UNIVERSIDAD NACIONAL DE COLOMBIA

Retardo Cosmológico Temporal en Teorías de Gravedad Modificada $f(R)$

Julián Orlando Jiménez Cárdenas

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Física
Bogotá D. C. , Colombia
2019

Retardo Cosmológico Temporal en Teorías de Gravedad Modificada $f(R)$

Julián Orlando Jiménez Cárdenas

Tesis o trabajo de grado presentada(o) como requisito parcial para optar al título de:
Físico

Director(a):
Ph.D. Leonardo Castañeda Colorado

Línea de Investigación:
Astrofísica, Gravitación y Cosmologia
Grupo de Investigación:
Grupo de Galaxias, Gravitación y Cosmologia

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Física
Bogotá D. C. ,Colombia
2019

Resumen

En la primera parte de este texto se presenta una introducción a la geometría diferencial como herramienta matemática para la Relatividad General. Se estudia la gravedad linealizada y el papel que esta desempeña en la radiación gravitacional, profundizando así en los conceptos de gauge, energía y contribución cuadrupolar. Seguidamente se presentan las ecuaciones de Einstein relajadas como una generalización para el estudio de la radiación gravitacional y se obtienen expresiones generales para la energía, el momentum lineal y angular. Posteriormente se muestra la relación entre las expresiones de flujo de energía, momentum lineal y angular con el tensor de Weyl.

Palabras clave: Radiación Gravitacional, Gravedad Linealizada, Ecuaciones de Einstein relajadas, Tensor de Weyl

Abstract

In the first part of this text an introduction to differential geometry is presented as a tool for General Relativity. The linearized gravity is studied and the role that this one plays in the gravitational radiation, deepening in the gauge, energy and quadrupolar contribution concepts. After this the relaxed Einstein equations are presented as a generalization for the study of gravitational radiation and general expression of energy, linear and angular momentum are obtained. Later it is shown the relation between the flux of energy, lineal and angular momentum with the Weyl tensor.

Keywords: Gravitational Radiation, Linearized Gravity, Relaxed Einstein field equations, Weyl tensor

Contenido

1 Relatividad General

1.1. Introducción

Este capítulo es una breve introducción a la teoría de la relatividad general, partiendo desde el concepto de variedad, el espacio tangente y cotangente, el concepto de curvatura y el papel que juega la gravedad en todas estas ideas matemáticas. Las referencias clave de este capítulo son [3, 6].

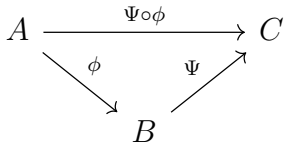
1.2. Variedades

1.2.1. Mapas

Definición 1.2.1 (Mapa)**.** *Dados dos conjuntos A y B, un mapa $\phi : M \to N$ es una relación que asigna cada elemento $x \in M$ a un único elemento $y \in N$. En este caso, se denota como $\phi(x) = y$.*

Definición 1.2.2 (Composición de Mapas)**.** *Con dos mapas $\phi : A \to B$ y $\Psi : B \to C$, se define la composición de ambos mapas, $\Psi \circ \phi : A \to C$, por su acción sobre los elementos de A:*

$$(\Psi \circ \phi)(a) = \Psi(\phi(a)).$$



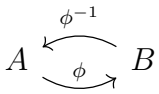
Un mapa $\phi : A \to B$ se dice inyectivo (uno a uno) si cada elemento de B tiene a lo sumo un elemento de A que es mapeado a él. Este mapa se dice sobreyectivo si cada elemento de B tiene al menos un elemento de A mapeado a él. A se conoce como el dominio del mapa ϕ, y su imagen es

$$Im\,\phi := \{y \in B : \exists x \in A \text{ tal que } \phi(x) = y\}.$$

La preimagen de un conjunto $U \subseteq B$ bajo la función ϕ se define como

$$\phi^{-1}(U) := \{x \in A : \exists y \in U \text{ tal que } \phi(x) = y\}.$$

Un mapa $\phi : A \to B$ que es inyectivo y sobreyectivo a la vez se conoce como invertible (biyectivo). En este caso, se define el mapa inverso $\phi^{-1} : B \to A$ de modo que se satisfaga que, para todo $y \in B$ $(\phi \circ \phi^{-1})(y) = y$.



Un mapa f de $\mathbb{R}^m$ a $\mathbb{R}^n$ toma una $m-$tupla $(x^1, x^2, \ldots, x^m)$ y la envía a una $n-$tupla $(y^1, y^2, \ldots, y^n)$, de modo que se puede pensar como una colección de n funciones ϕ^i de m variables:

$$y^i = \phi^i(x^1, \ldots, x^m) \text{ con } i = 1, \ldots, n,$$

de modo que

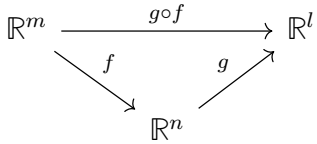
$$f(x^1, \ldots, x^m) = (\phi^1(x^1, \ldots, x^m), \ldots, \phi^n(x^1, \ldots, x^m)).$$

Se referirá a cada una de las funciones ϕ^i como C^p si son continuas y $p-$veces diferenciables, y al mapa entero $f : \mathbb{R}^m \to \mathbb{R}^n$ como C^p si cada uno de los campos escalares $\phi^i, i = 1, \ldots, n$ es al menos C^p.
Un mapa C^0 es continuo pero no necesariamente diferenciable, mientras que un mapa C^∞ es continuo y puede ser diferenciado cuantas veces se desee. Los mapas C^∞ se llaman suaves.

Definición 1.2.3 (Difeomorfismo)**.** *El mapa $\phi : A \to B$ se conoce como difeomorfismo si es biyectivo, y tanto él como su inversa son C^∞. Se dice entonces que los conjuntos A y B son difeomorfos.*

1.2.2. Regla de la cadena

Si tiene dos mapas $f : \mathbb{R}^m \to \mathbb{R}^n$ y $g : \mathbb{R}^n \to \mathbb{R}^l$, que se componen en $(g \circ f) : \mathbb{R}^m \to \mathbb{R}^l$, represente cada espacio en términos de coordenadas: x^a en $\mathbb{R}^m$, y^b en $\mathbb{R}^n$ y z^c en $\mathbb{R}^l$, donde los índices a, b, c varían sobre los valores apropiados.



La regla de la cadena relaciona las derivadas parciales de la composición $(g \circ f)$ con las derivadas parciales de los mapas f y g de la siguiente manera

$$\frac{\partial}{\partial x^a}(g \circ f)^c = \sum_{b=1}^{n} \frac{\partial f^b}{\partial x^a} \frac{\partial g^c}{\partial y^b}.$$

1.2.3. Variedades

En el capítulo de estabilidad se trató el concepto de variedad como un espacio métrico homeomorfo localmente a la bola abierta. En este capítulo se tomará una variedad más general: la variedad topológica, para lo cual se introducirá la idea de topología, y demás conceptos necesarios en términos de la topología de la variedad.

Definición 1.2.4 (Espacio topologico)**.** *Tome A como un conjunto arbitrario. τ es una topología para el conjunto A si satisface las siguientes condiciones:*

1. $\emptyset, A \in \tau$,
2. *Si $\{U_\alpha\}_{\alpha\in I} \subset \tau$ es una familia arbitraria de elementos de τ, entonces la unión de toda esta familia pertenece a τ, es decir, $\bigcup_{\alpha\in I} U_\alpha \in \tau$, y*
3. *Si $\{U_n\}_{n=1}^m \subset \tau$ es una familia finita de elementos de τ, entonces la intersección de todos sus elementos también es un elemento de τ, es decir, $\bigcap_{n=1}^m U_n \in \tau$.*

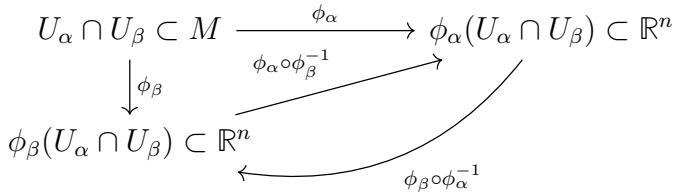
En este caso se dice que la pareja (A, τ) es un espacio topológico. Los elementos de τ se llaman abiertos y sus complementos se llaman cerrados.

Definición 1.2.5 (Carta o sistema coordenado)**.** *Considere un espacio topológico (M, τ). Una carta o sistema coordenado (U, ϕ) consiste de un conjunto abierto $U \subset M$, junto con un mapa inyectivo $\phi : U \to \mathbb{R}^n$, tal que $\phi(U)$ es abierto en $(\mathbb{R}^n, \tau_u)$*[1]*.*

Definición 1.2.6 (Atlas C^r)**.** *Un atlas C^r es una colección indexada de cartas $\{(U_\alpha.\phi_\alpha)\}_{\alpha\in I}$, con ϕ_α siendo al menos C^r, para todo $\alpha \in I$, que satisface las siguientes condiciones*

1. *$\bigcup_{\alpha\in I} U_\alpha = M$, es decir, $\{U_\alpha\}_{\alpha\in I}$ es un cubrimiento abierto para M y*
2. *si para algunos $\alpha, \beta \in I$ $(\alpha \neq \beta)$, $U_\alpha \cap U_\beta \neq \emptyset$, entonces el mapa $(\phi_\alpha \circ \phi_\beta^{-1}) : \phi_\beta(U_\alpha \cap U_\beta) \to \phi_\alpha(U_\alpha \cap U_\beta)$ toma puntos en $\phi_\beta(U_\alpha \cap U_\beta) \subseteq \mathbb{R}^n$ y los envía a puntos en $\phi_\alpha(U_\alpha \cap U_\beta)$, y viceversa. Ambas composiciones deben ser C^r. Si se satisface esta condición se dice que los mapas ϕ_α y ϕ_β son compatibles*

Un atlas se dice maximal si contiene todas las posibles cartas compatibles.



Definición 1.2.7 (C^r Variedad $n-$dimensional)**.** *Una C^r variedad $n-$dimensional es un espacio topológico (M, τ) junto con un atlas maximal C^r.*

[1] τ_u denota la topología usual sobre $\mathbb{R}^n$.

El hecho de que una variedad sea localmente como $\mathbb{R}^n$ (a través de las cartas) introduce la posibilidad de usar herramientas del cálculo real sobre ella. Tome por ejemplo dos C^∞ variedades (M, τ_M) y (N, τ_N) de dimensión m y n, respectivamente. Por simplicidad, pero sin pérdida de generalidad, tome $\phi : M \to \mathbb{R}^m$ y $\Psi : N \to \mathbb{R}^n$ como las cartas coordenadas de M y N, respectivamente. Si $f : M \to N$ es una función entre ambas variedades,

$$\begin{array}{ccc} M & \xrightarrow{\quad f \quad} & N \\ {\scriptstyle \phi^{-1}}\uparrow & & \downarrow{\scriptstyle \Psi} \\ \mathbb{R}^m & \xrightarrow{\Psi \circ f \circ \phi^{-1}} & \mathbb{R}^n \end{array}$$

se puede introducir el concepto de diferenciación sobre el mapa f, construyendo el mapa

$$(\Psi \circ f \circ \phi^{-1}) : \mathbb{R}^m \to \mathbb{R}^n,$$

de modo que el operador $\frac{\partial f}{\partial x^\mu}$ quede definido como

$$\frac{\partial f}{\partial x^\mu} := \frac{\partial}{\partial x^\mu}(\Psi \circ f \circ \phi^{-1}),$$

donde $\mu = 1, \ldots, m$.

