para toda partición finita A pues el supremo de funciones convexas resulta ser convexa (cf. Lema 4.1.8 del apéndice). En efecto, sea $\lambda \in [0,1]$ y ν_1,ν_2, μ_1 y μ_2 medidas de probabilidad sobre $(\Omega, \mathcal{B})$. Probaremos la convexidad en ambas entradas de manera conjunta: Nos apoyaremos en la siguiente desigualdad conocida como desigualdad suma-logaritmo[10] (que también puede ser usada para probar (1)):

$$\sum_{j=1}^{n} a_j \log\left(\frac{\sum_{j=1}^{n} a_j}{\sum_{j=1}^{n} b_j}\right) \leq \sum_{j=1}^{n} a_j \log\left(\frac{a_j}{b_j}\right),$$

con $a_1, ..., a_n$ $b_1, ..., b_n$ números reales no negativos, recordemos que convenimos que $\log(\frac{0}{b_j}) = -\infty$, $\log(\frac{a_j}{0}) = \infty$ y $\log(\frac{0}{0}) = 0$. Sean $\nu = \lambda\nu_1 + (1-\lambda)\nu_2$ y $\mu = \lambda\mu_1 + (1-\lambda)\mu_2$. Entonces

$$D(\nu||\mu) = \sum_{i=1}^{k} [\lambda\nu_1(A_i) + (1-\lambda)\nu_2(A_i)] \log\left(\frac{\lambda\nu_1(A_i) + (1-\lambda)\nu_2(A_i)}{\lambda\mu_1(A_i) + (1-\lambda)\mu_2(A_i)}\right).$$

Ahora, si definimos $a_1(A_i) = \lambda\nu_1(A_i)$, $a_2(A_i) = (1-\lambda)\nu_2(A_i)$, $b_1(A_i) = \lambda\mu_1(A_i)$ y $b_2(A_i) = (1-\lambda)\mu_2(A_i)$ y aplicamos la desigualdad anterior tenemos que

$$D(\nu||\mu) = \sum_{i=1}^{k} [a_1(A_i) + a_2(A_i)] \log\left(\frac{\sum_{j=1}^{2} a_j(A_i)}{\sum_{j=1}^{2} b_j(A_i)}\right)$$

$$= \sum_{i=1}^{k} \left[\sum_{j=1}^{2} a_j(A_i) \log\left(\frac{\sum_{j=1}^{2} a_j(A_i)}{\sum_{j=1}^{2} b_j(A_i)}\right)\right] \leq \sum_{i=1}^{k} \left[\sum_{j=1}^{2} a_j(A_i) \log\left(\frac{a_j(A_i)}{b_j(A_i)}\right)\right]$$

$$= \sum_{i=1}^{k} \lambda\nu_1(A_i) \log\left(\frac{\lambda\nu_1(A_i)}{\lambda\mu_1(A_i)}\right) + \sum_{i=1}^{k} (1-\lambda)\nu_2(A_i) \log\left(\frac{(1-\lambda)\nu_2(A_i)}{(1-\lambda)\mu_2(A_i)}\right)$$

$$= \lambda \sum_{i=1}^{k} \nu_1(A_i) \log\left(\frac{\nu_1(A_i)}{\mu_1(A_i)}\right) + (1-\lambda) \sum_{i=1}^{k} \nu_2(A_i) \log\left(\frac{\nu_2(A_i)}{\mu_2(A_i)}\right)$$

$$= \lambda D(\nu_1||\mu_1) + (1-\lambda) D(\mu_2||\nu_2).$$

□

[10] cf. Lema 4.1.9 del apéndice

En general puede ser complicado encontrar el supremo, por lo cual mostramos una manera más sencilla de computar la entropía relativa cuando $\nu \ll \mu$.

Lema 1.5.2. *Sean ν, μ dos medidas de probabilidad sobre $(\Omega, \mathcal{B})$ tales que $\nu \ll \mu$. Entonces se tiene que*

$$D(\nu||\mu) = \int_\Omega \log\left(\frac{d\nu}{d\mu}\right) d\nu. \tag{1.23}$$

En donde $\frac{d\nu}{d\mu}$ es la derivada de Radon-Nikodým de ν respecto a μ[11].

Prueba. Mostraremos las dos desigualdades en (1.23). Supongamos que $\nu \ll \mu$ entonces existe la derivada de Radon-Nikodým de ν respecto a μ, $\frac{d\nu}{d\mu}(\omega) = f(\omega)$. Sea B un evento tal que $\mu(B) > 0$ y consideremos a la función de distribución acumulada de la variable aleatoria f condicionada a que $\omega \in B$, es decir,

$$F_B(x) = \frac{\mu(\{f < x\} \cap B)}{\mu(B)} \quad x \in (-\infty, \infty).$$

Ahora, como $f(\omega) \geq 0$ casi seguramente con respecto a μ, entonces la esperanza de una variable aleatoria con distribución F_B es

$$\int_{-\infty}^{\infty} x \ dF_B(x) = \int_0^{\infty} x \ dF_B(x) = \frac{1}{\mu(B)} \int_{\{\omega | f(\omega) < \infty\} \cap B} f(\omega) \ d\mu(\omega)$$

$$= \frac{1}{\mu(B)} \int_B f(\omega) \ d\mu(\omega) = \frac{1}{\mu(B)} \int_B \frac{d\nu}{d\mu}(\omega) \ d\mu(\omega) = \frac{\nu(B)}{\mu(B)} < \infty.$$

En donde la primera identidad se tiene porque $f \geq 0$ implica $F_B(x) = 0$ para $x < 0$. Para cambiar de espacio en la tercera integral hacemos uso del teorema de cambio de variable para el push-forward[12] F_B inducido por f condicionada a estar en B, cf. Teorema 4.4.11 del apéndice aplicado a $\Psi = f$, la función identidad $g(x) = x$ y definiendo $\mu'(C) = \frac{\mu(C \cap B)}{\mu(B)}$ se tiene que $F_B = \mu' \circ \Psi^{-1} = \widetilde{\mu'}$. La tercera identidad se tiene ya que $f(\omega) < \infty$ para toda ω, es decir, el conjunto $\{\omega | f(\omega) < \infty\}$ tiene medida bajo μ igual a 1[13]. La última igualdad es la definición de la derivada de Radon-Nikodým. También observemos que

$$\int_0^{\infty} x \log x \ dF_B(x) = \frac{1}{\mu(B)} \int_B f(\omega) \log(f(\omega)) \ d\mu(\omega).$$

[11] cf. Teorema 4.4.8 del apéndice.
[12] cf. Definición 2.2.1 en el capítulo 2.
[13] cf. Lema 4.1.13 del apéndice.

Notemos que como $-e^{-1} \leq x \log x$ para toda $x \geq 0$ entonces

$$-\infty < -\frac{1}{e} = \int_0^\infty -\frac{1}{e} \ dF_B(x) \leq \int_0^\infty x \log x \ dF_B(x),$$

por lo que si aplicamos la desigualdad de Jensen (cf. Teorema 4.4.1 del apéndice) a la función convexa $x \log x$ se tiene

$$\frac{1}{\mu(B)} \int_B \log f \ d\nu = \frac{1}{\mu(B)} \int_B \log\left(\frac{d\nu}{d\mu}\right) \ d\nu$$

$$= \frac{1}{\mu(B)} \int_B \frac{d\nu}{d\mu} \log\left(\frac{d\nu}{d\mu}\right) \ d\mu = \frac{1}{\mu(B)} \int_B f \log f \ d\mu = \int_0^\infty x \log x \ dF_B(x)$$

$$\geq \left(\int_0^\infty x \ dF_B(x)\right) \log\left[\int_0^\infty x \ dF_B(x)\right] = \frac{\nu(B)}{\mu(B)} \log\left(\frac{\nu(B)}{\mu(B)}\right),$$

esto es, se tiene que para cualquier evento B que cumpla $Q(B) > 0$,

$$\frac{1}{\mu(B)} \int_B \log f \ d\nu \geq \frac{\nu(B)}{\mu(B)} \log\left(\frac{\nu(B)}{\mu(B)}\right).$$

Es decir,

$$\int_B \log f \ d\nu \geq \nu(B) \log\left(\frac{\nu(B)}{\mu(B)}\right) \quad \text{si} \quad \mu(B) > 0. \tag{1.24}$$

Ahora, sea $A = \{A_i\}_{i=1}^n$ una partición finita y medible de Ω, notemos que si $\nu(A_i) > 0$ entonces $\mu(A_i) > 0$ (esto ya que $\nu \ll \mu$), luego de (1.24) se sigue que

$$\int \log f \ d\nu = \sum_{i=1}^n \int_{A_i} \log f \ d\nu = \sum_{\{i|\nu(A_i)>0\}} \int_{A_i} \log f \ d\nu$$

$$\geq \sum_{\{i|\nu(A_i)>0\}} \nu(A_i) \log\left(\frac{\nu(A_i)}{\mu(A_i)}\right) = \sum_{i=1}^n \nu(A_i) \log\left(\frac{\nu(A_i)}{\mu(A_i)}\right).$$

Como la partición era arbitraria entonces se tiene que para toda partición medible $A \in \Theta$ de Ω se cumple que

$$D(\nu^A || \mu^A) = \sum_{i=1}^n \nu(A_i) \log\left(\frac{\nu(A_i)}{\mu(A_i)}\right) \leq \int \log f \ d\nu,$$

por lo tanto

$$D(\nu||\mu) = \sup_{A\in\Theta} D(\nu^A||\mu^A) \leq \int \log f \ d\nu.$$

Para probar la otra desigualdad en (1.23) vamos a definir lo que en [19] llaman cuantificadores. Definimos las siguientes funciones $q_n : \mathbb{R} \to \mathbb{R}$, $n = 1, 2, ...$

$$q_n(r) = \begin{cases} n & n \leq r \\ (k-1)2^{-n} & (k-1)2^{-n} \leq r < k2^{-n}; \quad k = 1, 2, ...n2^n \\ -(k-1)2^{-n} & -k2^{-n} \leq r < -(k-1)2^{-n}; \quad k = 1, 2, ...n2^n \\ -n & r < -n. \end{cases}$$

Fijando n notamos que q_n induce una partición en Ω, a saber la partición medible $\mathcal{A}_n$ con $2n2^n + 2$ conjuntos definidos como

$$\{\omega \in \Omega | q_n(\log f(\omega)) = n\},$$

$$\{\omega \in \Omega | q_n(\log f(\omega)) = (k-1)2^{-n}\}, \quad k = 1, ..., n2^n,$$

$$\{\omega \in \Omega | q_n(\log f(\omega)) = -(k-1)2^{-n}\}, \quad k = 1, ..., n2^n$$

$$\text{y} \quad \{\omega \in \Omega | q_n(\log f(\omega)) = -n\}.$$

Observemos que los $2n2^n$ conjuntos $\{\omega \in \Omega | q_n(\log f(\omega)) = (k-1)2^{-n}\}$, $k = 1, ..., n2^n$ y $\{\omega \in \Omega | q_n(\log f(\omega)) = -(k-1)2^{-n}\}$, $k = 1, ..., n2^n$ cumplen que para cualesquiera $\omega, \omega' \in A \in \mathcal{H}_n$ se tiene que

$$|\log f(\omega) - \log f(\omega')| \leq 2^{-n}. \tag{1.25}$$

A esta colección de conjuntos ajenos la denotaremos como $\mathcal{H}_n$. Por practicidad utilizaremos la siguiente notación:

$$\{\omega \in \Omega | q_n(\log f(\omega)) = n\} = \{n \leq \log f\}$$

análogamente

$$\{\omega \in \Omega | q_n(\log f(\omega)) = -n\} = \{\log f < -n\}.$$

Notemos que como $0 \leq \nu(\{n \leq \log f\}) \leq 1$ y $0 \leq \nu(\{\log f < -n\}) \leq 1$ entonces $0 \leq -\log[\nu(\{n \leq \log f\})]$ y $0 \leq -\log[\nu(\{\log f < -n\})]$ y así

$$\nu(\{n \leq \log f\}) \log\left[\frac{\nu(\{n \leq \log f\})}{\mu(\{n \leq \log f\})}\right] + \nu(\{\log f < -n\}) \log\left[\frac{\nu(\{\log f < -n\})}{\mu(\{\log f < -n\})}\right]$$

$\geq \nu(\{n \leq \log f\}) \log[\nu(\{n \leq \log f\})] + \nu(\{\log f < -n\}) \log[\nu(\{\log f < -n\})]$.

Además, como $\nu(\{n \leq \log f\})$ y $\nu(\{\log f < -n\})$ tienden a cero cuando n tiende a infinito y $x \log x$ tiende a cero cuando cuando x tiende a infinito entonces dada $\epsilon > 0$ existe N tal que si $n > N$ se tiene que

$$\nu(\{n \leq \log f\}) \log[\nu(\{n \leq \log f\})] + \nu(\{\log f < -n\}) \log[\nu(\{\log f < -n\})]$$

$$\geq -\epsilon.$$

Lo anterior nos lleva a la siguiente cota:

$$\sum_{A \in \mathcal{A}_n} \nu(A) \log \left[\frac{\nu(A)}{\mu(A)} \right]$$

$$= \sum_{A \in \mathcal{H}_n} \nu(A) \log \left[\frac{\nu(A)}{\mu(A)} \right] + \nu(\{n \leq \log f\}) \log \left[\frac{\nu(\{n \leq \log f\})}{\mu(\{n \leq \log f\})} \right]$$

$$+ \nu(\{\log f < -n\}) \log \left[\frac{\nu(\{\log f < -n\})}{\mu(\{\log f < -n\})} \right] \geq \sum_{A \in \mathcal{H}_n} \nu(A) \log \left[\frac{\nu(A)}{\mu(A)} \right] - \epsilon. \tag{1.26}$$

Ahora, tomemos $A \in \mathcal{H}_n$ y definimos $\bar{h} = \sup_{\omega \in A} \log f(\omega)$ y también $\underline{h} = \text{ínf}_{\omega \in A} \log f(\omega)$. Por (1.25) se tiene que

$$\bar{h} - \underline{h} = \sup_{\omega \in A} \log f(\omega) - \inf_{\omega' \in A} \log f(\omega') = \sup_{\omega \in A} \log f(\omega) + \sup_{\omega' \in A} -\log f(\omega')$$

$$= \sup_{\omega \in A} \log f(\omega) - \log f(\omega') \leq 2^{-n},$$

i.e.,

$$\underline{h} \geq \bar{h} - 2^{-n}. \tag{1.27}$$

También

$$\int_A \log f \ d\nu \leq \int_A \bar{h} \ d\nu = \bar{h}\nu(A); \tag{1.28}$$

además como la función exponencial es creciente se tiene que para toda $\omega \in A$ se cumple que $e^{\underline{h}} \leq e^{\log f(\omega)} = f(\omega)$ y así

$$\nu(A) = \int_A \frac{d\nu}{d\mu}\ d\mu = \int_A f\ d\mu \geq \int_A e^{\underline{h}}\ d\mu = e^{\underline{h}}\mu(A),$$

es decir,

$$\frac{1}{\nu(A)e^{-\underline{h}}} \leq \frac{1}{\mu(A)}. \tag{1.29}$$

Luego por (1.29), (1.27) y (1.28) se obtiene que

$$\nu(A)\log\left[\frac{\nu(A)}{\mu(A)}\right] \geq \nu(A)\log\left[\frac{\nu(A)}{\nu(A)e^{-\underline{h}}}\right] = \nu(A)\underline{h} \geq \nu(A)(\bar{h} - 2^{-n})$$

$$\geq \int_A \log f\ d\nu - \nu(A)2^{-n},$$

como lo anterior es para cualquier conjunto $A \in \mathcal{H}_n$ y por (1.26) se sigue que

$$\sum_{A\in\mathcal{A}_n} \nu(A)\log\left[\frac{\nu(A)}{\mu(A)}\right] \geq \sum_{A\in\mathcal{H}_n} \nu(A)\log\left[\frac{\nu(A)}{\mu(A)}\right] - \epsilon$$

$$\geq \sum_{A\in\mathcal{H}_n} \int_A \log f\ d\nu - 2^{-n} - \epsilon = \int_{\{\omega||\log(f(\omega))|\leq n\}} \log f\ d\nu - 2^{-n} - \epsilon.$$

Como ϵ es arbitrario, entonces se sigue que

$$\sup_{n\in\mathbb{N}} \sum_{A\in\mathcal{A}_n} \nu(A)\log\left[\frac{\nu(A)}{\mu(A)}\right] \geq \int \log f\ d\nu.$$

Recordando la Definición 1.5.1 se tiene que

$$D(\nu||\mu) = \sup_{A\in\Theta} \sum_{i=1}^{n} \nu(A_i)\log\left[\frac{\nu(A_i)}{\mu(A_i)}\right] \geq \sup_{n\in\mathbb{N}} \sum_{A\in\mathcal{A}_n} \nu(A)\log\left[\frac{\nu(A)}{\mu(A)}\right]$$

$$\geq \int \log f\ d\nu.$$

Esto es

$$D(\nu||\mu) \geq \int \log f\ d\nu.$$

□

Observación. *De la Definición 1.5.1 es claro que si el espacio Ω es discreto, es decir, tiene a lo más una cantidad numerable de elementos, entonces la entropía relativa es*

$$D(\nu||\mu) = \sum_{\omega\in\Omega} \nu(\omega)\log\left(\frac{\nu(\omega)}{\mu(\omega)}\right).$$

Ahora mostraremos la desigualdad de Pinsker que nos permite relacionar la entropía relativa y la distancia de variación total ($d(\cdot,\cdot)$). Además, es posible trabajar con un tipo de convergencia en el espacio de medidas de probabilidad llamada convergencia en información la cual se definirá después.

Lema 1.5.3. *(Desigualdad de Pinsker). Para cualesquiera dos medidas de probabilidad ν y μ sobre un espacio medible $(\Omega,\mathcal{B})$ se cumple que*

$$d(\nu,\mu) \leq \sqrt{2D(\nu||\mu)}.$$

Prueba. Supongamos que $D(\nu||\mu)$ es finito (de otro modo la demostración del lema es trivial). Sea

$$f(m) = p\log\left(\frac{p}{m}\right)+(1-p)\log\left(\frac{1-p}{1-m}\right)-2(p-m)^2.$$

Observemos que si $p \leq 1$ y $0 \leq m$ entonces $f(m) \geq 0$, ya que si $m = p$ entonces $f(m) = 0$; por otro lado la derivada de f respecto a m es

$$f'(m) = \frac{(1-2m)^2(p-m)}{(m-1)m} = (m-p)\frac{1-4m(1-m)}{m(1-m)},$$

que es negativa si $m < p$ y positiva si $p < m$. Además si $m = p$ entonces la derivada es cero lo cual nos dice que f decrece hacia su valor mínimo que es f(p) y luego crece, por lo que efectivamente $f(m) \geq f(p) = 0$. Ahora, como $d(\nu,\mu) = \sup_{B\in\mathcal{B}}|\nu(B)-\mu(B)|$[14] se tiene que

$$d(\nu,\mu)^2 = \left[\sup_{B\in\mathcal{B}}|\nu(B)-\mu(B)|\right]^2 = \sup_{B\in\mathcal{B}}\left[|\nu(B)-\mu(B)|^2\right]$$

$$\leq 4\sup_{B\in\mathcal{B}}\left[|\nu(B)-\mu(B)|^2\right].$$

Sea $\epsilon > 0$, por la propiedad del supremo se tiene que para 2ϵ existe un $B \in \mathcal{B}$ tal que

$$4\sup_{B\in\mathcal{B}}\left[|\nu(B)-\mu(B)|^2\right] - 2\epsilon \leq 4|\nu(B)-\mu(B)|^2,$$

de esta forma se tiene que

[14] cf. Definición 4.4.6 del apéndice.

$$d(\nu,\mu)^2 \leq 4(\nu(B)-\mu(B))^2 + 2\epsilon,$$

es decir,

$$-\frac{d(\nu,\mu)^2}{2} \geq -2[\nu(B)-\mu(B)]^2 - \epsilon.$$

Como $\{B, B^c\}$ es una partición de Ω y la entropía relativa es el supremo sobre las particiones finitas entonces se tiene que

$$D(\nu||\mu) - \frac{d(\nu,\mu)^2}{2}$$

$$\geq \nu(B)\log\left(\frac{\nu(B)}{\mu(B)}\right) + [1-\nu(B)]\log\left(\frac{1-\nu(B)}{1-\mu(B)}\right) - 2[\nu(B)-\mu(B)]^2 - \epsilon.$$

Gracias a la observación de la función f tenemos que el lado derecho de la desigualdad está acotado por abajo por $-\epsilon$, es decir,

$$D(\nu||\mu) - \frac{d(\nu,\mu)^2}{2} \geq -\epsilon;$$

lo cual nos da que

$$d(\nu,\mu)^2 \leq 2D(\nu||\mu) + 2\epsilon.$$

Como $\epsilon > 0$ era arbitrario entonces concluimos que

$$d(\nu,\mu) \leq \sqrt{2D(\nu||\mu)}.$$

□

En un principio uno se podría ver tentado a usar la entropía relativa como una métrica en el espacio de medidas de probabilidad sobre un espacio medible. Sin embargo, esto no es posible ya que, en general, no es simétrica, lo cual tiene sentido pues no esperamos perder la misma información al aproximar ρ con ν que al aproximar ν con ρ. A pesar de no ser una métrica la entropía relativa sí proporciona información acerca de la relación entre las medidas de probabilidad. De hecho, podemos definir un modo de convergencia con la entropía relativa fungiendo el papel de métrica cf. [19] y [11].

Definición 1.5.2. *Dada una sucesión de medidas de probabilidad* $\{\nu_n\}_{n\in\mathbb{N}}$ *sobre un espacio medible* $(\Omega, \mathcal{B})$ *decimos que la sucesión converge en información a* ν *si*

$$\lim_{n\to\infty} D(\nu_n||\nu) = 0.$$

La desigualdad de Pinsker mostrada en el Lema 1.5.3 implica que si una sucesión de medidas de probabilidad converge en información a alguna medida de probabilidad entonces la sucesión también converge en variación total a la misma medida de probabilidad.

Por último, nos gustaría comentar que la entropía relativa está relacionada con la transformada de Legendre-Fenchel (cf. Definición 1.2.1) mediante la identidad conocida como representación de Donsker-Varadhan cf. [15]. En [14] se prueban la mayoría de los resutados utilizando la representación de la entropía relativa como transformada de Legendre-Fenchel en lugar de la entropía relativa explícitamente.

El próximo capítulo está dedicado al uso de la entropía relativa para calcular proyecciones de medidas de probabilidad.

Capítulo 2

Representaciones de la I-proyección generalizada

En este capítulo introducimos los conceptos y definiciones necesarios para presentar los resutados de nuestro interés: Principalmente nos interesan dos conceptos fundamentales; a saber, el de I-proyección de una medida de probabilidad sobre un conjunto de medidas de probabilidad y el concepto de I-proyección generalizada de una medida de probabilidad sobre un conjunto de medidas de probabilidad. Probamos las condiciones para la existencia de las I-proyecciones, presentamos algunas propiedades y al final, en los últimos dos teoremas, obtenemos una descripción analítica de la I-proyección generalizada sobre dos conjuntos con ciertas características. Esta sección está basada principalmente en [11] y [10]. Otros excelentes materiales de consulta para algunos conceptos y resultados fueron [20] y [32]. Para todo lo concerniente a Análisis convexo puede consultarse [29] o [34].

Definición 2.0.1. *Sea $(S, \mathcal{B})$ un espacio medible arbitrario. Denotamos por $\Lambda(S)$ al conjunto de todas las medidas de probabilidad sobre $(S, \mathcal{B})$. Utilizaremos $\Lambda = \Lambda(S)$ cuando no exista motivo de confusión.*

Es de conocimiento general que muchos resultados sobre grandes desviaciones se deben al análisis convexo, por lo que es necesario poder hablar de convexidad en conjuntos de medidas, y para ello debe existir una estructura lineal, es decir, de espacio vectorial en el conjunto sobre cual queremos trabajar o en su defecto en un conjunto que contenga al conjunto que es de nuestro interés. Para poder trabajar debemos definir dos operaciones en Λ, a saber la suma y la multiplicación por escalares. Lo hacemos de la siguiente manera:

i) Suma: Dadas $P, Q \in \Lambda$ definimos $P + Q$ como

$$(P + Q)(B) = P(B) + Q(B) \quad \forall B \in \mathcal{B}.$$

ii) Multiplicación por escalares: Si $\lambda \in \mathbb{R}$ definimos $\lambda \cdot P$ como

$$(\lambda \cdot P)(B) = \lambda P(B) \quad \forall B \in \mathcal{B}.$$

Observamos que estas dos operaciones no son cerradas en Λ pues, por ejemplo, para la suma $(P+Q)(S) = P(S)+Q(S) = 1+1 = 2$ es decir $P+Q$ ya no es medida de probabilidad y para la multiplicación por escalares si multiplicamos por algún escalar negativo nuestra medida de probabilidad original ya no será mayor que cero, es decir, Λ no es un espacio vectorial. Sin embargo, el conjunto de las medidas con signo finitas denotado por $\mathbb{S}$ sí es un espacio vectorial sobre $\mathbb{R}$ con estas dos operaciones (cf. Lema 4.4.3 del apéndice) y en este caso sucede que $\Lambda \subset \mathbb{S}$. Más aún, Λ es un subconjunto convexo, lo cual mostramos a continuación.

Observación. Λ *es un conjunto convexo con las operaciones suma y multiplicación por escalares definidas en i) y ii): Si tomamos* $\lambda \in [0,1]$ *queremos ver que* $\lambda P + (1-\lambda)Q \in \Lambda$ *para toda* P *y para toda* Q *en* Λ. *En efecto, pues* $0 \leq \lambda P(B) + Q(B) - \lambda Q(B) \leq 1$ *para toda* $B \in \mathcal{B}$ *además* $\lambda P(S)+Q(S)-\lambda Q(S) = \lambda+1-\lambda = 1$ *y* $\lambda P(\emptyset)+Q(\emptyset)-\lambda Q(\emptyset) = 0+0-0 = 0$ *y claremente es sigma aditiva pues* P *y* Q *lo son.*

Con lo anterior ya estamos en condiciones de comenzar a discutir y trabajar en el espacio de medidas de probabilidad sobre un espacio medible. En lo subsiguiente la integral utilizada será siempre la integral de Lebesgue.

2.1. I-proyección generalizada

Haciendo una analogía con los espacios de Hilbert, nos gustaría saber bajo qué condiciones podemos proyectar una medida de probabilidad en un subespacio, es decir, encontrar una medida de probabilidad, dentro del conjunto sobre el cual estamos interesados, que sea lo más parecida a nuestra medida de probabilidad original; más aún, nos gustaría que dicha medida de probabilidad sea única. Al utilizar el término "parecida" estamos sugiriendo algún tipo de método de comparación entre dos medidas de probabilidad; podría ser, por ejemplo, una métrica como la distancia de variación total[1]. Sin embargo, algo más interesante sería utilizar la entropía relativa pues ya vimos que la convergencia en información implica la convergencia en variación total, es decir, lo resultados que se obtienen con la convergencia al usar la entropía relativa son más generales que los resultados al utilizar la métrica de variación total. Para ello introducimos los siguiente conceptos: Sea $Q \in \Lambda$. Definimos

$$D(\Pi||Q) = \inf_{P \in \Pi} D(P||Q)$$

[1] cf. Definición 4.4.6 del apéndice. Por comodidad durante el desarrollo de los resultados utilizaremos indistintamente $d(\nu,\mu)$ o bien $||\nu-\mu||_{VT}$ para denotar la distancia de variación total entre dos medidas de probabilidad ν y μ.

con el entendido de que si $D(P||Q) = \infty$ para toda $P \in \Pi$, entonces $D(\Pi||Q) = \infty$. En cualquier caso podemos observar que $D(\Pi||Q) \geq 0$, cf. (1) del Lema 1.5.1. Además, si existe $P \in \Pi$ tal que $D(P||Q) < \infty$ entonces $D(\Pi||Q) < \infty$.

Podemos notar que hay una clara analogía entre la definición anterior y la distancia de un punto a un subconjunto compacto en un espacio métrico: Si (X, d) es un espacio métrico y $A \subset X$ es compacto, se define $d(x, A) = \text{ínf}_{y \in A}\, d(x, y)$.

Definición 2.1.1. *Sean Q una medida de probabilidad y $\Pi \subset \Lambda$ tal que $D(\Pi||Q) < \infty$. Una medida de probabilidad P^* se dice que es la I-proyección de Q en Π si $P^* \in \Pi$ y $D(P^*||Q) = D(\Pi||Q)$.*

En general P^* no tiene por qué existir, sin embargo se puede asegurar que la I-proyección existe bajo ciertas condiciones que son enunciadas en el siguiente teorema.

Definimos $\Lambda_Q = \{P \in \Lambda|\ D(P||Q) < \infty\}$ y sea $\Pi \subset \Lambda$.

Teorema 2.1.1. *Sea $\Pi \subset \Lambda$ convexo y cerrado en la topología inducida por la métrica de variación total. Entonces se tiene que, para toda $Q \in \Lambda$ tal que $\Lambda_Q \cap \Pi \neq \emptyset$, existe P^* la I-proyección de Q en Π.*

Demostración. La intuición nos dice que la prueba es similar a la prueba de la existencia de la proyección sobre un conjunto cerrado y convexo en un espacio de Hilbert. En efecto, por ello debemos demostrar entonces un símil de la identidad del paralelogramo para la entropía relativa, lo cual se enuncia a continuación.

Afirmación. *Para toda R y para toda S en Λ se cumple que*

$$D(R||Q) + D(S||Q) = 2D\left(\frac{R+S}{2}\middle\|Q\right) + D\left(R\middle\|\frac{R+S}{2}\right) + D\left(S\middle\|\frac{R+S}{2}\right).$$

Prueba. La prueba es relativamente sencilla, sólo debemos escribir a la entropía relativa como integral[2] y hacer uso de las propiedades de la derivada de Radon-Nikodým (cf. el Teorema 4.4.8 y los resultados subsiguientes del apéndice). Tomamos S y R en Λ y observemos que

$$2D\left(\frac{R+S}{2}\middle\|Q\right) + D\left(R\middle\|\frac{R+S}{2}\right) + D\left(S\middle\|\frac{R+S}{2}\right) =$$

$$2\int \log\left(\frac{d[(R+S)/2]}{dQ}\right) d[(R+S)/2] + \int \log\left(\frac{dR}{d[(R+S)/2]}\right) dR$$

[2]Recordemos la representación integral de la entropía relativa mostrada en el Lema 1.5.2.

$$+\int \log\Big(\frac{dS}{d[(R+S)/2]}\Big)dS = \int \log\Big(\frac{d[(R+S)/2]}{dQ}\Big)dR$$
$$+\int \log\Big(\frac{d[(R+S)/2]}{dQ}\Big)dS + \int \log\Big(\frac{dR}{d[(R+S)/2]}\Big)dR$$

$$+\int \log\Big(\frac{dS}{d[(R+S)/2]}\Big)dS = \int \log\left[\Big(\frac{d[(R+S)/2]}{dQ}\Big)\Big(\frac{dR}{d[(R+S)/2]}\Big)\right]dR$$
$$+\int \log\left[\Big(\frac{d[(R+S)/2]}{dQ}\Big)\Big(\frac{dS}{d[(R+S)/2]}\Big)\right]dS$$

$$= \int \log\Big(\frac{dR}{dQ}\Big)dR + \int \log\Big(\frac{dS}{dQ}\Big)dS = D(R||Q) + D(S||Q).$$

□

Sea $\{P_n\}_{n\in\mathbb{N}}$ una sucesión de medidas de probabilidad tal que $D(P_n||Q)$ converge a $D(\Pi||Q)$ cuando n tiende a infinito, en donde $P_n \in \Pi$ y $D(P_n||Q) < \infty$ para toda $n \in \mathbb{N}$ (observamos que en particular se cumple que $P_n \ll Q$). Utilizando la afirmación anterior para P_m y P_n obtenemos que

$$D(P_m||Q) + D(P_n||Q) = 2D([P_m + P_n]/2||Q) + D(P_m||[P_m + P_n]/2)$$
$$+ D(P_n||[P_m + P_n]/2). \tag{2.1}$$

En donde $(P_m + P_n)/2 \in \Pi$ por convexidad. Entonces observamos que, por un lado, como $(P_m + P_n)/2 \in \Pi$ entonces $D(\Pi||Q) \leq D([P_m + P_n]/2||Q)$ y por otro lado como D es convexa se tiene que $D([P_m + P_n]/2||Q) \leq \frac{1}{2}\big(D(P_m||Q) + D(P_n||Q)\big)$. Tomando límite en ambas desigualdades obtenemos que

$$D(\Pi||Q) \leq \lim_{m,n\to\infty} D([P_m + P_n]/2||Q)$$

y que

$$\lim_{m,n\to\infty} D([P_m + P_n]/2||Q) \leq \lim_{m,n\to\infty} \frac{1}{2}\Big(D(P_m||Q) + D(P_n||Q)\Big) = D(\Pi||Q).$$

Es decir,

$$\lim_{m,n\to\infty} D([P_m + P_n]/2||Q) \to D(\Pi||Q).$$

Además tenemos que

$$\lim_{m,n\to\infty}[D(P_m||Q)+D(P_n||Q)]=2D(\Pi||Q).$$

De (2.1) se sigue que tanto $D(P_m||[P_m+P_n]/2)$ como $D(P_n||[P_m+P_n]/2)$ deben de converger a 0 cuando m y n tienden a infinito. Ahora, gracias a la desigualdad de Pinsker (cf. Lema 1.5.3) tenemos lo siguiente:

$$\left|\left|P_m-\frac{P_m+P_n}{2}\right|\right|_{VT}\leq\left(2D\Big(P_m\Big|\Big|\frac{P_m+P_n}{2}\Big)\right)^{\frac{1}{2}}\quad\forall m.$$

Utilizando la desigualdad del triángulo obtenemos que

$$\begin{aligned}||P_m-P_n||_{VT}\leq&\left|\left|P_m-\frac{P_m+P_n}{2}\right|\right|_{VT}+\left|\left|P_n-\frac{P_m+P_n}{2}\right|\right|_{VT}\\ \leq&\left(2D\Big(P_m\Big|\Big|\frac{P_m+P_n}{2}\Big)\right)^{\frac{1}{2}}+\left(2D\Big(P_n\Big|\Big|\frac{P_m+P_n}{2}\Big)\right)^{\frac{1}{2}}.\end{aligned}$$

Tomando límites de ambos lados obtenemos

$$\lim_{m,n\to\infty}||P_m-P_n||_{VT}$$

$$\leq\lim_{m,n\to\infty}\left[\left(2D\Big(P_m\Big|\Big|\frac{P_m+P_n}{2}\Big)\right)^{\frac{1}{2}}+\left(2D\Big(P_n\Big|\Big|\frac{P_m+P_n}{2}\Big)\right)^{\frac{1}{2}}\right]=0.$$

Por lo tanto $\{P_n\}_{n\in\mathbb{N}}$ es una sucesión de Cauchy y como Λ es un subconjunto cerrado de $(\mathbb{S},||\cdot||_{VT})$ que es un espacio de Banach (cf. Teorema 4.4.6 del apéndice y la observación posterior) entonces P_n converge a P_0 en variación cuando n tiende a infinito, con P_0 alguna medida de probabilidad, es decir,

$$\lim_{n\to\infty}P_n=P_0.$$

Como Π es cerrado tenemos que $P_0\in\Pi$. Por último, gracias al lema de Fatou (cf. Teorema 4.4.3 del apéndice) tenemos que

$$\begin{aligned}D(P_0||Q)=&\int\log\left(\frac{dP_0}{dQ}\right)dP_0=\int\liminf_{n\to\infty}\log\left(\frac{dP_n}{dQ}\right)dP_n\\ \leq&\liminf_{n\to\infty}\int\log\left(\frac{dP_n}{dQ}\right)dP_n=\lim_{n\to\infty}D(P_n||Q)=D(\Pi||Q),\end{aligned}$$

y como $P_0\in\Pi$, concluimos que $D(P_0||Q)=D(\Pi||Q)$, es decir, $P_0=P^*$ es la I-proyección de Q en Π. □

Observación. *Hasta ahora hemos mostrado la existencia de la I-proyección bajo ciertas condiciones pero nunca mencionamos la unicidad de esta. La prueba de la unicidad será simplificada mediante una nueva definición que enunciamos a continuación.*

El Teorema 2.1.1 garantiza que la I-proyección de una medida de probabilidad Q en un conjunto convexo Π existe si Π es cerrado; ahora, ¿qué sucede si Π no es cerrado? ¿Seguirá existiendo la I-proyección? La intución nos dice que no ya que se está trabajando con un proceso límite al tomar la convergencia de las entropías relativas de las medidas de probabilidad en Π respecto a Q hacia $D(\Pi||Q)$. En efecto, a continuación mostramos un contraejemplo.

Ejemplo. *Consideremos* $\Omega = \{a, b\}$*, un conjunto que consta de dos puntos. Sean* $\Pi = \left\{P \in \Lambda(\Omega) | 0 < P(a) < \frac{1}{2}\right\}$ *y* $Q \in \Lambda(\Omega)$ *tal que* $Q(a) = \frac{1}{2}$*. Observemos que* Π *es convexo pues si tomamos* $\lambda \in [0, 1]$ *y* $P_1, P_2 \in \Pi$ *entonces*

$$0 = 0 \cdot \lambda + 0 \cdot (1 - \lambda) < \lambda P_1(a) + (1 - \lambda)P_2(a) < \lambda\frac{1}{2} + (1 - \lambda)\frac{1}{2} = \frac{1}{2},$$

es decir, $\lambda P_1 + (1 - \lambda)P_2 \in \Pi$*. También notemos que* Π *no es cerrado en la topología inducida por la métrica de variación total. En efecto, tomemos la sucesión* $\{P_n\}_{n\in\mathbb{N}}$ *con* $P_n(a) = \frac{1}{2} - \frac{1}{n+2}$*, entonces* $P_n \in \Pi$ *para toda n y* P_n *converge a Q en variación total cuando n tiende a infinito ya que*

$$||P_n - Q||_{VT} = \frac{1}{2}\left[\left|\frac{1}{2} - \frac{1}{n+2} - \frac{1}{2}\right| + \left|1 - \left(\frac{1}{2} - \frac{1}{n+2}\right) - \frac{1}{2}\right|\right]$$

$$= \frac{1}{2}\left[\frac{1}{n+2} + \frac{1}{n+2}\right] = \frac{1}{n+2} \longrightarrow 0 \quad \textit{cuando} \quad n \longrightarrow \infty.$$

En donde en la primera igualdad hacemos uso de la representación de la distancia de variación total cf. Lema 4.4.4 del apéndice. Sin embargo, $Q \notin \Pi$*.*

Por otro lado, tomemos $P \in \Pi$ *entonces*

$$D(P||Q) = P(a)\log\left(\frac{P(a)}{Q(a)}\right) + P(b)\log\left(\frac{P(b)}{Q(b)}\right)$$

$$= P(a)\log\left(\frac{P(a)}{Q(a)}\right) + [1 - P(a)]\log\left(\frac{1 - P(a)}{1 - Q(a)}\right)$$

$$= P(a)\log(2P(a)) + [1 - P(a)]\log[2(1 - P(a))].$$

Observemos que la derivada de la función

$$f(x) = x\log(2x) + (1-x)\log[2(1-x)]$$

es $f'(x) = \log x - \log(1-x)$*, entonces* $f'(\frac{1}{2}) = 0$*; además* $f'(x) < 0$ *si* $x < \frac{1}{2}$ *y* $f'(x) > 0$ *si* $x > \frac{1}{2}$*, por lo tanto* $\frac{1}{2}$ *es el único punto mínimo de* f *en* $(0,1)$*. Así, el ínfimo en* Π *de* $D(P||Q)$ *es cero y se alcanza cuando* $P(a) = \frac{1}{2}$*, i.e., cuando* $P = Q$*, luego* $D(\Pi||Q) = D(Q||Q)$ *pero* $Q \notin \Pi$*.*

No todos los conjuntos son cerrados en la topología inducida por la métrica de variación total, y nos interesaría obtener siempre una medida de probabilidad que sea lo más parecida posible a otra medida $Q \in \Lambda$ en cualquier subconjunto convexo de Λ, es decir, nos interesa siempre poder obtener una proyección de Q en cualquier subconjunto convexo. Lo anterior nos lleva a formular una nueva definición.

Definición 2.1.2. *Sean* Q *y* Π *como en la Definición 2.1.1 . Una medida de probabilidad* P^* *(no necesariamente en* Π*) se dice que es la I-proyección generalizada de* Q *en* Π *si para toda sucesión de medidas de probabilidad* $\{P_n\}_{n\in\mathbb{N}} \subset \Pi$ *tal que* $D(P_n||Q)$ *converge a* $D(\Pi||Q)$ *cuando n tiende a infinito, se tiene que* P_n *coverge a* P^* *en variación total.*

Nos gustaría que esta nueva proyección siempre exista y que sea única. Afortunadamente lo anterior sucede y así, volviendo a la observación anterior, mostraremos que la I-proyección generalizada coincide con la I-proyección cuando esta última existe y por lo tanto la I-proyección también es única. Todo esto queda resumido en el siguiente teorema:

Teorema 2.1.2. *Sea* Π *un subconjunto convexo de* Λ *y supongamos que* $\Lambda_Q \cap \Pi \neq \emptyset$*. Entonces existe una única medida de probabilidad* $P^* \in \Lambda$ *(no necesariamente en* Π*) tal que* P_n *converge a* P^* *en variación total para toda sucesión* $\{P_n\}_{n\in\mathbb{N}} \subset \Pi$ *que cumpla que* $\lim_{n\to\infty} D(P_n||Q) = D(\Pi||Q)$*. Además, para cada* $P \in \Pi$ *se tiene*

$$D(P||P^*) + D(\Pi||Q) \leq D(P||Q). \tag{2.2}$$

La última desigualdad caracteriza a P^**, es decir, la I-proyección generalizada* P^* *es la única medida de probabilidad que cumple con (2.2) para toda* $P \in \Pi$*.*

Demostración. La primera parte de la demostración que corresponde a la existencia es completamente análoga a la del teorema anterior, de hecho en ese caso sólo utilizamos que el conjunto Π es cerrado para garantizar que la I-proyección pertenece a Π; en este caso no es de nuestro interés que la medida a la cual converge la sucesión $\{P_n\}_{n\in\mathbb{N}}$ pertenezca a nuestro conjunto Π. Dicho esto tenemos garantizado entonces que para toda $\{P_n\}_{n\in\mathbb{N}}$ tal que $D(P_n||Q)$ converge a $D(\Pi||Q)$ existe $P_0 \in \Lambda$ tal que $||P_n - P_0||_{VT}$ converge a 0

cuando n tiende a infinito. Hasta aquí tenemos la existencia, esto es, $P_0 = P^*$. Ahora proseguimos a mostrar la unicidad y lo haremos mostrando (2.2) y que esta desigualdad caracteriza a P^*, ya que si esto sucede tendríamos la unicidad.

Mostramos que la desigualdad efectivamente se cumple: tomamos $P \in \Pi$ y $\{P_n\}_{n\in\mathbb{N}} \subset \Pi$ tal que $n[D(\Pi||Q) - D(P_n||Q)]$ converge a 0 cuando n tiende a infinito, en particular $D(P_n||Q)$ converge a $D(\Pi||Q)$ y así P_n converge a P^* en variación total. Definimos la siguiente medida de probabilidad:

$$P'_n = \Big(1 - \frac{1}{n}\Big)P_n + \frac{1}{n}P.$$

Observamos que, por la convexidad de Π, $P'_n \in \Pi$ para toda n y por lo tanto tenemos que $D(\Pi||Q) \leq D(P'_n||Q)$ además por la propiedad de la entropía relativa, cf. Lema 4.1.10 del apéndice, se tiene que

$$D\left(\Big(1-\frac{1}{n}\Big)P_n+\frac{1}{n}P\Big|\Big|Q\right)=\Big(1-\frac{1}{n}\Big)D(P_n||Q)+\frac{1}{n}D(P||Q)-\Big(1-\frac{1}{n}\Big)D(P_n||P'_n)$$

$$-\frac{1}{n}D(P||P'_n).$$

Luego

$$D(\Pi||Q) \leq D(P'_n||Q) = D\left(\Big(1 - \frac{1}{n}\Big)P_n + \frac{1}{n}P\Big|\Big|Q\right)$$

$$=\Big(1 - \frac{1}{n}\Big)D(P_n||Q) + \frac{1}{n}D(P||Q)-\Big(1 - \frac{1}{n}\Big)D(P_n||P'_n) - \frac{1}{n}D(P||P'_n)$$

$$\leq D(P_n||Q) - \frac{1}{n}D(P_n||Q) + \frac{1}{n}D(P||Q) - \frac{1}{n}D(P||P'_n).$$

La última desigualdad se tiene ya que $-(1 - \frac{1}{n})D(P_n||P'_n) \leq 0$. Es decir,

$$D(\Pi||Q) \leq D(P_n||Q) - \frac{1}{n}D(P_n||Q) + \frac{1}{n}D(P||Q) - \frac{1}{n}D(P||P'_n),$$

de lo cual se sigue que

$$n[D(\Pi||Q) - D(P_n||Q)] + D(P_n||Q) + D(P||P'_n) \leq D(P||Q). \tag{2.3}$$

Por otro lado, dado que $D(\cdot||\cdot)$ es semi-continua inferiormente (cf. (2) del Lema 1.5.1) y P'_n converge a P^* en variación total tenemos que

$$D(P||P^*) \leq \liminf_{n\to\infty} D(P||P'_n).$$

Por lo cual, combinando esto con 2.3 y la elección original de P_n y la definición de P'_n,

$$\liminf_{n\to\infty}\big[n[D(\Pi||Q) - D(P_n||Q)] + D(P_n||Q) + D(P||P'_n)\big] \leq \liminf_{n\to\infty} D(P||Q),$$

es decir,

$$0 + D(\Pi||Q) + D(P||P^*) = D(\Pi||Q) + D(P||P^*) \leq D(P||Q).$$

Hemos mostrado que la desigualdad establecida en el Teorema 2.1.2 efectivamente se da pero falta mostrar que ésta caracteriza a P^*, es decir, que si existe alguna otra medida de probabilidad P' que cumpla la desigualdad entonces necesariamente $P' = P^*$.

Sea $P' \in \Lambda$ tal que

$$D(\Pi||Q) + D(P||P') \leq D(P||Q) \quad \forall P \in \Pi.$$

Sea $\{P_n\}_{n\in\mathbb{N}} \subset \Pi$ tal que $D(P_n||Q)$ converge a $D(\Pi||Q)$ cuando n tiende a infinito (de nuevo se tiene que P_n converge a P^* en variación total por lo que se acaba de mostrar), en particular como $P_n \in \Pi$ se tiene que

$$D(\Pi||Q) + D(P_n||P') \leq D(P_n||Q).$$

Tomando límite cuando n tiende a infinito obtenemos que $D(P_n||P')$ converge a 0 y como la convergencia en información implica la convergencia en variación total[3] entonces tenemos que P_n converge a P' en variación total pero también tenemos que P_n converge a P^* en variación total, por lo tanto concluimos que $P' = P^*$.

□

Notemos que P^* la I-proyección generalizada de Q en Π es la I-proyección si $P^* \in \Pi$. En efecto, si $\{P_n\}_{n\in\mathbb{N}} \subset \Pi$ es tal que $D(P_n||Q)$ converge a $D(\Pi||Q)$ cuando n tiende a infinito, entonces por la definición de I-proyección generalizada se tiene que P_n converge a P^* en variación total luego (análogamente al final de la demostración del Teorema 2.1.1) tenemos, por el Lema de Fatou, que

$$D(P^*||Q) \leq \lim_{n\to\infty} D(P_n||Q) = D(\Pi||Q).$$

Como $P^* \in \Pi$ entonces se sigue que $D(\Pi||Q) = D(P^*||Q)$. Por lo tanto, la I-proyección también es única (cuando existe).

En lo subsiguiente P^* denotará siempre a la I-proyección generalizada a menos de que se indique lo contrario.

[3]Recordemos la desigualdad de Pinsker, cf. Lema 1.5.3.

Otra desigualdad importante pero mucho más sencilla de probar es la siguiente: Si $Q \in \Lambda$ y $\Pi \subset \Lambda$ es un conjunto convexo tal que $D(\Pi||Q) < \infty$ entonces

$$D(P^*||Q) \leq D(\Pi||Q),$$

la cual se tiene por la semi-continuidad inferior de la entropía relativa: si $\{P_n\}_{n\in\mathbb{N}}$ es tal que $D(P_n||Q)$ converge a $D(\Pi||Q)$ cuando n tiende a infinito, entonces

$$D(P^*||Q) \leq \liminf_{n\to\infty} D(P_n||Q) = D(\Pi||Q).$$

2.2. Propiedades de la I-proyección generalizada

Ahora comenzaremos por mostrar las propiedades de la I-proyección generalizada que nos serán de gran utilidad para poder caracterizarla bajo ciertas circunstancias.

Lema 2.2.1. *Sean $\Pi \subset \Lambda$ un conjunto convexo tal que $\Lambda_Q \cap \Pi \neq \emptyset$ (i.e., $D(\Pi||Q) < \infty$) y P^* la I-proyección generalizada de Q en Π. Sea $P' \in \Lambda$ tal que $P' \ll Q$ y $P' \neq P^*$. Entonces*

$$\inf_{P\in\Pi\cap\Lambda_Q} \int \log\left(\frac{dP'}{dQ}\right) dP < \inf_{P\in\Pi\cap\Lambda_Q} \int \log\left(\frac{dP^*}{dQ}\right) dP = D(\Pi||Q). \tag{2.4}$$

En particular para cualquier función f medible en $(S, \mathcal{B})$ tal que $0 \leq \int f dP$ para cada $P \in \Pi \cap \Lambda_Q$, se tiene que

$$-\log \int e^f dQ \leq D(\Pi||Q), \tag{2.5}$$

y si se da la igualdad entonces

$$\frac{dP^*}{dQ}(s) = \frac{e^{f(s)}}{\int e^f dQ}. \tag{2.6}$$

Prueba. Gracias a (2.2) tenemos que para cada $P \in \Pi \cap \Lambda_Q$

$$D(\Pi||Q) \leq D(P||Q) - D(P||P^*) = \int \log\left(\frac{dP}{dQ}\right) dP - \int \log\left(\frac{dP}{dP^*}\right) dP$$

$$= \int \log\left(\frac{dP}{dQ}\right) - \log\left(\frac{dP}{dP^*}\right)\ dP = \int \log\left[\left(\frac{dP}{dQ}\right)\left(\frac{dP}{dP^*}\right)^{-1}\right]\ dP$$

$$= \int \log\left(\frac{dP^*}{dQ}\right) dP \leq D(P||Q).$$

La última desigualdad se da por el hecho de que $0 \leq D(P||Q)$ para toda P y para toda Q, así $D(P||Q) - D(P||P^*) \leq D(P||Q)$. También hicimos uso de algunas propiedades de la derivada de Radon-Nikodým, cf. Lema 4.4.8 del apéndice, en la tercera y cuarta identidad. En resumen tenemos que

$$D(\Pi||Q) \leq \int \log\left(\frac{dP^*}{dQ}\right) dP \leq D(P||Q) \quad \forall P \in \Pi \cap \Lambda_Q.$$

Observemos que $D(\Pi||Q) = \text{ínf}_{P\in\Pi} D(P||Q) = \text{ínf}_{P\in\Pi\cap\Lambda_Q} D(P||Q)$. Por lo que al tomar ínfimos obtenemos lo siguiente:

$$D(\Pi||Q) \leq \inf_{P\in\Pi\cap\Lambda_Q} \int \log\left(\frac{dP^*}{dQ}\right) dP \leq \inf_{P\in\Pi\cap\Lambda_Q} D(P||Q) = D(\Pi||Q),$$

es decir,

$$D(\Pi||Q) = \inf_{P\in\Pi\cap\Lambda_Q} \int \log\left(\frac{dP^*}{dQ}\right) dP. \tag{2.7}$$

Además como P^* queda únicamente determinada por (2.2) entonces si $P' \neq P^*$ se tiene que $D(P||Q) < D(P||P') + D(\Pi||Q)$ para toda $P \in \Pi \cap \Lambda_Q$ y ya que $P' \ll Q$, se sigue que

$$D(\Pi||Q) > D(P||Q) - D(P||P') = \int \log\left(\frac{dP}{dQ}\right) dP - \int \log\left(\frac{dP}{dP'}\right) dP$$

$$= \int \log\left[\left(\frac{dP}{dQ}\right)\left(\frac{dP}{dP'}\right)^{-1}\right] dP = \int \log\left(\frac{dP'}{dQ}\right) dP,$$

es decir,

$$\int \log\left(\frac{dP'}{dQ}\right) dP < D(\Pi||Q). \tag{2.8}$$

De (2.7) y (2.8) obtenemos que

$$\int \log\left(\frac{dP'}{dQ}\right) dP < D(\Pi||Q) = \inf_{P\in\Pi\cap\Lambda_Q} \int \log\left(\frac{dP^*}{dQ}\right) dP \quad \forall P \in \Pi \cap \Lambda_Q.$$

Así, al tomar ínfimos sobre $\Pi \cap \Lambda_Q$ obtenemos (2.4).

Sea f medible tal que $0 \leq \int f \ dP$ para toda $P \in \Pi \cap \Lambda_Q$. Para probar (2.5) debemos suponer que $\int e^f dQ < \infty$, de otro modo la afirmación sería trivial. Ahora, si definimos a P' de la siguiente manera:

$$P'(B) = \frac{1}{\int e^f dQ} \int_B e^f dQ \quad B \in \mathcal{B}$$

observamos que es una medida de probabilidad tal que $P' \ll Q$. En efecto, dado que $0 \leq e^f$ y por como está definida P' se sigue que $0 \leq P'(B) \leq 1$ para todo $B \in \mathcal{B}$, $P'(S) = 1$ y $P'(\emptyset) = 0$; la σ-aditividad la garantiza la σ-aditividad de la integral de Lebesgue. Para la continuidad absoluta observemos que si $Q(B) = 0$ entonces $\int_B e^f dQ = 0$ y se sigue que $P'(B) = 0$. De lo anterior se tiene que la derivada de Radon-Nikodým de P' respecto a Q se ve de la siguiente manera:

$$\frac{dP'}{dQ}(s) = \frac{e^{f(s)}}{\int e^f dQ}.$$

Ahora, como en un principio no sabemos si $P' \neq P^*$ entonces en (2.4) se puede dar la igualdad con P' definida hace unos instantes, es decir,

$$D(\Pi||Q) \geq \inf_{P \in \Pi \cap \Lambda_Q} \int \log\left(\frac{dP'}{dQ}\right) dP = \inf_{P \in \Pi \cap \Lambda_Q} \int \log\left(\frac{e^f}{\int e^f dQ}\right) dP$$

$$= \inf_{P \in \Pi \cap \Lambda_Q} \int \log e^f dP - \int \log\left(\int e^f dQ\right) dP = \inf_{P \in \Pi \cap \Lambda_Q} \int \log e^f dP$$

$$- \log \int e^f dQ = -\log \int e^f dQ + \inf_{P \in \Pi \cap \Lambda_Q} \int f \ dP \geq -\log \int e^f dQ.$$

La última desigualdad es cierta ya que

$$\int f \ dP \geq 0 \quad \forall P \in \Pi \cap \Lambda_Q;$$

además si sucede que $D(\Pi||Q) = -\log \int e^f dQ$ entonces necesariamente sucede que

$$D(\Pi||Q) = \inf_{P \in \Pi \cap \Lambda_Q} \int \log\left(\frac{dP'}{dQ}\right) dP,$$

lo cual por (2.4) sucede solamente si $P' = P^*$, i.e., $\frac{dP^*}{dQ}(s) = \frac{e^{f(s)}}{\int e^f dQ}$, con lo cual hemos probado (2.5) y (2.6). □

Lema 2.2.2. *Para subconjuntos convexos* $\Pi' \subset \Pi$ *de* Λ *tales que* $D(\Pi||Q) = D(\Pi'||Q) < \infty$ *se tiene que la I-proyección generalizada de* Q *en* Π *y la I-proyección generalizada de* Q *en* Π' *son la misma.*

Prueba. Se sigue de la definición de la I-proyección generalizada. Sea P^* la I-proyección generalizada de Q en Π. Si tomamos $\{P'_n\}_{n \in \mathbb{N}} \subset \Pi'$ tal que $D(P'_n||Q)$ converge a $D(\Pi'||Q) = D(\Pi||Q)$ cuando n tiende a infinito, en particular $\{P'_n\}_{n \in \mathbb{N}} \subset \Pi$ y por lo tanto P'_n converge a P^* en variación total. Y así las dos I-proyecciónes generalizadas de Q sobre cada conjunto coinciden. □

Observación. *El "regreso" no es válido, cf. Example 3.2 en [11]. Esto es, si* $\Pi' \subset \Pi \subset \Lambda$ *son conjuntos convexos tales que la I-proyección generalizada de* Q *en* Π *y la I-proyección generalizada de* Q *en* Π' *son la misma entonces no necesariamente es cierto que* $D(\Pi||Q) = D(\Pi'||Q)$. *Si la I-proyección generalizada de* Q *en* Π *es igual a la I-proyección generalizada de* Q *en* Π' *y además coincide con la I-proyección de* Q *en* Π' *entonces, en este caso, sí sucede que* $D(\Pi||Q) = D(\Pi'||Q)$. *En efecto, si denotamos a la I-proyección generalizada de* Q *en* Π' *como* P'^* *(que es también la I-proyección de* Q *en* Π'*) entonces* $P^* = P'^* \in \Pi' \subset \Pi$, *de esta manera* P^* *coincide con la I-proyección de* Q *en* Π. *Así*

$$D(\Pi||Q) = D(P^*||Q) = D(P'^*||Q) = D(\Pi'||Q).$$

Los lemas anteriores nos dan información acerca la I-proyección generalizada y nos serán de gran utilidad al intentar caracterizarla sobre un par de conjuntos con algunas propiedades mencionadas más adelante. Pero antes, es necesario mostrar un último lema concerniente a la entropía relativa. Éste nos da una propiedad interesante de la entropía relativa, a saber, la entropía relativa es invariante al cambiar de espacio mediante una función medible. A esta transformación se le conoce como push-forward; a continuación damos una definición formal.

