

Simulación Estocástica

Apuntes de clase.

Bruno Hernández: `bhernandez@dim.uchile.cl`

2019

Contents

1	Notación	3
2	Convergencia en Ley	4
2.1	Convergencia débil sobre espacios métricos	4
2.1.1	El Teorema Portmanteau	7
2.1.2	El Teorema del Mapeo.	10
2.2	Tensión (Tightness)	11
2.2.1	Relativamente compacto.	11
3	Función Característica	13
4	Teorema Central del Límite	20
4.1	Demostración del T.C.L.	21
5	Método de Monte Carlo y Simulación	22
5.1	Descripción del método	23
5.2	Simulación de variables aleatorias	24
5.3	Método de Aceptación-Rechazo	26
5.4	Reducción de varianza	27
5.4.1	Muestreo preferencial	27
5.4.2	Variable de control	29
5.4.3	Variables antitéticas	29
5.4.4	Método de Estratificación	30
6	Cadenas de Markov	31
6.1	Definiciones y propiedades elementales	31
6.2	Propiedad de Markov fuerte	33
6.3	Estados Recurrentes y Transientes	34
6.4	Casos Irreducibles y Recurrentes	36
6.5	El caso aperiódico	40
6.6	Cadenas de Markov Reversibles	46
6.7	Razón de convergencia al equilibrio	48
6.7.1	El caso reversible de finitos estados	48
6.7.2	El caso general	49
6.8	Estadísticas en cadenas de Markov	49
7	Algoritmos estocásticos usando cadenas de Markov	51
7.1	Markov chain Monte Carlo (M.C.M.C.)	51
7.2	Simulación de la distribución invariante	53
7.2.1	Simulación perfecta.	53
7.3	Simulated Annealing	55
8	Movimiento Browniano, proceso de difusión y aplicaciones	57
8.1	Movimiento Browniano	57

9	Martingalas	61
9.1	Tiempos de parada (t.d.p.)	63

1 Notación

- *i.i.d.*: 'Independientes e idénticamente distribuídas'.
- $X \perp\!\!\!\perp Y$: "La variable X es independiente a la variable Y ".
- (E, d) : Espacio métrico E , con métrica d .
- $\mathcal{B}(E)$: Conjunto de borelianos de E .
- $C_b(E) := \{f : E \rightarrow \mathbb{R} \mid f \text{ es continua y acotada}\}$
- $\mathcal{P}(E)$: Espacio de las medidas de probabilidades sobre E .
- *c.s.*: "Casi seguramente"
- **T.C.D.**: "Teorema de convergencia dominada".
- $\mathcal{N}(\mu, \sigma)$: "Distribución normal de media μ y desviación estándar de σ ", se hará un abuso de notación ya que también simbolizará la variable aleatoria que distribuya de esta forma.
- *M.M.C.*: "Método de Monte Carlo".
- *C.M.H.*: "Cadena de Markov homogénea".
- *M.C.M.C.*: "Markov chain Monte Carlo".
- *M.B.*: "Movimiento Browniano".
- **t.d.p.**: "Tiempo de parada".

2 Convergencia en Ley

Partiremos este capítulo enunciando uno de los teoremas más importante de la teoría de probabilidades, el teorema del límite central.

Teorema 1 (Teorema Central del Límite.) Dadas $(X_n)_{n \in \mathbb{N}}$ variables aleatorias i.i.d., con $\mu = \mathbb{E}(X_1)$ y $\sigma^2 = \text{Var}(X_1)$, finitas. Para el promedio n -ésimo:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad (2.0.1)$$

Se define la variable aleatoria Z_n como:

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \quad (2.0.2)$$

Entonces Z_n converge "**EN LEY**" a una variable normal estándar $\mathcal{N}(0, 1)$ a medida que n tiende a infinito.

Los resultados y usos que se desprenden de este teorema son bastante variados; desde el cálculo de probabilidades para una muestra aleatoria simple, estimación de parámetros, hasta nos permite el uso de test de hipótesis para distribuciones normales en ocasiones en que desconocemos por completo la distribución de los datos.¹

Y no es que sea raro encontrar este tipo de usos, puesto que el teorema no nos pide absolutamente nada con respecto al tipo de distribución de las variables aleatorias, dando la posibilidad de que éstas sean continuas, discretas o cualesquiera que se nos pudiesen ocurrir en diversos espacios de trabajo. Sin embargo el resultado es el mismo, convergencia a una distribución absolutamente continua.

Si Z_n puede tomar varias formas dada la colección de variables $(X_n)_{n \in \mathbb{N}}$, ¿por qué converge a una distribución continua? Para responder esta pregunta introduciremos las nociones de *convergencia débil de medida de probabilidad*.

2.1 Convergencia débil sobre espacios métricos

Para el conjunto de medidas de probabilidad sobre el espacio métrico E , es decir; $\mathcal{P}(E)$, definiremos la siguiente operación.

Definición 1 Para toda $\mu \in \mathcal{P}(E)$ y para toda $f : E \rightarrow \mathbb{R}$ función medible:

$$\langle \mu, f \rangle = \int_E f d\mu \quad (2.1.1)$$

Además diremos que una sucesión $(\mu_n)_{n \in \mathbb{N}}$ de elementos de $\mathcal{P}(E)$ converge **DÉBILMENTE** a $\mu \in \mathcal{P}(E)$ si:

$$\langle \mu_n, f \rangle \xrightarrow{n \rightarrow \infty} \langle \mu, f \rangle \quad \forall f \in C_b(E)$$

En tal caso anotamos: $\mu_n \Rightarrow \mu$.

Ejemplo 1: Tomemos el espacio métrico $E = \mathbb{R}$ con su métrica usual $d(x, y) = |x - y|$, y sea la sucesión de medidas de probabilidad $\mu_n = \delta_{\frac{1}{n}}$, donde δ_a representa la "Delta de Dirac" en el punto a :

$$\delta_a(x) = \begin{cases} +\infty & \text{si } x = a \\ 0 & \text{si } x \neq a \end{cases}$$

Veremos que esta distribución converge débilmente a la distribución δ_0 . En efecto, sea $f \in C_b(\mathbb{R})$, tenemos que:

$$\langle \mu_n, f \rangle = \int_{\mathbb{R}} f(x) \mu_n(dx) = \int_{\mathbb{R}} f(x) \delta_{\frac{1}{n}}(x) dx = f(1/n)$$

¹Todos estos usos están restringidos bajo aproximaciones.

Puesto que la función f es continua en todo \mathbb{R} , entonces:

$$f(1/n) \xrightarrow{n \rightarrow \infty} f(0) = \langle \mu, f \rangle$$

Por lo tanto, $\mu_n \Rightarrow \mu$.

Dado que el argumento utilizado en el ejemplo se sostuvo fuertemente de la continuidad de la función f , en general se tendrá que para cada sucesión de reales $(x_n)_n$ que converja a un punto \bar{x} , se tendrá la convergencia de las medidas de Dirac: $\delta_{x_n} \Rightarrow \delta_{\bar{x}}$. Se puede demostrar que esto es una doble implicancia, es decir; la convergencia débil de las medidas de Dirac implican la convergencia en \mathbb{R} de la sucesión de los puntos.

Para traducir el *Teorema Central del Límite* en conceptos de convergencia de medidas de probabilidad debemos asociar el concepto de *variable aleatoria* en función de la medida de probabilidad asociada.

Recordemos que una **variable aleatoria** a valores en E es una función $X : \Omega \rightarrow E$ medible, donde $(\Omega, \mathcal{F}, \mathbb{P})$ es un espacio de probabilidad (este espacio se encuentra usualmente implícito en la mayoría de las definiciones y cálculos de probabilidades).

Su **LEY** o **DISTRIBUCIÓN**, denotada como $\mathcal{L}(X) \in \mathcal{P}(E)$, se define como:

$$\mathcal{L}(X)(A) := \mathbb{P}(X^{-1}(A)) \quad \forall A \in \mathcal{B}(E) \quad (2.1.2)$$

Definición 2 Una sucesión $(X_n)_{n \in \mathbb{N}}$ de variables aleatorias en E se dice que converge **EN LEY** o **EN DISTRIBUCIÓN** a una variable aleatoria X si $(\mathcal{L}(X_n))_{n \in \mathbb{N}}$ converge débilmente a $\mathcal{L}(X)$, es decir:

$$\mathbb{E}[f(X_n)] \xrightarrow{n \rightarrow \infty} \mathbb{E}[f(X)] \quad \forall f \in C_b$$

En este caso anotamos: $X_n \xrightarrow{\mathcal{L}} X$ o $X_n \xrightarrow{d} X$.

Notamos que esta definición es idéntica a la anterior pero expresada en términos de la esperanza de la función f en vez de la integral de la función bajo la medida de probabilidad.

Ejemplo 2: Sea el espacio métrico $E = \mathbb{R}$ y una sucesión de variables aleatorias *i.i.d.* $(X_n)_{n \in \mathbb{N}}$, tal que para cada natural n , la variable tiene una distribución uniforme en el intervalo $(-1/n, 1/n)$, es decir; $X_n \sim \text{unif}(-\frac{1}{n}, \frac{1}{n}) \forall n \in \mathbb{N}$.

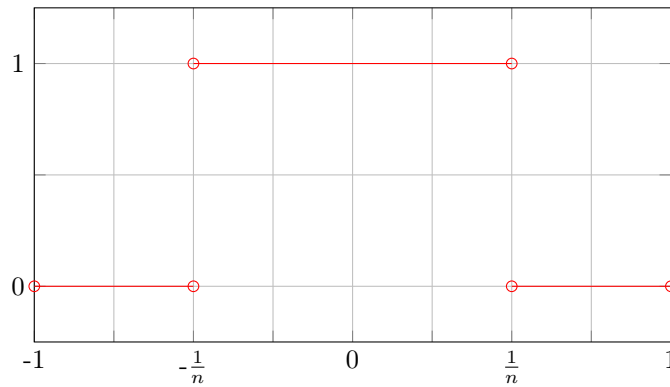


Figure 1: Densidad uniforme $(-1/n, 1/n)$.

Sea $f \in C_b(\mathbb{R})$ una función continua y acotada, entonces la esperanza $\mathbb{E}(f(X_n))$ está dada por:

$$\mathbb{E}(f(X_n)) = \int_{\mathbb{R}} f(x) \mu_n(dx)$$

donde $\mu_n(dx)$ está dado por la densidad de la distribución de X_n , es decir la uniforme:

$$\mu_n(dx) = \frac{n}{2} \mathbb{1}_{(-\frac{1}{n}, \frac{1}{n})}(x) dx$$

Luego:

$$\mathbb{E}(f(X_n)) = \int_{\mathbb{R}} f(x) \frac{n}{2} \mathbb{1}_{(-\frac{1}{n}, \frac{1}{n})}(x) dx = \frac{n}{2} \int_{-\frac{1}{n}}^{\frac{1}{n}} f(x) dx$$

Haciendo uso del **teorema del valor medio** en integrales y de la continuidad de f , sabemos de la existencia de un valor $\xi_n \in (-\frac{1}{n}, \frac{1}{n})$ tal que la integral anterior es de la forma:

$$\int_{-\frac{1}{n}}^{\frac{1}{n}} f(x) dx = \frac{2}{n} f(\xi_n)$$

Así,

$$\mathbb{E}(f(X_n)) = \frac{n}{2} \int_{-\frac{1}{n}}^{\frac{1}{n}} f(x) dx = f(\xi_n)$$

Por lo tanto la convergencia de $\mathbb{E}(f(X_n))$ estará dada por la convergencia de la nueva sucesión de puntos $(\xi_n)_{n \in \mathbb{N}}$. Para ver la convergencia de ξ_n notemos que:

- $\forall m \leq n$: $(-\frac{1}{n}, \frac{1}{n}) \subset (-\frac{1}{m}, \frac{1}{m})$, por lo tanto:

$$\xi_n \in \bigcap_{m=1}^n \left(-\frac{1}{m}, \frac{1}{m}\right)$$

Así la sucesión ξ_n es acotada.

- Luego, la sucesión tiene al menos un punto de acumulación. Supongamos que $\bar{\xi}$ es punto de acumulación de $(\xi_n)_n$, por lo dicho anteriormente debería suceder que:

$$\bar{\xi} \in \bigcap_{n \geq 1} \left(-\frac{1}{n}, \frac{1}{n}\right) = \{0\}$$

Finalmente se deduce que $\bar{\xi} = 0$.

Nuevamente, aplicando la continuidad de f :

$$\mathbb{E}(f(X_n)) = f(\xi_n) \rightarrow f(0) = \langle \delta_0, f \rangle = \mathbb{E}(f(0))$$

Así concluimos que:

- $\mu_n \Rightarrow \delta_0$, donde μ_n estaba dada por las distribuciones uniformes.
- $X_n \xrightarrow{\mathcal{L}} X$, donde $X \equiv 0$.

Ejemplo 3: Nuevamente tomamos como ejemplo $E = \mathbb{R}$, y definimos la siguiente medida:

$$\mu_n = \frac{1}{n+1} \sum_{k=0}^n \delta_{k/n}$$

De forma alternativa, definiendo $\mu_n = \mathcal{L}(X_n)$, donde X_n corresponde al experimento de escoger un valor al azar del conjunto $\{0, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n}{n}\}$ y sea $\mu = \mathcal{L}(X)$, con $X \sim \text{unif}(0, 1)$. Entonces $\forall f \in C_b(\mathbb{R})$:

$$\langle \mu_n, f \rangle = \frac{1}{n+1} \sum_{k=0}^n \langle \delta_{k/n}, f \rangle = \frac{1}{n+1} \sum_{k=0}^n f(k/n)$$

Notando que, para cada $n \in \mathbb{N}$ el conjunto $\{\frac{k}{n} | k = 0, \dots, n\}$ representa un refinamiento del intervalo $[0, 1]$ y f es una función acotada y continua en ese intervalo (por lo tanto es *Riemann-integrable*), entonces la suma anterior converge por ser suma de Riemann, y:

$$\langle \mu_n, f \rangle = \frac{1}{n+1} \sum_{k=0}^n f(k/n) \xrightarrow{n \rightarrow \infty} \int_0^1 f(x) dx = \langle \mu, f \rangle$$

Luego $\mu_n \Rightarrow \mu$, i.e. $X_n \xrightarrow{\mathcal{L}} X$.

2.1.1 El Teorema Portmanteau

El siguiente teorema nos proporciona útiles equivalencias a la convergencia débil; cada una de ellas sirve como una definición. Un conjunto $A \in \mathcal{B}(E)$ cuya frontera ∂A satisfaga que $\mu(\partial A) = 0$ se dirá conjunto μ -continuo (Note que ∂A es cerrado, por lo tanto pertenece a $\mathcal{B}(E)$).

Teorema 2 (Portmanteau) Sean $\mu_n, \mu \in \mathcal{P}(E)$. Las siguientes proposiciones son equivalentes:

- i) $\mu_n \Rightarrow \mu$
- ii) $\langle \mu_n, f \rangle \rightarrow \langle \mu, f \rangle$ para toda función f acotada y uniformemente continua.
- iii) $\langle \mu_n, f \rangle \rightarrow \langle \mu, f \rangle$ para toda función f , Lipschitz.
- iv) $\limsup_n \mu_n(F) \leq \mu(F) \forall F \subset E$ cerrado.
- v) $\liminf_n \mu_n(G) \geq \mu(G) \forall G \subset E$ abierto.
- vi) $\lim \mu_n(A) = \mu(A)$ para todo A , conjunto μ -continuo.

Notemos que, en el caso del **Ejemplo 1**, la desigualdad del punto iii) se satisface estrictamente si $F = \{0\}$, de la misma forma en el punto iv) si $G = \{0\}^c$. Si tomamos $A = \{0\}$, la convergencia no se alcanza en el punto vi), pero esto no contradice el teorema, ya que la medida límite de $\partial\{0\} = \{0\}$ es 1, no 0.

Demostración (Portmanteau): Dado que las funciones uniformemente continuas, en particular son continuas, la implicancia i) \rightarrow ii) se desprende trivialmente de la definición de convergencia débil en i). Así mismo la implicancia ii) \rightarrow iii) considerando las funciones Lipschitz. Veamos la implicancia iii) \rightarrow iv):

Sea F un conjunto cerrado. $\forall \epsilon > 0$, sea:

$$f(x) = \max \left(1 - \frac{d(x, F)}{\epsilon}, 0 \right) \quad (2.1.3)$$

Donde $d(x, F)$ está dado por la métrica del espacio E . Se puede probar que la función f definida de esta forma es $\frac{1}{\epsilon}$ -Lipschitz y vale 1 en todo F . Además:

$$\mathbb{1}_F \leq f \leq \mathbb{1}_{F^\epsilon} \quad \text{donde } F^\epsilon := \{x \in E \mid d(x, F) < \epsilon\} \quad (2.1.4)$$

Luego, integrando (bajo μ_n) en las dos primeras partes de la desigualdad y aplicando límite superior:

$$\limsup \mu_n(F) \leq \limsup \langle \mu_n, f \rangle$$

Pero, como habíamos dicho, al ser f una función Lipschitz, por la proposición iii); $\langle \mu_n, f \rangle$ converge, por lo tanto el límite superior en realidad es el límite, más aún, converge en términos de μ :

$$\limsup \langle \mu_n, f \rangle = \lim \langle \mu_n, f \rangle = \langle \mu, f \rangle \quad (2.1.5)$$

Ahora, nuevamente por la expresión en 2.1.4, integrando la segunda desigualdad (ahora bajo μ):

$$\langle \mu, f \rangle \leq \mu(F^\epsilon)$$

En resumen, tenemos que:

$$\limsup \mu_n(F) \leq \langle \mu, f \rangle \leq \mu(F^\epsilon)$$

Al ser F cerrado y $F^\epsilon \searrow F$, por continuidad de la medida se tiene que $\mu(F^\epsilon) \searrow \mu(F)$ al momento que $\epsilon \rightarrow 0$ concluyendo el resultado.

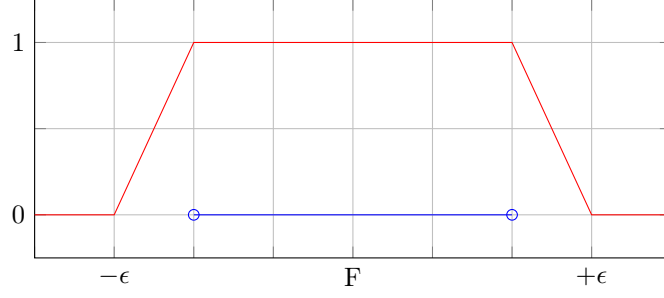


Figure 2: Esquema función f en 2.1.3

Es directo ver que la proposición $iv)$ y $v)$ son equivalentes, así que la implicancia $iv) \rightarrow v)$ es directa de tomar complemento y utilizar que μ_n y μ son medidas de probabilidad.

Veamos que $iv) \wedge v)$ implican a $vi)$: Sea A un conjunto μ -continuo. Dado que \bar{A} es cerrado, A° es abierto, ocupando la proposición $iv)$ y $A^\circ \subset A \subset \bar{A}$:

$$\mu(\bar{A}) \geq \limsup \mu_n(\bar{A}) \geq \limsup \mu_n(A) \geq \liminf \mu_n(A) \geq \liminf \mu_n(A^\circ) \geq \mu(A^\circ)$$

Además, como A es μ -continuo: $\mu(\partial A) = 0$, entonces $\mu(\bar{A}) = \mu(A^\circ) = \mu(A)$, por lo tanto todas las desigualdades anteriores son, en realidad, igualdades. Concluimos así que:

- $\liminf \mu_n(A) = \limsup \mu_n(A)$ por lo tanto el límite existe.
- $\mu(A) \geq \lim \mu_n(A) \geq \mu(A)$, por lo tanto $\lim \mu_n(A) = \mu(A)$.

Por último, para finalizar, veremos la implicancia $vi) \rightarrow i)$:

Sea $f \in C_b(E)$, sin pérdida de generalidad podemos suponer que $0 \leq f \leq 1$. En caso contrario, como f es una función acotada, su supremo e ínfimo son finitos, por lo tanto podríamos definir una normalización para f :

$$\tilde{f} = \frac{f - \inf_x f}{\sup_x f - \inf_x f}$$

Para la siguiente demostración haremos uso del siguiente resultado:

Lema 1 Sea g una función positiva, entonces:

$$\mathbb{E}_\mu(g(X)) = \langle \mu, g \rangle = \int_0^\infty \mu(g > t) dt \quad (2.1.6)$$

Demostración (Lema 1): La primera igualdad es la definición de esperanza:

$$\mathbb{E}_\mu(g(X)) = \int_0^\infty g(x) \mu(dx)$$

Al mismo tiempo, podemos reescribir la función g de la forma:

$$\mathbb{E}_\mu(g(X)) = \int_0^\infty \left(\int_0^{g(x)} dt \right) \mu(dx)$$

Ahora podemos extender la integral en dt a todo el intervalo positivo mediante una indicatriz que anule los valores superiores a $g(x)$, esto es:

$$\mathbb{E}_\mu(g(X)) = \int_0^\infty \left(\int_0^\infty \mathbb{1}_{\{g > t\}} dt \right) \mu(dx) = \int_0^\infty \int_0^\infty \mathbb{1}_{\{g > t\}} dt \mu(dx)$$

Ocupando el teorema de Fubini podemos intercambiar el orden de las integrales, entonces:

$$\mathbb{E}_\mu(g(X)) = \int_0^\infty \int_0^\infty \mathbb{1}_{\{g > t\}} dt \mu(dx) = \int_0^\infty \int_0^\infty \mathbb{1}_{\{g > t\}} \mu(dx) dt = \int_0^\infty \mu(g > t) dt$$

■

Entonces, volviendo a la demostración del teorema, tenemos que:

$$\langle \mu, f \rangle = \int_0^1 \mu(f > t) dt$$

$$\langle \mu_n, f \rangle = \int_0^1 \mu_n(f > t) dt$$

Donde el intervalo de integración lo podemos acotar, puesto que definimos el rango de la función f en $[0, 1]$. Además, como f es continua, es fácil ver que:

$$\partial\{f > t\} \subset \{f = t\}$$

Notemos que $\mu(f = t) = 0$ salvo para, a lo más numerables, valores de t 's. De lo contrario, la medida μ no sería finita. Luego, $dt - c.s.$ tenemos:

$$\mu(\partial\{f > t\}) \leq \mu(\{f = t\}) = 0$$

Es decir: de manera $dt - c.s.$ $\{f > t\}$ es μ -continuo. Por la proposición *vi*):

$$\mu_n(f > t) \xrightarrow{n} \mu(f > t) \quad dt - c.s.$$

Finalmente, ocupando el **T.C.D.**:

$$\langle \mu_n, f \rangle = \int_0^\infty \mu_n(f > t) dt \xrightarrow{\text{T.C.D.}} \int_0^\infty \mu(f > t) dt = \langle \mu, f \rangle$$

■

Habiendo establecido anteriormente, la equivalencia entre la convergencia débil de medidas de probabilidad y la convergencia en LEY de variables aleatorias, podemos enunciar el Teorema Portmanteau para variables aleatorias.

Teorema 3 (Portmanteau para variables aleatorias.) Sean $(X_n)_{n \in \mathbb{N}}$, X variables aleatorias en (E, d) espacio métrico. Las siguientes proposiciones son equivalentes:

- i) $X_n \xrightarrow{\mathcal{L}} X$
- ii) $\mathbb{E}(f(X_n)) \xrightarrow{n} \mathbb{E}(f(X))$ para toda función f acotada y uniformemente continua.
- iii) $\mathbb{E}(f(X_n)) \xrightarrow{n} \mathbb{E}(f(X))$ para toda función f Lipschitz.
- iv) $\limsup \mathbb{P}(X_n \in F) \leq \mathbb{P}(X \in F) \quad \forall F \subset E$ cerrado.
- v) $\liminf \mathbb{P}(X_n \in G) \geq \mathbb{P}(X \in G) \quad \forall G \subset E$ abierto.
- vi) $\lim \mathbb{P}(X_n \in A) = \mathbb{P}(X \in A)$ para todo $A \in \mathcal{B}(E)$ conjunto μ -continuo ($\mathbb{P}(X \in \partial A) = 0$).

Observación 1: Tomando $E = \mathbb{R}$, la condición *vi*) del teorema suele ser interesante, como en el siguiente ejemplo:

Sabemos de ejemplos anteriores que:

$$\delta_{1/n} \Rightarrow \delta_0$$

Sin embargo, para el conjunto $A = (0, \infty)$, $\forall n \in \mathbb{N}$:

$$\delta_{1/n}(A) = 1$$

$$\delta_0(A) = 0$$

Es decir que $\delta_{1/n}(A)$ no converge a $\delta_0(A)$, sin embargo esto no contradice la condición *vi*), pues el conjunto A no es δ_0 -continuo.

$$\delta_0(\partial A) = \delta_0(\{0\}) = 1$$

2.1.2 El Teorema del Mapeo.

Supongamos que h es una función de (E, d) a otro espacio métrico, (\hat{E}, \hat{d}) . Si h es $\mathcal{B}(E)/\mathcal{B}(\hat{E})$ -medible, entonces cada medida de probabilidad $\mu \in \mathcal{P}(E)$ induce a otra medida de probabilidad $\mu h^{-1} \in \mathcal{P}(\hat{E})$, definida de la manera usual, como $\mu h^{-1}(A) = \mu(h^{-1}(A))$.