1.2.4. Espacio tangente y cotangente

Tome $\mathcal{F}$ como el espacio de todas las funciones suaves $f : M \to \mathbb{R}$ ($\phi^{-1} \circ f$ es de clase C^∞, siendo ϕ la carta coordenada de M). Cada curva $\gamma : \mathbb{R} \to M$ que pasa por algún punto $p \in M$ define un operador sobre el espacio, la derivada direccional, que mapea f a

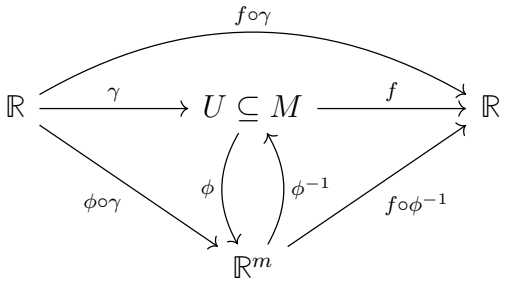
$$\frac{\mathrm{d}f}{d\lambda}\Big|_{\lambda:\gamma(\lambda)=p} := \frac{\mathrm{d}}{d\lambda}(f \circ \gamma)(\lambda)$$

(evaluada en p).

$$\begin{array}{ccccc} \mathbb{R} & \xrightarrow{\gamma} & M & \xrightarrow{f} & \mathbb{R} \\ & & \underset{f\circ\gamma}{\searrow} & & \end{array}$$

Definición 1.2.8 (Espacio tangente)**.** *El espacio tangente T_pM a un punto $p \in M$ es el espacio de los operadores derivadas direccionales dados por todas las curvas que pasan por el punto p. Este espacio resulta ser un espacio vectorial.*

El espacio tangente T_pM posee una base natural, $\{\partial_\mu\}$. Cada uno de estos operadores está definido en términos de la curva generada por la carta coordenada del punto p. Es decir, si (U, ϕ) es una carta coordenada tal que $p \in U$, se toma $(\phi^{-1})^\mu : \mathbb{R} \to M$ como la restricción de la función ϕ^{-1} a una única variable, x^μ, $\mu = 1, \ldots, m$, con el objetivo de que esta nueva función sea una curva sobre M que pase por p, para que defina la derivada direccional ∂_μ. Para ver que efectivamente es una base del espacio tangente T_pM, considere una variedad m-dimensional suave M, una carta coordenada (U, ϕ), una curva $\gamma : \mathbb{R} \to M$ y una función $f : M \to \mathbb{R}$.



Si λ es el parámetro de la curva γ, se expande el operador $\frac{\mathrm{d}}{\mathrm{d}\lambda}$ en términos de los operadores ∂_μ aplicando la regla de la cadena:

$$\frac{\mathrm{d}f}{\mathrm{d}\lambda} = \frac{\mathrm{d}}{\mathrm{d}\lambda}(f\circ\gamma) = \frac{\mathrm{d}}{\mathrm{d}\lambda}((f\circ\phi^{-1})\circ(\phi\circ\gamma)) = \frac{\mathrm{d}(\phi\circ\gamma)^\mu}{\mathrm{d}\lambda}\frac{\partial(f\circ\phi^{-1})}{\partial x^\mu} = \frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda}\partial_\mu f.$$

Como la función f es arbitraria,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} = \frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda}\partial_\mu,$$

con lo que los operadores derivada direccional $\{\partial_\mu\}$ son una base para T_pM, conocida como base coordenada. Además, esto implica que el espacio tangente T_pM tiene la misma dimensión de la variedad.

Una de las ventajas de este punto de vista de los vectores como operadores diferenciales es que la ley de transformación es inmediata. Como los vectores de la base son $\hat{e}_{(\mu)} = \partial_\mu$, los vectores de la base en un nuevo sistema coordenado $x^{\mu'}$ están dadas por la regla de la cadena [3]

$$\partial_{\mu'} = \frac{\partial x^\mu}{\partial x^{\mu'}}\partial_\mu$$

La ley de transformación de vectores se introduce de tal forma que un vector del espacio tangente $V = V^\mu\partial_\mu$ permanezca invariante bajo un cambio de base, es decir,

$$V^\mu\partial_\mu = V^{\mu'}\partial_{\mu'} = V^{\mu'}\frac{\partial x^\mu}{\partial x^{\mu'}}\partial_\mu,$$

y como la matriz $\frac{\partial x^{\mu'}}{\partial x^\mu}$ es la inversa de $\frac{\partial x^\mu}{\partial x^{\mu'}}$, la ley de transformación es

$$V^{\mu'} = \frac{\partial x^{\mu'}}{\partial x^\mu}V^\mu. \tag{1-1}$$

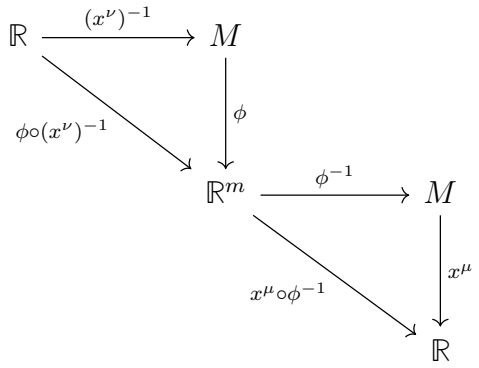
Definición 1.2.9 (Espacio cotangente)**.** *El espacio cotangente T_p^*M de una variedad M en un punto $p \in M$ es el conjunto de los mapas lineales $\omega : T_pM \to \mathbb{R}$. Los elementos de este espacio se conocen como 1-formas.*

El ejemplo canónico de 1-forma es el gradiente de una función $f : M \to \mathbb{R}$, denotado por $\mathrm{d}f$. Su acción sobre un vector $\frac{\mathrm{d}}{\mathrm{d}\lambda}$ del espacio tangente es exactamente la derivada direccional sobre la función f:

$$\mathrm{d}f\left(\frac{\mathrm{d}}{\mathrm{d}\lambda}\right) = \frac{\mathrm{d}f}{\mathrm{d}\lambda}\Big|_p.$$

Justo como las derivadas parciales a lo largo de los ejes coordenados proveen una base natural para el espacio tangente, los gradientes de las funciones coordenadas x^μ proveen una base natural para el espacio cotangente $\{\mathrm{d}x^\mu\}$, conocida como base dual. Observe que, al aplicar $\mathrm{d}x^\mu$ a ∂_η se obtiene que

$$\mathrm{d}x^\mu(\partial_\nu) = \frac{\partial x^\mu}{\partial x^\nu} = \frac{\partial}{\partial x^\nu}(x^\mu \circ (x^\nu)^{-1}) = \frac{\partial}{\partial x^\nu}((x^\mu \circ \phi^{-1}) \circ (\phi \circ (x^\nu)^{-1})).$$



Aplicando la regla de la cadena,

$$\frac{\partial}{\partial x^\nu}((x^\mu \circ \phi^{-1}) \circ (\phi \circ (x^\nu)^{-1})) = \frac{\partial(\phi \circ (x^\nu)^{-1})^\eta}{\partial x^\nu}\frac{\partial(x^\mu \circ \phi^{-1})}{\partial x^\eta}.$$

Intuitivamente, $x^\mu \circ \phi^{-1} = x^\mu$, donde x^μ del lado izquierdo de la igualdad es la μ-ésima coordenada de la carta, y al lado derecho es la μ-ésima componente de $\mathbb{R}^m$. Por otro lado, $(\phi \circ (x^\nu)^{-1})^\eta = x^\eta$, con lo que

$$\frac{\partial(\phi \circ (x^\nu)^{-1})^\eta}{\partial x^\nu}\frac{\partial(x^\mu \circ \phi^{-1})}{\partial x^\eta} = \frac{\partial x^\eta}{\partial x^\nu}\frac{\partial x^\mu}{\partial x^\eta} = \delta^\mu_\eta\frac{\partial x^\eta}{\partial x^\nu} = \frac{\partial x^\mu}{\partial x^\nu} = \delta^\mu_\nu.$$

En resumen,

$$\boxed{\mathrm{d}x^\mu(\partial_\nu) = \delta^\mu_\nu.} \tag{1-2}$$

Esta condición determina que $\{\mathrm{d}x^\mu\}$ es una base para el espacio cotangente T^*_pM [3]. De este modo, cualquier 1-forma ω se puede expandir en sus componentes: $\omega = \omega_\mu \mathrm{d}x^\mu$. Las propiedades de transformación de los vectores de la base dual y las componentes de una 1-forma se siguen de la misma forma que en el caso del espacio tangente:

$$\mathrm{d}x^{\mu'} = \frac{\partial x^{\mu'}}{\partial x^\mu}\mathrm{d}x^\mu;\ \omega_{\mu'} = \frac{\partial x^\mu}{\partial x^{\mu'}}\omega_\mu.$$

Definición 1.2.10 (Espacio producto cartesiano)**.** *Se define el espacio producto cartesiano* Π^k_l *respecto a un punto* $p \in M$ *de la variedad como:*

$$\Pi^k_l := \underbrace{T^*_pM \times \cdots \times T^*_pM}_{l-veces} \times \underbrace{T_pM \times \cdots \times T_pM}_{k-veces},\ \textit{es decir,}$$

$$\Pi^k_l = \{(\omega^1, \omega^2, \ldots, \omega^l, \mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_k) : \omega^i \in T^*_pM;\ \mathbb{X}_i \in T_pM\}.$$

Este espacio es un espacio vectorial con la suma y el producto usuales.

Definición 1.2.11 (Tensores)**.** *Un tensor* (k,l) $\mathbb{T} : \Pi_l^k \to \mathbb{R}$ *es un mapa multilineal (lineal en cada uno de sus argumentos). Este tensor se puede expandir en términos de las bases del espacio tangente y cotangente de la siguiente forma:*

$$\mathbb{T} = T^{\mu_1 \cdots \mu_k}{}_{\nu_1 \cdots \nu_l} \partial_{\mu_1} \otimes \cdots \otimes \partial_{\mu_k} \otimes \mathrm{d}x^{\nu_1} \otimes \cdots \otimes \mathrm{d}x^{\nu_l},$$

donde $T^{\mu_1 \cdots \mu_k}{}_{\nu_1 \cdots \nu_l}$ *son los coeficientes del tensor.*

De modo similar al caso de los vectores, los tensores transforman coordenadas en cada uno de sus índices de la siguiente forma:

$$T^{\mu_1' \cdots \mu_k'}{}_{\nu_1' \cdots \nu_l'} = \frac{\partial x^{\mu_1'}}{\partial x^{\mu_1}} \cdots \frac{\partial x^{\mu_k'}}{\partial x^{\mu_k}} \frac{\partial x^{\nu_1}}{\partial x^{\nu_1'}} \cdots \frac{\partial x^{\nu_l}}{\partial x^{\nu_l'}} T^{\mu_1 \cdots \mu_k}{}_{\nu_1 \cdots \nu_l}.$$

Infortunadamente, la derivada parcial de un tensor no es, en general, un tensor (no cumple esta regla de transformación de coordenadas). Esto motivará posteriormente la derivada covariante, que preservará el carácter tensorial tras aplicarse sobre un tensor.