Definición 2.2.1. *Sean* $(S_1, \mathcal{B}_1)$, $(S_2, \mathcal{B}_2)$ *espacios medibles y*

$$\Psi : (S_1, \mathcal{B}_1) \longrightarrow (S_2, \mathcal{B}_2)$$

una función medible. Sea μ *una medida sobre* $(S_1, \mathcal{B}_1)$. *Se define el push-forward de* μ *bajo* Ψ *como la medida* $\widetilde{\mu} : (S_2, \mathcal{B}_2) \to [0, \infty]$ *definida como* $\widetilde{\mu} = \mu \circ \Psi^{-1}$, *i.e.,* $\widetilde{\mu}(B) = \mu(\Psi^{-1}(B))$ *para cada* $B \in \mathcal{B}_2$.

Para ver que efectivamente es una medida debemos checar que cumple la definición. Por como está definida $\widetilde{\mu}$ se tiene que $\widetilde{\mu}(\emptyset) = 0$ y $0 \leq \widetilde{\mu}(B)$ para todo $B \in \mathcal{B}_2$. Para la σ-aditividad basta recordar que

$$\Psi^{-1}(\cup_{i=1}^{\infty} B_n) = \cup_{i=1}^{\infty} \Psi^{-1}(B_n)$$

y que

$$\Psi^{-1}(B_n \cap B_m) = \Psi^{-1}(B_n) \cap \Psi^{-1}(B_m).$$

Así, tenemos que si $\{B_n\}_{n\in\mathbb{N}}$ es una sucesión de elementos de $\mathcal{B}_2$ ajenos dos a dos entonces $\{\Psi^{-1}(B_n)\}_{n\in\mathbb{N}}$ es una sucesión de elementos de $\mathcal{B}_1$ también ajenos dos a dos y además

$$\widetilde{\mu}(\cup_{n\in\mathbb{N}} B_n) = \mu(\Psi^{-1}(\cup_{n\in\mathbb{N}} B_n)) = \mu(\cup_{n\in\mathbb{N}} \Psi^{-1}(B_n))$$

$$= \sum_{n\in\mathbb{N}} \mu(\Psi^{-1}(B_n)) = \sum_{n\in\mathbb{N}} \widetilde{\mu}(B_n).$$

Ejemplo. *Las distribuciones en Probabilidad:* Ψ *es una función medible (o variable aleatoria si* $(S, \mathcal{B})$ *son los reales con sus Borelianos) y en este caso* $\widetilde{P} = \widetilde{\mu}$ *es su ley o medida de distribución.*

Observamos también que el push-forward de una medida de probabilidad es de nuevo una medida de probabilidad. Por lo tanto, es natural preguntarse qué sucede con $D(\Pi||Q)$ al cambiar de espacio mediante Ψ. La respuesta a la pregunta anterior es enunciada en el siguiente lema:

Observación. *Usualmente se denota al push-forward de* P *bajo una función medible* Ψ *como* $\Psi_* P$ *o* $P\Psi^{-1}$*. En nuestro caso lo denotaremos como* $\widetilde{P}$*, siempre y cuando no exista motivo de confusión.*

Lema 2.2.3. *Sean* $(S_1, \mathcal{B}_1)$*,* $(S_2, \mathcal{B}_2)$ *espacios medibles y*

$$\Psi : (S_1, \mathcal{B}_1) \longrightarrow (S_2, \mathcal{B}_2)$$

una función medible. Sea Π *el conjunto de medidas de probabilidad sobre* S_1 *cuyo push-forward bajo* Ψ *pertenece a un conjunto convexo* $\widetilde{\Pi} \subset \Lambda(S_2)$*. Entonces, para* $Q \in \Lambda(S_1)$ *arbitraria y* $\widetilde{Q}$ *su push-forward bajo* Ψ*, se tiene que:*

$$D(\Pi||Q) = D(\widetilde{\Pi}||\widetilde{Q}). \tag{2.9}$$

Si $D(\Pi||Q) < \infty$ *entonces* P^**, la I-proyección generalizada de* Q *en* Π*, y* $\widetilde{P}^*$*, la I-proyección generalizada de* $\widetilde{Q}$ *en* $\widetilde{\Pi}$*, están relacionadas mediante la siguiente igualdad (que se da casi seguramente relativo a* Q*):*

$$\frac{dP^*}{dQ}(s) = \frac{d\widetilde{P}^*}{d\widetilde{Q}}(\Psi(s)). \tag{2.10}$$

Prueba. Primero observemos que toda partición finita de S_2 $\{A_i\}_{i=1}^m$ con $m \in \mathbb{N}$ y donde cada elemento de la particion pertenece a $\mathcal{B}_2$, genera una partición de S_1 donde cada elemento está en $\mathcal{B}_1$, esto ya que $S_1 = \Psi^{-1}(S_2) = \Psi^{-1}(\cup_{i=1}^m A_i) = \cup_{i=1}^m \Psi^{-1}(A_i)$ y además si $i \neq j$ entonces $\emptyset = \Psi^{-1}(A_i \cap A_j) = \Psi^{-1}(A_i) \cap \Psi^{-1}(A_j)$, por lo que si definimos $B_i = \Psi^{-1}(A_i)$ tenemos que $\{B_i\}_{i=1}^m$ es partición de S_1.

Sean

$$\Theta(S_1) = \{B = \{B_i\} | \text{B es partición medible y finita de } S_1\},$$

$$\Theta(S_2) = \{A = \{A_i\} | \text{A es partición medible y finita de } S_2\}$$

y

$$\Theta(S_1)^{\Psi} = \{B = \{B_i\} | B_i = \Psi^{-1}(A_i) \quad \text{con} \quad A = \{A_i\} \in \Theta(S_2)\}.$$

Entonces se tiene que $\Theta(S_1)^{\Psi} \subset \Theta(S_1)$. Luego, se sigue de la Definición 1.5.1 que

$$D(\widetilde{P}||\widetilde{Q}) \leq D(P||Q) \text{ con } P \in \Lambda(S_1) \text{ y } \widetilde{P} \text{ su push-forward bajo } \Psi. \quad (2.11)$$

Lo anterior ya que

$$D(\widetilde{P}||\widetilde{Q}) = \sup_{A\in\Theta(S_2)} \left\{ \sum_i \widetilde{P}(A_i) \log\left(\frac{\widetilde{P}(A_i)}{\widetilde{Q}(A_i)}\right)\right\} =$$

$$\sup_{A\in\Theta(S_2)} \left\{ \sum_i P\circ\Psi^{-1}(A_i) \log\left(\frac{P\circ\Psi^{-1}(A_i)}{Q\circ\Psi^{-1}(A_i)}\right)\right\}$$

$$= \sup_{B\in\Theta(S_1)^{\Psi}} \left\{ \sum_i P(B_i) \log\left(\frac{P(B_i)}{Q(B_i)}\right)\right\}$$

$$\leq \sup_{B\in\Theta(S_1)} \left\{ \sum_i P(B_i) \log\left(\frac{P(B_i)}{Q(B_i)}\right)\right\} = D(P||Q).$$

Más aún, para cualquier $\widetilde{P} \in \widetilde{\Pi}$ con $D(\widetilde{P}||\widetilde{Q}) < \infty$ (es decir $P \ll Q$) la medida de probabilidad $R \in \Lambda(S_1)$ definida por la densidad

$$\frac{dR}{dQ}(s) = \frac{d\widetilde{P}}{d\widetilde{Q}}(\Psi(s)) \quad s \in S_1,$$

tiene como push-forward bajo Ψ a $\widetilde{P}$ y satisface que

$$D(R||Q) = D(\widetilde{P}||\widetilde{Q}). \quad (2.12)$$

Esta última igualdad se sigue del teorema de cambio de variable aplicado a la derivada de Radon-Nikodým cf. Teorema 4.4.11 del apéndice pues

$$D(\widetilde{P}||\widetilde{Q}) = \int_{S_2} \log\left(\frac{d\widetilde{P}}{d\widetilde{Q}}\right) d\widetilde{P} = \int_{S_1} \log\left[\frac{d\widetilde{P}}{d\widetilde{Q}}\circ\Psi\right] \, dR$$

$$= \int_{S_1} \log\left(\frac{dR}{dQ}\right) dR = D(R||Q).$$

Para ver que efectivamente R tiene como push-forward bajo Ψ a $\widetilde{P}$ utilizamos el teorema de cambio de variable, la definición de R y la definición de derivada de Radon-Nikodým:

$$\int_{\Psi(B)} \frac{d\widetilde{P}}{d\widetilde{Q}} \, d\widetilde{Q} = \int_B \frac{d\widetilde{P}}{d\widetilde{Q}}\circ\Psi \, dQ = \int_B \frac{dR}{dQ} \, dQ = R(B) \quad \forall B \in \mathcal{B}_1.$$

Así, si $C \in \mathcal{B}_2$ entonces

$$R \circ \Psi^{-1}(C) = R(\Psi^{-1}(C)) = \int_{\Psi(\Psi^{-1}(C))} \frac{d\widetilde{P}}{d\widetilde{Q}} \; d\widetilde{Q} = \int_C \frac{d\widetilde{P}}{d\widetilde{Q}} \; d\widetilde{Q} = \widetilde{P}(C).$$

Por lo tanto, el push-forward bajo Ψ de R es $R \circ \Psi^{-1} = \widetilde{P}$, i.e., $R = P$. Ahora, de (2.12) obtenemos que $D(P||Q) = D(\widetilde{P}||\widetilde{Q})$ para toda $P \in \Pi$, tomando ínfimos obtenemos (2.9). De lo anterior se sigue que si $\{\widetilde{P}_n\}_{n\in\mathbb{N}} \subset \widetilde{\Pi}$ es tal que $D(\widetilde{P}_n||\widetilde{Q})$ converge a $D(\widetilde{\Pi}||\widetilde{Q})$ cuando n tiende a infinito, entonces las medidas de probabilidad $P_n \in \Pi$ definidas por:

$$\frac{dP_n}{dQ}(s) = \frac{d\widetilde{P}_n}{d\widetilde{Q}}(\Psi(s))$$

cumplen que $D(P_n||Q)$ converge a $D(\Pi||Q)$ cuando n tiende a infinito, esto ya que por (2.12) y (2.9) se tiene que

$$\lim_{n\to\infty} D(P_n||Q) = \lim_{n\to\infty} D(\widetilde{P}_n||\widetilde{Q}) = D(\widetilde{\Pi}||\widetilde{Q}) = D(\Pi||Q).$$

Por la definición de I-proyección generalizada tenemos que P_n converge a P^* y $\widetilde{P}_n$ converge a $\widetilde{P}^*$ en variación total, luego sucede que

$$\lim_{n\to\infty} \int_B \frac{dP_n}{dQ} \; dQ = P^*(B) \quad \forall B \in \mathcal{B}_1$$

y

$$\lim_{n\to\infty} \int_C \frac{d\widetilde{P}_n}{d\widetilde{Q}} \; d\widetilde{Q} = \widetilde{P}^*(C) \quad \forall C \in \mathcal{B}_2.$$

Por otro lado,

$$\lim_{n\to\infty} \int_C \frac{d\widetilde{P}_n}{d\widetilde{Q}} \; d\widetilde{Q} = \lim_{n\to\infty} \int_{\Psi^{-1}(C)} \frac{d\widetilde{P}_n}{d\widetilde{Q}} \circ \Psi \; dQ = \lim_{n\to\infty} \int_{\Psi^{-1}(C)} \frac{dP_n}{dQ} \; dQ$$

$$= \lim_{n\to\infty} P_n(\Psi^{-1}(C)) = P^*(\Psi^{-1}(C)).$$

Por lo tanto,

$$\widetilde{P}^*(C) = P^*(\Psi^{-1}(C)) \quad \forall C \in \mathcal{B}_2,$$

esto es,

$$\int_{\Psi^{-1}(C)} \frac{d\widetilde{P}^*}{d\widetilde{Q}} \circ \Psi \; dQ = \int_C \frac{d\widetilde{P}^*}{d\widetilde{Q}} \; d\widetilde{Q} = \int_{\Psi^{-1}(C)} \frac{dP^*}{dQ} \; dQ \quad \forall C \in \mathcal{B}_2.$$

Como la derivada de Radon-Nikodým de P^* respecto a Q es única casi seguramente relativo a Q entonces se sigue (2.10).

□

Finalmente mostraremos dos últimos lemas concernientes a la I-proyección y a la I-proyección generalizada y para ello es necesario definir algunos conjuntos especiales.

Sea $\mathcal{F}$ una familia de funciones reales y medibles sobre $(S, \mathcal{B})$.

Definición 2.2.2. *Designamos por* $\Pi(\mathcal{F})$ *y* $\Pi'(\mathcal{F})$ *a los conjuntos de todas las medidas de probabilidad* $P \in \Lambda$ *tales que* $\int f \ dP$ *existe y es no negativa, positiva respectivamente para toda* $f \in \mathcal{F}$*, es decir,*

$$\Pi(\mathcal{F}) = \left\{ P \in \Lambda \middle| \int f \ dP \geq 0, \quad \forall f \in \mathcal{F} \right\}$$

$$\Pi'(\mathcal{F}) = \left\{ P \in \Lambda \middle| \int f \ dP > 0, \quad \forall f \in \mathcal{F} \right\}.$$

Notemos que ambos conjuntos son convexos ya que si tomamos

$$R, S \in \Pi(\mathcal{F}) \ \ (R, S \in \Pi'(\mathcal{F}) \ \text{respectivamente}),$$

$f \in \mathcal{F}$ y consideramos una combinación lineal convexa $(1-\alpha)R + \alpha S$ con $\alpha \in [0, 1]$ obtenemos que

$$\int f \ d([1-\alpha]R + \alpha S) = (1-\alpha) \int f \ dR + \alpha \int f \ dS \geq 0$$

$$\int f \ d([1-\alpha]R + \alpha S) = (1-\alpha) \int f \ dR + \alpha \int f \ dS > 0$$

respectivamente, (cf. Lema 4.4.6 del apéndice).

Definición 2.2.3. *Sea* $K \in \mathcal{B}$*.* $\Pi(\mathcal{F}|K)$ *y* $\Pi'(\mathcal{F}|K)$ *denotan a los subconjuntos de* $\Pi(\mathcal{F})$ *y* $\Pi'(\mathcal{F})$ *respectivamente, que consisten en las medidas de probabilidad que cumplen* $P(K) = 1$*, esto es,*

$$\Pi(\mathcal{F}|K) = \left\{ P \middle| \int f \ dP \geq 0, \quad f \in \mathcal{F}, \quad P(K) = 1 \right\}$$

y

$$\Pi'(\mathcal{F}|K) = \left\{ P \middle| \int f \ dP > 0, \quad f \in \mathcal{F}, \quad P(K) = 1 \right\}.$$

Si $\mathcal{K} = \{K_i\}_{i \in \mathbb{N}}$ *es una sucesión de conjuntos tal que*

$$K_i \in \mathcal{B}, \ \ K_i \subset K_{i+1} \ \ \text{y cada} \ \ f \in \mathcal{F} \ \ \text{es acotada en} \ \ K_i \ \ \forall i. \tag{2.13}$$

En este caso escribiremos lo siguiente:

$$\Pi(\mathcal{F}|\mathcal{K}) = \bigcup_{i=1}^{\infty} \Pi(\mathcal{F}|K_i) \quad y \quad \Pi'(\mathcal{F}|\mathcal{K}) = \bigcup_{i=1}^{\infty} \Pi'(\mathcal{F}|K_i).$$

Observemos que tanto $\Pi(\mathcal{F}|K_i)$ y $\Pi'(\mathcal{F}|K_i)$ $i = 1, 2, ...$ como $\Pi(\mathcal{F}|\mathcal{K})$ y $\Pi'(\mathcal{F}|\mathcal{K})$ son conjuntos convexos, los dos primeros es claro; para los dos últimos notemos que como $K_i \subset K_{i+1}$ entonces $P(K_i) = 1$ implica que $P(K_{i+1}) = 1$ y así $\Pi(\mathcal{F}|K_i) \subset \Pi(\mathcal{F}|K_{i+1})$ y además ya observamos que cada $\Pi(\mathcal{F}|K_n)$ es convexo, por lo que si tomamos $P, Q \in \Pi(\mathcal{F}|\mathcal{K})$ entonces $P \in \Pi(\mathcal{F}|K_n)$ para algún n y $Q \in \Pi(\mathcal{F}|K_m)$ para algún m; podemos suponer sin pérdida de generalidad que $m < n$ y así $P, Q \in \Pi(\mathcal{F}|K_n)$ que es convexo, luego $(1-\lambda)P + \lambda Q \in \Pi(\mathcal{F}|K_n)$ para toda $\lambda \in [0, 1]$ y finalmente $(1-\lambda)P + \lambda Q \in \Pi(\mathcal{F}|\mathcal{K})$.

Las dos herramientas fundamentales para demostrar los dos teoremas principales concernientes a este capítulo son los siguientes dos lemas (2.2.4 y 2.2.5).

Lema 2.2.4. *Sea $Q \in \Lambda$. Si existe una $P_0 \in \Pi'(\mathcal{F})$ ($P_0 \in \Pi'(\mathcal{F}|\mathcal{K})$ respectivamente) tal que $D(P_0||Q) < \infty$ entonces se tiene que*

$$D(\Pi(\mathcal{F})||Q) = D(\Pi'(\mathcal{F})||Q)$$

$$(D(\Pi(\mathcal{F}|\mathcal{K})||Q) = D(\Pi'(\mathcal{F}|\mathcal{K})||Q)$$

respectivamente).

Prueba. Para mostrar lo anterior basta ver que para toda $P_1 \in \Pi(\mathcal{F})$, $P_1 \in \Pi(\mathcal{F}|\mathcal{K})$ existe $P' \in \Pi'(\mathcal{F})$ ($P' \in \Pi'(\mathcal{F}|\mathcal{K})$ respectivamente) tal que $D(P'||Q) \leq D(P_1||Q)$, ya que como $\Pi'(\mathcal{F}) \subset \Pi(\mathcal{F})$, $\Pi'(\mathcal{F}|\mathcal{K}) \subset \Pi(\mathcal{F}|\mathcal{K})$ las desigualdades

$$\inf_{P \in \Pi(\mathcal{F})} D(P||Q) \leq \inf_{P \in \Pi'(\mathcal{F})} D(P||Q)$$

y

$$\inf_{P \in \Pi(\mathcal{F}|\mathcal{K})} D(P||Q) \leq \inf_{P \in \Pi'(\mathcal{F}|\mathcal{K})} D(P||Q)$$

son triviales, y así sólo restaría mostrar las otras desigualdades.

Observemos que para toda $P_1 \in \Pi(\mathcal{F})$ ($P_1 \in \Pi(\mathcal{F}|\mathcal{K})$ respectivamente) se tiene que

$$P_\alpha = (1-\alpha)P_0 + \alpha P_1 \in \Pi'(\mathcal{F}) \text{ con } 0 < \alpha \leq 1$$

$$(P_\alpha = (1-\alpha)P_0 + \alpha P_1 \in \Pi'(\mathcal{F}|\mathcal{K}) \text{ con } 0 < \alpha \leq 1$$

respectivamente). Por otro lado, gracias a la convexidad de la entropía relativa tenemos que $D(P_\alpha||Q) \leq (1-\alpha)D(P_1||Q) + \alpha D(P_0||Q)$ y así si tomamos límites a ambos lados obtenemos que

$$\limsup_{\alpha\to 1} D(P_\alpha||Q) \leq \limsup_{\alpha\to 1}[(1-\alpha)D(P_1||Q) + \alpha D(P_1||Q)]$$

$$= \lim_{\alpha\to 1}[(1-\alpha)D(P_1||Q) + \alpha D(P_1||Q)] = D(P_1||Q).$$

Es decir,

$$\limsup_{\alpha\to 1} D(P_\alpha||Q) \leq D(P_1||Q).$$

Notemos que necesariamente existe una $P' \in \Pi'(\mathcal{F})$ ($\Pi'(\mathcal{F}|\mathcal{K})$ respectivamente) tal que $D(P'||Q) \leq \limsup_{\alpha\to 1} D(P_\alpha||Q)$. Ya que si no fuera así entonces para toda $P' \in \Pi'(\mathcal{F}), \Pi'(\mathcal{F}|\mathcal{K})$ se tendría que $\limsup_{\alpha\to 1} D(P_\alpha||Q) < D(P'||Q)$. Ahora, observemos que si tomamos una sucesión $\{\alpha_k\}_{k\in\mathbb{N}} \subset (0,1]$ tal que α_k converge a 1 nos genera una sucesión $\{D(P_{\alpha_k}||Q)\}_{k\in\mathbb{N}} \subset \mathbb{R}$ tal que

$$\limsup_{\alpha\to 1} D(P_\alpha||Q) = \limsup_{\alpha_k\to 1} D(P_{\alpha_k}||Q),$$

como $P_{\alpha_k} \in \Pi'(\mathcal{F})$ ($\Pi'(\mathcal{F}|\mathcal{K})$) para toda α_k y recordemos que el límite superior está dado por

$$\limsup_{\alpha_k\to 1} D(P_{\alpha_k}||Q) = \inf_{n\geq 0} \sup_{k\geq n}\{D(P_{\alpha_k}||Q)\},$$

entonces tendríamos que

$$\inf_{n\geq 0} \sup_{k\geq n}\{D(P_{\alpha_k}||Q)\} < D(P_{\alpha_k}||Q) \quad \forall k.$$

Luego, por ser un ínfimo y gracias a que la desigualdad es estricta entonces existe algún $n_0 \geq 0$ tal que

$$\sup_{k\geq n_0}\{D(P_{\alpha_k}||Q)\} < D(P_{\alpha_k}||Q),$$

lo cual es una contradicción. Entonces, efectivamente existe $P' \in \Pi'(\mathcal{F})$ ($\Pi'(\mathcal{F}|\mathcal{K})$ respectivamente) tal que $D(P'||Q) \leq \limsup_{\alpha\to 1} D(P_\alpha||Q)$ y así tenemos que $D(P'||Q) \leq D(P_1||Q)$. Como $P_1 \in \Pi(\mathcal{F})$ ($P_1 \in \Pi(\mathcal{F}|\mathcal{K})$ respectivamente) era arbitraria entonces tomamos ínfimos y obtenemos que

$$\inf_{P\in\Pi'(\mathcal{F})} D(P||Q) \leq \inf_{P\in\Pi(\mathcal{F})} D(P||Q)$$

y

$$\inf_{P\in\Pi'(\mathcal{F}|\mathcal{K})} D(P||Q) \leq \inf_{P\in\Pi(\mathcal{F}|\mathcal{K})} D(P||Q),$$

por lo tanto

$$D(\Pi(\mathcal{F})||Q) = D(\Pi'(\mathcal{F})||Q)$$

y

$$D(\Pi(\mathcal{F}|\mathcal{K})||Q) = D(\Pi'(\mathcal{F}|\mathcal{K})||Q).$$

□

En el siguiente lema se utilizan distintos modos de convergencia para las funciones medibles, lo cual se puede ver en la subsección 4.4.1. Modos de convergencia del apéndice.

Definición 2.2.4. *Un subconjunto Y de un espacio vectorial es un cono convexo si sucede que al tomar $n \in \mathbb{N}$, $y_i \in Y$ y $\alpha_i \geq 0$, i=1,...,n, entonces $\sum_{i=1}^{n} \alpha_i y_i \in Y$.*

Observación. *Ahora comenzaremos a trabajar en gran cantidad con afirmaciones relativas a una medida por lo cual utilizaremos la notación $[P]$ para decir que algo sucede casi seguramente relativo a P.*

Recordemos que dado un espacio vectorial V podemos trabajar con el espacio dual, i.e., el espacio de funcionales lineales acotados sobre V. Usualmente se denota al espacio dual de V como V^*; sin embargo, en nuestro caso lo denotaremos como V'.

Lema 2.2.5. *Sea $\mathcal{F}$ una familia de funciones reales y medibles sobre $(S, \mathcal{B})$ y supongamos que es un cono convexo. Entonces:*

a) Si $Q \in \Lambda$ es tal que $D(\Pi(\mathcal{F})||Q) < \infty$ y P^*, su I-proyección en $\Pi(\mathcal{F})$, existe entonces $\log(\frac{dP^*}{dQ}) - D(\Pi(\mathcal{F})||Q)$ pertenece a la $L_1(P^*)$-cerradura de $\mathcal{F}$.

b) Si $Q \in \Lambda$ es tal que $D(\Pi(\mathcal{F}|\mathcal{K})||Q) < \infty$ y P^* es su I-proyección generalizada en $\Pi(\mathcal{F}|\mathcal{K})$ entonces existe una sucesión $\{f_n\}_{n\in\mathbb{N}} \subset \mathcal{F}$ tal que

$$\log\Big(\frac{dP^*}{dQ}\Big) = D(\Pi(\mathcal{F}|\mathcal{K})||Q) + \lim_{n\to\infty} f_n \qquad [P^*]. \tag{2.14}$$

Observación. *El Lema 2.2.5 nos asegura que para la I-proyección de Q en $\Pi(\mathcal{F})$ hay una sucesión que converge a $\log(\frac{dP^*}{dQ}) - D(\Pi(\mathcal{F})||Q)$ en $L_1(P^*)$, mientras que para la I-proyección generalizada se tiene que existe una sucesión que converge casi donde sea a $\log(\frac{dP^*}{dQ}) - D(\Pi(\mathcal{F}|\mathcal{K})||Q)$ $[P^*]$.*

Prueba. a) Sea $\mathcal{F}_1 = \{f + g | f \in \mathcal{F}, \quad g \geq 0, \quad \text{g es medible y está acotada}\}$. $\mathcal{F}_1$ es un cono convexo ya que si $h_i \in \mathcal{F}_1$ y $\alpha_i \geq 0$, i=1,...,n entonces $h_i = f_i + g_i$ con $f_i \in \mathcal{F}$ y $g_i \geq 0$, acotada y medible, luego

$$\sum_{i=1}^{n} \alpha_i h_i = \sum_{i=1}^{n} \alpha_i f_i + \sum_{i=1}^{n} \alpha_i g_i.$$

Como $\mathcal{F}$ es un cono convexo entonces $\sum_{i=1}^{n} \alpha_i f_i \in \mathcal{F}$ y además

$$\sum_{i=1}^{n} \alpha_i g_i \geq 0$$

es acotada y medible, por lo tanto $\sum_{i=1}^{n} h_i \in \mathcal{F}_1$.

Sea

$$\varphi = \log\Big(\frac{dP^*}{dQ}\Big) - D(\Pi(\mathcal{F})||Q) \in L_1(P^*).$$

Para demostrar el lema primero mostraremos que si

$$P^* \in \Pi(\mathcal{F}) \;\; y \;\; D(P^*||Q) = D < \infty,$$

entonces φ pertenece a la $L_1(P^*)$-cerradura de $\mathcal{F}_1$.

En efecto, si suponemos lo contrario sucede que como φ es un punto entonces es convexo y además compacto entonces por el teorema de separación de Hahn-Banach[4] sabemos que φ puede ser separada de la cerradura de $\mathcal{F}_1$ en $L_1(P^*)$ por un hiperplano[5], i.e., por un funcional lineal acotado $\phi \in L_1(P^*)'$, en donde $L_1(P^*)' = L_\infty(P^*)$ (cf. Teorema 4.3.3 del apéndice). Es decir, se tiene que existen $\alpha_1, \alpha_2 \in \mathbb{R}$ tales que

$$\phi(\varphi) < \alpha_2 < \alpha_1 < \phi(f) \quad \forall f \text{ en la } L_1(P^*)\text{-cerradura de } \mathcal{F}_1.$$

En particular como $\mathcal{F}_1$ es subconjunto de su cerradura en $L_1(P^*)$ entonces

$$\phi(\varphi) < \alpha_2 < \alpha_1 < \phi(f) \quad \forall f \in \mathcal{F}_1.$$

Ahora, gracias al teorema de representación de Riesz[6], existe $h \in L_\infty(P^*)$ tal que

$$\int \varphi h \;\, dP^* = \phi(\varphi) < \alpha_2 < \alpha_1 < \phi(f) = \int f h \;\, dP^* \quad \forall f \in \mathcal{F}_1.$$

Como lo anterior es válido para toda $f \in \mathcal{F}_1$ obtenemos que

$$\int \varphi h \;\, dP^* < \inf_{f \in \mathcal{F}_1} \int f h \;\, dP^*. \tag{2.15}$$

Ahora, observemos que $0 \in \mathcal{F}_1$ por ser $\mathcal{F}_1$ un cono convexo, luego

$$\alpha_2 < \inf_{f \in \mathcal{F}_1} \int f h \;\, dP^* \leq \int 0 \cdot h \;\, dP^* = 0.$$

[4] cf. Teorema 4.3.1 del apéndice.
[5] cf. Definición 4.3.1 del apéndice.
[6] cf. Teorema 4.3.2 del apéndice.

Así, se tiene que $\alpha_2 < 0$. Por otro lado, notemos que tiene que suceder que $h \geq 0$ $[P^*]$ pues si suponemos lo contrario, es decir, que existe $B \in \mathcal{B}$ tal que $P^*(B) > 0$ y $h(s) < 0$ para toda $s \in B$, entonces $\int 1_B h \ dP^* \neq 0$, de hecho $\int 1_B h \ dP^* < 0$. Como $\mathcal{F}_1$ contiene a las funciones no negativas y acotadas entonces con $\alpha_2 < 0$ definimos

$$f = \frac{\alpha_2}{\int 1_B h \ dP^*} 1_B \in \mathcal{F}_1$$

(pues f es no negativa y acotada). Así, obtenemos que

$$\int fh \ dP^* = \int \frac{\alpha_2}{\int 1_B h \ dP^*} 1_B h \ dP^* = \frac{\alpha_2}{\int 1_B h \ dP^*} \int 1_B h \ dP^* = \alpha_2,$$

lo cual contradice el hecho de que $\alpha_2 < \phi(f)$ para toda f en la $L_1(P^*)$-cerradura de $\mathcal{F}_1$. La contradicción viene de suponer que h es negativa casi donde sea relativo a P^*, por lo tanto $h \geq 0$ $[P^*]$. Más aún, de hecho

$$0 \leq \phi(f) \quad \forall f \in \mathcal{F}_1$$

ya que si suponemos que existe una $f_0 \in \mathcal{F}_1$ tal que $a_{f_0} := \phi(f_0) < 0$ entonces como $\mathcal{F}_1$ es un cono convexo se tiene que $\phi(\mathcal{F}_1) = \{a \in \mathbb{R} | \phi(f) = a, \ \ f \in \mathcal{F}_1\}$ también es un cono convexo pues si $\alpha_i \geq 0$ y $a_i \in \phi(\mathcal{F}_1)$, $i = 1, ..., k$ entonces $a_i = \phi(f_i)$ con $f_i \in \mathcal{F}_1$, $i = 1, ..., k$ y

$$\sum_{i=1}^{k} \alpha_i a_i = \sum_{i=1}^{k} \alpha_i \phi(f_i) = \phi\left(\sum_{i=1}^{k} \alpha_i f_i \right) \in \phi(\mathcal{F}_1)$$

pues $\sum_{i=1}^{k} \alpha_i f_i \in \mathcal{F}_1$. Así, si $\alpha \geq 0$ tenemos que $\alpha a_{f_0} \in \phi(\mathcal{F}_1)$ y observamos que existe un $n \in \mathbb{N}$ tal que $n a_{f_0} \leq \alpha_2$ pues de lo contrario se tiene que para toda $n \in \mathbb{N}$ $\alpha_2 < n a_{f_0}$ es decir $n < \frac{\alpha_2}{a_{f_0}}$ lo cual no puede pasar pues los naturales no son acotados. Así, tenemos que $n a_{f_0} \leq \alpha_2$ y $n a_{f_0} \in \phi(\mathcal{F}_1)$, dicho de otro modo, tenemos que $n a_{f_0} = n\phi(f_0) = \phi(n f_0)$ y tomando a $f_1 = n f_0 \in \mathcal{F}_1$ entonces $\phi(f_1) \leq \alpha_2$ lo cual contradice el hecho de que $\alpha_2 < \phi(f)$ para toda $f \in \mathcal{F}_1$. Por lo tanto se tiene que

$$0 \leq \int fh \ dP^* \quad \forall f \in \mathcal{F}_1.$$

Además, como $0 \in \mathcal{F}_1$ entonces tenemos que

$$\int \varphi h \ dP^* < \inf_{f \in \mathcal{F}_1} \int fh \ dP^* = 0.$$

Más aún, podemos elegir h tal que $\int h \ dP^* = 1$ y que siga cumpliendo la desigualdad e igualdad anterior, esto ya que, suponiendo $h \neq 0$ (pues el caso h=0 no nos interesa) si definimos

$$h' = \frac{h}{\int h \ dP^*}$$

tenemos que $h' \in L_\infty$, $\int h' \ dP^* = 1$, $h' \geq 0$ y cumple que

$$0 \leq \frac{1}{\int h \ dP^*} \int fh \ dP^* = \int fh' \ dP^* \quad \forall f \in \mathcal{F}_1,$$

esto es,

$$\inf_{f \in \mathcal{F}_1} \int fh' \ dP^* = 0$$

(pues $0 \in \mathcal{F}_1$). Además, como teníamos que $\alpha_2 < 0$ entonces

$$\int \varphi h' \ dP^* = \frac{1}{\int h \ dP^*} \int \varphi h \ dP^* < \frac{1}{\int h \ dP^*} \ \alpha_2 < 0,$$

es decir,

$$\int \varphi h' \ dP^* < 0 = \inf_{f \in \mathcal{F}_1} \int fh' \ dP^*.$$

Ahora, como $\mathcal{F} \subset \mathcal{F}_1$ (ya que si $f \in \mathcal{F}$ entonces $f = f + g$ con $g = 0$), entonces (2.15) nos da que la medida de probabilidad P_0 definida por $\frac{dP_0}{dP^*} = h$ pertenece a $\Pi(\mathcal{F})$, esto ya que si $f \in \mathcal{F}$ entonces $f \in \mathcal{F}_1$ y además

$$0 = \inf_{f \in \mathcal{F}_1} \int fh \ dP^* \leq \int fh \ dP^* = \int f \ dP_0.$$

Por lo tanto $0 \leq \int f \ dP_0$ y así $P_0 \in \Pi(\mathcal{F})$.

También se tiene que $P_0 \in \Lambda_Q$ ya que $P^* \in \Lambda_Q$ y h es acotada, i.e., existe $M \in \mathbb{R}$ tal que $|h(s)| \leq M$ para toda $s \in S$ pero h es no negativa luego $h \leq M$ con $M \geq 0$, así

$$0 \leq D(P_0||Q) = \int \log\left(\frac{dP_0}{dQ}\right) dP_0 = \int h \log\left(h \frac{dP^*}{dQ}\right) \ dP^*$$

$$= \int h \log h \ dP^* + \int h \log\left(\frac{dP^*}{dQ}\right) \ dP^*$$

$$\leq M \log M + M \int \log\left(\frac{dP^*}{dQ}\right) dP^* = M \log M + M D(P^*||Q) < \infty.$$

Como por (2.4) tenemos que $D(\Pi(\mathcal{F})||Q) = \inf_{P \in \Pi(\mathcal{F}) \cap \Lambda_Q} \int \log\left(\frac{dP^*}{dQ}\right) dP$ entonces

$$\int \varphi h \ dP^* = \int \left[\log\left(\frac{dP^*}{dQ}\right) - D(\Pi(\mathcal{F})||Q)\right] \frac{dP_0}{dP^*} \ dP^*$$

$$= \int \log\left(\frac{dP^*}{dQ}\right) - D(\Pi(\mathcal{F})||Q)\ dP_0 = \int \log\left(\frac{dP^*}{dQ}\right) dP_0 - D(\Pi(\mathcal{F})||Q) \geq 0$$

(en la tercera identidad aplicamos el Lema 4.4.9 del apéndice). Es decir, $0 \leq \int \varphi h\ dP^*$, lo cual contradice (2.15). La contradicción viene de suponer que φ no pertenece a la $L_1(P^*)$-cerradura de $\mathcal{F}_1$ por lo que entonces este hecho si se da, i.e., existen f_n, g_n con $f_n \in \mathcal{F}$ y g_n acotada para toda $n \in \mathbb{N}$ tales que

$$\lim_{n\to\infty} f_n + g_n = \varphi \quad \text{en} \quad L_1(P^*). \tag{2.16}$$

Ahora, como $P^* \in \Pi(\mathcal{F})$ se tiene que $0 \leq \int f_n\ dP^*$ y además

$$\int \varphi\ dP^* = \int \log\left(\frac{dP^*}{dQ}\right) - D(\Pi(\mathcal{F})||Q)\ dP^*$$

$$= \int \log\left(\frac{dP^*}{dQ}\right) dP^* - D(\Pi(\mathcal{F})||Q) = 0.$$

Entonces, (2.16) nos da que

$$\lim_{n\to\infty} \left[\int f_n\ dP^* + \int g_n\ dP^*\right] = \int \varphi\ dP^* = 0$$

Para facilitar la notación definimos $||f|| = ||f||_{L_1(P^*)}$. De lo anterior se sigue que $||g_n|| \to 0$ cuando n tiende a infinito. Así,

$$||f_n - \varphi|| = ||f_n + g_n - \varphi - g_n|| \leq ||f_n + g_n - \varphi|| + ||g_n|| \quad \forall n \in \mathbb{N}.$$

Tomando límites de ambos lados de la desigualdad obtenemos

$$\lim_{n\to\infty} ||f_n - \varphi|| \leq \lim_{n\to\infty} ||f_n + g_n - \varphi|| + \lim_{n\to\infty} ||g_n|| = 0 + 0 = 0.$$

Lo que nos da como resultado que φ pertenece a la cerradura de $\mathcal{F}$ en la topología inducida por la métrica de $L_1(P^*)$.

Procedemos a probar b):

Denotemos ahora $D = D(\Pi(\mathcal{F}|\mathcal{K})||Q)$. En primer lugar tenemos que como $Q \in \Lambda$ es tal que $D(\Pi(\mathcal{F}|\mathcal{K})||Q) = D < \infty$ y $\Pi(\mathcal{F}|\mathcal{K}) = \cup_{i=1}^{\infty} \Pi(\mathcal{F}|K_i)$ con $K_i \subset K_{i+1}$ entonces existe $n_0 \in \mathbb{N}$ tal que $D_{n_0} := D(\Pi(\mathcal{F}|K_{n_0})||Q) < \infty$. Por otro lado, (2.13) implica que $\Pi(\mathcal{F}|K_n)$ es cerrado en la topología inducida por la métrica de variación total para toda n ya que si $\{P_m\}_{m\in\mathbb{N}} \subset \Pi(\mathcal{F}|K_n)$ es una sucesión de medidas de probabilidad tales que $||P_m - P||_{VT}$ converge a

0 cuando m tiende a infinito entonces $P \in \Pi(\mathcal{F}|K_n)$. En efecto, para mostrar lo anterior basta observar que como la convergencia en variación total es más fuerte que la convergencia fuerte (esto es, P_n converge fuertemente a R si $P_n(B)$ converge a R(B) para toda $B \in \mathcal{B}$), cf. Lema 4.4.5 del apéndice), entonces sucede que

$$P(K_n) = \lim_{m\to\infty} P_m(K_n) = \lim_{m\to\infty} 1 = 1.$$

Ahora, como f es acotada en K_i para toda $i = 1, 2, ...$ y para toda $f \in \mathcal{F}$ entonces

$$\lim_{m\to\infty} \int f \ dP_m = \int f \ dP \quad \forall f \in \mathcal{F},$$

pues la convergencia fuerte implica la convergencia anterior (débil), y como $0 \leq \int f \ dP_m$ para toda m se tiene que

$$0 \leq \lim_{m\to\infty} \int f \ dP_m = \int f \ dP \quad \forall f \in \mathcal{F}.$$

Así, $P \in \Pi(\mathcal{F}|K_n)$ lo que nos da que $\Pi(\mathcal{F}|K_n)$ es cerrado en la topología inducida por la métrica de variación total. Luego entonces existe la I-proyección de Q en $\Pi(\mathcal{F}|K_n)$ para toda $n > n_0$ denotada por P_n^*.

Sea $n > n_0$. Por el inciso a) tenemos que $\varphi_n = \log(\frac{dP_n^*}{dQ}) - D_n$ pertenece a la $L_1(P_n^*)$-cerradura de $\mathcal{F}$, en particular existen $f_n \in \mathcal{F}$ tales que

$$||\varphi_n - f_n||_{L_1(P_n^*)} < \frac{1}{n} \text{ excepto por a lo más un conjunto } A_n \text{ con}$$

$$P_n^*(A_n) < \frac{1}{n}. \tag{2.17}$$

Más aún, como $\Pi(\mathcal{F}|\mathcal{K}) = \cup_{i=1}^{\infty}\Pi(\mathcal{F}|K_i)$ y $K_i \subset K_{i+1}$ tenemos que

$$P_n^* \in \Pi(\mathcal{F}|\mathcal{K}). \tag{2.18}$$

Además, por la definición de I-proyección también se tiene

$$D(P_n^*||Q) = \inf_{P\in\Pi(\mathcal{F}|K_n)} D(P||Q)$$

y

$$D(\Pi(\mathcal{F}|\mathcal{K})||Q) = \inf_{P\in\Pi(\mathcal{F}|\mathcal{K})=\cup_{i=1}^{\infty}\Pi(\mathcal{F}|K_i)} D(P||Q).$$

Es decir, el conjunto sobre el cual tomamos el ínfimo en $D(\Pi(\mathcal{F}|\mathcal{K})||Q)$ contiene a $\Pi(\mathcal{F}|K_n)$, por lo cual, se tiene que

$$\lim_{n\to\infty} D(P_n^*||Q) = \lim_{n\to\infty} D_n = D = D(\Pi(\mathcal{F}|\mathcal{K})||Q).$$

Así, por la definición de I-proyección generalizada tenemos que

$$\lim_{n\to\infty} ||P_n^* - P^*||_{VT} = 0,$$

en particular se tiene que si $P_n^*(A_n)$ converge a 0 implica que $P^*(A_n)$ converge a 0 cuando n tiende a infinito. Entonces por (2.17) se sigue que

$$\varphi_n - f_n \to 0 \quad \text{en medida respecto a } P^*. \tag{2.19}$$

Por otro lado, como $||P_n^* - P^*||_{VT}$ converge a 0 y D_n converge a D cuando n tiende a infinito, se tiene entonces que

$$\lim_{n\to\infty} \varphi_n = \varphi = \log\Big(\frac{dP^*}{dQ}\Big) - D \tag{2.20}$$

(ya que $\varphi_n = \log(\frac{dP^*}{dQ}) - D_n$). De (2.19) y (2.20) obtenemos que f_n converge a φ en medida respecto a P^* cuando n tiende a infinito. Por último, como toda sucesión convergente en medida tiene una subsucesión que converge casi donde sea (cf. Teorema 4.4.7 del apéndice) entonces tenemos que existe $\{f_m\}_{m\in\mathbb{N}}$ con $f_m = f_{n_m}$ tal que

$$\log\Big(\frac{dP^*}{dQ}\Big) = D + \lim_{m\to\infty} f_m \qquad [P^*].$$

Es decir, hemos obtenido (2.14). □

2.3. Caracterizaciones de la I-proyección generalizada

La I-proyección generalizada de una medida de probabilidad Q en un conjunto convexo, como hemos podido observar, es una función de gran importancia (e.g. nos proporciona la velocidad de decaimiento en el teorema de Sanov[7]), por lo cual sería de gran ayuda poder caracterizarla (al menos para algunos conjuntos). Como se trata de encontrar un ínfimo, es decir, de optimizar las entropías relativas respecto a Q, no es tan sencillo describir analíticamente a dicho ínfimo ni a la I-proyección generalizada. Al caracterizarla para cierto tipo de conjuntos lograríamos saber cómo se comporta analíticamente y así podríamos trabajar con ella de manera explícita. Para ello presentamos dos representaciones para dos conjuntos con ciertas propiedades. Los conceptos, definiciones y algunos resultados son tomados de [6] y [7]. En cuanto a los conceptos y resultados de Análisis convexo pueden consultarse [29] y [34]. Para los conceptos de teoría de la probabilidad pueden

[7] cf. Teorema 1.3.1, Teorema 1.4.1 y Teorema 3.1.1.

consultarse [16] y [3]. Por último para el estudio de la teoría de los espacios vectoriales topológicos se pueden consultar [8] y [5].

Definición 2.3.1. *Un espacio vectorial topológico es localmente convexo si posee una base $\mathcal{V}$ de vecindades del vector 0 tal que cada una de ellas es un conjunto convexo. A un espacio vectorial topológico localmente convexo lo llamaremos espacio localmente convexo.*

En un enfoque más moderno se define a los espacios localmente convexos mediante seminormas (cf. el desarrollo del concepto en [31]). Sin embargo, en el desarrollo de los resultados aquí presentados utilizaremos la definición anterior pues bastará con la existencia de la base $\mathcal{V}$.

Sea V un espacio localmente convexo y consideremos ahora al espacio medible $(V, \mathcal{B})$ con $\mathcal{B}$ la σ-álgebra de Borel y a $\Lambda(V)$ como el conjunto de medidas de probabilidad sobre $(V, \mathcal{B})$.

Definición 2.3.2. *La esperanza o resultante de una medida de probabilidad $P \in \Lambda(V)$ se define como*

$$E(P) = v_0 \in V \quad si \quad \int \vartheta \ dP = \vartheta(v_0) \quad para\ cada\ \vartheta \in V' \tag{2.21}$$

siempre y cuando v_0 exista, de lo contrario $E(P)$ es indefinida. Otra terminología utilizada es el baricentro de P.

Observación. *Si $A \subset V$ es compacto, convexo y $P(A) = 1$ entonces $E(P)$ existe y $E(P) \in A$, cf. Teorema 4.3.6 del apéndice.*

Para $C \subset V$ utilizaremos conv(C) para denotar a la envoltura convexa de C, asimismo denotaremos cd(C) e int(C) a la cerradura de C y al interior de C respecto a la toplogía en V.

Por último $\equiv$ denota la continuidad absoluta mutua, esto es, $\nu \equiv \mu$ si y sólo si $\nu \ll \mu$ y $\mu \ll \nu$ (cf. Definición 4.4.5 del apéndice).

Teorema 2.3.1. *Sean $f_1, ..., f_k$ funciones medibles en $(S, \mathcal{B})$. Sea $\Pi \subset \Lambda$ definido de la siguiente manera*

$$\Pi = \left\{ P \Big| \int f_i \ dP \geq 0, \quad i = 1, ..., k \right\}. \tag{2.22}$$

Sean $Q \in \Lambda$ y $D = D(\Pi||Q)$. Entonces, se tiene que $D < \infty$ si y sólo si existe $P \in \Pi$ tal que $P \ll Q$. Ahora, si sucede lo anterior sea M el subespacio lineal más pequeño de $\mathbb{R}^k$ con la siguiente propiedad: $P(M) = 1$ para cada $P \in \Pi$ tal que $P \ll Q$. En este caso P^*, la I-proyección generalizada de Q en Π, está definida mediante la densidad

$$\frac{dP^*}{dQ}(s) = \begin{cases} e^{D+\sum_{i=1}^k \zeta_i^* f_i(s)} & si \quad s \in \{t|(f_1(t),...,f_k(t)) \in M\} \\ 0 & o.c. \end{cases} \tag{2.23}$$

en donde $\zeta^* = (\zeta_1^*,...,\zeta_k^*) \in \mathbb{R}_+^k$. *En el caso en que* $M{=}\mathbb{R}^k$, *(2.23) se sigue cumpliendo, es decir,* P^* *pertenece a la familia exponencial*

$$\left\{ P_\zeta \middle| \frac{dP_\zeta}{dQ} = \frac{e^{\sum_{i=1}^k \zeta_i f_i}}{\int e^{\sum_{i=1}^k \zeta_i f_i} dQ} \quad \zeta \in \Theta \right\} \tag{2.24}$$

con $\Theta = \{\zeta = (\zeta_1,...,\zeta_k) | \int e^{\sum_{i=1}^k \zeta_i f_i} dQ < \infty\}$ *si y sólo si existe* $P \in \Pi$ *tal que* $P \equiv Q$. *Con la última condición se tiene que*

$$D = \sup_{\zeta \in \mathbb{R}_+^k} \left[-\log \int e^{\sum_{i=1}^k \zeta_i f_i} dQ \right].$$

En donde el supremo se alcanza (i.e. es un máximo) si y sólo si $P_\zeta = P^*$.

Demostración. Primero observemos que siempre se tiene que si $D < \infty$ entonces existe P tal que $P \ll Q$, cf. la observación de la Definición 1.5.1. Observemos también que $(f_1(s),...f_k(s)) \in \mathbb{R}^k$. Consideremos la función $\Psi : S \to \mathbb{R}^k$, $\Psi(s) = (f_1(s),...f_k(s))$, y las proyecciones $\pi_i : \mathbb{R}^k \to \mathbb{R}$, $\pi_i((x_1....,x_k)) = x_i$, $i = 1,...,k$. Luego, como tenemos que

$$\Pi = \left\{ P \middle| \int f_i \ dP \geq 0, \quad i = 1,...k \right\},$$

si definimos $x = (x_1,...,x_k)$ y consideramos el siguiente conjunto

$$\widetilde{\Pi} = \left\{ P \middle| \int x_i \ dP(x) \geq 0, \quad i = 1,...k \right\}$$

con $x \in \mathbb{R}^k$, entonces se tiene que Π es el conjunto de medidas de probabilidad P tales que su push-forward bajo Ψ pertenece a $\widetilde{\Pi}$. En efecto, si $P \in \Pi$ entonces

$$\int_S f_i \ dP \geq 0.$$

Además tenemos que $(\pi_i \circ \Psi)(s) = \pi_i[(f_1(s),...,f_k(s))] = f_i(s)$, luego $\pi_i \circ \Psi = f_i$. Entonces con $\widetilde{P} = P \circ \Psi^{-1}$ y por el teorema de cambio de variable

$$\int_{\mathbb{R}^k} x_i \ d\widetilde{P}(x) = \int_{\mathbb{R}^k} \pi_i \ d\widetilde{P} = \int_S \pi_i \circ \Psi \ \ dP = \int_S f_i \ dP \geq 0.$$

Es decir, $\widetilde{P} \in \widetilde{\Pi}$ y de esta manera $\Pi \subset \{P|\widetilde{P} \in \widetilde{\Pi}\}$. Por otro lado, $\{P|\widetilde{P} \in \widetilde{\Pi}\} \subset \Pi$ ya que si tomamos P tal que $\widetilde{P} \in \widetilde{\Pi}$ por el mismo argumento anterior se tiene que

$$0 \leq \int_S f_i \ dP,$$

por lo que efectivamente Π consiste en todas las medidas de probabilidad cuyos push-forwards bajo Ψ pertenecen a $\widetilde{\Pi}$, o bien $\Pi = \{P|\widetilde{P} \in \widetilde{\Pi}\}$.

Así, por el Lema 2.2.3 tenemos que $D(\Pi||Q) = D(\widetilde{\Pi}||Q)$ y además

$$\frac{dP^*}{dQ}(s) = \frac{d\widetilde{P}^*}{d\widetilde{Q}}(\Psi(s)) \qquad [Q].$$

Entonces basta probar el teorema para el caso $S = \mathbb{R}^k$, $f_i(x_1, ..., x_k) = x_i$, $i = 1, ..., k$ y

$$\Pi = \left\{ P \Big| \int x_i \ dP(x) \geq 0, \quad i = 1, ..., k \right\}, \tag{2.25}$$

ya que lo anterior nos garantiza que si es cierto en este caso también lo es en el caso general.

Consideremos entonces a Π como en (2.25) y a una medida de probabilidad Q sobre $(\mathbb{R}^k, \mathcal{B} = \mathcal{B}(\mathbb{R}^k))$ tal que $P \ll Q$ para alguna $P \in \Pi$. Sean

$$K_n = \{x = (x_1, ..., x_k) | \ |x_i| \leq n, \quad i = 1, ..., k\} \tag{2.26}$$

y $\Pi_0 \subset \Pi$ definido de la siguiente forma:

$$\Pi_0 = \left\{ P \Big| P(K_n) = 1 \quad \text{para algún n y} \quad \int x_i \ dP(x) \geq 0, \quad i = 1, ..., k \right\}.$$

Es decir, Π_0 consiste en todas las medidas de probabilidad en $\mathbb{R}^k$ con soporte acotado.

Procederemos de la siguiente manera:

(1) Probaremos entonces que $D(\Pi_0||Q) < \infty$ y caracterizaremos a P_0^* la I-proyección generalizada de Q en Π_0.

(2) La demostración del teorema será completada al mostrar que

$$D(\Pi||Q) = D(\Pi_0||Q) \text{ y que } P^* = P_0^*$$

con P^* la I-proyección generalizada de Q en Π.

Mostramos (1):

Sea M el subespacio lineal más pequeño de $\mathbb{R}^k$ con la siguiente propiedad:

$$P(M) = 1 \quad \text{para cada} \quad P \in \Pi \quad \text{con} \quad P \ll Q. \tag{2.27}$$

Observamos que este subespacio siempre existe pues toda $P \in \Pi$ tal que $P \ll Q$ es medida de probabilidad sobre $\mathbb{R}^k$ luego $P(\mathbb{R}^k) = 1$, es decir, si no existe un subespacio propio de $\mathbb{R}^k$ que cumpla lo anterior entonces $M = \mathbb{R}^k$. También se tiene que $Q(M) > 0$ ya que si $Q(M) = 0$ y como $P \ll Q$ entonces $P(M) = 0$ lo cual no puede suceder. Ahora, consideremos el siguiente conjunto de medidas de probabilidad sobre $\mathbb{R}^k$:

$$\mathcal{P}_n = \left\{ P \Big| P \ll Q, \ \frac{dP}{dQ} \text{ es acotada}, \ M \cap K_n \subset \left\{ x | \frac{dP}{dQ}(x) > 0 \right\} \subset M \cap K_m \right\} \tag{2.28}$$

para alguna $m \geq n$ y todas las propiedades relativas a Q. Observemos que la última condición nos dice que si $P \in \mathcal{P}_n$ entonces su soporte es acotado.

Mostraremos que para toda n existe una $P_n \in \mathcal{P}_n \cap \Pi$. Para ello consideramos el siguiente conjunto:

$$E_n = \left\{ \left(\int x_1 \ dP(x), ..., \int x_k \ dP(x) \right) \in \mathbb{R}^k \Big| P \in \mathcal{P}_n \right\}.$$

Denotamos por $\mathbb{R}^k_+ = \{(y_1, ..., y_k) \in \mathbb{R}^k | y_i \geq 0 \quad \forall i = 1, ..., k\}$. Entonces observamos que es suficiente probar que $E_n \cap \mathbb{R}^k_+ \neq \emptyset$ para toda n pues si esto último sucede tenemos que existe una $P_n \in \mathcal{P}_n$ tal que

$$0 \leq \int x_i \ dP_n(x) \quad \forall i = 1, ..., k,$$

es decir, $P_n \in \Pi$. Sea Q_M la medida de probabilidad Q restringida a M y

$$sop(Q_M) = \{x \in M | \forall \text{U abierto tal que } \ x \in U \ \ Q_M(U) > 0\}$$

($sop(Q_M) \neq \emptyset$ pues $Q(M) > 0$). Sea $F \subset M$ la envoltura convexa cerrada de $sop(Q_M)$, entonces F es el subconjunto convexo cerrado más pequeño en M que contiene a $sop(Q_M)$ o representado también de la siguiente manera

$$F = \left\{ \sum_j \alpha_j x_j \Big| x_j \in sop(Q_M), \alpha_i \geq 0, \sum_j \alpha_j = 1 \right\}.$$

Sea aff(F) la envoltura afín de F, es decir, el conjunto más pequeño afín[8] que contiene a $sop(Q_M)$. Sea F_0 el interior de F relativo a aff(F), es decir,

$$F_0 = \{x \in F | \exists \epsilon > 0 \quad \text{tal que} \quad B_\epsilon(x) \cap aff(F) \subset F\}.$$

[8] cf. [29], pág. 3.