Necesitamos, entonces, condiciones bajo las cuales el hecho de tener una convergencia débil, $\mu_n \Rightarrow \mu$, en $\mathcal{P}(E)$, me implique una convergencia débil, $\mu_n h^{-1} \Rightarrow \mu h^{-1}$ para medidas en $\mathcal{P}(\hat{E})$.

Una de las condiciones pensadas para h podrías ser la continuidad, pero veremos que esta condición puede ser reemplazada por una más débil. Asumimos que h será una función $\mathcal{B}(E)/\mathcal{B}(\hat{E})$ -medible, y sea D_h el conjunto de los puntos de discontinuidad de la función h ($D_h \in \mathcal{B}(E)$). Tenemos el siguiente resultado; si $\mu_n \Rightarrow \mu$ y $\mu(D_h) = 0$, entonces $\mu_n h^{-1} \Rightarrow \mu h^{-1}$.

En otras palabras:

Teorema 4 (Del Mapeo.) Sea (\hat{E}, \hat{d}) otro espacio métrico y sea $h : E \rightarrow \hat{E}$, una función medible. $(X_n)_{n \in \mathbb{N}}$, X son variables aleatorias en E .

Si $X_n \xrightarrow{\mathcal{L}} X$ y $\mathbb{P}(X \in D_h) = 0$, entonces: $h(X_n) \xrightarrow{\mathcal{L}} h(X)$. Donde $D_h := \{x \in E \mid h \text{ es discontinua en } x\}$

Demostración: Sea $B \subset \hat{E}$, un conjunto $\mathcal{L}(h(X))$ -continua, es decir:

$$\mathbb{P}(h(X) \in \partial B) = 0 \quad (2.1.7)$$

Por el teorema Portmanteau en variables aleatorias, la convergencia en LEY es equivalente a decir que:

$$\mathbb{P}(h(X_n) \in B) \xrightarrow{n} \mathbb{P}(h(X) \in B)$$

Lo que a su vez, es equivalente a decir que:

$$\mathbb{P}(X_n \in h^{-1}(B)) \xrightarrow{n} \mathbb{P}(X \in h^{-1}(B))$$

Luego, basta con demostrar que $h^{-1}(B)$ es un conjunto $\mathcal{L}(X)$ -continuo, i.e.; falta por demostrar que $\mathbb{P}(X \in \partial h^{-1}(B)) = 0$.

Sea $x \notin D_h$, luego:

$$\begin{aligned} \text{si } x \in \overline{h^{-1}(B)} &\Rightarrow \exists x_n \in h^{-1}(B) \text{ tal que } x_n \rightarrow x \\ &\Rightarrow h(x) \in \overline{B} \end{aligned}$$

Similarmente, si $x \notin D_h$,

$$\begin{aligned} \text{si } x \in \text{int}(h^{-1}(B))^c &\Rightarrow \exists x_n \rightarrow x \text{ tal que } x_n \notin h^{-1}(B) \\ \text{i.e. : } h(x_n) &\notin B^\circ \end{aligned}$$

Luego, para $x \notin D_h$:

$$x \in \partial h^{-1}(B) \Rightarrow h(x) \in \partial B \quad \text{Es decir:}$$

$$\partial h^{-1}(B) \cap D_h^c \subset h^{-1}(\partial B) \cap D_h^c \subset h^{-1}(\partial B)$$

Así, y dado el hecho de que D_h^c es un conjunto de medida 1:

$$\mathbb{P}(X \in \partial h^{-1}(B)) = \mathbb{P}(X \in \partial h^{-1}(B) \cap D_h^c) \leq \mathbb{P}(X \in h^{-1}(\partial B)) = 0$$

De donde la última igualdad se tiene por la hipótesis 2.1.7. ■

Observación 1: La convergencia en ley de X_n a X no requiere que los X_n y X estén definidos en el mismo espacio de probabilidades, puesto que para cada $n \in \mathbb{N}$ el espacio $(\Omega, \mathcal{F}, \mu_n)$ va cambiando de medida, a diferencia de la convergencia c.s. o en probabilidad.

Proposición 1 Sean $(X_n)_{n \in \mathbb{N}}$ y X variables aleatorias definidas sobre el mismo espacio de probabilidades. Entonces:

$$X_n \xrightarrow{\text{c.s.}} X \Rightarrow X_n \xrightarrow{\mathbb{P}} X \Rightarrow X_n \xrightarrow{\mathcal{L}} X$$

Demostración: propuesto.

2.2 Tensión (Tightness)

2.2.1 Relativamente compacto.

Sea $M \subseteq \mathcal{P}(E)$ una familia de medidas de probabilidad sobre el espacio métrico (E, d) . Diremos que M es **relativamente compacto** si cada sucesión de elementos en M contiene una subsucesión que converge débilmente, es decir; si $\{\mu_n\}$ es una secuencia de elementos en M , entonces existe una subsucesión $\{\mu_{n_i}\}$ y una medida de probabilidad $\mu \in \mathcal{P}(E)$, tal que $\mu_{n_i} \Rightarrow \mu$.

Notamos que esta caracterización evita la descripción topológica que se encuentra subyacente en las definiciones de *compacidad* o *convergencia*. Se puede mostrar la existencia de una métrica que dota al espacio de tal topología pero no es de vital importancia para este curso, he ahí el motivo de su ausencia en estas notas.

Sin embargo, notamos que esta definición genera cierto problema para probar que un conjunto de medidas es relativamente compacto, dada la escasa gama de herramientas que poseemos en este momento. Es por esto, que se introduce la siguiente definición:

Definición 3 (Tensión) Una familia $M \subseteq \mathcal{P}(E)$ se dice **TENSA** si $\forall \epsilon > 0, \exists K \subseteq E$ compacto tal que:

$$\mu(K) > 1 - \epsilon \quad \forall \mu \in M$$

El siguiente resultado importante justifica la definición anterior:

Teorema 5 (de Prohorov) Sea $M \subseteq \mathcal{P}(E)$.

- i) Si M es tensa, entonces es relativamente compacta.
- ii) Supongamos que E es un espacio **polaco** (espacio métrico completo y separable). Si M es relativamente compacta, entonces es tensa.

Corolario 1 Si $\{\mu_n\}$ es tensa y cada subsucesión débilmente convergente, converge a la misma medida $\mu \in \mathcal{P}(E)$, entonces toda la secuencia μ_n converge débilmente a μ : $\mu_n \Rightarrow \mu$.

Demostración: Ver [1, págs. 59,69]

Ejemplo 4:

- Si E es compacto, entonces toda familia $M \subseteq \mathcal{P}(E)$ es tensa.
- En $E = \mathbb{R}$, sea $M = \{\delta_n\}$ el conjunto de las delta de Dirac en los puntos $n \in \mathbb{N}$. Claramente M no es tensión, cada vez que proponemos un conjunto compacto, al ser la sucesión $\{n \mid n \in \mathbb{N}\}$ no acotada, existirá un n suficientemente grande que se escape del compacto en \mathbb{R} .
- En $E = \mathbb{R}$, sea $M = \{\mathcal{N}(0, n)\}_{n \in \mathbb{N}}$. M no es tensión.

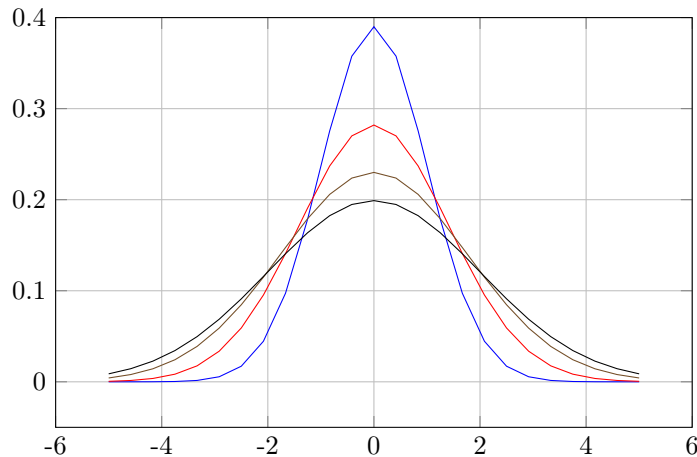


Figure 3: Sucesión de densidades $\mathcal{N}(0, n)$.

- Sin embargo, considerando $E = \overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$, entonces la sucesión $M = \{\delta_n\}_n$ es tensa pues: $\delta_n \Rightarrow \delta_{+\infty}$.
- De la misma forma, si $E = \overline{\mathbb{R}}$, la familia de medidas $M = \{\mathcal{N}(0, n)\}_n$, entonces:

$$\mathcal{N}(0, n) \Rightarrow \frac{1}{2} \delta_{-\infty} + \frac{1}{2} \delta_{+\infty}$$

3 Función Característica

Son variadas las transformaciones o funciones generadoras usadas en matemáticas, probabilidades y estadística. En general, todas ellas se basan en el uso de la función exponencial y su ventaja de convertir sumas en productos. En la siguiente sección se denotará como i ($= \sqrt{-1}$) a la componente imaginaria de un valor complejo.

Los siguientes son ejemplos de transformadas o funciones del tipo antes mencionado:

1. Función generadora de probabilidades: $g(s) = \mathbb{E}(s^X)$
2. Función generadora de momentos: $m(t) = \mathbb{E}(e^{tX})$
3. Transformada de Laplace: $\mathbb{L}(t) = \mathbb{E}(e^{-tX}) = \int e^{-tx} \mu(dx)$
4. Transformada de Fourier: $\mathbb{E}(e^{-itX}) = \int e^{-itx} \mu(dx)$

En lo que sigue, nuestro espacio de trabajo será el espacio métrico \mathbb{R} o \mathbb{R}^d dotados de las métricas usuales.

Definición 4 (Función Característica) Dada $\mu \in \mathcal{P}(\mathbb{R}^d)$, definimos su **FUNCIÓN CARACTERÍSTICA** como:

$$\begin{aligned} \hat{\mu} : \mathbb{R}^d &\longrightarrow \mathbb{C} \\ t &\longrightarrow \hat{\mu}(t) := \int_{\mathbb{R}^d} e^{it \cdot x} \mu(dx) \end{aligned} \quad (3.0.1)$$

Donde la notación:

$$t \cdot x = \sum_{j=1}^d t_j x_j$$

Equivalentemente, podemos enunciar la definición anterior para variables aleatorias, donde la función característica de la variable está dada por la función característica de su ley.

Definición 5 Dada X , variable aleatoria en \mathbb{R}^d , su **FUNCIÓN CARACTERÍSTICA** es:

$$\begin{aligned} \varphi_X(t) : \mathbb{R}^d &\longrightarrow \mathbb{C} \\ t &\longrightarrow \varphi_X(t) := \mathbb{E}(e^{it \cdot X}) \end{aligned} \quad (3.0.2)$$

Observación 1: Podemos notar que la función característica toma valores en el plano complejo, sin embargo; su cálculo se efectúa solamente en variables reales.

$$\varphi_X(t) = \mathbb{E}(e^{itX}) = \int e^{itx} \mu(dx) = \mathbb{E}(\cos(tX)) + i \mathbb{E}(\sin(tX))$$

Ejemplo 1: $X \sim \mathcal{N}(\mu, \sigma^2)$, donde, en este caso, μ y σ son valores reales. Entonces:

$$\varphi_X(t) = e^{i\mu t - \frac{\sigma^2 t^2}{2}}$$

La gran ventaja de la función característica por sobre las funciones como la transformada de Laplace, la funciones generadora de probabilidades o la función generadora de momentos es que garantiza la existencia de la esperanza en cualquier medida de probabilidad. Ya que, todo el cálculo se realiza al integrar sobre una función acotada; $|e^{itx}| \leq 1$ para todo valores $x, t \in \mathbb{R}$. Las siguientes proposiciones serán enunciadas para \mathbb{R} pero son ciertas en \mathbb{R}^d .

Proposición 2 Sea X variable aleatoria, $\varphi = \varphi_X$. Entonces:

- i) φ existe $\forall t$ y para cualquier distribución de X .
- ii) $\varphi(0) = 1$
- iii) $|\varphi(t)| \leq 1$ para todo t

- iv) φ es uniformemente continua, esto es; $\forall \epsilon > 0$, existe $\delta > 0$ tal que, $|\varphi(t) - \varphi(s)| \leq \epsilon$, para cualesquiera t, s tales que $|t - s| \leq \delta$
- v) $\varphi_{a+bX}(t) = e^{iat}\varphi(bt)$, para todos $a, b \in \mathbb{R}$.
- vi) la función característica de $-X$ es el valor conjugado de φ , esto es; $\varphi_{-X}(t) = \overline{\varphi(t)}$
- vii) Si $\varphi(t) \in \mathbb{R} \forall t \iff \mathcal{L}(X) = \mathcal{L}(-X)$

Demostración (proposición):

- i) Notemos que para cualquier x o t , la función $|e^{itx}| \leq 1$. Dado que las medidas de probabilidades son finitas, luego la función e^{itx} siempre es integrable.
- ii) Si $t = 0$, entonces $e^{itx} = 1$, luego $\mathbb{E}(e^0) = 1$.
- iii)

$$\begin{aligned} |\varphi(t)| &= |\mathbb{E}(e^{itX})| = \left| \int_{\mathbb{R}} e^{itx} \mu(dx) \right| \\ &\leq \int_{\mathbb{R}} |e^{itx}| \mu(dx) \leq 1 \cdot \mu(\mathbb{R}) \leq 1 \end{aligned}$$

- iv) Sea $h = t - s$, asumimos sin pérdida de generalidad $s < t$. Entonces:

$$\begin{aligned} |\varphi(t) - \varphi(s)| &= |\mathbb{E}(e^{itX}) - \mathbb{E}(e^{isX})| = |\mathbb{E}(e^{i(h+s)X}) - \mathbb{E}(e^{isX})| \\ &\leq |\mathbb{E}(e^{isX}(e^{ihX} - 1))| \leq \mathbb{E}(|e^{isX}| \cdot |e^{ihX} - 1|) \\ &\leq \mathbb{E}(|e^{ihX} - 1|) \end{aligned}$$

Cuando $h \rightarrow 0$, la función $e^{ihX} - 1$ converge a 0 para todo $\omega \in \Omega$. Es decir, $e^{ihX} - 1 \xrightarrow{h \rightarrow 0} 0$ c.s., como además esta función está acotada por 2 para cualquier valor de t o X , ocupando el **T.C.D.** concluimos que:

$$\lim_{h \rightarrow 0} \mathbb{E}(e^{ihX} - 1) = 0$$

- v) Por definición:

$$\varphi_{a+bX}(t) = \mathbb{E}(e^{it(a+bX)}) = \mathbb{E}(e^{iat}e^{i(bt)X}) = e^{iat}\mathbb{E}(e^{i(bt)X}) = e^{iat}\varphi(bt)$$

- vi) Recordamos que el conjugado del complejo $a + ib$ es $a - ib$. Entonces:

$$\mathbb{E}(e^{it(-X)}) = \mathbb{E}(\cos(-tX) + i\sin(-tX)) = \mathbb{E}(\cos(tX) - i\sin(tX)) = \mathbb{E}(\cos(tX)) - i\mathbb{E}(\sin(tX)) = \overline{\varphi(t)}$$

- vii) (\Leftarrow) Como φ está dado por la ley de la variable, tenemos que:

$$\mathcal{L}(X) = \mathcal{L}(-X) \Rightarrow \varphi_X(t) = \varphi_{-X}(t)$$

Pero, por lo mostrado en el punto iv), concluimos que $\varphi(t) = \overline{\varphi(t)} \forall t$ y eso sólo pasa cuando la función toma valores reales.

Para la implicancia (\Rightarrow) hemos de necesitar resultados posteriores y se mostrará más adelante.

■

Otra de las propiedades interesantes que posee la función característica tiene que ver con la medida de probabilidad generada por la convolución de medidas.

Definición 6 Dadas $\mu, \nu \in \mathcal{P}(\mathbb{R})$, su convolución $\mu * \nu \in \mathcal{P}(\mathbb{R})$, se define como:

$$\int f(z) \mu * \nu(dz) = \int \int f(x + y) \mu(dx) \nu(dy) \quad \forall f \text{ medible}$$

Proposición 3 $\forall \mu, \nu \in \mathcal{P}(\mathbb{R})$:

$$\widehat{\mu * \nu}(t) = \hat{\mu}(t)\hat{\nu}(t)$$

Equivalentemente, si $X \perp\!\!\!\perp Y$, son variables aleatorias en \mathbb{R} :

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t) \quad \forall t$$

El principal interés en la función característica es que describe de manera única a las distribuciones. Las probabilidades de intervalos se pueden recuperar mediante la función característica gracias al siguiente teorema de inversión.

Teorema 6 (Fórmula de Inversión.) Sea X una variable aleatoria en \mathbb{R} , Para todo $a < b$ se tiene que:

$$\mathbb{P}(a < X < b) + \frac{\mathbb{P}(X = a) + \mathbb{P}(X = b)}{2} = \lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \varphi_X(t) dt$$

Demostración (fórmula de Inversión): Sea $\mu = \mathcal{L}(X)$. Denotemos por I_T a la siguiente integral:

$$I_T = \int_{-T}^T \frac{1}{2\pi} \frac{e^{-ita} - e^{-itb}}{it} \varphi_X(t) dt$$

Reemplazamos la definición de función característica:

$$I_T = \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{2\pi it} \int_{\mathbb{R}} e^{itx} \mu(dx) dt = \int_{\mathbb{R}} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{2\pi it} e^{itx} dt \mu(dx)$$

Usando la identidad de Euler para números complejos; $e^{i\alpha} = \cos(\alpha) + i \sin(\alpha)$, tenemos que:

$$I_T = \int_{\mathbb{R}} \int_{-T}^T \left(\frac{\cos(t(x-a)) - \cos(t(x-b))}{2\pi it} + i \frac{\sin(t(x-a)) - \sin(t(x-b))}{2\pi it} \right) dt \mu(dx)$$

Usando que la función *coseno* es una función par, y la función identidad es impar, tenemos que la expresión:

$$\frac{\cos(t(x-a)) - \cos(t(x-b))}{t}$$

es una función impar para la variable t , por lo tanto; toda integral sobre un intervalo simétrico es nula.

$$\frac{1}{2\pi i} \int_{-T}^T \frac{\cos(t(x-a)) - \cos(t(x-b))}{t} dt = 0$$

Del mismo modo, la función *seno* y la función identidad son funciones impares, por lo tanto la expresión:

$$\frac{\sin(t(x-a)) - \sin(t(x-b))}{t}$$

es una función par, por lo tanto; para toda integral sobre un intervalo simétrico, tenemos que:

$$\frac{1}{2\pi i} \int_{-T}^T \frac{\sin(t(x-a)) - \sin(t(x-b))}{t} dt = \frac{2}{2\pi i} \int_0^T \frac{\sin(t(x-a)) - \sin(t(x-b))}{t} dt$$

Por lo tanto la integral I_T queda de la siguiente forma:

$$\begin{aligned} I_T &= \int_{\mathbb{R}} \frac{2i}{2\pi i} \int_0^T \frac{\sin(t(x-a)) - \sin(t(x-b))}{t} dt \mu(dx) = \int_{\mathbb{R}} \left(\frac{1}{\pi} \int_0^T \frac{\sin(t(x-a)) - \sin(t(x-b))}{t} dt \right) \mu(dx) \\ &= \int_{\mathbb{R}} \frac{1}{\pi} \left(\int_0^T \frac{\sin(t(x-a))}{t} dt - \int_0^T \frac{\sin(t(x-b))}{t} dt \right) \mu(dx) \end{aligned}$$

Para El siguiente paso ocuparemos el siguiente lema:

Lema 2 Sea $c \in \mathbb{R}$:

$$\lim_{T \rightarrow \infty} \frac{1}{\pi} \int_0^T \frac{\sin(ct)}{t} dt = \begin{cases} -\frac{1}{2} & \text{si } c < 0 \\ \frac{1}{2} & \text{si } c > 0 \\ 0 & \text{si } c = 0 \end{cases}$$

De esta forma, el límite de I_T dependerá de los valores de x respecto a a y b (tomando $c = x - a$ o $c = x - b$). Luego, si $a < b$, tenemos que:

$$\lim_{T \rightarrow \infty} \frac{1}{\pi} \left(\int_0^T \frac{\sin(t(x-a))}{t} dt - \int_0^T \frac{\sin(t(x-b))}{t} dt \right) = 0$$

Para todo $x \in (-\infty, a) \cup (b, \infty)$. Más aún, basta con verificar el signo de $x - a$ o $x - b$ en cada caso para probar que:

$$\lim_{T \rightarrow \infty} \frac{1}{\pi} \left(\int_0^T \frac{\sin(t(x-a))}{t} dt - \int_0^T \frac{\sin(t(x-b))}{t} dt \right) = \begin{cases} \frac{1}{2} & \text{si } x = a \\ \frac{1}{2} & \text{si } x = b \\ 1 & \text{si } x \in (a, b) \\ 0 & \text{si } x \in (-\infty, a) \cup (b, \infty) \end{cases}$$

Por lo tanto el límite de I_T se reduce a:

$$\begin{aligned} \lim_{T \rightarrow \infty} I_T &= \int_{\mathbb{R}} \lim_{T \rightarrow \infty} \frac{1}{\pi} \left(\int_0^T \frac{\sin(t(x-a))}{t} dt - \int_0^T \frac{\sin(t(x-b))}{t} dt \right) \mu(dx) \\ &= \int_{\{a\}} \frac{1}{2} \mu(dx) + \int_{\{b\}} \frac{1}{2} \mu(dx) + \int_{(a,b)} 1 \mu(dx) = \mu((a, b)) + \frac{\mu(\{a\}) + \mu(\{b\})}{2} \end{aligned}$$

■

Corolario 2 Si X e Y son variables aleatorias tales que $\varphi_X(t) = \varphi_Y(t)$ para todo valor de t , entonces $\mathcal{L}(X) = \mathcal{L}(Y)$.

Teorema 7 Sea $\mu \in \mathcal{P}(\mathbb{R})$:

- Si $\int |x|^n \mu(dx) < \infty$ para cierto $n \in \mathbb{N}$, entonces φ es n veces derivable y:

$$\varphi^{(n)}(t) = \int (ix)^n e^{itx} dt$$

En particular se tiene la igualdad: $\varphi^{(n)}(0) = i^n \int x^n \mu(dx)$.

- Si $\varphi^{(2n)}(0)$ existe para cierto $n \in \mathbb{N}$, entonces $\int x^{2n} \mu(dx) < \infty$.

Demostración: propuesto.

De manera recíproca, conocer la función característica, $\hat{\mu}$, nos proporciona conocimiento acerca de la forma en que la distribución reparte valores sobre el conjunto y, gracias al último teorema, es posible saber el valor de los momentos asociados en caso de existir. El siguiente resultado analiza el comportamiento de la función característica acercándose "a infinito" y vincula la integrabilidad de la función con la existencia de una densidad para μ con respecto a la medida de Lebesgue.

Teorema 8 (Transformada de Fourier para la función característica) Sea $\mu \in \mathcal{P}(\mathbb{R})$, $\varphi = \hat{\mu}$. Supongamos la condición:

$$\int_{\mathbb{R}} |\varphi(t)| dt < \infty$$

Entonces existe f , densidad de μ . Es decir; $\mu(dx) = f(x)dx$, y además se tiene que:

$$f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} \varphi(t) dt$$

Demostración: El primer paso para ver la existencia de una densidad es comprobar que la medida de probabilidad μ es *no atómica*, es decir, que para cada elemento $a \in \mathbb{R}$, la medida del singleton es cero: $\mu(\{a\}) = 0$.

Usando la fórmula de inversión tenemos que: para $a < b$:

$$\mu((a, b)) + \frac{\mu(\{a\}) + \mu(\{b\})}{2} = \lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt$$

Podemos reemplazar la expresión adentro de la integral de la siguiente forma:

$$\frac{e^{-ita} - e^{-itb}}{it} = \int_a^b e^{-itx} dx$$

Entonces, acotando:

$$\begin{aligned} \mu((a, b)) + \frac{\mu(\{a\}) + \mu(\{b\})}{2} &\leq \lim_{T \rightarrow \infty} \left| \frac{1}{2\pi} \int_{-T}^T \int_a^b e^{-itx} dx \varphi(t) dt \right| \\ &\leq \lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T \int_a^b |e^{-itx}| dx |\varphi(t)| dt = (b-a) \cdot \frac{1}{2\pi} \int_{\mathbb{R}} |\varphi(t)| dt \end{aligned}$$

Donde la última igualdad se tiene, pues $|e^{-itx}| = 1$ para todo valor de x o de t . Ahora, como la integral en todo el espacio de $|\varphi|$ es finita, y dada la continuidad de la medida, cuando hacemos tender $b \rightarrow a$:

$$\mu(\{a\}) \leq \lim_{b \rightarrow a} (b-a) \cdot \frac{1}{2\pi} \int_{\mathbb{R}} |\varphi(t)| dt = 0$$

Así probamos que la medida μ es *no atómica*, y además:

$$\begin{aligned} \mu((a, b)) &= \frac{1}{2\pi} \int_{\mathbb{R}} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt \\ &= \int_a^b \left(\frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} \varphi(t) dt \right) dx \end{aligned}$$

En palabras simples: integrando por la medida de Lebesgue, esa función, sobre ese conjunto, recupero la medida del conjunto. Lo que es la definición de densidad. ■

Hemos visto que una condición suficiente para la convergencia puntual de medidas de probabilidad en otro elemento de $\mathcal{P}(E)$ es que la sucesión tomada sea *tensa*. Del mismo modo, la convergencia puntual de funciones características de alguna sucesión de medidas μ_n , no necesariamente sería función característica, ni nos entrega información sobre la convergencia de μ_n . El siguiente teorema nos da condiciones bajo las cuales, los hechos anteriores se pueden asegurar.