Definición 1.2.12 (Tensor métrico)**.** *El tensor métrico* $g_{\mu\nu}$ *es un tensor simétrico* $(0,2)$*, cuya representación matricial tiene determinante no nulo (*$g = |g_{\mu\nu}| \neq 0$*), y satisface la relación*

$$g^{\mu\nu} g_{\nu\sigma} = \delta^{\mu}{}_{\sigma}. \tag{1-3}$$

La simetría de $g_{\mu\nu}$ *implica la simetría de* $g^{\mu\nu}$*, y la relación* (1-3) *permite que el tensor métrico se pueda usar para subir o bajar índices.*

Definición 1.2.13 (Elemento de línea)**.** *El elemento de línea se define de la siguiente forma*

$$\mathrm{d}s^2 = g_{\mu\nu} \mathrm{d}x^{\mu} \mathrm{d}x^{\nu}.$$

2 Estadística bayesiana

En este capítulo se hace una breve introducción a la teoría bayesiana, con el fin de obtener las herramientas necesarias para ajustar los parámetros de un modelo dado a través de un conjunto de datos experimentales. Algunas referencias clave de este capítulo son [1, 7, 8, 4, 2, 5].

2.1. Preliminares

Antes de enunciar la regla de Bayes, es necesario definir lo que es un espacio de probabilidad, y la probabilidad condicional. Esta sección busca fundamentar las bases de la teoría de la probabilidad, para comprender a cabalidad temas posteriores.

2.1.1. Espacio de probabilidad

La probabilidad es una teoría matemática que busca medir de alguna forma la posibilidad de que ocurra un evento contenido en un conjunto de posibles eventos, resultados todos de un experimento. Por supuesto, no se conoce de forma determinista cuál será el resultado tras la ejecución del experimento, de modo que sólo se puede hablar de posibilidades de que ocurra algún evento. Este tipo de experimentos se conocerán como experimentos aleatorios.

Definición 2.1.1 (Experimento Aleatorio)**.** *Un experimento se dice aleatorio si su resultado no se puede determinar de antemano.*

Definición 2.1.2 (Espacio de Muestra)**.** *El conjunto Ω de todos los posibles resultados de un experimento aleatorio se llama espacio de muestra. Un elemento $\omega \in \Omega$ se llama resultado o muestra. Ω se dice discreto si es finito o contable.*

Ahora se requiere definir lo que se entenderá por evento, para lo cual se definirá una estructura sobre el espacio de muestra, conocida como $\sigma-$álgebra, que dará cuenta de los eventos de interés tras la ejecución de un experimento aleatorio.

Definición 2.1.3 (σ-álgebra)**.** *Tome $\Omega \neq \emptyset$. Una colección $\Im$ de subconjuntos de Ω se llama $\sigma-$álgebra sobre Ω si:*

1. *$\Omega \in \Im$,*
2. *Si $A \in \Im$, entonces $A^c \in \Im$ y,*

3. Si $A_1, A_2, \ldots \in \Im$*, entonces* $\bigcup_{i=1}^{\infty} A_i \in \Im$*.*

Los elementos de $\Im$ *se llaman eventos.*

El siguiente teorema será de utilidad para construir una σ-álgebra a partir de un conjunto finito o contable de σ-álgebras.

Teorema 2.1.1. *Si* $\Omega \neq \emptyset$ *y* $\Im_1, \Im_2, \ldots$ *son* $\sigma-$*álgebras sobre* Ω*, entonces* $\bigcap_{i=1}^{\infty} \Im_i$ *es una* σ*-álgebra sobre* Ω*.*

Demostración. Como $\Omega \in \Im_j$, para $j = 1, 2, \ldots$, $\Omega \in \bigcap_{j=1}^{\infty} \Im_j$. Si $A \in \bigcap_{j=1}^{\infty} \Im_j$, $A \in \Im_j$, para $j = 1, 2, \ldots$, de modo que $A^c \in \Im_j$, y $A^c \in \bigcap_{j=1}^{\infty} \Im_j$. Por último, si

$$A_1, A_2, \ldots \in \bigcap_{j=1}^{\infty} \Im_j,$$

para todo $j = 1, 2, \ldots$, $A_1, A_2, \ldots \in \Im_j$, de modo que

$$\bigcup_{i=1}^{\infty} A_i \in \Im_j \text{ y } \bigcup_{i=1}^{\infty} A_i \in \bigcap_{j=1}^{\infty} \Im_j.$$

□

Con este teorema, se puede definir la σ-álgebra más pequeña[1] que contiene un subconjunto de Ω.

Definición 2.1.4 (σ-álgebra generada)**.** *Tome* $\Omega \neq \emptyset$ *y* $\mathcal{A}$ *como una colección de subconjuntos de* Ω*. Si* $\mathcal{M} := \{\Im : \Im$ *es una* $\sigma -$ *álgebra sobre* Ω *que contiene a* $\mathcal{A}\}$*,*

$$\sigma(\mathcal{A}) := \bigcap_{\Im \in \mathcal{M}} \Im$$

es la $\sigma-$*álgebra más pequeña sobre* Ω *que contiene a* $\mathcal{A}$*. Esta* $\sigma-$*álgebra se conoce como* σ*-álgebra generada por* $\mathcal{A}$*.*

Definición 2.1.5 (Espacio de medida)**.** *Tome* $\Omega \neq \emptyset$ *y sea* $\Im$ *una* σ*-álgebra sobre* Ω*. La pareja* $(\Omega, \Im)$ *se llama espacio de medida.*

Al evento $\emptyset$ se le conoce como evento imposible; Ω es el evento seguro y $\{\omega\}$, con $\omega \in \Omega$ es un evento simple. Se dice que el evento A ocurre después de llevar a cabo el experimento aleatorio si se obtiene un resultado en A, esto es, A ocurre si el resultado es algún $\omega \in A$.

1. El evento $A \cup B$ ocurre si y sólo si A ocurre, B pasa, o ambos ocurren.

[1] Es la más pequeña en el sentido de que es la que requiere menos elementos para satisfacer las condiciones necesarias para ser una σ-álgebra.

2. El evento $A \cap B$ ocurre si y sólo si A y B ocurren a la vez.

3. El evento A^c ocurre si y sólo si A no ocurre.

4. El evento $A - B$ ocurre si y sólo si A ocurre pero B no ocurre.

Si dos eventos no tienen eventos simples en común, se dirá que son eventos mutuamente excluyentes:

Definición 2.1.6 (Eventos mutuamente excluyentes)**.** *Dos eventos A y B se dicen mutuamente excluyentes si $A \cap B = \emptyset$.*

Antes de introducir la función de probabilidad, que medirá la posibilidad de que ocurra un evento de la $\sigma-$álgebra, es necesario definir por completez la frecuencia relativa, pues ella determina la posibilidad de que ocurra un evento al cabo de n repeticiones del experimento aleatorio.

Definición 2.1.7 (Frecuencia relativa)**.** *Para cada evento A, el número $f_r(A) := \frac{n(A)}{n}$ se llama la frecuencia relativa de A, donde $n(A)$ indica el número de veces que ocurre A en n repeticiones del experimento aleatorio.*

Cuando $n \to \infty$, se puede hablar intuitivamente de la probabilidad de que ocurra el evento A, normalizada de 0 a 1. Por supuesto, es imposible realizar infinitas veces un experimento aleatorio para determinar la probabilidad de ocurrencia de todos los eventos de la $\sigma-$álgebra, por lo que se introduce de antemano la función de probabilidad, suponiendo que ella da cuenta del comportamiento de la frecuencia relativa cuando $n \to \infty$.

Definición 2.1.8 (Espacio de probabilidad)**.** *Tome $(\Omega, \Im)$ como un espacio de medida. Una función real P sobre $\Im$ que satisface las siguientes condiciones:*

1. *$P(A) \geq 0$ para todo $A \in \Im$ (no negativa),*

2. *$P(\Omega) = 1$ (normalizada) y,*

3. *si $A_1, A_2, \ldots$ son eventos mutuamente excluyentes en $\Im$, esto es, si*

$$A_i \cap A_j = \emptyset \text{ para todo } i \neq j, \text{ entonces}$$

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i),$$

se llama medida de probabilidad sobre $(\Omega, \Im)$. La tripleta $(\Omega, \Im, P)$ se llama espacio de probabilidad.

El siguiente teorema caracteriza las propiedades más importantes de un espacio de probabilidad.

Teorema 2.1.2. *Si $(\Omega, \Im, P)$ es un espacio de probabilidad, entonces*

1. $P(\emptyset) = 0$.
2. *Si $A, B \in \Im$ y $A \cap B = \emptyset$, entonces $P(A \cup B) = P(A) + P(B)$.*
3. *Para todo $A \in \Im$, $P(A^c) = 1 - P(A)$.*
4. *Si $A \subseteq B$, entonces $P(A) \leq P(B)$ y $P(B - A) = P(B) - P(A)$. En particular, $P(A) \leq 1$ para todo $A \in \Im$.*
5. *Para todo $A, B \in \Im$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.*
6. *Tome $\{A_n\}_n \subseteq \Im$ como una sucesión creciente, esto es, $A_n \subseteq A_{n+1}, \forall n \in \mathbb{N}$; entonces*

$$P\left(\lim_{m\to\infty} A_n\right) = \lim_{n\to\infty} P(A_n), \text{ donde } \lim_{n\to\infty} A_n := \bigcup_{i=1}^{\infty} A_i.$$

7. *Tome $\{A_n\}_n \subseteq \Im$ como una sucesión decreciente, esto es, $A_n \supseteq A_{n+1}, \forall n \in \mathbb{N}$; entonces*

$$P\left(\lim_{m\to\infty} A_n\right) = \lim_{n\to\infty} P(A_n), \text{ donde } \lim_{n\to\infty} A_n := \bigcap_{i=1}^{\infty} A_i.$$

Demostración.