Antes de continuar es importante hacer las siguientes dos observaciónes, la segunda relaciona los soportes y la continuidad absoluta de dos medidas de probabilidad.

Observación. *Si $P \ll Q$ entonces $P \ll Q_M$.*

Observación. *Si $P \ll Q$ entonces $sop(P) \subset sop(Q)$. Si suponemos lo contrario tendríamos que existe $x \in sop(P)$ tal que $x \notin sop(Q)$ lo cual nos da que existe $\epsilon > 0$ tal que $Q(B_\epsilon(x)) = 0$ y sin embargo $P(B_\epsilon(x)) > 0$ lo cual es una contradicción pues $P \ll Q$.*

Ahora, si $P \in \mathcal{P}_n$ entonces $sop(P) \subset sop(Q_M)$ y tenemos que $sop(P)$ es acotado luego entonces $conv(sop(P)) \subset conv(sop(Q_M))$ y es acotado. Sea $H = cd[conv(sop(P))] \subset F$, entonces H es convexo compacto y $P(H) = 1$ pues $sop(P) \subset H$ luego por el teorema de Choquet (cf. Teorema 4.3.6 del apéndice) tenemos que

$$E(P) = \left(\int x_1 \ dP, ..., \int x_k \ dP \right) \in H \subset F,$$

por lo tanto $E_n \subset F$. Más aún, observemos que E_n es denso en F. En efecto, como $F \subset \cup_{n\in\mathbb{N}} K_n$ entonces si $x_0 = (x_0^1, ..., x_0^k) \in F$ se tiene que $x_0 \in K_m$ para algún m, como K_m es compacto y convexo entonces existe una medida de probabilidad μ sobre K_m cuyo soporte son los puntos extremos de K_m y tal que

$$x_0 = E(\mu) \quad \text{es decir} \quad x_0^i = \int x_i \ d\mu(x) \quad i = 1, ..., k$$

(cf. Teorema 4.3.7 del apéndice). Y notemos que para esta μ existe una sucesión $\{\mu_j\}_{j\in\mathbb{N}} \subset \mathcal{P}_n$ con $n < m$ tal que $\mu_j(B)$ converge a $\mu(B)$ cuando j tiende a infinito para toda $B \in \mathcal{B}$, lo que nos da que

$$x_0^i = \int x_i \ d\mu(x) = \lim_{j\to\infty} \int x_i \ d\mu_j(x) \quad i = 1, ..., k;$$

es decir, $\{x_j\}_{j\in\mathbb{N}} \subset E_n$ con $x_j = (\int x_1 \ d\mu_j(x), ..., \int x_k \ d\mu_j(x))$ es una sucesión tal que x_j converge a x_0 cuando j tiende a infinito. Por último, E_n es convexo ya que si

$$\left(\int x_1 \ d\nu(x), ..., \int x_k \ d\nu(x) \right), \left(\int x_1 \ d\mu(x), ..., \int x_k \ d\mu(x) \right) \in \ E_n$$

y $\lambda \in [0, 1]$, tenemos que

$$\lambda \left(\int x_1 \ d\nu(x), ..., \int x_k \ d\nu(x) \right) + (1 - \lambda) \left(\int x_1 \ d\mu(x), ..., \int x_k \ d\mu(x) \right)$$

$$= \left(\lambda \int x_1 d\nu(x) + (1-\lambda) \int x_1 d\mu(x), ..., \lambda \int x_k d\nu(x) + (1-\lambda) \int x_k d\mu(x) \right).$$

Y observemos que

$$\lambda \int x_i \ d\nu(x) + (1-\lambda) \int x_i \ d\mu(x) = \int x_i \ d[\lambda\nu(x)] + \int x_i \ d[(1-\lambda)\mu(x)]$$

$$= \int x_i \ d([\lambda\nu + (1-\lambda)\mu](x)) \quad \forall i = 1, ..., k.$$

Por lo que entonces basta ver que $\lambda\nu + (1-\lambda)\mu \in \mathcal{P}_n$. Claramente $\lambda\nu + (1-\lambda)\mu \ll Q$ pues si $Q(B) = 0$ para algún $B \in \mathcal{B}$ entoces $\nu(B) = \mu(B) = 0$; además

$$\frac{d(\lambda\nu + [1-\lambda]\mu)}{dQ} = \lambda \frac{d\nu}{dQ} + (1-\lambda)\frac{d\mu}{dQ}.$$

Como $\nu, \mu \in \mathcal{P}_n$ entonces sus respectivas derivadas de Radon-Nikodým respecto a Q son acotadas y así $\frac{d(\lambda\nu+[1-\lambda]\mu)}{dQ}$ es acotada, como $M \cap K_n \subset \{x | \frac{d\nu}{dQ} > 0\} \subset M \cap K_m$ y $M \cap K_n \subset \{x | \frac{d\mu}{dQ} > 0\} \subset M \cap K_m$ entonces si x es tal que

$$\frac{d(\lambda\nu + [1-\lambda]\mu)}{dQ}(x) > 0 \text{ si y sólo si } \lambda \frac{d\nu}{dQ}(x) + (1-\lambda)\frac{d\mu}{dQ}(x) > 0$$

$$\text{si y sólo si } \frac{d\nu}{dQ}(x) > 0 \quad \text{o} \quad \frac{d\mu}{dQ}(x) > 0.$$

Por lo tanto $\left\{ x \middle| \frac{d(\lambda\nu+[1-\lambda]\mu)}{dQ} > 0 \right\} = \{x | \frac{d\nu}{dQ} > 0\} \cup \{x | \frac{d\mu}{dQ} > 0\}$ y así tenemos $M \cap K_n \subset \left\{ x \middle| \frac{d[\lambda\nu+(1-\lambda)\mu]}{dQ} > 0 \right\} \subset M \cap K_m$; por lo que efectivamente E_n es convexo. Entonces, como $E_n \subset F$ es denso en F y además convexo se sigue que $F_0 \subset E_n$.

De este modo es suficiente mostrar que $F_0 \cap \mathbb{R}^k_+ \neq \emptyset$. Supongamos lo contrario, i.e., $F_0 \cap \mathbb{R}^k_+ = \emptyset$, en particular $F_0 \cap (M \cap \mathbb{R}^k_+) = \emptyset$ pues $M \cap \mathbb{R}^k_+ \subset \mathbb{R}^k_+$, claramente $M \cap \mathbb{R}^k_+$ es convexo (pues $\mathbb{R}^k_+$ es convexo y M también lo es al ser subespacio vectorial) y F_0 también lo es y además F_0 es abierto en M luego por el teorema de separación de Hahn-Banch (cf. Teorema 4.3.1 del apéndice) F_0 y $M \cap \mathbb{R}^k_+$ pueden ser separados por un hiperplano en M, i.e., un subespacio lineal $M_1 \subset M$ tal que $dim(M_1) = dim(M) - 1$. Lo anterior implica que para una medida de probabilidad P tal que $P \ll Q$, $P(M) = 1$ y $P \in \Pi$ necesariamente sucede que $sop(P) \subset M_1$ pues como

se observó se tiene que $P \ll Q_M$ y observemos que $Q_M(B) = 0$ para todo $B \subset M \backslash sop(Q_M)$ luego $P(B) = 0$ para todo $B \subset M \backslash sop(Q_M)$ en particular $P(intrel[M \cap \mathbb{R}^k_+]) = 0$, en donde $intrel(\cdot)$ denota al interior relativo; lo anterior nos dice que $sop(P) \subset cd(M \backslash \mathbb{R}^k_+)$ pero $P \in \Pi$, i.e., las entradas del vector esperanza de P son todas no negativas, es decir, tenemos una medida de probabilidad tal que si $x = (x_1, ..., x_k) \in sop(P)$ sucede que $x_i \leq 0$ para alguna $i = 1, ..., k$ (pueden ser varias e incluso todas) y $0 \leq \int x_i \ dP(x)$ para toda $i = 1, ..., k$. Esto sucede sólo si las r entradas $\{i_1, ..., i_r\}$ del vector x para las cuales $x_{i_j} \leq 0$, $j = 1, ..., r$ se tiene que $x_{i_j} = 0$, $j = 1, ..., r$ y $x_i \geq 0$ si $i \notin \{i_1, ..., i_r\}$, la razón es que el vector esperanza sea de entradas no negativas nos quiere decir que la masa de probabilidad está cargada hacia los vectores cuyas entradas son no negativas, es decir, la masa de probabilidad de P no puede estar concentrada en los vectores con entradas negativas. Notamos que lo anterior nos da que $sop(P)$ se encuentra contenido en algún subespacio vectorial propio M_2 de M (no puede ser M mismo ya que $sop(P) \subset cd[(M \cap \mathbb{R}^k_+)]$). Lo anterior nos daría que para toda $P \in \Pi$ que cumple $P \ll Q$ y $P(M) = 1$ se tiene que $sop(P) \subset M_2$, i.e., $P(M_2) = 1$, lo cual contradice la minimalidad de M. La contradicción viene de suponer que $F_0 \cap \mathbb{R}^k_+ = \emptyset$ entonces se tiene que $F_0 \cap \mathbb{R}^k_+ \neq \emptyset$ de lo cual se sigue que existe una medida de probabilidad $P_n \in \Pi$ tal que $P_n \in \mathcal{P}_n$.

Ahora, observemos que como P_n y Q son medidas de probabilidad entonces $\frac{dP_n}{dQ}$ es no negativa salvo en un conjunto de medida cero respecto a Q y como $K_n \cap M \subset \left\{x | \frac{dP_n}{dQ}(x) > 0\right\} \subset M \cap K_m$ tenemos que

$$1 = P_n(\mathbb{R}^k) = \int_{\mathbb{R}^k} \frac{dP_n}{dQ} \ dQ = \int_{\{x | \frac{dP_n}{dQ}(x) > 0\}} \frac{dP_n}{dQ} \ dQ + \int_{\{x | \frac{dP_n}{dQ}(x) = 0\}} \frac{dP_n}{dQ} dQ$$

$$+ \int_{\{x | \frac{dP_n}{dQ}(x) < 0\}} \frac{dP_n}{dQ} dQ = \int_{\{x | \frac{dP_n}{dQ}(x) > 0\}} \frac{dP_n}{dQ} dQ + \int_{\{x | \frac{dP_n}{dQ}(x) = 0\}} \frac{dP_n}{dQ} dQ + 0$$

$$= \int_{\{x | \frac{dP_n}{dQ}(x) > 0\}} \frac{dP_n}{dQ} dQ + 0 = \int_{\{x | \frac{dP_n}{dQ}(x) > 0\}} \frac{dP_n}{dQ} dQ = P_n\left(\left\{x | \frac{dP_n}{dQ}(x) > 0\right\}\right).$$

Así

$$1 = P_n\Big(\Big\{x | \frac{dP_n}{dQ}(x) > 0\Big\}\Big) \leq P_n(M \cap K_m) \leq P_n(K_m),$$

por lo que $P_n(K_m) = 1$ lo que nos da que $P_n \in \Pi_0$. Así, tenemos que $D(P_n||Q) < \infty$ pues $P_n \ll Q$ y $\frac{dP_n}{dQ}$ es acotada, es decir existe $M \in \mathbb{R}$ tal que $\frac{dP_n}{dQ}(x) \leq M$ para toda $x \in \mathbb{R}^k$ [Q], luego

$$D(P_n||Q) = \int \log\left(\frac{dP_n}{dQ}\right) dP_n \leq \int \log(M) dP_n = \log(M) < \infty.$$

Como $P_n \in \Pi_0$ se tiene que $D(\Pi_0||Q) < \infty$.

Sea $D_0 = D(\Pi_0||Q) < \infty$. Ahora caracterizaremos a la I-proyección generalizada de Q en Π_0:

Sea P_0^* la I-proyección generalizada de Q en Π_0, notemos que Π_0 es de la forma $\Pi(\mathcal{F}|\mathcal{K})$ con $\mathcal{F}$ el cono convexo (cf. Lema 4.1.11 del apéndice) de combinaciones lineales no negativas de las funciones $f_i = (x_1, ..., x_k) = x_i$, es decir, $\mathcal{F}$ consta de las funciones de la forma

$$f(x) = \sum_{i=1}^{k} \zeta_i x_i \quad \zeta = (\zeta_1, ..., \zeta_k) \in \mathbb{R}_+^k.$$

Tenemos la sucesión de cubos[9] $\{K_n\}_{n\in\mathbb{N}}$ que cumplen $K_i \in \mathcal{B}$, $K_i \subset K_{i+1}$ y cada $f \in \mathcal{F}$ está acotada en K_i, $i \in \mathbb{N}$. Así, $\Pi_0 = \cup_{i\in\mathbb{N}} \Pi(\mathcal{F}|K_i) = \Pi(\mathcal{F}|\mathcal{K})$. Por el Lema 2.2.5 tenemos que

$$\log\left(\frac{dP_0^*}{dQ}\right) = D_0 + \lim_{n\to\infty} f_n \qquad [P_0^*],$$

con $\{f_n\}_{n\in\mathbb{N}}$ una sucesión en $\mathcal{F}$. Observamos que el límite de cualquier sucesión $\{f_n\}_{n\in\mathbb{N}} \subset \mathcal{F}$ que converge casi donde sea es igual casi donde sea a una función en $\mathcal{F}$ ya que

$$f_n(x) = \sum_{i=1}^{k} \zeta_i^n x_i \to f \qquad [P_0^*] \quad \text{cuando} \quad n \to \infty,$$

es decir,

$$\zeta_i^n \to \zeta_i^* \geq 0 \quad \text{cuando} \quad n \to \infty.$$

Así $f(x) = \sum_{i=1}^{k} \zeta_i^* x_i$ con $\zeta^* = (\zeta_1^*, ..., \zeta_k^*) \in \mathbb{R}_+^k$ y $x = (x_1, ..., x_k) \in \mathbb{R}^k$. Por lo tanto

$$\log\left[\frac{dP_0^*}{dQ}\right](x) = D_0 + \sum_{i=1}^{k} \zeta_i^* x_i \quad [P_0^*]. \tag{2.29}$$

De acuerdo al plan de la demostración procedemos a mostrar (2):

Recordemos que la desigualdad (2.2) nos dice que

$$D(P||P_0^*) + D(\Pi_0||Q) \leq D(P||Q) \quad \forall P \in \Pi_0,$$

[9] cf. 2.26.

en particular

$$D(P_n||P_0^*) + D(\Pi_0||Q) \leq D(P_n||Q) \quad \forall n \in \mathbb{N},$$

lo cual implica que $P_n \ll P_0^*$ para toda $n \in \mathbb{N}$ (de lo contrario $D(P_n||P_0^*) = \infty$ y así $D(P_n||Q) = \infty$, lo cual no puede suceder pues $D(P_n||Q) < \infty$). Así, $P_n \ll P_0^*$ y como $P_n(M) = 1$ para toda n cf. (2.27), entonces se sigue que

$$\frac{dP_0^*}{dQ} \quad \text{es positiva en todo M} \quad [Q], \tag{2.30}$$

pues si n tiende a infinito entonces $M \cap K_n$ converge a M. Como $P_n \in \mathcal{P}_n$ entonces $M \cap K_n \subset \{x|\frac{dP_n}{dQ}(x) > 0\}$; así, existe un $n_0 \in \mathbb{N}$ tal que para toda $n \geq n_0$ $M \subset \{x|\frac{dP_n}{dQ}(x) > 0\}$, i.e, $\frac{dP_n}{dQ}$ es positiva en todo M $[Q]$. Ahora, si suponemos que existe algún $B \subset M$ tal que $Q(B) > 0$ y $\frac{dP_0^*}{dQ}(B) = 0$ entonces $P_0^*(B) = 0$ luego $P_n(B) = 0$, es decir,

$$P_n(B) = \int_B \frac{dP_n}{dQ}\ dQ = 0,$$

lo cual sucede sólo si $\frac{dP_n}{dQ}(x) = 0$ para toda $x \in B$ $(Q(B) > 0)$, que claramente es una contradicción. Luego $\frac{dP_0^*}{dQ}$ es positiva en todo M [Q].

Observemos que para toda $P \in \Pi$ tal que $P \ll Q$ satisface que $P \ll P_0^*$, esto ya que si $B \subset M$ es tal que $P_0^*(B) = 0$, luego si $Q(B) = 0$ entonces $P(B) = 0$ $(P \ll Q)$, si $Q(B) > 0$ entonces

$$P_0^*(B) = \int_B \frac{dP_0^*}{dQ}\ dQ = 0.$$

Por (2.30) se tiene que $\frac{dP_0^*}{dQ}$ es positiva en B, por lo cual este caso no puede suceder, y por ende $P(B) = 0$ si $P_0^*(B) = 0$, i.e., $P \ll P_0^*$. Como $\int x_i dP \geq 0$ con $i = 1, ..., k$, se sigue de (2.29) que

$$\int \log\Big(\frac{dP_0^*}{dQ}\Big)\ dP = \int D_0 + \sum_{i=1}^{k} \zeta_i^* x_i\ dP(x)$$

$$= D_0 + \sum_{i=1}^{k} \zeta_i^* \int x_i\ dP(x) \geq D_0;$$

donde $P \ll P_0^*$ nos permite quitar el símbolo $[P_0^*]$ al tomar la integral respecto a P. Por el Lema 2.2.1 se tiene que

$$D_0 \leq \inf_{P \in \Pi \cap \Lambda_Q} \int \log\Big(\frac{dP_0^*}{dQ}\Big)\ dP \leq D = D(\Pi||Q).$$

Si $P_0^* \neq P^*$ entonces la desigualdad es estricta y si $P_0^* = P^*$ entonces el ínfimo de las integrales es igual a $D(\Pi||Q)$. Como $\Pi_0 \subset \Pi$ entonces $D \leq D_0$

pues recordemos que $D = D(\Pi||Q) = \text{ínf}_{P\in\Pi} D(P||Q)$ y $D_0 = D(\Pi||Q) = \text{ínf}_{P\in\Pi_0} D(P||Q)$, así obtenemos que

$$D \leq D_0 \leq \inf_{P\in\Pi\cap\Lambda_Q} \int \log\Big(\frac{dP_0^*}{dQ}\Big) \quad dP \leq D,$$

luego entonces necesariamente $D = D_0$ y $P_0^* = P^*$. Ahora, de (2.29) y como todo lo anterior se hizo tomando en cuenta al subespacio M cf. (2.28), se sigue que

$$\frac{dP^*}{dQ}(x) = \begin{cases} e^{D+\sum_{i=1}^k \zeta_i^* x_i} & si \quad x \in M \\ 0 & \text{en otro caso.} \end{cases} \tag{2.31}$$

En donde $\zeta^* = (\zeta_1^*, ..., \zeta_k^*) \in \mathbb{R}_+^k$.

Observemos que si existe una medida de probabilidad $P \in \Pi$ tal que $P \equiv Q$, i.e., $P \ll Q$ y $Q \ll P$ entonces tenemos que $Q(M) = 1$ pues como $P(M) = 1$ entonces $P(\mathbb{R}^k \setminus M) = 0$ luego $Q(\mathbb{R}^k \setminus M) = 0$ y así $Q(M) = 1$. Observamos que todas las restricciones que pusimos respecto a M fueron para poder obtener que $\frac{dP^*}{dQ}$ es positiva en M y acotada, i.e., para que $\log \frac{dP^*}{dQ}$ esté definido en todo M; si ponemos $M = \mathbb{R}^k$ como $Q(M) = 1$ y $P \equiv Q$ entonces $\frac{dP^*}{dQ}$ es positiva en $\mathbb{R}^k$ luego todo lo anterior también es válido para $M = \mathbb{R}^k$. De esta forma se tiene que como $f(x) = \sum_{i=1}^k \zeta_i x_i$ es medible y $\int f dP = \int \sum_{i=1}^k \zeta_i x_i dP(x) = \sum_{i=1}^k \zeta_i \int x_i dP(x) \geq 0$ con $(\zeta_1, ..., \zeta_k) \in \mathbb{R}_+^k$ por el Lema 2.2.1 para cualquier $f \in \mathcal{F}$ se tiene que

$$-\log \int e^f \ dQ \leq D(\Pi||Q), \tag{2.32}$$

si la igualdad se da entonces

$$\frac{dP^*}{dQ}(x) = \left[\int e^f \ dQ\right]^{-1} e^{f(x)}.$$

Así, como $f(x) = \sum_{i=1}^k \zeta_i x_i$ con $\zeta = (\zeta_1, ...\zeta_k) \in \mathbb{R}_+^k$, entonces podemos manipular a $\vartheta = (\vartheta_1, ..., \vartheta_k)$ para que se de la igualdad en (2.32), pues variando los valores de las entradas de ϑ tenemos que la integral

$$\int e^{\sum_{i=1}^j \vartheta_i x_i} dQ(x)$$

toma todos los valores de $(0, 1]$ y $-\log(y)$ toma todos lo valores de $[0, \infty)$. De esta manera tenemos que P^* pertenece a la familia exponencial $\{P_\zeta | \zeta \in \Theta\}$ con

$$\frac{dP_\zeta}{dQ}(x) = \frac{e^{\sum_{i=1}^k \zeta_i x_i}}{\int e^{\sum_{i=1}^k \zeta_i x_i} dQ(x)}, \quad \Theta = \left\{\zeta \Big| \int e^{\sum_{i=1}^k \zeta_i x_i} dQ(x) < \infty\right\}. \tag{2.33}$$

Observamos que este último resultado es compatible con (2.31) ya que aplicando la función exponencial de ambos lados en (2.32) obtenemos que

$$\left[\int e^f \ dQ\right]^{-1} \leq e^D.$$

Y, como dijimos anteriormente, si se da la igualdad tenemos en (2.31) que

$$\frac{dP^*}{dQ}(x) = e^{D+\sum_{i=1}^k \zeta_i^* x_i} = e^D e^{\sum_{i=1}^k \zeta_i^* x_i} = \left[\int e^{\sum_{i=1}^k \zeta_i^* x_i} dQ(x)\right]^{-1} e^{\sum_{i=1}^k \zeta_i^* x_i}.$$

Es decir, P^* pertenece a la familia exponencial si $M = \mathbb{R}^k$. Ahora, para terminar esta parte de la demostración tenemos que mostrar que si P^* pertenece a la familia exponencial con las propiedades de (2.33) entonces existe $P \in \Pi$ tal que $P \equiv Q$, lo cual es sencillo pues si el subespacio lineal de R^k más pequeño que cumple que $P(M) = 1$ es R^k entonces P^* pertenece a la familia exponencial, es decir, esta familia es no vacía; así, existe un ζ tal que $\frac{dP_\zeta}{dQ}$ es positiva en todo $\mathbb{R}^k$ salvo en un conjunto de medida cero relativo a Q pues

$$\frac{dP_\zeta}{dQ}(x) = \frac{e^{\sum_{i=1}^k \zeta_i x_i}}{\int e^{\sum_{i=1}^k \zeta_i x_i} dQ(x)} > 0 \quad \forall \zeta \in \Theta.$$

Entonces, supongamos que Q no es absolutamente continua respecto a P_ζ, i.e., existe algún B medible tal que $P_\zeta(B) = 0$ y $Q(B) > 0$, esto nos da que

$$P_\zeta(B) = \int_B \frac{dP_\zeta}{dQ} \ dQ = 0,$$

pero como $Q(B) > 0$ entonces la única manera en que puede suceder lo anterior es que $\frac{dP_\zeta}{dQ}(B) = 0$ lo cual es una contradicción. Por lo que necesariamente existe una medida de probabilidad P tal que $P \ll Q$ y $Q \ll P$, i.e., $P \equiv Q$.

Notemos que (2.31) nos da que

$$D = -\log\left[\int e^{\sum_{i=1}^k \zeta_i^* x_i} dQ\right] \quad \text{siempre y cuando} \quad P_{\zeta^*} = P^*.$$

Aplicamos (2.5) del Lema 2.2.1 a la función $f(x) = \sum_{i=1}^k \zeta_i x_i$ con $\zeta = (\zeta_1, ..., \zeta_k) \in \mathbb{R}_+^k$ y obtenemos que

$$-\log\left[\int e^{\sum_{i=1}^{k}\zeta_i x_i}dQ\right] \leq D \quad \text{para cada} \quad \zeta \in \mathbb{R}^k_+.$$

Además la desigualdad es estricta a menos que $P_\zeta = P^*$ y en este caso se da la igualdad, por lo tanto tenemos que si $P \equiv Q$ para alguna $P \in \Pi$ entonces

$$D = \sup_{\zeta\in\mathbb{R}^k_+}\left\{-\log\left[\int e^{\sum_{i=1}^{k}\zeta_i x_i}dQ\right]\right\},$$

donde el supremo se alcanza si y sólo si $P_\zeta = P^*$. Por lo tanto hemos probado el teorema para el caso $S = \mathbb{R}^k$ y $f_i(x_1, ..., x_k) = x_i$, $i = 1, ..., k$ luego, por la observación hecha al principio de la demostración, hemos probado el teorema de manera general.

□

El Teorema 2.3.1 que acabamos de demostrar caracteriza a la I-proyección generalizada de cualquier medida de probabilidad Q tal que $P \ll Q$ para un conjunto de medidas de probabilidad que cumplen que dada una colección finita de funciones medibles la integral de cualquiera de estas funciones es no negativa. Ahora nuestro cometido es caracterizar la I-proyección generalizada en cierto tipo de conjuntos pero trabajando ahora con S como un espacio vectorial topológico localmente convexo y $\mathcal{B}$ la σ-algebra de Borel. Antes de enunciar este segundo teorema es necesario dar una definición adicional.

Definición 2.3.3. *Una medida de probabilidad $P \in \Lambda(V)$ es convexa-tensa*[10] *si existen subconjuntos medibles $\{K_n\}_{n\in\mathbb{N}} \subset V$ tales que $K_n \subset K_{n+1}$, K_n es compacto y convexo para toda n y $P(K_n)$ converge a 1 cuando n tiende a infinito.*

Observación. *Una observación importante es que si V es un espacio de Fréchet, es decir, es un espacio vectorial topológico metrizable y completo ([6] pág. 353) y además es separable entonces toda medida de probabilidad sobre V es convexa-tensa, cf. [11].*

Ahora ya estamos en condiciones de enunciar el segundo teorema de gran importancia en este capítulo.

Teorema 2.3.2. *Sean V un espacio localmente convexo, Q una medida de probabilidad sobre $(V, \mathcal{B})$ convexa-tensa y $C \subset V$ convexo tal que*

$$int(C) \cap conv(sop(Q)) \neq \emptyset.$$

[10] cf. El concepto de medida regular ó medida tensa en [16].

Sea $\{K_n\}_{n\in\mathbb{N}}$ *una sucesión de conjuntos que hacen a* Q *convexa-tensa y consideremos los siguiente conjuntos:*

$$\Pi(C) = \{P|E(P) \in C\} \text{ y } \Pi_0(C) = \{P|P(K_n) = 1 \text{ para algún } n\} \cap \Pi(C). \tag{2.34}$$

Sea $D = D(\Pi_0(int(C))||Q)$. *Entonces se tiene que*

$$D(\Pi(C)||Q) = D(\Pi_0(int(C))||Q) = D < \infty \tag{2.35}$$

y P^*, *la I-proyección generalizada en común de* Q *en* $\Pi(C)$ *y* $\Pi_0(int(C))$ *(cf. Lema 2.2.2), pertenece a la familia exponencial* $\{P_\vartheta|\vartheta \in \Theta\}$ *definida mediante las densidades*

$$\frac{dP_\vartheta}{dQ}(v) = \frac{e^{\vartheta(v)}}{\int e^{\vartheta} dQ}, \quad \Theta = \left\{\vartheta|\vartheta \in V', \ \int e^{\vartheta} \ dQ < \infty\right\}. \tag{2.36}$$

Más aún,

$$D = \sup_{\vartheta\in V'}\left[\inf_{v\in C} \vartheta(v) - \log\left(\int e^{\vartheta} \ dQ\right)\right]. \tag{2.37}$$

En donde el supremo se alcanza si y sólo si $P_\vartheta = P^*$.

Observación. *Observemos que el teorema es invariante bajo traslaciones. En efecto, si* C *es un conjunto convexo definimos*

$$C' = C + v_0 = \{v' \in V|v' = v + v_0 \quad v \in C\}$$

con $v_0 \in V$ *fijo y la medida de probabilidad* $Q'(B) = Q(B - v_0)$ *para todo* $B \subset V$ *medible (*Q' *está bien definida y es una medida de probabilidad). Observemos que* $sop(Q') = sop(Q) + v_0$. *Se tiene que si existe* $P \ll Q$ *entonces* $P' \ll Q'$ *con* $P'(B) = P(B - v_0)$ *y así se tiene la siguiente relación*

$$\int_{B-v_0} \frac{dP}{dQ} \ dQ = P(B - v_0) = P'(B) = \int_B \frac{dP'}{dQ'} \ dQ'.$$

Como $Q(B - v_0) = Q'(B)$ *entonces tenemos que la función* f *medible definida como*

$$f(v) = \frac{dP}{dQ}(v - v_0)$$

cumple con la definición de derivada de Radon-Nikodým de P' *respecto a* Q', *como es única relativa a* Q' *se tiene entonces que*

$$\frac{dP'}{dQ'}(v) = \frac{dP}{dQ}(v - v_0) \quad [Q'].$$

Ahora, observemos que $E(P) = v$ *si y sólo si* $E(P') = v + v_0$ *pues si* $E(P) = v$ *para cada* $\vartheta \in V'$ *se tiene*

$$\vartheta(v) = \int \vartheta \ dP,$$

luego

$$\int \vartheta \ dP' = \int \vartheta(x) \ dP'(x) = \int \vartheta(x) \ dP(x - v_0) = \int \vartheta(y + v_0) \ dP(y)$$

$$= \int \vartheta(y) dP(y) + \int \vartheta(v_0) dP(y) = \vartheta(v) + \vartheta(v_0) = \vartheta(v + v_0).$$

De manera similar si $E(P') = v + v_0$ *se tiene que* $E(P) = v$*, es decir, tenemos que* $E(P) \in C$ *si y sólo si* $E(P') \in C + v_0$*. Por lo anterior se tiene entonces que*

$$D(\Pi(C)||Q) = D(\Pi(C + v_0)||Q')$$

y

$$D(\Pi_0(int(C))||Q) = D(\Pi_0(int(C + v_0))||Q')$$

pues

$$D(P||Q) = \int \log\Big(\frac{dP}{dQ}\Big) dP = \int \log\Big(\frac{dP'}{dQ'}\Big) dP' = D(P'||Q')$$

para toda $P \in \Pi(C)$ *y para toda* $P' \in \Pi(C + v_0)$*. Por último, se tiene que* P^* *pertenece a* $\{P_\vartheta | \vartheta \in \Theta\}$ *definido por*

$$\frac{dP_\vartheta}{dQ}(v) = \frac{e^{\vartheta(v)}}{\int e^{\vartheta} dQ}, \quad \Theta = \left\{ \vartheta | \vartheta \in V', \quad \int e^{\vartheta} \ dQ < \infty \right\},$$

si y sólo si $P'^* \in \{P'_\vartheta | \vartheta \in \Theta\}$*, definido por*

$$\frac{dP'_\vartheta}{dQ'}(v) = \frac{e^{\vartheta(v)}}{\int e^{\vartheta} dQ'}, \quad \Theta = \left\{ \vartheta | \vartheta \in V', \quad \int e^{\vartheta} \quad dQ' < \infty \right\},$$

ya que si $P'^* \in \{P_\vartheta | \vartheta \in \Theta\}$ *entonces*

$$\frac{dP'^*}{dQ}(v) = \frac{dP^*}{dQ}(v - v_0) = \frac{e^{\vartheta(v - v_0)}}{\int e^{\vartheta(x)} dQ(x)} = \frac{e^{\vartheta(v)} e^{-\vartheta(v_0)}}{\int e^{\vartheta(y - v_0)} dQ'(y)}$$

$$= \frac{e^{-\vartheta(v_0)}}{e^{-\vartheta(v_0)}} \frac{e^{\vartheta(v)}}{\int e^{\vartheta(y)} dQ'(y)} = \frac{e^{\vartheta(v)}}{\int e^{\vartheta(y)} dQ'(y)},$$

es decir, P'^* pertenece a la familia $\{P'_\vartheta|\vartheta \in \Theta\}$. De manera similar se tiene que si $P'^* \in \{P'_\vartheta|\vartheta \in \Theta\}$ entonces $P^* \in \{P_\vartheta|\vartheta \in \Theta\}$.

Todo lo anterior nos da que efectivamente el teorema es invariante bajo traslaciones. Esto se utilizará (más adelante) en la demostración.

Demostración. Primero probaremos que

$$D(\Pi_0(int(C))||Q) = D < \infty \text{ y } P^* \equiv Q, \tag{2.38}$$

donde P^* denota la I-proyección generalizada de Q en $\Pi_0(int(C))$.

Por hipótesis tenemos que $int(C) \cap conv(sop(Q)) \neq \emptyset$, entonces existe $v_0 \in int(C)$ tal que $v_0 \in conv(sop(Q))$, i.e.,

$$v_0 = \sum_{i=1}^{k} \alpha_i v_i, \quad \alpha_i \geq 0 \quad i = 1, ...k, \quad \sum_{i=1}^{k} \alpha_i = 1, \quad v_i \in sop(Q) \quad \forall i.$$

Observamos que como v_i está en el soporte de Q entonces toda vecindad de v_i tiene Q-medida positiva, i.e., $Q(v_i + U) > 0$ con U vecindad del vector 0 en V. Tomemos ahora una vecindad cerrada y convexa U de 0 tal que $v_0 + U \subset int(C)$, observemos que como $Q(K_n) \to 1$ y $v_i \in sop(Q)$ para toda i entonces existe una n_0 para la cual los conjuntos

$$A_i = (v_i + U) \cap K_{n_0}$$

tienen Q-medida positiva para toda i. Además, los conjuntos A_i son compactos pues $A_i \subset K_{n_0}$, A_i es cerrado y K_{n_0} compacto, también son conjuntos convexos al ser intersección de conjuntos convexos. Consideremos la medida de probabilidad

$$P_0 = \sum_{i=1}^{k} \alpha_i R_i \text{ con } R_i(\cdot) = \frac{Q(\cdot \cap A_i)}{Q(A_i)}.$$

Como es combinación lineal convexa de medidas de probabilidad en efecto es medida de probabilidad. Entonces para $n \geq n_0$

$$E(P_0) \in int(C), \quad D(P_0||Q) < \infty \quad \text{y} \quad P_0(K_n) = 1. \tag{2.39}$$

La primera propiedad se sigue de lo siguiente: $E(R_i) \in v_i + U$ ya que $v_i + U$ es un subconjunto cerrado de un conjunto compacto $(v_i + U \subset K_n)$ para algún n, por lo que $v_i + U$ es compacto y además es convexo. Por otro lado

$$R_i(v_i + U) = \frac{Q(v_i + U \cap A_i)}{Q(A_i)} = \frac{Q(A_i)}{Q(A_i)} = 1,$$

por lo tanto se tiene que $E(R_i) \in v_i + U$ (cf. Observación de la Definición 2.3.3) luego observemos que $E(P_0) = \sum_{i=1}^{k} \alpha_i E(R_i)$ y como $v_0 = \sum_{i=1}^{k} \alpha_i v_i$ se tiene que $E(P_0) = \sum_{i=1}^{k} \alpha_i E(R_i) \in v_0 + U$.

Para lo segundo mostraremos que $P_0 \ll Q$, para ello consideramos $B \subset V$ medible tal que $Q(B) = 0$, por como está definida P_0 hay que mostrar que $R_i(B) = 0$ para toda i y por como está definida cada R_i hay que mostrar $Q(B \cap A_i) = 0$ para toda i, lo cual es claro pues $B \cap A_i \subset B$ y

$$Q(B \cap A_i) \leq Q(B) = 0.$$

Ahora, tenemos que

$$D(P_0||Q) = D\left(\sum_{i=1}^{k} \alpha_i R_i \Big|\Big| Q\right) \leq \sum_{i=1}^{k} \alpha_i D(R_i||Q) < \infty$$

pues $D(R_i||Q) < \infty$ para cada $i = 1, ..., k$, esto último ya que como $R_i \ll Q$ se tiene

$$D(R_i||Q) = \int \log\left(\frac{dR_i}{dQ}\right) dR_i = \frac{1}{Q(A_i)} \int_{A_i} \log\left(\frac{dR_i}{dQ}\right) dQ$$

$$\leq \frac{1}{Q(A_i)} \int_{A_i} \frac{dR_i}{dQ}\ dQ = \frac{R_i(A_i)}{Q(A_i)} = \frac{1}{Q(A_i)} < \infty \quad \forall i = 1, ..., k.$$

La última propiedad se da ya que

$$P_0(K_n) = \sum_{i=1}^{k} \alpha_i R_i(K_n) = \sum_{i=1}^{k} \alpha_i \frac{Q(K_n \cap (v_i + U) \cap K_{n_0})}{Q((v_i + U) \cap K_{n_0})}$$

$$\sum_{i=1}^{k} \alpha_i \frac{Q((v_i + U) \cap K_{n_0})}{Q((v_i + U) \cap K_{n_0})} = \sum_{i=1}^{k} \alpha_i = 1.$$

Con lo anterior tenemos una medida de probabilidad P_0 tal que $P_0(K_n) = 1$ con $n \geq n_0$, $P_0 \in \Pi(int(C))$ ($E(P_0) \in int(C)$), es decir, $P_0 \in \Pi_0(int(C))$ y además $D(P_0||Q) < \infty$, por lo tanto $D(\Pi_0(int(C))||Q) < \infty$ luego existe P^* la I-proyección generalizada de Q en $\Pi_0(int(C))$. Para terminar la prueba de (2.38) restaría mostrar que $P^* \equiv Q$; ya se tiene que $P^* \ll Q$ por la definición de I-proyección generalizada pues $D(P^*||Q) = D(\Pi_0(int(C))||Q)$, falta mostrar que $Q \ll P^*$ para ello definimos las siguientes medidas de probabilidad:

$$P_n = (1 - \beta_n)P_0 + \beta_n Q_n, \quad \text{con} \quad Q_n(\cdot) = \frac{Q(\cdot \cap K_n)}{Q(K_n)}, \quad n \geq n_0.$$

Observemos que si $\beta_n \in [0,1]$ es suficientemente pequeña, es decir si β_n converge a 0 cuando n tiende a infinito, P_n satisface las propiedades enunciadas en (2.39) para algún n. En efecto, en primer lugar

$$P_n(K_n) = (1-\beta_n)P_0(K_n) + \beta_n Q_n(K_n) = (1-\beta_n)\cdot 1 + \beta_n \frac{Q(K_n)}{Q(K_n)} = 1.$$

En segundo lugar observemos que $E(P_n)$ siempre existe y $E(P_n) \in K_n$ pues $P_n(K_n) = 1$ y K_n es compacto y convexo, entonces se tiene que si $\vartheta \in V'$

$$\vartheta(E(P_n)) = \int \vartheta \ dP_n = \int \vartheta \ d([1-\beta_n]P_0 + \beta_n Q_n)$$

$$= (1-\beta_n)\int \vartheta \ dP_0 + \beta_n \int \vartheta \ dQ_n = (1-\beta_n)\vartheta(E(P_0)) + \beta_n \int \vartheta \ dQ_n,$$

luego si β_n converge a 0 cuando cuando n tiende a infinito se tiene que

$$\lim_{n\to\infty} \vartheta(E(P_n)) = \lim_{n\to\infty} \int \vartheta \ dP_n = \lim_{n\to\infty} (1-\beta_n)\vartheta(E(P_0))$$

$$+ \lim_{n\to\infty} \beta_n \int \vartheta \ dQ_n = \vartheta(E(P_0)) + \lim_{n\to\infty} \beta_n \int \vartheta \ dQ_n.$$

Es necesario verificar que $\beta_n \int \vartheta \ dQ_n$ converge a 0 cuando n tiende a infinito, para ello es suficiente mostrar que el límite de $\int \vartheta \ dQ_n$ cuando n tiende a infinito existe y que dicho límite no es infinito. Lo cual es sencillo de ver pues por como está definida Q_n se tiene que Q_n converge a Q de manera fuerte cuando n tiende a infinito por lo que $\int \vartheta \ dQ_n$ converge a $\int \vartheta \ dQ$ cuando n tiende a infinito y como Q es convexa-tensa entonces $E(Q)$ siempre existe pues por definición de convexa-tensa tenemos que $Q(K_n)$ converge a 1 cuando n tiende a infinito, por lo que existe un K_n que es convexo y compacto tal que $|Q(K_n) - 1| < \epsilon$ luego $E(Q)$ existe y $E(Q) \in K_n$, así $\int \vartheta \ dQ_n$ converge a $\int \vartheta \ dQ = \vartheta(E(Q)) < \infty$, i.e., $\beta_n \int \vartheta \ dQ_n$ converge a 0 cuando n tiende a infinito. Por lo tanto tenemos que para toda $\epsilon > 0$ existe $N \in \mathbb{N}$ tal que para toda $n > N$

$$|\vartheta(E(P_n)) - \vartheta(E(P_0))| < \epsilon.$$

Ahora como $E(P_0) \in int(C)$ que es abierto entonces existe $\epsilon' > 0$ para el cual $B_{\epsilon'}(E(P_0)) \subset int(C)$. Por lo tanto, para este ϵ' existe $N_{\epsilon'}$ tal que $E(P_n) \in B_{\epsilon'}(E(P_0))$ para toda $n > N_{\epsilon'}$, es decir, $E(P_n) \in int(C)$ si $n > N_{\epsilon'}$.

Por último, como la entropía relativa es convexa se tiene que

$$D(P_n||Q) = D([1-\beta_n]P_0 + \beta_n Q_n||Q) \leq (1-\beta_n)D(P_0||Q) + \beta_n D(Q_n||Q) < \infty$$

pues se probó que $D(P_0||Q) < \infty$ y $D(Q_n||Q) < \infty$, la prueba de esto último es análoga a la prueba del hecho $D(R_i||Q) < \infty$. En efecto, como $Q_n \ll Q$ se tiene

$$D(Q_n||Q) = \int \log\left(\frac{dQ_n}{dQ}\right) dQ_n = \frac{1}{Q(K_n)} \int_{K_n} \log\left(\frac{dQ_n}{dQ}\right) dQ$$

$$\leq \frac{1}{Q(K_n)} \int_{K_n} \frac{dQ_n}{dQ}\ dQ = \frac{Q_n(K_n)}{Q(K_n)} = \frac{1}{Q(K_n)} < \infty \quad \forall n \geq n_0.$$

Por lo que efectivamente se tiene que $E(P_n) \in int(C)$, $D(P_n||Q) < \infty$ y $P_n(K_n) = 1$ para alguna n, a saber para toda $n > M = \text{máx}\{n_0, N_{\epsilon'}\}$. Ahora, como $E(P_n) \in int(C)$ y $P_n(K_n) = 1$ entonces $P_n \in \Pi_0(int(C))$ por (2.2) tenemos que

$$D(P_n||Q) \geq D(P_n||P^*) + D(\Pi_0(int(C))||Q),$$

lo que nos da que $P_n \ll P^*$ ya que de lo contrario $D(P_n||P^*) = \infty$ y entonces $D(P_n||Q) = \infty$ lo cual, como acabamos de ver, no puede suceder. Además, si $P_n(B) = 0$ entonces necesariamente $Q_n(B) = 0$, i.e., $Q_n \ll P_n$. De lo anterior se obtiene que

$$Q_n \ll P_n \ll P^* \ll Q \quad \text{para cada} \quad n \geq M.$$

Como $Q(K_n) \to 1$ cuando n tiende a infinito tenemos entonces que $Q_n \to Q$ de manera fuerte cuando n tiende a infinito, es decir, tenemos que

$$Q = \lim_{n\to\infty} Q_n \ll P^* \ll Q.$$

Por lo tanto $P^* \equiv Q$ como se quería demostrar.

Hasta ahora hemos mostrado que $D(\Pi_0(int(C))||Q) < \infty$. Ahora debemos mostrar que $D(\Pi(C)||Q) = D(\Pi_0(int(C))||Q)$. Para ello mostraremos que se dan las dos desigualdades, y utilizaremos una aproximación a $D(\Pi(C)||Q)$ mediante $\Pi_0(cd(C))$. Notemos que gracias a la observación hecha al principio de la demostración del teorema podemos suponer sin pérdida de generalidad que $0 \in int(C)$, más aún, podemos suponer que $E(P_0) = 0$ ya que si no es así, es decir, $E(P_0) = v_0 \in int(C)$ basta considerar la traslación $int(C) - v_0$. Ahora, por el teorema biopolar (cf. Teorema 4.3.4 del apéndice) tenemos que

$$cd(C) = C^{\circ\circ} = \{v | \vartheta(v) \leq 1 \quad \forall \vartheta \in C^\circ\},$$

con

$$C^\circ = \{\vartheta \in V' | \vartheta(v) \leq 1 \quad \forall v \in C\}. \tag{2.40}$$

Éste último conjunto es el conjunto polar de C respecto a la dualidad $\langle V, V' \rangle$[11]. El teorema de Alaoglu-Bourbaki[12] nos garantiza que C° es compacto en la topología débil de V', i.e., en la topología de la convergencia puntual de los funcionales lineales (cf. Lema 4.3.1 y Teorema 4.2.1 del apéndice).

Para poder utilizar el Lema 2.2.5 debemos recordar la notación utilizada en la sección anterior: Recordemos que si $\mathcal{F}$ es una familia de funciones medibles y $\mathcal{K} = \{K_i\}_{i=1}^{\infty}$ una sucesión de conjuntos tales que K_i es medible, $K_i \subset K_{i+1}$ y cada $f \in \mathcal{F}$ estaba acotada en K_i $i = 1, 2, ...$ entonces designábamos

$$\Pi(\mathcal{F}|\mathcal{K}) = \bigcup_{i=1}^{\infty} \Pi(\mathcal{F}|K_i) \quad \text{y} \quad \Pi'(\mathcal{F}|\mathcal{K}) = \bigcup_{i=1}^{\infty} \Pi'(\mathcal{F}|K_i), \tag{2.41}$$

con

$$\Pi(\mathcal{F}|K_i) = \left\{ P \middle| \int f \ dP \geq 0 \quad \forall f \in \mathcal{F} \quad \text{y} \quad P(K_i) = 1 \right\}$$

y

$$\Pi'(\mathcal{F}|K_i) = \left\{ P \middle| \int f \ dP > 0 \quad \forall f \in \mathcal{F} \quad \text{y} \quad P(K_i) = 1 \right\}.$$

Observemos que la condición $E(P) \in cd(C)$ es equivalente a pedir que

$$a - a \int \vartheta \ dP \geq 0 \quad \forall \vartheta \in C^\circ, \quad a \geq 0 \tag{2.42}$$

y $E(P) \in int(C)$ es equivalente a

$$a - a \int \vartheta \ dP > 0 \quad \forall \vartheta \in C^\circ, \quad a \geq 0. \tag{2.43}$$

Para ver que esto sucede recordemos que

$$E(P) = v_0 \Leftrightarrow \vartheta(v_0) = \int \vartheta \ dP \quad \forall \vartheta \in V',$$

entonces si $a \geq 0$ y $v_0 = E(P) \in cd(C)$ por (2.40) se tiene que $v_0 \in C^{\circ\circ}$, i.e., $\vartheta(v_0) \leq 1$ para todo $\vartheta \in C^\circ$. Así, $v_0 = E(P) \in cd(C)$ si y sólo si $\int \vartheta \ dP \leq 1$ para todo $\vartheta \in C^\circ$ si y sólo si $-a \int \vartheta \ dP \geq -a$ si y sólo si $a - a \int \vartheta \ dP \geq 0$

[11] cf. Definición 4.3.4 y el desarrollo posterior en el apéndice.
[12] cf. Teorema 4.3.5 del apéndice.

para todo $\vartheta \in C^\circ$. Para el caso en que $v_0 = E(P) \in int(C)$ la prueba es completamente análoga sólo que ahora se tiene

$$int(C) = \{v | \vartheta(v) < 1 \ \ \forall \vartheta \in C^\circ\}.$$

Observamos que si $f \in \{f | f = a(1-\vartheta), \ \ a \geq 0 \ \ y \ \ \vartheta \in C^\circ\}$ entonces

$$\int f \ \ dP = \int a(1-\vartheta) \ \ dP = a - a \int \vartheta \ \ dP.$$

Ahora, notemos que $\{f | f = a(1-\vartheta), \quad a \geq 0 \quad \vartheta \in C^\circ\}$ es una familia de funciones medibles y además es un cono convexo ya que si

$$\{f_i\}_{i=1}^k \subset \{f | f = a(1-\vartheta), \ \ a \geq 0 \ \ \vartheta \in C^\circ\} \ \ y \ \ \alpha_i \geq 0, \ \ i = 1, ..., k$$

se tiene que

$$\sum_{i=1}^k \alpha_i f_i = \sum_{i=1}^k \alpha_i a_i (1 - \vartheta_i) = \sum_{i=1}^k \alpha_i a_i - \sum_{i=1}^k \alpha_i a_i \vartheta_i =$$

$$\sum_{i=1}^k \alpha_i a_i \left[1 - \sum_{i=1}^k \frac{\alpha_i a_i}{\sum_{j=1}^k \alpha_j a_j} \vartheta_i \right]$$

y observemos que

$$\sum_{i=1}^k \frac{\alpha_i a_i}{\sum_{j=1}^k \alpha_j a_j} \vartheta_i(v) \leq \sum_{i=1}^k \frac{\alpha_i a_i}{\sum_{j=1}^k \alpha_j a_j} = 1 \quad \forall v \in C,$$

por lo que si definimos

$$a = \sum_{i=1}^k \alpha_i \quad \text{y} \quad \vartheta = \sum_{i=1}^k \frac{\alpha_i a_i}{\sum_{j=1}^k \alpha_j a_j} \vartheta_i,$$

entonces $a \geq 0$, $\vartheta \in C^\circ$ y

$$\sum_{i=1}^k \alpha_i f_i = a(1 - \vartheta).$$

Tomemos $\mathcal{F} = \{f | f = a(1-\vartheta), \quad a \geq 0 \quad \vartheta \in C^\circ\}$ y $\mathcal{K}$ una sucesión de conjuntos que hacen a Q convexa-tensa (enunciada en las hipotésis). Notemos que cada $f \in \mathcal{F}$ es acotada en cada K_i pues K_i es compacto para cada i; como las condiciones $E(P) \in cd(C)$ y $E(P) \in int(C)$ son equivalentes a (2.42) y (2.43) respectivamente, entonces

$$\Pi_0(cd(C)) = \{P | E(P) \in cd(C), \ \ P(K_n) = 1 \ \text{ p.a. } \ n\}$$

$$=\left\{P\Big|\int f\ dP\geq 0\quad \forall f\in\mathcal{F},\ P(K_n)=1\quad \text{p.a. } n\right\}$$

y

$$\Pi_0(int(C))=\{P|E(P)\in int(C),\quad P(K_n)=1\quad \text{p.a. } n\}$$

$$=\left\{P\Big|\int f\ dP>0\quad \forall f\in\mathcal{F},\ P(K_n)=1\quad \text{p.a. } n\right\}.$$

Así, $\Pi_0(cd(C))$ y $\Pi_0(int(C))$ pueden ser representados de la siguiente manera (cf.(2.41)):

$$\Pi_0(cd(C))=\Pi(\mathcal{F}|\mathcal{K})\quad \text{y}\quad \Pi_0(int(C))=\Pi'(\mathcal{F}|\mathcal{K}). \tag{2.44}$$

Recordemos que $D(\Pi(\mathcal{F}|\mathcal{K})||Q)=D(\Pi'(\mathcal{F}|\mathcal{K})||Q)=D$ (cf. Lema 2.2.4) por el Lema 2.2.2 se tiene que la I-proyección generalizada de Q en $\Pi_0(cd(C))$ y en $\Pi_0(int(C))$ es la misma, a la cual denotaremos como P^*. Ahora, el Lema 2.2.5 nos garantiza que

$$\log\left(\frac{dP^*}{dQ}\right)=D+\lim_{n\to\infty}a_n(1-\vartheta_n)\qquad [P^*],$$

pero ya sabemos que $P^*\equiv Q$ (cf. (2.38)) por lo que es lo mismo que

$$\log\left(\frac{dP^*}{dQ}\right)=D+\lim_{n\to\infty}a_n(1-\vartheta_n)\qquad [Q]. \tag{2.45}$$

Donde $a_n\geq 0$, $\vartheta_n\in C^\circ$, $n=1,2,...$

Queremos mostrar que P^* pertenece a la familia exponencial (2.36) para ello debemos analizar que sucede con el límite en (2.45). Primero observemos que como $\vartheta_n\in C^\circ$ para toda n y este conjunto es compacto entonces existe $\{\vartheta_{n_k}\}_k$ una subsucesión que converge, digamos a $\vartheta_0\in C^\circ$, en la topología débil (que es la topología de convergencia puntual). Así,

$$\lim_{n\to\infty}a_n(1-\vartheta_n(v))=\lim_{k\to\infty}a_{n_k}(1-\vartheta_{n_k}(v))\qquad [Q]$$

Primero notemos que si $\lim_{k\to\infty}a_{n_k}=0$ entonces

$$\lim_{k\to\infty}a_{n_k}(1-\vartheta_{n_k}(v))=0\qquad [Q].$$

Por lo que

$$\log\left(\frac{dP^*}{dQ}\right)=D\Leftrightarrow\frac{dP^*}{dQ}=e^D,$$

luego

$$1 = P^*(V) = \int_V \frac{dP^*}{dQ} \quad dQ = \int e^D \quad dQ = e^D \Leftrightarrow D = 0;$$

lo cual implica que

$$\frac{dP^*}{dQ} = 1.$$

Es decir, $P^* = Q$ y así P^* pertenece a la familia exponencial con $\vartheta = 0$. Ahora, como el producto $a_{n_k}(1-\vartheta_{n_k}(v))$ converge y ϑ_{n_k} converge a ϑ_0 cuando k tiende a infinito $[Q]$, entonces la única manera[13] en que puede suceder que el límite cuando k tiende a infinito de a_{n_k} no exista o sea igual a infinito es si $1 - \vartheta_{n_k}$ converge a 0 $[Q]$, i.e., si $\vartheta_0(v) = 1$ $[Q]$ pero esto último no puede suceder. En efecto, supongamos que sucede, esto es, $\vartheta_0(v) = 1$ para todo $v \in V \setminus N$ en donde $Q(N) = 0$, pero observemos que como $0 = E(P_0)$ y por la definición de resultante se tiene que

$$\int \vartheta \ dP_0 = \vartheta(0) \quad \forall \vartheta \in V',$$

en particular se cumple para ϑ_0, i.e.,

$$\int \vartheta_0 \ dP_0 = \vartheta_0(0).$$

Como $P_0 \ll Q$ ($D(P_0||Q) < \infty$, cf.(3.42)) y $Q(N) = 0$ entonces $P_0(N) = 0$, además $\vartheta_0(0) = 0$ por ser funcional lineal, así

$$0 = \vartheta_0(0) = \int_V \vartheta_0 \ dP_0 = \int_{V\setminus N} \vartheta_0 \ dP_0 = \int_{V\setminus N} 1 \ dP_0 = 1$$

lo cual es una contradicción. Entonces a_{n_k} converge a $a_0 \geq 0$ cuando k tiende a infinito. Así, se tiene que

$$\lim_{n\to\infty} a_n(1 - \vartheta_n) = \lim_{k\to\infty} a_{n_k}(1 - \vartheta_{n_k}) = a_0(1 - \vartheta_0) \qquad [Q],$$

es decir,

$$\log\left(\frac{dP^*}{dQ}\right) = D + a_0(1 - \vartheta_0) \qquad [Q] \quad a_0 \geq 0, \quad \vartheta_0 \in C^\circ; \tag{2.46}$$

y así,

$$\frac{dP^*}{dQ} = e^D e^{a_0(1-\vartheta_0)}.$$

[13] cf. Lema 4.1.12 del apéndice.