Teorema 9 (Continuidad de Lévy) Sea $(\mu_n)_{n \in \mathbb{N}}$ una sucesión en $\mathcal{P}(\mathbb{R})$:

i) Si $\exists \mu \in \mathcal{P}(\mathbb{R})$ tal que $\mu_n \Rightarrow \mu$, entonces:

$$\hat{\mu}_n(t) \xrightarrow{n} \hat{\mu}(t) \quad \forall t \in \mathbb{R}$$

ii) Si $\hat{\mu}_n(t) \rightarrow \varphi(t) \quad \forall t \in \mathbb{R}$, para alguna función φ continua en 0, entonces $\exists \mu \in \mathcal{P}(\mathbb{R})$ tal que $\varphi = \hat{\mu}$ y $\mu_n \Rightarrow \mu$.

Demostración:

i) Para la demostración de la primera parte, basta notar que la función $f(x) = e^{itx} \in C_b(\mathbb{R})$, por lo tanto, por definición de convergencia débil:

$$\hat{\mu}_n(t) = \mathbb{E}_{\mu_n}(e^{itX}) = \langle \mu_n, f \rangle \xrightarrow{n} \langle \mu, f \rangle = \mathbb{E}_{\mu}(e^{itX}) = \hat{\mu}(t)$$

ii) Si la sucesión $(\mu_n)_n$ fuese relativamente compacta, entonces:

- $\exists \mu \in \mathcal{P}(\mathbb{R})$ y una subsucesión $\{n_i\}_{i \in \mathbb{N}}$ tal que $\mu_{n_i} \Rightarrow \mu$.
- Como $\hat{\mu}_n \rightarrow \varphi$, en particular la convergencia se tiene para cualquier subsucesión de $\hat{\mu}_n$, por lo tanto; de haber sub sucesión $(\mu_{n_i})_i$ débil-convergente a μ , entonces $\hat{\mu} = \varphi$, y del hecho de que la función característica es propia de la medida de probabilidad, se concluye que todas las subsucesiones débilmente convergentes de μ_n convergen al mismo elemento $\mu \in \mathcal{P}(\mathbb{R})$. (Ver corolario 1, teorema de Prohorov)

Así aseguramos la existencia de $\mu \in \mathcal{P}(\mathbb{R})$ tal que $\hat{\mu} = \varphi$

Por lo tanto basta probar que la secuencia $(\mu_n)_{n \in \mathbb{N}}$ es relativamente compacta. O, equivalentemente, gracias al Teorema de Prohorov, mostrar que es tensa. Para ello, veamos que: $\forall T > 0$,

$$\frac{1}{2T} \int_{-T}^T \hat{\mu}_n(t) dt = \frac{1}{2T} \int_{-T}^T \int_{\mathbb{R}} e^{itx} \mu_n(dx) dt = \int_{\mathbb{R}} \frac{1}{2T} \int_{-T}^T e^{itx} dt \mu_n(dx)$$

Luego, ocupando la caracterización de la exponencial compleja:

$$\int_{-T}^T e^{itx} dt = \int_{-T}^T \cos(tx) dt + i \int_{-T}^T \sin(tx) dt = 2 \int_0^T \cos(tx) dt = \frac{2}{x} \sin(Tx)$$

Entonces, reemplazando en la ecuación anterior:

$$\frac{1}{2T} \int_{-T}^T \hat{\mu}_n(t) dt = \int_{\mathbb{R}} \frac{\sin(Tx)}{Tx} \mu_n(dx)$$

Para cualquier valor $l \in \mathbb{R}$ podemos separar la integral en dos intervalos disjuntos:

$$\frac{1}{2T} \int_{-T}^T \hat{\mu}_n(t) dt = \int_{\mathbb{R}} \frac{\sin(Tx)}{Tx} \mu_n(dx) \leq \int_{|x| < l} \left| \frac{\sin(Tx)}{Tx} \right| \mu_n(dx) + \int_{|x| \geq l} \left| \frac{\sin(Tx)}{Tx} \right| \mu_n(dx)$$

Luego, usando las siguientes cotas: para $|x| < l$:

$$\left| \frac{\sin(Tx)}{Tx} \right| \leq 1$$

mientras que para $|x| \geq l$:

$$\left| \frac{\sin(Tx)}{Tx} \right| \leq \frac{1}{Tl}$$

Por lo tanto:

$$\begin{aligned} \frac{1}{2T} \int_{-T}^T \hat{\mu}_n(t) dt &\leq \int_{|x| < l} 1 \cdot \mu(dx) + \int_{|x| \geq l} \frac{1}{Tl} \mu(dx) = \mu_n(|x| < l) + \frac{1}{Tl} \mu(|x| \geq l) \\ &\leq 1 - \mu_n(|x| \geq l) + \frac{1}{Tl} \mu_n(|x| \geq l) \end{aligned}$$

De esta forma, y tomando $l = 2/T$, tenemos que:

$$\frac{1}{2} \mu_n(|x| \geq 2/T) \leq 1 - \frac{1}{2T} \int_{-T}^T \hat{\mu}_n(t) dt = \frac{1}{2T} \int_{-T}^T (1 - \hat{\mu}_n(t)) dt$$

Como $\hat{\mu}_n(t) \rightarrow \varphi(t)$ para todo t . Ocupando el **T.C.D.**:

$$\limsup_n \frac{1}{2} \mu_n(|x| \geq \frac{2}{T}) \leq \frac{1}{2T} \int_{-T}^T (1 - \varphi(t)) dt$$

Tomando límite superior, cuando $T \rightarrow 0$: ²

$$\limsup_{T \rightarrow 0} \limsup_n \frac{1}{2} \mu_n(|x| \geq \frac{2}{T}) \leq (1 - \varphi(0))$$

De donde tenemos que, por la convergencia puntual de los $\hat{\mu}_n$ a φ , como $\hat{\mu}_n(0) = 1$ para todo $n \in \mathbb{N}$, entonces necesariamente $\varphi(0) = 1$. Entonces:

$$\limsup_{T \rightarrow \infty} \limsup_n \frac{1}{2} \mu_n(|x| \geq \frac{2}{T}) = 0$$

En otras palabras; dado $\epsilon > 0$, $\exists 0 < T < \delta$, y $\exists n_o \in \mathbb{N}$ tal que:

$$\frac{1}{2} \mu_n(|x| \geq \frac{2}{T}) \leq \epsilon \quad \forall n \geq n_o$$

Luego concluimos que $(\mu_n)_{n \in \mathbb{N}}$ es tensa. ■

²El lado derecho de la siguiente desigualdad se puede ver como una uniforme en $(-1/\xi, 1/\xi)$ cuando $T = 1/\xi$, que ya vimos anteriormente que converge débilmente a δ_0 .

Teorema 10 (Cramér-Wold) Sean $X^{(n)} = (X_1^{(n)}, X_2^{(n)}, \dots, X_d^{(n)})$ y $X = (X_1, X_2, \dots, X_d)$, variables aleatorias en \mathbb{R}^d . Entonces:

$$X^{(n)} \xrightarrow{\mathcal{L}} X \iff \forall \lambda \in \mathbb{R}^d : \lambda \cdot X^{(n)} \xrightarrow{\mathcal{L}} \lambda \cdot X$$

Este teorema es de gran utilidad, por un lado tenemos convergencia en ley en \mathbb{R}^d y por el otro, convergencia en \mathbb{R} , gracias al teorema esto será equivalente bajo algunas condiciones. En particular, podemos deducir que si una sucesión de variables aleatorias en \mathbb{R}^d converge en ley, lo harán cada una de sus coordenadas por si solas.

Demostración:

(\Rightarrow) $\forall \lambda \in \mathbb{R}^d$ definamos la función:

$$\begin{aligned} f_\lambda : \mathbb{R}^d &\rightarrow \mathbb{R} \\ x &\rightarrow \lambda \cdot x \end{aligned}$$

Entonces f_λ es continua en todo el espacio, en particular la medida de los puntos de discontinuidad es cero (vacío). Como $X^{(n)}$ converge en ley a X , por teorema del mapeo tenemos que también lo harán $f_\lambda(X^{(n)})$ a $f_\lambda(X)$.

(\Leftarrow) Tomemos los vectores canónicos en \mathbb{R}^d , $\lambda_i = (0, \dots, 1, \dots, 0)$, se tiene que $\lambda_i \cdot X^{(n)} = X_i^{(n)}$, por lo tanto $\mathcal{L}(X_i^{(n)}) \Rightarrow \mathcal{L}(X_i)$. En particular $(\mathcal{L}(X_i^{(n)}))_{n \in \mathbb{N}}$ es tensa $\forall i$, luego $(\mathcal{L}(X^{(n)}))_{n \in \mathbb{N}}$ también es tensa³. Aplicando el teorema de Prohorov, la sucesión es relativamente compacta.

Existe $\mu \in \mathcal{P}(\mathbb{R}^d)$ que es límite débil de alguna subsucesión de $\mathcal{L}(X^{(n)})$. Por hipótesis, como para todo $t \in \mathbb{R}^d$, $t \cdot X^{(n)} \xrightarrow{\mathcal{L}} t \cdot X$, entonces:

$$\mathbb{E}(e^{it \cdot X^{(n)}}) \rightarrow \mathbb{E}(e^{it \cdot X}) = \varphi_X(t)$$

En particular, esta convergencia se tiene para cualquier sub sucesión $\{n_i, i \in \mathbb{N}\}$, por lo tanto necesariamente $\hat{\mu} = \varphi_X$ y toda subsucesión de $\mathcal{L}(X^{(n)})$ converge a μ (por la unicidad de la función generadora). Luego $\mu = \mathcal{L}(X)$.

■

³Esta conclusión se dejará como ejercicio.

4 Teorema Central del Límite

Nuestro objetivo es probar el primer teorema enunciado en este curso. Aquel que dice que a suma de variables aleatorias independientes, cuando se estandarizan, converge en ley a una distribución normal estándar. La demostración usual se basa en el uso de la función generadora de momentos. Sin embargo, ésta sólo existe cuando se garantiza la existencia de todos los momentos, para cada orden, por lo tanto, un resultado más general que sólo necesite varianza y esperanza finita requiere el uso de la función característica de la variable.

Los siguientes lemas son necesarios, previos a la demostración:

Lema 3 Sea X , variable aleatoria en \mathbb{R} tal que $\mathbb{E}(X^2) < +\infty$. Entonces:

$$\varphi_X(t) = 1 + it\mathbb{E}(X) - \frac{t^2}{2}\mathbb{E}(X^2) + o(t^2)$$

Demostración: Usando la expansión de Taylor de la función e^x y evaluarla en ix , tenemos que:

$$e^{ix} = 1 + ix - \frac{x^2}{2} + R_2(ix)$$

Donde R_k es el *resto integral* de la función $f(x) = e^x$:

$$R_k(ix) = \int_0^{ix} \frac{f^{(k)}(t)}{k!} (ix - t)^k dt = \int_0^x \frac{f^{(k)}(it)}{k!} (ix - it)^k dt = i^k \int_0^x \frac{f^{(k)}(it)}{k!} (x - t)^k dt$$

Basta tomar $k = 2$ y $f^{(k)}(ix) = e^{ix}$ entonces:

$$|R_2(ix)| = \left| -\frac{1}{2} \int_0^x e^{it} (x - t)^2 dt \right| \leq \frac{1}{2} \int_0^{|x|} (|x| - t)^2 dt \leq \frac{|x|^3}{6}$$

Por otro lado, integrando por partes:

$$\begin{aligned} R_2(ix) &= -\frac{1}{2} \int_0^x e^{it} (x - t)^2 dt = -\frac{1}{2} \left[\frac{e^{it}}{i} (x - t)^2 \Big|_{t=0}^{t=x} + 2 \int_0^x \frac{e^{it}}{i} (x - t) dt \right] \\ \Rightarrow |R_2(ix)| &\leq \frac{1}{2} \left[x^2 + 2 \int_0^{|x|} (|x| - t) dt \right] = x^2 \end{aligned}$$

Así, obtenemos que:

$$|R_2(ix)| \leq \min\{x^2, \frac{|x|^3}{6}\}$$

Luego:

$$\begin{aligned} \varphi_X(t) &= \mathbb{E}(e^{itX}) = \mathbb{E} \left(1 + itX - \frac{t^2}{2}X^2 + R_2(itX) \right) \\ &= 1 + it\mathbb{E}(X) - \frac{t^2}{2}\mathbb{E}(X^2) + \mathbb{E}(R_2(itX)) \end{aligned}$$

Con esto, sólo basta probar que $\mathbb{E}(R_2(itX)) = o(t^2)$. Para esto necesitamos ver que $|R_2(itX)|/t^2 \xrightarrow{t \rightarrow 0} 0$. En efecto:

$$\begin{aligned} \frac{|\mathbb{E}(R_2(itX))|}{t^2} &\leq \frac{\mathbb{E} \left(\min\{t^2 X^2, \frac{t^3 |X|^3}{6}\} \right)}{t^2} \\ &\leq \mathbb{E} \left(\min \left\{ X^2, \frac{t|X|^3}{6} \right\} \right) \end{aligned}$$

Acá podemos notar que la esperanza está bien definida, puesto que el mínimo siempre es menor o igual a la variable X^2 , cuya esperanza es finita. Sin embargo, a medida que $t \rightarrow 0$, el término $\frac{t|X|^3}{6}$ se irá acercando a cero, minorando a X^2 y arrastrando la esperanza a cero vía el **T.C.D.**. Por lo tanto $\mathbb{E}(R_2(itX)) = o(t^2)$, concluyendo. ■

Lema 4 $\forall w_i, z_i \in \mathbb{C}$, tal que $|w_i| \leq 1, |z_i| \leq 1$ para todo $i = 1, \dots, n$. Se tiene que:

$$\left| \prod_{i=1}^n z_i - \prod_{i=1}^n w_i \right| \leq \sum_{i=1}^n |z_i - w_i|$$

Demostración:

$$\begin{aligned} \left| \prod_{i=1}^n z_i - \prod_{i=1}^n w_i \right| &= \left| \prod_{i=1}^n z_i - \prod_{i=1}^n w_i + w_n \prod_{i=1}^{n-1} z_i - w_n \prod_{i=1}^{n-1} w_i \right| = \left| (z_n - w_n) \prod_{i=1}^{n-1} z_i + w_n \left(\prod_{i=1}^{n-1} z_i - \prod_{i=1}^{n-1} w_i \right) \right| \\ &\leq |z_n - w_n| \prod_{i=1}^{n-1} |z_i| + |w_n| \cdot \left| \prod_{i=1}^{n-1} z_i - \prod_{i=1}^{n-1} w_i \right| \leq |z_n - w_n| + \left| \prod_{i=1}^{n-1} z_i - \prod_{i=1}^{n-1} w_i \right| \end{aligned}$$

De donde la última desigualdad sale del hecho de que, tanto los w_i como los z_i tenían módulo menor o igual a 1. Finalmente, iterando este desarrollo n veces, se concluye el resultado por inducción. ■

Recordemos la manera en que se enuncia el teorema central del límite.

Teorema 11 (Teorema Central del Límite) Sean X_1, X_2, \dots variables aleatorias i.i.d., con $\mathbb{E}(X_1) = \mu \in \mathbb{R}$ y $\text{Var}(X_1) = \sigma^2 < +\infty$. Entonces:

$$Z_n = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n (X_i - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

4.1 Demostración del T.C.L.

Sea $Z \sim \mathcal{N}(0, 1)$, luego $\varphi_Z(t) = e^{-\frac{t^2}{2}}$. Por el teorema de continuidad de Lévy, basta probar que:

$$\varphi_{Z_n}(t) \xrightarrow{n} \varphi_Z(t) \quad \forall t$$

Sea $\varphi = \varphi_{\frac{X_i - \mu}{\sigma}}$, como $\frac{X_i - \mu}{\sigma}$ tiene esperanza 0 y varianza 1, luego por el primer lema anterior:

$$\varphi(t) = 1 - \frac{t^2}{2} + o(t^2)$$

Además:

$$\begin{aligned} \varphi_{Z_n}(t) &= \mathbb{E}(e^{it \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mu}{\sigma}}) \stackrel{i.i.d.}{=} \prod_{i=1}^n \mathbb{E}\left(e^{\frac{it}{\sqrt{n}} \left(\frac{X_i - \mu}{\sigma}\right)}\right) \\ &= \prod_{i=1}^n \varphi\left(\frac{t}{\sqrt{n}}\right) = \varphi\left(\frac{t}{\sqrt{n}}\right)^n = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \end{aligned}$$

Con esto:

$$\left| \varphi_{Z_n}(t) - e^{-t^2/2} \right| \leq \left| \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n - \left(1 - \frac{t^2}{2n}\right)^n \right| + \left| \left(1 - \frac{t^2}{2n}\right)^n - e^{-t^2/2} \right|$$

Usando el segundo lema, con $z_i = 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)$ y $w_i = 1 - \frac{t^2}{2n}$, para todo i , tenemos que:

$$\begin{aligned} \left| \varphi_{Z_n}(t) - e^{-t^2/2} \right| &\leq \sum_{i=1}^n |o(t^2/n)| + \left| \left(1 - \frac{t^2}{2n}\right)^n - e^{-t^2/2} \right| = n|o(t^2/n)| + \left| \left(1 - \frac{t^2}{2n}\right)^n - e^{-t^2/2} \right| \\ &\leq \frac{|o(t^2/n)|}{\frac{1}{n}} + \left| \left(1 - \frac{t^2}{2n}\right)^n - e^{-t^2/2} \right| \rightarrow 0 \end{aligned}$$

Donde el segundo sumando converge a 0 por la definición de límite de la función exponencial⁴. Con esto demostramos que φ_{Z_n} converge a esa exponencial, que es justamente la función característica de la variable normal estándar. ■

⁴Además queda como propuesto mostrar que cuando n tiende a infinito, para cualquier t , el primer sumando converge a 0, utilizando la definición de la función $o(t^2/n)$.

5 Método de Monte Carlo y Simulación

Para introducir el método de Monte Carlo, consideremos el problema de la integración numérica. Existen diversos métodos numéricos para aproximar (computacionalmente) integrales de la forma:

$$\int_{[0,1]^d} f(x)dx,$$

basados en fórmulas del tipo $\sum_{i=1}^n w_i f(x_i)$, cuando $f : \mathbb{R}^d \rightarrow \mathbb{R}$ es una función conocida y, usualmente, la suma de los términos w_i es uno y los términos x_i son puntos del conjunto $[0,1]^d$. Por ejemplo, $w_i = 1/k$ y la malla de $[0,1]^d$ compuesta por:

$$\left\{0, \frac{1}{k}, \frac{2}{k}, \dots, \frac{k}{k}\right\}^d \quad \text{Para algún } k \text{ natural.}$$

La cual contiene $n = (k+1)^d$ ($\approx k^d$) puntos. Si el método es de orden 1, el error estará asociado al orden del salto de la malla, en este caso es de $1/k$, por lo tanto:

$$\text{error} \sim \frac{1}{k} \approx \frac{1}{n^{1/d}}$$

Luego, si quisiéramos ajustarnos a un error de un orden pequeño, $\varepsilon > 0$, necesariamente tendríamos que tomar una cantidad de datos proporcionalmente, de la forma; $n \approx 1/\varepsilon^d$, lo que es aceptable para un ε pequeño y un d pequeño (por ejemplo $d \leq 3$). El problema es que para variables de dimensión mayor a eso, el n asociado crece de manera abrupta.

Los métodos de Monte Carlos ofrecen una alternativa a este problema, reformulando la integral, vista ahora como la esperanza de una variable aleatoria. La discretización se forma de manera natural al elegir $w_i = 1/n$ y x_i como la realización de variables aleatorias uniformes en $[0,1]^d$. Luego, la convergencia de $\sum_{i=1}^n w_i f(x_i)$ estará garantizada por la Ley de los Grandes Números.

Teorema 12 (Ley de los Grandes Números (L.G.N.)) Sean $(X_n)_{n \in \mathbb{N}}$ variables aleatorias i.i.d., con $\mathbb{E}(X_1) = \mu < +\infty$. Se define el promedio finito:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

Entonces:

$$\bar{X}_n \xrightarrow{n} \mu \quad \text{c.s. y en } L^1$$

Observación 1

- Si $X_1 \in L^p$, entonces la convergencia es en L^p .
- Si $\sigma^2 = \text{Var}(X_1) < +\infty$, entonces:

$$\mathbb{E}(|\bar{X}_n - \mu|) \leq \frac{\sigma}{\sqrt{n}}$$

Además de esto, la tasa de convergencia es del orden de $n^{1/2}$ y está dada por el Teorema Central del Límite. Claramente, la tasa de convergencia podría parecer algo lenta en comparación con las tasas de otros métodos numéricos de integración en dimensión 1. Pero todos esos métodos colapsan cuando nos escapamos a dimensiones mayores.

Siguiendo con el ejemplo anterior, llamemos I a la integral que queríamos aproximar:

$$I = \int_{[0,1]^d} f(x)dx$$

Si definimos la variable $U \sim \text{unif}([0,1]^d)$, es decir $U = (U_1, U_2, \dots, U_d)$, con $U_i \sim \text{unif}(0,1)$ i.i.d., entonces la integral anterior la podemos escribir como $I = \mathbb{E}(f(U))$. Considerando $U^{(1)}, U^{(2)}, \dots, U^{(m)}$ realizaciones i.i.d. de la variable U , por **L.G.N.**:

$$\frac{1}{m} \sum_{i=1}^m f(U^{(i)}) \approx I$$

5.1 Descripción del método

Para usar el método de Monte Carlo, es necesario poder escribir la cantidad de interés como el valor esperado de una variable aleatoria "*eficientemente simulable en un computador*". Esto suele ser fácil, como es el caso de la aproximación de una integral, pero podría ganar mucha más complejidad, como cuando deseamos resolver alguna ecuación diferencial parabólica o elíptica.

Sea un valor de interés, $\alpha \in \mathbb{R}$, los pasos a seguir son:

1. Escribir $\alpha = \mathbb{E}(X)$, donde X es una variable aleatoria.
2. Para algún n suficientemente grande, simular X_1, X_2, \dots, X_n copias *i.i.d* de X (muestra).
3. Aproximar $\alpha \approx \frac{1}{n} \sum_{i=1}^n X_i$.

Es claro, por **L.G.N.** que a medida que nuestra muestra aumenta, la estimación será más cercana. Por lo tanto hay que tener una idea de lo que puede ser "suficientemente grande" en cada caso de estimación.

Ventajas de *M.M.C* (método de Monte Carlo):

- Simple de implementar.
- Escala moderadamente bien, a pesar de la dimensión.

Desventajas de *M.M.C*:

- El resultado es aleatorio (para cada realización diferente, el resultado de la estimación variará).
- El error es del orden $1/\sqrt{n}$, lo que en la práctica no suele ser muy bueno.

Observación 2 Si $\sigma^2 = \text{Var}(X_1) < \infty$, por el **T.C.L.**:

$$\frac{\bar{X}_n - \alpha}{\sigma/\sqrt{n}} \approx Z, \quad Z \sim \mathcal{N}(0, 1)$$

Así,

$$|\bar{X}_n - \alpha| \approx |Z| \frac{\sigma}{\sqrt{n}}$$

Luego, es de utilidad aproximar (estimar) σ para tener una idea del error cometido.

Ejemplo 1 Dada Y variable aleatoria en \mathbb{R} y $A \subseteq \mathbb{R}$. Si me interesa calcular (estimar):

$$p = \mathbb{P}(Y \in A) = \mathbb{E}(\mathbb{1}_{Y \in A})$$

Lo natural es definir $X = \mathbb{1}_{Y \in A}$. Sean X_1, X_2, \dots, X_n copias *i.i.d.* de X , luego:

$$p \approx \frac{X_1 + X_2 + \dots + X_n}{n} =: \hat{p}$$

Notemos que $\sigma^2 = \text{Var}(X) = \text{Var}(\text{Bernoulli}(p)) = p(1-p) \leq 1/4$ (ver figura).