1. $1 = P(\Omega \cup \emptyset \cup \emptyset \cup \cdots) = P(\Omega) + P(\emptyset) + P(\emptyset) + \cdots = 1 + P(\emptyset) + \cdots \implies P(\emptyset) = 0$.
2. $P(A \cup B) = P(A \cup B \cup \emptyset \cup \emptyset \cup \cdots) = P(A) + P(B)$.
3. $P(A) + P(A^c) = P(A \cup A^c) = P(\Omega) = 1 \implies P(A^c) = 1 - P(A)$.
4. Si $A \subseteq B$, $B = A \cup (B - A)$, de modo que $P(B) = P(A) + P(B - A)$. Como $P \geq 0$, $P(B) \geq P(A)$ y $P(B - A) = P(B) - P(A)$. Si $B = \Omega$, $P(A) \leq 1$.
5. Use el hecho de que $A \cup B = [A - (A \cap B)] \cup [B - (A \cap B)] \cup [A \cap B]$.
6. Tome la sucesión $C_1 = A_1, C_2 = A_2 - A_1, \ldots, C_r = A_r - A_{r-1}, \ldots$. Es claro que

$$\bigcup_{i=1}^{\infty} C_i = \bigcup_{i=1}^{\infty} A_i.$$

Más aún, como $C_i \cap C_j = \emptyset \ \forall i \neq j$, se sigue que

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\bigcup_{n=1}^{\infty} C_n\right) = \sum_{n=1}^{\infty} P(C_n) = \lim_{n\to\infty} \sum_{k=1}^{n} P(C_k)$$

$$= \lim_{n\to\infty} P\left(\bigcup_{k=1}^{n} C_k\right) = \lim_{n\to\infty} P\left(A_n\right).$$

7. Tome la sucesión $\{B_n = A_n^c\}_n$ y aplique el resultado anterior.

□

Aplicando el teorema anterior de forma inductiva, para algunos eventos $A_1, A_2, \ldots, A_n \in \Im$:

$$\begin{aligned} P(A_1 \cup A_2 \cup \cdots \cup A_n) = &\sum_{i=1}^{n} P(A_i) - \sum_{i_1<i_2} P(A_{i_1} \cap A_{i_2}) + \cdots \\ &+ (-1)^{r+1} \sum_{i_1<i_2<\cdots<i_r} P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_r}) \\ &+ \cdots + (-1)^{n+1} P(A_1 \cap A_2 \cap \cdots \cap A_n). \end{aligned}$$

Por otro lado, tome $(\Omega, \Im, P)$ como un espacio de probabilidad con Ω finito o contable y $\Im = \mathbb{P}(\Omega)$. Tome $\emptyset \neq A \in \Im$. Es claro que

$$A = \bigcup_{\omega \in A} \{\omega\}, \text{ de modo que}$$

$$P(A) = \sum_{\omega \in A} P(\omega), \text{ donde } P(\omega) := P(\{\omega\}).$$

Así, P queda completamente definido por $p_j := P(\omega_j)$, donde $\omega_j \in \Omega$. El vector $|\Omega|-$dimensional $p := (p_1, p_2, \ldots)$ satisface las siguientes condiciones:

- $p_j \geq 0$ y
- $\sum_{j=1}^{\infty} p_j = 1$.

Un vector que satisface las anteriores condiciones se llama **vector de probabilidad**.

2.1.2. Probabilidad condicional

Tome B como un evento cuya opción de ocurrir debe ser medida bajo la suposición de que otro evento A fue observado. Si el experimento se repite n veces bajo las mismas circunstancias, entonces la frecuencia relativa de B bajo la condición A se define como

$$f_r(B|A) := \frac{n(A \cap B)}{n(A)} = \frac{\frac{n(A \cap B)}{n}}{\frac{n(A)}{n}} = \frac{f_r(A \cap B)}{f_r(A)}, \text{ si } n(A) > 0.$$

Esto motiva la definición de probabilidad condicional, como el comportamiento de esta frecuencia relativa cuando $n \to \infty$.

Definición 2.1.9 (Probabilidad condicional)**.** *Tome $(\Omega, \Im, P)$ como un espacio de probabilidad. Si $A, B \in \Im$, con $P(A) > 0$, entonces la probabilidad del evento B bajo la condición A se define como sigue*

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

El siguiente teorema provee algunas propiedades de la probabilidad condicional.

Teorema 2.1.3 (Medida de probabilidad condicional)**.** *Tome* $(\Omega, \Im, P)$ *como un espacio de probabilidad y* $A \in \Im$*, con* $P(A) > 0$*. Entonces:*

1. $P(\cdot|A)$ *es una medida de probabilidad sobre* Ω *centrada en* A*, esto es,* $P(A|A) = 1$*.*
2. *Si* $A \cap B = \emptyset$*, entonces* $P(B|A) = 0$*.*
3. $P(B \cap C|A) = P(B|A \cap C)P(C|A)$ *si* $P(A \cap C) > 0$*.*
4. *Si* $A_1, A_2, \ldots, A_n \in \Im$*, con* $P(A_1 \cap A_2 \cap \cdots \cap A_{n-1}) > 0$*, entonces*

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}).$$

Demostración. 1. Las tres propiedades de una medida de probabilidad deben ser verificadas.

a) Claramente, $P(B|A) \geq 0$ para todo $B \in \Im$.

b) $P(\Omega|A) = \frac{P(\Omega \cap A)}{P(A)} = \frac{P(A)}{P(A)} = 1$. También se tiene que $P(A|A) = 1$.

c) Tome $A_1, A_2, \ldots \in \Im$ una sucesión de conjuntos disyuntos. Entonces

$$P\left(\bigcup_{i=1}^{\infty} A_i|A\right) = \frac{P\left(A \cap \bigcup_{i=1}^{\infty} A_i\right)}{P(A)} = \frac{P\left(\bigcup_{i=1}^{\infty} A \cap A_i\right)}{P(A)}$$
$$= \sum_{i=1}^{\infty} \frac{P(A \cap A_i)}{P(A)} = \sum_{i=1}^{\infty} P(A_i|A).$$

2. Si $A \cap B = \emptyset$, $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(\emptyset)}{P(A)} = 0$.

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

4. $P(A_1 \cap \cdots \cap A_n) = \frac{P(A_1 \cap \cdots \cap A_n)}{P(A_1 \cap \cdots \cap A_{n-1})} \frac{P(A_1 \cap \cdots \cap A_{n-1})}{P(A_1 \cap \cdots \cap A_{n-2})} \cdots \frac{P(A_1 \cap A_2)}{P(A_1)} P(A_1) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1})$.

□

El siguiente teorema permitirá probar la regla de Bayes.

Teorema 2.1.4 (Teorema de probabilidad total)**.** *Tome* $A_1, A_2, \ldots$ *como una partición finita o contable de* Ω*, esto es,* $A_i \cap A_j = \emptyset, \forall i \neq j$ *y* $\bigcup_{i=1}^{\infty} A_i = \Omega$*, tal que* $P(A_i) > 0$*, para todo* $A_i \in \Im$*. Entonces, para todo* $B \in \Im$*:*

$$P(B) = \sum_i P(B|A_i)P(A_i).$$

Demostración. Observe que

$$B = B \cap \Omega = B \cap \left(\bigcup_{i=1}^{\infty} A_i \right) = \bigcup_{i=1}^{\infty} B \cap A_i,$$

de modo que

$$P(B) = P\left(\bigcup_{i=1}^{\infty} B \cap A_i \right) = \sum_{i=1}^{\infty} P(B \cap A_i) = \sum_{i=1}^{\infty} P(B|A_i)P(A_i).$$

□

Matemáticamente, este teorema se puede interpretar como que la probabilidad de que ocurra B se puede medir en términos de una partición de Ω en el sentido de que B, como subconjunto de Ω puede ocurrir cuando ocurran algunos elementos de la partición, los cuales tendrán mayor peso en el término $P(B|A_i)$ de la suma.
Como corolario del teorema anterior, se obtiene la **regla de Bayes**, que constituye la base para la **teoría Bayesiana**.

Corolario 2.1.1 (Regla de Bayes)**.** *Tome $A_1, A_2, \ldots$ como una partición finita o contable de Ω con $P(A_i) > 0$, para todo i; entonces, para todo $B \in \Im$ con $P(B) > 0$:*

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_j P(B|A_j)P(A_j)}, \forall i.$$

Demostración.

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B|A_i)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_j P(B|A_j)P(A_j)}.$$

Con la partición $A_1 = A, A_2 = A^c$ se obtiene la forma usual de la regla de Bayes. □

A continuación se definen las distribuciones a priori y a posteriori, que hacen referencia a la probabilidad de que ocurran ciertos eventos de una partición de Ω antes de que ocurra un evento B y al cabo de que este evento B ocurra.

Definición 2.1.10 (Distribuciones a priori y a posteriori)**.** *Tome $A_1, A_2, \ldots$ como una partición finita o contable de Ω, con $P(A_i) > 0$, para todo i. Si $P(B) > 0$, con $B \in \Im$, entonces $\{P(A_n)\}_n$ se llama distribución a priori (antes de que B ocurra), y $\{P(A_n|B)\}_n$ se llama distribución a posteriori (después de que B ocurra).*

Algunas veces, la ocurrencia de un evento B no afecta la probabilidad de un evento A, es decir,

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

En este caso, se dice que el evento A es independiente del evento B. Esto motiva la siguiente definición.

Definición 2.1.11 (Eventos independientes)**.** *Dos eventos A y B se dicen independientes si y sólo si*

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

Si esta condición no se tiene, se dice que los eventos son dependientes.

En algunos casos, es necesario analizar la independencia de dos o más eventos. Para ello, se dan las siguientes definiciones.

Definición 2.1.12 (Familia independiente)**.** *Una familia de eventos $\{A_i : i \in I\}$ se dice independiente si*

$$P\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} P(A_i),$$

para cualquier subconjunto no vacío $J \subseteq I$.

Definición 2.1.13 (Eventos independientes par a par)**.** *Una familia de eventos $\{A_i : i \in I\}$ se dice par a par independiente si*

$$P(A_i \cap A_j) = P(A_i)P(A_j), \text{ para todo } i \neq j.$$

2.1.3. Variables aleatorias

En un experimento aleatorio, generalmente hay mayor interés en determinar ciertos valores numéricos asociados a los resultados del experimento aleatorio, que al resultado mismo del experimento aleatorio. Con esto en mente se define la variable aleatoria.

Definición 2.1.14 (Variable aleatoria)**.** *Tome $(\Omega, \Im, P)$ como un espacio de probabilidad. Una variable aleatoria es un mapa $X : \Omega \to \mathbb{R}$ tal que, para todo $A \in \mathbb{B}$, $X^{-1}(A) \in \Im$, donde $\mathbb{B}$ es la σ-álgebra de Borel sobre $\mathbb{R}$ (σ-álgebra más pequeña que contiene todos los intervalos de la forma $(-\infty, a]$).*
El conjunto de posibles valores de X es $\mathbb{S} := \{x \in \mathbb{R} : \exists \omega \in \Omega \text{ tal que } X(\omega) = x\}$, conocido como ***soporte de la variable aleatoria*** *X.*

Si X es una variable aleatoria definida sobre un espacio de probabilidad $(\Omega, \Im, P)$, se introduce la notación

$$\{X \in B\} := \{\omega \in \Omega : X(\omega) \in B\}, \text{ con } B \in \mathbb{B}.$$

Definición 2.1.15 (Variable aleatoria discreta)**.** *Una variable aleatoria X se dice discreta cuando el soporte $\mathbb{S}$ de X es un subconjunto finito o contable de $\mathbb{R}$. Para $x \in \mathbb{S}$, la función $f(x) = P(X = x)$ se llama función de densidad de probabilidad (pdf para abreviar).*

Definición 2.1.16 (Variable aleatoria continua)**.** *Una variable aleatoria X se dice continua si el soporte $\mathbb{S}$ de X es la unión de uno o más intervalos y si existe una función no negativa y real $f(x)$ tal que $P(X \leq x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$. La función $f(x)$ se llama función de densidad de probabilidad (pdf).*

Algunas propiedades de la *pdf* discreta son las siguientes:

1. $f(x) \geq 0, \forall x \in \mathbb{S}$ y $f(x) = 0, \forall x \notin \mathbb{S}$.
2. $\sum_{x \in \mathbb{S}} f(x) = 1$.
3. $P(X \in B) = \sum_{x \in B} f(x)$.

Análogamente, para la *pdf* continua:

1. $f(x) \geq 0, \forall x \in \mathbb{S}$ y $f(x) = 0, \forall x \notin \mathbb{S}$.
2. $\int_{\mathbb{S}} f(x)\mathrm{d}x = 1$.
3. $P(X \in B) = \int_{B} f(x)\mathrm{d}x$.