Por (2.5) del Lema 2.2.1, para la función medible $f = a_0(1-\vartheta_0)$ se tiene que

$$D \geq -\log\left[\int e^{a_0(1-\vartheta_0)}dQ\right]. \tag{2.47}$$

Si probamos que de hecho se da la igualdad en (2.47) entonces por el Lema 2.2.1 tendríamos que P^* pertenece a la familia exponencial $\{P_\vartheta|\vartheta \in \Theta\}$ con $\vartheta = -a_0\vartheta_0$, i.e., $\frac{dP^*}{dQ} = \frac{e^\vartheta}{\int e^\vartheta dQ}$. Procedemos entonces a probar que se da la igualdad en (2.47). En efecto, pues de (2.46) se sigue que

$$a_0(1-\vartheta_0) = \log\left(\frac{dP^*}{dQ}\right) - D \qquad [Q],$$

como la igualdad se da casi donde sea relativo a Q se tiene que

$$-\log\left[\int e^{a_0(1-\vartheta_0)}dQ\right] = -\log\left[\int e^{\log(\frac{dP^*}{dQ})-D}dQ\right] = -\log\left[e^{-D}\int \frac{dP^*}{dQ}\ dQ\right]$$

$$= -\log[e^{-D}\cdot 1] = D.$$

Por lo tanto, P^* pertence a la familia exponencial mencionada.

Observamos que como $\vartheta_0 \in C^\circ$ entonces $\vartheta_0(v) \leq 1$ para toda $v \in C$, cf. (2.40), luego $-a_0 \leq -a_0\vartheta_0(v)$ para toda $v \in C$, si ponemos $\vartheta^* = -a_0\vartheta_0$ tenemos que $-a_0 \leq \vartheta^*(v)$ para toda $v \in C$ y así

$$D = -\log\left[\int e^{a_0(1-\vartheta_0)}dQ\right] = -a_0 - \log\left[\int e^{\vartheta^*}dQ\right]$$

$$\leq \inf_{v\in C}\vartheta^*(v) - \log\left[\int e^{\vartheta^*}dQ\right],$$

esto es,

$$D \leq \inf_{v\in C}\vartheta^*(v) - \log\left[\int e^{\vartheta^*}dQ\right]. \tag{2.48}$$

Por otro lado, consideremos $\vartheta \in V'$ que cumpla

$$-\infty < \inf_{v\in C}\vartheta(v)$$

y al aplicar (2.5) a la función medible $f_\vartheta = \vartheta - \text{ínf}_{v\in C}\,\vartheta(v)$ obtenemos

$$D(\Pi(C)||Q) \geq -\log\left[\int e^{\vartheta - \text{ínf}_{v\in C}\,\vartheta(v)} dQ\right] = \inf_{v\in C}\vartheta(v) - \log\left[\int e^{\vartheta} dQ\right]. \tag{2.49}$$

Esta desigualdad se da para toda $\vartheta \in V'$ que cumpla $-\infty < \text{ínf}_{v\in C}\,\vartheta(v)$, en particular para ϑ^*, por lo que (2.48) y (2.49) nos dan que

$$D(\Pi_0(int(C))||Q) = D \leq D(\Pi(C)||Q),$$

con la desigualdad trivial $D(\Pi(C)||Q) \leq D$ $(\Pi_0(int(C)) \subset \Pi(C))$ llegamos a que $D(\Pi(C)||Q) = D(\Pi_0(int(C))||Q) < \infty$, como se quería probar.

El Lema 2.2.2 nos da que la I-proyección generalizada de Q en $\Pi(C)$ y la I-proyección generalizada de Q en $\Pi_0(int(C))$ son la misma. Más aún, observamos que (2.49) nos indica que si la igualdad se da entonces $D(\Pi(C)||Q)$ es un máximo, i.e.,

$$D(\Pi(C)||Q) = \max_{\vartheta\in V'}\left\{\inf_{v\in C}\vartheta(v) - \log\left[\int e^{\vartheta} dQ\right]\right\}.$$

Entonces, (2.6) del Lema 2.2.1 nos da que $\frac{dP_\vartheta}{dQ}$ es en realidad la Q-densidad de la I-proyección generalizada de Q en $\Pi(C)$, es decir, $P_\vartheta = P^*$. Y viceversa, si $P^* = P_\vartheta$ entonces la igualdad se da y $D(\Pi(C)||Q)$ es un máximo, es decir, de nuevo se tiene que

$$D(\Pi(C)||Q) = \max_{\vartheta\in V'}\left\{\inf_{v\in C}\vartheta(v) - \log\left[\int e^{\vartheta} dQ\right]\right\}.$$

□

Antes de pasar al último capítulo nos gustaría hacer algunas observaciones. Primero, es interesante notar que fue necesario utilizar los dos conjuntos $\Pi_0(int(C))$ y $\Pi_0(cd(C))$, el primero para la desigualdad trivial $D(\Pi(C)||Q) \leq D$ utilizada en la última parte de la demostración pues no hay manera de obtenerla comparando con $\Pi_0(cd(C))$ ya que este último conjunto no nos puede proporcionar dicha información. Por otro lado, $\Pi_0(cd(C))$ fue útil al utilizar el teorema bipolar, ya que nos proporciona una descripción analítica de este conjunto mediante el espacio dual de V y la manipulación de los funcionales lineales. Con lo anterior vemos por qué fue necesario obtener estás dos propiedades juntas, así como la igualdad $D(\Pi_0(int(C))||Q) = D(\Pi_0(cd(C))||Q)$.

Capítulo 3

Teoremas límite

Una de las suposiciones más fuertes a la hora de trabajar con procesos aleatorios es la de independencia, por ejemplo: hacer la suposición de que el proceso no dependa de todo lo que ha ocurrido antes o lo que ocurrirá después. Claramente no todos los fenómenos se comportan de tal manera por lo que es necesario comenzar a estudiar procesos con algún tipo de dependencia. En primer lugar se comienza por estudiar los procesos de Markov cuya dependencia está definida sólo en términos del pasado inmediato. Otro tipo de dependencia es el concepto de cuasiindependecia asintótica que se define más adelante. Los resultados principales que presentamos son concernientes a este concepto, a la propiedad de Sanov (definida también más adelante y cuya base es el teorema de Sanov) y a la representación de la I-proyección generalizada. Los resultados y conceptos son tomados principalmente de [11]. Haremos un uso extensivo del concepto de topología débil en el conjunto de medidas de probabilidad por lo cual referimos a la Definición 4.3.4 y a todo el desarrollo ulterior del apéndice. Para los conceptos de teoría de la probabilidad y teoría de la medida referimos al lector a [4] y [25].

Comencemos con un poco de terminología y definiciones que utilizaremos a lo largo de este capítulo. De nuevo sea $(S, \mathcal{B})$ un espacio de medida arbitrario, el cual se dejará fijo en lo consecutivo y Λ el conjunto de todas las medidas de probabilidad sobre $(S, \mathcal{B})$.

Definición 3.0.1. *Un conjunto $A \in \mathcal{B}$ es un átomo si $\mu(A) > 0$ y para cualquier subconjunto medible $B \subset A$ con $\mu(B) < \mu(A)$ sucede que $\mu(B) = 0$.*

Definición 3.0.2. *Denotamos por $\Lambda_f \subset \Lambda$ al conjunto de todas las medidas de probabilidad atómicas con un número finito de átomos, i.e., $P \in \Lambda_f$ si y sólo si para cada $B \in \mathcal{B}$ se tiene*

$$P(B) = \sum_{i=1}^{k} \alpha_i 1_B(s_i) \quad s_i \in S, \quad \alpha_i > 0, \quad i = 1, ..., k \quad y \quad \sum_{i=1}^{k} \alpha_i = 1,$$

cf. Lemna 4.4.11 del apéndice.

Consideremos $\{X_n\}_{n\in\mathbb{N}}$, una sucesión de variables aleatorias independientes[1] que toman valores en S y tienen distribución común P_X. Nos fijamos en el n-ésimo producto cartesiano de $(S, \mathcal{B}, P_X)$ como el espacio muestral del vector aleatorio $X^n = (X_1, ..., X_n)$, en donde $\mathcal{B}^n$ es la σ-álgebra producto y las probabilidades de los eventos en términos de X^n determinados formalmente por la medida producto $P_X^n = \otimes_{i=1}^n P_X$. Por otro lado, ya hemos introducido la medida empírica al principio de la sección 1.3 limitándonos al caso infinito numerable; ahora, siguiendo un camino similar, debemos dar una definición acorde a nuestros propósitos.

Definición 3.0.3. *Consideremos la siguiente función* $\hat{P}_n : S^n \times \mathcal{B} \to \mathbb{R}^+ \cup \{0\}$ *definida por*

$$\hat{P}_n(x, B) = \frac{1}{n}\sum_{i=1}^{n} 1_B(x_i) \quad x = (x_1, ..., x_n) \in S^n, \quad B \in \mathcal{B}.$$

Si tomamos $\hat{P}_n(X^n, \cdot)$ *a la función se le conoce como medida empírica o distribución empírica del vector aleatorio* X^n *y es un elemento aleatorio de* Λ*, si consideramos la muestra* $s = (s_1, ..., s_n) \in S^n$*, a la función* $\hat{P}_n(s, \cdot) \in \Lambda$ *definida por*

$$\hat{P}_n(s, B) = \frac{1}{n}\sum_{i=1}^{n} 1_B(s_i) \quad B \in \mathcal{B}$$

se le conoce como la media muestral.

Denotaremos a la medida empírica y a la media muestral de la siguiente manera: $\hat{P}_n(X^n, \cdot) = \hat{P}_{X^n}(\cdot)$ y $\hat{P}_n(s, \cdot) = \hat{P}_{s,n}(\cdot)$ respectivamente. Considerando a la muestra $s = (s_1, ..., s_n)$ notemos que para cualquier conjunto de medidas de probabilidad $\xi \subset \Lambda$ la probabilidad de que la distribución empírica $\hat{P}_{X^n}$ se encuentre en ξ es por definición:

$$\mathbb{P}(\hat{P}_{X^n} \in \xi) = P_X^n(A_n) \quad A_n = \{s | \hat{P}_n(s, \cdot) \in \xi\}.$$

La última probabilidad está bien definida siempre y cuando $A_n \in \mathcal{B}^n$. Si $A_n \notin \mathcal{B}^n$ consideraremos lo que se conoce como la probabilidad superior y la probabilidad inferior:

$$\bar{\mathbb{P}}(\hat{P}_X \in \xi) = \text{ínf}\{P_X^n(C) | C \in \mathcal{B}^n, \quad A_n \subset C\}$$

y

$$\underline{\mathbb{P}}(\hat{P}_X \in \xi) = \sup\{P_X^n(C) | C \in \mathcal{B}^n, \quad C \subset A_n\}.$$

[1] Para la definición general de variables aleatorias independientes véase [16], pág. 252.

Observación. *Notemos que lo anterior son la medida exterior e interior en el conjunto potencia de S^n inducidas o generadas por P_X^n. Esto ya que toda cuasimedida genera una medida exterior y una medida interior, en particular una medida de probabilidad es una cuasimedida. cf. [18] capítulo 7.*

3.1. Teorema de Sanov (Versión general)

Hasta ahora hemos obtenido el teorema de Sanov para medidas de probabilidad sobre espacios a lo más infinito numerables; además, nos preguntamos sobre la probabilidad de que la medida empírica se encuentre fuera de la bola de radio a y centro en ρ. Como podemos notar nos hemos restringido a casos muy sencillos. Ahora es momento de analizar dos cuestiones y obtener el resultado correspendiente: la primera es, naturalmente, ¿qué sucede con las medidas sobre espacios medibles arbitrarios? (incluyendo aquellos cuya cardinalidad es mayor a la cardinalidad de los naturales) y la segunda ¿qué sucede si consideramos conjuntos más generales en lugar de bolas? Es decir ¿qué sucede si la medida empírica cae fuera de un conjunto arbitrario que contenga a la medida en la que estamos interesados? Lo que veremos ahora es que el resultado obtenido también se cumple con algunas modificaciones para estos casos. El teorema es enunciado de una manera ligeramente diferente, además utilizaremos la Definición (1.5.1) de la entropía relativa. Para poder enunciar el teorema de la manera en que nos interesa debemos introducir algunas definiciones ya que ahora al enfrentarnos a espacios arbitrarios no será posible identificar al conjunto de medidas de probabilidad sobre dichos espacios con algún simplex ni con algún subconjunto de $\mathbb{R}^n$ o $\mathbb{R}^\infty$ (en donde la topología euclideana es muy amigable) lo anterior nos pone en una situación un poco más desafiante ya que entonces es necesario definir una nueva topología en nuestro conjunto de medidas de probabilidad. El siguiente teorema, la idea general de la demostración y los conceptos concernientes a ellos son tomados de [12].

Recordemos que Λ denota al conjunto de todas las medidad de probabilidad sobre S, y Θ denota al conjunto de todas las particiones finitas y medibles de S, i.e.,

$$\Theta = \{A = (A_1, ..., A_k) | \text{A es partición medible de S}, \quad k \in \mathbb{N}\}$$

Ahora debemos equipar a Λ con una topología que haga continuas a todas las funciones

$$\Gamma_B :\to [0,1], \ \ \Gamma_B(P) = P(B) \ \text{ con } \ B \in \mathcal{B} \ \text{ y } \ P \in \Lambda;$$

la topología más gruesa que hace esto es la topología débil respecto a la

familia de proyecciónes $\{\Gamma_B | B \in \mathcal{B}\}$[2]. Llamaremos a esta topología τ[3]. Esta topología es definida mediante los abiertos básicos de la siguiente manera (cf. [11]):

Definición 3.1.1. *Para $P \in \Lambda$, $A \in \Theta$ y $\epsilon > 0$ definimos $U(P, A, \epsilon) = \{Q \in \Lambda \big| |P(A_i) - Q(A_i)| < \epsilon,\ i = 1, ..., k\}$. Se define la topología τ como la generada por $\{U(P, A, \epsilon) | P \in \Lambda,\ A \in \Theta\ y\ \epsilon > 0\}$, es decir, el anterior conjunto es una base para esta topología.*

Sin embargo, es posible trabajar con una topología más débil y lo cual nos será de gran utilidad.

Definición 3.1.2. *Se define la topología τ_0 en Λ como la topología generada por los abiertos*

$$U_0(P, A, \epsilon)$$
$$= \{Q \in \Lambda \big| |P(A_i) - Q(A_i)| < \epsilon,\ Q(A_i) = 0\ si\ P(A_i) = 0,\ i = 1, ..., k\},$$

es decir, $\{U_0(P, A, \epsilon) | P \in \Lambda,\ A \in \Theta\ y\ \epsilon > 0\}$ es una base para τ_0.

Claramente $U_0(P, A, \epsilon) \in \tau$ para toda partición medible A y para todo $\epsilon > 0$ por lo que $\tau_0 \subset \tau$. Denotaremos al interior y a la cerradura de un conjunto $\xi \subset \Lambda$ respecto a las topologías τ y τ_0 como $int_\tau(\xi)$, $cd_\tau(\xi)$ e $int_{\tau_0}(\xi)$, $cd_{\tau_0}(\xi)$ respectivamente.

Teorema 3.1.1. *(Sanov) Sean $\{X_n\}_{n\in\mathbb{N}}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas sobre S con distribución común P_X y $\xi \subset \Lambda$. Entonces*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\hat{P}_{X^n} \in \xi) \geq - \inf_{P \in int_\tau(\xi)} D(P||P_X),$$

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\hat{P}_{X^n} \in \xi) \leq - \inf_{P \in cd_\tau(\xi)} D(P||P_X)$$

y $\bar{B}_\epsilon(Q) = \{P \in \Lambda | D(P||Q) \leq \epsilon\}$ ($\epsilon > 0$), las bolas inducidas por la entropía relativa, son conjuntos compactos en la topología τ.

Observemos que, considerando la muestra $s = (s_1, ..., s_n)$, ambas cotas se ven de la siguiente manera:

$$\liminf_{n\to\infty} \frac{1}{n} \log P_X^n\{s = (s_1, ..., s_n) | \hat{P}_{s,n} \in \xi\} \geq - \inf_{P \in int_\tau(\xi)} D(P||P_X) \quad (3.1)$$

$$\limsup_{n\to\infty} \frac{1}{n} \log P_X^n\{s = (s_1, ..., s_n) | \hat{P}_{s,n} \in \xi\} \leq - \inf_{P \in cd_\tau(\xi)} D(P||P_X) \quad (3.2)$$

[2] cf. Definición 4.3.5 y el desarrollo posterior del apéndice.

[3] En [24] se estudia esta topología y se hace una comparación respecto a otras topologías iniciales.

Observación. *Si* $\{s = (s_1, ..., s_n)|\hat{P}_{s,n} \in \xi\}$ *no se encuentra dentro de la* σ*-álgebra entonces al igual que antes tomamos la probabilidad superior respecto a* P_X *en (3.1) y la probabilidad inferior respecto a* P_X *en (3.2), de este modo tenemos una versión todavía más general del teorema de Sanov.*

Observación. *En (3.2) la cota inferior puede ser reducida sustituyendo* $int_\tau(\xi)$ *por* $int_{\tau_0}(\xi)$*, ya que como* $int_\tau(\xi) \subset int_{\tau_0}(\xi)$ *entonces*

$$\inf_{P \in int_{\tau_0}(\xi)} D(P||P_X) \leq \inf_{P \in int_\tau(\xi)} D(P||P_X),$$

es decir,

$$-\inf_{P \in int_\tau(\xi)} D(P||P_X) \leq -\inf_{P \in int_{\tau_0}(\xi)} D(P||P_X).$$

La prueba que mostramos incluye este hecho.

Antes de comenzar con la demostración necesitamos dos lemas estándares en teoría de la información, ambos son lemas combinatorios y usualmente se demuestran para probar la versión finita del teorema de Sanov, en nuestro caso lo hicimos de una manera distinta y por ello procedemos a probarlos en este instante. Los resultados son tomados de [14].

Consideremos por un instante un conjunto finito $K = \{a_1, ..., a_k\}$ y consideremos al conjunto $\mathcal{L}_n(K)$ de distribuciones de probabilidad sobre K que son de la siguiente manera

$$P = \left(\frac{n_1}{n}, ..., \frac{n_k}{n}\right), \quad n_i \in \mathbb{Z}^+ \cup \{0\}.$$

Para una $P \in \mathcal{L}_n(K)$ denotamos por $\mathcal{T}_n(P)$ al conjunto de sucesiones de longitud n compuestas por elementos de K en donde cada $a_i \in \{a_1, ..., a_k\}$ aparece n_i veces, es decir, $\mathcal{T}_n(P) = \{x \in K^n|\ L_n(x, \cdot) = P(\cdot)\}$[4].

Observación. *Los conjuntos* $\mathcal{L}_n(K)$ *representan los histogramas posibles asociados al conjunto K.*

Observación. *Para facilitar la notación durante la prueba de los siguientes dos lemas denotaremos* $L_{x,n}(\cdot) = L_n(x, \cdot)$ *con* $x = (x_1, ..., x_n) \in K^n$ *fijo. Recordemos también la notación utilizada en el primer capítulo(cf. Definición 1.3.1): si* $X^n = (X_1, ..., X_n)$ *es un vector aleatorio denotamos* $L_n(X^n, \cdot) = L_n(\cdot)$.

Una observación importante es que si damos cualquier vector en el conjunto $\mathcal{T}_n(P)$ entonces dicho conjunto consta de todas las posibles permutaciones de las entradas del vector dado, esto ya que precisamente nos interesa que $L_{x,n} = P$ es decir que cada $a_i \in K$ aparezca n_i veces, entonces para

[4] c.f. la Definición 1.3.1 del capítulo 1.

obtener todo $\mathcal{T}_n(P)$ basta obtener un vector que este en dicho conjunto y permutar sus entradas.

Lema 3.1.1. $|\mathcal{L}_n(K)| \leq (n+1)^k$

Prueba. Observemos que si $P \in \mathcal{L}_n$ entonces cada entrada de P pertenece al conjunto $\{\frac{0}{n}, \frac{1}{n}, ..., \frac{n}{n} = 1\}$ cuya cardinalidad es n+1, como P es medida de probabilidad sobre K (cuya cardinalidad es k) entonces dicha cardinalidad determina cuántas entradas tendrá el vector de probabilidad P, en este caso k. Así, tenemos que en la primera entrada hay n+1 posibilidades, para la segunda lo mismo y así hasta la entrada k, por lo tanto

$$|\mathcal{L}_n(K)| \leq (n+1)^k.$$

□

Antes de enunciar el segundo lema introducimos un concepto utilizado en gran medida dentro de la teoría de la información, a saber la entropía de una medida de probabilidad. En nuestro caso, como sólo lo utilizaremos en la prueba del siguiente lema, definiremos la entropía solamente para vectores de probabilidad. Sobre el concepto y sus propiedades referimos al lector a [14]. Para facilitar la notación utilizaremos la siguiente nomeclatura: si P es una medida de probabilidad sobre un conjunto M y $\{X_n\}_{n\in\mathbb{N}}$ una sucesión de variables aleatorias sobre M con distribución común P entonces si p es una propiedad $P(\{x|x \in p\}) = P(p)$. Por ejemplo, consideramos a $M = K^n$ $P^n(\{\omega \in M|X^n(\omega) = x\}) = P^n(X^n = x)$.

Observemos que utilizamos $P^n(X^n = x)$ en lugar de $\mathbb{P}(X^n = x)$ ya que la medida de probabilidad $\mathbb{P}$ es la inducida por las variables aleatorias y su distribución P.

Definición 3.1.3. *Para un vector de probabilidad ν sobre $K = \{a_1, ..., a_k\}$ $\nu = (\nu_1 = \nu(a_1), ..., \nu_k = \nu(a_k))$. Definimos su entropía como*

$$H(\nu) = -\sum_{i=1}^{k} \nu_i \log(\nu_i).$$

Para facilitar la notación denotamos al soporte de ν como

$$\Delta_\nu = \{a_i \in K|\nu(a_i) > 0\} \subset K.$$

Lema 3.1.2. *Para $P \in \mathcal{L}_n(K)$ y cualquier otra medida de probabilidad Q sobre K se tiene que*

$$(n+1)^{-k} e^{-nD(P||Q)} \leq Q^n(\mathcal{T}_n(P)) \leq e^{-nD(P||Q)}.$$

Prueba. Antes que nada debemos verificar dos cosas:

(i) Si $X^n = (X_1, ..., X_n)$ es un vector aleatorio tal que $X \in K^n$ y X_i son variables aleatorias independientes con distribución común Q, $x \in \mathcal{T}_n(P)$ con $P \in \mathcal{L}_n(K)$ entonces se tiene

$$Q^n(X^n = x) = e^{-n[H(P)+D(P||Q)]}.$$

(ii)Para cada $P \in \mathcal{L}_n$

$$(n+1)^{-k} e^{nH(P)} \leq |\mathcal{T}_n(P)| \leq e^{nH(P)}.$$

Para (i) notamos que la medida empírica L_n se concentra en medidas de probabilidad $P \in \mathcal{L}_n$ tales que $\Delta_P \subset \Delta_Q$ ya que si en algún momento existe algún $a_j \in K$ tal que $Q(a_j) = 0$ entonces lo eliminamos del conjunto pues no produce ninguna información extra y consideramos al conjunto $\{a_1, ..., a_k\} \setminus \{a_j\}$, así cada entrada de L_n debe de ser distinta de cero con probabilidad 1. Por lo que podemos asumir sin pérdida de generalidad que $L_{x,n} = P$ para alguna $P \in \mathcal{L}_n$ con $\Delta_P \subset \Delta_Q$; además, esto último nos garantiza que $P \ll Q$ luego $D(P||Q) < \infty$. Observemos que como $x \in \mathcal{T}_n(P)$ entonces $L_{x,n} = P = (\frac{n_1}{n}, ..., \frac{n_k}{n})$, es decir, $x_i = a_j$ para algún $a_j \in K$, $i = 1, ..., n$ y cada $a_i \in K$ aparece n_i veces en la secuencia, así

$$Q^n((X_1, ..., X_n) = (x_1, ..., x_n)) = Q(X_1 = x_1)...Q(X_n = x_n)$$

$$= Q(a_1)^{n_1} Q(a_2)^{n_2} ... Q(a_k)^{n_k} = \prod_{i=1}^{k} Q(a_i)^{n_i} = \prod_{i=1}^{k} Q(a_i)^{n\frac{n_i}{n}} = \prod_{i=1}^{k} Q(a_i)^{nP(a_i)}$$

$$= e^{-n[H(P)+D(P||Q)]}.$$

En donde la última igualdad se sigue de la siguiente identidad:

$$H(P) + D(P||Q) = -\sum_{i=1}^{k} P(a_i) \log(P(a_i)) + \sum_{i=1}^{k} P(a_i) \log\left(\frac{P(a_i)}{Q(a_i)}\right)$$

$$= -\sum_{i=1}^{k} P(a_i) \log(P(a_i)) + \sum_{i=1}^{k} P(a_i) \log(P(a_i)) - \sum_{i=1}^{k} P(a_i) \log(Q(a_i))$$

$$= -\sum_{i=1}^{k} P(a_i) \log(Q(a_i)).$$

Así, tenemos que

$$e^{-n[H(P)+D(P||Q)]} = e^{-n[-\sum_{i=1}^{k} P(a_i)\log(Q(a_i))]} = \prod_{i=1}^{k} e^{\log(Q(a_i))^{nP(a_i)}}$$

$$= \prod_{i=1}^{k} Q(a_i)^{nP(a_i)},$$

por lo tanto

$$Q^n(X^n = x) = e^{-n[H(P)+D(P||Q)]}.$$

El particular, como $D(Q||Q) = 0$ se sigue que para cualquier vector aleatorio $X^n = (X_1, ..., X_n)$ en donde las X_i son independientes, tienen distribución común $Q \in \mathcal{L}_n$ y $x \in \mathcal{T}_n(Q)$ se tiene que

$$Q^n(X^n = x) = e^{-nH(Q)}. \tag{3.3}$$

Ahora, para probar (ii) primero observemos que todas las posibles realizaciones en $\mathcal{T}_n(P)$ tienen la misma probabilidad de ocurrir (de hecho es claro al ver que $x \in \mathcal{T}_n(P)$ era arbitrario y $Q^n(X^n = x)$ no depende de x); así, si $x \in \mathcal{T}_n(P)$ utilizando (3.3) se tiene que

$$e^{-nH(P)}|\mathcal{T}_n(P)| = P^n(X^n = x)|\mathcal{T}_n(P)| = P^n[X^n = (X_1, ..., X_n) \in \mathcal{T}_n(P)]$$

$$= P^n(L_n = P) \leq 1,$$

luego

$$|\mathcal{T}_n(P)| \leq e^{nH(P)}.$$

De esta forma obtenemos la cota superior.

Para probar la cota inferior damos $P' \in \mathcal{L}_n$ tal que $\Delta_{P'} \subset \Delta_P$, observemos que $P^n(L_n = P') > 0$ sólo si $P' \in \mathcal{L}_n$ y $\Delta_{P'} \subset \Delta_P$ ya que si $P' \notin \mathcal{L}_n$ entonces siempre sucede que $L_n \neq P$ pues L_n es un elemento aleatorio de $\mathcal{L}_n$. Por otro lado, si $\Delta_{P'}$ no es subconjunto de Δ_P entonces existe a_i tal que $P'(a_i) > 0$ y $P(a_i) = 0$, luego $L_n \neq P'$ ya que cada X_i se distribuye bajo P y así nunca aparecerá a_i en la secuencia. Por el mismo argumento podemos suponer sin pérdida de generalidad que $\Delta_P = \{1, ..., k\} = K$. Así, tenemos que

$$\frac{P^n(L_n = P)}{P^n(L_n = P')} = \frac{P^n(X \in \mathcal{T}_n(P))}{P^n(X \in \mathcal{T}_n(P'))} = \frac{|\mathcal{T}_n(P)|e^{-nH(P)}}{|\mathcal{T}_n(P')|e^{-nH(P')}}$$

$$= \frac{|\mathcal{T}_n(P)| \prod_{i=1}^k P(a_i)^{nP(a_i)}}{|\mathcal{T}_n(P')| \prod_{i=1}^k P(a_i)^{nP'(a_i)}} = \prod_{i=1}^k \frac{(nP'(a_i))!}{(nP(a_i))!} P(a_i)^{nP(a_i)-nP'(a_i)}.$$

Esta última igualdad ya que como $\mathcal{T}_n(P)$ consiste de todos los vectores x de longitud n que cumplen que $L_{x,n} = P$ y $P \in \mathcal{L}_n$ es decir $P = (\frac{n_1}{n}, ..., \frac{n_k}{n})$, entonces cada a_i que es entrada de x se repite n_i veces. Entonces tenemos que las permutaciones de $x \in \mathcal{T}_n(P)$ son las permutaciones de un conjunto múltiple de n elementos (tenemos n entradas en el vector) donde cada elemento a_i tiene multiplicidad n_i, dicha cantidad nos la da el coeficiente multinomial, es decir,

$$|\mathcal{T}_n(P)| = \frac{n!}{n_1!...n_k!}.$$

De lo anterior se sigue entonces que

$$\frac{|\mathcal{T}_n(P)|}{|\mathcal{T}_n(P')|} = \frac{\frac{n!}{\prod_{i=1}^k (n_i!)}}{\frac{n!}{\prod_{i=1}^k (n_i'!)}} = \prod_{i=1}^k \frac{n_i'!}{n_i!},$$

en donde $n_i' = P'(a_i)$. Así,

$$\frac{|\mathcal{T}_n(P)| \prod_{i=1}^k P(a_i)^{nP(a_i)}}{|\mathcal{T}_n(P')| \prod_{i=1}^k P(a_i)^{nP'(a_i)}} = \prod_{i=1}^k \frac{n_i'!}{n_i!} P(a_i)^{nP(a_i)-nP'(a_i)}$$

$$= \prod_{i=1}^k \frac{(n\frac{n_i'}{n})!}{(n\frac{n_i}{n})!} P(a_i)^{nP(a_i)-nP'(a_i)} = \prod_{i=1}^k \frac{(nP'(a_i))!}{(nP(a_i))!} P(a_i)^{nP(a_i)-nP'(a_i)}.$$

En esto último el factor del producto es una expresión de la forma $\frac{r!}{s!}(\frac{s}{n})^{s-r}$ con $r = (nP'(a_i))$ y $s = (nP(a_i))$. Para proseguir mostraremos que $s^{r-s} \leq \frac{r!}{s!}$ $\forall r, s \in \mathbb{Z}_+$, para ello consideramos dos casos:

Caso 1. $s \leq r$.

Como $s \leq r$ entonces existe $k \in \mathbb{Z}_+ \cup \{0\}$ tal que $k < s$ y $s = r - k$, luego

$$\frac{r!}{s!} = \frac{r(r-1)...(r-k)!}{(r-k)!} = \prod_{i=0}^{k-1}(r-i) \geq \prod_{i=0}^{k-1} s = s^k = s^{r-(r-k)} = s^{r-s}.$$

Caso 2. $r \leq s$.

Análogamente, como $r \leq s$ entonces existe $k \in \mathbb{Z}_+ \cup \{0\}$ tal que $k < r$ y $r = s - k$, luego

$$\frac{r!}{s!} = \frac{(s-k)!}{s(s-1)...(s-k)!} = \frac{1}{\prod_{i=0}^{k-1}(s-i)} \geq \frac{1}{\prod_{i=0}^{k-1} s} = \frac{1}{s^k} = \frac{1}{s^{s-r}} = s^{r-s}$$

Dicho esto tenemos que $\frac{r!}{s!}(\frac{s}{n})^{s-r} \geq s^{r-s}(\frac{s}{n})^{s-r} = \frac{1}{n^{s-r}} = n^{r-s}$ y así se sigue que

$$\frac{P^n(L_n = P)}{P^n(L_n = P')} = \prod_{i=1}^{k} \frac{(nP'(a_i))!}{(nP(a_i))!} P(a_i)^{nP(a_i)-nP'(a_i)} \geq \prod_{i=1}^{k} n^{nP'(a_i)-nP(a_i)}$$

$$= n^{n[\sum_{i=1}^{k} P'(a_i) - \sum_{i=1}^{k} P(a_i)]} = n^{n(1-1)} = 1.$$

Lo anterior lleva a que para toda $P, P' \in \mathcal{L}_n$ tal que $\Delta_{P'} \subset \Delta_P$ se tiene que

$$P^n(L_n = P') \leq P^n(L_n = P),$$

entonces

$$1 = \sum_{P' \in \mathcal{L}_n} P^n(L_n = P') = |\mathcal{L}_n| P^n(L_n = P') \leq |\mathcal{L}_n| P^n(L_n = P)$$

$$= |\mathcal{L}_n| e^{-nH(P)} |\mathcal{T}_n(P)|,$$

es decir,

$$|\mathcal{L}_n|^{-1} e^{nH(P)} \leq |\mathcal{T}_n(P)|.$$

Por el Lema 3.1.1 tenemos la cota inferior:

$$(n+1)^{-k} e^{nH(P)} \leq |\mathcal{T}_n(P)|.$$

Ahora, continuando con la prueba del lema, tenemos que por (i)

$$Q^n(L_n = P) = |\mathcal{T}_n(P)| Q^n(X^n = x) = |\mathcal{T}_n(P)| e^{-n[H(P)+D(P||Q)]}$$

con $x \in \mathcal{T}_n(Q)$, y de (ii) se sigue que

$$(n+1)^{-k} e^{nH(P)} e^{-n[H(P)+D(P||Q)]} \leq Q^n(L_n = P) \leq e^{nH(P)} e^{-n[H(P)+D(P||Q)]},$$

esto es,

$$(n+1)^{-k} e^{-nD(P||Q)} \leq Q^n(L_n = P) \leq e^{-nD(P||Q)}.$$

Como $Q^n(L_n = P) = Q^n(\{x | x \in \mathcal{T}_n(P)\}) = Q^n(\mathcal{T}_n(P))$ entonces se sigue el resultado.

□

Habiendo mostrado los dos lemas anteriores proseguimos con la demostración del teorema.

Demostración (Sanov). Durante la demostración utilizaremos la representación original de la entropía relativa definida en el capítulo 1[5]. A saber,

$$D(P||Q) = \sup_{A\in\Theta} D(P^A||Q^A).$$

En donde $P^A = (P(A_1), ..., P(A_n))$ y $Q^A = (Q(A_1), ..., Q(A_n))$. También, para facilitar la notación, tomaremos al conjunto K utilizado en los lemas anteriores como $K = \{1, ..., k\}$. La estrategia para la demostración es utilizar los resultados que ya conocemos para conjuntos finitos, a saber los lemas recien mostrados, mediante las particiones finitas $A \in \Theta$, identificando a una partición $A = (A_1, ..., A_k)$ con el conjunto finito $K = \{1, ..., k\}$ para todo $k \in \mathbb{N}$; de este modo obtener desigualdades de tal manera que al realizar el proceso límite[6] obtengamos las cotas de nuestro interés.

Procederemos de la siguiente manera:

(i) Mostramos la cota inferior de (3.1) con int_{τ_0}.

(ii) Mostramos la compacidad en la topología τ de las bolas inducidas por la entropía relativa.

(iii) Finalmente mostramos la cota superior de (3.2) haciendo uso de (ii).

(i) Para lo primero afirmamos que sucede lo siguiente:

$$-D(P||P_X) \leq \liminf_{n\to\infty} \frac{1}{n} \log P_X^n(\{s|\hat{P}_{s,n} \in \xi\}), \tag{3.4}$$

para toda $P \in int_{\tau_0}(\xi)$ o equivalentemente para toda $P \in \Lambda$ tal que $U_0(P, A, \epsilon) \subset \xi$ para alguna $A = (A_1, ..., A_k) \in \Theta$ y $\epsilon > 0$. Para ver que esto sucede tomamos P con las características anteriores y sea $\{\tilde{P}_n\}_{n\in\mathbb{N}}$ una sucesión tal que $\tilde{P}_n \in \mathcal{L}_n(K)$ para cada n, $\tilde{P}_n(j) \to P(A_j)$ y que $\tilde{P}_n(j) = 0$ cuando $P(A_j) = 0$, $j = 1, ..., k$ (esto es posible ya que $P(A_j) \in [0, 1]$ para toda j y $\frac{n_j}{n} \in \mathbb{Q}^+$ el cual es denso en $\mathbb{R}^+ \cup \{0\}$, en particular es denso en $[0, 1]$). Entonces como $\tilde{P}_n(j) \to P(A_j)$ tenemos que

$$|\tilde{P}_n(j) - P(A_j)| < \epsilon_n \quad i = 1, ..., k.$$

En donde $\epsilon_n \to 0$. Luego entonces para toda n tal que $\epsilon_n \leq \epsilon$ se tiene que $\{s|\hat{P}_{s,n} \in U_0(P, A, \epsilon_n)\} \subset \{s|\hat{P}_{s,n} \in \xi\}$ ya que si $\hat{P}_{s,n} \in U_0(P, A, \epsilon_n)$ entonces

[5] cf. Definición 1.5.1.

[6] (Observamos que la entropía relativa es, de hecho, un proceso límite pues se toma un supremo.)

$\hat{P}_{s,n} \in U_0(P, A, \epsilon) \subset \xi$. De esta manera

$$P_X^n(\{s|\hat{P}_{s,n} \in \xi\}) \geq P_X^n(\{s|\hat{P}_{s,n} \in U_0(P, A, \epsilon_n)\}) \geq P_X^n(\{s|\hat{P}_{s,n}^A = \tilde{P}_n\}). \tag{3.5}$$

La última desigualdad ya que por como se escogió $\tilde{P}_n$ tenemos que $\tilde{P}_n \in U_0(P, A, \epsilon_n)$ y si agregamos la condición de que $\hat{P}_X^A = \tilde{P}_n$ entonces sucede que $\{s|\hat{P}_{s,n}^A = \tilde{P}_n\} \subset \{s|\tilde{P}_{s,n} \in U_0(P, A, \epsilon_n)\}$. Ahora bien, podemos identificar a $\{A_1, ..., A_k\}$ con $\{1, ..., k\}$ y por otra parte también podemos mandar a cada s_i a un $j \in \{1, ..., k\}$ mediante A_j, es decir, $s_i = j$ si $s_i \in A_j$, $j = 1, ..., k$, como A es una partición entonces la anterior función está bien definida. Observemos que $\hat{P}_{s,n}^A = \tilde{P}_n$ es lo siguiente

$$\hat{P}_{s,n}^A = \left(\frac{1}{n}\sum_{i=1}^n 1_{A_1}(s_i), ..., \frac{1}{n}\sum_{i=1}^n 1_{A_k}(s_i)\right) = \left(\frac{n_1}{n}, ..., \frac{n_k}{n}\right).$$

Ya que $\sum_{i=1}^n 1_{A_j}(s_i)$ cuenta las veces que $s_i \in A_j$ $j = 1, ..., k$ tenemos que cada $s_i \in \{1, ..., k\}$ aparece en la sucesión $s = (s_1, ..., s_n)$ n_j veces $j = 1, ..., k$. Es decir, s es una sucesión de longitud n donde cada elemento pertenece a $\{1, ..., k\}$ y cada $j \in \{1, ..., k\}$ aparece en la sucesión n_j veces. Por lo tanto $\{s|\hat{P}_{s,n}^A = \tilde{P}_n\} = \mathcal{T}_n(\tilde{P}_n)$ (Bajo la identificación y al mandar a cada elemento de s a un elento de K claramente). Por lo tanto tenemos que

$$P_X^n(\{s|\hat{P}_{s,n}^A = \tilde{P}_n\}) = (P_X^A)^n(\mathcal{T}_n(\tilde{P}_n)) \geq (n+1)^{-k} e^{-nD(\tilde{P}_n||P_X^A)}. \tag{3.6}$$

En donde la última desigualdad se da por el Lema 3.1.2. También aclaramos que al hacer la identificación de A con K y de los elementos de $s = (s_1, ..., s_n)$ con los elementos de K entonces utilizamos la notación $j \in K$ o A_j indistintamente, además hacemos la observación de que mediante la identificación P_X^A es una medida de probabilidad sobre K ($P_X^A(j) = P_X(A_j)$ con $j = 1, ..., k$). Más aún, guarda la información de la medida de probabilidad original P_X^n respecto al conjunto $\{s|\hat{P}_{s,n}^A = \tilde{P}_n\}$, es decir,

$$(P_X^A)^n(\mathcal{T}_n(\tilde{P}_n)) = (P_X^A)^n(\{y \in K^n | y \in \mathcal{T}_n(\tilde{P}_n)\}) = P_X^n(\{s \in S^n | \hat{P}_{s,n}^A = \tilde{P}_n\}).$$

Así, (3.5) y (3.6) nos dan que

$$P_X^n(\{s|\hat{P}_{s,n} \in \xi\}) \geq (n+1)^{-k} e^{-nD(\tilde{P}_n||P_X^A)}.$$

Entonces, al tomar logaritmo y multiplicar por n^{-1},

$$\frac{1}{n}\log P_X^n(\{s|\hat{P}_{s,n} \in \xi\}) \geq \frac{1}{n}\log\Big[(n+1)^{-k} e^{-nD(\tilde{P}_n||P_X^A)}\Big]. \tag{3.7}$$

Ahora, como $\tilde{P}_n(A_j) \to P(A_j)$ (recordemos la identificación $A_j = j$) tenemos que

$$\lim_{n\to\infty} D(\tilde{P}_n||P_X^A) = \lim_{n\to\infty} \sum_{j=1}^{k} \tilde{P}_n(A_j) \log\Big(\frac{\tilde{P}_n(A_j)}{P_X(A_j)}\Big) = \sum_{j=1}^{k} P(A_j) \log\Big(\frac{P(A_j)}{P_X(A_j)}\Big)$$

$$= D(P^A||P_X^A),$$

es decir, $D(\tilde{P}_n||P_X^A) \to D(P^A||P_X^A)$. Por último, como $D(P||P_X)$ es el supremo sobre todas las particiones se tiene que $D(P^A||P_X^A) \leq D(P||P_X)$, i.e., $-D(P||P_X) \leq -D(P^A||P_X^A)$. Así, tomando límite inferior en (3.7) tenemos que

$$\liminf_{n\to\infty} \frac{1}{n} \log P_X^n(\{s|\hat{P}_{s,n} \in \xi\}) \geq \liminf_{n\to\infty} \frac{1}{n} \left(\log[(n+1)^{-k}] + \log\big[e^{-nD(\tilde{P}_n||P_X^A)}\big] \right)$$

$$= \liminf_{n\to\infty} \frac{1}{n} \log[(n+1)^{-k}] + \liminf_{n\to\infty} -D(\tilde{P}_n||P_X^A) = -D(P^A||P_X^A) \geq -D(P||P_X).$$

Así, hemos obtenido (3.4) para toda $P \in int_{\tau_0}(\xi)$ por lo cual al tomar supremo obtenemos

$$\liminf_{n\to\infty} \frac{1}{n} \log P_X^n(\{s|\hat{P}_{s,n} \in \xi\}) \geq \sup_{P \in\ int_{\tau_0}(\xi)} -D(P||P_X)$$

$$= - \inf_{P \in int_{\tau_0}(\xi)} D(P||P_X).$$

Es decir, obtenemos (3.1).

(ii) Procedemos ahora a probar la compacidad de las bolas inducidas por la entropía relativa.

Consideremos a $\mathcal{G}$ el conjunto de funciones aditivas[7] y finitas sobre $(S, \mathcal{B})$ que toman valores en $[0, 1]$ y que cumplen que la imagen de S es 1, es decir,

$$\mathcal{G} = \{g : \mathcal{B} \to [0, 1] \big|\ \text{g es finita, aditiva y } g(S) = 1\}.$$

Notemos que si dotamos a $[0, 1]^{\mathcal{B}}$, el conjunto de todas las funciones de $\mathcal{B}$ en $[0, 1]$, con la topología producto tenemos que $\mathcal{G}$ es un subconjunto cerrado de $[0, 1]^{\mathcal{B}}$ ya que si tomamos $\{g_n\}_{n\in\mathbb{N}} \subset \mathcal{G}$ una sucesión tal que g_n converge a g cuando n tiende a infinito entonces $g \in \mathcal{G}$[8]. En efecto, basta mostrar que g es finita y aditiva pues $g(S) = 1$ ya que $1 = g_n(S) \to g(S)$ para toda n. Lo primero es fácil de ver ya que como g_n converge a g entonces para toda $\epsilon > 0$

[7] $g : \mathcal{B} \to [0, 1]$ es aditiva si $g(B_1 \cup B_2) = g(B_1) + g(B_2)$ con $B_1 \cap B_2 = \emptyset$.

[8] Recordemos que la convergencia en $[0, 1]^{\mathcal{B}}$ está dada por la convergencia puntual cf. Teorema 4.2.1 del apéndice.

existe $N \in \mathbb{N}$ tal que si $n > N$ entonces $|g_n(B) - g(B)| < \epsilon$ con $B \in \mathcal{B}$ y como cada g_n es finita entonces $g(B)$ no puede tomar los valores ∞ o $-\infty$ ya que de lo contrario tendríamos $\infty < \epsilon$ lo cual es una contradicción, luego g es finita. Para ver que g es aditiva tomamos $B_1, B_2 \in \mathcal{B}$ tales que $B_1 \cap B_2 = \emptyset$ y observamos que como g_n es aditiva para cada n entonces

$$g_n(B_1) + g_n(B_2) = g_n(B_1 \cup B_2) \to g(B_1 \cup B_2) \quad \text{cuando } n \to \infty.$$

Por otro lado

$$g_n(B_1) + g_n(B_2) \to g(B_1) + g(B_2) \quad \text{cuando } n \to \infty.$$

Así, $g(B_1 \cup B_2) = g(B_1) + g(B_2)$ y por lo tanto $g \in \mathcal{G}$, es decir, $\mathcal{G}$ es cerrado. Ahora, gracias al teorema de Tychonoff[9] sabemos que $[0,1]^{\mathcal{B}}$ es compacto por ser producto de compactos y entonces $\mathcal{G}$ es compacto por ser un subconjunto cerrado de un compacto.

Ahora, observemos que podemos extender la definición de la entropía relativa para medidas de probabilidad a $\mathcal{G}$, es decir, si $P, Q \in \mathcal{G}$ entonces definimos la entropía relativa de P respecto a Q como

$$D(P||Q) = \sup_{A \in \Theta} D(P^A||Q^A).$$

En donde

$$D(P^A||Q^A) = \sum_{i=1}^{m} P(A_i) \log\left(\frac{P(A_i)}{Q(A_i)}\right).$$

De esta forma observamos que para cualquier partición $A \in \Theta$ fija el subconjunto $\{P \in \mathcal{G} | D(P^A||Q^A) \leq \alpha\}$ de $\mathcal{G}$ es cerrado, esto ya que si tomamos $\{P_n\}_{n \in \mathbb{N}} \subset \{P \in \mathcal{G} | D(P^A||Q^A) \leq \alpha\}$ tal que $P_n(B) \to P(B)$ con $B \in \mathcal{B}$ entonces $D(P^A||Q^A) \leq \alpha$ pues $D(P_n^A||Q^A) \leq \alpha$ para toda P_n y así

$$\lim_{n \to \infty} D(P_n^A||Q^A) \leq \alpha.$$

Por otra parte

$$\lim_{n \to \infty} D(P_n^A||Q^A) = \lim_{n \to \infty} \sum_{i=1}^{m} P_n(A_i) \log\left(\frac{P_n(A_i)}{Q(A_i)}\right)$$

$$= \sum_{i=1}^{m} \lim_{n \to \infty} P_n(A_i) \log\left(\frac{P_n(A_i)}{Q(A_i)}\right) = \sum_{i=1}^{m} P(A_i) \log\left(\frac{P(A_i)}{Q(A_i)}\right) = D(P^A||Q^A),$$

[9] cf. Teorema 4.2.2 del apéndice.

esto es, $D(P^A||Q^A) \leq \alpha$ y por lo tanto $\{P \in \mathcal{G}|D(P^A||Q^A) \leq \alpha\}$ es cerrado para toda $A \in \Theta$, luego es compacto para toda $A \in \Theta$. Ahora, observemos que

$$\mathcal{Y} := \{P \in \mathcal{G}|D(P||Q) \leq \alpha\} = \bigcap_{A\in\Theta} \{P \in \mathcal{G}|D(P^A||Q^A) \leq \alpha\},$$

esto ya que si $P \in \mathcal{Y}$ entonces $P \in \{P \in \mathcal{G}|D(P^A||Q^A) \leq \alpha\}$ para toda $A \in \Theta$ pues

$$D(P^A||Q^A) \leq \sup_{A\in\Theta} D(P^A||Q^A) = D(P||Q) \leq \alpha.$$

Luego, si $P \in \{P \in \mathcal{G}|D(P^A||Q^A) \leq \alpha\}$ para toda $A \in \Theta$ entonces $D(P^A||Q^A) \leq \alpha$ para toda $A \in \Theta$ y de esta manera

$$D(P||Q) = \sup_{A\in\Theta} D(P^A||Q^A) \leq \alpha.$$

Como intersección arbitraria de compactos es compacto (pues $[0,1]^{\mathcal{B}}$ es Hausdorff al ser producto arbitrario de espacios Hausdorff[10]), tenemos que $\mathcal{Y}$ es compacto. Lo anterior se cumple para cualesquiera $P, Q \in \mathcal{G}$, en particular para $Q \in \Lambda \subset \mathcal{G}$. Afirmamos que cuando esto sucede, es decir, $Q \in \Lambda$ entonces necesariamente $P \in \Lambda$. Para probarlo es suficiente mostrar que P es $\sigma - aditiva$. En efecto, supongamos que P no es $\sigma - aditiva$ entonces no es continua[11] en el vacío, i.e., existe $\{B_n\}_{n\in\mathbb{N}}$ una sucesión de eventos tales que $B_{n+1} \subset B_n$, $\cap_{n\in\mathbb{N}} B_n = \emptyset$ que cumple que $\lim_{n\to\infty} P(B_n) > 0$ (cf. Corolario 4.4.1 del apéndice); en tanto que por la $\sigma - aditividad$ de Q (es continua en $\emptyset$) se tiene que $Q(B_n)$ converge a 0 cuando n tiende a infinito. Consideremos entonces las particiones de S de la forma $A_n = (B_n, B_n^c)$, se sigue de lo anterior que $D(P^{A_n}||Q^{A_n})$ tiende a ∞ y por lo tanto tenemos que $D(P||Q) = \infty$, luego $P \notin \mathcal{Y}$. Entonces si $P \in \mathcal{Y}$ necesariamente es $\sigma - aditiva$ y por lo tanto $\mathcal{Y} \subset \Lambda$ cuando $Q \in \Lambda$. Así $\mathcal{Y} = \bar{B}_\epsilon(Q)$ por la definición de $\bar{B}_\epsilon(Q)$. Como $\mathcal{Y} \subset \Lambda \subset \mathcal{G} \subset [0,1]^{\mathcal{B}}$ es compacto en la topología de subespacio que hereda de $[0,1]^{\mathcal{B}}$ entonces $\bar{B}_\epsilon(Q)$ es compacto en la topología τ ya que la topología producto y la topología τ coinciden (cf. Definición 4.3.6 del apéndice y su desarrollo ulterior). Esto prueba la compacidad de las bolas inducidas por la entropía relativa en la topología τ.

(iii) Procedemos ahora a mostrar la cota superior de (3.2).

Retomamos algunos argumentos mencionados en la prueba del Teorema 1.3.1. Recordemos que ξ es un subconjunto arbitrario de Λ. Sea $A \in \Theta$ $A = (A_1, ..., A_k)$ una partición, definimos los siguientes conjuntos:

[10] cf. Lema 4.2.2 y Lema 4.2.3 del apéndice.
[11] cf. Lema 4.4.1 del apéndice.

$$\xi^A = \{P^A | P \in \xi\} \quad \text{y} \quad \xi(A) = \{P \in \Lambda | P^A \in \xi^A\}.$$

Observación. *Notemos que $\xi \subset \xi(A)$ ya que por la definición si $P \in \xi$ entonces $P^A \in \xi^A$ y así $P \in \xi(A)$. No obstante, si $Q^A \in \xi^A$ no necesariamente $Q \in \xi$ ya que Q^A es un vector de probabilidad $Q^A = (Q(A_1), ..., Q(A_k))$, entonces se puede tener que $P(A_j) = Q(A_j)$ para toda j con $P \in \xi$ pero $Q \notin \xi$, por lo tanto $\xi(A)$ no es subconjunto de ξ.*

De la observación anterior se sigue que

$$P_X^n(\{s|\hat{P}_{s,n} \in \xi\}) \leq P_X^n(\{s|\hat{P}_{s,n} \in \xi(A)\}).$$

Ahora, notemos que $P_X^n(\{s|\hat{P}_{s,n} \in \xi(A)\}) = P_X^n(\{s|\hat{P}_{s,n}^A \in \xi^A \cap \mathcal{L}_n(K)\})$ pues si $\hat{P}_{s,n} \in \xi(A)$ entonces claramente $\hat{P}_{s,n}^A \in \xi^A$ y viceversa. Además, $P_X^n(\{s|\hat{P}_{s,n}^A \in \mathcal{L}_n(K)\}) = 1$ (la medida de la intersección de un conjunto B con un conjunto de medida 1 es igual a medida del conjunto B[12]), por lo que efectivamente se da la igualdad.

Por otro lado, observemos que si $\widetilde{P}_1 \neq \widetilde{P}_2$ los eventos

$$\{s|\hat{P}_{s,n}^A = \widetilde{P}_1, \ \ \widetilde{P}_1 \in \xi^A \cap \mathcal{L}_n(K)\} \ \text{y} \ \{s|\hat{P}_{s,n}^A = \widetilde{P}_2, \ \ \widetilde{P}_2 \in \xi^A \cap \mathcal{L}_n(K)\}$$

son ajenos pues si $\widetilde{P}_1 \neq \widetilde{P}_2$ entonces difieren en alguna entrada.

Por último, recordemos que $P_X^n(\{s|\hat{P}_{s,n}^A = \widetilde{P}\}) = (P_X^n)^A(\mathcal{T}_n(\widetilde{P}))$ bajo la identificación, cf.(3.6). Así, tenemos que

$$P_X^n(\{s|\hat{P}_{s,n} \in \xi\}) \leq P_X^n(\{s|\hat{P}_{s,n} \in \xi(A)\}) = P_X^n(\{s|\hat{P}_{s,n}^A \in \xi^A \cap \mathcal{L}_n(K)\})$$

$$= P_X^n(\{s|\hat{P}_{s,n}^A = \widetilde{P}, \quad \widetilde{P} \in \xi^A \cap \mathcal{L}_n(K)\}) = P_X^n\left(\bigcup_{\widetilde{P} \in \xi^A \cap \mathcal{L}_n(K)} \{s|\hat{P}_{s,n}^A = \widetilde{P}\}\right)$$

$$= \sum_{\widetilde{P} \in \xi^A \cap \mathcal{L}_n(K)} P_X^n(\{s|\hat{P}_{s,n}^A = \widetilde{P}\}) \leq |\xi^A \cap \mathcal{L}_n(K)| \max_{\widetilde{P} \in \xi^A \cap \mathcal{L}_n(K)} P_X^n(\{s|\hat{P}_{s,n}^A = \widetilde{P}\})$$

$$\leq |\mathcal{L}_n(K)| \max_{\widetilde{P} \in \xi^A \cap \mathcal{L}_n(K)} P_X^n(\{s|\hat{P}_{s,n}^A = \widetilde{P}\}) = |\mathcal{L}_n(K)| \max_{\widetilde{P} \in \xi^A \cap \mathcal{L}_n(K)} (P_X^A)^n(\mathcal{T}_n(\widetilde{P}))$$

$$\leq (n+1)^k \max_{\widetilde{P} \in \xi^A \cap \mathcal{L}_n(K)} (P_X^A)^n(\mathcal{T}_n(\widetilde{P})) \leq (n+1)^k \max_{\widetilde{P} \in \xi^A \cap \mathcal{L}_n(K)} e^{-nD(\widetilde{P}||P_X^A)}.$$

[12] cf. Lema 4.1.13 del apéndice.