Supongamos que queremos un intervalo de confianza al 95%, con error $\pm \varepsilon$.

$$0.95 \stackrel{!}{=} \mathbb{P}(p \in [\hat{p} - \varepsilon, \hat{p} + \varepsilon]) = \mathbb{P}(|\hat{p} - p| \leq \varepsilon) = \mathbb{P}\left(\frac{|\hat{p} - p|}{\sigma/\sqrt{n}} \leq \frac{\varepsilon\sqrt{n}}{\sigma}\right)$$

Nuevamente, ocupando **T.C.L.**:

$$\begin{aligned} 0.95 &\approx \mathbb{P}\left(|\mathcal{N}(0, 1)| \leq \frac{\varepsilon\sqrt{n}}{\sigma}\right) \Rightarrow \frac{\varepsilon\sqrt{n}}{\sigma} \approx 1.96 \\ \Rightarrow \sqrt{n} &\approx \frac{1.96\sigma}{\varepsilon} \Rightarrow n \approx \frac{1.96^2\sigma^2}{\varepsilon^2} \leq \frac{4 \cdot 1/4}{\varepsilon^2} = \frac{1}{\varepsilon^2} \end{aligned}$$

Así, ganamos una relación entre la cantidad de muestra que necesitamos para controlar el error asociado a la estimación. Por ejemplo, si queremos un error no mayor a $\varepsilon = 0.01$, bastaría con tomar $n = 10000$ observaciones para lograrlo con 95% de confianza.

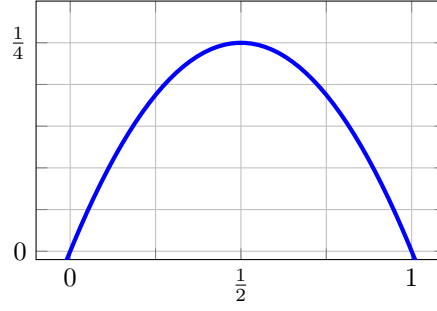


Figure 4: Gráfico de $p(1-p)$, cuyo vértice (punto max) se encuentra en $(1/2, 1/4)$

5.2 Simulación de variables aleatorias

En la actualidad, todos los lenguajes de programación modernos poseen generadores de números *pseudoaleatorios*⁵. Tales programas producen como salida, una secuencia perfectamente determinista (y, a veces, periódica), pero cuyas propiedades estadísticas se parecen lo suficiente a las de una secuencia de realizaciones independientes de una variable aleatoria $unif(0, 1)$. El problema de crear un "buen generador" de números aleatorios es crear una fórmula de recurrencia que, en un tiempo razonable, produzca una secuencia de números que se parezca lo más posible a una secuencia de realizaciones independientes $unif(0, 1)$ con el periodo lo más largo posible.⁶ Con este recurso computacional en mente, el objetivo ahora es recrear otras variables aleatorias con distribuciones diferentes que sean de interés, a partir de realizaciones uniformes en el intervalo $(0, 1)$.

1. **Simulación de una variable aleatoria Bernoulli:** Sea $0 < p < 1$. Si U es una variable aleatoria $unif(0, 1)$. Entonces $X = \mathbb{1}_{\{U \leq p\}}$ es una variable *Bernoulli* de parámetro p .
2. **Simulación de una variable aleatoria Binomial:** Sea $n \in \mathbb{N}$, U_1, U_2, \dots, U_n variables aleatorias *i.i.d.* $unif(0, 1)$, entonces:

$$X = \mathbb{1}_{\{U_1 \leq p\}} + \mathbb{1}_{\{U_2 \leq p\}} + \dots + \mathbb{1}_{\{U_n \leq p\}}$$

es una variable aleatoria binomial de parámetros n y p , $B(n, p)$.

3. **Simulación de una variable aleatoria Geométrica:**

$$X = \inf\{k \geq 1; U_k \leq p\}$$

Cuando $(U_n)_n$ son variables aleatorias *i.i.d.* con distribución $unif(0, 1)$, entonces $X \sim geom(p)$.

Para proceder a una simulación más eficiente nos basaremos en el siguiente lema.

Lema 5 Sea X una variable aleatoria, y F su función de distribución (i.e. $F(x) = \mathbb{P}(X \leq x)$). Para $0 \leq t \leq 1$, definimos:

$$F^{-1}(t) = \inf\{x; F(x) > t\}$$

Entonces, si U tiene ley $unif(0, 1)$, $F^{-1}(U)$ tiene ley $\mathcal{L}(X)$.

Demostración: Notemos que si F es una función invertible, entonces su inversa coincide con F^{-1} , además tendremos que F^{-1} será una función creciente, al igual que F . Entonces:

$$\mathbb{P}(F^{-1}(U) \leq t) = \mathbb{P}(U \leq F(t)) = \int_0^{F(t)} dx = F(t) = \mathbb{P}(X \leq t)$$

⁵La palabra marca una diferencia respecto al término *aleatorio*, pero no se dará enfoque a este concepto en este curso.

⁶ver [3, cap. 1]

Donde la integral sale del hecho de que U es una variable uniforme entre $[0, 1]$ y $F(t)$ siempre pertenece a tal intervalo por ser una probabilidad.

¿Qué pasa cuando la función F no es invertible? Notemos que $\forall x \in \mathbb{R}$:

$$\{t : F^{-1}(t) \leq x\} \cup \{F(x)\} = \{t : t \leq F(x)\} \quad (5.2.1)$$

En efecto: claramente $F(x) \in \{t : t \leq F(x)\}$, además:

$$F^{-1}(t) \leq x \iff \inf\{y : F(y) > t\} \leq x \iff \exists y^* \leq x, \exists y_n \searrow y^*, t.q. F(y_n) > t \forall n$$

Dado que F es una función de distribución, es continua por la derecha, entonces $F(y^*) \geq t$. Ocupando que F es creciente, tenemos:

$$y^* \leq x \wedge F(y^*) \geq t \Rightarrow F(x) \geq F(y^*) \geq t$$

Esto prueba la inclusión \subseteq en 5.2.1. Recíprocamente, si $F(x) \geq t$ hay dos casos posibles:

- Si $t = F(x)$, entonces el lado izquierdo de 5.2.1 corresponde al singleton.
- Si $t < F(x)$, por definición de ínfimo:

$$x \geq \inf\{y : F(y) > t\} = F^{-1}(t)$$

Por lo tanto t pertenece al lado izquierdo de 5.2.1 y se concluye la segunda inclusión.

Entonces, usando 5.2.1; sea $Y = F^{-1}(U)$:

$$\mathbb{P}(Y \leq y) = \mathbb{P}(F^{-1}(U) \leq y)$$

Como U es una variable aleatoria absolutamente continua los singleton tienen medida de probabilidad igual a cero. Así:

$$\mathbb{P}(F^{-1}(U) \leq y) = \mathbb{P}(F^{-1}(U) \leq y \vee U = F(y)) \stackrel{5.2.1}{=} \mathbb{P}(U \leq F(y)) = F(y)$$

Luego la función de probabilidad acumulada de Y es F , es decir $Y \sim \mathcal{L}(X)$. ■

En los casos en que la función de distribución F tiene inversa de forma explícita, el lema anterior nos permita, cómodamente, simular realizaciones de la variable X .

Ejemplo 2 (variable aleatoria exponencial.) Para $X \sim \exp(\lambda)$, $F(x) = 1 - e^{-x} \forall x \geq 0$, luego:

$$F^{-1}(t) = -\frac{\log(1-t)}{\lambda}$$

Por el lema anterior; para $U \sim \text{unif}(0, 1)$.

$$F^{-1}(U) = -\frac{\log(1-U)}{\lambda} \sim \exp(\lambda)$$

Finalmente, si $U \sim \text{unif}(0, 1)$ entonces la variable aleatoria $1 - U$ también es uniforme en el intervalo $[0, 1]$, por lo tanto:

$$F^{-1}(U) \stackrel{d}{=} -\frac{\log(U)}{\lambda} \sim \exp(\lambda)$$

Una de las variables más importantes de simular son las variables aleatorias normales. Lamentablemente, el lema anterior no nos es de utilidad pues no conocemos una forma explícita de la función de distribución, sin embargo podemos hacer lo siguiente.

Si $U, V \sim \text{unif}(0, 1)$ independientes, entonces:

$$X := \sqrt{-2\log(U)}\cos(2\pi V) \quad , \quad Y := \sqrt{-2\log(U)}\sin(2\pi V)$$

son variables aleatoria con distribución $\mathcal{N}(0, 1)$ independientes (**ejercicio, cambio de variable**). Luego, para generar $\mathcal{N}(\mu, \sigma^2)$, basta ponderar y trasladar de la forma $\sigma X + \mu$.⁷

⁷Los software ocupan algoritmos de mayor eficiencia para generar variables aleatorias normales, por ejemplo el algoritmo de Ziggurat.

5.3 Método de Aceptación-Rechazo

Supongamos que conocemos X , una variable aleatoria con función de densidad f conocida, pero sin F^{-1} explícita. El objetivo es poder simular realizaciones de X , ¿cómo podemos simular X ? Supongamos que existe Y , una variable aleatoria eficientemente simulable, con función de densidad g que cumpla lo siguiente:

$$\begin{aligned} \exists k > 0, \quad f(x) &\leq kg(x) \quad \forall x \\ g(x) > 0 &\iff f(x) > 0 \end{aligned}$$

Entonces, $\forall x$ tal que $g(x) > 0$, definimos:

$$\alpha(x) = \frac{f(x)}{kg(x)} \quad (5.3.1)$$

Proposición 4 Sean $(Y_n, U_n)_n$ una secuencia de realizaciones independientes, tal que $U_n \sim \text{unif}(0, 1)$, $Y_n \stackrel{d}{=} Y$ y U_n es independiente de Y_n para todo n . Definimos la variable aleatoria N como:

$$N = \inf\{k \in \mathbb{N} : U_k \leq \alpha(Y_k)\}$$

Entonces $N < \infty$ c.s. y $Y_N \stackrel{d}{=} X$.

Demostración: Veamos que $N < \infty$ \mathbb{P} -c.s. Para ello, notemos que:

$$\mathbb{P}(N = \infty) = \mathbb{P}(\forall m, U_m > \alpha(Y_m)) = \lim_{n \rightarrow \infty} \mathbb{P}(\forall m = 1, \dots, n : U_m > \alpha(Y_m))$$

Como las realizaciones son *i.i.d.* tenemos que:

$$\lim_{n \rightarrow \infty} \mathbb{P}(\forall m = 1, \dots, n : U_m > \alpha(Y_m)) = \lim_{n \rightarrow \infty} [\mathbb{P}(U_m > \alpha(Y_m))]^n = 0$$

Donde la última igualdad sale de que $\mathbb{P}(U > \alpha(Y)) < 1$ (con la desigualdad estricta).

En efecto, sea $p = \mathbb{P}(U \leq \alpha(Y))$, usando probabilidades totales:

$$\begin{aligned} p &= \mathbb{P}(U \leq \alpha(Y)) = \int_{\mathbb{R}} \mathbb{P}(U \leq \alpha(y))g(y)dy = \int_{\mathbb{R}} \left(\int_0^{\alpha(y)} dx \right) g(y)dy \\ &= \int_{\mathbb{R}} \alpha(y)g(y)dy = \int_{\mathbb{R}} \frac{f(y)}{kg(y)}g(y)dy = \frac{1}{k} \int_{\mathbb{R}} f(y)dy = \frac{1}{k} > 0 \end{aligned}$$

Como $p > 0$, entonces $1 - p = \mathbb{P}(U > \alpha(Y)) < 1$ y concluimos que $N < \infty$ \mathbb{P} -c.s.

Ahora veamos que $Y_N \stackrel{d}{=} X$. $\forall A \subseteq \mathcal{B}(\mathbb{R})$ tenemos que:

$$\begin{aligned} \mathbb{P}(Y_N \in A) &= \sum_{n=1}^{\infty} \mathbb{P}(Y_n \in A, N = n) = \sum_{n=1}^{\infty} \mathbb{P}(Y_n \in A, U_n \leq \alpha(Y_n), U_m > \alpha(Y_m), \forall m = 1, \dots, n-1) \\ &= \sum_{n=1}^{\infty} (1-p)^{n-1} \mathbb{P}(Y_n \in A, U_n \leq \alpha(Y_n)) \\ &= \sum_{n=1}^{\infty} (1-p)^{n-1} \int_{\mathbb{R}} \mathbb{P}(y \in A, U_n \leq \alpha(y))g(y)dy \end{aligned}$$

Recordando que U_n es uniforme, y (por lo visto al principio de la demostración) $p = 1/k$:

$$\begin{aligned} \mathbb{P}(Y_N \in A) &= \sum_{n=1}^{\infty} (1-p)^{n-1} \int_{\mathbb{R}} \mathbb{P}(y \in A, U_n \leq \alpha(y))g(y)dy \\ &= \sum_{n=1}^{\infty} (1-p)^{n-1} \int_{\mathbb{R}} \mathbb{1}_{y \in A} \alpha(y)g(y)dy \end{aligned}$$

$$\begin{aligned}
&= \sum_{n=1}^{\infty} (1-p)^{n-1} \int_{\mathbb{R}} \mathbb{1}_{y \in A} \frac{f(y)}{kg(y)} g(y) dy \\
&= \sum_{n=1}^{\infty} (1-p)^{n-1} p \int_{\mathbb{R}} \mathbb{1}_{y \in A} f(y) dy \\
&= \mathbb{P}(X \in A) \sum_{n=1}^{\infty} p(1-p)^{n-1} = \mathbb{P}(X \in A)
\end{aligned}$$

■

Entonces, definiendo $\alpha(x) = \frac{f(x)}{kg(x)}$ para todo x tal que $g(x) > 0$. Explicamos el método de Aceptación-Rechazo mediante el siguiente algoritmo:

Algoritmo Aceptación-Rechazo: para simular la variable X .

1. Simular $U \sim \text{unif}(0, 1)$, simular $Y \sim g$. Ambas (U e Y) independientes entre si.
2. Si $U \leq \alpha(Y)$, $X = Y$ es un muestreo de X .
3. Si $U > \alpha(Y)$, volver al paso 1 e iterar.

Observación: De la demostración previa notamos que; $N \sim \text{geom}(1/k)$. Entonces el valor esperado de la cantidad de iteraciones necesarias para simular un valor de X es $\mathbb{E}(N) = k$. Luego, para que el método termine en un tiempo razonable, es necesario ajustar el valor k lo más cercano a 1 posible, es decir; elegir una densidad g suficientemente parecida a f .

Observación: El método anterior es válido, aún tomando densidades con respecto a una medida μ cualquiera. Es decir, para f y g densidades de X e Y , respectivamente, que cumplen las condiciones:

$$\mathbb{P}(X \in A) = \int_A f(x) \mu(dx) \leq \int_A kg(x) \mu(x) = k \mathbb{P}(Y \in A)$$

En particular, si X es una variable aleatoria discreta, digamos en \mathbb{N} , entonces $f(x) = \mathbb{P}(X = x)$ es la densidad de X con respecto a la medida *cuenta-puntos*. Luego, si Y es otra variable aleatoria discreta que cumple las hipótesis con su densidad, es posible aplicar el método.

5.4 Reducción de varianza

Recordemos que el método de Monte Carlo (M.M.C.) aproxima un valor

$$\alpha = \mathbb{E}(X) \approx \frac{1}{n} \sum_{i=1}^n X_i \quad (5.4.1)$$

donde X_1, \dots, X_n son copias *i.i.d.* de X . Además sabíamos que el error es del orden de $\frac{\sigma}{\sqrt{n}}$, con $\sigma^2 = \text{Var}(X)$.

Notamos que el valor por estimar en 5.4.1 sólo depende de la esperanza de la variable X y no en la variable en sí o su distribución, por lo tanto tenemos la libertad de cambiar la variable a utilizar por una que tenga menor varianza, así estimaremos el mismo valor con un error menor.

5.4.1 Muestreo preferencial

Sea X con densidad f . Sea el valor a estimar

$$\alpha = \mathbb{E}(g(X))$$

donde $g : \mathbb{R} \rightarrow \mathbb{R}$ es una función dada.

Ahora, supongamos que tenemos Y , otra variable aleatoria con densidad $\tilde{f} > 0$, entonces:

$$\alpha = \mathbb{E}(g(X)) = \int g(x) f(x) dx = \int \frac{g(x) f(x)}{\tilde{f}(x)} \tilde{f}(x) dx$$

Si definimos la función h como:

$$h(x) = \frac{g(x)f(x)}{\tilde{f}(x)}$$

Entonces α es la esperanza de la función h bajo la densidad \tilde{f} , es decir $\alpha = \mathbb{E}(h(Y))$. Además:

$$\begin{aligned} \text{Var}(h(Y)) &= \mathbb{E}(h(Y)^2) - \alpha^2 \\ &= \int \frac{g(x)^2 f(x)^2}{\tilde{f}(x)^2} \tilde{f}(x) dx - \alpha^2 \\ &= \int \frac{f(x)^2 g(x)^2}{\tilde{f}(x)} dx - \alpha^2 \end{aligned}$$

Si $g \geq 0$, entonces si escogemos

$$\tilde{f}(x) = \frac{g(x)f(x)}{\alpha} \quad (5.4.2)$$

entonces, obtenemos $\text{Var}(h(Y)) = 0$.

Evidentemente, 5.4.2 no funciona en la práctica, pues no conocemos el valor de α . Sin embargo, nos sugiere que sería óptimo tomar una función \tilde{f} suficientemente parecido, en cuanto a forma, a $g(x)f(x)$ (diferenciado por una ponderación constante).

Esto motiva el siguiente procedimiento heurístico, escogeremos \tilde{f} cumpliendo:

- $\tilde{f}(x)$ "parecido" a $|g(x)f(x)|$, normalizado para que integre 1.
- Sea fácil de simular una variable Y , con densidad \tilde{f} .

Luego basta aproximar $\alpha \approx \frac{1}{n} \sum_{i=1}^n h(Y_i)$, con Y_1, \dots, Y_n copias *i.i.d.* de Y .

Ejemplo: Queremos aproximar:

$$I = \int_0^1 \cos\left(\frac{\pi x}{2}\right) dx = \frac{2}{\pi}$$

Suponiendo que no podemos calcular esta integral, definiendo:

$$I = \mathbb{E}(g(X)), \quad g(x) = \cos\left(\frac{\pi x}{2}\right)$$

con X una variable aleatoria uniforme en el intervalo $[0, 1]$, podemos calcular I mediante *M.M.C.*

En este caso, podríamos reemplazar la función *coseno* por un polinomio de grado 2. Dado que el interior de la integral es 0 en $x = 1$ y 1 en $x = 0$, es natural elegir $\tilde{f}(x)$ de la forma $\lambda(1 - x^2)$. Si normalizamos la función en este intervalo, tenemos que $\tilde{f}(x) = 3(1 - x^2)/2$. Al calcular las varianzas, podemos notar que el método reduce la varianza en un factor de 100.

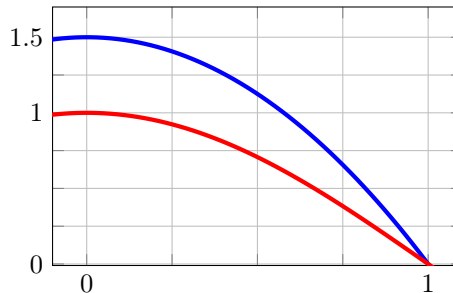


Figure 5: En rojo la función $\cos(\pi x/2)$, en azul la función $3(1 - x^2)/2$ mayorándola.

5.4.2 Variable de control

Este método involucra escribir $\mathbb{E}(f(X))$ de la forma:

$$\mathbb{E}(f(X)) = \mathbb{E}(f(X) - h(X)) + \mathbb{E}(h(X))$$

Donde $\mathbb{E}(h(X))$ puede ser calculado explícitamente, y $\text{Var}(f(X) - h(X))$ es significativamente menor que $\text{Var}(f(X))$. Así, empleamos el método de Monte Carlo para el cálculo de $\mathbb{E}(f(X) - h(X))$ y el cálculo directo de $\mathbb{E}(h(X))$.

Ejemplo: Supongamos que deseamos calcular la siguiente integral:

$$I = \int_0^1 e^x dx = e - 1$$

Aplicando el *M.M.C.*, con $I = \mathbb{E}(e^X) = \mathbb{E}(g(X))$, con $X \sim \text{unif}(0, 1)$. Notamos que la varianza de la función es $\text{Var}(g(X)) \approx 0,242$.

Dado que, cerca de $x = 0$, $e^x \approx 1 + x$, podemos escribir:

$$\int_0^1 e^x dx = \int_0^1 (e^x - 1 - x) dx + \frac{3}{2}$$

En este caso, elegimos $h(x) = 1 + x$. Es fácil comprobar que $\text{Var}(g(x) - h(x)) \approx 0.0437$. Es decir, hemos reducido la varianza de la estimación en un factor de 5.5 aproximadamente.

5.4.3 Variables antitéticas

Supongamos que deseamos calcular la siguiente integral;

$$I = \int_0^1 f(x) dx$$

Dado que el cambio de variable: $x \rightarrow 1 - x$ es invariante ante la medida dx en $[0, 1]$. Es decir,

$$\int_0^1 f(x) dx = \int_0^1 f(1 - x) dx$$

Entonces es posible reescribir I de la siguiente forma:

$$I = \frac{1}{2} \int_0^1 (f(x) + f(1 - x)) dx$$

Otra forma de plantear el problema, es definir la variable aleatoria $X \sim \text{unif}(0, 1)$. Así, $1 - X \sim \text{unif}(0, 1)$, entonces:

$$I = \mathbb{E}(f(X)) = \mathbb{E}\left(\frac{f(X) + f(1 - X)}{2}\right)$$

y obtenemos la misma conclusión.

Ahora, podemos calcular I con *M.M.C.*. Simulamos n variables aleatorias *i.i.d.* uniformes en $[0, 1]$, U_1, \dots, U_n , y aproximamos I por:

$$\begin{aligned} I_{2n} &= \frac{1}{n} \left(\frac{1}{2} (f(U_1) + f(1 - U_1)) + \dots + (f(U_n) + f(1 - U_n)) \right) \\ &= \frac{1}{2n} (f(U_1) + f(1 - U_1) + \dots + f(U_n) + f(1 - U_n)) \end{aligned}$$

A grandes razgos, este método es equivalente a emplear el método tradicional de Monte Carlo, con la variable $\frac{f(U) + f(1 - U)}{2}$. Podemos observar que su varianza:

$$\text{Var}\left(\frac{f(U) + f(1 - U)}{2}\right) = \frac{1}{4} \text{Var}(f(U) + f(1 - U))$$

$$= \frac{1}{4} (Var(f(U)) + 2Cov(f(U), f(1-U)) + Var(f(1-U)))$$

Dado que $X \stackrel{d}{=} 1 - X$:

$$Var\left(\frac{f(U) + f(1-U)}{2}\right) = \frac{1}{2} [Var(f(U)) + Cov(f(U), f(1-U))]$$

Así, si $Cov(f(U), f(1-U)) \leq 0$, se reduce la varianza de la estimación en al menos un factor de $1/2$.

En general, tenemos que la aproximación mejora siempre que:

$$\mathbb{E}(f(U)f(U-1)) < \mathbb{E}(f(U)^2)$$

Lema 6 Si f es una función monótona, entonces:

$$Cov(f(U), f(1-U)) \leq 0$$

5.4.4 Método de Estratificación

Este método es bien conocido en el diseño de muestras de encuestas. Supongamos que se busca calcular

$$I = \mathbb{E}(g(X)) = \int g(x)f(x)dx$$

Donde X tiene ley $f(x)dx$. Consideremos D_1, \dots, D_m , una partición del soporte de la densidad f . Luego, podemos descomponer I en

$$I = \sum_{i=1}^m I_i = \sum_{i=1}^m \mathbb{E}(\mathbb{1}_{\{X \in D_i\}} g(X)) \quad (5.4.3)$$

Este método propone estimar I_i con *M.M.C.*, usando n_i simulaciones. Definimos como $\sigma_i^2 = Var(\mathbb{1}_{\{X \in D_i\}} g(X))$. Entonces, la varianza de la aproximación es

$$\sum_{i=1}^m \frac{\sigma_i^2}{n_i} \quad (5.4.4)$$

Si minimizamos 5.4.4, bajo la restricción de un número de simulaciones $\sum_{i=1}^m n_i = n$ dado, obtenemos $n_i = n\sigma_i / \sum_{i=1}^m \sigma_i$, con un valor mínimo de la varianza de

$$\frac{(\sum_{i=1}^m \sigma_i)^2}{n}$$

Puede probarse que esta varianza es menor que la obtenida con n simulaciones bajo el método estándar de Monte Carlo directamente en 5.4.3. Por supuesto, rara vez es posible calcular el valor de σ_i , lo que limita el uso de esta técnica (pero bien uno puede estimarlos vía su estimador insesgado de varianza mínima $\frac{1}{n-1} \sum (x_i - \bar{x})^2$, o bajo el propio *M.M.C.*).

6 Cadenas de Markov

Una cadena de Markov es una sucesión de variables aleatorias $\{X_n; n = 0, 1, 2, \dots\}$, definidas sobre algún espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$, tomando valores en un conjunto E arbitrario, pero que supondremos que será finito o numerable, y que posee la propiedad de Markov. Intuitivamente, una cadena de Markov posee la cualidad de que, conociendo el estado presente X_n , sin necesidad de conocer los estados anteriores puede hacer buenas predicciones acerca de los estados futuros. Los siguientes temas presentan variadas aplicaciones de las cadenas de Markov. Tenga en cuenta que restringiremos el contenido a cadenas de Markov *homogéneas*, aunque el caso no homogéneo sea necesario en muchas aplicaciones.