Definición 2.1.17 (Función de distribución acumulativa)**.** *La función de distribución acumulativa (CDF, para abreviar) de una variable aleatoria se define como la función $F(x) = P(X \leq x)$.*

El siguiente teorema resume algunas propiedades importantes de una *CDF*.

Teorema 2.1.5. *Si X es una variable aleatoria, con CDF $F(x)$, entonces:*

1. $\lim_{x \to -\infty} F(x) = 0$ *y* $\lim_{x \to \infty} F(x) = 1$.
2. *$F(x)$ es no decreciente; esto es, $F(x) \leq F(y)$, siempre que $x \leq y$.*
3. *$F(x)$ es continua por derecha.*
4. $P(a < X \leq b) = F(b) - F(a)$.

Demostración. 1. **Cuando $\mathbf{x \to -\infty}$**, para todo $n \in \mathbb{N}$, se satisface que

$$\{X \leq -n\} \supseteq \{X \leq -(n+1)\}, \text{ y en adición,}$$

$$\emptyset = \bigcap_{n=1}^{\infty} \{X \leq -n\}, \text{ con lo que}$$

$$\lim_{x \to -\infty} F(x) = \lim_{n \to \infty} F(-n) = P(\emptyset) = 0.$$

Cuando $\mathbf{x \to \infty}$, para todo $n \in \mathbb{N}$, se satisface que

$$\{X \leq n\} \subseteq \{X \leq n+1\}, \text{ y,}$$

$$\Omega = \bigcup_{n=1}^{\infty} \{X \leq n\}, \text{ de modo que}$$

$$\lim_{x\to\infty} F(x) = \lim_{n\to\infty} = P(\Omega) = 1.$$

2. Si $x \leq y$, entonces

$$\{X \leq x\} \subseteq \{X \leq y\}, \textit{ con lo que}$$

$$F(x) = P(X \leq x) \leq P(X \leq y) = F(y).$$

3. Tome $x \in \mathbb{R}$ fijo. Suponga que $\{x_n\}_{n\in\mathbb{N}}$ es una sucesión decreciente de números reales, cuyo límite va a x. Se puede ver que

$$\{X \leq x_1\} \supseteq \{X \leq x_2\} \supseteq \cdots, \text{ y}$$

$$\bigcap_{n=1}^{\infty} \{X \leq x_n\} = \{X \leq x\}, \text{ de modo que}$$

$$\lim_{y\to x^+} F(y) = \lim_{n\to\infty} F(x_n) = \lim_{n\to\infty} P(X \leq x_n) = P(X \leq x) = F(x).$$

4. Dado que $\Omega = \{X \leq a\} \cup \{a < X \leq b\} \cup \{X > b\}$,

$$1 = P(X \leq a) + P(a < X \leq b) + P(X > b),$$

$$\therefore P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a).$$

□

Los teoremas posteriores indican cómo determinar la *pdf* a partir de la *CDF* de una variable aleatoria.

Teorema 2.1.6. *Si X es una variable aleatoria discreta con CDF $F(x)$ y soporte $\mathbb{S} = \{x_0, x_1, \dots\}$, con $x_0 < x_1 < \cdots$, entonces, para $x_k \in \mathbb{S}$,*

$$f(x_k) = F(x_k) - F(x_{k-1}).$$

Demostración. Para $k \geq 1$,

$$f(x_k) = P(X = x_k) = P(x_{k-1} < X \leq x_k) = F(x_k) - F(x_{k-1}).$$

□

Teorema 2.1.7. *Para una variable aleatoria continua, $f(x) = \frac{\mathrm{d}F}{\mathrm{d}x}, \forall x \in \mathbb{R}$.*

Demostración. La prueba se sigue directamente del teorema fundamental del cálculo. □

2.1.4. Vectores aleatorios

En la mayoría de los análisis estadísticos, más de una variable debe ser analizada al cabo de un experimento aleatorio. Cada observación se puede representar como un vector de observaciones, conocido como vector aleatorio.

Definición 2.1.18 (Vector aleatorio)**.** *Un vector aleatorio $\vec{X} = (X_1, X_2, \ldots, X_k)$ es un vector $k-$dimensional, donde $X_1, \ldots, X_k$ son variables aleatorias. Un vector aleatorio se dice discreto cuando cada una de las variables aleatorias que lo conforman son discretas, y continuo cuando son continuas.*

Definición 2.1.19 (Variable aleatoria bivariada)**.** *Un vector aleatorio bidimensional $\vec{X} = (X_1, X_2)$ se llama variable aleatoria bivariada.*

De modo similar al caso de las variables aleatorias, los vectores aleatorios tienen *pdf*, un soporte y una *CDF*. El soporte de un vector aleatorio $k-$dimensional es el conjunto de valores que puede tomar, denotado por $\mathbb{S}_{\vec{X}} \subseteq \mathbb{R}^k$.

Definición 2.1.20 (Función de densidad de probabilidad adjunta discreta)**.** *Tome $\vec{X}$ como un vector aleatorio discreto $k-$dimensional. La pdf adjunta de $\vec{X}$ se define como*

$$f(\vec{x}) := f(x_1, x_2, \ldots, x_k) = P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k)$$

para $\vec{x} = (x_1, x_2, \ldots, x_k) \in \mathbb{S}_{\vec{X}}$.

La *pdf* adjunta discreta tiene las siguientes propiedades:

1. $0 \leq f(x_1, x_2, \ldots, x_k) \leq 1, \forall \vec{x} \in \mathbb{S}_{\vec{X}}$.
2. $\sum_{\vec{x} \in \mathbb{S}_{\vec{X}}} f(\vec{x}) = 1$.
3. Para cualquier subconjunto $B \subseteq \mathbb{S}_{\vec{X}}$, $P(\vec{X} \in B) = \sum_{\{\vec{x} \in \mathbb{S}_{\vec{X}} : \vec{x} \in B\}} f(\vec{x})$.

Definición 2.1.21 (Función de densidad de probabilidad adjunta continua)**.** *Tome $\vec{X}$ como un vector aleatorio k-dimensional continuo. La pdf continua de $\vec{X}$ se define como cualquier función no negativa $f(\vec{x})$ que satisfaga las siguientes propiedades:*

1. *$f(x_1, \ldots, x_k) > 0, \forall \vec{x} \in \mathbb{S}_{\vec{X}}$.*
2. *$\int_{\mathbb{S}_{\vec{X}}} f(x_1, \ldots, x_k) \mathrm{d}x_1 \cdots \mathrm{d}x_k = 1$.*
3. *Para cualquier subconjunto $B \subset \mathbb{S}_{\vec{X}}$, $P(\vec{X} \in B) = \int_B f(x_1, \ldots, x_k) \mathrm{d}x_1 \cdots \mathrm{d}x_k$.*

Definición 2.1.22 (Función de distribución acumulativa adjunta)**.** *Tome $\vec{X} = (X_1, X_2, \ldots, X_k)$ como un vector aleatorio $k-$dimensional. La CDF adjunta de $\vec{X}$ se define como*

$$F(x_1, x_2, \ldots, x_k) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_k \leq x_k), \ \forall (x_1, \ldots, x_k) \in \mathbb{R}^k.$$

Las componentes de un vector aleatorio $\vec{X}$ son variables aleatorias, por lo que las *pdf* de cada variable aleatoria X_i de $\vec{X}$ se pueden derivar de la *pdf* adjunta de $\vec{X}$.

Definición 2.1.23 (Función de densidad de probabilidad marginal)**.** *Tome $\vec{X} = (X_1, \ldots, X_k)$ como un vector aleatorio $k-$dimensional. La función de densidad de probabilidad marginal de la variable aleatoria X_i es*

$$f_i(x_i) = \underbrace{\sum_{x_1 \in \mathbb{S}_X} \cdots \sum_{x_k \in \mathbb{S}_{X_k}}}_{\text{quitando la suma sobre } x_i} f(x_1, \ldots, x_k) \qquad \text{cuando } \vec{X} \text{ es discreta y}$$

$$f_i(x_i) = \underbrace{\int_{x_1 \in \mathbb{S}_{X_1}} \cdots \int_{x_k \in \mathbb{S}_{X_k}}}_{\text{quitando la integral sobre } x_i} f(x_1, \ldots, x_k) \prod_{n \neq i} \mathrm{d}x_n \quad \text{cuando } \vec{X} \text{ es continua.}$$

También se puede definir la distribución condicional dada una variable aleatoria de forma similar al caso de la probabilidad condicional.

Definición 2.1.24 (Función de densidad de probabilidad condicional)**.** *Tome $\vec{X}(X_1, \ldots, X_k)$ como un vector aleatorio $k-$dimensional. Para un valor fijo de x_i, donde $f_i(x_i) > 0$, la función de densidad de probabilidad condicional para $\vec{Y}|X_i$; donde $\vec{Y}$ es un vector aleatorio $(k-1)-$dimensional con todas las variables aleatorias de $\vec{X}$, a excepción de X_i; es*

$$f(\vec{y}|x_i) = \frac{f(x_1, \ldots, x_k)}{f_i(x_i)},$$

donde $\vec{y} \in \mathbb{S}_{\vec{Y}}$.

Esta última definición motiva, como en el caso de los eventos independientes, el concepto de variables aleatorias independientes.

Definición 2.1.25 (Colección independiente de variables aleatorias)**.** *Una colección de variables aleatorias $\{X_1, X_2, \ldots, X_k\}$ se dice independiente cuando*

$$F(x_1, x_2, \ldots, x_k) = \prod_{i=1}^{k} F_i(x_i), \forall \vec{x} \in \mathbb{R}^k,$$

donde $F_i(x_i)$ es la CDF marginal de la variable aleatoria X_i (determinada a partir de la pdf marginal: $F_i(x_i) := P(X_i \leq x_i)$).
También se puede definir la independencia entre variables aleatorias usando las pdf, en el sentido de que la misma colección de variables aleatorias se dice independiente cuando

$$f(x_1, x_2, \ldots, x_k) = \prod_{i=1}^{k} f_i(x_i), \forall \vec{x} \in \mathbb{S}_{\vec{X}},$$

donde $f_i(x_i)$ es la pdf marginal asociada a la variable aleatoria X_i.

Definición 2.1.26 (Colección de variables aleatorias independientes idénticamente distribuidas)**.** *Una colección de variables aleatorias* $\{X_1, X_2, \ldots, X_k\}$ *se dice independiente e idénticamente distribuidas (iid, para abreviar) si y sólo si* $X_1, X_2, \ldots, X_k$ *son variables aleatorias independientes y la pdf de cada variable aleatoria es idéntica.*

2.2. Función de verosimilitud

2.2.1. Estadística

En la realidad, uno se encuentra en presencia de un experimento aleatorio del cual desconoce su función de densidad de probabilidad, dependiendo de las variables aleatorias asociadas a este. A través de una muestra lo suficientemente amplia, se desearía es reconstruir el modelo probabilístico que generó dicha muestra. Sin embargo, el problema resulta casi imposible de resolver sin especificar la forma de la función de densidad de probabilidad, por lo que se suele hacer es escoger una *pdf* con ciertos parámetros a determinar, buscando el mejor ajuste posible respecto a la muestra. Las componentes de esta estimación paramétrica son las siguientes [7]

1. Un modelo probabilístico $f(x, \theta)$, especificado con los valores de los parámetros desconocidos.

2. Un conjunto de posibles valores de θ bajo consideración, llamado el espacio de parámetros, denotado por Θ.

3. Una muestra aleatoria de n observaciones del modelo probabilístico.

4. Un conjunto de estimadores puntuales para los valores de los parámetros desconocidos, basados en la información contenida en la muestra aleatoria.