La penúltima desigualdad por el Lema 3.1.1 y la última por el Lema 3.1.2. Como $\widetilde{P} \in \xi^A$ entonces $\widetilde{P} = P^A$ para alguna $P \in \xi$. De lo anterior se sigue que

$$P_X^n(\{s|\hat{P}_{s,n} \in \xi\}) \leq (n+1)^k \max_{P^A \in \xi^A \cap \mathcal{L}_n(K)} e^{-nD(P^A||P_X^A)}. \tag{3.8}$$

Ahora, observemos que

$$\max_{P^A \in \xi^A \cap \mathcal{L}_n(k)} e^{-nD(P^A||P_X^A)} \leq \sup_{P^A \in \xi^A} e^{-nD(P^A||P_X^A)} = \sup_{P \in \xi} e^{-nD(P^A||P_X^A)},$$

como $f(y) = e^{-ny}$ es decreciente entonces

$$\sup_{P \in \xi} e^{-nD(P^A||P_X^A)} = e^{-n \inf_{P \in \xi} D(P^A||P_X^A)},$$

de esta manera y por (3.8) se tiene que

$$P_X^n(\{s|\hat{P}_{s,n} \in \xi\}) \leq (n+1)^k e^{-n \inf_{P \in \xi} D(P^A||P_X^A)}.$$

Así, obtenemos

$$\frac{1}{n} \log[P_X^n(\{s|\hat{P}_{s,n} \in \xi\})] \leq \frac{1}{n} \log[(n+1)^k] + \frac{1}{n} \log\left(e^{-n \inf_{P \in \xi} D(P^A||P_X^A)}\right),$$

es decir,

$$\frac{1}{n} \log[P_X^n(\{s|\hat{P}_{s,n} \in \xi\})] \leq \frac{1}{n} \log[(n+1)^k] - \inf_{P \in \xi} D(P^A||P_X^A).$$

Tomando el límite superior obtenemos

$$\limsup_{n \to \infty} \frac{1}{n} \log[P_X^n(\{s|\hat{P}_{s,n} \in \xi\})] \leq - \inf_{P \in \xi} D(P^A||P_X^A) \quad \text{con} \ \ A \in \Theta;$$

como $A \in \Theta$ era arbitaria entonces

$$\limsup_{n \to \infty} \frac{1}{n} \log[P_X^n(\{s|\hat{P}_{s,n} \in \xi\})] \leq \inf_{A \in \Theta} \left\{ - \inf_{P \in \xi} D(P^A||P_X^A) \right\}$$

$$= - \sup_{A \in \Theta} \inf_{P \in \xi} D(P^A||P_X^A),$$

es decir,

$$\limsup_{n \to \infty} \frac{1}{n} \log[P_X^n(\{s|\hat{P}_{s,n} \in \xi\})] \leq - \sup_{A \in \Theta} \inf_{P \in \xi} D(P^A||P_X^A). \tag{3.9}$$

De acuerdo a (3.9) para probar la cota superior en (3.2) basta mostrar que

$$\sup_{A\in\Theta} \operatorname{ínf}_{P\in\xi} D(P^A||P_X^A) \geq \operatorname*{ínf}_{P\in cd_\tau(\xi)} D(P||P_X). \tag{3.10}$$

Para probarlo podemos suponer que el lado izquierdo es finito pues de lo contrario la prueba es trivial.

Afirmación. *Para cualquier* $\alpha > 0$ *que cumpla*

$$\sup_{A\in\Theta} \operatorname{ínf}_{P\in\xi} D(P^A||P_X^A) < \alpha \tag{3.11}$$

se tiene que

$$cd_\tau(\xi) \cap \bar{B}_\alpha(P_X) \neq \emptyset. \tag{3.12}$$

Basta probar la afirmación para tener (3.10), ya que si suponemos que la afirmación es cierta y que

$$\sup_{A\in\Theta} \operatorname{ínf}_{P\in\xi} D(P^A||P_X^A) < \operatorname*{ínf}_{P\in cd_\tau(\xi)} D(P||P_X),$$

si $\alpha' = \operatorname{ínf}_{P\in cd_\tau(\xi)} D(P||P_X)$, entonces

$$cd_\tau(\xi) \cap \bar{B}_{\alpha'}(P_X) \neq \emptyset,$$

es decir, existe $Q \in cd_\tau(\xi)$ tal que $D(Q||P_X) \leq \alpha' = \operatorname{ínf}_{P\in cd_\tau(\xi)} D(P||P_X)$, lo cual sólo puede suceder si $D(Q||P_X) = \operatorname{ínf}_{P\in cd_\tau(\xi)} D(P||P_X)$, i.e.,

$$\sup_{A\in\Theta} \operatorname{ínf}_{P\in\xi} D(P^A||P_X^A) < D(Q||P_X)$$

entonces $\operatorname{ínf}_{P\in\xi} D(P^A||P_X^A) < D(Q||P_X)$ para toda $A \in \Theta$, así existe una $Q' \in \xi$ tal que $D(Q'^A||P_X^A) < D(Q||P_X)$ para toda $A \in \Theta$ luego

$$D(Q'||P_X) = \sup_{A\in\Theta} D(Q'^A||P_X^A) < D(Q||P_X) = \operatorname*{ínf}_{P\in cd_\tau(\xi)} D(P||P_X),$$

lo cual no puede suceder pues $Q' \in \xi \subset cd_\tau(\xi)$. Por lo tanto tenemos que efectivamente la afirmación implica la desigualdad

$$\sup_{A\in\Theta} \operatorname{ínf}_{P\in\xi} D(P^A||P_X^A) \geq \operatorname*{ínf}_{P\in cd_\tau(\xi)} D(P||P_X).$$

Para probar la afirmación procederemos de la siguiente manera:

(1) Mostramos que

$$cd_\tau(\xi) = \bigcap_{A\in\Theta} cd_\tau(\xi(A)).$$

(2) Mostramos que

$$\left[\bigcap_{A\in\Theta} cd_\tau(\xi(A))\right] \cap \bar{B}_\alpha(P_X) \neq \emptyset.$$

Observamos que de (1) y (2) obtenemos (3.12). Procedemos entonces a probar (1): Como ya se había mencionado anteriormente tenemos que, por la definición de $\xi(A)$, $\xi \subset \xi(A)$ para cada partición $A \in \Theta$. De esta forma $cd_\tau(\xi) \subset cd_\tau(\xi(A))$ para cada $A \in \Theta$ y así

$$cd_\tau(\xi) \subset \bigcap_{A\in\Theta} cd_\tau(\xi(A)).$$

Para mostrar la otra contención damos una P en la intersección de los conjuntos $cd_\tau(\xi(A))$, es decir, $P \in cd_\tau(\xi(A))$ para cada $A \in \Theta$; debemos mostrar que para toda vecindad $U(P, A, \epsilon)$ de P con $\epsilon > 0$ y $A \in \Theta$ se tiene que $U(P, A, \epsilon) \cap \xi \neq \emptyset$, para ello damos un $\epsilon > 0$ fijamos una partición arbitraria $A = (A_1, ..., A_k) \in \Theta$ y observamos que $U(P, A, \epsilon) \cap \xi(A) \neq \emptyset$ pues $P \in cd_\tau(\xi(A))$, de este modo podemos tomar una medida de probabilidad $P' \in U(P, A, \epsilon) \cap \xi(A)$. $P' \in \xi(A)$ y por la definición de $\xi(A)$ se tiene que existe una $\widetilde{P} \in \xi$ tal que $\widetilde{P}(A_i) = P'(A_i)$ $i = 1, ..., k$ lo cual nos da que $\widetilde{P} \in U(P, A, \epsilon)$ pues $P' \in U(P, A, \epsilon)$ y recordamos que

$$U(P, A, \epsilon) = \left\{Q \in \Lambda \middle|\ |Q(A_i) - P(A_i)| < \epsilon, \quad i = 1, ..., k\right\},$$

por lo tanto $\widetilde{P} \in U(P, A, \epsilon) \cap \xi$, es decir, $U(P, A, \epsilon) \cap \xi \neq \emptyset$ para toda $A \in \Theta$ y para todo $\epsilon > 0$; así, $P \in\ cd_\tau(\xi)$, i.e.,

$$\bigcap_{A\in\Theta} cd_\tau(\xi(A)) \subset cd_\tau(\xi),$$

con lo cual queda demostrado (1).

Procedemos ahora a mostrar (2): Primero observemos que

$$\left[\bigcap_{A\in\Theta} cd_\tau(\xi(A))\right] \cap \bar{B}_\alpha(P_X) = \bigcap_{A\in\Theta} [cd_\tau(\xi(A)) \cap \bar{B}_\alpha(P_X)].$$

Como $cd_\tau(\xi(A)) \cap \bar{B}_\alpha(P_X) \subset \bar{B}_\alpha(P_X)$ y este último es compacto si mostramos que $cd_\tau(\xi(A)) \cap \bar{B}_\alpha(P_X)$ es cerrado para toda $A \in \Theta$ y que $\{cd_\tau(\xi(A)) \cap \bar{B}_\alpha(P_X) | A \in \Theta\}$ tiene la propiedad de intersección finita, es decir, para toda $n \in \mathbb{N}$

$$\bigcap_{j=1}^{n} [cd_\tau(\xi(A^j)) \cap \bar{B}_\alpha(P_X)] \neq \emptyset \quad , A^j \in \Theta \quad j = 1, ..., n,$$

se seguiría entonces que

$$\bigcap_{A\in\Theta}[cd_\tau(\xi(A))\cap\bar{B}_\alpha(P_X)]\neq\emptyset,$$

cf. Lema 4.2.4 del apéndice, de esta forma obtendríamos (2).

Lo primero se sigue del hecho de que las bolas $\bar{B}_\alpha(Q)$ son compactas en la topología τ pues entonces $cd_\tau(\xi(A))\cap\bar{B}_\alpha(P_X)$ es intersección de un subconjunto cerrado y un conjunto compacto y por lo tanto es cerrado. Resta mostrar la propiedad de intersección finita, para ello primero mostramos que para cada $A=(A_1,...,A_k)\in\Theta$ se tiene que

$$\xi(A)\cap\bar{B}_\alpha(P_X)\neq\emptyset. \tag{3.13}$$

En efecto, (3.11) nos da que para toda $A\in\Theta$

$$\inf_{P\in\xi}D(P^A||P_X^A)<\alpha,$$

lo cual implica que existe $\widetilde{P}\in\xi$ tal que $D(\widetilde{P}^A||P_X^A)<\alpha$, mediante esta medida de probabilidad construimos una nueva medida de probabilidad de la siguiente manera:

$$P(B)=\sum_{i=1}^{k}\frac{\widetilde{P}(A_i)}{P_X(A_i)}P_X(B\cap A_i),\quad B\in\mathcal{B}.$$

En donde definimos $\frac{\widetilde{P}(A_i)}{P_X(A_i)}=0$ si $P_X(A_i)=0$ (ya que $\widetilde{P}(A_i)=0$ siempre que $P_X(A_i)=0$ pues $D(\widetilde{P}^A||P_X^A)<\infty$, i.e., $\widetilde{P}^A\ll P_X^A$). Efectivamente es una medida de probabilidad pues $\frac{P_X(\cdot\cap A_i)}{P_X(A_i)}$ es medida de probabilidad para cada i y $0\leq\tilde{P}(A_i)\leq 1$, $\sum_{i=1}^{k}\tilde{P}(A_i)=1$, es decir, P es una combinación lineal convexa de medidas de probabilidad.

Por otro lado, observemos que $P^A=\widetilde{P}^A$ ya que

$$P(A_j)=\sum_{i=1}^{k}\frac{\widetilde{P}(A_i)}{P_X(A_i)}P_X(A_j\cap A_i)=\widetilde{P}(A_j),\quad j=1,...,k.$$

Así, $P^A\in\xi^A$ luego $P\in\xi(A)$. Además, si $D(P||P_X)=D(\widetilde{P}^A||P_X^A)<\alpha$ entonces $P\in\xi(A)\cap\bar{B}_\alpha(P_X)$ y obtendríamos (3.13); para justificar esta última igualdad observemos que como $D(P||P_X)$ es el supremo sobre todas las particiones entonces basta mostrar que para cualquier refinamiento $E=(E_1,...,E_l)$ de A, es decir, cualquier partición E de S tal que cada elemento E_m, $m=1,...,l$ cumple que $E_m\subset A_i$ para alguna $i=1,...,k$, se tiene que $D(P^E||P_X^E)=D(\widetilde{P}^A||P_X^A)$. Para ello es suficiente observar que como E refina a A entonces cada elemento A_i de A puede ser expresado como una unión de elementos de E, concretamente,

$$A_i = \bigcup_{m \in M_i} E_m \text{ con } M_i = \{m | E_m \subset A_i\}.$$

Para facilitar las cosas podemos suponer sin pérdida de generalidad que el refinamiento E está ordenado de tal manera que $A_1 = \cup_{s=1}^{r_1} E_s$, $A_2 = \cup_{s=r_1+1}^{r_2} E_s$,..., $A_k = \cup_{s=r_{k-1}+1}^{l} E_s$ con $1 \leq r_1 < r_2 < ... < r_{k-1} \leq l$ podemos suponerlo ya que sólo estamos reordenando la partición E. Por como definimos P se tiene que

$$D(P^E||P_X^E) = \sum_{j=1}^{l} \log \left[\frac{\sum_{i=1}^{k} \frac{\widetilde{P}(A_i)}{P_X(A_i)} P_X(E_j \cap A_i)}{P_X(E_j)} \right] \sum_{i=1}^{k} \frac{\widetilde{P}(A_i)}{P_X(A_i)} P_X(E_j \cap A_i).$$

Ahora, como se observó cada E_j es subconjunto de un y sólo un A_i, es decir, $E_j \subset A_i^j$, y observemos que es posible que $A_i^j = A_i^r$ con $j \neq r$ pues $A_i = \cup_{m \in M_i} E_m$, así $\sum_{i=1}^{k} \frac{\widetilde{P}(A_i)}{P_X(A_i)} P_X(E_j \cap A_i) = \frac{\widetilde{P}(A_i^j)}{P_X(A_i^j)} P_X(A_i^j)$ para cada j. Esto ya que $P_X(E_j \cap A_i^r) = 0$ con $j \neq r$ pues $E_j \cap A_i^r = \emptyset$ y además $P_X(E_j \cap A_i^j) = P_X(E_j)$ pues $E_j \subset A_i^j$. Luego

$$\sum_{j=1}^{l} \log \left[\frac{\sum_{i=1}^{k} \frac{\widetilde{P}(A_i)}{P_X(A_i)} P_X(E_j \cap A_i)}{P_X(E_j)} \right] \sum_{i=1}^{k} \frac{\widetilde{P}(A_i)}{P_X(A_i)} P_X(E_j \cap A_i)$$

$$= \sum_{j=1}^{l} \log \left[\frac{\frac{\widetilde{P}(A_i^j)}{P_X(A_i^j)} P_X(E_j)}{P_X(E_j)} \right] \frac{\widetilde{P}(A_i^j)}{P_X(A_i^j)} P_X(E_j)$$

$$= \sum_{j=1}^{l} \log \left[\frac{\widetilde{P}(A_i^j)}{P_X(A_i^j)} \right] \frac{\widetilde{P}(A_i^j)}{P_X(A_i^j)} P_X(E_j) = \sum_{i=1}^{k} \log \left[\frac{\widetilde{P}(A_i)}{P_X(A_i)} \right] \frac{\widetilde{P}(A_i)}{P_X(A_i)} P_X(A_i)$$

$$= \sum_{i=1}^{k} \log \left[\frac{\widetilde{P}(A_i)}{P_X(A_i)} \right] \widetilde{P}(A_i) = D(\widetilde{P}^A||P_X^A).$$

En donde la igualdad

$$\sum_{j=1}^{l} \log \left[\frac{\widetilde{P}(A_i^j)}{P_X(A_i^j)} \right] \frac{\widetilde{P}(A_i^j)}{P_X(A_i^j)} P_X(E_j) = \sum_{i=1}^{k} \log \left[\frac{\widetilde{P}(A_i)}{P_X(A_i)} \right] \frac{\widetilde{P}(A_i)}{P_X(A_i)} P_X(A_i)$$

se da ya que como $A_i = \cup_{m \in M_i} E_m$ entonces $A_i^1 = ... = A_i^{r_1} = A_1$, $A_i^{r_1+1} = ... = A_i^{r_2} = A_2$,..., $A_i^{r_{k-1}+1} = ... = A_i^l = A_k$ y así

$$\log\left[\frac{\widetilde{P}(A_i^1)}{P_X(A_i^1)}\right]\frac{\widetilde{P}(A_i^1)}{P_X(A_i^1)}P_X(E_1)+\log\left[\frac{\widetilde{P}(A_i^2)}{P_X(A_i^2)}\right]\frac{\widetilde{P}(A_i^2)}{P_X(A_i^2)}P_X(E_2)+...$$

$$...+\log\left[\frac{\widetilde{P}(A_i^l)}{P_X(A_i^1)}\right]\frac{\widetilde{P}(A_i^l)}{P_X(A_i^l)}P_X(E_l)$$

$$=\log\left[\frac{\widetilde{P}(A_1)}{P_X(A_1)}\right]\frac{\widetilde{P}(A_1)}{P_X(A_1)}[P_X(E_1)+...+P_X(E_{r_1})]$$

$$+\log\left[\frac{\widetilde{P}(A_2)}{P_X(A_2)}\right]\frac{\widetilde{P}(A_2)}{P_X(A_2)}[P_X(E_{r_1+1})+...+P_X(E_{r_2})]+...$$

$$+\log\left[\frac{\widetilde{P}(A_k)}{P_X(A_k)}\right]\frac{\widetilde{P}(A_k)}{P_X(A_k)}[P_X(E_{r_{k-1}+1})+...+P_X(E_l)]$$

$$=\log\left[\frac{\widetilde{P}(A_1)}{P_X(A_1)}\right]\frac{\widetilde{P}(A_1)}{P_X(A_1)}[P_X(\cup_{m\in M_1}E_m=A_1)]$$

$$+\log\left[\frac{\widetilde{P}(A_2)}{P_X(A_2)}\right]\frac{\widetilde{P}(A_2)}{P_X(A_2)}[P_X(\cup_{m\in M_2}E_m=A_2)]+...$$

$$+\log\left[\frac{\widetilde{P}(A_k)}{P_X(A_k)}\right]\frac{\widetilde{P}(A_k)}{P_X(A_k)}[P_X(\cup_{m\in M_k}E_m=A_k)]$$

$$=\sum_{i=1}^{k}\log\left[\frac{\widetilde{P}(A_i)}{P_X(A_i)}\right]\frac{\widetilde{P}(A_i)}{P_X(A_i)}P_X(A_i).$$

Por lo que efectivamente $D(P^E||P_X^E)=D(\widetilde{P}^A||P_X^A)$.

Ahora mostramos la propiedad de intersección finita: Consideremos una colección finita y arbitraria de particiones $\{A^i\}_{i=1}^n\subset\Theta$. Entonces siempre existe una partición $A\in\Theta$ que refina a cada $A^i=(A_1^i,...,A_k^i)$[13]. Sea $A\in\Theta$ la partición que refina a cada partición de $\{A^i\}_{i=1}^n$ luego $\xi(A)\subset\xi(A^i)$ para cada i pues si $P\in\xi(A)$ entonces $P^A\in\xi^A$, es decir, existe $Q\in\xi$ tal que $P(A_m)=Q(A_m)$ para toda m, como A refina a A^i se tiene que

$$P(A_j^i)=P\left(\bigcup_{m\in M_j}A_m\right)=\sum_{m\in M_j}P(A_m)=\sum_{m\in M_j}Q(A_m)$$

$$=Q\left(\bigcup_{m\in M_j}A_m\right)=Q(A_j^i),$$

[13]cf. Lema 4.1.14 del apéndice.

entonces $P^{A^i} \in \xi^{A^i}$ luego $P \in \xi(A^i)$, por lo que efectivamente $\xi(A) \subset \xi(A^i)$ para cada $i = 1, ..., n$; (3.13) implica que

$$\bigcap_{i=1}^{n} [\xi(A^i) \cap \bar{B}_\alpha(P_X)] \neq \emptyset,$$

y como $\xi(A^i) \cap \bar{B}_\alpha(P_X) \subset cd_\tau(\xi(A^i)) \cap \bar{B}_\alpha(P_X)$ por lo tanto

$$\bigcap_{i=1}^{n} [cd_\tau(\xi(A^i)) \cap \bar{B}_\alpha(P_X)] \neq \emptyset.$$

De esta forma hemos probado la propiedad de intersección finita para $\{cd_\tau(\xi(A)) \cap \bar{B}_\alpha(P_X) | A \in \Theta\}$ lo cual da por concluida la prueba.

□

Ahora que ya tenemos el teorema de Sanov en cualquier espacio medible podemos formular nueva terminología. En la siguiente sección veremos la relación que hay entre la denominada propiedad de Sanov y una nueva forma de dependencia entre variables aleatorias conocida como cuasiindependencia asintótica, ambas se definirán más adelante. Mostraremos algunos lemas para demostrar dos resultados principales concernientes a estas dos nuevas propiedades y la caracterización de la I-proyección generalizada.

3.2. Propiedad de Sanov y cuasiindependencia asintótica

Definición 3.2.1. *Sea $\{X_i\}_{i\in\mathbb{N}}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas sobre el espacio de medida $(S, \mathcal{B})$ con distribución común P_X. Diremos que un conjunto Π de medidas de probabilidad sobre $(S, \mathcal{B})$ tiene la propiedad de Sanov si la distribución empírica $\hat{P}_{X^n}$ del vector $X^n = (X_1, ..., X_n)$ satisface que*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\hat{P}_{X^n} \in \Pi) = - \inf_{P\in\Pi} D(P||P_X). \tag{3.14}$$

Si el evento $\{s|\hat{P}_{s,n} \in \Pi\}$ no se encuentra en la σ-álgebra diremos que el conjunto Π tiene la propiedad de Sanov si

$$\lim_{n\to\infty} \frac{1}{n} \log \bar{\mathbb{P}}(\hat{P}_{X^n} \in \Pi) = -D(\Pi||P_X)$$

y

$$\lim_{n\to\infty} \frac{1}{n} \log \underline{\mathbb{P}}(\hat{P}_{X^n} \in \Pi) = -D(\Pi||P_X).$$

Observación. *Notemos que el Teorema 3.1.1 (Sanov versión general) no garantiza que un conjunto posea la propiedad de Sanov pues en la cota inferior el ínfimo se toma sobre el interior del conjunto respecto a la topología τ, asimismo en la cota superior el ínfimo se toma sobre la cerradura del conjunto en la misma topología. Si el conjunto es abierto y cerrado al mismo tiempo entonces el conjunto sí posee la propiedad de Sanov.*

Ahora es momento de presentar una nueva forma de dependencia llamada cuasiindependencia asintótica, este tipo de dependencia está motivada por la manera en que se comportan en el límite ciertos tipos de sucesiones de variables aleatorias dependientes.

Definición 3.2.2. *Para cada subconjunto $A \in \mathcal{B}^n$ de S^n con $P_X^n(A) > 0$ denotamos por $P_{X_i|A}$, $P_{X^n|A}$ a la distribución condicional de X_i $i = 1, ..., n$ y $X^n = (X_1, ..., X_n)$ respectivamente, dado que $X^n \in A$. $P_{X^n|A}$ es la medida de probabilidad sobre $(S^n, \mathcal{B}^n)$ definida por:*

$$P_{X^n|A}(E) = \frac{P_X^n(E \cap A)}{P_X^n(A)}, \quad E \in \mathcal{B}^n$$

y $P_{X_i|A}$ es su i-ésima distribución marginal. En el caso en que $P_{X_1|A} = ... = P_{X_n|A}$, esta medida de probabilidad será denotada simplemente como $P_{X|A}$.

Así como extendimos la propiedad de Sanov al caso en que el evento que nos interesa no se encuentra en la σ-álgebra, así hacemos lo análogo para esta medida de probabilidad cuando $A \notin \mathcal{B}^n$, es decir, $P_{X^n|A} = P_{X^n|\bar{A}}$ en donde $A \subset \bar{A} \in \mathcal{B}^n$ y $P_X^n(\bar{A}) = \text{mín}\{P_X^n(C)|C \in \mathcal{B}^n, \ A \subset C\}$. Esto nos lleva a una definición sin ambigüedades de la distribución condicional siempre y cuando A tenga P_X^n-medida exterior positiva. En particular, las distribuciones

$$P_{X^n|\hat{P}_n \in \Pi} = P_{X^n|A_n} \quad \text{y} \quad P_{X|\hat{P}_n \in \Pi} = P_{X|A_n}$$

están bien definidas siempre y cuando $\bar{\mathbb{P}}(\{s|\hat{P}_{s,n} \in \Pi\}) > 0$, con A_n como se definió al principio de este capítulo, es decir,

$$\mathbb{P}(\hat{P}_{X^n} \in \Pi) = P_X^n(A_n) \quad A_n = \{s|\hat{P}_{s,n} \in \Pi\}.$$

Definición 3.2.3. *Sean $\{X_n\}_{n\in\mathbb{N}}$ variables aleatorias que toman valores en S con distribución conjunta $P^{(n)}$ $n = 1, 2, ...$ y $A_n \subset S^n$ conjuntos tales que $P_{X^n|A_n} = P^{(n)}$, $n = 1, 2, ...$ Si sucede que*

$$\lim_{n\to\infty} \frac{1}{n} D(P^{(n)}||Q^n) = 0$$

para alguna medida de probabilidad $Q \in \Lambda$ decimos que $X_1, ..., X_n$ son asintóticamente cuasiindependientes bajo la condición $X^n \in A_n$ con distribución límite Q.

La terminología anterior está motivada por el hecho de que, en el límite, "el conjunto de variables aleatorias se comportan como variables independientes con distribución común Q".

Definición 3.2.4. *Sean $(X, \mathcal{A})$ y $(Y, \mathcal{B})$ dos espacios medibles. Un kernel de Markov o kernel de transición que parte de $(X, \mathcal{A})$ y tiene como destino o espacio de estados a $(Y, \mathcal{B})$ es una función $\nu : X \times \mathcal{B} \to [0, 1]$ que cumple las siguientes características:*

(1) La función $\nu(\cdot, B)$ es $\mathcal{A}$-medible para cada $B \in \mathcal{B}$.

(2) La función $\nu(x, \cdot)$ es una medida de probabilidad sobre $(Y, \mathcal{B})$ para cada $x \in X$.

Ejemplo. *Observemos que $\hat{P}_n(\cdot, \cdot)$ cumple con la definición de kernel de Markov ya que $\hat{P}_n : S^n \times \mathcal{B} \to [0, 1]$ y cumple que $\hat{P}_n(\cdot, B)$ es medible para cada $B \in \mathcal{B}$ y $\hat{P}_n(s, \cdot)$ es medida de probabilidad para cada $s \in S^n$.*

Observación. *Notemos que toda medida de probabilidad induce un kernel de Markov de la siguiente manera: Sea $P \in \Lambda(Y)$ definimos*

$$\nu : X \times \mathcal{B} \to [0, 1] \ \text{ como } \ \nu(x, \cdot) = P(\cdot) \ \ \forall x \in X.$$

Claramente $\nu(x, \cdot)$ es una medida de probabilidad sobre Y y $\nu(\cdot, B)$ es medible pues es una constante para cada $B \in \mathcal{B}$. Por lo tanto ν es un kernel de Markov de $(X, \mathcal{A})$ a $(Y, \mathcal{B})$. Más aún, observemos que dadas dos medidas de probabilidad $P_1, P_2 \in \Lambda(Y)$ y un conjunto medible $A \subset X$ podemos definir

$$\nu(x, \cdot) = \begin{cases} P_1(\cdot) & si \quad x \in A \\ P_2(\cdot) & si \quad x \notin A. \end{cases}$$

$\nu(x, \cdot)$ es una medida de probabilidad para todo $x \in X$ y $\nu(\cdot, B)$ es $\mathcal{A}$-medible para cada $B \in \mathcal{B}$. En efecto, si $P_1(B) = P_2(B)$ es una constante y es $\mathcal{A}$-medible. Si $P_1(B) \neq P_2(B)$ tomamos un $E \in \mathcal{B}(\mathbb{R})$, $E \subset [0, 1]$.

Caso 1: $P_1(B), P_2(B) \in E$.

Entonces $\nu^{-1}(E) = A \cup A^c \cup \emptyset \in \mathcal{A}$.

Caso 2: $P_1(B), P_2(B) \notin E$.

Entonces $\nu^{-1}(E) = \emptyset \in \mathcal{A}$.

Caso 3: $P_1(B) \in E$ y $P_2(B) \notin E$.

Entonces $\nu^{-1}(E) = A \cup \emptyset \in \mathcal{A}$

Caso 4: $P_1(B) \notin E$ y $P_2(B) \in E$. Análogamente al caso 3 se tiene que $\nu^{-1}(E) = A^c \cup \emptyset \in \mathcal{A}$.

Tomemos un espacio de probabilidad arbitrario $(\Omega, \mathcal{A}, \mu)$ y un kernel de Markov ν de $(\Omega, \mathcal{A})$ a $(S, \mathcal{B})$ notemos que la función conjuntista $\mu\nu : \mathcal{B} \to \mathbb{R}$ definida de la siguiente forma:

$$\mu\nu(B) = \int \nu(\omega, B) \ d\mu(\omega), \quad B \in \mathcal{B},$$

es una medida de probabilidad. En efecto, como $\nu(\omega, \cdot)$ es medida de probabilidad para cada $\omega \in \Omega$ claramente $0 \leq \mu\nu(B) \leq 1$ pues $0 \leq \nu(\omega, B) \leq 1$ para cada $\omega \in \Omega$ y $\mu\nu(\emptyset) = 0$ y $\mu\nu(S) = 1$. Por último, si $\{B_i\}_{i\in\mathbb{N}}$ es una familia de conjuntos medibles ajenos dos a dos entonces se tiene que

$$\nu\big(\omega, \cup_{i\in\mathbb{N}} B_i\big) = \sum_{i\in\mathbb{N}} \nu(\omega, B_i) \quad \text{para cada} \quad \omega \in \Omega,$$

luego

$$\mu\nu\left(\bigcup_{i\in\mathbb{N}} B_i\right) = \int \nu\big(\omega, \cup_{i\in\mathbb{N}} B_i\big) \ d\mu(\omega) = \sum_{i\in\mathbb{N}} \int \nu(\omega, B_i) \ d\mu(\omega)$$

$$= \sum_{i\in\mathbb{N}} \mu\nu(B_i).$$

En donde aplicamos el teorema de convergencia dominada de Lebesgue a la sucesión $f_n(\omega) = \sum_{i=1}^{n} \nu(\omega, B_i)$ y $f(\omega) = \sum_{i\in\mathbb{N}} \nu(\omega, B_i)$, con $f_n(\omega) \to f(\omega)$ para toda ω; f_n es dominada por la función constante 1, i.e., $|f_n(\omega)| \leq 1$ para toda ω.

Esta construcción de una nueva medida de probabilidad a través de un kernel de Markov nos permite definir lo siguiente.

Definición 3.2.5. *Un conjunto de medidas de probabilidad $\Pi \subset \Lambda$ se dice que es completamente convexo si para cada espacio de probabilidad $(\Omega, \mathcal{A}, \mu)$ y cada kernel de Markov ν de $(\Omega, \mathcal{A})$ a $(S, \mathcal{B})$ tal que $\nu(\omega, \cdot) \in \Pi$ para cada $\omega \in \Omega$ se tiene que $\mu\nu \in \Pi$. Un conjunto convexo de medidas de probabilidad $\Pi \subset \Lambda$ es casi completamente convexo si existen subconjuntos completamente convexos $\Pi_1 \subset \Pi_2 \subset ...$ tales que $\Pi_i \subset \Pi$ para toda i y $\Pi \cap \Lambda_f \subset \bigcup_{i\in\mathbb{N}} \Pi_i$.*

Ejemplo. *Una clase muy amplia de ejemplos de este tipo de conjuntos (tanto completamente convexos como casi completamente convexos) es mostrada en el Lema 3.3.1.*

Notemos que la definición de convexidad completa generaliza la noción de convexidad en conjuntos de medidas de probabilidad. En efecto, tomemos $P_1, P_2 \in \Pi \subset \Lambda$ luego $P_\lambda = \lambda P_1 + (1-\lambda)P_2 \in \Lambda$, $\lambda \in [0, 1]$ por lo que si Π es un conjunto completamente convexo entonces tomamos un espacio de probabilidad $(\Omega, \mathcal{A}, \mu)$ y un conjunto $A \subset \Omega$ medible tal que $\mu(A) = \lambda$, por la observación anterior tenemos que

$$\nu(\omega,\cdot)=\begin{cases}P_1(\cdot) & si \quad \omega\in A\\ P_2(\cdot) & si \quad \omega\notin A\end{cases}$$

es un kernel de Markov de $(\Omega,\mathcal{A})$ a $(S,\mathcal{B})$, además cumple que $\nu(\omega,\cdot)\in\Pi$ para cada $\omega\in\Omega$; como Π es completamente convexo entonces $\mu\nu\in\Pi$. Y

$$\mu\nu(B)=\int \nu(\omega,B)\ d\mu(\omega)=\int_A \nu(\omega,B)\ d\mu(\omega)+\int_{A^c}\nu(\omega,B)\ d\mu(\omega)$$

$$=\int_A P_1(B)\ d\mu(\omega)+\int_{A^c}P_2(B)\ d\mu(\omega)=P_1(B)\mu(A)+P_2(B)\mu(A^c)$$

$$=\lambda P_1(A)+(1-\lambda)P_2(B)=P_\lambda(B)\quad \text{con}\ \ B\in\mathcal{B}.$$

Por lo tanto $P_\lambda\in\Pi$, es decir, Π es convexo. Ahora, si Π es casi completamente convexo entonces Π es convexo por definición. Si Π es completamente convexo entonces es casi completamente convexo pues tomamos a $\Pi_i=\Pi$ para toda i y observemos que cumple con la definición de conjunto casi completamente convexo.

Por otro lado, observemos que a pesar de que Λ_f es convexo[14] éste no es completamente convexo: Tomemos μ una medida de probabilidad sin átomos y consideremos al espacio de probabilidad $(S^n,\mathcal{B}^n,\mu^n)$ con μ^n la medida producto $\otimes_{i=1}^n\mu$. Sea $\hat{P}_n\in\Lambda_f$ y consideremos a $\nu=\hat{P}_n$ como kernel de Markov de $(S^n,\mathcal{B}^n)$ a $(S,\mathcal{B})$, $\hat{P}_n(s,\cdot)\in\Lambda_f$ para cada $s\in S$. No obstante,

$$\mu\nu(B)=\int\nu(s,B)\ d\mu^n(s)=\int_{S^n}\hat{P}_n(s,B)\ d\mu^n(s)$$

$$=\int_{S^n}\frac{1}{n}\sum_{i=1}^n 1_B(s_i)\ d\mu^n(s)=\frac{1}{n}\sum_{i=1}^n\int_{S^n}1_B(s_i)\ d\mu^n(s)$$

$$=\frac{1}{n}\sum_{i=1}^n\int_{S^n}1_B(s_i)\prod_{j\neq i}1_S(s_j)\ d\mu^n(s)$$

$$=\frac{1}{n}\sum_{i=1}^n\int_S 1_B(s_i)\ d\mu(s_i)\prod_{j\neq i}\int_S 1_S(s_j)\ d\mu(s_j)$$

$$=\frac{1}{n}\sum_{i=1}^n\mu(B)=\mu(B).$$

[14] cf. Lema 4.4.12 del apéndice.

Por lo que $\mu\nu$ es una medida de probabilidad sin átomos, i.e., $\mu\nu \notin \Lambda_f$. Entonces efectivamente Λ_f no es completamente convexo, ni siquiera es casi completamente convexo pues si lo fuera tendríamos que $\Lambda_f = \Lambda_f \cap \Lambda_f \subset \cup_{i\in\mathbb{N}}\Pi_i$ con $\Pi_i \subset \Lambda_f$, $\Pi_i \subset \Pi_{i+1}$ para toda i, lo que nos daría que $\Lambda_f = \cup_{i\in\mathbb{N}}\Pi_i$, es decir, que Λ_f es completamente convexo lo cual acabamos de ver que no puede suceder.

Hasta ahora lo que hemos mostrado es lo siguiente: utilizando la abreviación CX para convexo, CC para completamente convexo y CCC para casi completamente convexo tenemos las implicaciónes mostrada en el siguiente diagrama

$$CC \to CCC \to CX, \quad CX \nrightarrow CCC \quad \text{y} \quad CX \nrightarrow CC.$$

En el Lema 3.3.1 se obtiene una clase de ejemplos de subconjuntos que son casi completamente convexos pero no completamente convexos; con lo cual obtenemos entonces que

$$CC \to CCC \to CX \quad \text{y} \quad CX \nrightarrow CCC \nrightarrow CC.$$

Con la terminología anterior ya contamos con las herramientas necesarias para enunciar el primer resultado de este capítulo.

Teorema 3.2.1. *Sea $\Pi \subset \Lambda$ un conjunto casi completamente convexo, supongamos que $D(\Pi||P_X) < \infty$ y sea P^* la I-proyección generalizada de P_X en Π. Entonces*

$$\frac{1}{n}\log \bar{\mathbb{P}}(\hat{P}_{X^n} \in \Pi) \leq -D(\Pi||P_X) \quad \textit{para cada } n \in \mathbb{N}. \tag{3.15}$$

Además, para cada $\Pi' \subset \Pi$ tal que $\bar{\mathbb{P}}(\hat{P}_{X^n} \in \Pi') > 0$ se tiene que

$$\frac{1}{n}\log \bar{\mathbb{P}}(\hat{P}_{X^n} \in \Pi') \leq -D(\Pi||P_X) - \frac{1}{n}D(P_{X^n|\hat{P}_n\in\Pi'}||P^{*n}). \tag{3.16}$$

Por último, si Π' es tal que

$$D(int_{\tau_0}\Pi'||P_X) = D(\Pi||P_X) < \infty \tag{3.17}$$

entonces Π y Π' tienen la propiedad de Sanov y $X_1, ..., X_n$ son asintóticamente cuasiindependientes bajo la condición $\hat{P}_{X^n} \in \Pi'$ con distribución límite P^*.

Para poder demostrar este teorema es necesario mostrar antes tres lemas que enunciamos a continuación. Las técnicas de demostración del primer lema son muy similiares a las utilizadas en la demostración de (i) en el Teorema 3.1.1 demostrado al principio de este capítulo.

Consideremos al espacio topológico (Λ, τ_0), recordemos que un conjunto $\xi \subset \Lambda$ es abierto relativo en $\xi \cup \Lambda_f$ si $\xi \cap (\xi \cup \Lambda_f)$ es abierto en $(\xi \cup \Lambda_f, \tau_{\xi \cup \Lambda_f})$ en donde $\tau_{\xi \cup \Lambda_f}$ denota la topología relativa en $\xi \cup \Lambda_f$ (cf. Definición 4.2.3 del apéndice). Como $\xi \subset \xi \cup \Lambda_f$ entonces ξ debe ser abierto en la topología relativa, lo anterior significa que para toda $P \in \xi$ existe $U_0(P, A, \epsilon)$ tal que $U_0(P, A, \epsilon) \cap (\xi \cup \Lambda_f) \subset \xi$, pero

$$U_0(P, A, \epsilon) \cap (\xi \cup \Lambda_f) = [U_0(P, A, \epsilon) \cap \xi] \cup [U_0(P, A, \epsilon) \cap \Lambda_f],$$

esto es, $U_0(P, A, \epsilon) \cap \Lambda_f \subset \xi$.

Para facilitar la notación definimos $\{s|\hat{P}_{s,n} \in \Pi\} = \{\hat{P}_{s,n} \in \Pi\}$. En algunas ocasiones utilizaremos indistintamente $\{s|\hat{P}_{s,n} \in \Pi\}$ o $\{\hat{P}_{s,n} \in \Pi\}$.

Lema 3.2.1. *Si $\Pi \subset \Lambda$ cumple alguna de las siguientes dos condiciones:*

(1) Es un abierto relativo de $\Pi \cup \Lambda_f$ respecto a la topología τ_0.

(2) $D(\Pi||P_X) = D(int_{\tau_0}\Pi||P_X)$.

Entonces

$$-D(\Pi||P_X) \leq \liminf_{n\to\infty} \frac{1}{n} \log \underline{\mathbb{P}}(\hat{P}_{X^n} \in \Pi). \tag{3.18}$$

Prueba. Observemos que es suficiente mostrar que (1) implica (3.18) pues para mostrar que (2) también lo hace aplicamos (3.18) al conjunto $int_{\tau_0}\Pi$ (que es abierto en la topología τ_0 y por lo tanto es abierto relativo). En efecto, primero si s es tal que $\hat{P}_{s,n} \in int_{\tau_0}\Pi$ entonces $\hat{P}_{s,n} \in \Pi$ luego

$$\{\hat{P}_{s,n} \in int_{\tau_0}\Pi\} \subset \{\hat{P}_{s,n} \in \Pi\},$$

así

$$\underline{\mathbb{P}}(\hat{P}_{X^n} \in int_{\tau_0}\Pi) \leq \underline{\mathbb{P}}(\hat{P}_{X^n} \in \Pi)$$

por lo que

$$\liminf_{n\to\infty} \frac{1}{n} \log \underline{\mathbb{P}}(\hat{P}_{X^n} \in int_{\tau_0}\Pi) \leq \liminf_{n\to\infty} \frac{1}{n} \log \underline{\mathbb{P}}(\hat{P}_{X^n} \in \Pi).$$

Aplicando (3.18) al conjunto $int_{\tau_0}\Pi$ tenemos que

$$-D(\Pi||P_X) = -D(int_{\tau_0}\Pi||P_X) \leq \liminf_{n\to\infty} \frac{1}{n} \log \underline{\mathbb{P}}(\hat{P}_{X^n} \in int_{\tau_0}\Pi)$$

$$\leq \liminf_{n\to\infty} \frac{1}{n} \log \underline{\mathbb{P}}(\hat{P}_{X^n} \in \Pi),$$

es decir,

$$-D(\Pi||P_X) \leq \liminf_{n\to\infty} \frac{1}{n} \log \underline{\mathbb{P}}(\hat{P}_{X^n} \in \Pi).$$

Entonces mostremos que efectivamente se da (3.18). Es necesario suponer que $D(\Pi||P_X) < \infty$ de lo contrario la afirmación sería trivial. Sea $\delta > 0$, como $D(\Pi||P_X)$ es un ínfimo entonces existe $P \in \Pi$ tal que

$$D(P||P_X) < D(\Pi||P_X) + \delta. \tag{3.19}$$

Como Π es abierto relativo de $\Pi \cup \Lambda_f$ respecto a la topología τ_0 para P existen $A = (A_1, ...A_k) \in \Theta$ y $\epsilon > 0$ tales que $U_0(P, A, \epsilon) \cap \Lambda_f \subset \Pi$. Así,

$$\mathbb{P}(\hat{P}_{X^n} \in U_0(P, A, \epsilon)) = \underline{\mathbb{P}}(\hat{P}_{X^n} \in U_0(P, A, \epsilon)) = \underline{\mathbb{P}}(\hat{P}_{X^n} \in U_0(P, A, \epsilon) \cap \Lambda_f)$$

$$\leq \underline{\mathbb{P}}(\hat{P}_{X^n} \in \Pi),$$

es decir,

$$\mathbb{P}(\hat{P}_{X^n} \in U_0(P, A, \epsilon)) \leq \underline{\mathbb{P}}(\hat{P}_{X^n} \in \Pi)$$

Ahora si tomamos $\tilde{P}_n = (r_1^n = \frac{n_1^n}{n}, ..., r_k^n = \frac{n_k^n}{n})$ ocurre que $\tilde{P}_n(i)$ converge a $P(A_i)$ cuando n tiende a infinito, $\tilde{P}_n(i) = 0$ si $P(A_i) = 0$ y la misma identificación de la partición $A = (A_1, ..., A_k)$ con el conjunto $K = \{1, ..., k\}$, cf. (i) del Teorema 3.1.1. Entonces existe un $N \in \mathbb{N}$ tal que si $n > N$ sucede que

$$(n+1)^{-k} e^{-nD(\tilde{P}_n||P_X^A)} \leq (P_X^A)^n(\mathcal{T}_n(\tilde{P}_n)) = P_X^n(\{s|\hat{P}_{s,n}^A = \tilde{P}_n\})$$

$$= \mathbb{P}(\hat{P}_{X^n}(A_i) = r_i^n, \quad i = 1, ..., k) \leq \mathbb{P}(\hat{P}_{X^n} \in U_0(P, A, \epsilon))$$

En donde la primera desigualdad se da por el Lema 3.1.2. Para mostrar el porqué de la segunda desigualdad afirmamos que

$$\{s|\hat{P}_{s,n}(A_i) = r_i^n \;\; i = 1, ..., k\} \subset \{s|\hat{P}_{s,n} \in U_0(P, A, \epsilon)\}.$$

En efecto, como $\tilde{P}_n(A_i)$ converge a $P(A_i)$ entonces para ϵ existe un $N \in \mathbb{N}$ tal que si $n > N$ se tiene que

$$|r_i^n - P(A_i)| < \epsilon, \quad i = 1, ..., k. \tag{3.20}$$

Así, si s es tal que $\hat{P}_{s,n}(A_i) = r_i^n$ entonces $|\hat{P}_{s,n}(A_i) - P(A_i)| < \epsilon$. Además, siempre sucede que $\hat{P}_{s,n}(A_i) = 0$ si $P(A_i) = 0$, luego $\hat{P}_{s,n} \in U_0(P, A, \epsilon)$ por lo que efectivamente $\{s|\hat{P}_{s,n}(A_i) = r_i^n \quad i = 1, ..., k\} \subset \{s|\hat{P}_{s,n} \in U_0(P, A, \epsilon)\}$. En suma, tenemos que

$$\underline{\mathbb{P}}(\hat{P}_{X^n} \in \Pi) \geq (n+1)^{-k} e^{-nD(\tilde{P}_n||P_X^A)}, \tag{3.21}$$

por (3.19)

$$\liminf_{n\to\infty} D(\tilde{P}_n||P_X^A) = \lim_{n\to\infty} \sum_{i=1}^{k} r_i^n \log \frac{r_i^n}{P_X(A_i)} = \sum_{i=1}^{k} P(A_i) \log \frac{P(A_i)}{P_X(A_i)}$$

$$= D(P^A||P_X^A) \leq D(P||P_X) \leq D(\Pi||P_X) + \delta,$$

es decir,

$$\liminf_{n\to\infty} -D(\tilde{P}_n||P_X^A) \geq -D(\Pi||P_X) - \delta. \tag{3.22}$$

Por último, de (3.21) y (3.22) se sigue

$$\liminf_{n\to\infty} \frac{1}{n} \log \underline{\mathbb{P}}(\hat{P}_{X^n} \in \Pi) \geq \liminf_{n\to\infty} \frac{1}{n} \log \Big[(n+1)^{-k} e^{-nD(\tilde{P}_n||P_X^A)} \Big]$$

$$= \liminf_{n\to\infty} \frac{1}{n} \log \Big[(n+1)^{-k} \Big] + \liminf_{n\to\infty} \frac{1}{n} \log \Big[e^{-nD(\tilde{P}_n||P_X^A)} \Big]$$

$$= 0 + \liminf_{n\to\infty} -D(\tilde{P}_n||P_X^A) \geq -D(\Pi||P_X) - \delta,$$

es decir,

$$\liminf_{n\to\infty} \frac{1}{n} \log \underline{\mathbb{P}}(\hat{P}_{X^n} \in \Pi) \geq -D(\Pi||P_X) - \delta. \tag{3.23}$$

Como δ era arbitrario de (3.23) se obtiene la cota deseada:

$$\liminf_{n\to\infty} \frac{1}{n} \log \underline{\mathbb{P}}(\hat{P}_{X^n} \in \Pi) \geq -D(\Pi||P_X).$$

□

Lema 3.2.2. *Si* $\Pi \subset \Lambda$ *es completamente convexo y* $\Pi' \subset \Pi$ *es tal que* $\bar{\mathbb{P}}(\hat{P}_{X^n} \in \Pi') > 0$ *entonces* $P_{X|\hat{P}_n \in \Pi'} \in \Pi$.

Prueba. Sea

$$A' = \{s | \hat{P}_n(s, \cdot) \in \Pi'\}. \tag{3.24}$$

Como las variables $X_1, ..., X_n$ son independientes entonces

$$P_{X|\hat{P}_n \in \Pi'} = P_{X_i|A'}$$

es la distribución marginal uno-dimensional (que no depende de i) de

$$P_{X^n|\hat{P}_n \in \Pi'} = P_{X^n|A'};$$

esto último está definido como se hizo en la Definición 3.2.2, i.e.,

$$P_{X^n|A'}(B) = \frac{P_X^n(B \cap A')}{P_X^n(B)}, \quad B \in \mathcal{B}^n$$

siempre y cuando $A' \in \mathcal{B}^n$ y si no sucede esto entonces tomamos la probabilidad superior, es decir, $P_{X^n|A'} = P_{X^n|A}$ en donde $A \in \mathcal{B}^n$, $A' \subset A$ y $P_X^n(A)$ es mínima sujeta a estas condiciones, i.e.,

$$P_X^n(A) = \bar{\mathbb{P}}(\hat{P}_{X^n} \in \Pi').$$

Ahora, recordemos que la medida empírica está definida a partir de la función

$$\hat{P}_n(s, B) = \frac{1}{n} \sum_{i=1}^{n} 1_B(s_i), \quad B \in \mathcal{B}.$$

Como $\hat{P}_n(\cdot, B)$ es medible para cualquier $B \in \mathcal{B}$ y

$$1_B(s_i) = 1_B(s_i) \prod_{j \neq i} 1_S(s_j)$$

entonces al integrar respecto a $P_{X^n|A'}$ nos da que para cualquier $B \in \mathcal{B}$ sucede lo siguiente

$$\int_{S^n} \hat{P}_n(s, B) \; dP_{X^n|A'}(s) = \int_{S^n} \frac{1}{n} \sum_{i=1}^{n} 1_B(s_i) \; dP_{X^n|A'}(s)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \int_{S^n} 1_B(s_i) \; dP_{X|A'}^n(s) = \frac{1}{n} \sum_{i=1}^{n} \int_{S^n} 1_B(s_i) \prod_{j \neq i} 1_S(s_j) \; dP_{X|A'}^n(s)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \int_S 1_B(s_i) \; dP_{X|A'}(s_i) \prod_{j \neq i} \int_S 1_S(s_j) \; dP_{X|A'}(s_j)$$

$$= \frac{1}{n} \sum_{i=1}^{n} P_{X|A'}(B) = P_{X|A'}(B) = P_{X|\hat{P}_n \in \Pi'}(B).$$

En donde hacemos uso del teorema de Fubini[15] para la medida producto $P_{X|A'}^n$. En resumen, se tiene que

$$\int \hat{P}_n(s, B) \; dP_{X^n|A'}(s) = P_{X|\hat{P}_n \in \Pi'}(B) \tag{3.25}$$

Consideremos dos casos.

Caso 1: $A' \in \mathcal{B}^n$.

[15] cf. Teorema 4.4.10 del apéndice.

En este caso la integral puede ser restringida a A' pues la medida $P_{X^n|\hat{P}_n\in\Pi'}$ está restringida a este conjunto, es decir, $P_{X^n|\hat{P}_n\in\Pi'}(A') = 1$. Recordemos que $\hat{P}_n(\cdot,\cdot)$ cumple con la definición de kernel de Markov[16], entonces $\hat{P}_n(s,\cdot)$ es medida de probabilidad para cada $s \in S^n$ en particular para cada $s \in A'$. Luego como (3.24) nos da que $\hat{P}_n(s,\cdot) \in \Pi' \subset \Pi$ con $s \in A'$ entonces $\hat{P}_n(s,\cdot) \in \Pi$ que es completamente convexo. Así, por definición (cf. Definición 3.2.5), se tiene que para el espacio de probabilidad $(S^n, \mathcal{B}^n, P_{X^n|A'})$ y el kernel de Markov $\hat{P}_n$ la medida

$$\int \hat{P}_n(s,B)\ \ dP_{X|A'}(s), \quad B \in \mathcal{B}$$

pertenece a Π. Gracias a (3.25) obtenemos que $P_{X|\hat{P}_n\in\Pi'} \in \Pi$.

Caso 2: $A' \notin \mathcal{B}^n$.

En este caso por el Lema 4.1.15 del apéndice se tiene que

$$\mathcal{O} = \{F = E \cap A' | E \in \mathcal{B}^n\}$$

es una σ-álgebra de subconjuntos de A' y $(A', \mathcal{O}, P')$ es un espacio de probabilidad con $P'(F) = \bar{P}_{X^n|A'}(E)$ y $F = E \cap A'$.

Gracias a esto tenemos que

$$\int_{A'} \hat{P}_n(s,B)\ \ dP'(s) = \int \hat{P}_n(s,B)\ \ dP_{X^n|A'}(s).$$

De (3.25) se sigue

$$\int_{A'} \hat{P}_n(s,B)\ \ dP'(s) = P_{X|\hat{P}_n\in\Pi'}(B).$$

Análogamente se tiene que para el espacio de probabilidad $(A', \mathcal{O}, P')$ y el kernel de Markov $\hat{P}_n$

$$\int_{A'} \hat{P}_n(\cdot,B)\ \ dP' \in \Pi,$$

es decir, $P_{X|\hat{P}_n\in\Pi'} \in \Pi$.

□

Lema 3.2.3. *Sean $P_i \in \Lambda$, $Q_i \in \Lambda$, i=1,...,n medidas de probabilidad sobre $(S, \mathcal{B})$. Sea $P^{(n)}$ una medida de probabilidad sobre $(S^n, \mathcal{B}^n)$ tal que cada P_i es su i-ésima marginal. Entonces*

$$D(P^{(n)}||Q_1 \otimes ... \otimes Q_n) = D(P^{(n)}||P_1 \otimes ... \otimes P_n) + \sum_{i=1}^{n} D(P_i||Q_i). \quad (3.26)$$

[16]cf. Definición 3.2.4.

Prueba. Primero es necesario suponer que $P^{(n)} \ll P_1 \otimes ... \otimes P_n$ y $P_i \ll Q_i$ para cada $i = 1, ..., n$, ya que de lo contrario la prueba es trivial pues ambos lados de la igualdad son ∞. Con la suposición anterior se tiene que si $s = (s_1, ..., s_n) \in S^n$ entonces

$$\frac{dP^{(n)}}{d(P_1 \otimes ... \otimes P_n)}(s) \prod_{i=1}^{n} \frac{dP_i}{dQ_i}(s_i) = \frac{dP^{(n)}}{d(P_1 \otimes ... \otimes P_n)}(s) \frac{d(P_1 \otimes ... \otimes P_n)}{d(Q_1 \otimes ... \otimes Q_n)}(s)$$

$$= \frac{dP^{(n)}}{d(Q_1 \otimes ... \otimes Q_n)}(s).$$

La anterior igualdad se sigue de la propiedad de la derivada de Radon-Nikodým de la medida producto, cf. Lema 4.4.10 del apéndice. Resumiendo, se tiene que

$$\frac{dP^{(n)}}{d(Q_1 \otimes ... \otimes Q_n)}(s) = \frac{dP^{(n)}}{d(P_1 \otimes ... \otimes P_n)}(s) \prod_{i=1}^{n} \frac{dP_i}{dQ_i}(s_i).$$

Tomando logaritmo de ambos lados e integrando respecto a la medida $P^{(n)}$ obtenemos

$$\int \log \left[\frac{dP^{(n)}}{d(Q_1 \otimes ... \otimes Q_n)} \right] dP^{(n)} = \int \log \left[\frac{dP^{(n)}}{d(P_1 \otimes ... \otimes P_n)} \prod_{i=1}^{n} \frac{dP_i}{dQ_i} \right] dP^{(n)}$$

$$= \int \log \left[\frac{dP^{(n)}}{d(P_1 \otimes ... \otimes P_n)} \right] dP^{(n)} + \int \log \left[\prod_{i=1}^{n} \frac{dP_i}{dQ_i} \right] dP^{(n)}$$

$$= D(P^{(n)} || P_1 \otimes ... \otimes P_n) + \int \sum_{i=1}^{n} \log \left[\frac{dP_i}{dQ_i} \right] dP^{(n)}$$

$$= D(P^{(n)} || P_1 \otimes ... \otimes P_n) + \sum_{i=1}^{n} \int \log \left[\frac{dP_i}{dQ_i} \right] dP_i$$

$$= D(P^{(n)} || P_1 \otimes ... \otimes P_n) + \sum_{i=1}^{n} D(P_i || Q_i),$$

es decir,

$$D(P^{(n)} || Q_1 \otimes ... \otimes Q_n) = D(P^{(n)} || P_1 \otimes ... \otimes P_n) + \sum_{i=1}^{n} D(P_i || Q_i).$$

□

Proseguimos con la demostración del Teorema 3.2.1.