6.1 Definiciones y propiedades elementales

Definición 7 Sea E un conjunto finito o numerable. Un proceso estocástico $\{X_n; n \in \mathbb{N}\}$ es llamado Cadena de Markov si, para todo $n \in \mathbb{N}$, la ley condicional de X_{n+1} dado X_0, X_1, \dots, X_n , es igual a la ley condicional de X_{n+1} dado X_n .

Esto es; $\forall x_0, \dots, x_n \in E$:

$$\mathbb{P}(X_{n+1} = x_{n+1} \mid X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n)$$

La cadena se dirá 'Homogénea', si lo anterior no depende del índice n . En cuyo caso, la matriz $P = (P_{xy})_{x,y \in E}$, con $P_{x,y} = \mathbb{P}(X_{n+1} = y \mid X_n = x)$

Además, llamamos $\mu \in \mathcal{P}(E)$ a $\mu_x = \mathbb{P}(X_0 = x)$, denominada 'Distribución inicial'. Así, decimos que $(X_n)_{n \in \mathbb{N}}$ es una (μ, P) -cadena de Markov homogénea.

Ejemplo 3 En $E = \mathbb{Z}$, si Z_0, Z_1, Z_2, \dots son las ganancias de apuestas modeladas como variables aleatorias i.i.d., entonces la ganancia total acumulada $X_n = \sum_{i=0}^n Z_i$ es una cadena de Markov (C.M.).

Un simple criterio, que nos permitirá en muchos casos verificar si un proceso dado es C.M., está dado por el siguiente lema.

Lema 7 Sean E y F dos conjuntos numerables, y sea f un mapeo de $\mathbb{N} \times E \times F$ hacia E . Sean X_0, Y_1, Y_2, \dots , son variables aleatorias mutuamente independientes, siendo X_0 valor en E e Y_n valores en F .

Sean $\{X_n; n \geq 1\}$ un proceso en E , definido como:

$$X_{n+1} = f(n, X_n, Y_{n+1}), \quad n \in \mathbb{N}$$

Entonces $\{X_n; n \in \mathbb{N}\}$ es cadena de Markov.

Demostración:

$$\mathbb{P}(X_{n+1} = x_{n+1} \mid X_0 = x_0, \dots, X_n = x_n) = \frac{\mathbb{P}(X_0 = x_0, \dots, X_n = x_n, X_{n+1} = x_{n+1})}{\mathbb{P}(X_0 = x_0, \dots, X_n = x_n)}$$

$$= \sum_{\{z; f(n, x_n, z) = x_{n+1}\}} \frac{\mathbb{P}(X_0 = x_0, \dots, X_n = x_n, Y_{n+1} = z)}{\mathbb{P}(X_0 = x_0, \dots, X_n = x_n)}$$

Dado que Y_{n+1} es una variable independiente a todos los valores $\{X_0, \dots, X_n\}$, entonces:

$$\begin{aligned} &= \sum_{\{z; f(n, x_n, z) = x_{n+1}\}} \frac{\mathbb{P}(X_0 = x_0, \dots, X_n = x_n, Y_{n+1} = z)}{\mathbb{P}(X_0 = x_0, \dots, X_n = x_n)} \\ &= \sum_{\{z; f(n, x_n, z) = x_{n+1}\}} \frac{\mathbb{P}(X_0 = x_0, \dots, X_n = x_n) \mathbb{P}(Y_{n+1} = z)}{\mathbb{P}(X_0 = x_0, \dots, X_n = x_n)} \\ &= \sum_{\{z; f(n, x_n, z) = x_{n+1}\}} \mathbb{P}(Y_{n+1} = z) \end{aligned}$$

$$= \frac{\mathbb{P}(X_n = x_n, X_{n+1} = x_{n+1})}{\mathbb{P}(X_n = x_n)} = \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n)$$

■

Una cadena de Markov es el análogo de una sucesión determinista determinada por una relación de recurrencia del tipo

$$x_{n+1} = f(n, x_n),$$

opuesto al sistema 'con memoria', del tipo

$$x_{n+1} = f(n, x_n, x_{n-1}, \dots, x_0).$$

Acá la función $f(n, \cdot)$ es reemplazada por la matriz de transición

$$P_{xy} = \mathbb{P}(X_{n+1} = y | X_n = x).$$

Hasta el momento, podemos asumir que la matriz $P = (P_{xy} : x, y \in E)$ es independiente de la variable tiempo n , una vez dicho que la cadena es *homogénea*.

Lema 8 Sea $f : E \times [0, 1] \rightarrow E$ medible, sea X_0 variable aleatoria en E , Y_1, Y_2, \dots , variables aleatorias uniformes en $[0, 1]$, todas independientes entre sí (y de X_0). Entonces $(X_n)_{n \in \mathbb{N}}$ definido recursivamente por $X_{n+1} = f(X_n, Y_{n+1})$ es una C.M.H. (cadena de Markov homogénea).

La matriz P es llamada *Markoviana* (o *estocástica*), en el sentido en que tiene la siguiente propiedad; para cada $x \in E$, el vector fila $(P_{xy} ; y \in E)$ es una medida de probabilidad en E , en otras palabras;

$$P_{xy} \geq 0, \quad \forall y \in E; \quad \sum_{y \in E} P_{xy} = 1$$

El vector de distribución inicial, definido por $\mu_x = \mathbb{P}(X_0 = x)$, se entiende como un vector fila, de modo que $(\mu P)_y = \sum_{x \in E} \mu_x P_{xy}$.

Proposición 5 Para una (μ, P) -C.M.H. se cumplen:

- i) $\mathbb{P}(X_0 = x_0, \dots, X_n = x_n) = \mu_{x_0} P_{x_0 x_1} P_{x_1 x_2} \dots P_{x_{n-1} x_n}$. (Esta propiedad es equivalente a la definición de (μ, P) - C.M.H.)
- ii) $\mathbb{P}(X_{n+1} = x_{n+1}, \dots, X_{n+m} = x_{n+m} | X_0 = x_0, \dots, X_n = x_n) = P_{x_{n+1} x_{n+2}} \dots P_{x_{n+m-1} x_{n+m}}$.
- iii) $\mathbb{P}(X_n = y | X_0 = x) = (P^n)_{xy}$ (Donde la notación anterior hace referencia a la potencia matricial).
- iv) $\mathbb{P}(X_n = y) = (\mu P^n)_y$
- v) Para $g : E \rightarrow \mathbb{R}$,

$$\mathbb{E}(g(X_n) | X_0 = x) = (P^n g)_x = \sum_{y \in E} P_{xy}^n g(y)$$

$$vi) \mathbb{E}(g(X_n)) = \mu P^n g = \sum_{x, y \in E} \mu_x P_{xy}^n g_y$$

(En esta proposición hacemos uso de la notación $g_y = g(y)$).

Demostración: [3, págs. 20,21]

6.2 Propiedad de Markov fuerte

Para la mejor comprensión de esta parte, recordemos la propiedad de Markov. Sea $\{X_n; n \in \mathbb{N}\}$ una cadena de Markov a valores en E definida sobre el espacio de probabilidades $(\Omega, \mathcal{F}, \mathbb{P})$. Dada una medida de probabilidad μ sobre E , usaremos la notación \mathbb{P}_μ para denotar cualquier probabilidad sobre (Ω, \mathcal{F}) tal que, bajo \mathbb{P}_μ , la sucesión $\{X_n; n \geq 0\}$ es una cadena de Markov con ley inicial μ ; en otras palabras, μ es la ley de X_0 , esto es,

$$\mathbb{P}_\mu(X_0 = x) = \mu_x, \quad x \in E.$$

para el caso $\mu = \delta_x$, escribiremos \mathbb{P}_x en vez de \mathbb{P}_{δ_x} .

\mathbb{P}_x puede ser interpretado como la ley condicional de X , dado $X_0 = x$. Para cualquier $n \geq 0$, definimos \mathcal{F}_n como la sigma-algebra de los eventos que están determinados por las variables X_0, X_1, \dots, X_n , esto es,

$$\mathcal{F}_n = \left\{ \{\omega; (X_0(\omega), \dots, X_n(\omega)) \in B_n\}; B_n \in 2^{E^{n+1}} \right\}$$

Donde 2^F denota la colección de todos los subconjuntos del conjunto F .

Teorema 13 Sea $\{X_n; n \geq 0\}$ una (μ, P) cadena de Markov. Entonces, para cada $n \in \mathbb{N}$, $x \in E$, condicionalmente a $\{X_n = x\}$, $\{X_{n+p}; p \geq 0\}$ es una (δ_x, P) -cadena de Markov independiente de (X_0, X_1, \dots, X_n) . En otras palabras, para todo $A \in \mathcal{F}_n$ y todo $m > 0$, $x_1, \dots, x_m \in E$,

$$\begin{aligned} \mathbb{P}(A \cap \{X_{n+1} = x_1, \dots, X_{n+m} = x_m\} | X_n = x) \\ = \mathbb{P}(A | X_n = x) \mathbb{P}_x(X_1 = x_1, \dots, X_m = x_m) \end{aligned}$$

Demostración: Es suficiente probar el resultado en el caso donde $A = \{X_0 = y_0, X_1 = y_1, \dots, X_n = y_n\}$ (A es unión finita o numerable de conjuntos disjuntos de esta forma, y el resultado en el caso general saldrá de la σ -aditividad de \mathbb{P}). Es suficiente considerar el caso $y_n = x$. Así, el lado izquierdo de la igualdad es equivalente a

$$\frac{\mathbb{P}(X_0 = y_0, \dots, X_n = x, X_{n+1} = x_1, \dots, X_{n+m} = x_m)}{\mathbb{P}(X_n = x)},$$

así, aplicando la propiedad anterior iterativamente, esto es igual a

$$\frac{\mathbb{P}(A)}{\mathbb{P}(X_n = x)} \times P_{xx_1} \times P_{x_1x_2} \times \dots \times P_{x_{m-1}x_m},$$

o, en otras palabras,

$$\mathbb{P}(A | X_n = x) \mathbb{P}_x(X_1 = x_1, \dots, X_m = x_m).$$

■

El último resultado nos dice en particular que, el pasado y el futuro de una cadena son condicionalmente independientes, dada la posición de la cadena en el tiempo presente n .

Ahora, nos gustaría expandir la propiedad de Markov, reemplazando el tiempo dado n por un tiempo aleatorio (pero no cualquier tiempo aleatorio).

Definición 8 Una variable aleatoria T que toma valores en el conjunto $\mathbb{N} \cup \{+\infty\}$ es llamada tiempo de parada (t.d.p.) si, para todo $n \in \mathbb{N}$,

$$\{T = n\} \in \mathcal{F}_n.$$

En otras palabras, las observaciones X_1, X_2, \dots, X_n , la trayectoria de la cadena en el tiempo n , es suficiente para decidir si T es igual a n .

Ejemplo 4 1. Para cada $x \in E$, el tiempo de llegada al estado x ,

$$S_x = \begin{cases} \inf\{n \geq 0; X_n = x\} & , \text{ si existe } n \text{ que cumpla.} \\ +\infty & , \text{ otro caso.} \end{cases}$$

y el tiempo del primer retorno al estado x ,

$$T_x = \begin{cases} \inf & , \text{ si existe } n \text{ que lo cumpla.} \\ +\infty & , \text{ en otro caso.} \end{cases}$$

son tiempo de parada⁸. En el caso de T_x esto se cumple porque

$$\{T_x = n\} = \{X_1 \neq x\} \cap \cdots \cap \{X_{n-1} \neq x\} \cap \{X_n = x\}.$$

2. $\forall A \subseteq E$, $T_A := \inf\{n \geq 0 : X_n \in A\}$ es el tiempo de llegada al conjunto A . T_A es t.d.p.
3. $L_A := \sup\{n \geq 1 : X_n \in A\}$ es la última vez que la cadena visita A . No es tiempo de parada, puesto que necesitamos conocer la trayectoria luego del tiempo n para decidir si $L_A = n$, o no.

Dado un tiempo de parada T , anotamos

$$\mathcal{F}_T := \{B \in \mathcal{F} \mid B \cap \{T = n\} \in \mathcal{F}_n, \}$$

Donde $(\Omega, \mathcal{F}, \mathbb{P})$ es el espacio de probabilidades subyacente.

(**Propuesto:** \mathcal{F}_T es una σ -álgebra)

Teorema 14 (Propiedad de Markov-fuerte.) Sea $\{X_n; n \geq 0\}$ una (μ, P) -C.M., y T un tiempo de parada. Condicionado a $\{T < \infty\} \cap \{X_T = x\}$, $\{X_{T+n}; n \geq 0\}$ es una (δ_x, P) -C.M. que es independiente de \mathcal{F}_T .

En otras palabras; para todo $A \in \mathcal{F}_T$ y todo $m > 0$, $x_1, \dots, x_m \in E$,

$$\begin{aligned} & \mathbb{P}(A \cap \{X_{T+1} = x_1, \dots, X_{T+m} = x_m\} \mid X_T = x, T < \infty) \\ &= \mathbb{P}(A \mid X_T = x, T < \infty) \times \mathbb{P}_x(X_1 = x_1, \dots, X_m = x_m). \end{aligned}$$

Demostración: Es suficiente mostrar que, para todo $n \in \mathbb{N}$,

$$\begin{aligned} & \mathbb{P}(A \cap \{T = n\} \cap \{X_{T+1} = x_1, \dots, X_{T+m} = x_m\} \mid X_T = x) \\ &= \mathbb{P}(A \cap \{T = n\} \mid X_T = x) \mathbb{P}_x(X_1 = x_1, \dots, X_m = x_m), \end{aligned}$$

que es resultado directo de la propiedad markoviana.

■

6.3 Estados Recurrentes y Transientes

Definimos $T_x = \inf\{n \geq 1; X_n = x\}$ como en el ejemplo anterior.

Definición 9 $x \in E$ se dice recurrente si $\mathbb{P}_x(T_x < \infty) = 1$, y transiente en el caso contrario (es decir, si $\mathbb{P}_x(T_x < \infty) < 1$).

Definimos además, el número de retornos al estado x :

$$N_x = \sum_{n \geq 1} \mathbb{1}_{\{X_n = x\}}.$$

Proposición 6 a) Si x es recurrente, entonces

$$\mathbb{P}_x(N_x = +\infty) = 1.$$

⁸Con la convención de que el ínfimo de un conjunto vacío es $+\infty$, basta con escribir: $T_x = \inf\{n \geq 1; X_n = x\}$.

b) Si x es transiente, entonces

$$\mathbb{P}_x(N_x = k) = (1 - p_x) p_x^k, \quad k \geq 0,$$

donde $p_x = \mathbb{P}_x(T_x < \infty)$ (en particular, $N_x < \infty$, \mathbb{P}_x -c.s.).

Demostración: Sea

$$\begin{aligned} T_x^2 &= \inf\{n > T_x; X_n = x\} \\ &= T_x + \inf\{n \geq 1; X_{T_x+n} = x\}. \end{aligned}$$

No es difícil ver que T_x^2 es un tiempo de parada:

$$\begin{aligned} \mathbb{P}_x(T_x^2 < \infty) &= \mathbb{P}_x(T_x^2 < \infty | T_x < \infty) \mathbb{P}_x(T_x < \infty) \\ &= \sum_{n=1}^{\infty} \mathbb{P}_x(T_x^2 = T_x + n | T_x < \infty) \mathbb{P}_x(T_x < \infty). \end{aligned}$$

Pero, por la propiedad de Markov fuerte, deducimos que

$$\begin{aligned} &{}_x(T_x^2 = T_x + n | T_x < \infty) \\ &= \mathbb{P}_x(X_{T_x+1} \neq x, \dots, X_{T_x+n-1} \neq x, X_{T_x+n} = x | T_x < \infty) \\ &= \mathbb{P}_x(X_1 \neq x, \dots, X_{n-1} \neq x, X_n = x) \\ &= \mathbb{P}_x(T_x = n). \end{aligned}$$

Finalmente,

$$\mathbb{P}_x(T_x^2 < \infty) = (\mathbb{P}_x(T_x < \infty))^2$$

o

$$\mathbb{P}_x(N_x \geq 2) = (\mathbb{P}_x(T_x < \infty))^2.$$

Así, iterando el mismo argumento, deducimos que ambas afirmaciones de la proposición se deducen fácilmente de esta identidad.

■

Corolario 3 x es recurrente si y sólo si

$$\sum_{n=0}^{\infty} (P^n)_{xx} = +\infty.$$

Demostración: Notemos que

$$\begin{aligned} \mathbb{E}_x(N_x) &= \mathbb{E}_x\left(\sum_{n \geq 1} \mathbb{1}_{\{X_n = x\}}\right) = \sum_{n \geq 1} \mathbb{P}_x(X_n = x) \\ &= \sum_{n \geq 1} (P^n)_{xx}. \end{aligned}$$

Si x es recurrente, entonces $N_x = \infty$ \mathbb{P}_x -c.s., luego $\mathbb{E}(N_x) = \infty$, probando (\Rightarrow).

Recíprocamente, si x es transiente, por la parte **b)** de la proposición anterior, se tiene que $\mathbb{E}_x(N_x) < \infty$, con lo cual se prueba (\Leftarrow).

■

Definición 10 Decimos que un estado y es accesible desde el estado x (y anotamos $x \rightarrow y$), siempre que exista $n \geq 0$ tal que $\mathbb{P}_x(X_n = y) > 0$. Decimos que x e y se comunican (y escribimos $x \leftrightarrow y$) cuando $y \rightarrow x$, además de $x \rightarrow y$.

La relación $x \leftrightarrow y$ es una relación de equivalencia, entonces podemos particionar el conjunto E en clases de equivalencia definidas por la relación \leftrightarrow .
 Notamos que $x \leftrightarrow y \iff \exists n \geq 0$ tal que $(P^n)_{xy} > 0$, dado que $\mathbb{P}_x(X_n = y) = (P^n)_{xy}$.

Teorema 15 Sea $C \subset E$ una clase de equivalencia de la relación \leftrightarrow . Entonces todos los estados de C son recurrentes, o bien, son todos transientes.

Demostración: [3, págs. 26- Teo.4.5]

Luego, las clases de equivalencia definidas por \leftrightarrow , se les llaman clases de recurrencia.

Definición 11 Una (μ, P) -C.M. se dice irreducible cuando E consiste sólo de una clase de equivalencia. Se dirá irreducible y recurrente si es reducible y además esa clase de equivalencia es recurrente.

Proposición 7 Toda cadena de Markov irreducible sobre un espacio de estados E finito, es recurrente.

Demostración: Dado que E es finito, al menos un estado debe ser visitado infinitas veces con probabilidad positiva, por lo tanto, casi seguramente, por la proposición anterior, este estado (y todos los de la clase) es (son) recurrente.

■

6.4 Casos Irreducibles y Recurrentes

En esta sección, asumiremos que una cadena es irreducible y recurrente. Empezaremos estudiando las *excursiones* de una cadena entre dos retornos sucesivos al mismo estado x :

$$\varepsilon_k = (X_{T_x^k}, X_{T_x^k+1}, \dots, X_{T_x^{k+1}}), \quad k \geq 0.$$

Tales excursiones son secuencias aleatorias de largo aleatorio, al menos 2 y finito, compuesta de elementos $E \setminus \{x\}$, excepto por el primero y el último, que son iguales a x . Denotamos como U al conjunto de secuencias

$$u = (x, x_1, \dots, x_n, x),$$

con $n \geq 0$, $x_l \neq x$, $1 \leq l \leq n$. U es numerable, y es el conjunto de todas las posibles excursiones $\varepsilon_0, \varepsilon_1, \dots$. Por lo tanto, estas variables aleatorias toman valores en un conjunto numerable, y sus leyes de probabilidad están caracterizadas por las cantidades

$$\mathbb{P}(\varepsilon_k = u), \quad u \in U.$$

Proposición 8 Bajo \mathbb{P}_x , la secuencia $(\varepsilon_0, \varepsilon_1, \dots)$ de excursiones es i.i.d., en otras palabras, existen probabilidades $\{p_u; u \in U\}$ sobre U tal que, para todo $k > 0$, $u_0, u_1, \dots, u_k \in U$,

$$\mathbb{P}_x(\varepsilon_0 = u_0, \varepsilon_1 = u_1, \dots, \varepsilon_k = u_k) = \prod_{l=0}^k p_{u_l}.$$

Demostración: Esto es consecuencia de la propiedad fuerte de Markov. En efecto, $\{\varepsilon_0 = u_0\} \in \mathcal{F}_{T_x}$, y el evento

$$\{\varepsilon_1 = u_1, \dots, \varepsilon_k = u_k\}$$

es de la forma

$$\{X_{T_x+1} = x_1, \dots, X_{T_x+p} = x_p\},$$

pára algún $p > 0$, $x_1, \dots, x_p \in E$. En consecuencia,

$$\begin{aligned} & \mathbb{P}_x(\varepsilon_0 = u_0, \varepsilon_1 = u_1, \dots, \varepsilon_k = u_k) \\ &= \mathbb{P}_x(\{\varepsilon_0 = u_0\} \cap \{X_{T_x+1} = x_1, \dots, X_{T_x+p} = x_p\} | T_x < \infty) \end{aligned}$$

$$\begin{aligned}
& \mathbb{P}_x(\varepsilon_0 = u_0) \mathbb{P}_x(X_1 = u_1, \dots, X_p = x_p) \\
&= \mathbb{P}_x(\varepsilon_0 = u_0) \mathbb{P}_x(\varepsilon_0 = u_1, \dots, \varepsilon_{k-1} = u_k) \\
&= \mathbb{P}_x(\varepsilon_0 = u_0) \mathbb{P}_x(\varepsilon_0 = u_1) \cdots \mathbb{P}_x(\varepsilon_{k-1} = u_k) \\
&= p_{u_0} p_{u_1} \cdots p_{u_k},
\end{aligned}$$

donde $\{p_u; u \in U\}$ es la ley de ε_0 bajo \mathbb{P}_x . ■

Una medida (no necesariamente de probabilidad) sobre el conjunto E , es un 'vector fila' $\{\gamma_x; x \in E\}$ tal que $0 \leq \gamma_x < \infty$, para todo x . Cuando la medida es finita, $\sum_{x \in E} \gamma_x < \infty$, podemos normalizarla, haciéndola medida de probabilidad sobre E , $(\gamma_x / \sum_{z \in E} \gamma_z; x \in E)$.

Una medida γ se dice *invariante* (respecto a la matriz de transición P) cuando

$$\gamma P = \gamma,$$

esto es,

$$\sum_{y \in E} \gamma_y P_{yx} = \gamma_x, \quad \forall x \in E.$$

Una medida γ se dice *estríctamente positiva* si $\gamma_x > 0$, para todo $x \in E$.

Una medida γ es invariante si y sólo si la cadena (γ, P) cumple $\mathcal{L}(X_n) = \gamma, \forall n \in \mathbb{N}$. Esto es, γ es la ley de X_n para todo valor de n .

Dado que es para todo n , $\{X_{n+m}; m \in \mathbb{N}\}$ también es (γ, P) -C.M.

Observación: Una **probabilidad** invariante es una medida de probabilidad π , que satisface $\pi P = \pi$, o equivalentemente, para todo $x \in E$,

$$\sum_{y \neq x} \pi_y P_{yx} = \pi_x (1 - P_{xx})$$

esto es,

$$\mathbb{P}(X_n \neq x, X_{n+1} = x) = \mathbb{P}(X_n = x, X_{n+1} \neq x)$$

lo que significa que en el equilibrio, el número medio de salidas desde el estado x entre los tiempo n y $n+1$ es igual al número medio de llegadas al estado x entre los mismos tiempos. Esta es la intuición que caracteriza a la probabilidad invariante.

Teorema 16 Sea $\{X_n; n \in \mathbb{N}\}$ una C.M. irreducible y recurrente. Entonces existe una medida invariante γ estrictamente positiva, cual es única salvo constantes multiplicativas.

Demostración: Para probar la existencia, sea γ_y^x el número esperado de visitas al estado y durante la excursión ε_0 partiendo de x , esto es,

$$\begin{aligned}
\gamma_y^x &= \mathbb{E}_x \left(\sum_{n=1}^{T_x} \mathbb{1}_{X_n=y} \right) \\
&= \sum_{n=1}^{\infty} \mathbb{P}_x(X_n = y, n \leq T_x) \\
&= \sum_{z \in E} \sum_{n=1}^{\infty} \mathbb{P}_x(\{X_{n-1} = z, n-1 < T_x\} \cap \{X_n = y\}) \\
&= \sum_{z \in E} \left(\sum_{n=2}^{\infty} \mathbb{P}_x(X_{n-1} = z, n-1 \leq T_x) \right) P_{zy} \\
&= (\gamma^x P)_y
\end{aligned}$$

Notamos que es necesario usar la recurrencia para llegar a la penúltima igualdad. Ahora haremos uso de la irreducibilidad de la cadena. Existen n, m tal que $(P^n)_{xy} > 0$, $(P^m)_{yx} > 0$. Por lo tanto, dado que $\gamma_x^x = 1$,

$$0 < (P^n)_{xy} = \gamma_x^x (P^n)_{xy} \leq (\gamma^x P^n)_y = \gamma_y^x,$$

$$\gamma_y^x (P^n)_{yx} \leq (\gamma^x P^m)_x = \gamma_x^x = 1.$$

En consecuencia, γ^x es una medida estrictamente positiva, que satisface $\gamma_x^x = 1$.