5. Las propiedades específicas de los estimadores que permiten evaluar la precisión y eficiencia del estimador.

A continuación se define lo que se entenderá por muestra.

Definición 2.2.1 (Muestra)**.** *Una colección de variables aleatorias* $X_1, \ldots, X_k$ *se llama muestra de tamaño* n*. Una muestra de* n *variables aleatorias independientes* $X_1, \ldots, X_n$ *se llama muestra aleatoria.*

Con una muestra dada, se pueden estimar algunos parámetros de una población. Esta estimación se conoce como estadística.

Definición 2.2.2 (Estadística y estimador)**.** *Dada una muestra* $X_1, X_2, \ldots, X_n$*, una estadística* $T = T(X_1, \ldots, X_n)$ *es una función de la muestra que no depende de ningún otro*

parámetro desconocido. Un estimador es una estadística que se usa para determinar una cantidad desconocida, y el estimado es el valor observado del estimador (evaluando la función en la muestra).

El comportamiento de un estimador y su efectividad para estimar un parámetro se puede medir a través de la distribución de probabilidad del estimador, conocida como distribución muestral.

Definición 2.2.3 (Distribución muestral)**.** *Para una muestra $X_1, \ldots, X_n$ y una estadística $T = T(X_1, \ldots, X_n)$, la distribución muestral de la estadística T es la distribución de probabilidad asociada a la variable aleatoria T. La pdf de la distribución muestral se denota como $f_T(t; \theta)$.*

Antes de continuar, es necesario definir el concepto de valor esperado o media, requerido para medir la eficiencia de un estimador.

Definición 2.2.4 (Valor esperado)**.** *Tome X como una variable aleatoria con pdf $f(x)$ en $\mathbb{S}_X$. El valor esperado de la variable aleatoria X, denotado por $E(X)$, se define como*

$$E(X) = \sum_{x \in \mathbb{S}_X} x f(x)$$

cuando X es una variable aleatoria discreta, y como

$$E(X) = \int_{x \in \mathbb{S}_X} x f(x) \mathrm{d}x$$

cuando X es una variable aleatoria continua.

Cuando una estadística T es usada para estimar un parámetro θ, se espera que la media de dicha estadística sea cercano a θ. Cuando se da la igualdad, T se llama estimador imparcial del parámetro θ.

Definición 2.2.5 (Estimador imparcial)**.** *Una estadística T se dice estimador imparcial de un parámetro θ cuando $E(T) = \theta$, $\forall \theta \in \Theta$. Una estadística se conoce como estimador parcial de θ cuando $E(T) \neq \theta$, y la parcialidad de una estadística T para estimar un parámetro θ se define como $Bias(T; \theta) = E(T) - \theta$.*

En algunos casos, cuando el estimador es parcial, se puede despreciar la parcialidad tomando una muestra lo suficientemente grande. Un estimador cuya parcialidad va a cero cuando $n \to \infty$ se conoce como estimador asintóticamente imparcial.

Definición 2.2.6 (Estimador asintóticamente imparcial)**.** *Una estadística $T_n = T(X_1, \ldots, X_n)$ se conoce como estimador asintóticamente imparcial de un parámetro θ cuando*

$$\lim_{n \to \infty} Bias(T_n; \theta) = 0.$$

A pesar de que es importante que un estimador se aproxime a su respectivo parámetro, el valor esperado (la media) no mide la precisión ni la exactitud del estimador T. Para medir la precisión se introduce el **error estándar de una estadística** como la desviación estándar de la misma, es decir,

$$SE(T) := \sqrt{E((T - E(T))^2)} := \sqrt{Var(T)}.$$

Por otro lado, para medir la exactitud de un estimador T usado para estimar un parámetro θ se introduce el **error cuadrado medio** asociado a T.

$$MSE(T;\theta) = E((T-\theta)^2).$$

Cuando se realiza una estimación paramétrica, la información acerca del parámetro desconocido $\theta \in \Theta$ está contenida en una muestra aleatoria de tamaño n seleccionada a partir de una *pdf* común $f(x;\theta)$. Una estadística que contiene toda la información relevante acerca de θ en una muestra se conoce como estadística suficiente [7].

Definición 2.2.7 (Estadística suficiente)**.** *Tome $X_1, \ldots, X_n$ como una muestra de variables aleatorias iid con pdf común $f(x;\theta)$, para $\theta \in \Theta \subseteq \mathbb{R}^d$. Un vector de estadísticas $\vec{S}(\vec{X}) := (S_1(\vec{X}), \ldots, S_k(\vec{X}))$ se dice que es una estadística suficiente $k-$dimensional para un parámetro θ si y sólo si la distribución condicional de $\vec{X}$ dado $S = s$ no depende de θ, para ningún valor de s.*

Esta última definición no es útil en la práctica para determinar la suficiencia de una estadística. Para esta labor se introduce la función de verosimilitud, que junto con el teorema de factorización de Neyman-Fisher puede determinar si una estadística es suficiente.

Definición 2.2.8 (Función de verosimilitud)**.** *Para una muestra $X_1, \ldots, X_n$, la función de verosimilitud $L(\theta|\vec{X})$ es la pdf adjunta de $\vec{X} = (X_1, \ldots, X_n)$, es decir,*

$$L(\theta|\vec{X}) = f(x_1, \ldots, x_n; \theta)$$

La función logarítmica de verosimilitud $\ell(\theta)$ se define como el logaritmo de la función de verosimilitud.

Cuando $X_1, \ldots, X_n$ es una muestra de variables aleatorias *iid*, se puede escribir la función de verosimilitud como

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta),$$

Teorema 2.2.1 (Teorema de factorización de Neyman-Fisher)**.** *Tome $X_1, \ldots, X_n$ como una muestra de variables aleatorias iid con pdf $f(x;\theta)$, y espacio de parámetros Θ. Una estadística $S(\vec{X})$ es suficiente para θ si y sólo si $L(\theta)$ se puede factorizar como*

$$L(\theta) = g(S(\vec{x});\theta)h(\vec{x}),$$

donde $g(S(\vec{x});\theta)$ no depende de $\vec{x} = (x_1, \ldots, x_n)$, excepto a través de $S(\vec{x})$, y $h(\vec{x})$ no depende de θ.

En un modelo paramétrico $f(\vec{x};\vec{\theta})$, la función de verosimilitud conecta los datos observados con dicho modelo, de tal modo que se puede hacer inferencias estadísticas sobre $\vec{\theta}$. Su importancia se resume en la ley de verosimilitud.
Ley de verosimilitud: Tome $X_1,\ldots,X_n$ como una muestra de variables aleatorias *iid* con *pdf* común $f(x;\vec{\theta})$ y espacio de parámetros Θ. Para $\vec{\theta}\in\Theta$, mientras mayor sea el valor de $L(\vec{\theta})$, el modelo probabilístico con parámetro $\vec{\theta}$ se ajusta más a los datos observados. Entonces, el grado con el cual la información de la muestra da soporte a un parámetro $\vec{\theta}_0\in\Theta$, en comparación con otro parámetro $\vec{\theta}_1\in\Theta$ es igual a la razón entre sus verosimilitudes

$$\Lambda(\vec{\theta}_0,\vec{\theta}_1)=\frac{L(\vec{\theta}_0)}{L(\vec{\theta}_1)}.$$

En particular, la información en la muestra coincide mejor con $\vec{\theta}_1$ que con $\vec{\theta}_0$ cuando $\Lambda<1$, y viceversa cuando $\Lambda>1$.
Para encontrar el parámetro $\vec{\theta}$ para el cual la función de verosimilitud alcanza su mayor valor se introduce la función de Score.

Definición 2.2.9 (Función de Score)**.** *Tome $X_1,\ldots,X_n$ como una muestra de variables aleatorias con función de verosimilitud $L(\vec{\theta})$, para $\vec{\theta}\in\Theta$. Si la función de verosimilitud logarítmica $\ell(\vec{\theta})$ es diferenciable, la función de Score se define como*

$$Sc(\vec{\theta})=\nabla_{\vec{\theta}}\,\ell(\theta),$$

de tal modo que una condición necesaria para que $\vec{\theta}\in\Theta$ sea un máximo es que $Sc(\theta)=\vec{0}$.

2.2.2. Estimación bayesiana

En la estimación paramétrica puntual bayesiana, el parámetro θ se trata como una variable aleatoria, con su propia *pdf* $\pi(\theta;\lambda)$. Esta distribución recibe el nombre de distribución previa y λ se llama hiperparámetro de la distribución previa. Cuando θ es una variable aleatoria, el modelo paramétrico $f(x;\theta)$ que genera la muestra aleatoria es la distribución condicional de X dado θ, por lo que la *pdf* de X se denotará como $f(x|\theta)$.
Las inferencias de θ en la aproximación bayesiana están basadas en la distribución de θ dados los valores observados de una muestra aleatoria $\vec{x}=(x_1,\ldots,x_n)$, llamada distribución posterior, y denotada por $f(\theta|\vec{x})$ [7].
Usando el teorema de Bayes y el teorema de la probabilidad total, en el caso de que θ es una variable aleatoria continua,

$$f(\theta|\vec{x})=\frac{f(\vec{x},\theta;\lambda)}{f_{\vec{X}}(\vec{x})}=\frac{f(\vec{x}|\theta)\pi(\theta;\lambda)}{\int_{\mathbb{S}_\theta}f(\vec{x}|\theta)\pi(\theta;\lambda)\mathrm{d}\theta}.$$

De modo similar, cuando θ es una variable aleatoria discreta,

$$f(\theta|\vec{x}) = \frac{f(\vec{x}|\theta)\pi(\theta;\lambda)}{\sum_{\theta\in\mathbb{S}_\theta} f(\vec{x}|\theta)\pi(\theta;\lambda)}.$$

La distribución posterior combina la información disponible de θ en la distribución previa y la función de verosimilitud para producir una distribución actualizada que contiene toda la información disponible de θ.
El siguiente teorema indica que la distribución posterior depende de la muestra $\vec{x}$ sólo bajo una estadística suficiente para θ.