Demostración. Sea $\Pi' \subset \Pi$ tal que $\bar{\mathbb{P}}(\hat{P}_n(s,\cdot) \in \Pi') > 0$, como Π es casi completamente convexo entonces existen $\{\Pi_i\}_{i=1}^{\infty}$ conjuntos completamente convexos con $\Pi_i \subset \Pi$ para toda i tales que $\Pi_i \subset \Pi_{i+1}$ y $\Pi \cap \Lambda_f \subset \cup_{i=1}^{\infty}\Pi_i$. Ahora, consideremos los conjuntos $\Pi_i' = \Pi_i \cap \Pi'$, como $\Pi' \subset \Pi$ y

$$\bar{\mathbb{P}}(\hat{P}_n(s,\cdot) \in \Pi') > 0$$

(recordemos que $\hat{P}_n(s,\cdot) \in \Lambda_f$) entonces existe al menos una i para la cual $\Pi_i' \neq \emptyset$ además $\Pi_i' \subset \Pi_{i+1}'$. Observemos que se da lo siguiente:

$$\{s|\hat{P}_n(s,\cdot) \in \Pi'\} = \bigcup_{i=1}^{\infty}\{s|\hat{P}_n(s,\cdot) \in \Pi_i'\}, \tag{3.27}$$

ya que si $s \in \{s|\hat{P}_n(s,\cdot) \in \Pi'\}$ entonces

$$s \in \{s|\hat{P}_n(s,\cdot) \in \Pi \cap \Lambda_f\} \subset \{s|\hat{P}_n(s,\cdot) \in \cup_{i=1}^{\infty}\Pi_i\},$$

es decir, $\hat{P}_n(s,\cdot) \in \Pi_i$ para algún i y además $\hat{P}_n(s,\cdot) \in \Pi'$ luego $\hat{P}_n(s,\cdot) \in \Pi_i'$, i.e., $s \in \{s|\hat{P}_n(s,\cdot) \in \Pi_i'\}$ para algún i por lo tanto

$$s \in \cup_{i=1}^{\infty}\{s|\hat{P}_n(s,\cdot) \in \Pi_i'\}.$$

La otra contención es clara pues si $s \in \cup_{i=1}^{\infty}\{s|\hat{P}_n(s,\cdot) \in \Pi_i'\}$ entonces $\hat{P}_n(s,\cdot) \in \Pi_i'$ para algún i, en particular $\hat{P}_n(s,\cdot) \in \Pi'$, luego

$$s \in \{s|\hat{P}_n(s,\cdot) \in \Pi'\}.$$

Fijemos n y consideremos a los conjuntos $A \in \mathcal{B}^n$, $A_i \in \mathcal{B}^n$, i=1,2,... que cumplen $\{s|\hat{P}_n(s,\cdot) \in \Pi'\} \subset A$, $\{s|\hat{P}_n(s,\cdot) \in \Pi_i'\} \subset A_i$ y su medida bajo P_X^n es mínima bajo estas condiciones, i.e.,

$$P_X^n(A) = \text{mín}\{P_X^n(B)|\{s|\hat{P}_n(s,\cdot) \in \Pi'\} \subset B \quad B \in \mathcal{B}^n\}$$

y

$$P_X^n(A_i) = \text{mín}\{P_X^n(B)|\{s|\hat{P}_n(s,\cdot) \in \Pi_i'\} \subset B \quad B \in \mathcal{B}^n\}.$$

Como $\Pi_i' \subset \Pi_{i+1}'$ y gracias a (3.27) estos conjuntos pueden ser elegidos de tal manera que cumplan que

$$A_i \subset A_{i+1} \quad \text{y} \quad A = \bigcup_{i=1}^{\infty} A_i, \tag{3.28}$$

como $\{A_i\}_{i=1}^{\infty}$ es una sucesión creciente de eventos entonces se tiene que

$$\lim_{i\to\infty} P_X^n(A_i) = P_X^n(A) = \bar{\mathbb{P}}(\hat{P}_{X^n} \in \Pi'). \tag{3.29}$$

Podemos asumir sin pérdida de generalidad que $P_X^n(A_i) > 0$ para cada i ya que si no sucede esto entonces encontramos el primer conjunto A_i para el

cual sucede esto, eliminamos los anteriores y comenzamos a contar a partir de este; lo anterior lo podemos hacer pues sabemos que

$$\bar{\mathbb{P}}(\hat{P}_{X^n} \in \Pi') > 0$$

y por (3.27) se tiene que existe un $k \in \mathbb{N}$ para el cual $P_X^n(A_k) > 0$. Consideremos las siguientes medidas de probabilidad[17]:

$$P_{X^n|\hat{P}_n \in \Pi'} = P_{X^n|A} \quad \text{y} \quad P_{X^n|\hat{P}_n \in \Pi_i'} = P_{X^n|A_i}. \tag{3.30}$$

Aplicando la identidad (3.26) con $P^{(n)} = P_{X^n|A_i}$, $Q_i = P_X$ para toda i y como $P_{X|A_i}$ es la marginal de dimensión 1 de $P_{X^n|A_i}$ obtenemos

$$D(P_{X^n|A_i}||P_X^n) = D(P_{X^n|A_i}||P_{X|A_i}^n) + \sum_{j=1}^{n} D(P_{X|A_i}||P_X)$$

$$= D(P_{X^n|A_i}||P_{X|A_i}^n) + nD(P_{X|A_i}||P_X) \geq nD(P_{X|A_i}||P_X).$$

Por otro lado, ya que $P_{X^n|A}(B) = \frac{P_X^n(B \cap A)}{P_X^n(A)}$ tenemos que

$$D(P_{X^n|A_i}||P_X^n) = \int \log\left[\frac{dP_{X^n|A_i}}{dP_X^n}\right] dP_{X^n|A_i} = \frac{1}{P_X^n(A_i)} \int_{A_i} \log\left[\frac{dP_{X^n|A_i}}{dP_X^n}\right] dP_X^n$$

$$= \frac{1}{P_X^n(A_i)} \int_{A_i} \log\left[\frac{1}{P_X^n(A_i)}\right] + \log\left[\frac{dP_X^n}{dP_X^n}\right] \; dP_X^n = -\log P_X^n(A_i).$$

De este modo

$$-\log P_X^n(A_i) \geq nD(P_{X|A_i}||P_X)$$

o escrito de otra manera:

$$\log P_X^n(A_i) \leq -nD(P_{X|A_i}||P_X). \tag{3.31}$$

Ahora, como Π_i es completamente convexo, $\Pi_i' \subset \Pi_i$ y $\bar{\mathbb{P}}(\hat{P}_{X^n} \in \Pi_k') = P_X^n(A_i) > 0$ y recordando la notación de (3.30), i.e., $P_{X|A_i} = P_{X|\hat{P}_n \in \Pi_i'}$ (ya que es la marginal de $P_{X^n|A_i}$) entonces por el Lema 3.2.2 se sigue que

$$P_{X|A_i} = P_{X|\hat{P}_n \in \Pi_i'} \in \Pi_i \subset \Pi, \tag{3.32}$$

lo cual nos da que $D(\Pi||P_X) \leq D(P_{X|A_i}||P_X)$, i.e.,

$$-D(P_{X|A_i}||P_X) \leq -D(\Pi||P_X).$$

[17] cf. Definición 3.2.2.

Por (3.29) y (3.31) se tiene que

$$\frac{1}{n}\log\bar{\mathbb{P}}(\hat{P}_{X^n}\in\Pi') = \lim_{i\to\infty}\frac{1}{n}\log P_X^n(A_i) = \lim_{i\to\infty}\frac{-n}{n}D(P_{X|A_i}||P_X)$$

$$\leq -D(\Pi||P_X),$$

es decir,

$$\frac{1}{n}\log\bar{\mathbb{P}}(\hat{P}_{X^n}\in\Pi') \leq -D(\Pi||P_X). \tag{3.33}$$

Como Π' es cualquier subconjunto de Π y no es necesariamente propio entonces en particular se cumple para $\Pi' = \Pi$, es decir,

$$\frac{1}{n}\log\bar{\mathbb{P}}(\hat{P}_{X^n}\in\Pi) \leq -D(\Pi||P_X),$$

con lo cual hemos probado (3.15). Para mostrar el resto del teorema utilizamos de nuevo la identidad (3.26) pero ahora con $Q_i = P^*$ para toda i, en donde P^* es la I-proyección generalizada de P_X en Π (su existencia está asegurada por la hipótesis $D(\Pi||P_X) < \infty$), es decir, se tiene que

$$D(P_{X^n|A_i}||P^{*n}) = D(P_{X^n|A_i}||P^n_{X|A_i}) + \sum_{j=1}^{n} D(P_{X|A_i}||P^*)$$

$$= D(P_{X^n|A_i}||P^n_{X|A_i}) + nD(P_{X|A_i}||P^*),$$

es decir,

$$D(P_{X^n|A_i}||P^{*n}) = D(P_{X^n|A_i}||P^n_{X|A_i}) + nD(P_{X|A_i}||P^*). \tag{3.34}$$

Ahora, por (3.32) y la identidad (2.2) del Teorema 2.1.2 se tiene que

$$D(P_{X|A_i}||P_X) \geq D(P_{X|A_i}||P^*) + D(\Pi||P_X). \tag{3.35}$$

De nuevo como

$$-\log P_X^n(A_i) = D(P_{X^n|A_i}||P_X^n) = D(P_{X^n|A_i}||P^n_{X|A_i}) + nD(P_{X|A_i}||P_X),$$

por (3.35) y (3.34) se tiene que

$$-\log P_X^n(A_i) = D(P_{X^n|A_i}||P^n_{X|A_i}) + nD(P_{X|A_i}||P_X)$$

$$\geq D(P_{X^n|A_i}||P^n_{X|A_i}) + nD(P_{X|A_i}||P^*) + nD(\Pi||P_X)$$

$$= D(P_{X^n|A_i}||P^{*n}) + nD(\Pi||P_X);$$

al multiplicar por -1 se sigue que

$$\log P_X^n(A_i) \leq -D(P_{X^n|A_i}||P^{*n}) - nD(\Pi||P_X). \tag{3.36}$$

Por último, gracias a (3.28) y de nuevo por cómo está definida $P_{X^n|A}$, tenemos que

$$D(P_{X^n|A_i}||P^{*n}) = \int \log\left[\frac{dP_{X^n|A_i}}{dP^{*n}}\right] dP_{X^n|A_i}$$

$$= \frac{1}{P_X^n(A_i)} \int_{A_i} \log\left[\frac{1}{P_X^n(A_i)}\right] + \log\left[\frac{dP_X^n}{dP^{*n}}\right] dP_X^n = -\log P_X^n(A_i) + D(P_X^n||P^{*n}).$$

Analógamente se tiene que $-\log P_X^n(A) + D(P_X^n||P^{*n}) = D(P_{X^n|A}||P^{*n})$. Así,

$$\lim_{i\to\infty} D(P_{X^n|A_i}||P^{*n}) = \lim_{i\to\infty} -\log P_X^n(A_i) + D(P_X^n||P^{*n})$$

$$= -\log P_X^n(A) + D(P_X^n||P^{*n}) = D(P_{X^n|A}||P^{*n}),$$

por lo que tenemos que $D(P_{X^n|A_i}||P^{*n})$ converge a $D(P_{X^n|A}||P^{*n})$ cuando i tiende a infinito, gracias a esto y a (3.36) se obtiene que

$$\lim_{i\to\infty} \frac{1}{n} \log P_X^n(A_i) \leq \lim_{i\to\infty} -\frac{1}{n} D(P_{X^n|A_i}||P^{*n}) - D(\Pi||P_X)$$

$$= -\frac{1}{n} D(P_{X^n|A}||P^{*n}) - D(\Pi||P_X).$$

Recordando (3.29) tenemos entonces que

$$\frac{1}{n} \log \bar{\mathbb{P}}(\hat{P}_n \in \Pi') \leq -\frac{1}{n} D(P_{X^n|A}||P^{*n}) - D(\Pi||P_X),$$

escribiendo la desigualdad anterior con la notación de (3.30) se tiene

$$\frac{1}{n} \log \bar{\mathbb{P}}(\hat{P}_{X^n} \in \Pi') \leq -\frac{1}{n} D(P_{X^n|\hat{P}_n \in \Pi'}||P^{*n}) - D(\Pi||P_X) \tag{3.37}$$

que es la identidad (3.16) que se quería demostrar. Para terminar la demostración del teorema observemos que de (3.37) se sigue que

$$\frac{1}{n} \log \bar{\mathbb{P}}(\hat{P}_{X^n} \in \Pi') \leq -D(\Pi||P_X) \quad \forall n \in \mathbb{N},$$

es decir,

$$\limsup_{n\to\infty} \frac{1}{n} \log \bar{\mathbb{P}}(\hat{P}_{X^n} \in \Pi') \leq -D(\Pi||P_X).$$

Además, la hipótesis (3.17) nos da lo siguiente:

$$D(\Pi'||P_X) \leq D(int_{\tau_0}\Pi'||P_X) = D(\Pi||P_X) \leq D(\Pi'||P_X),$$

ya que $int_{\tau_0}\Pi' \subset \Pi'$ y $\Pi' \subset \Pi$, por lo tanto $D(\Pi||P_X) = D(\Pi'||P_X)$ y así se tiene que

$$\limsup_{n\to\infty} \frac{1}{n} \log \bar{\mathbb{P}}(\hat{P}_{X^n} \in \Pi') \leq -D(\Pi'||P_X). \tag{3.38}$$

Además, $D(\Pi'||P_X) = D(int_{\tau_0}\Pi'||P_X)$, luego por el Lema 3.2.1 se tiene que

$$\liminf_{n\to\infty} \frac{1}{n} \log \underline{\mathbb{P}}(\hat{P}_n \in \Pi') \geq -D(\Pi'||P_X). \tag{3.39}$$

Observemos que (3.38) y (3.39) nos dan que Π' tiene la propiedad de Sanov. En efecto, como $\underline{\mathbb{P}} \leq \bar{\mathbb{P}}$ entonces se sigue que

$$-D(\Pi'||P_X) \leq \liminf_{n\to\infty} \frac{1}{n} \log \underline{\mathbb{P}}(\hat{P}_{X^n} \in \Pi') \leq \liminf_{n\to\infty} \frac{1}{n} \log \bar{\mathbb{P}}(\hat{P}_{X^n} \in \Pi')$$

$$\leq \limsup_{n\to\infty} \frac{1}{n} \log \bar{\mathbb{P}}(\hat{P}_{X^n} \in \Pi') \leq -D(\Pi'||P_X),$$

i.e.,

$$\lim_{n\to\infty} \frac{1}{n} \log \bar{\mathbb{P}}(\hat{P}_{X^n} \in \Pi') = -D(\Pi'||P_X).$$

También

$$-D(\Pi'||P_X) \leq \liminf_{n\to\infty} \frac{1}{n} \log \underline{\mathbb{P}}(\hat{P}_{X^n} \in \Pi') \leq \limsup_{n\to\infty} \frac{1}{n} \log \underline{\mathbb{P}}(\hat{P}_{X^n} \in \Pi')$$

$$\leq \limsup_{n\to\infty} \frac{1}{n} \log \bar{\mathbb{P}}(\hat{P}_{X^n} \in \Pi') \leq -D(\Pi'||P_X),$$

i.e.,

$$\lim_{n\to\infty} \frac{1}{n} \log \underline{\mathbb{P}}(\hat{P}_{X^n} \in \Pi') = -D(\Pi'||P_X).$$

Ahora, por (3.15) se tiene que

$$\limsup_{n\to\infty} \frac{1}{n} \log \bar{\mathbb{P}}(\hat{P}_{X^n} \in \Pi) \leq -D(\Pi||P_X), \tag{3.40}$$

como $\Pi' \subset \Pi$ entonces $int_{\tau_0}\Pi' \subset int_{\tau_0}\Pi$, luego (de nuevo bajo la hipótesis (3.17)) se tiene $D(\Pi||P_X) \leq D(int_{\tau_0}\Pi||P_X) \leq D(int_{\tau_0}\Pi'||P_X) = D(\Pi||P_X)$, i.e., $D(int_{\tau_0}\Pi||P_X) = D(\Pi||P_X)$, por lo cual gracias al Lema 3.2.1 se tiene que

$$\liminf_{n\to\infty} \frac{1}{n} \log \underline{\mathbb{P}}(\hat{P}_{X^n} \in \Pi) \geq -D(\Pi||P_X). \tag{3.41}$$

De manera análoga por (3.40) y (3.41) se obtiene que Π tiene la propiedad de Sanov. Por último, bajo la misma hipótesis (3.17) se tiene que (3.16) conlleva a la cuasiindependencia asintótica de las variables aleatorias $X_1, ..., X_n$ bajo la condición $\hat{P}_n \in \Pi'$ con distribución límite P^* pues

$$\lim_{n\to\infty} \frac{1}{n} \log \bar{\mathbb{P}}(\hat{P}_{X^n} \in \Pi') \leq \lim_{n\to\infty} -\frac{1}{n} D(P_{X^n|\hat{P}_n\in\Pi'}||P^{*n}) - D(\Pi||P_X),$$

pero como Π' tiene la propiedad de Sanov tenemos que

$$\lim_{n\to\infty} \frac{1}{n} \log \bar{\mathbb{P}}(\hat{P}_{X^n} \in \Pi') = -D(\Pi||P_X)$$

por ende

$$\lim_{n\to\infty} -\frac{1}{n} D(P_{X^n|\hat{P}_n\in\Pi'}||P^{*n}) = 0$$

o lo que es lo mismo

$$\lim_{n\to\infty} \frac{1}{n} D(P_{X^n|\hat{P}_n\in\Pi'}||P^{*n}) = 0.$$

□

Este teorema es de gran importancia por varios factores, el primero es que (3.15) nos proporciona la ya conocida cota del teorema de Sanov pero no lo hace en el límite sino que acota dicha probabilidad para cada n. En segundo lugar, observemos que la identidad (3.26) del Lema 3.2.3 con $P_1 = P_2 = ... = P_n = P$ y $Q_1 = Q_2 = ... = Q_n = Q$ nos da que

$$nD(P||Q) \leq D(P^{(n)}||P^n) + nD(P||Q) = D(P^{(n)}||P^n) + \sum_{i=1}^{n} D(P||Q)$$

$$= D(P^{(n)}||Q^n).$$

Así,

$$D(P||Q) \leq \frac{1}{n} D(P^{(n)}||Q^n).$$

Ahora, en esta desigualdad tomando $P = P_{X|\hat{P}_n\in\Pi'}$, $Q = P^*$ y $P^{(n)} = P_{X^n|\hat{P}_n\in\Pi'}$ y multiplicando por -1 obtenemos

$$-\frac{1}{n}D(P_{X^n|\hat{P}_n\in\Pi'}||P^{*n}) \leq -D(P_{X|\hat{P}_n\in\Pi'}||P^*).$$

Esta última desigualdad sumada con (3.16), es decir,

$$\frac{1}{n}\log\bar{\mathbb{P}}(\hat{P}_{X^n}\in\Pi') \leq -D(\Pi||P_X) - \frac{1}{n}D(P_{X^n|\hat{P}_n\in\Pi'}||P^{*n}),$$

nos dan lo siguiente:

$$\begin{aligned}\frac{1}{n}\log\bar{\mathbb{P}}(\hat{P}_{X^n}\in\Pi') &\leq -D(\Pi||P_X) - \frac{1}{n}D(P_{X^n|\hat{P}_n\in\Pi'}||P^{*n}) \\ &\leq -D(\Pi||P_X) - D(P_{X|\hat{P}_n\in\Pi'}||P^*),\end{aligned}$$

esto es,

$$D(P_{X|\hat{P}_n\in\Pi'}||P^*) \leq -\frac{1}{n}\log\bar{\mathbb{P}}(\hat{P}_{X^n}\in\Pi') - D(\Pi||P_X).$$

Con esto obtenemos una relación entre la velocidad de convergencia del teorema de Sanov y la velocidad de convergencia en información de la medida de probabilidad $P_{X|\hat{P}_n\in\Pi'}$ hacia P^* que es la I-proyección generalizada de la distribución común P_X de las variables aleatorias $\{X_n\}_{n\in\mathbb{N}}$. En tercer y último lugar se ha establecido la cuasiindependecia asintótica de las variables aleatorias bajo la condición de que su medida empírica se encuentre en Π' con la I-proyección generalizada de P_X como distribución límite. Es decir, por un lado ya se sabía que la I-proyección generalizada está intrinsecamente relacionada con la propiedad de Sanov pues es con aquella que se alcanza el ínfimo de las entropías relativas respecto a Q en el conjunto Π. Por otro lado, también se ha establecido una relación directa entre la cuasiindependencia asintótica de las variables aleatorias y la I-proyección generalizada.

3.3. Un teorema límite para un estadístico

En este último capítulo mostaremos un último teorema que es consecuencia del Teorema 2.3.2 y del Teorema 3.2.1. Para ello es necesario enunciar y demostrar un último lema.

Lema 3.3.1. *Sean $\mathcal{F}$ una familia de funciones reales y medibles sobre $(S,\mathcal{B})$ y $\mathcal{K}=\{K_i\}_{i=1}^{\infty}$ una sucesión de subconjuntos de S que cumplen las propiedades enunciadas en (2.13), esto es, $K_i\in\mathcal{B}$, $K_i\subset K_{i+1}$ y cada $f\in\mathcal{F}$ está acotada en K_i, $i=1,2,....$ Entonces los conjuntos de medidas de probabilidad $\Pi(\mathcal{F}|\mathcal{K})$ y $\Pi'(\mathcal{F}|\mathcal{K})$ son casi completamente convexos.*

Prueba. Sea $\mu\nu$ como en la Definición 3.2.5, es decir, $(\Omega,\mathcal{A},\mu)$ es un espacio de probabilidad y ν es un kernel de Markov de $(\Omega,\mathcal{A})$ a $(S,\mathcal{B})$ tal que $\nu(\omega,\cdot)\in\Pi(\mathcal{F}|K_i)$ para cada $\omega\in\Omega$ y recordemos que

$$\mu\nu(B) = \int \nu(\omega, B)\ d\mu(\omega), \quad B \in \mathcal{B}$$

Como $\Pi(\mathcal{F}|\mathcal{K}) = \cup_{i=1}^{\infty}\Pi(\mathcal{F}|K_i)$ debemos mostrar que $\Pi(\mathcal{F}|K_i)$ es completamente convexo para cada i pues

$$\Pi(\mathcal{F}|\mathcal{K}) \cap \Lambda_f = \bigcup_{i=1}^{\infty} \Pi(\mathcal{F}|K_i) \cap \Lambda_f \subset \bigcup_{i=1}^{\infty} \Pi(\mathcal{F}|K_i).$$

Para ello observemos que, definiendo $\nu_\omega = \nu(\omega, \cdot)$, si f es medible y acotada se tiene que

$$\int_S f(s)\ d\mu\nu(s) = \int_\Omega \left[\int_S f(s)\ d\nu_\omega(s)\right]\ d\mu(\omega), \tag{3.42}$$

cf. Lema 4.4.13 del apéndice. Como $\nu(\omega, \cdot) \in \Pi(\mathcal{F}|K_i)$ entonces

$$\int f(s)\ d\nu_\omega(s) \geq 0 \quad \forall \omega \in \Omega,$$

por lo tanto de (3.42) se sigue que

$$\int f\ d\mu\nu \geq 0.$$

Además,

$$\mu\nu(K_i) = \int \nu(\omega, K_i)\ d\mu(\omega) = \int 1\ d\mu(\omega) = 1$$

por lo cual $\mu\nu \in \Pi(\mathcal{F}|K_i)$, es decir, $\Pi(\mathcal{F}|K_i)$ es completamente convexo para cada i y por lo tanto $\Pi(\mathcal{F}|\mathcal{K})$ es casi completamente convexo. La prueba para $\Pi'(\mathcal{F}|\mathcal{K})$ es análoga.

□

Observemos que gracias a este resultado hemos obtenido una clase muy amplia de conjuntos completamente convexos que quedan determinados por la familia de funciones $\mathcal{F}$, a saber

$$\Pi(\mathcal{F}|K_i) = \left\{P \Big| \int f\ dP \geq 0, \quad f \in \mathcal{F},\ \text{y}\ P(K_i) = 1\right\}$$

y

$$\Pi'(\mathcal{F}|K_i) = \left\{P \Big| \int f\ dP > 0, \quad f \in \mathcal{F}\ \text{y}\ P(K_i) = 1\right\}, \quad i = 1, 2, ...$$

Con ello también hemos obtenido una una clase muy amplia de conjuntos casi completamente convexos, a saber $\Pi(\mathcal{F}|\mathcal{K})$ y $\Pi'(\mathcal{F}|\mathcal{K})$.

Recordemos a los conjuntos

$$\Pi(\mathcal{F}) = \left\{ P \in \Lambda \Big| \int f \quad dP \geq 0 \quad f \in \mathcal{F} \right\}$$

y

$$\Pi'(\mathcal{F}) = \left\{ P \in \Lambda \Big| \int f \quad dP > 0 \quad f \in \mathcal{F} \right\},$$

observemos que si las funciones $f \in \mathcal{F}$ no son acotadas en S entonces $\Pi(\mathcal{F})$ y $\Pi'(\mathcal{F})$ no son completamente convexos ya que en general $\int f \ d\mu\nu$ puede no existir para cualquier kernel de Markov ν. Por otro lado, notemos que si existe una sucesión $\mathcal{K} = \{K_i\}_{i\in\mathbb{N}}$ tal que $K_i \in \mathcal{B}$, $K_i \subset K_{i+1}$, f es acotada en cada K_i para toda i y $S = \cup_{i\in\mathbb{N}} K_i$ entonces

$$\Pi(\mathcal{F}) = \bigcup_{i\in\mathbb{N}} \Pi(\mathcal{F}|K_i) = \Pi(\mathcal{F}|\mathcal{K}) \text{ y } \Pi'(\mathcal{F}) = \bigcup_{i\in\mathbb{N}} \Pi'(\mathcal{F}|K_i) = \Pi'(\mathcal{F}|\mathcal{K}),$$

por el Lema 3.3.1 se tiene que $\Pi(\mathcal{F})$ y $\Pi'(\mathcal{F})$ son casi completamemte convexos. Lo anterior nos muestra que[18]

$$CCC \twoheadrightarrow CC.$$

Para terminar este capítulo mostraremos el siguiente resultado que es una aplicación de los teoremas 2.3.2 y 3.2.1 a un estadístico, esto es, una función medible sobre un espacio vectorial. En este caso consideraremos un espacio localmente convexo.

Teorema 3.3.1. *Sean $(V, \mathcal{C})$ un espacio medible en donde V es un espacio localmente convexo, $\Psi : (S, \mathcal{B}) \to (V, \mathcal{C})$ una función medible, Q la medida de probabilidad generada por P_X mediante Ψ (i.e., el push-forward de P_X) y supongamos que Q es convexa-tensa. Sean $C \subset V$ convexo tal que*

$$int(C) \cap conv(sop(Q)) \neq \emptyset,$$

y

$$D = \sup_{\vartheta \in V'} \left[\inf_{v\in C} \vartheta(v) - \log\Big(\int e^{\vartheta} dQ \Big) \right]. \tag{3.43}$$

Entonces $D < \infty$ y

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} \Psi(X_i) \in C \right) = -D. \tag{3.44}$$

[18] cf. Definición 3.2.5 y el desarrollo posterior.

Más aún, si en (3.42) tenemos que el supremo es un máximo, sean

$$A_n = \left\{ (s_1, ..., s_n) \Big| \frac{1}{n} \sum_{i=1}^{n} \Psi(s_i) \in C \right\},$$

$\vartheta^* \in V'$ *un funcional lineal con el cual se alcanza el máximo en (3.43) y* P^* *definida mediante la densidad*

$$\frac{dP^*}{dP_X}(s) = ce^{\vartheta^*(\Psi(s))}, \quad s \in S \ \ y \ \ c \in \mathbb{R}.$$

Entonces $X_1, ..., X_n$ *son asintóticamente cuasiindependientes bajo la condición*

$$\frac{1}{n} \sum_{i=1}^{n} \Psi(X_i) \in C, \tag{3.45}$$

con distribución límite P^**, es decir,*

$$\lim_{n\to\infty} \frac{1}{n} D(P_{X^n|A_n} || P^{*n}) = 0.$$

En el caso en que A_n *no sea elemento de* $\mathcal{B}^n$ *entonces (3.44) se interpreta como el hecho de que se cumple la igualdad para la probabilidad superior y para la probabilidad inferior*[19]*.*

Demostración. Por hipótesis tenemos que Q, el push-forward de P_X bajo Ψ, es convexa-tensa[20] por lo cual existen $\mathcal{K} = \{K_i\}_{i=1}^{\infty}$ $(K_i \subset V)$ conjuntos medibles, convexos y compactos tales que $K_i \subset K_{i+1}$ y

$$Q\left(\bigcup_{i=1}^{\infty} K_i \right) = P_X\left(\left\{ s | \Psi(s) \in \bigcup_{i=1}^{\infty} K_i \right\} \right) = 1.$$

Podemos suponer sin pérdida de generalidad que $\Psi(s) \in \cup_{i=1}^{\infty} K_i$ para toda $s \in S$ pues de lo contrario $\Psi(s)$ no estaría en el soporte y no nos proporciona ninguna información. Sean Π, Π' y Π'' los conjuntos de medidas de probabilidad cuyo push-forward bajo Ψ pertenece a los conjuntos $\Pi_0(cd(C))$, $\Pi_0(C)$ y $\Pi_0(int(C))$ respectivamente, cf. (2.34). Esto es,

$$\Pi = \{P \in \Lambda | \widetilde{P} \in \Pi_0(cd(C))\}, \quad \Pi' = \{P \in \Lambda | \widetilde{P} \in \Pi_0(C)\}$$

y

$$\Pi'' = \{P \in \Lambda | \widetilde{P} \in \Pi_0(int(C))\}.$$

Recordemos que

$$\Pi_0(cd(C)) = \Pi(\mathcal{F}|\mathcal{K}), \quad \Pi_0(int(C)) = \Pi'(\mathcal{F}|\mathcal{K})$$

[19] cf. Definición 3.2.1.
[20] cf. Definición 2.3.3.

y $D(\Pi(\mathcal{F}|\mathcal{K})) = D(\Pi'(\mathcal{F}|\mathcal{K}))$ con $\mathcal{F} = \{f|f = a(1-\vartheta),\ \ a \geq 0\ \ \vartheta \in C^{\circ}\}$, es decir, $D(\Pi_0(cd(C))||Q) = D(\Pi_0(int(C))||Q)$ (cf. (2.44)).

Por un lado, observemos que el Lema 2.2.3 nos proporciona que

$$D(\Pi||P_X) = D(\Pi_0(cd(C))||Q) = D(\Pi_0(int(C))||Q) = D(\Pi''||P_X)$$

y

$$D(\Pi'||P_X) = D(\Pi_0(C)||Q).$$

Por otro lado, podemos aplicar el Teorema 2.3.2 al conjunto C y a Q pues éstos cumplen las hipótesis, luego (2.35) nos lleva a que

$$D(\Pi_0(int(C))||Q) < \infty,$$

como $D(\Pi_0(cd(C))||Q) \leq D(\Pi_0(C)||Q) \leq D(\Pi_0(int(C))||Q)$ entonces

$$D(\Pi_0(cd(C))||Q) = D(\Pi_0(C)||Q) = D(\Pi_0(int(C))||Q),$$

i.e.,

$$D(\Pi||P_X) = D(\Pi'||P_X) = D(\Pi''||P_X).$$

Por (2.36) se tiene que

$$D(\Pi_0(int(C))||Q) = \sup_{\vartheta \in V'}\left[\inf_{v \in C} \vartheta(v) - \log\Big(\int e^{\vartheta} dQ\Big)\right],$$

por lo cual se sigue que

$$D = D(\Pi_0(int(C))||Q) = D(\Pi''||P_X) = D(\Pi'||P_X) = D(\Pi||P_X).$$

Además, por (2.37) se tiene que $\widetilde{P}^*$ la I-proyección generalizada de Q en $\Pi_0(C)$ y $\Pi_0(int(C))$ (recordamos que es la misma por el Lema 2.2.2) está dada por la densidad

$$\frac{d\widetilde{P}^*}{dQ}(v) = \frac{e^{\vartheta^*(v)}}{\int e^{\vartheta^*} dQ}.$$

De nuevo por el Lema 2.2.3, específicamente por (2.10), se tiene que

$$\frac{dP^*}{dP_X}(s) = \frac{d\widetilde{P}^*}{dQ}(\Psi(s)) = \frac{e^{\vartheta^*(\Psi(s))}}{\int e^{\vartheta^*} dQ} = ce^{\vartheta^*(\Psi(s))} \quad s \in S.$$

En donde P^* es la I-proyección generalizada de P_X en Π, $\vartheta^* \in V'$ es con el cual se alcanza el máximo para D y $c = \frac{1}{\int e^{\vartheta^*} dQ} < \infty$.

Ahora, observemos que el evento

$$\left\{ s = (s_1, ..., s_n) \Big| \frac{1}{n} \sum_{i=1}^{n} \Psi(s_i) \in C \right\} \tag{3.46}$$

es igual al evento $\{s|\hat{P}_{s,n} \in \Pi'\}$. En efecto, en primer lugar el push-forward de $\hat{P}_{s,n}$ bajo Ψ es

$$\widetilde{\hat{P}}_{s,n}(E) = \frac{1}{n}\sum_{i=1}^{n} 1_{\Psi^{-1}(E)}(s_i) = \frac{1}{n}\sum_{i=1}^{n} 1_E(\Psi(s_i)),$$

luego $\widetilde{\hat{P}}_{s,n}(K_j) = 1$ para alguna j ya que como $\Psi(s_i) \in \cup_{j\in\mathbb{N}} K_j$ entonces cada $s_i \in K_{m_i}$, tomamos a $j = \text{máx}\{m_1, ..., m_n\}$ y así $\Psi(s_i) \in K_j$ para toda $i = 1, ..., n$, de esta manera

$$\widetilde{\hat{P}}_{s,n}(K_j) = \frac{1}{n}\sum_{i=1}^{n} 1_{K_j}(\Psi(s_i)) = 1.$$

Y en segundo lugar las afirmaciones

$$\frac{1}{n}\sum_{i=1}^{n} \Psi(s_i) \in C \quad \text{y} \quad E\Big(\widetilde{\hat{P}}_{s,n}\Big) \in C$$

son equivalentes.

Gracias a esta observación se tiene que

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} \Psi(X_i) \in C\right) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\hat{P}_{X^n} \in \Pi'),$$

si Π' tiene la propiedad de Sanov entonces

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\hat{P}_{X^n} \in \Pi') = -D(\Pi'||P_X) = -D.$$

Así, obtenemos (3.42) y (3.44). Por lo que basta mostrar que Π' posee la propiedad de Sanov y que las variables $X_1, ..., X_n$ son asintóticamente cuasiindependientes bajo la condición $\hat{P}_{X^n} \in \Pi'$ con distribución límite P^*. Observemos lo siguiente: suponiendo que Π es casi completamente convexo y $\Pi'' = int_{\tau_0}(\Pi'')$ se tiene que

$$D(\Pi'||P_X) \leq D(int_{\tau_0}(\Pi')||P_X) \leq D(int_{\tau_0}(\Pi'')||P_X) = D(\Pi''||P_X)$$

$$= D(\Pi||P_X) = D(\Pi'||P_X),$$

es decir, $D(\Pi||P_X) = D(int_{\tau_0}(\Pi')||P_X)$. Aplicando el Teorema 3.2.1 a Π y Π' obtenemos el resultado buscado. Entonces basta mostrar que Π es casi completamente convexo y que $\Pi'' = int_{\tau_0}(\Pi'')$, i.e., que Π'' es abierto en la topología τ_0.

Primero mostramos que Π es casi completamente convexo: Notemos que como $\Pi_0(cd(C)) = \Pi(\mathcal{F}|\mathcal{K})$ el Lema 3.3.1 implica que $\Pi_0(cd(C))$ es casi

completamente convexo. Esto es, tenemos subconjuntos $\widetilde{\Pi}_1 \subset \widetilde{\Pi}_2 \subset ... \subset \Pi_0(cd(C))$ que son completamente convexos y $\Pi_0(cd(C)) \cap \Lambda_f \subset \cup_{i=1}^{\infty}\widetilde{\Pi}_i$; si logramos mostrar que existen subconjuntos $\Pi_1 \subset \Pi_2 \subset ... \subset \Pi$ completamente convexos tales que $\Pi \cap \Lambda_f \subset \cup_{i=1}^{\infty}\Pi_i$ entonces habremos terminado. Nuestros candidatos naturales son los conjuntos Π_i de medidas de probabilidad cuyo push-forward bajo Ψ pertenece a $\widetilde{\Pi}_i$, esto es,

$$\Pi_i = \{P \in \Lambda | \widetilde{P} \in \widetilde{\Pi_i}\}.$$

Observemos que $\Pi_i \subset \Pi_{i+1}$ pues si $P \in \Pi_i$ entonces $\widetilde{P}$ su push-forward bajo Ψ pertenece a $\widetilde{\Pi}_i \subset \widetilde{\Pi}_{i+1}$, luego $P \in \Pi_{i+1}$. Probemos que los conjuntos Π_i, i=1,2,... son completamente convexos: Sea $\nu : (\Omega, \mathcal{A}) \to (S, \mathcal{B})$ un kernel de Markov con $(\Omega, \mathcal{A}, \mu)$ un espacio de probabilidad tal que $\nu(\omega, \cdot) \in \Pi_i$ para cada $\omega \in \Omega$, entonces $\widetilde{\nu}(\omega, \cdot) \in \widetilde{\Pi}_i$ que como es completamente convexo se tiene que la medida de probabilidad definida por

$$\widetilde{\mu\nu}(A) = \int \widetilde{\nu}(\omega, A)\ d\mu(\omega), \quad A \in \mathcal{C},$$

pertenece a $\widetilde{\Pi}_i$, y observemos que la medida de probabilidad definida por

$$\mu\nu(B) = \int \nu(\omega, B)\ d\mu(\omega), \quad B \in \mathcal{B}$$

es, de hecho, la medida de probabilidad cuyo push-forward bajo Ψ es $\widetilde{\mu\nu}$ por como se define $\widetilde{\nu}$ y $\widetilde{\mu\nu}$, i.e.,

$$\widetilde{\mu\nu}(A) = \int \widetilde{\nu}(\omega, A)\ d\mu(\omega) = \int \nu(\omega, \Psi^{-1}(A))\ d\mu(\omega) = \mu\nu(\Psi^{-1}(A)).$$

Por lo que en efecto Π_i es completamente convexo para cada i. Ahora, si $P \in \Pi \cap \Lambda_f \subset \Pi$ entonces tenemos por definición de Π que $\widetilde{P}$ el push-forward bajo Ψ de P pertenece a $\widetilde{\Pi}$. Más aún, pertenece a $\widetilde{\Pi} \cap \Lambda_f(V)$. En efecto, pues el push-forward bajo Ψ de una medida de probabilidad atómica con un número finito de átomos sobre $(S, \mathcal{B})$ es una medida de probabilidad atómica con un número finito de átomos sobre $(V, \mathcal{C})$. Por lo tanto $\widetilde{P} \in \widetilde{\Pi} \cap \Lambda_f \subset \cup_{i=1}^{\infty}\widetilde{\Pi}_i$, es decir, $\widetilde{P} \in \widetilde{\Pi}_i$ para algún i y por lo tanto $P \in \Pi_i$, esto es, $P \in \cup_{i=1}^{\infty}\Pi_i$. Entonces $\Pi \cap \Lambda_f \subset \cup_{i=1}^{\infty}\Pi_i$, es decir, Π es casi completamente convexo.

Para mostrar que Π'' es abierto en la topología τ_0 primero observamos lo siguiente: Sea

$$\mho_n := \{R | R \text{ es medida de probabilidad sobre } (V, \mathcal{C}) \text{ y } R(K_n) = 1\},$$

notemos que $\mho_n \subset \mho_{n+1}$. Ahora, la función $E : \mho_n \to V$ definida mediante $R \to E(R)$ es continua en la topología débil-$*$ (topología vaga) para toda n (cf. Lema 4.3.2 del apéndice).

Sea $P_0 \in \Pi''$, queremos mostrar que existe una vecindad de P_0 en la topología τ_0 que esté contenida en Π'', i.e., $U_0(P_0, A, \epsilon) \subset \Pi''$ para algún $\epsilon > 0$ y alguna partición, finita y medible $A = (A_1, ..., A_m)$. Recordemos que

$$U_0(P_0, A, \epsilon) = \{R \big| |R(A_j) - P_0(A_j)| < \epsilon \text{ y } R(A_j) = 0 \text{ si } P_0(A_j) = 0\}$$

Por la definición de Π'' tenemos que $\widetilde{P}_0$, el push-forward de P_0 bajo Ψ, cumple que $\widetilde{P}_0 \in \Pi_0(int(C))$, así $E(\widetilde{P}_0) \in int(C)$ y $\widetilde{P}_0(K_n) = 1$ para algún n. Como $E(\widetilde{P}_0) \in int(C)$ existe U una vecindad del 0 en V tal que $E(\widetilde{P}_0) + U \subset int(C)$. Entonces si logramos encontrar una $\epsilon > 0$ y una partición, finita y medible A tal que para toda $P \in U_0(P_0, A, \epsilon)$ se tiene que $E(\widetilde{P}) \in E(\widetilde{P}_0) + U$ y $\widetilde{P}(K_n) = 1$ entonces habremos terminado. Como la topología vaga es más débil que la topología τ (cf. [26]) entonces E tambien es continua en $\mho_n$ con la topolgía τ, i.e., existe $\epsilon' > 0$ y $A' = (A'_1, ..., A'_k)$ una partición medible de K_n tal que para toda $R \in U(\widetilde{P}_0, A', \epsilon')$ se tiene que $E(R) \in E(\widetilde{P}_0) + U$, por lo tanto basta mostrar que existen una $\epsilon > 0$ y A una partición medible de S tal que para toda $P \in U_0(P_0, A, \epsilon)$ se tiene que $\widetilde{P} \in U(\widetilde{P}_0, A', \epsilon')$. Consideremos $\epsilon = \epsilon'$ y $A = (A_1 = \Psi^{-1}(A'_1), ..., A_n = \Psi^{-1}(A'_n), A_{n+1} = \Psi^{-1}(K_n^c))$, notemos que A es partición medible de S. Afirmamos que $U_0(P_0, A, \epsilon)$ es la vecindad buscada. En efecto, sea $P \in U_0(P_0, A, \epsilon)$ y consideremos $\widetilde{P}$ su push-forward bajo Ψ y observemos que $\widetilde{P} \in U(\widetilde{P}_0, A', \epsilon')$ ya que, en primer lugar, como $\widetilde{P}_0(K_n) = 1$ entonces $\widetilde{P}_0(K_n^c) = 0$ y así $\widetilde{P}(K_n^c) = 0$ y por lo tanto $\widetilde{P} \in \mho_n$. En segundo lugar, se tiene que

$$|\widetilde{P}(A'_i) - \widetilde{P}_0(A'_i)| = |P(A_i) - P_0(A_i)| < \epsilon = \epsilon' \quad i = 1, ..., n.$$

Por lo que efectivamente $\widetilde{P} \in U(\widetilde{P}_0, A', \epsilon')$. Con lo cual damos por concluida la demostración.

□

Notemos que el Teorema 3.3.1 no es más que una consecuencia de los teoremas 2.3.2 y 3.2.1. En efecto, la demostración del resultado se logró al aplicar los dos teoremas anteriores a los conjuntos Π, Π' y Π'' generados por la función medible Ψ. Observemos que una de las claves para probar este teorema, así como se hizo en el Teorema 2.3.2, es la igualdad $\Pi_0(C) = \Pi(\mathcal{F}|\mathcal{K})$ y $\Pi_0(int(C)) = \Pi'(\mathcal{F}|\mathcal{K})$ lo cual fue posible gracias al teorema bipolar del ánalisis funcional. El Teorema 3.3.1 es un teorema que establece una gran desviación para la sucesión de variables aleatorias $\{\Psi(X_n)\}_{n\in\mathbb{N}}$ independientes con distribución común Q, con Ψ un estadístico y C un conjunto convexo que interseca al soporte de Q. Más aún, se establece la cuasiindependencia asintótica de las variables aleatorias $\{X_n\}_{n\in\mathbb{N}}$ bajo la condición de que el promedio bajo Ψ se encuentre en C con distibución límite dada por la I-proyección generalizada de su distribución común P_X en donde aquella queda caracterizada por su derivada de Radon-Nikodỳm respecto a P_X y esta derivada pertenece a una familia de funciones exponenciales.

Capítulo 4

Apéndice

Este apartado está dividido en 4 secciones: en la primera probamos los hechos que se fueron utilizando a lo largo de los capítulos anteriores y que por cuestiones prácticas y estéticas no se mencionó prueba alguna en aquellos momentos. En la segunda, tercera y cuarta sección encontramos todo lo referente a topología, análisis real y análisis funcional, y teoría de la medida respectivamente.

4.1. Algunos resultados útiles

4.1.1. Modos de convergencia

Empezaremos enunciando los distintos modos de convergencia que existen en teoría de la medida y teoría de probabilidad. No pretendemos dar todos los tipos de convergencia sólo enunciaremos los que son utilizados en este trabajo. Para un desarrollo completo y sus respectivas implicaciones pueden consultarse [18] y [16].

Definición 4.1.1. *Sean $(S, \mathcal{B}, \mu)$ un espacio de medida y $\{f_n\}_{n\in\mathbb{N}}$ una sucesión de funciones tales que $f_n : S \to \mathbb{R}$ y cada f_n es medible. Se dice que:*

i) $\{f_n\}_{n\in\mathbb{N}}$ converge casi donde sea a una función real y medible f si existe N un conjunto medible tal que $\mu(N) = 0$ y

$$\lim_{n\to\infty} f_n(s) = f(s) \quad si \quad s \in S \setminus N.$$

ii) $\{f_n\}_{n\in\mathbb{N}}$ converge en medida a una función real y medible f si

$$\lim_{n\to\infty} \mu(\{s \in S \big| |f_n(s) - f(s)| \geq \epsilon\}) = 0 \quad \forall \epsilon > 0.$$

A este tipo de convergencia se le suele denotar de la siguiente manera: $f_n \xrightarrow{\mu} f$.

iii) $\{f_n\}_{n\in\mathbb{N}}$ *converge casi uniformemente a una función real y medible* f *si para todo* $\delta > 0$ *existe* $B \in \mathcal{B}$ *tal que* $\mu(B) < \delta$ *y* $\{f_n\}_{n\in\mathbb{N}}$ *converge uniformemente a* f *en* $S \setminus B$. *A este tipo de convergencia se le suele denotar de la siguiente manera:* $f_n \xrightarrow{c.u.} f$.

Por último si $1 \leq p < \infty$ *y* $f_n \in L_p(\mu)$[1] *para toda* n *entonces se dice que: iv)* $\{f_n\}_{n\in\mathbb{N}}$ *converge en media* p *a una función real y medible* f *si*

$$\lim_{n\to\infty} ||f_n - f||_p = 0.$$

A este tipo de convergencia se le suele denotar de la siguiente manera: $f_n \xrightarrow{L_p} f$.

Observación. *A veces se realizan afirmaciones considerando a la medida* μ, *es decir, que algo sucede salvo en un conjunto de medida cero relativo a* μ. *En este caso se dice que algo sucede relativo a* μ *y se denota mediante* $[\mu]$. *Por ejemplo, la convergencia casi donde sea es la convergencia puntual salvo en un conjunto de medida cero relativo a* μ. *Así se suele denotar la convergencia en i) de la siguiente manera:* $f_n \to f \quad [\mu]$.

Observación. *Cuando* μ *es una medida de probabilidad a la convergencia en i) se le suele llamar convergencia casi segura y a veces se abrevia c.s.; a la convergencia en ii) se le suele llamar convergencia en probabilidad y a la convergencia en iv) se le suele llamar de la misma manera, i.e., convergencia en media.*

Ahora, si consideramos una sucesión de variables aleatorias reales $\{X_n\}_{n\in\mathbb{N}}$ y consideramos sus funciones de distribución $\{F_n\}_{n\in\mathbb{N}}$. Se dice que la sucesión de variables aleatorias converge en distribución a X si

$$\lim_{n\to\infty} F_n(x) = F(x).$$

para toda $x \in \mathbb{R}$ en la cual F es continua. Todos los demás tipos de convergencia implican este último.

Proseguimos con algunos resultados utilizados a lo largo de este trabajo.

Lema 4.1.1. *Sea* X *una variable aleatoria que toma valores en* $\mathbb{R}$ *y con función generadora de momentos* $\varphi(t) = \mathbb{E}(e^{tX})$. *Sea* $A = \{t|\varphi(t) < \infty\}$ *y supongamos que* $int(A)$ *es no vacío, entonces* φ *es infinitamente diferenciable en* $int(A)$, $X^k e^{tX}$ *es integrable para cada* $t \in int(A)$ *y*

$$\varphi^{(k)}(t) = \mathbb{E}(X^k e^{tX}) \quad \forall t \in int(A).$$

[1] cf. el desarrollo anterior al Teorema 4.3.2.

Prueba. Como $int(A) \neq \emptyset$ entonces podemos tomar $t_0 \in int(A)$ y sea $\epsilon > 0$ tal que $t_0 + 2\epsilon, t_0 - 2\epsilon \in A$. Sean $\{t_n\}_{n\in\mathbb{N}}$ tales que t_n converge a t_0 cuando n tiende a infinito y $|t_n - t_0| \leq \epsilon$ para toda n. Observemos que para cada n se tiene que

$$\frac{\varphi(t_n) - \varphi(t_0)}{t_n - t_0} = \frac{\mathbb{E}(e^{t_n X}) - \mathbb{E}(e^{t_0 X})}{t_n - t_0} = \mathbb{E}(Y_n),$$

en donde

$$Y_n = \frac{e^{t_n X} - e^{t_0 X}}{t_0 - t_n}.$$

Observemos que si n tiende a infinito

$$Y_n \to \frac{d}{dt} e^{tX}\bigg|_{t=t_0} = X e^{t_0 X}.$$

Definimos $Y = X e^{t_0 X}$. Ahora, por el teorema del valor medio se tiene que

$$e^{t_n X} - e^{t_0 X} = (t_n - t_0) X e^{X t^*}$$

para algún t^* entre t_n y t_0, claramente este punto depende de n y del ω en que evaluamos, i.e., $X(\omega)$. Luego,

$$|Y_n| = \left| \frac{e^{t_n X} - e^{t_0 X}}{t_0 - t_n} \right| = |X| e^{X t^*} \leq |X| \Big(e^{X(t_0 - \epsilon)} + e^{X(t_0 + \epsilon)} \Big).$$

Notemos que

$$|X| = \frac{1}{\epsilon} \epsilon |X| \leq \frac{1}{\epsilon} \left(1 + \epsilon |X| + \frac{\epsilon^2 |X|^2}{2!} + \frac{\epsilon^3 |X|^3}{3!} + ...) \right)$$

es la serie de Taylor de la función exponencial. Así,

$$|X| \leq \frac{1}{\epsilon} e^{\epsilon |X|} \leq \frac{1}{\epsilon} \left(e^{-\epsilon X} + e^{\epsilon X} \right),$$

por lo tanto se tiene que

$$|Y_n| \leq \frac{1}{\epsilon} \left(e^{-\epsilon X} + e^{\epsilon X} \right) \left(e^{(t_0 - \epsilon) X} + e^{(t_0 + \epsilon) X} \right)$$

$$= \frac{1}{\epsilon} \left(e^{(t_0 - 2\epsilon) X} + 2 e^{t_0 X} + e^{(t_0 + 2\epsilon) X} \right),$$

está última variable aleatoria la definimos como Z. Y observamos que

$$\mathbb{E}(Z) = \frac{1}{\epsilon}\varphi(t_0 - 2\epsilon) + 2\varphi(t_0) + \varphi(t_0 + 2\epsilon) < \infty.$$

Por lo tanto construimos una variables aleatoria Z tal que $\mathbb{E}(Z) < \infty$ y $|Y_n| \leq Z$ para toda n. Luego, por el teorema de convergencia dominada de Lebesgue, se tiene que Y_n y Y son integrables y además $\mathbb{E}(Y_n) \to \mathbb{E}(Y)$. Esto es Xe^{t_0X} es integrable y $\varphi'(t_0) = \mathbb{E}(Xe^{t_0X})$ para toda $t_0 \in int(A)$. Considerando a la variables aleatoria $X' = Xe^{t_0X}$ con el mismo argumento se obtiene que $X^2e^{t_0X}$ es integrable y $\varphi''(t_0) = \mathbb{E}(X^2e^{t_0X})$ para toda $t_0 \in int(A)$. Iterando el argumento se obtiene el resultado deseado.

□

Teorema 4.1.1. *(Desigualdad de Markov-Chebyshev). Sea X una variable aleatoria real y no negativa y $a > 0$ entonces se tiene que $\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$.*

Demostración.

$$\mathbb{E}(X) = \int_{\mathbb{R}} x \ dF(x) = \int_0^a x \ dF(x) + \int_a^\infty x \ dF(x) \geq \int_a^\infty x \ dF(x)$$

$$\geq \int_a^\infty a \ dF(x) = a\int_a^\infty dF(x) = a\mathbb{P}(X \geq a).$$

□

Lema 4.1.2. *Sea $\mathcal{M}_1(\Gamma) = \{\nu | \nu$ es medida de probabilidad sobre $(\Gamma, \mathcal{B})\}$ con $|\Gamma| \leq \aleph_0$. Entonces*

1) Si $|\Gamma| = n$ se tiene que

$$\mathcal{M}_1(\Gamma) = \left\{ x = (x_1, ..., x_n) \in \mathbb{R}^n \Big| \sum_{s=1}^n x_n = 1 \quad y \quad x_s \geq 0 \quad \forall s = 1, ..., n \right\} \tag{4.1}$$

2) Si $|\Gamma| = \aleph_0$ se tiene que

$$\mathcal{M}_1(\Gamma) = \left\{ x = (x_1, ..., x_n, ...) \in \mathbb{R}^\infty \Big| \sum_{s=1}^\infty x_n = 1 \quad y \quad x_s \geq 0 \quad \forall s \right\} \tag{4.2}$$

Prueba. Primero observamos que como $|\Gamma| \leq \aleph_0$ entonces $\Gamma = \{\gamma_1, ..., \gamma_n\}$ para algún $n \in \mathbb{N}$ o $\Gamma = \{\gamma_1, \gamma_2, ...\}$ y así para cualquier $B \subset \Gamma$ tal que $B \in \mathcal{B}$ se ve de la siguiente forma: $B = \cup_{s \in A}\{\gamma_s\}$ con $A = \{1, ..., n\}$ o bien $A = \{1, 2, ...\}$, observamos que dicha unión es ajena pues cada elemento es un singulete distinto. Así si $\mu \in \mathcal{M}_1(\Gamma)$ entonces

$$\mu(B) = \mu\left(\bigcup_{s \in A} \gamma_s\right) = \sum_{s \in A} \mu(\gamma_s)$$

por lo que la medida de probabilidad μ queda únicamente determinada por los valores $\mu(\gamma_s)$ por lo que para cada $\nu \in \mathcal{M}_1(\Gamma)$ definimos $\nu(\gamma_s) = \nu_s \in \mathbb{R}^+ \cup \{0\}$. Ahora, si $|\Gamma| = n$ sea

$$\Psi : \mathcal{M}_1(\Gamma) \to \left\{ x = (x_1, ..., x_n) \in \mathbb{R}^n \Big| \sum_{s=1}^{n} x_s = 1 \ \ y \ \ x_s \geq 0 \quad \forall s = 1, ..., n \right\}$$

definida por

$$\Psi(\nu) = (\nu_1, ..., \nu_n).$$

Si $|\Gamma| = \aleph_0$ sea

$$\Psi : \mathcal{M}_1(\Gamma) \to \left\{ x = (x_1, ..., x_n, ...) \in \mathbb{R}^\infty \Big| \sum_{s=1}^{\infty} x_s = 1 \ \ y \ \ x_s \geq 0 \quad \forall s \right\}$$

definida por

$$\Psi(\nu) = (\nu_s)_{s=1}^{\infty}.$$

Ψ es biyectiva: Si tomamos dos medidas de probabiliad $\nu, \mu \in \mathcal{M}_1(\Gamma)$ tales que $\nu \neq \mu$, i.e,. $\nu(\gamma_s) \neq \mu(\gamma_s)$ para algúna s entonces $\Psi(\nu) \neq \Psi(\mu)$ pues $\nu_s \neq \mu_s$, en ambos casos. Si tomamos

$$x \in \left\{ x = (x_1, ..., x_n) \in \mathbb{R}^n \Big| \sum_{s=1}^{n} x_n = 1 \ \ y \ \ x_s \geq 0 \quad \forall s = 1, ..., n \right\}$$

o

$$x \in \left\{ x = (x_1, ..., x_n, ...) \in \mathbb{R}^\infty \Big| \sum_{s=1}^{\infty} x_s = 1 \ \ y \ \ x_s \geq 0 \ \ \forall s \right\}$$

entonces $\nu \in \mathcal{M}_1(\Gamma)$ definida como $\nu(\gamma_s) = x_s$ es tal que $\Psi(\nu) = x$, en ambos casos.