Para asegurar la unicidad, sea λ una medida invariante tal que $\lambda_x = 1$. Debemos mostrar primero que $\lambda \geq \gamma^x$, luego $\lambda = \gamma^x$. Notemos que esta parte de la demostración usa sólo la irreducibilidad (y no la recurrencia). Tenemos

$$\begin{aligned} \lambda_y &= P_{xy} + \sum_{z_1 \neq x} \lambda_{z_1} P_{z_1 y} \\ &= P_{xy} + \sum_{z_1 n \in qx} P_{xz_1} P_{z_1 y} + \sum_{z_1, z_2 \neq x} \lambda_{z_2} P_{z_2 z_1} P_{z_1 y} \\ &\geq \sum_{n=0}^{\infty} \sum_{z_1, \dots, z_n \neq x} P_{xz_n} P_{z_n z_{n-1}} \cdots P_{z_1 y} \\ &= \sum_{n=0}^{\infty} \mathbb{P}_x(X_{n+1} = y, T_x \geq n+1) \\ &= \gamma_y^x. \end{aligned}$$

Por lo tanto, definamos $\mu = \lambda - \gamma^x$ que también es una medida invariante, y $\mu_x = 0$. Sea $y \in E$, y n tal que $(P^n)_{yx} > 0$. Entonces

$$0 = \mu_x = \sum_{z \in E} \mu_z (P^n)_{zx} \geq \mu_y (P^n)_{yx}.$$

Por lo tanto $\mu_y = 0$, y esto se cumple para todo $y \in E$. ■

Sea $m_x = \mathbb{E}_x(T_x)$. Si esta cantidad es finita, entonces el estado x se dirá *recurrente positivo*, en otro caso se dirá *recurrente nulo*.

Teorema 17 Sea $(X_n)_{n \in \mathbb{N}}$ una C.M. irreducible. Un estado x es recurrente positivo si y sólo si todos los estados son recurrentes positivos, si y sólo si existe π una probabilidad invariante.

Además, π cumple que

$$\pi = (\pi_x = m_x^{-1}, x \in E).$$

Demostración: Notemos que

$$m_x = \sum_{y \in E} \gamma_y^x.$$

Por lo tanto si x es recurrente positivo, entonces la probabilidad $\pi = (\pi_y = \gamma_y^x / m_x, y \in E)$ es invariante.

Por otro lado, si π es una probabilidad invariante, de la irreducibilidad, π es estrictamente positiva⁹, por lo tanto si x es un estado arbitrario, $\lambda = (\lambda_y = \pi_y / \pi_x, y \in E)$ es una medida invariante que satisface $\lambda_x = 1$. De la irreducibilidad y de la demostración de unicidad del teorema anterior,

$$m_x = \sum_{y \in E} \gamma_y^x = \sum_{y \in E} \frac{\pi_y}{\pi_x} = \frac{1}{\pi_x} < \infty.$$

Por lo tanto, x es recurrente positiva, al igual que todos los demás estados. ■

⁹Ver demostración de teorema anterior.

Los dos último teoremas marcan una clara dicotomía en el siguiente sentido: en el caso *irreducible y recurrente*, la cadena es *recurrente positiva* cuando existe una *probabilidad invariante*, y será *recurrente nula* si una (por ende, todas) *medida invariante* tiene masa total infinita ($\sum_i \pi_i = +\infty$). En particular, si $|E| < \infty$, entonces no existe estado recurrente nulo, más bien, cualquier estado recurrente es recurrente positivo.

Corolario 4 Sea $\{X_n\}$ una cadena de Markov irreducible que es recurrente positiva. Para cualquier $x \in E$, asociamos $T_x = \inf\{n > 0; X_n = x\}$. Entonces para todo $y \in E$,

$$\mathbb{E}_y(T_x) < \infty.$$

Demostración: Notemos que

$$T_x \geq T_x \mathbb{1}_{T_y < T_x},$$

de donde, tomando esperanza con respecto a \mathbb{P}_x ,

$$m_x \geq \mathbb{E}_x(T_x | T_y < T_x) \mathbb{P}_x(T_y < T_x).$$

Pero, usando la propiedad fuerte de Markov; $\mathbb{E}_x(T_x | T_y < T_x) > \mathbb{E}_y(T_x)$, y de la irreducibilidad que $\mathbb{P}_x(T_y < T_x) > 0$, y concluimos. ■

Observación: (caso no-irreducible) Por simplicidad, consideremos sólo el caso $|E| < \infty$. Existe al menos una clase recurrente (que es recurrente positiva), por lo tanto existe al menos una probabilidad invariante. Cualquier probabilidad invariante cobra sólo estados recurrentes. Si sólo existe una clase recurrente, entonces la cadena posee sólo una probabilidad invariante. En otro caso, podemos asociarle a cada clase recurrente una única probabilidad invariante con soporte en tal clase, y toda medida invariante son combinaciones convexas de estas, que conformarán puntos extremos de las combinaciones. Por lo tanto, si existen al menos dos clases recurrentes diferentes, existirá una cantidad no numerable de probabilidades invariantes.

Restrinjámonos, nuevamente, al caso irreducible. Ahora podemos enunciar el teorema ergódico para C.M.H., cual es una generalización de la L.G.N.

Teorema 18 Sea $(X_n)_{n \in \mathbb{N}}$ una C.M.H. irreducible y recurrente positiva. Sea $\pi = (\pi_x, x \in E)$ su única probabilidad invariante. Si $f : E \rightarrow \mathbb{R}$ es una función acotada, entonces

$$\frac{1}{n} \sum_{k=1}^n f(X_k) \xrightarrow{n \rightarrow \infty} \sum_{x \in E} f(x) \pi_x = \int_E f d\pi, \quad \mathbb{P} - c.s.$$

Demostración: Como f es acotada, existe c tal que $|f(x)| \leq c$, para cada $x \in E$. Sea

$$N_x(n) = \sum_{1 \leq k \leq n} \mathbb{1}_{X_k=x}$$

el número de retornos al estado x antes del tiempo n . Deseamos estudiar el límite, cuando $n \rightarrow \infty$ de

$$\frac{N_x(n)}{n}.$$

Sea $S_x^0, S_x^1, \dots, S_x^k, \dots$ los largos de las excursiones $\varepsilon_0, \varepsilon_1, \dots, \varepsilon_k, \dots$ partiendo del estado x . Tenemos

$$S_x^0 + \dots + S_x^{N_x(n)-1} \leq n < S_x^0 + \dots + S_x^{N_x(n)}.$$

Por lo tanto

$$\frac{S_x^0 + \dots + S_x^{N_x(n)-1}}{N_x(n)} \leq \frac{n}{N_x(n)} \leq \frac{S_x^0 + \dots + S_x^{N_x(n)}}{N_x(n)}.$$

Pero, dado que las variables aleatorias ε_k son *i.i.d.* (por lo tanto también lo son S_x^k), cuando $n \rightarrow \infty$,

$$\frac{S_x^0 + \dots + S_x^{N_x(n)}}{N_x(n)} \rightarrow \mathbb{E}_x(T_x) = m_x, \quad \mathbb{P}_x - c.s.$$

dado que $N_x(n) \rightarrow +\infty$, \mathbb{P}_x -casi seguramente. Entonces, nuevamente por la ley de los grandes números,

$$\frac{n}{N_x(n)} \rightarrow m_x, \quad \mathbb{P}_x - c.s.,$$

esto es,

$$\frac{N_x(n)}{n} \rightarrow \frac{1}{m_x}, \quad \mathbb{P}_x - c.s.$$

Esta convergencia también es cierta \mathbb{P}_μ -c.s., para toda ley inicial μ , dado que el límite de $N_x(n)/n$ es el mismo para las cadenas $\{N_n; n \geq 0\}$ y $\{X_{T_x+n}; n \geq 0\}$.

Ahora sea $F \subset E$. Definimos $\bar{f} = \sum_{x \in E} \pi_x f(x)$, $c = \sup_x |f(x)|$. Tenemos que

$$\begin{aligned} \left| \frac{1}{n} \sum_{k=1}^n f(X_k) - \bar{f} \right| &= \left| \sum_{x \in E} \left(\frac{N_x(n)}{n} - \pi_x \right) f(x) \right| \\ &\leq c \sum_{x \in F} \left| \frac{N_x(n)}{n} - \pi_x \right| + c \sum_{x \notin F} \left(\frac{N_x(n)}{n} + \pi_x \right) \\ &= c \sum_{x \in F} \left| \frac{N_x(n)}{n} - \pi_x \right| + c \sum_{x \in F} \left(\pi_x - \frac{N_x(n)}{n} \right) + 2c \sum_{x \notin F} \pi_x \\ &\leq 2c \sum_{x \in F} \left| \frac{N_x(n)}{n} - \pi_x \right| + 2c \sum_{x \notin F} \pi_x. \end{aligned}$$

Basta con elegir un conjunto finito F tal que $\sum_{x \notin F} \pi_x \leq \epsilon/4c$, y un $N(\omega)$ tal que, para todo $n \geq N(\omega)$,

$$\sum_{x \in F} \left| \frac{N_x(n)}{n} - \pi_x \right| \leq \frac{\epsilon}{4c},$$

así probamos el resultado. ■

6.5 El caso aperiódico

Ya hemos visto que en el caso irreducible y recurrente positivo;

$$\frac{1}{n} \sum_{k=1}^n \mathbb{1}_{\{X_k=y\}} \rightarrow \pi_y \quad c.s.,$$

cuando $n \rightarrow \infty$. Tomando esperanza bajo \mathbb{P}_x , deducimos que

$$\frac{1}{n} \sum_{k=1}^n (P^k)_{xy} \rightarrow \pi_y, \quad \forall x, y \in E. \quad (6.5.1)$$

La pregunta natural que surge es; si mantenemos las mismas hipótesis, ¿será que

$$(P^n)_{xy} \xrightarrow{n \rightarrow \infty} \pi_y \quad (6.5.2)$$

para todo $x, y \in E$?

En general, este resultado no es cierto.

Consideremos el paseo aleatorio sobre $E = \mathbb{Z}/N$, donde N es un entero par (identificamos a N con 0),

$$X_n = X_0 + Y_1 + \cdots + Y_n,$$

con Y_n variables *i.i.d.* tomando valores en $\{-1, 1\}$, con igual probabilidad ($\mathbb{P}(Y_n = 1) = \mathbb{P}(Y_n = -1) = 1/2$); en otras palabras,

$$X_n = (X_0 + Y_1 + \cdots + Y_n) \bmod N.$$

Esta cadena es irreducible, y recurrente positiva dado que E es inito. Pero $(P^{2k+1})_{xx} = 0$, para todo $x \in E$. En el caso particular de $N = 2$, tenemos que $P^{2k} = I$ y $P^{2k+1} = P$. Es decir, $(P^n)_{xx}$ no converge, aunque 6.5.1 si se cumple. Para obtener 6.5.2 necesitamos descartar este tipo de comportamientos.

Para que la convergencia deseada en 6.5.2 sea cierta, necesitamos una condición adicional.

Definición 12 Un estado $x \in E$ se dice *aperiódico* si existe N tal que

$$(P^n)_{xx} > 0, \quad \forall n \geq N.$$

Lema 9 Si P es irreducible y existe un estado aperiódico x , entonces para todos $y, z \in E$, existe M tal que $(P^n)_{yz} > 0$, para todo $n \geq M$. En particular, todos los estados son aperiódicos.

Demostración: De la irreducibilidad, existen $r, s \in \mathbb{N}$ tales que $(P^r)_{yx} > 0$, $(P^s)_{xz} > 0$. Más aún,

$$(P^{r+n+s})_{yz} \geq (P^r)_{yx}(P^n)_{xx}(P^s)_{xz} > 0$$

si $n \geq N$. Entonces tenemos la propiedad deseada con $M = N + r + s$. ■

Observación: Supongamos el caso irreducible y recurrente positivo. Sea π la probabilidad invariante, entonces $\pi_y > 0$, para todo $y \in E$. Por lo tanto, el hecho de que exista N tal que, para todo $n \geq N$, $(P^n)_{xy} > 0$ es una condición necesaria para establecer la convergencia $(P^n)_{xy} \rightarrow \pi_y$. Ahora veremos que, además, es condición suficiente.

Teorema 19 Supongamos que P es irreducible, recurrente positiva y aperiódica. Sea π la única medida de probabilidad invariante. Si $\{X_n; n \in \mathbb{N}\}$ es una (μ, P) -C.M., para todo $y \in E$,

$$\mathbb{P}(X_n = y) \rightarrow \pi_y, \quad n \rightarrow \infty;$$

en otras palabras,

$$(\mu P^n)_y \rightarrow \pi_y,$$

para cualquier ley inicial μ . En particular, para todos $x, y \in E$,

$$(P^n)_{xy} \rightarrow \pi_y.$$

Demostración: Debemos usar un par de argumentos. Sea $\{Y_n; n \in \mathbb{N}\}$ una (π, P) -C.M., independiente de $\{X_n; n \in \mathbb{N}\}$, y $x \in E$ arbitrario. Sea

$$T = \inf\{n \geq 0; X_n = Y_n = x\}.$$

- **Paso 1:** Mostrar que $\mathbb{P}(T < \infty) = 1$.

$\{W_n = (X_n, Y_n); n \in \mathbb{N}\}$ es una cadena da Markov a valores en $(E \times E)$, con ley inicial λ (donde $\lambda_{(x,u)} = \mu_x \pi_u$) y matriz de transición $\tilde{P}_{(x,u)(y,v)} = P_{xy} P_{uv}$. Dado que P es per-iódica, para todo x, u, y, v , para todo n suficientemente grande,

$$(\tilde{P}^n)_{(x,u)(y,v)} = (P^n)_{xy}(P^n)_{uv} > 0.$$

Porque \tilde{P} es irreducible. Más aún, \tilde{P} posee probabilidad invariante

$$\tilde{\pi}_{(x,u)} = \pi_x \pi_u.$$

Por lo tanto, \tilde{P} es recurrente positiva. T es el momento del primer encuentro de la cadena $\{W_n\}$ con el punto (x, x) ; y es finito casi seguramente.

- **Paso 2:** Definimos

$$Z_n = \begin{cases} X_n & , \quad n \leq T; \\ Y_n & , \quad n > T. \end{cases}$$

Por la propiedad fuerte de Markov, ambos procesos $\{X_{T+n}; n \geq 0\}$ y $\{Y_{T+n}; n \geq 0\}$ son (δ_x, P) -C.M., independientes de (X_0, \dots, X_T) . En consecuencia, $\{Z_n; n \in \mathbb{N}\}$ es, como $\{X_n\}$, una (μ, P) -C.M.

- **Paso 3:** Concluyendo. con estas tres identidades

$$\mathbb{P}(Z_n = y) = \mathbb{P}(X_n = y),$$

$$\mathbb{P}(Y_n = y) = \pi_y,$$

$$\mathbb{P}(Z_n = y) = \mathbb{P}(X_n = y, n \leq T) + \mathbb{P}(Y_n = y, n > T).$$

Por lo tanto,

$$|\mathbb{P}(X_n = y) - \pi_y| = |\mathbb{P}(Z_n = y) - \mathbb{P}(Y_n = y)| \leq \mathbb{P}(n < T) \rightarrow 0,$$

cuando $n \rightarrow \infty$.

■

Observación: Podemos definir el periodo del estado $x \in E$ como el máximo común divisor de los enteros n tales que $(P^n)_{xx} > 0$. Uno puede mostrar, con un argumento muy cercano al lema anterior, que cuando P es reducible, todos los estados tienen el mismo periodo. Un estado se dice *aperiódico* si su periodo es 1.

Ahora precisamos de la velocidad, o tasa de convergencia del teorema anterior, bajo una suposición adicional, llamada **condición de Doeblin**: existe $n_0 \in \mathbb{N}$, $\beta > 0$ y una probabilidad ν sobre E tal que

$$(D) \quad (P^{n_0})_{xy} \geq \beta \nu_y, \quad \forall x, y \in E.$$

Observación: La condición (D) es equivalente a la condición

$$\exists x \in E, n_0 \geq 1 \text{ tal que } \inf_{y \in E} (P^{n_0})_{yx} > 0.$$

Esto indica que el estado x es aperiódico. Pero esto no implica la irreducibilidad. Queda como ejercicio mostrar que esto implica la existencia de una única clase recurrente y de una única probabilidad invariante.

Lema 10 Si P es irreducible y aperiódica, y E es finito, entonces la condición (D) se satisface.

Demostración: Elegimos $x \in E$. Para cada $y \in E$, existe un n_y tal que $n \geq n_y \implies (P^n)_{yx} > 0$. Sea

$$\bar{n} = \sup_{y \in E} n_y, \quad \alpha = \inf_y (P^{\bar{n}})_{yx}.$$

Entonces $\alpha > 0$, y para todo $y \in E$,

$$(P^{\bar{n}})_{yx} \geq \alpha.$$

Por lo tanto la condición (D) se satisface con $n_0 = \bar{n}$, $\beta = \alpha$, $\nu = \delta_x$.

■

Por otro lado, la condición de Doeblin es raramente satisfecha en el caso $|E| = +\infty$, dado que, típicamente, para todo $n \in \mathbb{N}$, $y \in E$,

$$\inf_{x \in E} (P^n)_{xy} = 0.$$

Para Obtener una tasa de convergencia de $(P^n)_{xy}$ a π_y , necesitamos una forma de cuantificar esa convergencia.

Para ello utilizaremos la norma $l^1(\mathcal{P}(E))$ entre μ y ν , medidas de probabilidad en E .

Por ejemplo,

$$\|\mu - \nu\|_1 := \sum_{x \in E} |\mu_x - \nu_x|$$

Ahora introduciremos una herramienta que será de gran utilidad en las siguientes demostraciones.

Definición 13 Un **coupling** de dos probabilidades p y q sobre E es un par (X, Y) de variables aleatorias a valores en E , tales que p es la ley de X y q es la ley de Y .

Lema 11 Sean p y q , dos probabilidades en E .

$$\|p - q\|_1 = 2 \inf_{(X,Y) \text{ coupling de } p,q} \mathbb{P}(X \neq Y)$$

Demostración: Primero, notemos que si (X, Y) es coupling de p y q ,

$$\mathbb{P}(X = Y) = \sum_{x \in E} \mathbb{P}(X = Y = x) \leq \sum_{x \in E} p_x \wedge q_x$$

Entonces,

$$\mathbb{P}(X \neq Y) \geq 1 - \sum_{x \in E} p_x \wedge q_x = \sum_{x \in E} (p_x - q_x)^+$$

y

$$\|p - q\|_1 = \sum_{x \in E} |p_x - q_x| \leq 2\mathbb{P}(X \neq Y).$$

Por otro lado, definiendo $\alpha = \sum_{x \in E} p_x \wedge q_x$. Si ξ, U, V y W variables aleatorias mutuamente independientes tales que; $\xi \sim \text{Bernoulli}(\alpha)$, la ley de U es r , definido por $r_x = \alpha^{-1} p_x \wedge q_x$, la ley de V es \bar{p} definido por $\bar{p}_x = (1 - \alpha)^{-1} (p_x - q_x)^+$, y la ley de W , \bar{q} , definida como $\bar{q}_x = (1 - \alpha)^{-1} (q_x - p_x)^+$, entonces

$$\begin{aligned} X &= \xi U + (1 - \xi) V, \\ Y &= \xi U + (1 - \xi) W \end{aligned}$$

es un coupling (X, Y) de p y q .

En efecto, calculemos $\mathcal{L}(X)$:

$$\begin{aligned} \mathbb{P}(X = x) &= \mathbb{P}(X = x | \xi = 1) \mathbb{P}(\xi = 1) + \mathbb{P}(X = x | \xi = 0) \mathbb{P}(\xi = 0) \\ &= \mathbb{P}(U = x) \mathbb{P}(\xi = 1) + \mathbb{P}(V = x) \mathbb{P}(\xi = 0) \\ &= \frac{p_x \wedge q_x}{\alpha} \alpha + \frac{(p_x - q_x)^+}{1 - \alpha} (1 - \alpha) \\ &= p_x \wedge q_x + (p_x - q_x)^+ = p_x \end{aligned}$$

Por lo tanto $\mathcal{L}(X) = p$. De la misma forma se muestra que $\mathcal{L}(Y) = q$.

Veamos que $\mathbb{P}(X \neq Y) = (1 - \alpha)$:

$$\begin{aligned} \mathbb{P}(X \neq Y) &= \mathbb{P}(X \neq Y | \xi = 1) \mathbb{P}(\xi = 1) + \mathbb{P}(X \neq Y | \xi = 0) \mathbb{P}(\xi = 0) \\ &= \cancel{\mathbb{P}(U \neq U)} \mathbb{P}(\xi = 1) + \mathbb{P}(V \neq W) \mathbb{P}(\xi = 0) \\ &= [1 - \mathbb{P}(V = W)] (1 - \alpha) \\ &= \left[1 - \sum_{x \in E} \mathbb{P}(V = x, W = x) \right] (1 - \alpha) \\ &= \left[1 - \sum_{x \in E} \mathbb{P}(V = x) \mathbb{P}(W = x) \right] (1 - \alpha) \\ &= \left[1 - \sum_{x \in E} \frac{(p_x - q_x)^+ (q_x - p_x)^+}{(1 - \alpha)^2} \right] (1 - \alpha) \\ &= (1 - \alpha) \end{aligned}$$

De donde la última igualdad sale de que el producto $(p_x - q_x)^+ (q_x - p_x)^+ = 0, \forall x \in E$.

Luego,

$$\begin{aligned} \|p - q\|_1 &= \sum_{x \in E} |p_x - q_x| = \sum_{x \in E} [p_x + q_x - 2(p_x \wedge q_x)] \\ &= 2 - 2 \sum_{x \in E} p_x \wedge q_x = 2 - 2\alpha = 2(1 - \alpha) = 2\mathbb{P}(X \neq Y). \end{aligned}$$

■

Observación: Sea (E, d) un espacio métrico cualquiera (E no necesariamente finito o numerable), $\forall \mu, \nu$, probabilidades sobre E , podemos definir

$$W_d(\mu, \nu) := \inf_{\text{coupling de } \mu, \nu} \mathbb{E}[d(X, Y)]$$

Esto define una distancia sobre el espacio de medidas de probabilidad de E , llamada *distancia de Wasserstein*. Notar que el caso en que la distancia del espacio sea $d(x, y) = \mathbb{1}_{x \neq y}$, da lugar a $\mathbb{E}[d(x, y)] = \mathbb{P}(X \neq Y)$.

El siguiente resultado entrega una tasa de convergencia hacia la probabilidad invariante.

Teorema 20 *Supongamos que P es irreducible y satisface la condición de Doeblin (D) . Entonces P es aperiódica, recurrente positiva, y si π denota una probabilidad invariante,*

$$\sum_{y \in E} |(P^n)_{xy} - \pi_y| \leq 2(1 - \beta)^{[n/n_0]}, \quad \forall x \in E, n \in \mathbb{N},$$

donde $[n/n_0]$ es la parte entera de n/n_0 .

Demostración: Como la cadena es irreducible, la condición de Doeblin (D) claramente implica que es aperiódica.

Paso 1. Primero probar que para cualesquiera dos probabilidades μ y ν en E ,

$$\|\mu P^n - \nu P^n\|_1 \leq 2(1 - \beta)^{[n/n_0]}. \quad (6.5.3)$$

Para probar esto, gracias al lema anterior, es suficiente contruir un coupling (X_n, Y_n) de las probabilidades μP^n y νP^n tal que

$$\mathbb{P}(X_n \neq Y_n) \leq (1 - \beta)^{[n/n_0]}.$$

Supongamos que $n = kn_0 + m$, con $m < n_0$. Dados (X_{-0}, Y_0) con la ley $\mu \times \nu$ sobre $E \times E$, para $l = 0, 1, 2, \dots, k-1$, definimos $(X_{(l+1)n_0}, Y_{(l+1)n_0})$ en términos de (X_{ln_0}, Y_{ln_0}) de la siguiente manera. Sea $\{\xi_l, U_l, V_l; l \geq 0\}$ una sucesión de variables aleatorias mutuamente independientes, donde ξ_l es una variable *Bernoulli* con $\mathbb{P}(\xi_l = 1) = \beta = 1 - \mathbb{P}(\xi_l = 0)$, la ley de U_l es $\tilde{m} = \beta^{-1}m$ y V_l es uniforme en $[0, 1]$. Definimos

$$Q_{xy} = (-1\beta)^{-1}((P^{n_0})_{xy} - m_y)$$

y $f : E \times [0, 1] \rightarrow E$ tal que, para todo $x, y \in E$, $\{u; f(x, u) = y\}$ es un boreliano subconjunto de $[0, 1]$, y dado que V es uniforme en $[0, 1]$, la ley de $f(x, V)$ es Q_x , $x \in E$. Ahora sean

$$X_{(l+1)n_0} = \xi_l U_l + (1 - \xi_l) f(X_{ln_0}, V_l),$$

$$Y_{(l+1)n_0} = \xi_l U_l + (1 - \xi_l) f(Y_{ln_0}, V_l).$$

Notamos que, realmente, hemos contruido un coupling (X_{ln_0}, Y_{ln_0}) de μP^{ln_0} y νP^{ln_0} , para $l = 0, \dots, k$, tales que

$$\mathbb{P}(X_{ln_0} \neq Y_{ln_0}) \leq \mathbb{P}(\cap_{m=0}^l \xi_m = 0) = (1 - \beta)^l.$$

Resta por construir un coupling (X_n, Y_n) de μP^n y νP^n , tal que $\{X_n \neq Y_n\} \subset \{X_{kn_0} \neq Y_{kn_0}\}$, lo que es fácil.