Teorema 2.2.2. *Si $X_1, \ldots, X_n$ es una muestra de variables independientes iid con pdf común $f(x|\theta)$, S es una estadística suficiente para θ, y $\pi(\theta;\lambda)$ una distribución previa para θ, entonces la distribución posterior de θ dado $\vec{X}$ depende de la muestra sólo a través de una estadística suficiente S.*

Demostración. La prueba se hará únicamente para el una distribución previa discreta, puesto que en el caso continuo la prueba es similar, reemplazando la sumatoria por la integral.
Tome $X_1, \ldots, X_n$ como una muestra de variables aleatorias *iid* con *pdf* común $f(x|\theta)$, S como una estadística suficiente para θ, y $\pi(\theta;\lambda)$ como la distribución previa de θ. Entonces, como S es suficiente para θ, por el teorema de factorización de Fisher-Neyman, la distribución adjunta de $X_1, \ldots, X_n$ puede ser factorizada como $f(\vec{x}|\theta) = f(S;\theta)h(\vec{x})$, para algunas funciones g y h.
Entonces, la *pdf* de la distribución posterior es

$$f(\theta|\vec{x}) = \frac{f(\vec{x}|\theta)\pi(\theta)}{\sum_{\theta\in\mathbb{S}_\theta} f(\vec{x}|\theta)\pi(\theta)} = \frac{g(S;\theta)h(\vec{x})\pi(\theta)}{\sum_{\theta\in\mathbb{S}_\theta} g(S;\theta)h(\vec{x})\pi(\theta)} = \frac{g(S;\theta)\pi(\theta)}{\sum_{\theta\in\mathbb{S}_\theta} g(S;\theta)\pi(\theta)},$$

que es una función de S y θ únicamente. Entonces, $f(\theta|\vec{x})$ depende de $\vec{X}$ sólo a través de la estadística suficiente S. □

2.3. Cadenas de Markov Monte Carlo

La teoría comentada en la sección anterior estipula el número de muestras necesarias para tener una estadística suficiente, y da a conocer posibles funciones de costo para estimar la precisión y exactitud de un estimador dado. Por supuesto, se requiere un estimador para obtener los valores de los parámetros de un modelo probabilístico. Para ello, se usarán algoritmos relacionados con cadenas de Markov Monte Carlo (*MCMC* para abreviar).
Para empezar, recuerde que el teorema de Bayes estipula que, en el caso en el cual θ es una variable aleatoria continua:

$$P(\theta) = \frac{L(\theta)\pi(\theta)}{\int_{\theta\in\mathbb{S}_\theta} L(\theta)\pi(\theta)\mathrm{d}\theta} = \frac{L(\theta)\pi(\theta)}{Z}, \tag{2-1}$$

donde $P(\theta) := f(\theta|\vec{x})$ es la distribución posterior, $L(\theta) = f(\vec{x}|\theta)$ es la función de verosimilitud, $\pi(\theta)$ es la distribución previa y la constante Z se conoce como evidencia.
Considere ahora una función $f(\theta)$ del parámetro o parámetros del modelo que va a $\mathbb{R}$. El valor esperado de esta función sobre todo el soporte de θ es

$$\underbrace{E_P[f(\theta)]}_{\text{valor esperado respecto a } P} = \int_{\theta\in\mathbb{S}_\theta} f(\theta)P(\theta)\mathrm{d}\theta. \tag{2-2}$$

Esta integral se puede aproximar usando rejillas. Por ejemplo, en el caso en el que el modelo tiene un único parámetro $\theta \in \mathbb{S}_\theta \subseteq \mathbb{R}$, se puede construir una partición del soporte de θ (finita o infinita). Acá se supondrá un soporte acotado, pero el razonamiento para un soporte no acotado es similar. Considere entonces la partición $\mathcal{P} = \{\theta_1 < \theta_2 < \cdots < \theta_{n+1}\}$. Definiendo $\Delta\theta_i = \theta_{i+1} - \theta_i$ como el desplazamiento entre los elementos de la partición y $\overline{\theta_i} = \frac{\theta_{i+1}+\theta_i}{2}$ como el punto medio entre θ_{i+1} y θ_i, el valor esperado de $f(\theta)$ se puede aproximar de la siguiente forma:

$$E_P[f(\theta)] \approx \sum_{i=1}^{n} f(\overline{\theta_i})P(\overline{\theta_i})\Delta\theta_i. \tag{2-3}$$

La generalización a más dimensiones es directa: se descompone el soporte $\mathbb{S}_\theta \subseteq \mathbb{R}^N$ en n cuboides $N-$dimensionales. La contribución de cada uno de estos cuboides es proporcional al producto del peso $f(\overline{\theta_i})P(\overline{\theta_i})$ (donde $\overline{\theta_i}$ es el centro geométrico del $i-$ésimo cuboide) y al volumen

$$\Delta\theta_i = \prod_{j=1}^{N} \Delta\theta_{i,j},$$

donde $\Delta\theta_{i,j}$ es el ancho del $i-$ésimo cuboide en la j-ésima dimensión. Así, la ecuación para el valor esperado de $f(\theta)$ tiene la misma forma que (2-3). Más aún, escribiendo el valor esperado de la siguiente forma

$$E_P[f(\theta)] = \int_{\theta\in\mathbb{S}_\theta} f(\theta)P(\theta)\mathrm{d}\theta = \frac{\int_{\theta\in\mathbb{S}_\theta} f(\theta)P(\theta)\mathrm{d}\theta}{\underbrace{\int_{\theta\in\mathbb{S}_\theta} P(\theta)\mathrm{d}\theta}_{1}} = \frac{\int_{\theta\in\mathbb{S}_\theta} f(\theta)ZP(\theta)\mathrm{d}\theta}{\int_{\theta\in\mathbb{S}_\theta} ZP(\theta)\mathrm{d}\theta}$$

y tomando $\tilde{P}(\theta) = ZP(\theta)$, se puede aproximar la evidencia a través del mismo procedimiento, dado que $Z = \int_{\theta\in\mathbb{S}_\theta} \tilde{P}(\theta)\mathrm{d}\theta$:

$$E_P[f(\theta)] = \frac{\int_{\theta\in\mathbb{S}_\theta} f(\theta)\tilde{P}(\theta)\mathrm{d}\theta}{\int_{\theta\in\mathbb{S}_\theta} \tilde{P}(\theta)\mathrm{d}\theta} \approx \frac{\sum_{i=1}^{n} f(\overline{\theta_i})\tilde{P}(\overline{\theta_i})\Delta\theta_i}{\sum_{i=1}^{n} \tilde{P}(\overline{\theta_i})\Delta\theta_i}.$$

Esta sustitución a la distribución posterior no normalizada $\tilde{P}(\theta)$ es crucial para calcular valores esperados en la práctica, puesto que es posible calcular directamente $\tilde{P}(\theta) = L(\theta)\pi(\theta)$,

sin conocer Z. En el caso de que se necesite Z, se puede aproximar numéricamente a través de rejillas, como se vio anteriormente.
Una desventaja de la aproximación a este problema por medio de rejillas es el crecimiento exponencial de la cantidad de cubos necesarios para cubrir el soporte cuando aumenta su dimensión. Otra desventaja es que, al no conocer la forma de la distribución posterior, la contribución de cada parte de la rejilla puede ser altamente inexacta, dependiendo de su estructura. Si no se escogen bien los cuboides de la rejilla se podría terminar con muchos puntos localizados en regiones donde $\tilde{P}(\theta)$ o $f(\theta)\tilde{P}(\theta)$ son relativamente pequeños, lo que implicaría que su suma podría estar dominada por un pequeño número de puntos con pesos muy grandes. Para resolver esta desventaja, se busca incrementar la resolución de la rejilla en regiones donde la distribución posterior es grande y disminuirla en las otras regiones para evitar este efecto [8].
Para resolver esta última desventaja (teóricamente), se introducirá la idea de **media pesada muestral** de un conjunto de $\{f_1, \ldots, f_n\}$ observaciones con peso $\{\omega_1, \ldots, \omega_n\}$ de la siguiente forma:

$$f_{mean} = \frac{\sum_{i=1}^{n} \omega_i f_i}{\sum_{i=1}^{n} \omega_i}. \tag{2-4}$$

Si se toma $f_i := f(\overline{\theta_i})$ y $\omega_i := \tilde{P}(\overline{\theta_i})\Delta\theta_i$, se observa que el valor esperado se puede escribir de manera aproximada como una media pesada muestral. De esta forma, es necesario encontrar una forma eficiente de calcular dicha media pesada muestral para evitar la desventaja comentada. El **tamaño de muestra efectivo** n_{eff} es una primera aproximación, basada en el hecho de que no todas las muestras dan la misma información. En teoría, uno puede encontrar una manera de aproximar $E_P[f(\theta)]$ de mejor o igual forma de la que se tiene en términos de una media pesada muestral de tamaño n usando un número más pequeño de muestras n_{eff} si se es capaz de localizarlos más eficientemente.
De manera formal, se define n_{eff} del siguiente modo[4]:

$$n_{eff} = \frac{\left(\sum_{i=1}^{n} \omega_i\right)^2}{\sum_{i=1}^{n} \omega_i^2}. \tag{2-5}$$

Intuitivamente, el mejor caso es cuando todos los pesos son iguales ($\omega_i = \omega$), donde

$$n_{eff}^{best} = \frac{(n\omega)^2}{n\omega^2} = n,$$

y el peor caso es cuando todo el peso está concentrado en una única muestra, ($\omega_j = \omega$, para algún j y $\omega_i = 0$ en otro caso):

$$n_{eff}^{worst} = \frac{\omega^2}{\omega^2} = 1.$$

El mejor caso hace referencia a cuando todos los elementos de la rejilla tienen aproximadamente la misma contribución en la integral, mientras que el peor caso hace referencia a cuando la integral entera está contenida en un único cuboide de la rejilla.

Se debe procurar entonces que los pesos tiendan a ser una constante. En teoría, si se conoce la distribución posterior lo suficientemente bien, para n lo suficientemente grande, se podría ajustar $\Delta\theta_i$ para que los pesos $\omega_i = \tilde{P}(\overline{\theta_i})\Delta\theta_i$ sean uniformes a cierto nivel de precisión. Esta uniformidad ocurre cuando

$$\Delta\theta_i \propto \frac{1}{\tilde{P}(\overline{\theta_i})}, \text{ para todo } i.$$

Cuando $n \to \infty$, el espaciamiento $\Delta\theta$ cambia como función de θ. Esto motiva la definición de la densidad de puntos $Q(\theta)$, conocida como **distribución propuesta**, basada en la resolución variable $\Delta\theta(\theta)$ en la rejilla infinita como función de θ:

$$Q(\theta) \propto \frac{1}{\Delta\theta(\theta)}.$$

Usando $Q(\theta)$, se puede reescribir el valor esperado como

$$E_P[f(\theta)] = \frac{\int_{\theta\in\mathbb{S}_\theta} f(\theta)\tilde{P}(\theta)\mathrm{d}\theta}{\int_{\theta\in\mathbb{S}_\theta} \tilde{P}(\theta)\mathrm{d}\theta} = \frac{\int_{\theta\in\mathbb{S}_\theta} f(\theta)\frac{\tilde{P}(\theta)}{Q(\theta)}Q(\theta)\mathrm{d}\theta}{\int_{\theta\in\mathbb{S}_\theta} \frac{\tilde{P}(\theta)}{Q(\theta)}Q(\theta)\mathrm{d}\theta} = \frac{E_Q[f(\theta)\tilde{P}(\theta)/Q(\theta)]}{E_Q[\tilde{P}(\theta)/Q(\theta)]}.$$

En palabras, la rejilla de n elementos en el límite de infinita resolución se manifiesta en una nueva distribución $Q(\theta)$, con la cual se puede escribir el valor esperado $E_P[f(\theta)]$ en términos de los valores esperados $E_Q[f(\theta)\tilde{P}(\theta)/Q(\theta)]$ y $E_Q[\tilde{P}(\theta)/Q(\theta)]$. La practicidad de esto radica en el hecho de que se pueden calcular estas últimos valores esperados generando una muestra aleatoria de n elementos a partir de $Q(\theta)$.
Como no se sabe la forma exacta de $P(\theta)$ de antemano, se desconoce cuál rejilla proveerá un estimado óptimo para $E_P[f(\theta)]$. Una de las formas de calcular este valor esperado es usando la distribución propuesta $Q(\theta)$, generando muestras a partir de ella. Esto motiva a escoger $Q(\theta)$ de manera que se puedan generar muestras de manera fácil y directa. Generando n muestras $\{\theta_1, \ldots, \theta_n\}$ de esta distribución, con pesos asociados q_i y definiendo

$$f(\theta_i) = f_i, \quad \tilde{\omega_i} := \tilde{\omega}(\theta_i) = \tilde{P}(\theta_i)/Q(\overline{\theta_i}),$$

el valor esperado se puede aproximar como

$$E_P[f(\theta)] = \frac{E_Q[f(\theta)\tilde{P}(\theta)/Q(\theta)]}{E_Q[\tilde{P}(\theta)/Q(\theta)]} \approx \frac{\sum_{i=1}^n f_i\tilde{\omega_i}q_i}{\sum_{i=1}^n \tilde{\omega_i}q_i}.$$

Si además se toma $Q(\theta)$ de modo que las muestras sean *iid*, los correspondientes pesos q_i se reducen a $1/n$, de manera que

$$E_P[f(\theta)] \approx \frac{n^{-1}\sum_{i=1}^n f_i\tilde{\omega_i}}{n^{-1}\sum_{i=1}^n \tilde{\omega_i}}.$$

El denominador de la última expresión es nuevamente una aproximación directa de la evidencia,

$$Z = \int_{\theta \in \mathbb{S}_\theta} \tilde{P}(\theta) \mathrm{d}\theta \approx n^{-1} \sum_{i=1}^{n} \tilde{\omega_i}.$$

De este modo, los pasos a seguir para calcular el valor esperado son los siguientes.