□

Teorema 4.1.2. *(Glivenko-Cantelli). Sean $\{X_n\}_{n\in\mathbb{N}}$ variables aleatorias independientes e idénticamente distribuidas con función de distribución F. Recordamos la notación $L_n(\cdot) = L(X^n, \cdot)$. Entonces*

$$\lim_{n\to\infty} d(L_n, F) = 0 \quad \textit{casi seguramente.}$$

Demostración. [33], pág. 266. □

4.1.2. Órdenes de crecimiento

Definición 4.1.2. *Se dice que una función $g : \mathbb{R} \to \mathbb{R}$ es O(f(x)) pronunciado "o grande de f" si existen $M \in \mathbb{R}$ y $n_0 \in \mathbb{N}$ tales que para toda $x > n_0$ se tiene que $|g(x)| < M|f(x)|$.*

Lema 4.1.3. *Sea $k \in K_n$ entonces $\sum_{s=1}^{r} \left(\frac{-O(\log k_s)}{n}\right) + \frac{O(\log n)}{n} = O\left(\frac{\log n}{n}\right)$.*

Prueba. Primero observamos que por definción tenemos lo siguiente:

$|O(\log k_s)| \leq M \log k_s$ con $M_s \in \mathbb{R}$ para toda $s \in \{1, ..., r\}$, también $|O(\log n)| \leq M_1 \log n$ para algún $M_1 \in \mathbb{R}$ entonces $|\frac{O(\log k_s)}{n}| \leq \frac{M_s}{n} \log k_s$ y $|\frac{O(\log n)}{n}| \leq \frac{M_1}{n} \log n$. Además, como $\sum_{s=1}^{r} k_s = n$ entonces $k_s \leq n$ para toda $s = 1, ..., r$. Así, $\log k_s \leq \log n$ para toda $s = 1, ..., r$. Por lo tanto

$$\left|\sum_{s=1}^{r}\left(\frac{-O(\log k_s)}{n}\right) + \frac{O(\log n)}{n}\right| \leq \sum_{s=1}^{r}\left(\frac{|O(\log k_s)|}{n}\right) + \frac{|O(\log n)|}{n}$$

$$\leq \sum_{s=1}^{r}\left(\frac{M_s}{n} \log k_s\right) + \frac{M_1}{n} \log n \leq \frac{1}{n}\left(\sum_{s=1}^{r}(M_s \log n) + M_1 \log n\right)$$

$$= \left[\left(\sum_{s=1}^{r} M_s\right) + M_1\right]\frac{\log n}{n} = M\frac{\log n}{n} \ \ con \ \ M = \left(\sum_{s=1}^{r} M_s\right) + M_1.$$

□

Lema 4.1.4. *$|K_n| = \binom{n+r-1}{r-1}$ y $\binom{n+r-1}{r-1} = O(n^{r-1})$.*

Prueba. Para la primera afirmación notemos que K_n consiste de vectores que cumplen

$$\sum_{i=1}^{r} x_i = n. \tag{4.3}$$

Primero mostraremos que hay $\binom{n-1}{r-1}$ vectores con r entradas enteras positivas que cumplen con (4.3); para ello consideremos n objetos (digamos unos) acomodados consecutivamente en un renglón y nos fijamos en los $n-1$ espacios que hay entre ellos entonces observamos que cualquier elección de $r-1$ espacios entre los unos de los $n-1$ que hay en total (definiendo x_i como el número de unos que hay entre el espacio $i-1$ y el espacio i, $i \in \{2, ...r-1\}$; x_1 es el número de unos que hay antes del primer espacio elegido y x_r es el número de unos que hay después del último espacio elegido es decir del $r-1$ espacio) nos da una solución para (4.3). Por ejemplo si tenemos $r = 4$

y $n = 10$ entonces nuestro arreglo de unos y espacios entre los unos queda de la siguiente manera:

$$11{,}111{,}1{,}1111 \quad \Rightarrow \quad x_1 = 2, \quad x_2 = 3, \quad x_3 = 1 \quad y \quad x_4 = 4 \quad \sum_{i=1}^{4} x_i = 10.$$

Por otro lado, si tenemos una solución $(x_1, ..., x_r)$ a (4.3) entonces es claro que podemos encontrar una elección de $r-1$ espacios entre los unos que nos de como resultado dicho vector con x_i como se definió arriba. Con lo anterior mostramos que existe una correspondencia biunívoca entre las soluciones a (4.3) y todas las posibles elecciones diferentes de $r-1$ espacios entre los unos del total de $n-1$ espacios entre los unos. Es decir, el número de vectores con entradas enteras positivas que cumplen (4.3) es $\binom{n-1}{r-1}$.

Ahora, para encontrar el número de vectores con entradas enteras no negativas observamos que la cantidad de soluciones enteras no negativas a (4.3) es exactamente la misma cantidad de soluciones enteras positivas a la ecuación $\sum_{i=1}^{r} y_i = n+r$ mediante la transformación $y_i \to x_i + 1$, $i \in 1, ..., r$; por lo anterior tenemos que dicha cantidad es $\binom{n+r-1}{r-1}$.

Para mostrar la segunda afirmación del lema observemos que

$$\binom{n+r-1}{r-1} = \frac{(n+r-1)!}{n!(r-1)!} = \frac{\prod_{i=1}^{r-1} n+r-i}{(r-1)!} \leq \prod_{i=1}^{r-1} n+r = (n+r)^{r-1}.$$

Es claro que existen $M_r \in \mathbb{R}$ y $n_0 \in \mathbb{N}$ tales que $|(n+r)^{r-1}| = (n+r)^{r-1} \leq Mn^{r-1}$ para toda $n \geq n_0$. □

Lema 4.1.5. $O(f(n)) + O(f(n)) = O(f(n))$.

Prueba. Por definición de O existen M_1, M_2, n_1 y n_2 tales que $|O(f(n))| \leq M_1 f(n)$ para toda $n \geq n_1$ y $|O(f(n))| \leq M_2 f(n)$ para toda $n \geq n_2$. Sea $n_0 = \text{máx}\{n_1, n_2\}$ y $M = M_1 + M_2$ entonces $|O(f(n)) + O(f(n))| \leq |O(f(n))| + |O(f(n))| \leq M_1 f(n) + M_2 f(n) = Mf(n)$ para toda $n > n_0$. □

Lema 4.1.6. *Sean $r > 1$ y $M \in \mathbb{R}$ fijo entonces $\frac{1}{n}\log M + \frac{r-1}{n}\log n = O(\frac{\log n}{n})$.*

Prueba. Como M es fijo y $\frac{1}{n}$ tiende a cero cuando n tiende infinito entonces se tiene que existe $L \in \mathbb{R}$ y $n_0 \in \mathbb{N}$ tal que para toda $n > n_0$ $|\frac{1}{n}\log M| < L\frac{\log n}{n}$. Por lo tanto $\frac{1}{n}\log M = O(\frac{\log n}{n})$.

Por otro lado $\frac{r-1}{n}\log n = (r-1)\frac{\log n}{n} = O(\frac{\log n}{n})$. Así, se tiene que

$$\left|\frac{1}{n}\log M + \frac{r-1}{n}\log n\right| \leq \left|\frac{1}{n}\log M\right| + \left|\frac{r-1}{n}\log n\right| = O\left(\frac{\log n}{n}\right) + O\left(\frac{\log n}{n}\right)$$

$$= O\left(\frac{\log n}{n}\right).$$

La última igualdad se sigue del Lema 4.1.5.

□

Dada por concluida esta subsección continuamos con otros resultados utilizados para probar algunas propiedades de la entropía relativa en la sección 1.5 del primer capítulo.

Lema 4.1.7. *El supremo de funciones semicontinuas inferiormente es una función semicomtinua inferiormente. Es decir, si consideramos el siguiente conjunto:*

$$\{f_t | f_t : \mathbb{R} \to \mathbb{R} \ \textit{es una función semicontinua inferiormente}, \ t \in \mathbb{R}\}$$

entonces

$$\sup_{t\in\mathbb{R}}\{f_t\}$$

es una función semicontinua inferiormente.

Prueba. Recordemos que una función f es semicontinua inferiormente si dado $x \in \mathbb{R}$ y $r \in \mathbb{R}$ tal que $r < f(x)$ entonces existe una vecindad U de x tal que $r < f(y)$ para toda $y \in U$. Sea $f(x) := \sup_{t\in\mathbb{R}}\{f_t(x)\}$, $x \in \mathbb{R}$. Sea $x \in \mathbb{R}$ tenemos que para cualquier r tal que $r < f(x)$ como f es el supremo entonces debe existir algún $t_0 \in \mathbb{R}$ tal que $r < f_{t_0}(x)$ y como cada f_t es semicontinua inferiormente entonces existe una vecindad U de x tal que $r < f_{t_0}(y)$ $\forall y \in U$, así tenemos que $r < f_{t_0}(y) \leq f(y)$ para toda $y \in U$ ya que f es el supremo. Y por lo tanto $r < f(y)$ para toda $y \in U$ con U vecindad de x, es decir, f es semicontinua inferiormente. □

Lema 4.1.8. *El supremo de funciones convexas es una función convexa. Es decir, si consideramos el conjunto*

$$\{g_t | g_t : \mathbb{R} \to \mathbb{R} \ \textit{es una función convexa}, \ t \in \mathbb{R}\}$$

entonces

$$\sup_{t\in\mathbb{R}}\{g_t\}$$

es una función convexa.

Prueba. Sea $g = \sup_{t\in\mathbb{R}}\{g_t\}$ entonces tenemos que si $\lambda \in [0, 1]$

$$g_t((1-\lambda)z + \lambda w) \leq (1-\lambda)g_t(z) + \lambda g_t(w) \leq (1-\lambda)g(z) + \lambda g(w) \quad \forall t.$$

Es decir,

$$g_t((1-\lambda)z + \lambda w) \leq (1-\lambda)g(z) + \lambda g(w) \quad \forall t$$

como es para toda $t \in \mathbb{R}$ entonces podemos tomar el supremo de ambos lados de la desigualdad y como g ya no depende de t obtenemos

$$g((1-\lambda)z + \lambda w) \leq (1-\lambda)g(z) + \lambda g(w)$$

y por lo tanto g es convexa. □

Lema 4.1.9. *(Desigualdad suma-logaritmo). Para números reales no negativos $a_1, ..., a_k, b_1, ..., b_k$ se tiene la siguiente desigualdad*

$$\sum_{j=1}^{n} a_j \log\left(\frac{\sum_{j=1}^{n} a_j}{\sum_{j=1}^{n} b_j}\right) \leq \sum_{j=1}^{n} a_j \log\left(\frac{a_j}{b_j}\right).$$

Prueba. Sean $a = \sum_{j=1}^{k} a_j$ y $b = \sum_{j=1}^{k} b_j$ y consideremos la función $f(x) = x \log x$ que claramente es convexa, y observemos que $\frac{b_j}{b} \geq 0$ para toda j y $\sum_{j=1}^{k} \frac{b_j}{b} = 1$, entonces

$$\sum_{j=1}^{n} a_j \log\left(\frac{a_j}{b_j}\right) = \sum_{j=1}^{n} b_j f\left(\frac{a_j}{b_j}\right) = b \sum_{j=1}^{n} \frac{b_j}{b} f\left(\frac{a_j}{b_j}\right) \geq bf\left(\sum_{j=1}^{k} \frac{b_j}{b} \frac{a_j}{b_j}\right)$$

$$= \sum_{j=1}^{k} a_j \log\left(\frac{\sum_{j=1}^{k} a_j}{\sum_{j=1}^{k} b_j}\right).$$

□

Lema 4.1.10. *Sean $\lambda_1, ..., \lambda_n$ tales que $0 \leq \lambda_i$ para toda $i = 1, ..., n$ y $\sum_{i=1}^{n} \lambda_i = 1$. Sean $P_1, ..., P_n$ medidas de probabilidad sobre un espacio medible $(S, \mathcal{B})$. Entonces*

$$D\left(\sum_{i=1}^{n} \lambda_i P_i \middle|\middle| Q\right) = \sum_{i=1}^{n} \lambda_i D(P_i||Q) - \sum_{i=1}^{n} \lambda_i D\left(P_i \middle|\middle| \sum_{i=1}^{n} \lambda_i P_i\right).$$

Prueba. [32], Lema 7. □

Lema 4.1.11. $\mathcal{F} = \left\{g | g(x) = \sum_{i=1}^{k} \zeta_i f_i(x),\ f_i(x) = x_i \ y\ \zeta_i \geq 0\right\}$ *es un cono convexo.*

Prueba. Sean $\{\lambda\}_{j=1}^{n}$ reales no negativos y $\{g_j\}_{j=1}^{n} \subset \mathcal{F}$

$$\sum_{j=1}^{n} \lambda_j g_j(x) = \sum_{j=1}^{n} \lambda_j \sum_{i=1}^{k} \zeta_i^j f_i(x) = \sum_{j=1}^{n} \lambda_j \sum_{i=1}^{k} \zeta_i^j x_i = \sum_{i=1}^{k} \sum_{j=1}^{n} \lambda_j \zeta_i^j x_i = \sum_{i=1}^{k} \mu_i x_i$$

$$= \sum_{i=1}^{k} \mu_i f_i(x),$$

con $\mu_i = \sum_{j=1}^{n} \lambda_j \zeta_i^j \geq 0$.

□

Lema 4.1.12. *Sean $\{a_n\}_{n\in\mathbb{N}}$ y $\{b_n\}_{n\in\mathbb{N}}$ sucesiones de números reales tales que*

$$\lim_{n\to\infty} a_n \cdot b_n = x \quad y \quad \lim_{n\to\infty} b_n = y \neq 0$$

entonces se tiene que $\{a_n\}_{n\in\mathbb{N}}$ converge.

Prueba. Como $y \neq 0$ esto quiere decir que existe N tal que para toda $n > N$ $b_n \neq 0$, consideremos $n > N$

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} \frac{a_n b_n}{b_n} = \frac{\lim_{n\to\infty} a_n b_n}{\lim_{n\to\infty} b_n} = \frac{x}{y}.$$

□

Lema 4.1.13. *Sean A y B dos eventos de un espacio de probabilidad $(S, \mathcal{B}, P)$. Si $P(B) = 1$ entonces $P(A \cap B) = P(A)$.*

Prueba. Observemos que podemos suponer que $P(A) > 0$ pues en caso contrario se tiene que $P(A \cap B) \leq P(A) = 0$ y se sigue el resultado.

Caso 1: $A \cap B = \emptyset$ entonces $P(A \cap B) = 0$, como $P(A) > 0$ entonces $P(A \cup B) = P(A) + P(B) = P(A) + 1 > 1$ lo cual contradice que P sea medida de probabilidad, por lo que tenemos que este caso no puede suceder.

Caso 2: Si $A \cap B \neq \emptyset$ entonces $P(A \cap B) > 0$ si suponemos lo contrario, i.e., que $P(A \cap B) \neq P(A)$ entonces $P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + 1 - P(A \cap B)$.

Subcaso 2.1: Si $P(A \cap B) < P(A)$ entonces $P(A \cup B) > 1$ lo cual es una contradicción.

Subcaso 2.2: Si $P(A) < P(A \cap B)$ entonces $P(B) \leq P(A \cap B) < 1$ lo cual, de nuevo, es una contradicción. □

Definición 4.1.3. *Dado un conjunto S y una partición $A = \{A_j | j \in J\}$ de S, se dice que $B = \{B_i | i \in I\}$ refina a A o simplemente B es un refinamiento de A si:*

(1) B es una partición de S.

(2) $B_i \subset A_j$ para algún $j \in J$ y para cada $i \in I$.

Observación. *Si B es un refinamiento de A y C es un refinamiento de B entonces C es un refinamiento de A.*

Lema 4.1.14. *Sea S un conjunto y $\{A^m\}_{m=1}^n$ una colección finita de particiones finitas. Entonces siempre existe A una partición de S que refina a cada A^m.*

Prueba. La prueba es por inducción sobre la cantidad de particiones que se tienen.

(i) Base inductiva: Sea n=2, por lo que $\{A^1, A^2\}$ es nuestra colección de particiones de S, en donde $A^1 = (A_1^1, ..., A_k^1)$ y $A^2 = (A_1^2, ..., A_l^2)$ con $k, l \in \mathbb{N}$. Consideremos a la partición

$$A = \{A_i^1 \cap A_j^2 | A_i^1 \cap A_j^2 \neq \emptyset, \;\; i = 1, ..., k \;\; j = 1, ..., l\}.$$

Notemos que $A_i^1 \cap A_j^2 = \emptyset$ para toda i y para toda j no puede suceder pues lo anterior implicaría que $A_i^1 = \emptyset$ para algún i o $A_j^2 = \emptyset$ para algún j lo cual, en ambos casos, es una contradicción. Además, es fácil ver que es una partición pues en primer lugar se tiene que

$$\bigcup_{i,j} [A_i^1 \cap A_j^2] = S,$$

la contención $\cup_{i,j}[A_i^1 \cap A_j^2] \subset S$ es clara, la contención $S \subset \cup_{i,j}[A_i^1 \cap A_j^2]$ se tiene pues si $s \in S$ entonces $s \in A_i^1$ y $s \in A_j^2$ para alguna i y alguna j. En segundo lugar, es claro que $[A_i^1 \cap A_j^2] \cap [A_r^1] \cap A_p^2] = \emptyset$ con $i \neq r$ o $j \neq p$.

Afirmamos que A es la partición buscada. En efecto, ya que $A_i^1 \cap A_j^2 \subset A_i^1$ y $A_i^1 \cap A_j^2 \subset A_j^2$ para toda i y toda j.

(ii) Hipótesis de inducción: Supongamos que es válido para una cantidad q de particiones.

(iii) Paso inductivo: Verifiquemos que entonces es válido para q+1 particiones. Sea $\{A^m\}_{m=1}^{q+1}$ una colección de particiones finitas de S. Por hipótesis de inducción tenemos que para $\{A^m\}_{m=1}^q$ existe una partición A' que refina a cada A^m, $m = 1, ..., q$; consideremos las dos particiones A' y A^{q+1}, por la base inductiva tenemos que existe A un refinamiento de A' y de A^{q+1}, luego A al ser refinamiento de un refinamiento de cada A^m, $m = 1, ..., q$ es refinamiento de A^m para cada $m = 1, ..., q$ y es un refinamiento de A^{q+1} por lo que efectivamente existe una partición A que refina a cada A^m, $m = 1, ..., q + 1$. □

Lema 4.1.15. *Sea $(S, \mathcal{B})$ un espacio medible. Sea $B \in \mathcal{B}$ y consideremos a la medida de probabilidad definida mediante $P_B(E) = \frac{P(E \cap B)}{P(B)}$ con $E \in \mathcal{B}$. Ahora, si $B \notin \mathcal{B}$ consideremos a la medida exterior*

$$\bar{P}(B) = \text{ínf}\{P(C) | B \subset C \;\; y \;\; C \in \mathcal{B}\}.$$

Entonces $\mathcal{O} = \{F = E \cap B | E \in \mathcal{B}\}$*, la familia de subconjuntos de B, es una* σ*-álgebra de subconjunto de B y* $(B, \mathcal{O}, P')$ *con* $P'(F) = \frac{\bar{P}(E \cap B)}{\bar{P}(B)}$ *y* $F = E \cap B$ *es un espacio de probabilidad.*

Prueba. Veamos primero que $\mathcal{O}$ es una σ-álgebra de subconjunto de B:

(i) Usualmente se demuestra que si $F \in \mathcal{O}$ entonces $F^c \in \mathcal{O}$, en este caso utilizaremos la definición de [18], i.e., demostaremos que si $F_1, F_2 \in \mathcal{O}$ entonces $F_1 \setminus F_2 \in \mathcal{O}$: $F_1 \setminus F_2 = E_1 \cap B \setminus E_2 \cap B = E_1 \cap B \cap [E_2 \cap B]^c = E_1 \cap B \cap [E_2^c \cup B^c] = [E_1 \cap B \cap E_2^c] \cup [E_1 \cap B \cap B^c] = E_1 \cap B \cap E_2^c \cup \emptyset = E_1 \setminus E_2 \cap B$. Como $E_1 \setminus E_2 \in \mathcal{B}$ entonces $F_1 \setminus F_2 \in \mathcal{O}$.

(ii) Dada $\{E_i \cap B\}_{i=1}^{\infty}$ entonces $\cup_{i=1}^{\infty} E_i \cap B = [\cup_{i=1}^{\infty} E_i] \cap B$ y de nuevo como $\cup_{i=1}^{\infty} E_i \in \mathcal{B}$ entonces se sigue que $\cup_{i=1}^{\infty} E_i \cap B \in \mathcal{B}$.

(iii) Claramente $B \in \mathcal{O}$ pues $B = B \cap S$.

Dicho esto consideremos la medida probabilidad P' sobre $(B, \mathcal{O})$ definida como $P'(F) = \frac{\bar{P}(E \cap B)}{\bar{P}(B)} = \bar{P}_B(E)$ con $F = E \cap B$. $P'(F)$ es medida de probabilidad. En efecto, $0 \leq P'(F) \leq 1$ por como está definida, claramente $P'(A') = 1$ y $P'(\emptyset) = 0$ por lo mismo. Por último, la σ-aditividad se obtiene gracias a lo siguiente:

Sea $\{F_i\}_{i=1}^{\infty}$ una sucesión de elementos de $\mathcal{O}$ ajenos dos a dos. Observemos que esta sucesión nos genera una sucesión de conjutos $\{E_i\}_{i=1}^{\infty}$. Ahora, como $B \in \mathcal{O}$ y cada E_i es $\bar{P}$-medible entonces por el Teorema B de [21] (pág. 46) se tiene que

$$\bar{P}\left(\bigcup_{i=1}^{\infty} E_i \cap B\right) = \sum_{i=1}^{\infty} \bar{P}(E_i \cap B),$$

lo que nos da que P' es σ-aditiva.

Observemos que esta medida de probabilidad es única pues si existe P_1' que cumple que $P_1'(F) = \bar{P}_B(E)$ entonces $P_1' = P'$.

□

4.2. Topología

Una de los resultados de gran relevancia utilizados en la demostración del Teorema 3.1.1 es el hecho de que el producto arbitrario de conjuntos compactos es de nuevo un conjunto compacto en la topología producto, conocido como el teorema de Tychonoff. Además, se hizo un uso extensivo de las propiedades de la topología producto por lo cual presentamos una ligera introducción a estos conceptos y resultados los cuales son tomados de [27].

Sea $\{(X_i, \tau_i)\}_{i \in I}$ una familia de espacios topológicos indexados a un conjunto arbitario I. Consideremos su producto

$$\prod_{i\in I} X_i.$$

Sea $\pi_j : \prod_{i\in I} X_i \to X_j$ definida por $\pi_j((x_i)_{i\in I}) = x_j$, la proyección asociada al índice j.

Definición 4.2.1. *Sea* $\Phi_j = \{\pi_j^{-1}(U_j) | U_j$ *es abierto en* $X_j\}$ *y consideremos la unión de estos conjuntos, es decir,*

$$\Phi = \bigcup_{j\in I} \Phi_j$$

La topología τ_Φ *generada por la subbase* Φ *se conoce como la topología producto y a*

$$\left(\prod_{i\in I} X_i, \tau_\Phi\right)$$

se le conoce como espacio producto.

Observación. *La topología producto* τ_Φ *en* $\prod_{i\in I} X_i$ *tiene como base a todos los conjuntos de la forma* $\prod_{i\in I} U_i$*, en donde* U_i *es abierto en* X_i *para cada* $i \in I$ *y además* $U_i = X_i$ *excepto por un número finito de i's.*

Recordemos que el producto $\prod_{i\in I} X$, con (X, τ) un espacio topológico, también puede ser visto como el conjunto de todas las funciones de I en X, es decir,

$$\prod_{i\in I} X = X^I = \{f | f : I \to X \text{ es función}\}.$$

Ahora, para cada $x \in X$ tenemos la proyección $\pi_x : Y^X \to X$ definida por $\pi_x(f) = f(x)$. De este modo podemos dar una subbase explícita para este caso.

Definición 4.2.2. *Sean* Y *un conjunto y* X *un espacio topológico. Dado cualquier* $y \in Y$ *y* $U \subset X$ *un abierto, entonces definimos*

$$W(y, U) = \{f \in X^Y | f(y) \in U\}.$$

El conjunto $\mathcal{W} = \{W(y, U) | y \in Y$ *y* $U \subset Y$ *es abierto*$\}$ *forma una subbase para la topología producto en* $\prod_{i\in I} X$*.*

Es claro que $\mathcal{W} = \Phi$ en este caso ya que para todo $y \in Y$ y $U \subset X$ abierto se tiene que $\pi_y^{-1}(U) = \{f \in X^Y | f(y) \in U\} = W(y, U)$.

Lema 4.2.1. *(Propiedades de la topología producto). Consideremos de nuevo un conjunto arbitrario* Y*, un espacio topológico* X *y para cada* $y \in Y$ *la proyección* π_y*. Entonces*

(1) Cada π_y es continua en la topología producto.

(2) La topología producto es la topología más débil o gruesa en la cual todas las funciones π_y son continuas.

Prueba. La primera afirmación es muy sencilla de probar pues como se observó hace unos instantes $\pi_y^{-1}(U) = W(y, U)$, es decir, los elementos de la subbase (que son abiertos) son preimágenes de abiertos, por lo que π_y es continua para cada $y \in Y$. Para la segunda afirmación supongamos que existe τ una topología en X^Y que hace continuas a todas las proyecciones π_y, debemos mostrar que $\tau_\Phi \subset \tau$. En efecto, damos un elemento básico $V \in \tau_\Phi$ entonces V es intersección finita de elementos subbásicos, es decir,

$$V = \bigcap_{i=1}^{n} W(y, U) = \bigcap_{i=1}^{n} \pi_y^{-1}(U),$$

en donde $U \subset Y$ es abierto. Como π_y es continua respecto a τ entonces $\pi_y(U)$ es abierto y ya que intersección finita de abiertos es abierto entonces V es un abierto en τ, como era un elemento básico tenemos entonces que $\tau_\Phi \subset \tau$.

□

Teorema 4.2.1. *La convergencia en la topología producto está dada por la convergencia puntual. Es decir, sean $\{f_n\}_n \subset X^Y$ y $f \in X^Y$. Entonces f_n converge a f en la topología producto si y sólo si f_n converge puntualmente a f.*

Demostración. Sea $\{f_n\}_n$ tal que f_n converge a f en la topología producto, si U es una vecindad de f(y) entonces $W(y, U)$ es una vecindad de f en X^Y de esta manera $f_n \in W(y, U)$ para toda n excepto por un número finito de ellas y así $f_n(y) \in U$ para toda n excepto por un número finito de ellas, es decir, $f_n(y)$ converge a $f(x)$. El regreso es completamente análogo: si f_n converge puntualmente a f y $W(y, U)$ es una vecindad de f en X^Y entonces U es vecindad de $f(y)$ en X, luego $f_n(y) \in U$ para toda n excepto por un número finito de ellas, de esta forma $f_n \in W(y, U)$ para un número finito de ellas, es decir, f_n converge a f en la topología producto.

□

Teorema 4.2.2. *(Tychonoff). El producto arbitrario de conjuntos compactos es un espacio compacto en la topología producto. Es decir, si X_i es compacto para cada $i \in I$ entonces*

$$\prod_{i \in I} X_i$$

es compacto.

Demostración. [27], pág. 234. □

Lema 4.2.2. *El producto de espacios Hausdorff es un espacio Haudorff. Es decir, si X_i es Hausdorff para cada $i \in I$ entonces $\prod_{i\in I} X_i$ es Haudorff.*

Prueba. Sean $x = (x_i)_i$ y $y = (y_i)_i$ elementos de $\prod_{i\in I} X_i$ tales que $x \neq y$, luego existe al menos un índice, digamos j, para el cual $x_j \neq y_j$. Ahora, como X_i es Haudorff para cada $i \in I$ entonces tenemos que para este índice j existen $U_j, V_j \subset X_j$ abiertos tales que $x_j \in U_j$, $y_j \in V_j$ y $U_j \cap V_j = \emptyset$. Consideremos $U = \prod_{i\in I} U_i$ y $V = \prod_{i\in I} V_i$ con $U_i = X_i$ y $V_i = X_i$ excepto para el índice j, luego U y V son abiertos en $\prod_{i\in I} X_i$ tales que $x \in U$, $y \in V$ y además $U \cap V = \emptyset$, por lo tanto $\prod_{i\in I} X_i$ es Hausdorff.

□

Lema 4.2.3. *La intersección arbitraria de espacios compactos es un espacio compacto. Es decir, sea X un espacio Hausdorff, si $\{X_i\}_{i\in I}$ es una colección de subespacios de X en donde cada X_i es compacto entonces*

$$\bigcap_{i\in I} X_i$$

es compacto.

Prueba. Recordemos que dado X un espacio topológico Hausdorff y $A, B \subset X$ tales que $B \subset A$, A es compacto y B cerrado entonces B es compacto. Por lo anterior basta probar que $\cap_{i\in I} X_i$ es cerrado ya que $\cap_{i\in I} X_i \subset X_i$ y X_i es compacto. Claramente $\cap_{i\in I} X_i$ es cerrado ya que es intersección arbitraria de cerrados pues cada X_i es cerrado al ser compacto.

□

Observación. *Es importante notar que la hipótesis de que X sea Hausdorff es necesaria pues que todo subconjunto compacto sea cerrado requiere la propiedad de ser Hausdorff.*

Lema 4.2.4. *Un conjunto C es compacto si y sólo si toda colección de conjuntos cerrados $\{C_i\}_{i\in I}$, con $C_i \subset C$ para cada i, que tiene la propiedad de intersección finita tiene intersección no vacía, i.e.,*

$$\bigcap_{i\in I} C_i \neq \emptyset.$$

Prueba. [27], pág. 169. □

Definición 4.2.3. *Sea (X, τ) un espacio topológico. Para $Y \subset X$ se define la topología relativa o topología de subespacio en Y como*

$$\tau_Y = \{U \cap Y | U \in \tau\}.$$

4.3. Análisis real y Análisis funcional

Como pudimos notar muchos resultados del análisis (real y funcional) fueron esenciales durante el desarrollo de los capítulos anteriores, notablemente, el teorema de Hahn-Banach y el teorema de Banach-Alaoglu. Es por ello que ahora desarrollamos de manera breve un poco esta teoría. Comenzamos con el análisis de los hiperplanos.

Consideremos un espacio vectorial real X y una variedad lineal en X, es decir, un conjunto de la forma $\{x + Y\}$ donde $x \in X$ está fijo y Y es un subespacio lineal de X.

Definición 4.3.1. *Una variedad lineal tal que Y es subespacio propio maximal*[2] *de X se conoce como hiperplano y lo denotaremos como H.*

Observemos que dado un hiperplano H éste queda determinado por algún funcional lineal y un real, es decir, para cada hiperplano H existe ϕ_H un funcional lineal y $\lambda_H \in \mathbb{R}$ tal que $H = \phi_H^{-1}(\lambda_H) = \{x \in X | \phi_H(x) = \lambda_H\}$. En efecto, notemos que H puede o no ser subespacio vectorial de X, si H no es subespacio vectorial de X entonces en primer lugar $x \notin Y$ pues Y sí es subespacio vectorial y en segundo lugar

$$span(Y \cup \{x\}) := \{\alpha y_1 + \beta y_2 | y_1, y_2 \in Y \cup \{x\} \ \ y \ \ \alpha, \beta \in \mathbb{R}\} = X$$

ya que Y es subespacio vectorial maximal, es decir, si suponemos que

$$span(Y \cup \{x\}) \neq X$$

se tiene que $span(Y \cup \{x\})$ es un subespacio vectorial de X que contiene a Y y es distinto de X, de este modo Y no es maximal, lo cual es una contradicción. Además $span(Y \cup \{x\}) = \{y + \lambda x | y \in Y, \ \lambda \in \mathbb{R}\}$, de esta forma definimos $\phi : X \to \mathbb{R}$ como $\phi(y + \lambda x) = \lambda$ claramente ϕ es lineal y además $\phi^{-1}(1) = \{x + y | y \in Y\} = Y + x = H$ y tenemos el hiperplano. Ahora, si H es un subespacio vectorial de X tenemos que $H = Y$ y así tomamos cualquier $x \in X$ tal que $x \notin Y$ y por el mismo argumento de nuevo se tiene que $span(Y \cup \{x\}) = X$ y definimos ϕ de la misma manera, en este caso $\phi^{-1}(0) = Y = H$ y de nuevo tenemos a nuestro hiperplano.

Para facilitar la notación utilizaremos $H = \phi^{-1}(\lambda)$.

Observación. *En el caso en que X sea un espacio vectorial de dimensión n entonces cualquier hiperplano H en X tiene dimensión* $n - 1$.

Definición 4.3.2. *Dado un hiperplano* $H = \phi^{-1}(\lambda)$ *definimos las componentes inducidas por H como los subconjuntos de X* $\{x \in X | \phi(x) \leq \lambda\}$ *y* $\{x \in X | \phi(x) \geq \lambda\}$. *Si las desigualdades son estrictas entonces las llamamos componentes abiertas.*

[2] Y es subespacio propio máximal de X si el único subespacio de X que contiene propiamente a Y es X mismo.

Definición 4.3.3. *Dos subconjuntos A,B de X son separados por H si están contenidos en diferentes componentes, es decir, si* $A \subset \{x \in X | \phi(x) \leq \lambda\}$ *y* $B \subset \{x \in X | \phi(x) \geq \lambda\}$ *o vicerversa. Se dice que son separados de manera estricta si son las componentes abiertas.*

Teorema 4.3.1. *(Hahn-Banach). Sean A,B dos subconjuntos no vacíos, ajenos y convexos de un espacio vectorial topológico X.*

(a) Si A es abierto entonces existen $\phi \in X^*$ *y* $\alpha \in \mathbb{R}$ *tales que*

$$Re(\phi(a)) < \alpha \leq Re(\phi(b))$$

para toda $a \in A$ *y para toda* $b \in B$.

(b) Si A es compacto, B es cerrado y X es localmente convexo entonces existen $\phi \in X^*$, $\alpha_1, \alpha_2 \in \mathbb{R}$ *tales que*

$$Re(\phi(a)) < \alpha_1 < \alpha_2 < Re(\phi(b))$$

para toda $a \in A$ *y para toda* $b \in B$.

Demostración. [31], pág. 59. □

Observación. *Si el campo con el que se está trabajando sobre el espacio vectorial es* $\mathbb{R}$ *entonces claramente* $Re(\phi) = \phi$.

Durante la demostración del Lema 2.2.5 hicimos uso de algunos hechos de gran importancia que se enuncian a continuación.

Recordemos que si tenemos un espacio de medida fijo $(S, \mathcal{B}, \mu)$, f es una función medible sobre S y $0 < p < \infty$, se definen la norma

$$||f||_p = \left[\int |f|^p \quad d\mu \right]^{\frac{1}{p}}$$

y el espacio $L_p(S, \mathcal{B}, \mu) = \{f : S \to \mathbb{C} | \text{f es medible y } ||f||_p < \infty\}$. Para simplificar la notacion denotaremos por L_p a $L_p(S, \mathcal{B}, \mu)$ siempre y cuando no exista motivo de confusión y escribiremos $L_p(\mu)$ para hacer referencia a la medida con la cual estamos trabajando. Recordemos también que

$$||f||_\infty = \text{ínf}\{a \geq 0 | \mu(\{x : |f(x)| > a\}) = 0\}$$

$||\cdot||_\infty$ se conoce como el supremo esencial de $|f|$ y es una norma. Ahora, definimos el espacio

$$L_\infty = L_\infty(S, \mathcal{B}, \mu) = \{f : S \to \mathbb{C} | \text{f es medible y } ||f||_\infty < \infty\}.$$

Además, como tenemos una identificación en estos conjuntos, es decir, $f = g$ si $f(s) = g(s)$ para toda $s \in S$ excepto por un conjunto de medida

cero relativo a μ, entonces resulta que $f \in L_\infty$ si y sólo si existe una función medible y acotada g tal que $f = g$ casi donde sea. Notemos que como μ sólo determina los conjuntos de medida cero entonces si μ y ν son mutuamente absolutamente continuas[3] entonces $L_p(\mu) = L_p(\nu)$. Por último, recordemos que $(L_p, ||\cdot||_p)$ con $0 < p \leq \infty$ es un espacio de Banach.

Denotaremos al espacio dual de L_p, es decir, el espacio de funcionales lineales acotados sobre L_p, como $(L_p)'$.

Teorema 4.3.2. *(Representación de Riesz). Sean p y q exponentes conjugados, i.e.,* $\frac{1}{p} + \frac{1}{q} = 1$. *Si* $1 < p < \infty$ *para cada* $\varphi \in (L_p)'$ *existe* $g \in L_q$ *tal que*

$$\varphi(f) = \int fg \ d\mu \quad \forall f \in L_p.$$

Si μ es σ-finita entonces lo anterior también se tiene para p=1.

Demostración. [17], pág. 190. □

Teorema 4.3.3. *(El espacio dual de L_1). Sea $(S, \mathcal{B})$ un espacio medible y sea μ una medida σ-finita entonces se tiene que $[L_1(\mu)]'$, el espacio dual de $L_1(\mu)$, es $L_\infty(\mu)$.*

Demostración. [17], pág. 190. □

En la demostración del Teorema 2.3.2 usamos algunos hechos que tienen que ver con un concepto conocido como dualidad, así como los conjuntos polares y el teorema de Banach-Alaoglu. Comenzamos definiendo la dualidad entre dos espacios vectoriales reales. Los conceptos son tomados de [6] y [7].

Definición 4.3.4. *Sean X y Y dos espacios vectoriales reales. Una dualidad entre X y Y es un una función bilineal $\Psi : X \times Y \to \mathbb{R}$. Una dualidad es llamada estricta o no degenerada si para cada $x \in X$, $x \neq 0$ existe un $y \in Y$ tal que $\Psi(x, y) \neq 0$ y para cada $y \in Y$, $y \neq 0$ existe un $x \in X$ tal que $\Psi(x, y) \neq 0$, es decir, para cada $(x_0, y_0) \in X \times Y$ $x_0 \neq 0$ $y_0 \neq 0$ las funciones $x \mapsto \Psi(x, y_0)$ y $y \mapsto \Psi(x_0, y)$ son distintas de cero. En algunas ocasiones se denota a $\Psi(x, y) = \langle x, y \rangle$ si no hay motivo de confusión. Por lo que si X y Y están en dualidad lo denotaremos como $\langle X, Y \rangle$.*

Observación. *Una dualidad induce las funciones $\psi : X \to Y'$ y $\phi : Y \to X'$ definidas por $x \mapsto \psi_x$ y $y \mapsto \phi_y$ en donde $\psi_x(y) = \langle x, y \rangle$ y $\phi_y(x) = \langle x, y \rangle$.*

Observación. *Si X es un espacio vectorial entonces X y X' siempre están en dualidad mediante $\langle x, \varphi \rangle = \varphi(x)$. Más aún, si X es un espacio vectorial topológico localmente convexo por el teorema de Hahn-Banach X y X' están en dualidad estricta.*

[3] cf. Definición 4.4.5.

Toda dualidad entre dos espacio vectoriales X y Y nos induce dos topología en dichos espacios, a saber la topología débil en X y Y denotadas por $\sigma(X,Y)$ y $\sigma(Y,X)$ respectivamente. Se define cada una como la topología más gruesa o débil que hace continuas a la familia de funciones $\{\phi_y|y \in Y\}$ y $\{\psi_x|x \in X\}$ respectivamente. Una base de abiertos alrededor del 0 para la topología $\sigma(X,Y)$ es la conformada por lo conjuntos $\{x \in X \big| |\langle x,y\rangle| < \epsilon\}$ con $y \in Y$ y $\epsilon > 0$.

También es posible definir la topología débil de manera más general, es decir, sin la necesidad de que los espacios sean espacios vectoriales y sin necesidad de una dualidad.

Sea X un conjunto distinto del vacío y $\{(X_i,\tau_i)\}_{i\in I}$ una familia de espacios topológicos indexada por I. Consideremos

$$\mathcal{F} = \{f_i : X \to X_i\},$$

una familia de funciones, con esta familia podemos generar una topología en X, a saber, la topología débil generada por $\mathcal{F}$.

Definición 4.3.5. *La topología débil en X es la topología generada por la subbase*

$$\{f_i^{-1}(U_i)|U_i \in \tau_i,\ \ i \in\ I\}.$$

Esta topología es la topología más gruesa o débil que hace a todas las funciones de $\mathcal{F}$ continuas.

Observación. *Claramente si consideramos a X y Y espacios vectoriales topológicos y la dualidad $\langle X,Y\rangle$ entonces ambas definiciones de la topología débil coinciden. En efecto, consideremos a $\mathcal{F} = \{\phi_y : X \to \mathbb{R}_y|y \in Y\}$, en donde $\phi_y = \langle x,y\rangle$, $(\mathbb{R}_y,\tau_y) = (\mathbb{R},\tau)$ para toda $y \in Y$, con τ la topología usual en $\mathbb{R}$; la topología generada por $\mathcal{F}$ es la misma que $\sigma(X,Y)$. Análogamente para $\mathcal{F} = \{\psi_x : Y \to \mathbb{R}_x|x \in X\}$, $\psi_x = \langle x,y\rangle$ y $\sigma(Y,X)$.*

Lema 4.3.1. *(La topología débil y la topología producto coinciden[4]). Sea (X,τ) un espacio topológico. Consideremos el producto $\prod_{i\in I} X$ y la proyección $\pi_i : X^I \to X$, $\pi_j((x)_{i\in I}) = x_j$. Entonces la topología en X generada por $\mathcal{F} = \{\pi_i|i \in I\}$ coincide con la topología producto en $\prod_{i\in I} X$.*

Prueba. Recordemos que X es el conjunto de funciones de I en X_i, i.e.,

$$\prod_{i\in I} X_i = \{x : I \to X_i|x(i) \in X_i\} = \{(x_i)_{i\in I}|x_i \in X_i\},$$

la topología generada por $\mathcal{F} = \{\pi_i|i \in I\}$ tiene como subbase a

$$\{\pi_i^{-1}(U_i)|U_i \in \tau_i,\ \ i \in I\} = \Phi$$

y Φ es la subbase que genera a la topología producto.

□

[4]Comparese con (2) del Lema 4.2.1.

Una observación interesante es la siguiente:

Observación. *Si X es un espacio vectorial topológico entonces $X' \subset \mathbb{R}^X$ si dotamos a este último de la topología producto entonces observamos que $(X', \sigma(X', X))$ en donde (por la observación anterior) $\sigma(X', X)$ es la topología inducida por las funciones $\{\psi_x : X' \to \mathbb{R}_\varphi | \varphi \in X'\}$ con $\psi_x = \langle x, \varphi \rangle = \varphi(x)$, es decir, ψ_x es la función evaluación que de hecho es $\pi_x(\varphi)$. Luego por el lema anterior se tiene que $\sigma(X', X)$ es en realidad la topología heredada de la topología producto en $\mathbb{R}^X$.*

De hecho la observación anterior aplica para cualquier familia de funciones que sean las evaluaciones independientemente de si es un espacio vectorial o no. En particular, si consideramos $(S, \mathcal{B})$ un espacio medible, en donde $\mathcal{B}$ no es obligatoriamente la σ-álgebra de Borel, entonces $\Lambda \subset [0,1]^{\mathcal{B}}$, i.e., el espacio de medidas de probabilidad sobre S es un subconjunto de las funciones que van de la σ-álgebra $\mathcal{B}$ a $[0,1]$, luego si consideramos $\Gamma_B : \Lambda \to [0,1]$ $\Gamma_B(P) = P(B)$, $B \in \mathcal{B}$ es la función evaluación o la proyección de P en $[0,1]_B$, por lo que la topología más gruesa que hace a cada función de la familia $\mathcal{F} = \{\Gamma_B | B \in \mathcal{B}\}$ continua es la topología débil que es la topología heredada de la topología producto y al mismo tiempo es la topología de convergencia puntual.

Por otro lado, también se sabe que cualquier espacio vectorial X puede ser encajado en su doble dual $(X')'$ (denotado por X'') de la siguiente forma $e : X \to X''$, $x \mapsto e_x$ con $e_x(\varphi) = \varphi(x)$; claramente es una función inyectiva. Naturalmente tenemos una familia de funciones con la cual generar una topología débil en X', a saber, $\mathcal{F} = \{e_x : X' \to \mathbb{R} | x \in X''\}$.

Definición 4.3.6. *Dada la familia $\mathcal{F} = \{e_x : X' \to \mathbb{R}\}$ definimos la topología débil-$*$ (pronunciado débil estrella) como la topología más gruesa que hace a cada función de la familia $\mathcal{F}$ continua.*

Observamos que esta topología es más débil o gruesa que la topología débil en X'. Si X es reflexivo entonces $X'' = X$ y ambas topologías coinciden. A veces se define a la topología débil-$*$ como la topología $\sigma(X', X)$, es decir, la topología débil en X' inducida por la dualidad $\langle X, X' \rangle$; en este caso se consideran a la topología débil y a la topología débil-$*$ en X' como la misma topología (cf. [23]).

También tenemos que si X es un espacio topológico vectorial arbitrario cualquier funcional lineal sobre $C(X)$, i.e., cualquier $\varphi \in C(X)'$ (en donde C(X) es el espacio de las funciones reales y continuas sobre X) se ve de la siguiente manera:

$$\varphi(f) = \int f \ d\mu = \langle f, \mu \rangle.$$

En donde m pertenece al espacio de las medidas generalizadas sobre X denotado por $M(X)$. De hecho, se tiene que $C(X)^* = M(X)$ (cf. [26] y [1]); así, se tiene la dualidad $\langle M(X), C(X)\rangle$ y la topología débil en $M(X)$ es la topología más gruesa que hace a las funciones $\{\psi_f | f \in C(X)\}$ continuas, en donde $\psi_f(\mu) = \langle f, \mu\rangle = \langle f, \varphi\rangle = \varphi(f)$. Está topología es la heredada de la topología producto en $\mathbb{R}^{C(X)}$ pues por lo anterior podemos identificar a cada $\mu \in M(X)$ con un funcional lineal, es decir, $M(X)$ es un subespacio de $\mathbb{R}^{C(X)}$. Como $\Lambda(X) \subset M(X)$ ésta es la topología que hereda $\Lambda(X)$. Una base alrededor de una medida de probabilidad $\nu \in \Lambda(X)$ para esta topología es la conformada por los elementos de la siguiente forma:

$$\left\{\mu \in \Lambda(X) \middle| \quad \left|\int f_i \ d\mu - \int f_i \ d\nu\right| < \epsilon \quad i = 1, ..., k\right\}.$$

Observación. *Algunos autores llaman topología vaga a la topología débil-$*$ (cf. [6], [17]) mientras que otros hacen una distinción (cf. [26]) ya que depende del espacio que tomemos para las funciones f, es decir, si se pide que sean continuas, acotada, con soporte compacto, etc.*

Un análisis más profundo alrededor de estos conceptos y resultados puede consultarse en [5], [23] y [26].

Definición 4.3.7. *Sean X y Y dos espacios vectoriales que están en dualidad. Para un subconjunto $E \subset X$ el conjunto polar o simplemente el polar de E se define de la siguiente manera:*

$$E^\circ = \{y \in Y | \langle x, y\rangle \leq 1 \quad \forall x \in E\} \subset Y.$$

Definición 4.3.8. *Sean X y Y dos espacios vectoriales que están en dualidad. Para un subconjunto $E \subset X$ el conjunto bipolar o simplemente el bipolar de E se define de la siguiente manera*

$$E^{\circ\circ} = (E^\circ)^\circ = \{x \in X | \langle x, y\rangle \leq 1 \quad \forall x \in E^\circ\}.$$

Teorema 4.3.4. *(Teorema Bipolar). Sean X y Y dos espacios vectoriales que están en dualidad estricta. Usando la topología $\sigma(X, Y)$ en X, para un subconjunto $E \subset X$ se tiene que*

$$E^{\circ\circ} = cd(conv(E \cup \{0\}))$$

Demostración. [7], pág. 51. □

Teorema 4.3.5. *(Banach-Alaoglu-Bourbaki). Sean X un espacio topológico vectorial y*

$$K = \{\varphi \in X' | \varphi(x) \leq 1 \ \text{ para cada } x \in V\},$$

en donde V es una vecindad del 0. Entonces K es compacto en la topología débil.

Demostración. [31], pág. 68. □

Observación. *En ocasiones el teorema de Banach-Alaoglu es llamado teorema de Alaoglu-Bourbaki ya que en primera lugar fue demostrado para espacios de Banach y después se dio la prueba en su versión más general que fue la presentada aquí.*

Teorema 4.3.6. *(Existencia de un resultante). Sea V un espacio localmente convexo y sea P una medida de probabilidad sobre $(V, \mathcal{B})$ con $\mathcal{B}$ la σ-álgebra de Borel, sea $E \subset V$ compacto, convexo tal que $P(E) = 1$. Entonces $E(P)$, la esperanza o resultante de P, existe, es única y además $E(P) \in E$.*

Demostración. [7], pág. 115. □

Teorema 4.3.7. *(Choquet). Sea V un espacio localmente convexo, consideremos el espacio medible $(V, \mathcal{B})$ con $\mathcal{B}$ la σ-álgebra de Borel. Sean $X \subset V$ un conjunto metrizable, convexo y compacto y $x_0 \in X$. Entonces existe P una medida de probabilidad sobre X tal que $E(P) = x_0$ y el soporte de P son los puntos extremos de X.*

Demostración . [28], pág. 14. □

Lema 4.3.2. *Sea V un espacio localmente convexo, consideremos el espacio medible $(V, \mathcal{B})$ con $\mathcal{B}$ la σ-álgebra de Borel. Entonces la función*

$$E : \mho_n \to S, \quad \mu \mapsto E(\mu)$$

es continua en la topología débil-$*$ (también conocida como topología vaga) para toda n. Lo anterior quiere decir que para cualquier $R \in \mho_n$ y para cualquier U vecindad del 0 en V existen $f_1, ..., f_k$ funciones continuas sobre K_n y $\delta_i > 0$ tales que si $P \in \mho_n$ y

$$\left| \int_V f_i \ dP - \int_V f_i \ dR \right| < \delta_i, \quad i = 1, ..., k \ \text{ entonces } \ E(P) \in E(R) + U.$$

Prueba. [7], pág. 115. □

4.4. Teoría de la medida

Todos los conceptos utilizados en esta subsección dedicada a teoría de la medida se pueden encontrar en [18], [17]. Otro material de consulta fueron [30], [4] y [21].

Una medida sobre un espacio medible $(S, \mathcal{B})$ es una función conjuntista $\mu : \mathcal{B} \to [0, \infty]$ tal que

1. $\mu(\emptyset) = 0$

2. Si $\{E_i\}_{i=1}^{\infty}$ es una sucesión de conjuntos ajenos tales que $E_i \in \mathcal{B}$ entonces

$$\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i).$$

A 2. se le conoce como aditividad contable o σ-aditividad. Es claro que esta implica la aditividad finita, es decir, si μ es σ-aditiva entonces es finitamente aditiva. Recordemos también que una medida se dice que es σ-finita si $S = \cup_{n\in\mathbb{N}} E_i$ con $E_i \in \mathcal{B}$ y $\mu(E_i) < \infty$ para toda i.

Definición 4.4.1. *Sea $(S, \mathcal{B})$ un espacio medible. Decimos que una medida finitamente aditiva μ es continua por abajo si para $\{E_i\}_{i=1}^{\infty} \subset \mathcal{B}$ tal que $E_i \subset E_{i+1}$ para toda i se tiene que*

$$\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \lim_{i\to\infty} \mu(E_i).$$

Análogamente decimos que es continua por arriba si para $\{E_i\}_{i=1}^{\infty} \subset \mathcal{B}$ tal que $E_{i+1} \subset E_i$ para toda i y $\mu(E_1) < \infty$ se tiene que

$$\mu\left(\bigcap_{i=1}^{\infty} E_i\right) = \lim_{i\to\infty} \mu(E_i).$$

Diremos que μ es continua si es continua por abajo y continua por arriba.

Es de conocimiento general que si μ es σ-aditiva entonces es continua por abajo y continua por arriba, y por lo tanto es continua. Ahora probaremos que si μ es una medida finitamente aditiva que es continua en $\emptyset$ (lo cual se define a continuación) entonces es continua.

Definición 4.4.2. *Diremos que μ es continua en $\emptyset$ si μ es continua por arriba en $\emptyset$, es decir, si $\{E_i\}_{i=1}^{\infty}$ es una sucesión de conjuntos tales que $E_i \in \mathcal{B}$, $E_{i+1} \subset E_i$ y $\cap_{i\in\mathbb{N}} E_i = \emptyset$ entonces*

$$\mu\left(\bigcap_{i\in\mathbb{N}} E_i\right) = \lim_{i\to\infty} \mu(E_i).$$

Claramente una medida finitamente aditiva μ siempre es continua por abajo en $\emptyset$ pues la única manera en que sucede $\cup_{i\in\mathbb{N}} E_i = \emptyset$ con $E_i \in \mathcal{B}$ y $E_i \subset E_{i+1}$ para toda i es cuando $E_i = \emptyset$ para toda i y así tenemos que

$$\mu\left(\bigcup_{i\in\mathbb{N}} E_i\right) = \mu(\emptyset) = 0 = \lim_{i\to\infty} 0 = \lim_{i\to\infty} \mu(E_i).$$

Lema 4.4.1. *Sea $(S, \mathcal{B})$ un espacio medible y μ una medida aditivamente finita. Supongamos que μ es continua en $\emptyset$ entonces es σ-aditiva y en consecuencia es continua.*

Prueba. Sean $\{A_i\}_{i\in\mathbb{N}}$ conjuntos ajenos tales que $A_i \in \mathcal{B}$ para toda i. Debemos mostrar que

$$\mu\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} \mu(A_i).$$

Para ello tomemos $A = \cup_{i\in\mathbb{N}} A_i$ y definamos $B_n = \cup_{i=1}^{n} A_i$ y $C_n = \cup_{i=n+1}^{\infty} A_i$. Ahora, es claro que $A = B_n \cup C_n$ para toda n, observemos que por la aditividad finita de μ se tiene que

$$\mu(A) = \mu(B_n \cup C_n) = \mu(B_n) + \mu(C_n) = \sum_{i=1}^{n} \mu(A_i) + \mu(C_n).$$

Por otro lado, tenemos que $C_{n+1} \subset C_n$, además

$$\cap_{n\in\mathbb{N}} C_n = \cap_{n\in\mathbb{N}} (\cup_{i=n+1}^{\infty} A_i) = \emptyset.$$

La continuidad por abajo de μ nos da que $\mu(C_n)$ converge a 0 cuando n tiende a infinito, por lo cual tenemos que para toda $\epsilon > 0$ existe $N \in \mathbb{N}$ tal que si $n \geq N$ entonces $0 \leq \mu(C_n) < \epsilon$. Tomemos ahora una M fija tal que $M > N$ luego

$$\mu(A) = \sum_{i=1}^{M} \mu(A_i) + \mu(C_M) < \sum_{i=1}^{M} \mu(A_i) + \epsilon \leq \sum_{i=1}^{\infty} \mu(A_i) + \epsilon.$$

Como lo anterior se cumple para toda $\epsilon > 0$ hacemos tender ϵ a 0 y obtenemos la σ-subaditividad de μ:

$$\mu(A) \leq \sum_{i=1}^{\infty} \mu(A_i).$$

Ahora, sólo resta mostrar que

$$\sum_{i=1}^{\infty} \mu(A_i) \leq \mu(A),$$

para lo cual observamos que como

$$\sum_{i=1}^{n} \mu(A_i) \to \sum_{i=1}^{\infty} \mu(A_i) \quad \text{cuando} \quad n \to \infty$$

entonces para toda $\epsilon > 0$ existe $N \in \mathbb{N}$ tal que si $n \geq N$ se tiene que

$$\left|\sum_{i=1}^{\infty}\mu(A_i)-\sum_{i=1}^{n}\mu(A_i)\right|=\sum_{i=1}^{\infty}\mu(A_i)-\sum_{i=1}^{n}\mu(A_i)<\epsilon.$$

Es decir,

$$\sum_{i=1}^{\infty}\mu(A_i)<\sum_{i=1}^{n}\mu(A_i)+\epsilon.$$

Como $\cup_{i=1}^{n}A_i\subset\cup_{i\in\mathbb{N}}A_i=A$ tenemos que

$$\mu(A)=\mu\left(\bigcup_{i=1}^{\infty}A_i\right)=\sum_{i=1}^{\infty}\mu(A_i)<\sum_{i=1}^{n}\mu(A_i)+\epsilon\leq\sum_{i=1}^{\infty}\mu(A_i)+\epsilon,$$

Ya que, de nuevo, lo anterior fue para toda $\epsilon>0$ y hacemos tender ϵ a 0 obtenemos que

$$\mu(A)\leq\sum_{i=1}^{\infty}\mu(A_i).$$

Por lo tanto μ es σ-aditiva, luego es continua.