Paso 2. Ahora probaremos que cualquier probabilidad μ sobre E , $\{\mu P^n; n \geq 0\}$ es una sucesión de Cauchy en el espacio de Banach, $l^1(E)$. Si $\nu = \mu P^m$, por 6.5.3

$$\|\mu P^{n+m} - \mu P^n\|_1 = \|\nu P^n - \mu P^n\|_1 \leq 2c^{n-n_0},$$

donde $c = (1 - \beta)^{1/n_0}$. Se concluye.

Paso 3. Del paso anterior tenemos que la secuencia de probabilidades $\{\mu P^n; n \geq 0\}$ converge en $l^1(E)$, hacia π , probabilidad en E . Pero

$$\pi P = \lim_{n \rightarrow \infty} \mu P^{n+1} = \pi,$$

por lo tanto π es invariante, y la cadena es recurrente positiva. En consecuencia, de 6.5.3, para cada probabilidad μ de E ,

$$\|\mu P^n - \pi\|_1 \leq 2(1 - \beta)^{[n/n_0]},$$

lo que establece la razón de convergencia buscado, y la aperiodicidad.

■

Ahora enunciaremos un teorema central del límite para cadenas de Markov irreducibles, recurrentes positivas. Si tal cadena, además satisface

$$\sum_{y \in E} |(P^n)_{xy} - \pi_y| \leq M t^n, \quad x \in E, n \in \mathbb{N}$$

con $M \in \mathbb{R}$ y $0 < t < 1$, es llamada *uniformemente ergódica*. Sólomente debemos mostrar que la condición de Doeblin implica uniforme ergodicidad. Esta propiedad implica el teorema central del límite.

Teorema 21 Sea $\{X_n; n \in \mathbb{N}\}$ es una cadena de Markov a valores de E , con matriz de de transición P irreducible, que además es uniformemente ergódica y aperiódica. Sea π la única distribución invariante de la cadena, y $f : E \rightarrow \mathbb{R}$, tal que

$$\sum_{x \in E} \pi_x f^2(x) < \infty \quad y \quad \sum_{x \in E} \pi_x f(x) = 0.$$

Entonces, cuando $n \rightarrow \infty$,

$$\frac{1}{\sqrt{n}} \sum_{k=1}^n f(X_k) \xrightarrow{\mathcal{L}} \sigma_f Z,$$

donde $Z \sim \mathcal{N}(0, 1)$ y

$$\begin{aligned} \sigma_f^2 &= \sum_{x \in E} \pi_x (Qf)_x^2 - \sum_x \pi_x (PQf)_x^2 \\ &= 2 \sum_x \pi_x (Qf)_x f_x - \sum_x \pi_x f_x^2, \end{aligned}$$

con

$$(Qf)_x = \sum_{n=0}^{\infty} \mathbb{E}_x [f(X_n)], \quad x \in E.$$

Note que la propiedad uniforme ergodicidad implica que la serie que define al operador Q converja.

6.6 Cadenas de Markov Reversibles

Consideremos el caso irreducible y recurrente positivo. La propiedad de Markov (aquella que dice que el pasado y el futuro de la cadena son independientes condicionado al valor presente) nos dice que, cuando $\{X_n; n \in \mathbb{N}\}$ es una cadena de Markov, para cada N , $\{\hat{X}_n^N = X_{N-n}; 0 \leq n \leq N\}$ también es cadena de Markov, a menos que $\{X_n\}$ inicie con probabilidad invariante π .

Proposición 9 Sea $\{X_n; n \in \mathbb{N}\}$ una cadena de Markov con matriz de transición P , que suponemos irreducible, y π su distribución invariante. Entonces la cadena a tiempo-reverso $\{\hat{X}_n^N; 0 \leq n \leq N\}$ es una (π, \hat{P}) -C.M., con

$$\pi_y \hat{P}_{yx} = \pi_x P_{xy}, \quad x, y \in E.$$

Demostración:

$$\begin{aligned} & \mathbb{P}(\hat{X}_{p+1} = x \mid \hat{X}_p = y) \\ &= \mathbb{P}(X_n = x \mid X_{n+1} = y) \\ &= \mathbb{P}(X_{n+1} = y \mid X_n = x) \times \frac{\mathbb{P}(X_n = x)}{\mathbb{P}(X_{n+1} = y)} \end{aligned}$$

■

Decimos que la cadena $\{X_n; n \in \mathbb{N}\}$ es *reversible* si $P = \hat{P}$, lo cual se satisface si y sólo si se satisfacen las *ecuaciones de balance detallado*:

$$\pi_x P_{xy} = \pi_y P_{yx}, \quad x, y \in E,$$

donde π es la probabilidad invariante. Es fácil verificar que si π satisface esa relación, entonces es P -invariante. La proposición contraria no necesariamente es cierta.

Observación: Si π es la distribución invariante de una cadena de Markov irreducible (y, por lo tanto, recurrente positiva), la cadena no necesariamente es reversible. Suponga que $\text{card}(E) \geq 3$. Entonces podría existir $x \neq y$ tal que $P_{xy} = 0 \neq P_{yx}$. En consecuencia, $\pi_x P_{xy} = 0 \neq \pi_y P_{yx}$. La transición de y a x de la cadena original corresponde a la transición de x a y de la cadena a tiempo inverso, por lo tanto $P_{yx} \neq 0 \implies \hat{P}_{xy} \neq 0$, de donde se concluye que $\hat{P} \neq P$.

Observación: Dada una matriz de transición P de una cadena de Markov irreducible, recurrente positiva, entonces una de las cosas que se busca es el cálculo de la probabilidad invariante. Este problema no es simple resoluble.

Otro problema, es determinar una matriz de transición P , cuya cadena de Markov asociada admita una probabilidad π , invariante.

El segundo problema es bastante más fácil de resolver. En efecto, siempre existen varias soluciones. La manera más fácil de resolverlo es buscar P tal que la cadena asociada es reversible con respecto a π . En otras palabras, es suficiente encontrar una matriz de transición P , irreducible, tal que la cantidad $\pi_x P_{xy}$ sea simétrica en x, y .

A fin de resolver el primer problema, uno puede tratar de encontrar π tal que

$$\pi_x P_{xy} = \pi_y P_{yx}, \quad \forall x, y \in E,$$

lo que es diferente a resolver $\pi P = \pi$, no implica ninguna suma con respecto a x . Pero esta ecuación tiene solución sólo si la cadena es reversible con respecto a su única medida de probabilidad invariante, lo que podría no ser el caso.

Supongamos ahora que, dado el par (P, π) , deseamos verificar si π es, o no, una probabilidad invariante con respecto a la matriz de transición P . Si la cantidad $\pi_x P_{xy}$ es simétrica en x, y , entonces la respuesta es que sí lo es, y ganamos una propiedad adicional, la reversibilidad. Si no es el caso, uno necesita verificar (o no) la igualdad $\pi P = \pi$. La siguiente proposición puede ser de utilidad para el problema de verificación.

Proposición 10 Sea P una matriz de transición irreducible, y π una probabilidad en E , estrictamente positiva. Para cada par $x, y \in E$, definimos

$$\hat{P}_{xy} = \begin{cases} \frac{\pi_y}{\pi_x} P_{yx}, & \text{si } x \neq y, \\ P_{xx}, & \text{si } x = y. \end{cases}$$

π es probabilidad invariante de la cadena definida por la matriz de transición P , y \hat{P} es la matriz de transición de la cadena a tiempo reverso si y sólo si, para cada $x \in E$,

$$\sum_{y \in E} \hat{P}_{xy} = 1.$$

6.7 Razón de convergencia al equilibrio

Suponga el caso irreducible, recurrente positivo y aperiódico. Entonces, sabemos que para cada $x, y \in E$, $(P^n)_{x,y} \rightarrow p_{xy}$ cuando $n \rightarrow \infty$, donde π es la única medida de probabilidad invariante. Más generalmente, esperamos que para una gran clase de funciones $f : E \rightarrow \mathbb{R}$, $(P^n f)_x \rightarrow \langle f, \pi \rangle$ cuando $n \rightarrow \infty$ para todo $x \in E$, donde, de ahora en adelante,

$$\langle f, \pi \rangle = \sum_{x \in E} f(x) \pi_x.$$

En esta sección discutiremos la razón a la cual esta convergencia se mantiene.

6.7.1 El caso reversible de finitos estados

En primer lugar, consideremos el caso simple. Asumamos que E es finito (y escribimos $d = |E|$) y el proceso es reversible. Notemos que podemos identificar $L^2(\pi)$ con \mathbb{R}^d , dotado del producto escalar

$$\langle f, g \rangle_\pi = \sum_{x \in E} f(x) g(x) \pi_x.$$

Ahora, la reversibilidad de P es equivalente a decir que P es, como elemento de $\mathcal{L}(L^2(\pi))$, un operador auto-adjunto, en el sentido que

$$\begin{aligned} \langle Pf, g \rangle_\pi &= \sum_{x, y \in E} P_{x,y} f(y) g(x) \pi_x \\ &= \sum_{x, y \in E} P_{y,x} f(y) g(x) \pi_y \\ &= \langle f, Pg \rangle_\pi, \end{aligned}$$

donde usamos la ecuación de balance detallado para la segunda identidad. Ahora veremos que la norma del operador P , como elemento de $\mathcal{L}(L^2(\pi))$, es al menos 1. En efecto, si $\|\cdot\|_\pi$ denota la norma usual en $L^2(\pi)$,

$$\begin{aligned} \|Pf\|_\pi^2 &= \sum_{x \in E} [(Pf)_x]^2 \pi_x \\ &= \sum_{x \in E} (\mathbb{E}[f(X_t) | X_0 = x])^2 \pi_x \\ &\leq \mathbb{E}[f^2(X_t) | X_0 = x] \pi_x \\ &= \sum_{x \in E} f^2(x) \pi_x, \end{aligned}$$

donde se usó la desigualdad de Schwarz (o, equivalentemente, la desigualdad de Jensen) para la parte en que aparece la desigualdad, y la invarianza de π para la última identidad.

A fin de poder trabajar en \mathbb{R}^d , dotado de la norma euclídeana, introduciremos una nueva matriz $d \times d$

$$\tilde{P}_{x,y} := \sqrt{\frac{\pi_x}{\pi_y}} P_{x,y}.$$

En notación matricial, $\tilde{P} = \Pi^{1/2} P \Pi^{-1/2}$, donde $\Pi_{x,y} = \delta_{x,y} \pi_x$ es una matriz diagonal. Más aún, si $\|\cdot\|$ es la norma euclídeana en \mathbb{R}^d , para cada $f : E \rightarrow \mathbb{R}$ (es decir, f es una colección de números reales indexada por los d elementos de E , en otras palabras un elemento de \mathbb{R}^d), anotando como $g = \Pi^{-1/2} f$, tenemos

$$\|\tilde{P}f\|^2 = \sum_{x \in E} (P \Pi^{-1/2} f)_x^2 = \|Pg\|_\pi^2 \leq \|g\|_\pi^2 = \|f\|^2.$$

Primero, notemos que f es un vector propio de \tilde{P} si y sólo si $g = \Pi^{-1/2} f$ es un vector propio derecho de P , y $g' = \Pi^{1/2} f$ es un vector propio izquierdo de P asociados al mismo valor propio. Tenemos que \tilde{P} es una $d \times d$ matriz simétrica, cuya norma está acotada por 1. Por lo tanto, por resultado elementales de álgebra lineal, \tilde{P} admite los valores propios $-1 \leq \lambda_d \leq \lambda_{d-1} \leq \lambda_2 \leq \lambda_1 \leq 1$. Establecemos el siguiente lema.

Lema 12 *Tenemos que $\lambda_2 < \lambda_1 = 1$ y $-1 < \lambda_d$.*

Demostración: Ver [3, Pag. 41]

Sean g_1, \dots, g_d la base ortonormal de $L^2(\pi)$ hecha por vectores propios derechos de P , correspondientes respectivamente a los valores propios $1, \lambda_1, \dots, \lambda_d$. Para cada $f \in L^2(\pi)$, dado $g_1 = (1, \dots, 1)$,

$$\begin{aligned} f - \langle f, \pi \rangle &= \sum_{l=2}^d \langle f, g_l \rangle_{\pi} g_l, \\ Pf - \langle f, \pi \rangle &= \sum_{l=2}^d \lambda_l \langle f, g_l \rangle_{\pi} g_l, \\ P^n f - \langle f, \pi \rangle &= \sum_{l=2}^d \lambda_l^n \langle f, g_l \rangle_{\pi} g_l, \\ \|P^n f - \langle f, \pi \rangle\|_{\pi}^2 &= \sum_{l=2}^d \lambda_l^{2n} \langle f, g_l \rangle_{\pi}^2, \\ &\leq \sup_{2 \leq l \leq d} \lambda_l^{2n} \|f - \langle f, \pi \rangle\|_{\pi}^2, \end{aligned}$$

por lo tanto tenemos la siguiente proposición.

Proposición 11

$$\|P^n f - \langle f, \pi \rangle\|_{\pi} \leq (1 - \beta)^n \|f - \langle f, \pi \rangle\|_{\pi},$$

donde $\beta := (1 - \lambda_2) \wedge (1 + \lambda_d)$ es la brecha espectral.

6.7.2 El caso general

Más generalmente, lo mismo es cierto para

$$\beta := 1 - \sup_{f \in L^2(\pi), \|f\|_{\pi}=1} \|Pf - f, \pi\|_{\pi}.$$

En efecto, con este β , considerando sólo el caso $f \neq 0$, dado que todas las desigualdades a continuación son ciertas para $f = 0$, tenemos

$$\begin{aligned} \|Pf - \langle f, \pi \rangle\|_{\pi} &= \left\| P \left(\frac{f}{\|f\|_{\pi}} \right) - \left\langle \frac{f}{\|f\|_{\pi}}, \pi \right\rangle \right\|_{\pi} \times \|f\|_{\pi} \\ &\leq (1 - \beta) \|f\|_{\pi}. \end{aligned}$$

Finalmente, la proposición anterior se mantiene en el caso general para esta nueva definición de β . Note que

$$\begin{aligned} \|P^{n+1} f - \langle f, \pi \rangle\|_{\pi} &= \|P[P^n f - \langle f, \pi \rangle]\|_{\pi} \\ &\leq (1 - \beta) \|P^n f - \langle f, \pi \rangle\|_{\pi}. \end{aligned}$$

El resultado se concluye mediante inducción.

6.8 Estadísticas en cadenas de Markov

El objetivo de esta sección es introducir las nociones básicas de la estimación de los parámetros de una cadena de Markov.

Hemos visto que, para cada $n \geq 0$, la ley del vector aleatorio (X_0, X_1, \dots, X_n) depende sólo de la ley inicial μ y de la matriz de transición P . Estamos interesados en las condiciones bajo las cuales uno puede estimar el par (μ, P) , dado un vector de observaciones de (X_0, X_1, \dots, X_n) , y de forma que el error tienda a cero cuando $n \rightarrow \infty$.

Primero, analicemos el estimador de la probabilidad invariante μ . Para cada $x \in E$,

$$\hat{\mu}_x^n = \frac{1}{n+1} \sum_{l=0}^n \mathbb{1}_{\{X_l=x\}}$$

es un estimador consistente de μ_x , esto es consecuencia directa del teorema ergódico:

Proposición 12 *Para cada $x \in E$, $\hat{\mu}_x^n \rightarrow \mu_x$, casi seguramente cuando $n \rightarrow \infty$.*

Ahora observemos el estimador de P_{xy} , $x, y \in E$. Elegimos el estimador

$$\hat{P}_{xy}^n = \frac{\sum_{l=0}^{n-1} \mathbb{1}_{\{X_l=x, X_{l+1}=y\}}}{\sum_{l=0}^{n-1} \mathbb{1}_{\{X_l=x\}}}.$$

Tenemos la siguiente proposición.

Proposición 13 *Para cada $x, y \in E$, $\hat{P}_{xy}^n \rightarrow P_{xy}$, casi seguramente cuando $n \rightarrow \infty$.*

Demostración: Ver [3, Pag. 43]

7 Algoritmos estocásticos usando cadenas de Markov

El objetivo de este capítulo es presentar algunos algoritmos basados en la simulación de cadenas de Markov. Supondremos en todo momento que todas las cadenas de Markov toman valores en un conjunto E , finito.

7.1 Markov chain Monte Carlo (M.C.M.C.)

Se desea simular una variable aleatoria con ley $\pi \in \mathcal{P}(E)$, en muchas situaciones esto es muy costoso computacionalmente. Por ejemplo, si E es finito, pero muy grande, y π se conoce salvo constante multiplicativa, siquiera podría ser posible calcular la constante $\sum_{x \in E} \pi_x$ que normaliza π .

En tal caso se propone: simular una cadena de Markov con alguna matriz de transición P , que posea a π como probabilidad invariante.

Si se desea aproximar $\sum_{x \in E} f(x)\pi_x$, y $(X_n)_n$ es una P -C.M., gracias al teorema ergódico, tenemos

$$\frac{1}{N} \sum_{n=1}^N f(X_n) \approx \sum_{x \in E} f(x)\pi_x, \quad (M.C.M.C.)$$

Para encontrar P que tenga a π como probabilidad invariante, es suficiente imponer las ecuaciones de balanceo detallado:

$$\pi_x P_{xy} = \pi_y P_{yx}, \quad \forall x, y \in E \quad (7.1.1)$$

Con esto en mente, se propone el siguiente método: Dada R , una matriz de transición markoviana arbitraria sobre E . Entonces las fórmulas

$$\begin{cases} P_{xy} = R_{xy} \wedge \left(\frac{\pi_y}{\pi_x} R_{yx} \right), & x \neq y, \\ P_{xx} = 1 - \sum_{y \neq x} P_{xy} \end{cases} \quad (7.1.2)$$

(donde $a \wedge b = \inf(a, b)$), definen una matriz Markoviana P , tal que cumpla 7.1.1, para cada $x, y \in E$. La irreducibilidad de R no es suficiente para asegurar la de P , para eso necesitamos, para cada $x \neq y$ que exista $n \geq 1$ y $\{x_0, \dots, x_n\} \subset E$ con $x_0 = x$ y $x_n = y$ tal que

$$R_{x_{k-1}x_k} \wedge R_{x_k x_{k-1}} > 0, \quad \forall 1 \leq k \leq n.$$

¿Cómo elegir R en la práctica? Se escoge un grafo no dirigido G sobre E , tal que para todos $x, y \in E$, existen $n \in \mathbb{N}$, x_1, \dots, x_{n+1} , tal que $x_1 = x$, $x_{n+1} = y$ y para todo $1 \leq k \leq n$, $(x_k, x_{k+1}) \in G$ (camino de x a y), y elegimos R de la siguiente forma:

$$R_{xy} > 0 \iff (x, y) \in G.$$

Entonces la matriz P definida como 7.1.2 es irreducible.

Existen dos elecciones clásicas de R para un grafo G dado:

- **Gibbs sampler:**

$$R_{xy} = \begin{cases} \left(\sum_{\{z; (x,z) \in G\}} \pi_z \right)^{-1} \pi_y, & \text{si } (x, y) \in G, \\ 0, & \text{si } (x, y) \notin G. \end{cases}$$

- **Algoritmo Metrópolis:**

$$R_{xy} = \begin{cases} (n_x)^{-1}, & \text{si } (x, y) \in G \\ 0, & \text{si } (x, y) \notin G, \end{cases}$$

donde $n_x = |\{z; (x, z) \in G\}|$.

Observación: Para que el método sea práctico, debe ser computacionalmente eficiente sumlar las transiciones de R .

Todo lo anterior da lugar al siguiente algoritmo para simular la P -CMH.:

1. Elegir grafo G y la matriz R cumpliendo lo recién descrito.
2. Escoger punto de partida x_0 a gusto.
3. Iterar. Dado $X_n = x$, escoger $y = Y_n$ de acuerdo a la ley de probabilidad R_x .
4. Elegir X_{n+1} de la siguiente forma

$$X_{n+1} = \begin{cases} Y_n, & \text{con probabilidad } \frac{\pi_y R_{yx}}{\pi_x R_{xy}} \wedge 1, \\ X_n, & \text{con probabilidad } 1 - \left(\frac{\pi_y R_{yx}}{\pi_x R_{xy}} \wedge 1 \right). \end{cases}$$

Una forma de hacer lo anterior es considerar una variable aleatoria U_n uniforme en $[0, 1]$ e independiente de las demás variables, y definir

$$X_{n+1} = \mathbb{1}_{\{U_n \leq \pi_y R_{yx} / \pi_x R_{xy}\}} Y_n + \mathbb{1}_{\{U_n > \pi_y R_{yx} / \pi_x R_{xy}\}} X_n.$$

Observación: La dependencia de π es mediante el cociente π_y / π_x , luego basta conocer π salvo constante multiplicativa.

Ejemplo 5 (Modelo Ising) Es uno de los más populares modelos de la física estadística. Dado $N \in \mathbb{N}$ (N lo supondremos grande), sea

$$\Lambda = \Lambda_N = \{-N, \dots, N\}^2 \subset \mathbb{Z}^2,$$

cuya frontera es $\partial\Lambda = \Lambda_N \setminus \Lambda_{N-1}$, y definimos el espacio de configuraciones como

$$E = \{-1, 1\}^\Lambda.$$

De modo que $\forall m \in \Lambda$ y $\forall x \in E$, $x(m) \in \{-1, 1\}$ denota el spin del sitio $m \in \Lambda$.

Para $x \in E$, definimos

$$H(x) = \frac{1}{2} \sum_{\substack{m, m' \in \Lambda \\ |m - m'| = 1}} |x(m) - x(m')|^2.$$

Sea

$$E^+ = \{x \in E; x(m) = 1, \forall m \in \partial\Lambda\}.$$

Supongamos que queremos simular la siguiente distribución:

$$\pi_x = \frac{e^{-\beta H(x)}}{Z(\beta)} \quad \forall x \in E,$$

donde $\beta > 0$ es un parámetro ($1/\beta$ es la temperatura), y

$$Z(\beta) = \sum_{x \in E} e^{-\beta H(x)}.$$

Note que, incluso para un valor de N no tan grande, el cálculo de $Z(\beta)$ es imposible. Por ejemplo, para $N = 10$:

$$\begin{aligned} |E| &= |\{-1, 1\}^\Lambda| = 2^{|\Lambda|} \\ &= 2^{(2(N-1)+1)^2} \approx 2^{4N^2} = 2^{400} \\ &= (2^{10})^{40} \approx (10^3)^{40} = 10^{120} \end{aligned}$$

Usaremos MCMC. sobre el siguiente grafo G : $xy \in G$ ssi x e y difieren en exáctamente un sitio, esto es, existe $m_0 \in \Lambda \setminus \Lambda$ tal que $x(m_0) \neq y(m_0)$ y $x(m) = y(m) \forall m \neq m_0$. Usando el algoritmo metrópolis:

$$R_{xy} = \begin{cases} |\Lambda \setminus \partial \Lambda|^{-1} = \frac{1}{(2N-1)^2}, & xy \in G \\ 0, & xy \notin G \end{cases}$$

En el paso 4) del algoritmo, la expresión de $\frac{\pi_y R_{yx}}{\pi_x R_{xy}}$ se simplifica bastante:

$$\forall xy \in G, \frac{\pi_y R_{yx}}{\pi_x R_{xy}} = \frac{e^{-\beta H(y)} / Z(\beta)}{e^{-\beta H(x)} / Z(\beta)} \frac{1/(2N-1)^2}{1/(2N-1)^2} = e^{-\beta(H(y)-H(x))}$$

7.2 Simulación de la distribución invariante

Uno de los problemas en algoritmos MCMC. es la elección del número de veces en que debemos iterar la cadena de Markov. La diferencia con el método estándar de Monte Carlo es que la cadena parte de un punto arbitrario, esto es, la cadena no parte según su distribución de probabilidad invariante. En este sentido, uno podría pensar que existe una "fase inicial" del algoritmo, durante la cual la ley de la cadena toma valores cercanos a la distribución invariante. Entonces, durante la segunda fase del algoritmo, podríamos controlar la tasa de convergencia en el teorema ergódico, lo cual se puede hacer con ayuda del teorema del límite central.

Asumiremos que $|E| < \infty$ y, para aligerar la notación $E = \{1, \dots, N\}$.