1. Se debe generar n muestras *iid* $\{\theta_1, \ldots, \theta_n\}$ a partir de $Q(\theta)$.
2. Se calculan sus correspondientes pesos $\tilde{\omega_i} = \tilde{P}(\theta_i)/Q(\theta_i)$.
3. Se estima $E_P[f(\theta)]$ aproximando $E_Q[f(\theta)\tilde{P}(\theta)/Q(\theta)]$ y $E_Q[\tilde{P}(\theta)/Q(\theta)]$ a través de los pesos de las muestras.

Los métodos *MCMC* buscan generar muestras de tal modo que los pesos asociados $\{\tilde{\omega_1}, \ldots, \tilde{\omega_n}\}$ son constantes. $Q(\theta)$ juega un papel fundamental para lograr este cometido, y para ilustrar, considere los siguientes casos.

- Tome $Q(\theta) = Q^{unif}(\theta)$, definida sobre un cuboide de volumen V, de la siguiente manera

$$Q^{unif}(\theta) = \begin{cases} 1/V, & \text{si } \theta \text{ está dentro del cuboide o} \\ 0 & \text{en otro caso.} \end{cases}$$

  Los pesos en este caso serán proporcionales a la distribución posterior:

$$\tilde{\omega_i}^{unif} = \frac{\tilde{P}(\theta_i)}{Q^{unif}(\theta_i)} = V\tilde{P}(\theta_i) \propto P(\theta_i).$$

- Tome $Q(\theta) = Q^{prior}(\theta) = \pi(\theta)$ como la distribución previa de θ. Los pesos en este caso se pueden calcular mediante la función de verosimilitud.

$$\tilde{\omega_i}^{prior} = \frac{\tilde{P}(\theta_i)}{Q^{prior}(\theta_i)} = \frac{ZP(\theta_i)}{\pi(\theta_i)} = \frac{L(\theta_i)\pi(\theta_i)}{\pi(\theta_i)} = L(\theta_i).$$

- Tome $Q(\theta) = Q^{post}(\theta) = P(\theta)$ como la distribución posterior de θ, de modo que los pesos serán constantes e iguales a la evidencia Z:

$$\tilde{\omega_i}^{post} = \frac{\tilde{P}(\theta_i)}{Q^{post}(\theta_i)} = \frac{ZP(\theta_i)}{P(\theta_i)} = Z.$$

Siguiendo la idea del último caso, si uno desea que sus pesos sean constantes, se debe procurar que $Q(\theta)$ sea lo más cercana posible a $P(\theta)$. Los modelos *MCMC* buscan generar muestras con pesos proporcionales a la distribución posterior, para obtener un estimado óptimo del valor esperado.
Los modelos *MCMC* logran esto creando una cadena de valores de parámetros correlacionados $\{\theta_1 \to \cdots \to \theta_n\}$ al cabo de n iteraciones de tal modo que el número $m(\theta)$ de iteraciones

hechas en cada región particular δ_θ, centrada en θ es proporcional a la densidad posterior $P(\theta)$. En otras palabras, la densidad de muestras generadas por el modelo *MCMC*

$$\rho(\theta) := \frac{m(\theta)}{n}$$

en la posición θ integrada sobre δ_θ es aproximadamente

$$\int_{\theta\in\delta_\theta} P(\theta)\mathrm{d}\theta \approx \int_{\theta\in\delta_\theta} \rho(\theta)\mathrm{d}\theta \approx n^{-1}\sum_{j=1}^{n} \mathbb{1}[\theta_j \in \delta_\theta],$$

donde $\mathbb{1}[\cdot]$ es la función indicadora, equivalente a 1 si la condición a la que está siendo evaluada es verdadera, y cero si es falsa. La densidad de muestras se puede aproximar de esta forma contando el número de muestras dentro de δ_θ y normalizando por el número total de muestras n.
Cuando $n \to \infty$, se garantiza que $\rho(\theta) \to P(\theta)$ en cualquier punto θ [2]. Con una aproximación razonable de $\rho(\theta)$, se pueden usar las muestras $\{\theta_1 \to \cdots \to \theta_n\}$ generadas por $\rho(\theta)$ para estimar la evidencia

$$Z = \int_{\theta\in\mathbb{S}_\theta} \frac{\tilde{P}(\theta)}{\rho(\theta)}\rho(\theta)\mathrm{d}\theta = E_\rho[\tilde{P}(\theta)/\rho(\theta)] \approx n^{-1}\sum_{i=1}^{n} \frac{\tilde{P}(\theta_i)}{\rho(\theta_i)}.$$

Además, como el modelo *MCMC* produce una serie de n muestras de la distribución posterior, el valor esperado de $f(\theta)$ se reduce a

$$E_P[f(\theta)] \approx \frac{n^{-1}\sum_{i=1}^{n} f_i\tilde{\omega}_i}{n^{-1}\sum_{i=1}^{n} \tilde{\omega}_i} = \frac{n^{-1}\sum_{i=1}^{n} f_i}{n^{-1}\sum_{i=1}^{n} 1} = n^{-1}\sum_{i=1}^{n} f_i,$$

que es la expresión del promedio aritmético de los valores $f_i = f(\theta_i)$.

2.3.1. Algoritmo de Metropolis-Hastings

Se desea generar muestras $\theta_i \to \theta_{i+1}$ de modo que la distribución de las muestras finales $\rho(\theta)$ sea estacionaria cuando $n \to \infty$ (que converja) y sea igual a $P(\theta)$. La primera condición se puede satisfacer usando el **balance detallado**, que refiere a la idea de que la probabilidad sea conservada cuando uno se mueve de una posición a otra (es decir, el proceso es reversible). Formalmente, esto implica que

$$M(\theta_{i+1}|\theta_i)M(\theta_i) = M(\theta_{i+1},\theta_i) = M(\theta_i|\theta_{i+1})M(\theta_{i+1}),$$

donde $M(\theta_{i+1}|\theta_i)$ es la probabilidad de moverse de θ_i a θ_{i+1} y $M(\theta_{i+1}|\theta_i)$ es la probabilidad de moverse de θ_{i+1} a θ_i. Reescribiendo esta última igualdad:

$$\frac{M(\theta_{i+1}|\theta_i)}{M(\theta_i|\theta_{i+1})} = \frac{M(\theta_{i+1})}{M(\theta_i)} = \frac{P(\theta_{i+1})}{P(\theta_i)}. \tag{2-6}$$

La última desigualdad hace referencia al hecho de que la distribución con la que se buscan generar las muestras es la posterior, $P(\theta)$.
Es necesario construir un procedimiento que permita moverse a una nueva posición calculando esta probabilidad M. Para ello, se propone una nueva posición $\theta_i \to \theta'_{i+1}$ usando la distribución propuesta $Q(\theta'_{i+1}|\theta_i)$. Se decide si aceptar $(\theta_{i+1} = \theta'_{i+1})$ o rechazar $(\theta_{i+1} = \theta_i)$ esta nueva posición con una **probabilidad de transición** $T(\theta'_{i+1}|\theta_i)$. Combinando ambas distribuciones, se obtiene la probabilidad de moverse a una nueva posición

$$M(\theta_{i+1}|\theta_i) = Q(\theta_{i+1}|\theta_i)T(\theta_{i+1}|\theta_i).$$

El problema estaría completamente determinado si se conoce la probabilidad de transición $T(\theta'_{i+1}|\theta_i)$. Para encontrarla, se hace uso de la condición de balance detallado (2-6).

$$\frac{T(\theta_{i+1}|\theta_i)}{T(\theta_i|\theta_{i+1})} = \frac{P(\theta_{i+1})}{P(\theta_i)}\frac{Q(\theta_i|\theta_{i+1})}{Q(\theta_{i+1}|\theta_i)}.$$

El criterio de Metropolis [5]:

$$T(\theta_{i+1}|\theta_i) = \text{mín}\left[1, \frac{P(\theta_{i+1})}{P(\theta_i)}\frac{Q(\theta_i|\theta_{i+1})}{Q(\theta_{i+1}|\theta_i)}\right]$$

satisface esta condición, por lo que el algoritmo para generar la muestra es el siguiente:

1. Se propone una nueva posición $\theta_i \to \theta'_{i+1}$, generando una muestra de la distribución propuesta $Q(\theta'_{i+1}|\theta_i)$.

2. Se calcula la probabilidad de transición $T(\theta'_{i+1}|\theta_i)$.

3. Se genera un número aleatorio u_{i+1} a partir de una distribución uniforme entre 0 y 1.

4. Si $u_{i+1} \leq T(\theta'_{i+1}|\theta_i)$, se acepta el movimiento y se toma $\theta_{i+1} = \theta_i$. Si $u_{i+1} > T(\theta'_{i+1}|\theta_i)$, se rechaza el movimiento y se toma $\theta_{i+1} = \theta_i$.

5. Se incrementa $i = i + 1$ y se repite el proceso.

Bibliografía

[1] L. Blanco, V. Arunachalam, and Dharmaraja S. *Introduction to Probability and Stochastic Processes with Applications*. Wiley, 2012.

[2] Steve Brooks, Andrew Gelman, and Galin L. Jones. *Handbook of Markov chain Monte Carlo*. CRC Press, 2011.

[3] S. Carroll. *Lecture Notes on General Relativity*. Institute of Theoretical Physics, University of California, 1997.

[4] Leslie Kish. *Survey sampling*. Wiley, 1995.

[5] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[6] J. Munkres. *Topology*. Pearson Education Limited, 2014.

[7] R. Rossi. *Mathematical statistics: an introduction to likelihood based inference*. John Wiley Sons, Inc., 2018.

[8] Joshua S. Speagle. A conceptual introduction to markov chain monte carlo methods, 2019.

[9] K. Thorne, J. Wheeler, and C. Misner. *Gravitation*. W.H. Freeman and Company, 1973.