□

Corolario 4.4.1. *μ una medida aditivamente finita es σ-aditiva si y sólo si es continua en $\emptyset$*

Prueba. Se sigue del Lema 4.4.1. □

Durante los tres capítulos anteriores hicimos un uso extensivo de varios resultados de teoría de la medida que son ampliamente conocidos y utilizados, a saber: la desigualdad de Jensen, el teorema de convergencia monótona, el teorema de convergencia dominada de Lebesgue (CDL), el lema de Fatou, el teorema de Radon-Nikodým y algunos hechos realacionados o derivados de estos resultados. A continuación los enunciamos y demostramos en algunos casos. Continuaremos trabajando con un espacio de medida $(S,\mathcal{B},\mu)$.

Teorema 4.4.1. *(Desigualdad de Jensen). Sea μ una medida positiva sobre $(S,\mathcal{B})$ tal que $\mu(S)=1$. Si f es una función real tal que $f\in L_1(\mu)$, $a<f(s)<b$ para toda $s\in S$ y φ es convexa en (a,b) entonces*

$$\varphi\left(\int_S f d\mu\right)\leq\int_S(\varphi\circ f)\ d\mu.$$

Los casos $a=\infty$ y $b=-\infty$ están considerados.

Demostración. [30], pág. 63. □

Teorema 4.4.2. *(Convergencia monónota). Sean $\{f_n\}_{n\in\mathbb{N}}$ una sucesión de funciones tales que $f_n : S \to [0,\infty]$, $f_n \leq f_{n+1}$, f_n es medible para toda $n \in \mathbb{N}$ y supongamos que*

$$f(s) = \lim_{n\to\infty} f_n(s) \qquad [\mu].$$

Entonces f es medible y

$$\lim_{n\to\infty} \int f_n \ d\mu = \int f \ d\mu.$$

Demostración. [30], pág. 22. □

Teorema 4.4.3. *(Lema de Fatou). Sea $\{f_n\}_{n\in\mathbb{N}}$ una sucesión de funciones medibles sobre $(S,\mathcal{B})$ entonces*

$$\int \liminf_{n\to\infty} f_n \ d\mu \leq \liminf_{n\to\infty} \int f_n \ d\mu.$$

Demostración. [17], pág 52. □

Teorema 4.4.4. *(Convergencia dominada de Lebesgue (CDL)). Sea $\{f_n\}_{n\in\mathbb{N}} \subset L_1(\mu)$ una sucesión de funciones tales que:*

(a)f_n converge a f casi donde sea relativo a μ.

(b)Existe una función no negativa $g \in L_1(\mu)$ tal que $|f_n| \leq g$ para toda n, casi donde sea relativo a μ.

Entonces $f \in L_1(\mu)$ y

$$\int f \ d\mu = \lim_{n\to\infty} \int f_n \ d\mu$$

Demostración. [30], pág. 54. □

Durante el desarrollo de los resultados trabajamos un poco con las medidas con signo por lo cual revisamos algunos resultados.

Definición 4.4.3. *Sea $(S,\mathcal{B})$ un espacio medible. Una medida con signo es una función conjuntista $\nu :\to \mathbb{R} \cup \{-\infty,\infty\}$ tal que*

(i) $\nu(\emptyset) = 0$.

(ii) ν toma a lo más uno de los valores extendidos ∞ o $-\infty$.

(iii) ν es σ-aditiva.

Claramente toda medida es una medida con signo pero no vicerversa.

Teorema 4.4.5. *(Descomposición de Jordan). Sean $(S,\mathcal{B})$ un espacio medible y μ una medida con signo. Entonces existen dos medidas μ^+, μ^- llamadas la variación positiva y la variación negativa (respectivamente), alguna de ellas finita tales que $\mu = \mu^+ - \mu^-$*

Demostración. [18], pág. 127. □

Definición 4.4.4. *Sean $(S, \mathcal{B})$ un espacio medible y ν, μ dos medidas sobre $(S, \mathcal{B})$. Decimos que ν es absolutamente continua respecto a μ denotado por $\nu \ll \mu$ si para $B \in \mathcal{B}$ sucede que $\nu(B) = 0$ siempre que $\mu(B) = 0$.*

Definición 4.4.5. *Sea $(S, \mathcal{B})$ un espacio medible y ν, μ dos medidas sobre $(S, \mathcal{B})$. Decimos que ν y μ son mutuamente absolutamente continuas denotado por $\nu \equiv \mu$ si sucede que $\nu \ll \mu$ y $\mu \ll \nu$.*

Lema 4.4.2. *Sea $(S, \mathcal{B})$ un espacio medible y ν, μ medidas con signo tales que $|\mu|(S) < \infty$ entonces $\mu \ll \nu$ si y sólo si para todo $\epsilon > 0$ existe un $\delta > 0$ que depende de ϵ tal que $|\mu(B)| < \epsilon$ para todo $B \in \mathcal{B}$ tal que $|\nu|(B) < \delta$*

Prueba. [18], pág. 130. □

Lema 4.4.3. *(Medidas con signo es un espacio vectorial). Sean $(S, \mathcal{B})$ un espacio medible y*

$$\mathbb{S} = \{\mu | \mu \text{ es medida con signo finita sobre } (S, \mathcal{B})\}.$$

Entonces $\mathbb{S}$ es un espacio vectorial con las siguientes operaciones binarias:

- $+ : \mathbb{S} \times \mathbb{S} \to \mathbb{S}$, *$\nu + \mu$ es la medida con signo finita que cumple $(\nu + \mu)(B) = \nu(B) + \mu(B)$ para toda $B \in \mathcal{B}$.*
- $\cdot : \mathbb{R} \times \mathbb{S} \to \mathbb{S}$, *$\lambda \cdot \mu$ es la medida con signo finita que cumple $(\lambda \cdot \mu)(B) = \lambda\mu(B)$ para toda $\lambda \in \mathbb{R}$ y para toda $B \in \mathcal{B}$.*

Prueba. Ambas operaciones son cerradas pues ν y μ son finitas. Las propiedades de asociatividad, conmutatividad, existencia del elemento neutro y existencia del elemento inverso para la suma son heredadas de $\mathbb{R}$, asimismo la existencia del elemento identidad respecto a la multiplicación escalar, la compatibilidad de la multiplicación escalar respecto a la multiplicación en el campo, la distributividad de la multiplicación escalar respecto a la suma de vectores y la distributividad de la multiplicación escalar respecto a la suma en el campo también son heredadas de $\mathbb{R}$ como espacio vectorial sobre $\mathbb{R}$.
□

Observación. *El conjunto de medidas de probabilidad sobre $(S, \mathcal{B})$ denotado por Λ es un subconjunto convexo*[5] *de $\mathbb{S}$.*

Por otro lado también nos interesa medir distancias entre dos medidas, específicamente, nos interesa saber cuando difieren una medida de probabilidad respecto a otra o en otras palabras establecer una métrica en $\mathbb{S}$ que cumpla ciertas condiciones como son, por ejemplo, el ser un espacio completo

[5] cf. La observación de la Definición 2.0.1 en el capítulo 2.

con dicha métrica. Una de las métricas más utilizadas que cumple con esta propiedad es la métrica de variación total que, de hecho, es producto de una norma. A continuación damos una definición de dicha norma y probamos algunas de sus propiedades.

Definición 4.4.6. *La distancia de variación total entre dos medidas finitas ν, μ sobre un espacio medible $(S, \mathcal{B})$ se define de la siguiente manera*

$$d_{VT}(\nu, \mu) = ||\nu - \mu||_{VT} = \sup_{A \in \mathcal{B}} |\nu(A) - \mu(A)|.$$

Denotaremos a la distancia de variación total entre dos medidas de probabilidad ν, μ simplemente como $d(\nu, \mu)$.

Teorema 4.4.6. *$(\mathbb{S}, ||\cdot||_{VT})$ es un espacio de Banach.*

Demostración. Sea $\{\mu_n\}_{n\in\mathbb{N}} \subset \mathbb{S}$ una sucesión de Cauchy luego dada $\epsilon > 0$ tenemos que existe $N \in \mathbb{N}$ tal que si $n, m > N$ entonces $||\mu_n - \mu_m||_{VT} < \epsilon$. En particular, se tiene que $|\mu_n(B) - \mu_m(B)| < \epsilon$ para toda $B \in \mathcal{B}$, i.e., $\{\mu_n(B)\}_{n\in\mathbb{N}}$, con $B \in \mathcal{B}$ fijo es una sucesión de reales que es de Cauchy como $\mathbb{R}$ es completo entonces existe $a_B \in \mathbb{R}$ tal que $\mu_n(B)$ converge a a_B para cada $B \in \mathcal{B}$. Así, de manera natural, definimos a $\mu : \mathcal{B} \to \mathbb{R}$ como

$$\mu(B) = \lim_{n\to\infty} \mu_n(B), \quad B \in \mathcal{B}.$$

Primero mostramos que $\mu \in \mathbb{S}$: Es claro que $\mu(\emptyset) = 0$ pues $\mu(\emptyset) = \lim_{n\to\infty} \mu_n(\emptyset) = 0$ por lo que basta mostrar que μ es finita y que dados $\{E_i\}_{i\in\mathbb{N}}$ ajenos tales que $E_i \in \mathcal{B}$ para toda i se tiene que

$$\mu\left(\bigcup_{i\in\mathbb{N}} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i).$$

En donde lo anterior se entiende de la siguiente manera, si $|\mu(\cup_{i\in\mathbb{N}} E_i)| < \infty$ entonces la serie converge absolutamente. Claramenete μ es finita pues $|\mu_n(B) - \mu(B)| < \epsilon$ para toda $B \in \mathcal{B}$ con $n > N$. Para lo segundo mostraremos que μ es continua en $\emptyset$ lo cual nos da que μ es σ-aditiva. Sea $\{F_i\}_{i\in\mathbb{N}}$ una sucesión de conjuntos decreciente, i.e. tal que $F_{i+1} \subset F_i$, $F_i \in \mathcal{B}$ para toda i y $\cap_{i\in\mathbb{N}} F_i = \emptyset$, debemos mostrar que $\mu(F_i)$ converge a 0 cuando i tiende a infinito, lo cual es sencillo de ver ya que como μ_n es continua en $\emptyset$ para cada n entonces $\mu_n(F_i)$ converge a 0 para toda n, luego dado $\epsilon > 0$ tomamos $N \in \mathbb{N}$ tal que $|\mu_n(F_i) - \mu(F_i)| < \frac{\epsilon}{2}$ para toda $n > N$ y $M \in \mathbb{N}$ tal que $|\mu_n(E_i)| < \frac{\epsilon}{2}$ para toda $i > M$, así si tomamos $N_0 = \text{máx}\{N, M\}$, $n > N_0$ e $i > N_0$ entonces se tiene que

$$|\mu(F_i)| = |\mu(F_i) - \mu_n(F_i) + \mu_n(F_i)| \leq |\mu(F_i) - \mu_n(F_i)| + |\mu_n(F_i)|$$

$$|\mu_n(F_i) - \mu(F_i)| + |\mu_n(F_i)| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Por lo tanto tenemos que μ es continua en $\emptyset$, luego es σ-aditiva.

Ahora mostramos que $||\mu_n - \mu||_{VT}$ converge a 0 cuando n tiende a infinito:

$$\lim_{n\to\infty} ||\mu_n - \mu||_{VT} = \lim_{n\to\infty} \sup_{B\in\mathcal{B}} |\mu_n(B) - \mu(B)| = \sup_{B\in\mathcal{B}} |\lim_{n\to\infty} \mu_n(B) - \mu(B)|$$

$$= \sup_{B\in\mathcal{B}} |0| = 0.$$

Por lo tanto tenemos que $(\mathbb{S}, ||\cdot||_{VT})$ es un espacio normado completo, es decir, es un espacio de Banach.

□

Observación. *Sea* $\Lambda = \{P | P$ *es medida de probabilidad sobre* $(S, \mathcal{B})\}$. *Entonces* Λ *es un subconjunto convexo y cerrado de* $(\mathbb{S}, ||\cdot||_{VT})$. *Que es convexo se probó al inicio del capítulo 2. Para ver que es cerrado tomemos una sucesión* $\{P_n\}_{n\in\mathbb{N}} \subset \Lambda$ *tal que* P_n *converge a* P *en variación total, para ver que* $P \in \Lambda$ *basta mostrar que* $0 \leq P(B) \leq 1$ *para toda* $B \in \mathcal{B}$ *y* $P(S) = 1$ *pues por lo anterior se tiene que* P *es una medida finita. Si* $B \in \mathcal{B}$ *claramente* $0 \leq P(B) \leq 1$ *pues*

$$P(B) = \lim_{n\to\infty} P_n(B) \leq 1.$$

Analogámente

$$0 \leq \lim_{n\to\infty} P_n(B) = P(B).$$

También tenemos que

$$P(S) = \lim_{n\to\infty} P_n(S) = 1,$$

por lo que efectivamente $P \in \Lambda$.

La idea de la distancia de variación total nos dice cual es la máxima variación entre las probabilidades del mismo evento. Intuitivamente nos interesa saber que tanto difieren las probabilidades para un mismo evento y al tomar el supremo de estas probabilidades nos da la máxima diferencia que, de cierta manera, nos da información de que tan "cercanas" son estas dos medidas de probabilidad (en términos de las probabilidades que asignan a cada evento). A veces (y casi siempre) es bastante complicado trabajar con supremos e ínfimos ya que es difícil conocer por completo el conjunto sobre el cual estamos optimizando, por ello damos a continuación una representación más amigable de la distancia de variación total cuando el espacio medible es a lo más infinito numerable.

Lema 4.4.4. *(Representación de la distancia de variación total). Sea Ω un conjunto a lo más infinito numerable. Sean μ y ν dos medidas de probabilidad sobre $(\Omega, \mathcal{B})$. Entonces*

$$||\nu - \mu||_{VT} = \frac{1}{2}\sum_{\omega\in\Omega} |\nu(\omega) - \mu(\omega)|.$$

Prueba. Sean $B = \{\omega \in \Omega | \mu(\omega) \leq \nu(\omega)\}$ y A cualquier evento de Ω entonces

$$\nu(A) - \mu(A) \leq \nu(A \cap B) - \mu(A \cap B) \leq \nu(B) - \mu(B). \tag{4.4}$$

La primera desigualdad se sigue ya que para todo $\omega \in B^c$ sucede que $\nu(\omega) - \mu(\omega) < 0$, en particular para todo $\omega \in A \cap B^c$; así, al no considerar estos elementos, es decir, considerar a $A \cap B$ se tiene que $\nu(A \cap B) - \mu(A \cap B)$ no puede disminuir comparado con $\nu(A) - \mu(A)$ pues le quitamos a A todos los elemento que aportaban una cifra negativa. Y la segunda desigualdad ya que $A \cap B \subset B$, por como está definido B la diferencia de probabilidades no puede de nuevo disminuir al aumentar elementos de B. De manera completamente análoga se tiene que

$$\mu(A) - \nu(A) \leq \mu(A \cap B^c) - \nu(A \cap B^c) \leq \mu(B^c) - \nu(B^c). \tag{4.5}$$

Observamos que las dos cotas superiores en (4.4) y (4.5) en realidad son la misma ya que

$$\mu(B^c) - \nu(B^c) = \mu(\Omega \setminus B) - \nu(\Omega \setminus B) = \mu(\Omega) - \mu(B) - \nu(\Omega) + \nu(B)$$

$$= \nu(B) - \mu(B).$$

Más aún, de hecho cuando tomamos $A = B$ o $A = B^c$ tenemos que $|\nu(A) - \mu(A)| = \nu(B) - \mu(B)$. Es decir, se tiene que

$$|\nu(A) - \mu(A)| \leq \nu(B) - \mu(B) \quad \forall A \subset \Omega$$

y la igualdad se da cuando $A = B$ o $A = B^c$, por lo que hemos encontrado al supremo que estabamos buscando. Entonces

$$||\nu - \mu||_{VT} = \sup_{A\subset\Omega} |\nu(A) - \mu(A)| = \nu(B) - \mu(B)$$

$$= \frac{1}{2}[\nu(B) - \mu(B) + \mu(B^c) - \nu(B^c)]$$

$$= \frac{1}{2}\left[\sum_{\omega\in B} |\nu(\omega) - \mu(\omega)| + \sum_{\omega\in B^c} |\nu(\omega) - \mu(\omega)|\right] = \frac{1}{2}\sum_{\omega\in\Omega} |\nu(\omega) - \mu(\omega)|.$$

□

También es importante notar que la convergencia en variación total es más fuerte que la convergencia de manera fuerte, la cual definimos a continuación.

Definición 4.4.7. *Sea $\{\mu\}_{n\in\mathbb{N}}$ una sucesión de medidas sobre un espacio medible $(S,\mathcal{B})$ decimos que μ_n converge de manera fuerte a una medida μ si*

$$\lim_{n\to\infty}\mu_n(B)=\mu(B)\quad\forall B\in\mathcal{B}.$$

Lema 4.4.5. *(La convergencia en VT implica la convergencia de manera fuerte). Sea $\{P_n\}_{n\in\mathbb{N}}$ una sucesión de medidas de probabilidad sobre un espacio medible $(S,\mathcal{B})$ tales que*

$$\lim_{n\to\infty}||P_n-P||_{VT}=0.$$

Entonces $P_n(B)$ converge a $P(B)$ para toda $B\in\mathcal{B}$.

Prueba. Como $\lim_{n\to\infty}||P_n-P||_{VT}=0$ tenemos que para toda $\epsilon>0$ existe $N\in\mathbb{N}$ tal que si $n>N$ se tiene que

$$\sup_{B\in\mathcal{B}}|P_n(B)-P(B)|<\epsilon,$$

en particular se tiene que $|P_n(B)-P(B)|<\epsilon$ para toda $B\in\mathcal{B}$, es decir, $P_n(B)$ converge a $P(B)$ para toda $B\in\mathcal{B}$ cuando n tiende a infinito.

□

Además observemos que si P_n converge a P en variación y f es una función medible y acotada se tiene que

$$\int f\ dP_n\to\int f\ dP\quad n\to\infty.$$

Lo anterior es fácil de ver ya que para toda función medible y acotada f existe una sucesión de funciones simples $\{s_m\}_{m\in\mathbb{N}}$ que converge de manera uniforme a f y también se tiene lo siguiente para cada s_m:

$$\lim_{n\to\infty}\int s_m\ dP_n=\lim_{n\to\infty}\sum_{i=1}^{k}\alpha_i^m P_n(B_i^m)=\sum_{i=1}^{k}\alpha_i^m\lim_{n\to\infty}P_n(B_i^m)$$

$$=\sum_{i=1}^{k}\alpha_i^m P(B_i^m)=\int s_m dP<\infty,$$

luego se tiene que

$$\lim_{n\to\infty}\int f\ dP_n=\lim_{n\to\infty}\int\lim_{m\to\infty}s_m\ dP_n=\lim_{n\to\infty}\lim_{m\to\infty}\int s_m\ dP_n$$

$$= \lim_{n\to\infty}\lim_{m\to\infty}\sum_{i=1}^{k}\alpha_i^m P_n(B_i^m) = \lim_{m\to\infty}\sum_{i=1}^{k}\alpha_i^m \lim_{n\to\infty} P_n(B_i^m) = \lim_{m\to\infty}\sum_{i=1}^{k}\alpha_i^m P(B_i^m)$$

$$\lim_{m\to\infty}\int s_m \ dP = \int \lim_{m\to\infty} s_m \ dP = \int f \ dP.$$

Lema 4.4.6. *(Integración respecto a combinación lineal positiva). Sean $(S,\mathcal{B})$ un espacio medible, ν y μ dos medidas finitas sobre S, y $\alpha,\beta \in [0,\infty)$; como las medidas con signo finitas son un espacio vectorial se tiene que $\alpha\nu+\beta\mu$ es una medida con signo finita (positiva). Entonces sucede que $f \in L_1(\nu)\cap \in L_1(\mu)$ si y sólo si $f \in L_1(\alpha\nu+\beta\mu)$ y se tiene que*

$$\int f \ d(\alpha\nu+\beta\mu) = \alpha\int f \ d\nu + \beta\int f \ d\mu.$$

Prueba. Primero recordemos la definición de la integral de Lebesgue de una función medible g respecto a una medida ρ

$$\int g \ d\rho = \int g^+ d\rho - \int g^- d\rho.$$

En donde

$$\int g^+ d\rho = \sup\left\{\int h \ d\rho \Big| h \in \mathcal{S}_-(g^+)\right\},$$

$$\int g^- d\rho = \sup\left\{\int h \ d\rho \Big| h \in \mathcal{S}_-(g^-)\right\}.$$

Con $\mathcal{S}_-(g^+) = \{h|h$ es medible y simple, $h \geq 0$ y $h \leq g^+\}$ y $\mathcal{S}_-(g^-) = \{h|h$ es medible y simple, $h \geq 0$ y $h \leq g^-\}$. Además si h es simple y no negativa se tiene que $\int h \ d\rho = \sum_{i=1}^n \lambda_i\rho(B_i)$. Entonces

$$\int g^+ d\rho = \sup\left\{\sum_{i=1}^n \lambda_i\rho(B_i) \Big| \sum_{i=1}^n \lambda_i\rho(B_i) = h \in \mathcal{S}_-(g^+)\right\}.$$

Gracias a la definición tenemos que

$$\int g^+ d(\alpha\nu+\beta\mu)$$

$$= \sup\left\{\sum_{i=1}^n \lambda_i(\alpha\nu+\beta\mu)(B_i) \Big| \sum_{i=1}^n \lambda_i(\alpha\nu+\beta\mu)(B_i) = h \in \mathcal{S}_-(g^+)\right\}$$

$$= \sup\left\{\sum_{i=1}^{n}\lambda_i[\alpha\nu(B_i)+\beta\mu(B_i)]\Big|\sum_{i=1}^{n}\lambda_i[\alpha\nu(B_i)+\beta\mu(B_i)] = h \in \mathcal{S}_-(g^+)\right\}$$

$$= \alpha\sup\left\{\sum_{i=1}^{n}\lambda_i\nu(B_i)\Big|h \in \mathcal{S}_-(g^+)\right\} + \beta\sup\left\{\sum_{i=1}^{n}\lambda_i\mu(B_i)\Big|h \in \mathcal{S}_-(g^+)\right\}$$

$$= \alpha\int g^+ d\nu + \beta\int g^+ d\mu.$$

Análogamente se tiene que

$$\int g^- d(\alpha\nu+\beta\mu) = \alpha\int g^- d\nu + \beta\int g^- d\mu.$$

Por lo anterior $f \in L_1(\nu)\cap \in L_1(\mu)$ si y sólo si $f \in L_1(\alpha\nu+\beta\mu)$. Además se tiene que

$$\int f\ d(\alpha\nu+\beta\mu) = \int f^+ d(\alpha\nu+\beta\mu) - \int f^- d(\alpha\nu+\beta\mu)$$

$$= \alpha\int f^+ d\nu + \beta\int f^+ d\mu - \left[\alpha\int f^- d\nu + \beta\int f^- d\mu\right]$$

$$= \alpha\int f^+ - f^- d\nu + \beta\int f^+ - f^- d\mu = \alpha\int f\ d\nu + \beta\int f\ d\mu.$$

□

En particular, se tiene que si $\lambda \in [0,1]$ entonces

$$\int f\ d(\lambda\nu+(1-\lambda)\mu) = \lambda\int f\ d\nu + (1-\lambda)\int f\ d\mu.$$

Lema 4.4.7. *(Derivación bajo el signo de integral). Sean $(S,\mathcal{B})$ un espacio medible y $X \subset \mathbb{R}$ un subconjunto abierto. Sea $f : X \times S \to \mathbb{R}$ una función que satisface lo siguiente:*

1. $f(t,s)$ es Lebesgue integrable respecto a s para cada $t \in X$.

2. f es derivable respecto a t para casi todo $s \in S$ y para toda $t \in X$

3. Existe una función integrable $g : S \to \mathbb{R}$ tal que $|\frac{\partial}{\partial t}f(t,s)| \leq g(s)$ para casi todo $s \in S$, para toda $t \in X$.

Entonces

$$\frac{d}{dt}\int f(t,s)\ ds = \int \frac{\partial}{\partial t}f(t,s)\ ds.$$

Prueba. Primero observemos que la definición de derivada está en términos de un límite, es decir:

$$\frac{d}{dt}\int f(t,s)\ ds = \lim_{h\to 0}\frac{\int f(t+h,s)\ ds - \int f(t,s)\ ds}{h}.$$

También

$$\frac{\partial}{\partial t}f(t,s) = \lim_{h\to 0}\frac{f(t+h,s)-f(t,s)}{h},$$

luego gracias al teorema CDL[6] y la linealidad de la integral tenemos que

$$\int \frac{\partial}{\partial t}f(t,s)\ ds = \int \lim_{h\to 0}\frac{f(t+h,s)-f(t,s)}{h}\ ds$$

$$= \lim_{h\to 0}\int \frac{f(t+h,s)-f(t,s)}{h}\lim_{h\to 0}\frac{\int f(t+h,s)\ ds - \int f(t,s)\ ds}{h}$$

$$= \frac{d}{dt}\int f(t,s)\ ds.$$

□

Definición 4.4.8. *Sean $(S,\mathcal{B},\mu)$ un espacio de medida y $\{f_n\}_{n\in\mathbb{N}}$ una sucesión de funciones reales y medibles. Se dice que $\{f_n\}_{n\in\mathbb{N}}$ es de Cauchy en medida si para todo $\epsilon > 0$ se tiene que*

$$\lim_{n,m\to\infty}\mu(\{s\in S\big||f_m(s)-f-n(s)|\geq\epsilon\}) = 0.$$

Teorema 4.4.7. *(F. Riesz - H. Weyl). Sean $(S,\mathcal{B},\mu)$ un espacio de medida y $\{f_n\}_{n\in\mathbb{N}}$ una sucesión de funciones medibles que es Cauchy en medida. Entonces existe f una función medible y una subsucesión $\{f_{n_k}\}_{k\in\mathbb{N}}$ de $\{f_n\}_{n\in\mathbb{N}}$ tal que:*

(i) f_{n_k} converge a f casi donde sea relativo a μ.

(ii) f_n converge a f en medida[7].

Demostración. [18], pág. 110. □

Observación. *Si $\{f_n\}_{n\in\mathbb{N}}$ es tal que f_n converge a f en medida entonces f_n es de Cauchy en medida.*

[6] cf. Teorema 4.4.4.

[7] cf. ii) de la Definición 4.1.1.

Teorema 4.4.8. *(Radon-Nikodým). Sean $(S, \mathcal{B})$, un espacio medible, $\mu : \mathcal{B} \to \mathbb{R} \cup \{\pm\infty\}$, una medida σ-finita, y $\nu : \mathcal{B} \to \mathbb{R} \cup \{\pm\infty\}$, una medida con signo σ-finita tal que $\nu \ll \mu$. Entonces existe $f : S \to \mathbb{R}$, función medible, tal que*

$$\nu(E) = \int_E f \ d\mu \quad \forall E \in \mathcal{B}.$$

Si existe $\widetilde{f} : S \to \mathbb{R}$ medible tal que

$$\nu(E) = \int_E \widetilde{f} \ d\mu \quad \forall E \in \mathcal{B}$$

entonces $f = \widetilde{f}$ casi donde sea relativo a μ.

Demostración. [18], pág. 132. □

Definición 4.4.9. *Sean $(S, \mathcal{B})$ un espacio medible y μ, ν con las mismas hipotesis del Teorema 4.4.8. La única función f que satisface*

$$\nu(E) = \int_E f \ d\mu \quad \forall E \in \mathcal{B}$$

se le conoce como derivada de Radon-Nikodým de ν respecto a μ y se denota, por obvias analogías, de la siguiente manera:

$$f = \frac{d\nu}{d\mu} \qquad [\mu].$$

En donde recordamos que $[\mu]$ significa casi donde sea relativo a μ.

A continuación enunciamos algunas propiedades de esta función.

Lema 4.4.8. *Sean $(S, \mathcal{B})$ un espacio medible, ν, ν_1, ν_2, ρ y $\mu : S \to \mathbb{R} \cup \{\pm\infty\}$ medidas σ-finitas.*

1. Si $\nu_1 \ll \mu$ y $\nu_2 \ll \mu$ entonces

$$\nu_1 \pm \nu_2 \ll \mu \ \ y \ \ \frac{d(\nu_1 \pm \nu_2)}{d\mu} = \frac{d\nu_1}{d\mu} \pm \frac{d\nu_2}{d\mu} \qquad [\mu].$$

2. Si $\nu \ll \mu$ y $\lambda \in \mathbb{R}$ $\lambda \neq 0$ entonces

$$\lambda\nu \ll \mu \ \ y \ \ \frac{d(\lambda\nu)}{d\mu} = \lambda\frac{d\nu}{d\mu} \qquad [\mu].$$

3. Si $\rho \ll \nu$ y $\nu \ll \mu$ entonces

$$\rho \ll \mu \ \ y \ \ \frac{d\rho}{d\mu} = \left(\frac{d\rho}{d\nu}\right)\left(\frac{d\nu}{d\mu}\right) \qquad [\mu].$$

4. Si $\nu \equiv \mu$, i.e., $\nu \ll \mu$ y $\mu \ll \nu$ entonces

$$\nu\left(\left\{s \in S \middle| \left(\frac{d\nu}{d\mu}\right)(s) = 0\right\}\right) = 0 \ \ y \ \ \frac{d\mu}{d\nu} = \left(\frac{d\nu}{d\mu}\right)^{-1} \qquad [\mu].$$

Prueba. [18], pág. 136. □

Lema 4.4.9. *Sean $(S, \mathcal{B})$ un espacio medible, ν,μ medidas σ-finitas tales que $\nu \ll \mu$, $f \in L_1(\nu)$ y $f\frac{d\nu}{d\mu} \in L_1(\mu)$. Entonces*

$$\int f\left(\frac{d\nu}{d\mu}\right)d\mu = \int f \ d\nu.$$

Prueba. (i) Supongamos primero que $f\frac{d\nu}{d\mu} \in L_1(\mu)$ es tal que $f\frac{d\nu}{d\mu} \geq 0$ luego tenemos que $f \geq 0$ y existe $\{f_n\}_{n\in\mathbb{N}}$ una sucesión de funciones simples tales que $0 \leq f_n \leq f_{n+1} \leq f$ y $\sup_{n\in\mathbb{N}} f_n = f$ y por supuesto tenemos que $0 \leq f_n\frac{d\nu}{d\mu} \leq f_{n+1}\frac{d\nu}{d\mu} \leq f\frac{d\nu}{d\mu}$ y $\sup_{n\in\mathbb{N}} \frac{d\nu}{d\mu}f_n = \frac{d\nu}{d\mu}f$, luego por el teorema de convergencia monótona[8] tenemos que

$$\int f \ d\nu = \sup_{n\in\mathbb{N}} \int f_n \ d\nu \ \text{ y } \ \int f\left(\frac{d\nu}{d\mu}\right)d\mu = \sup_{n\in\mathbb{N}} \int f_n\left(\frac{d\nu}{d\mu}\right)d\mu.$$

Por otro lado, como $f_n = \sum_{i=1}^{n} \lambda_i 1_{B_i}$ para alguna $n \in \mathbb{N}$ con $\lambda_i \in \mathbb{R}$, $B_i \in \mathcal{B}$, $i = 1, ..., n$ tenemos que

$$\int f_n \ d\nu = \sum_{i=1}^{n} \lambda_i \int 1_{B_i} \ d\nu = \sum_{i=1}^{n} \lambda_i \nu(B_i) = \sum_{i=1}^{n} \lambda_i \int 1_{B_i}\left(\frac{d\nu}{d\mu}\right)d\mu$$

$$= \int f_n\left(\frac{d\nu}{d\mu}\right)d\mu,$$

luego

$$\int f \ d\nu = \sup_{n\in\mathbb{N}} \int f_n \ d\nu = \sup_{n\in\mathbb{N}} \int f_n\left(\frac{d\nu}{d\mu}\right)d\mu = \int f\left(\frac{d\nu}{d\mu}\right)d\mu.$$

(ii) Ahora, si f es cualquier función arbitraria en $L_1(\nu)$ entonces consideremos $f = f^+ - f^-$, además como ν es una medida con signo podemos dar una descomposición de Jordan $\nu = \nu^+ - \nu^-$ con ν^+ y ν^- medidas sobre S, luego

$$\int f\left(\frac{d\nu}{d\mu}\right)d\mu = \int f^+ - f^-\left(\frac{d\nu}{d\mu}\right)d\mu = \int f^+\left(\frac{d\nu}{d\mu}\right)d\mu - \int f^-\left(\frac{d\nu}{d\mu}\right)d\mu.$$

[8] cf. Teorema 4.4.2.

Observamos lo siguiente

$$\int f^{\pm}\Big(\frac{d\nu}{d\mu}\Big)d\mu = \int f^{\pm}\left[\frac{d(\nu^+ - \nu^-)}{d\mu}\right]d\mu = \int f^{\pm}\left[\frac{d\nu^+}{d\mu} - \frac{d\nu^-}{d\mu}\right]d\mu$$

$$= \int f^{\pm}\left(\frac{d\nu^+}{d\mu}\right)d\mu - \int f^{\pm}\left(\frac{d\nu^-}{d\mu}\right)d\mu = \int f^{\pm}d\nu^+ - \int f^{\pm}d\nu^-.$$

La última igualdad por (i). Por lo tanto tenemos que

$$\int f\Big(\frac{d\nu}{d\mu}\Big)d\mu = \int f^+d\nu^+ - \int f^+d\nu^- - \left[\int f^-d\nu^+ - \int f^-d\nu^-\right]$$

$$\int f^+ - f^-d\nu^+ - \int f^+ - f^-d\nu^- = \int f^+ - f^-d(\nu^+ - \nu^-) = \int f\ d\nu.$$

□

Durante el desarrollo de los capítulos precedentes especialmente dentro del capítulo 3 hicimos uso de algunos resultados que involucran a la medida de producto, es por ello que hacemos un breve recodatorio de dichos conceptos ahora. Para un desarrollo más profundo referimos al lector a [21], [4] y [18].

Definición 4.4.10. *Sean $(X, \mathcal{A})$ y $(Y, \mathcal{B})$ dos espacios medibles. Se define el espacio medible producto denotado $(X \times Y, \mathcal{A} \otimes \mathcal{B})$ en donde $\mathcal{A} \otimes \mathcal{B}$ es la σ-álgebra de subconjuntos de $X \times Y$ generada por $\mathcal{A} \times \mathcal{B}$.*

Definición 4.4.11. *Sea $B \subset X \times Y$. Para $x \in X$ se define la x-sección de B denotado por B_x como el subconjunto de Y definido por*

$$B_x = \{y \in Y | (x, y) \in B\}.$$

Análogamente se define la y-sección de B denotado por B^y.

Teorema 4.4.9. *(La medida producto). Sean $(X, \mathcal{A}_1, \nu)$ y $(Y, \mathcal{A}_2, \mu)$ dos espacios de medida σ-finita. Entonces la función conjuntista $\nu \otimes \mu : \mathcal{A}_1 \otimes \mathcal{A}_2 \to \mathbb{R}$ definida por*

$$(\nu \otimes \mu)(B) = \int \nu(B_x)\ d\mu = \int \nu(B^y)\ d\mu$$

es una medida σ-finita y es la única medida sobre $\mathcal{A}_1 \otimes \mathcal{A}_2$ con la siguiente propiedad:

$$(\nu \otimes \mu)(A \times B) = \nu(A)\mu(B) \quad \forall A \times B \in \mathcal{A}_1 \times \mathcal{A}_2.$$

Demostración. [18], pág. 151. □

Teorema 4.4.10. *(Tonelli-Fubini). Sean $(X, \mathcal{A}_1, \nu)$ y $(Y, \mathcal{A}_2, \mu)$ espacios de medida σ-finitos. Si h es una función integrable en $X \otimes Y$ entonces casi toda sección de h es integrable. Si definimos*

$$f(x) = \int h(x,y)\ d\nu(y) \quad g(y) = \int h(x,y)\ d\mu(x)$$

se tiene que f y g son integrables y

$$\int h\ d(\nu \otimes \mu) = \int f\ d\mu = \int g\ d\nu.$$

Demostración. [21], pág. 148. □

Lema 4.4.10. *Sean ν_i, μ_i medidas σ-finitas sobre un espacio medible $(X, \mathcal{A})$ con $\nu_i \ll \mu_i$ $i = 1, 2$. Sean $\nu = \nu_1 \otimes \nu_2$ y $\mu = \mu_1 \otimes \mu_2$ las correspondientes medidas producto sobre $(X \times X, \mathcal{A} \otimes \mathcal{A})$ entonces $\nu \ll \mu$ y*

$$\frac{d\nu}{d\mu}(x,y) = \frac{d\nu_1}{d\mu_1}(x)\frac{d\nu_2}{d\mu_2}(y) \qquad [\mu].$$

Prueba. Primero tomemos $B \in \mathcal{A} \otimes \mathcal{A}$, $x \in X$ y $B_x = \{y \in X | (x,y) \in B\}$. Se tiene que

$$\nu(B) = \int_X \nu_2(B_x)\ d\nu_1(x) \text{ y } \mu(B) = \int_X \mu_2(B_x)\ d\mu_1(x).$$

Supongamos que $\mu(B) = 0$, entonces $\mu_2(B_x) = 0$ $[\mu_1]$ luego, como $\nu_2 \ll \mu_2$, se sigue que $\nu_2(B_x) = 0$ $[\mu_1]$ y como $\nu_1 \ll \mu_1$ se tiene que $\nu_2(B_x) = 0$ $[\nu_1]$. Así,

$$\nu(B) = \int_X \nu_2(B_x)\ d\nu_1(x) = 0,$$

por lo que efectivamente $\nu \ll \mu$. Ahora, tomemos de nuevo $B \in \mathcal{A} \otimes \mathcal{A}$ y observemos lo siguiente:

$$\begin{aligned}
\nu(B) &= \int_{X\times X} 1_B(x,y)\ d\nu = \int_X \int_X 1_{B_x}(y)\ d\nu_2(y) d\nu_1(x) \\
&= \int_X \left[\int_X 1_{B_x}(y) \frac{d\nu_2}{d\mu_2}(y)\ d\mu_2(y) \right] d\nu_1(x) \\
&= \int_X \left[\int_X 1_{B_x}(y) \frac{d\nu_2}{d\mu_2}(y)\ d\mu_2(y) \right] \frac{d\nu_1}{d\mu_1}(x)\ d\mu_1(x) \\
&= \int_X \int_X 1_{B_x}(y) \frac{d\nu_2}{d\mu_2}(y) \frac{d\nu_1}{d\mu_1}(x)\ d\mu_2(y) d\mu_1(x)
\end{aligned}$$

$$= \int_{X\times X} 1_B(x,y)\frac{d\nu_2}{d\mu_2}(y)\frac{d\nu_1}{d\mu_1}(x)\ d\mu(x,y) = \int_B \frac{d\nu_2}{d\mu_2}(y)\frac{d\nu_1}{d\mu_1}(x)\ d\mu(x,y).$$

Como $\nu \ll \mu$ entonces por el teorema de Radon-Nikodým se tiene que

$$\nu(B) = \int_{X\times X} \frac{d\nu}{d\mu}\ d\mu,$$

gracias a la unicidad de la derivada de Radon-Nikodým se sigue que

$$\frac{d\nu}{d\mu}(x,y) = \frac{d\nu_1}{d\mu_1}(x)\frac{d\nu_2}{d\mu_2}(y) \qquad [\mu].$$

□

Todo lo anterior se generaliza para n factores, es decir, para el producto cartesiano de n espacios medibles. Véase la sección 37 de [21].

Teorema 4.4.11. *(Cambio de variable respecto al pushforward). Sean $(S_2, \mathcal{B}_2)$ un espacio medible, $(S_1, \mathcal{B}_1, \mu)$ un espacio de medida y $\Psi : S_1 \to S_2$ una función medible. Consideremos el pushforward de μ bajo Ψ, i.e., $\widetilde{\mu} = \mu \circ \Psi^{-1}$. Entonces $f \in L_1(S_2, \mathcal{B}_2, \widetilde{\mu})$ si y sólo si $f \circ \Psi \in L_1(S_1, \mathcal{B}_1, \mu)$ y en este caso sucede que*

$$\int_{S_2} f\ d\widetilde{\mu} = \int_{S_1} f \circ \Psi\ d\mu.$$

Demostración. [4], pág. 190. □

Lema 4.4.11. *Sea $(S, \mathcal{B})$ un espacio medible entonces $P \in \Lambda_f(S) = \Lambda_f$ si y sólo si para cada $B \in \mathcal{B}$ se tiene que*

$$P(B) = \sum_{i=1}^{k} \alpha_i 1_B(s_i) \quad s_i \in S,\ \ \alpha_i > 0,\ \ i = 1, ..., k \quad y \quad \sum_{i=1}^{k} \alpha_i = 1.$$

Demostración. ⇐] Supongamos que para cada $B \in \mathcal{B}$

$$P(B) = \sum_{i=1}^{k} \alpha_i 1_B(s_i) \quad s_i \in S, \quad \alpha_i > 0, \quad i = 1, ..., k \quad y \quad \sum_{i=1}^{k} \alpha_i = 1.$$

Entonces P tiene un número finito de atómos que son $B_1, ..., B_k$ tales que $P(B_i) = \alpha_i$.

⇒] Supongamos que $P \in \Lambda_f$. Sea A el conjunto de átomos de P, luego A es finito, dígamos $A = \{B_1, ..., B_k\}$. Sea $B \in \mathcal{B}$ y T el conjunto de átomos que son subconjunto de B, se tiene entonces que (cf. [2])

$$P(B) = \sum_{A \in T} P(A).$$

Como $T \subset A$ entonces $T = \{B_{i_j}\}_{j=1}^{n}$ con $n \leq k$, tomemos s_i tal que $s_i \in B_i$ y $s_i \neq s_j$ con $i \neq j$, $i = 1, ..., k$. Luego, es claro que si definimos $\alpha_i = P(B_i)$, $i = 1, ..., k$ se tiene que

$$P(B) = \sum_{j=1}^{n} P(B_{i_j}) = \sum_{j=1}^{n} \alpha_{i_j} 1_B(s_{i_j}) = \sum_{i=1}^{k} \alpha_i 1_B(s_i).$$

Además

$$\sum_{i=1}^{k} \alpha_i = 1.$$

□

Lema 4.4.12. *Sea $(S, \mathcal{B})$ un espacio medible. Entonces $\Lambda_f(S) = \Lambda_f$ el espacio de medidas de probabilidad atómicas con un número finito de átomos es convexo.*

Prueba. Sean $P, Q \in \Lambda_f$ entonces por el Lema 4.4.11 para toda $B \in \mathcal{B}$

$$P(B) = \sum_{i=1}^{n} \alpha_i 1_B(s_i) \quad \text{y} \quad Q(B) = \sum_{i=1}^{m} \alpha'_i 1_B(s'_i)$$

con $\alpha_i > 0$, $\alpha'_i > 0$ para toda i $\sum_{i=1}^{n} \alpha_i = 1$ y $\sum_{i=1}^{m} \alpha'_i = 1$. Sea $\lambda \in [0, 1]$.

Caso 1: Todos los átomos son distintos, i.e., $s_i \neq s'_j$ para toda i, j. Definimos $r_1 = s_1, ..., r_n = s_n, r_{n+1} = s'_i, ..., r_{n+m} = s'_i$ y $\beta_1 = \lambda\alpha_1, ..., \beta_n = \lambda\alpha_n, \beta_{n+1} = (1-\lambda)\alpha'_1, ..., \beta_{n+m} = (1-\lambda)\alpha'_m$. Claramente $\beta_i > 0$ para toda i. Aademás

$$\sum_{i=1}^{n+m} \beta_i = \sum_{i=1}^{n} \lambda\alpha_i + \sum_{i=1}^{m} (1-\lambda)\alpha'_i = \lambda + (1-\lambda) = 1;$$

así, para toda $B \in \mathcal{B}$

$$\lambda P(B) + (1-\lambda)Q(B) = \sum_{i=1}^{n} \lambda\alpha_i 1_B(s_i) + \sum_{i=1}^{m} (1-\lambda)\alpha'_i 1_B(s'_i)$$

$$\sum_{i=1}^{n} \beta_i 1_B(r_i) + \sum_{i=n+1}^{m} \beta_i 1_B(r_i) = \sum_{i=1}^{n+m} \beta_i 1_B(r_i),$$

es decir, $\lambda P + (1-\lambda)Q \in \Lambda_f$.

Caso 2: Todos los átomos son iguales. Podemos suponer sin pérdida de generalidad que $s_i = s'_i$ para toda i. En este caso definimos $r_i = s_i$ para toda i y $\beta_i = \lambda\alpha_i + (1-\lambda)'_i$ para toda i. Claramente $\beta_i > 0$ para toda i y

$$\sum_{i=1}^{n} \beta_i = \sum_{i=1}^{n} \lambda\alpha_i + \sum_{i=1}^{n}(1-\lambda)\alpha'_i = \lambda + (1-\lambda) = 1.$$

De esta forma, para toda $B \in \mathcal{B}$

$$\lambda P(B) + (1-\lambda)Q(B) = \sum_{i=1}^{n} \lambda\alpha_i 1_B(s_i) + \sum_{i=1}^{n}(1-\lambda)\alpha'_i 1_B(s_i)$$

$$\sum_{i=1}^{n}[\lambda\alpha_i + (1-\lambda)\alpha'_i]1_B(r_i) = \sum_{i=1}^{n} \beta_i 1_B(r_i),$$

es decir, $\lambda P + (1-\lambda)Q \in \Lambda_f$.

Caso 3: $s_i = s'_j$ para algunas $i = 1, ..., n$, $j = 1, ..., m$ pero no todas. Supongamos sin pérdida de generalidad que $n < m$ y $s_1 = s'_1, ..., s_k = s'_k$ para algún $0 < k < n$. En este caso definimos

$$r_1 = s_1, ..., r_k = s_k, r_{k+1} = s_{k+1}, ..., r_n = s_n, r_{n+1} = s'_{k+1}, ..., r_{n+m-k} = s'_m$$

y

$$\beta_1 = \lambda\alpha_1 + (1-\lambda)\alpha'_1, ..., \beta_k = \lambda\alpha_k + (1-\lambda)\alpha'_k, \beta_{k+1} = \lambda\alpha_{k+1}, ..., \beta_n = \lambda\alpha_n,$$

$$\beta_{n+1} = (1-\lambda)\alpha'_{k+1}, ..., \beta_{n+m-k} = (1-\lambda)\alpha'_m.$$

Nuevamente $\beta_i > 0$ para toda i y

$$\sum_{i=1}^{n+m-k} \beta_i = \sum_{i=1}^{k} \beta_i + \sum_{i=k+1}^{n} \beta_i + \sum_{i=n+1}^{n+m-k} \beta_i$$

$$= \sum_{i=1}^{k}[\lambda\alpha_i + (1-\lambda)\alpha'_i] + \sum_{i=k+1}^{n} \lambda\alpha_i + \sum_{i=k+1}^{m} (1-\lambda)\alpha'_i$$

$$= \sum_{i=1}^{n} \lambda\alpha_i + \sum_{i=1}^{m}(1-\lambda)\alpha'_i = \lambda + (1-\lambda) = 1.$$

Así,

$$\lambda P(B) + (1-\lambda)Q(B) = \sum_{i=1}^{n} \lambda\alpha_i 1_B(s_i) + \sum_{i=1}^{m}(1-\lambda)\alpha'_i 1_B(s'_i)$$

$$= \sum_{i=1}^{k}[\lambda\alpha + (1-\lambda)\alpha'_i]1_B(s_i) + \sum_{i=k+1}^{n} \lambda\alpha_i 1_B(s_i) + \sum_{i=k+1}^{m} (1-\lambda)\alpha'_i 1_B(s'_i)$$

$$= \sum_{i=1}^{k} \beta_i 1_B(r_i) + \sum_{i=k+1}^{n} \beta_i 1_B(r_i) + \sum_{i=n+1}^{n+m-k} \beta_i 1_B(r_i) = \sum_{i=1}^{n+m-k} \beta_i 1_B(r_i),$$

es decir, $\lambda P + (1-\lambda)Q \in \Lambda_f$.

□

Lema 4.4.13. *Sean $(S, \mathcal{B})$ un espacio medible, $(\Omega, \mathcal{A}, \mu)$ un espacio de probabilidad arbitrario y $\nu : \Omega \times \mathcal{B} \to [0,1]$ un kernel de Markov. Sea $f : S \to \mathbb{R}$ una función medible y acotada, usando la notación $\nu_\omega(\omega, \cdot)$ se tiene que*

$$\int_\Omega \nu(\omega, B) \ d\mu(\omega) = \int_\Omega \left(\int_S f(s) \ d\nu_\omega(s) \right) \ d\mu(\omega).$$

Prueba. La prueba es sencilla: primero se prueba la igualdad para funciones características, después para funciones simples y por último para cualquier función medible. Sea $B \in \mathcal{B}$, por un lado se tiene que

$$\int_S 1_B(s) \ d\mu\nu(s) = \mu\nu(B) = \int_\Omega \nu(\omega, B) \ d\mu(\omega).$$

Por otro lado,

$$\int_\Omega \left(\int_S 1_B(s) \ d\nu_\omega(s) \right) \ d\mu(\omega) = \int_\Omega \nu(\omega, B) \ d\mu(\omega).$$

Ahora, sea g una función simple, i.e., $g(s) = \sum_{i=1}^n \alpha_i 1_{B_i}(s)$ con $B_i \in \mathcal{B}$ y $\alpha_i \in \mathbb{R}$.

$$\int_S g(s) \ d\mu\nu(s) = \int_S \sum_{i=1}^n \alpha_i 1_{B_i}(s) \ d\mu\nu(s) = \sum_{i=1}^n \alpha_i \int_S 1_{B_i}(s) \ d\mu\nu(s)$$

$$= \sum_{i=1}^n \alpha_i \int_\Omega \nu(\omega, B_i) \ d\mu(\omega).$$

Por otra parte,

$$\int_\Omega \left(\int_S g(s) \ d\nu_\omega(s) \right) \ d\mu(\omega) = \int_\Omega \left(\int_S \sum_{i=1}^n \alpha_i 1_{B_i}(s) \ d\nu_\omega(s) \right) \ d\mu(\omega)$$

$$= \int_\Omega \left(\sum_{i=1}^n \alpha_i \int_S 1_{B_i}(s) \ d\nu_\omega(s) \right) \ d\mu(\omega) = \int_\Omega \sum_{i=1}^n \alpha_i \nu(\omega, B_i) \ d\mu(\omega)$$

$$= \sum_{i=1}^{n} \alpha_i \int_{\Omega} \nu(\omega, B_i) \ d\mu(\omega).$$

Por último, sea f una función medible y acotada, se tiene que existe una sucesión de funciones simples $\{g_n\}_{n\in\mathbb{N}}$ que converge de manera uniforme a f (cf. Grabinsky Lema 2.6). Si definimos $G_n(\omega) = \int_S g_n(s) \ d\nu_\omega(s)$ entonces G_n converge a $\int_S f(s) \ d\nu_\omega(s)$ de manera uniforme. En efecto, como g_n converge a f de manera uniforme entonces para toda $\epsilon > 0$ existe $N \in \mathbb{N}$ tal que si $n > N$ y para toda $s \in S$ se tiene que $|g_n(s) - f(s)| < \epsilon$. Sea $n > N$ y sea $\omega \in \Omega$ entonces

$$\left| G_n(\omega) - \int_S f(s) \ d\nu_\omega(s) \right| = \left| \int_S g_n(s) \ d\nu_\omega(s) - \int_S f(s) \ d\nu_\omega(s) \right|$$

$$= \left| \int_S g_n(s) - f(s) \ d\nu_\omega(s) \right| \leq \int_S |g_n(s) - f(s)| \ d\nu_\omega(s) = \int_S \epsilon \ d\nu_\omega(s) = \epsilon.$$

Así,

$$\int_S f(s) \ d\mu\nu(s) = \int_S \lim_{n\to\infty} g_n(s) \ d\mu\nu(s) = \lim_{n\to\infty} \int_S g_n(s) \ d\mu\nu(s)$$

$$= \lim_{n\to\infty} \int_{\Omega} \left(\int_S g_n(s) \ d\nu_\omega(s) \right) \ d\mu(\omega) = \int_{\Omega} \left(\int_S f(s) \ d\nu_\omega(s) \right) \ d\mu(\omega).$$

□

Bibliografía

[1] Alexandroff, A. Additive set-functions in abstract spaces. *Rec. Math. [Mat. Sbornik] N.S. 13*, 2-3 (1943), 169–238.

[2] Azram, M., Elfaki Faiz, A., and J.I., D. Classification of atoms. *Australian Journal of Basic and Applied Sciences 5*, 5 (2011), 5–8.

[3] Billingsley, P. *Convergence in Probability Measures*, 2nd ed. John Wiley & Sons, 1999.

[4] Bogachev, V. *Measure Theory*, vol. 2. Springer-Verlag, 2007.

[5] Bourbaki, N. *Topological Vector Spaces*. Springer-Verlag, 1987.

[6] Choquet, G. *Lectures on Analysis*, vol. 1. Benjamin, 1969.

[7] Choquet, G. *Lectures on Analysis*, vol. 2. Benjamin, 1969.

[8] Conway, J. *A Course in Functional Analysis*, 2nd ed. Springer-Verlag, 1990.

[9] Cover, Thomas, M., and Thomas, Joy, A. *Elements of Information Theory*, 2nd ed. John Wiley & Sons, 2006.

[10] Csiszár, I. I-divergence geometry of probability and minimization problems. *Ann. Probab. 3*, 1 (1975), 146–158.

[11] Csiszár, I. Sanov property, generalized I-projection and a conditional limit theorem. *Ann. Probab. 12*, 3 (1984), 568–593.

[12] Csiszár, I. A simple proof of Sanov's theorem. *Bull. Braz Math Soc, New Series 37* (2006), 453–459.

[13] Csiszár, I., and Shields, P. Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory. 1* (2004), 417–528.

[14] Dembo, A., and Zeitouni, O. *Large Deviations Techniques and Applications*, 2nd ed. Springer-Verlag, 1998.

[15] Donsker, M., and Varadhan, S. Asymptotic evaluation of certain markov process expectations for large time, iv. *Communications on Pure and Applied Mathematics 36* (1983), 183–212.

[16] Dudley, R. *Real Analysis and Probability.* Cambridge University Press, 2002.

[17] Folland, G. *Real Analysis*, 2nd ed. John Wiley & Sons, 1999.

[18] Grabinsky, G. *Teoría de la Medida.* Facultad de Ciencas, UNAM, 2016.

[19] Gray, R. M. *Entropy and Information Theory*, 2nd ed. Springer, 2011.

[20] Groeneboom, P., Oosterhoff, J., and F.H., R. Large deviation theorems for empirical probability measures. *Ann. Probab. 7*, 4 (1979), 553–586.

[21] Halmos, Paul, M. *Measure Theory*, vol. 2. Springer-Verlag, 1974.

[22] Hollander, F. D. *Large Deviations.* Fields Institute monographs. American Mathematical Society, 2000.

[23] Hong, L. The linear topology associated with weak convergence of probability measures. *Missouri J. Math. Sci. 26*, 2 (2014), 168–172.

[24] Kallianpur, G. The topology of weak convergence of probability measures. *Journal of Mathematics and Mechanics 10*, 6 (1961), 947–969.

[25] Klenke, A. *Probability Theory. A comprehensive course*, 2nd ed. Springer-Verlag, 2008.

[26] Merkle, M. Topics in weak convergence of probability measures. *Zbornik Radova*, 17 (2000), 235–274.

[27] Munkres, James, R. *Topology*, 2nd ed. Prentice-Hall, 2000.

[28] Phelps, R. *Lectures on Choquet's theorem.* Springer, 1966.

[29] Rockafellar, Tyrrell, R. *Convex Analysis.* Princeton University Press, 1972.

[30] Rudin, W. *Real and Complex Analysis*, 2nd ed. McGraw-Hill, 1978.

[31] Rudin, W. *Functional Analysis*, 2nd ed. McGraw-Hill, 1991.

[32] Topsoe, F. Information theoretical optimization techniques. *Kybernetika 15*, 1 (1979), 8–27.

[33] Vaart, A. v. d. *Asymtotic Statistics*, first ed. Cambridge University Press, 1998.

[34] Zălinescu, C. *Convex Analysis in General Vector Spaces*. World Scientific, 2002.