7.2.1 Simulación perfecta.

La idea de MCMC. es simular un P -CMH. $(X_n)_{n \in \mathbb{N}}$, donde P tiene a π como probabilidad invariante, de modo que

$$\frac{1}{n} \sum_{k=1}^N f(X_k) \approx \sum_{x \in E} f(x) \pi_x$$

El problema subyacente es que, a veces, se necesita tomar n muy grande para que lo anterior funcione, especialmente si X_0 tiene ley muy alejada de π .

Veremos, a continuación, una alternativa a MCMC, llamada *simulación perfecta*. Obtendremos una simulación exacta de π (e lugar de una aproximación como en MCMC) calzando una cantidad aleatoria de pasos.

Supondremos que P cumple

$$\beta = \beta(P) = \sum_{y \in E} \inf_{x \in E} P_{xy} > 0 \quad (7.2.1)$$

Gracias a las suposiciones que hicimos con respecto a E (finito, $E = [N]$), 7.2.1 es la condición de Doeblin (D) con $n_0 = 1$. Notar que $\beta(P) \leq 1$.

Sea $\nu \in \mathcal{P}(E)$ dada por:

$$\nu_y := \frac{\inf_{x \in E} P_{xy}}{\beta}, \quad y \in E,$$

Observación: Uno podría elegir otro par (β, ν) , con $\beta > 0$, y ν probabilidad sobre E tal que $P_{xy} \geq \beta \nu_y$, pero la elección anterior es óptimo, en el sentido que maximiza β .

Observación: La afirmación $\beta(P) > 0$ implica la existencia de una única clase recurrente, por lo tanto P posee una única probabilidad invariante, a la que llamaremos π .

Sean los siguientes objetos independientes cada uno del otro, $\forall n \in \mathbb{N}$:

- $\xi_n \sim \text{Bernoulli}(\beta)$
- $Z_n \sim \nu$

- $U_n \sim \text{unif}(0, 1)$

Sea Q la matriz Markoviana definida por

$$Q_{xy} = \frac{1}{1-\beta}(P_{xy} - \beta\nu_y),$$

y $f : E \times [0, 1] \rightarrow E$, función de transición de Q , es decir: $\forall x \in E, \forall U \sim \text{unif}(0, 1)$, se tiene que $f(x, U) \sim Q_x$.

Dado X_0 , independiente de todo lo anterior, definamos $(X_n)_{n \in \mathbb{N}}$ mediante la recursión

$$X_n = \xi_n Z_n + (1 - \xi_n) f(X_{n-1}, U_n) \quad (7.2.2)$$

Proposición 14 *La recurrencia definida en 7.2.2 es una P-C.M.*

Demostración: Condicionando a los valores de ξ_n y ocupando probabilidades totales:

$$\begin{aligned} \mathbb{P}(X_1 = y \mid X_0 = x) &= \mathbb{P}_x(X_1 = y) \\ &= \mathbb{P}_x(X_1 = y \mid \xi = 1) \mathbb{P}(\xi = 1) + \mathbb{P}_x(X_1 = y \mid \xi = 0) \mathbb{P}(\xi = 0) \\ &= \mathbb{P}(Z = y) \mathbb{P}(\xi = 1) + \mathbb{P}_x(f(X_0, U) = y) \mathbb{P}(\xi = 0) \\ &= \nu_y \beta + \mathbb{P}(f(x, U) = y) (1 - \beta) \\ &= \nu_y \beta + Q_{xy} (1 - \beta) \\ &= \nu_y \beta + \frac{(P_{xy} - \beta\nu_y)}{(1 - \beta)} (1 - \beta) \\ &= P_{xy} \end{aligned}$$

Así probamos que la matriz de transición es P , falta por demostrar que $(X_n)_n$ es una cadena de Markov, es decir $\mathbb{P}(X_n = y \mid X_{n-1} = x) = \mathbb{P}(X_1 = y \mid X_0 = x) = P_{xy} \forall n \in \mathbb{N}$, resultado que es directo del Lema 8.

■

Tenemos lo siguiente:

Proposición 15 *Sea*

$$T = \inf\{n \geq 1 \mid \xi_n = 1\}$$

Entonces $T \sim \text{geom}(\beta)$, $X_T \sim \nu$, además X_T es independiente con T .

Demostración:

$$\begin{aligned} \mathbb{P}(X_T = x \mid T = n) &= \mathbb{P}(\xi_1 = 0, \dots, \xi_{n-1} = 0, \xi_n = 1, Z_n = x) \\ &= (1 - \beta)^{n-1} \beta \nu_x \end{aligned}$$

■

La idea, ahora, es construir una P-C.M. estacionaria. Para construir $(X_n)_{n \in \mathbb{Z}}$, consideremos $(\xi_n)_{n \in \mathbb{Z}}$, $(Z_n)_{n \in \mathbb{Z}}$, $(U_n)_{n \in \mathbb{Z}}$, al igual que antes, y sea

$$\tau(n) = \max\{k \leq n : \xi_k = 1\}$$

Para n tal que $\xi_n = 1$, se tiene que $n = \tau(n)$ y definimos $X_n := Z_n$.

Mientras que para n tal que $\xi_n = 0$, se tiene $\tau(n) < n$, y definimos iterativamente:

$$\begin{aligned} X_{\tau(n)+1} &:= f(X_{\tau(n)}, U_{\tau(n)+1}) \\ X_{\tau(n)+2} &:= f(X_{\tau(n)+1}, U_{\tau(n)+2}) \end{aligned}$$

$$X_n := f(X_{n-1}, U_n)$$

Básicamente, es la misma construcción anterior, sólo que partiendo con ley ν en el instante aleatorio $\tau(0) \leq 0$.

Notar que $-\tau(0) + 1 \sim \text{geom}(\beta)$.

Proposición 16 $(X_n)_{n \in \mathbb{Z}}$ es estacionario, es decir, $\forall l, k \in \mathbb{Z}$, $(X_{l+1}, \dots, X_{l+k}) \stackrel{d}{=} (X_1, \dots, X_k)$. En particular, $X_0 \sim \pi$.

Demostración: Sea $\mu = \mathcal{L}(X_0)$. Basta probar que $\mu P = \mu$.

$$\begin{aligned} \mu_x = \mathbb{P}(X_0 = x) &= \sum_{k=0}^{\infty} \mathbb{P}(X_0 = x \mid \tau(0) = -k) \mathbb{P}(\tau(0) = -k) \\ &= \sum_{k=0}^{\infty} (\nu Q^k)_x (1 - \beta)^k \beta \end{aligned}$$

Como $P_{xy} = (1 - \beta)Q_{xy} + \beta\nu_y$, tenemos

$$\begin{aligned} (\mu P)_y &= \sum_{x \in E} \mu_x P_{xy} = \sum_{x \in E} \sum_{k=0}^{\infty} (\nu Q^k)_x (1 - \beta)^k \beta P_{xy} \\ &= \beta \sum_{k=0}^{\infty} (1 - \beta)^{k+1} \sum_{x \in E} (\nu Q^k)_x Q_{xy} + \beta^2 \nu_y \sum_{k=0}^{\infty} (1 - \beta)^k \sum_{x \in E} (\nu Q^k)_x \\ &= \beta \sum_{k=0}^{\infty} (1 - \beta)^{k+1} (\nu Q^{k+1})_y + \beta^2 \nu_y \sum_{k=0}^{\infty} (1 - \beta)^k \\ &= \beta \sum_{k=1}^{\infty} (1 - \beta)^k (\nu Q^k)_y + \beta \\ &= \beta \sum_{k=1}^{\infty} (1 - \beta)^k (\nu Q^k)_y + \beta (\nu Q^0)_y (1 - \beta)^0 \\ &= \sum_{k=0}^{\infty} \beta (1 - \beta)^k (\nu Q^k)_y = \mu_y \end{aligned}$$

■

7.3 Simulated Annealing

La búsqueda de máximos globales de alguna función es uno de los problemas más importantes de la matemática aplicada. En el caso de funciones diferenciables definidas sobre \mathbb{R}^d , uno puede empezar por un punto arbitrario y moverse en la dirección determinada por el gradiente hasta que la función detenga su crecimiento. Desafortunadamente, tales métodos conducen a mínimos locales, los cuales podrían no ser mínimos globales. En el caso en que la función esté definida sobre un conjunto finito E , uno podría calcular cada valor de la función $f(x)$ en cada punto $x \in E$, pero la eficiencia de este procedimiento depende bastante del tamaño que tenga el conjunto E , en casos que le el número de elementos sea demasiado grande esto se hace imposible de implementar.

Usaremos aleatoriedad para escapar de óptimos locales, buscando en zonas distintas del espacio E , ojalá convergiendo al óptimo global. A medida que el algoritmo progresa, se irán reduciendo las perturbaciones aleatorias.

Específicamente: sea E finito, sea $U : E \rightarrow \mathbb{R}_-$, y supongamos

$$\max_{x \in E} U_x = 0.$$

Buscamos $x \in E$ tal que $U_x = 0$. Para $\beta > 0$ ($1/\beta$ es la *temperatura*), sea $\pi^\beta \in \mathcal{P}(E)$ dada por

$$\pi_x^\beta = \frac{e^{\beta U_x}}{Z_\beta}, \quad \text{con}^{10} \quad Z_\beta := \sum_{x \in E} e^{\beta U_x}$$

Notamos que π_x^β favorece a los $x \in E$ cuyos valores de U_x sean más altos. Más aún:

$$\pi^\beta \xrightarrow{\beta \rightarrow \infty} \lambda$$

donde λ es una medida uniforme sobre el conjunto de los máximos de U .

Para cada $\beta > 0$, consideremos una matriz markoviana irreducible y aperiódica, tal que tiene a π^β como distribución invariante.

Sea G un grafo no dirigido sobre E , conexo, y sea

$$\begin{aligned} n_x &= |\{y \in E : xy \in G\}| \\ &= \#(\{\text{vecinos de } x\}), \end{aligned}$$

y definimos; para $x \neq y$

$$P_{xy}^\beta = \begin{cases} 1/n_x (e^{\beta(U_y - U_x)} \wedge 1), & xy \in G \\ 0, & \sim \end{cases}$$

y para $x = y$

$$P_{xx}^\beta = 1 - \sum_{y \neq x} P_{xy}^\beta.$$

Vemos que P^β hace improbables a las transiciones de estados que hacen disminuir el valor de U . Es fácil ver que π^β es invariante para P^β . Luego, si $(X_n^\beta)_{n \in \mathbb{N}}$ es una P^β -CMH, su ley converge a π^β cuando $n \rightarrow \infty$.

SIMULATED ANNEALING: Tomar $\beta = \beta_n$, con $\beta_n \rightarrow \infty$, lentamente.

Usualmente

$$\beta_n = \frac{1}{c} \ln(n + e),$$

con $c > 0$, una constante a calibrar. Esto genera

$$(X_n)_{n \geq 0} = (X_n^{\beta_n})_{n \geq 0},$$

Lo cual es una cadena de Markov *no-homogénea*.

La idea es tomar c suficientemente grande, de manera que X_n converga a un máximo de U , pero no tanto para que la convergencia ocurra en un tiempo razonable.

Ejemplo 6 Si

$$c > M := \max_{x \in E} (-U_x),$$

la convergencia está garantizada.

Observaciones: Típicamente, el algoritmo "recuerda" el $\hat{x} \in E$ visitado por $(X_n)_{n \geq 0}$ en que $U_{\hat{x}}$ es el máximo entre los visitados. Cuando se decide detener la cadena, digamos en el n_* , se retorna \hat{x} , no X_{n_*} .

8 Movimiento Browniano, proceso de difusión y aplicaciones

8.1 Movimiento Browniano

La idea general es poder escoger una función (o una asignación) en la variable $t \in \mathbb{R}^+$, de manera de poder asegurar la continuidad y mantener la aleatoriedad. En palabras simples, elegir aleatoriamente una función continua.

Una manera de hacer lo anterior es mediante un *paseo aleatorio*:

- Sean X_1, X_2, X_3, \dots variables aleatorias *i.i.d.*, tales que

$$\mathbb{P}(X_i = -1) = \mathbb{P}(X_i = 1) = \frac{1}{2}, \quad \forall i \in \mathbb{N}. \quad ^{11}$$

- $\forall n \in \mathbb{N}$,

$$S_n := \sum_{i=1}^n X_i, \quad \text{con } S_0 = 0.$$

- $\forall t \in \mathbb{R}^+$, S_t es la interpolación lineal de la colección $(S_n)_{n \in \mathbb{N}}$.

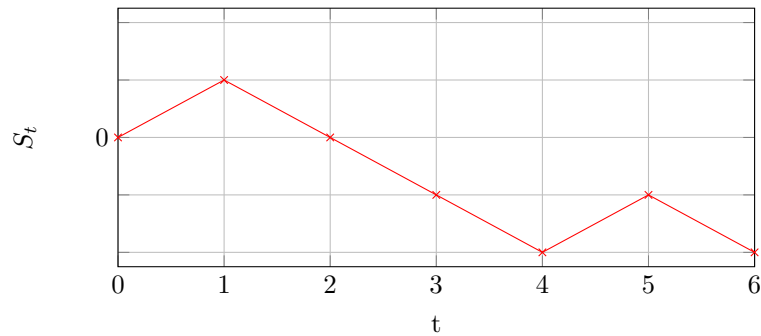


Figure 6: Ejemplo de *Paseo Aleatorio*.

Para valores de t suficientemente grande, y escalando tiempo y espacio adecuadamente, el gráfico de S_t toma la forma de la Figura 7.

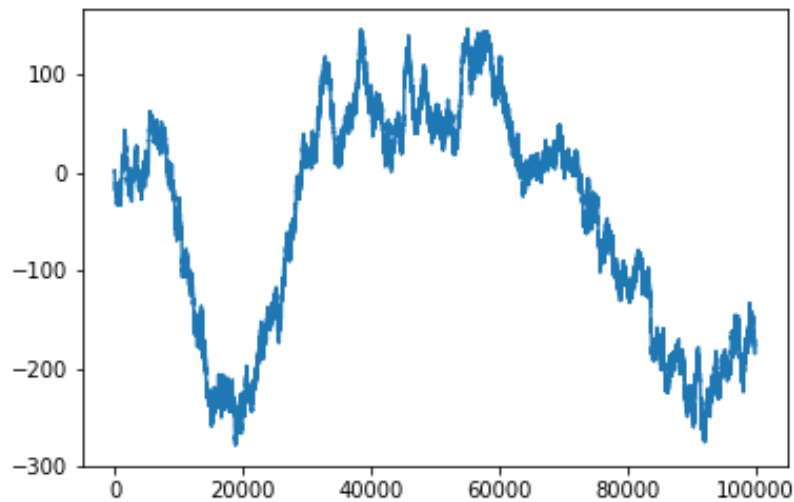


Figure 7: Simulación de movimiento Browniano.

¹¹En general, cualquier variable aleatoria con $\mathbb{E}(X) = 0$ y $\text{Var}(X) = 1$, puede servir.

Específicamente, el escalamiento corresponde al siguiente: $\forall a > 0$, definimos

$$B_t^{(a)} := \frac{1}{a} S_{ta^2}$$

Estudiaremos la distribución de $B_t^{(a)}$: para $a > 0$ fijo y $n \in \mathbb{N}$ (ambos suficientemente grandes), tomar $t = \frac{n}{a^2}$. Luego

$$B_t^{(a)} = \sqrt{\frac{t}{n}} S_n = \sqrt{t} \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$$

Ocupando *T.C.L.*, y recordando que $\mathbb{E}(X_i) = 0$ y $Var(X_i) = 1$, notamos que:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \approx Z \sim \mathcal{N}(0, 1),$$

por lo tanto, concluimos que:

$$B_t^{(a)} \approx \sqrt{t} \mathcal{N}(0, 1) = \mathcal{N}(0, t).$$

También es directo ver que, para valores enteros $m > n$, la variable $S_m - S_n$ es independiente de S_n (puesto que la colección $(X_i)_i$ era *i.i.d.*), es decir; $(S_n)_{n \in \mathbb{N}}$ posee la propiedad de **Incrementos Independientes**, lo cual se manifiesta también en $(B_t^{(a)})_{t \geq 0}$. (ver [4, cap. 2])

Lo anterior motiva la siguiente definición.

Definición 14 (Movimiento Browniano) *Un Movimiento Browniano es un proceso estocástico¹², denotado $(B_t)_{t \geq 0}$, que satisface:*

- i. $B_0 = 0$, c.s.
- ii. Posee incrementos independientes: $\forall t \geq 0$, $(B_{t+s} - B_t)_{s \geq 0}$ es independiente de $(B_s)_{0 \leq s \leq t}$.
- iii. Posee incrementos normales: $\forall t, s \geq 0$, $B_{t+s} - B_t \sim \mathcal{N}(0, s)$.
- iv. Es un proceso continuo; con probabilidad 1, la función $t \rightarrow B_t$ es continua en t .

La ley del proceso $(B_t^{(a)})_{t \leq 0}$ converge, cuando $a \rightarrow \infty$, a la ley del movimiento browniano $(B_t)_{t \geq 0}$, en un sentido adecuado (Teorema de Donsker, ver [4, cap.2, Teo.4.2]).

$(B_t)_{t \geq 0}$ también es llamado **Proceso de Wiener**. El nombre "browniano" fue puesto en honor al botánico escocés, Robert Brown, quien observó el movimiento errático de partículas de polen en el agua en 1827. Posteriormente, en 1905, el físico Albert Einstein explicaría este movimiento como resultado de muchas pequeñas colisiones de polen con las moléculas de agua circundantes.

Pero, ¿realmente existe un proceso tal que cumpla i, ii, iii, y iv, de la definición anterior?

Teorema 22 *El Movimiento Browniano existe.*

Demostración: ver [4, cap. 2].

Proposición 17 *Sea $(B_t)_{t \geq 0}$ Movimiento Browniano (M.B.), $\forall 0 = t_0 < t_1 < t_2 < \dots < t_n$, la colección*

$$\left(\frac{B_{t_i} - B_{t_{i-1}}}{\sqrt{t_i - t_{i-1}}} \right)_{i=1, \dots, n}$$

son i.i.d. $\mathcal{N}(0, 1)$.

¹²Es decir, una colección de variables aleatorias definidas en algún espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$, indexada por la variable $t \geq 0$.

Demostración: Directo de la definición de $M.B.$

■

Esto da lugar a un método sencillo para generar un $M.B.$ discretizado: para un horizonte $T > 0$, y $n \in \mathbb{N}$, sea $\Delta t = T/n$, y $t_i = i\Delta t$. Dado Z_1, \dots, Z_n , *i.i.d.* $\mathcal{N}(0, 1)$, entonces el proceso $(Y_t)_{t \in [0, T]}$, definido como

$$Y_{t_i} := \sqrt{\Delta t} \sum_{j=1}^i Z_j$$

interpolando linealmente entre los t_i 's para los valores en $[0, T] \setminus \{t_0, \dots, t_n\}$, es una aproximación de $(B_t)_{t \in [0, T]}$ (de hecho, tiene la misma ley en la malla $0 = t_0 < t_1 < \dots < t_n = T$).

Proposición 18 Sea $(B_t)_{t \geq 0}$ un $M.B.$:

- a. $\forall t_0 > 0$, el proceso $X_t := B_{t+t_0} - B_{t_0}$ es un $M.B.$. (Invarianza bajo shift)
- b. $\forall c > 0$, el proceso $X_t := \frac{1}{\sqrt{c}} B_{ct}$ es un $M.B.$ (Invarianza bajo escalamiento)
- c. El proceso $X_t := B_1 - B_{1-t}$ es un $M.B.$ en $[0, 1]$. (Propiedad de tiempo reverso)
- d. El proceso $X_t := tB_{1/t}$ es un $M.B.$ (con $X_0 = 0$). (Inversión del tiempo)
- e. el proceso $X_t := -B_t$ es un $M.B.$ (Simetría)

Demostración: ver [4, cap. 2].

Corolario 5 $\forall t, s \geq 0$:

$$\text{Cov}(B_t, B_s) = t \wedge s,$$

donde $t \wedge s = \min\{t, s\}$.

Demostración: Sin pérdida de generalidad podemos asumir que $s \leq t$, entonces:

$$\text{Cov}(B_t, B_s) = \mathbb{E}(B_t B_s) - \mathbb{E}(B_t) \mathbb{E}(B_s).$$

Recordando que, tanto B_t como B_s con movimientos brownianos estándar, por lo tanto $\mathbb{E}(B_t) = \mathbb{E}(B_s) = 0$, de esta forma tenemos:

$$\begin{aligned} \text{Cov}(B_t, B_s) &= \mathbb{E}(B_t B_s) = \mathbb{E}(B_t B_s - B_s^2 + B_s^2) \\ &= \mathbb{E}((B_t - B_s) B_s) + \mathbb{E}(B_s^2). \end{aligned}$$

Como B_s es de esperanza 0, entonces $\mathbb{E}(B_s^2) = \text{Var}(B_s) = s$, además notamos que la resta $B_t - B_s$ equivale (en distribución) a un movimiento browniano de la forma B_{t-s} , con s fijo y para todo $t \geq s$ (Proposición 18), además por la Proposición 17, $B_t - B_s$ es independiente de $B_s - 0 = B_s - B_0$. De esta forma tenemos:

$$\begin{aligned} \text{Cov}(B_t, B_s) &= \mathbb{E}((B_t - B_s) B_s) + \mathbb{E}(B_s^2) \\ &= \mathbb{E}(B_t - B_s) \mathbb{E}(B_s) + s \\ &= s. \end{aligned}$$

Donde se ha vuelto a ocupar que B_s es de esperanza 0.

■

A pesar de ser continuas, las trayectorias de un movimiento browniano son bastante irregulares. La *trayectoria* de un $M.B.$ es la función (aleatoria)

$$t \rightarrow B_t$$

Es posible probar que, para todo punto del espacio muestral $\omega \in \Omega$, la trayectoria dada por $B_t(\omega)$ es de variación acotada en cualquier intervalo finito $[0, T]$, sin embargo esta es **no-monótona** en cualquier intervalo, implicando la *no-diferenciabilidad* de la función. [4]

Teorema 23 (Paley-Wiener-Zygmund (1933)) Para casi todo $\omega \in \Omega$, la trayectoria del movimiento Browniano $B_t(\omega)$ es no diferenciable. Es decir,

$$\mathbb{P}(t \geq 0 : B_t \text{ es derivable en } t) = 0$$

El siguiente resultado es la famosa *Ley del logaritmo iterado*, que describe las oscilaciones del movimiento Browniano cerca de $t = 0$ y para $t \rightarrow +\infty$

Teorema 24 (Ley del Logaritmo Iterado (A.Hinêin-1933)) $\mathbb{P} - c.s.$ se cumple que:

i.

$$\limsup_{t \rightarrow 0^+} \frac{B_t}{\sqrt{2t \log(\log(1/t))}} = 1$$

ii.

$$\liminf_{t \rightarrow 0^+} \frac{B_t}{\sqrt{2t \log(\log(1/t))}} = -1$$

iii.

$$\limsup_{t \rightarrow +\infty} \frac{B_t}{\sqrt{2t \log(\log(t))}} = 1$$

iv.

$$\liminf_{t \rightarrow +\infty} \frac{B_t}{\sqrt{2t \log(\log(t))}} = -1$$

9 Martingalas

Dado un espacio de probabilidades $(\Omega, \mathcal{F}, \mathbb{P})$, se define una *filtración* como una colección $(\mathcal{F}_t)_{t \geq 0}$ de sub σ -álgebras de \mathcal{F} , crecientes respecto a la inclusión. Esto es:

$$\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}, \quad \forall 0 \leq s \leq t$$

A la tupla $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$ la llamamos *espacio de probabilidades filtrado*. Intuitivamente podemos pensar en \mathcal{F}_t como la información acumulada en el proceso estocástico involucrado en \mathcal{F} hasta el momento $t \geq 0$.

Dado un proceso $(X_t)_{t \geq 0}$, podemos definir la *filtración natural* del proceso como:

$$\mathcal{F}_t^X = \sigma(\{X_s : s \in [0, t]\}).$$

Donde (recordemos) la notación $\sigma(\mathcal{A})$ corresponde a la σ -álgebra generada por el conjunto \mathcal{A} . Un buen ejercicio sería probar que \mathcal{F}_t^X es efectivamente una filtración del espacio $(\Omega, \mathcal{F}, \mathbb{P})$.

Definición 15 El proceso estocástico X_t se dice *adaptado* a la filtración $\{\mathcal{F}_t\}$ si, para cada $t \geq 0$, X_t es una variable aleatoria \mathcal{F}_t -medible.

Es fácil notar que todo proceso estocástico $(X_t)_{t \geq 0}$ es adaptado a su filtración natural \mathcal{F}_t^X . Desde ahora, y salvo que se diga antes, siempre trabajaremos con procesos adaptados.

Dado un espacio filtrado¹³ $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$ y un movimiento browniano, $(B_t)_{t \geq 0}$ con respecto a este espacio. Se puede probar que B_t es un proceso adaptado a este espacio, cumpliendo *i*, *iii* y *iv* de la definición 14, y donde es posible reemplazar la condición *ii* por

$$(B_{t+s} - B_t)_{s \geq 0} \perp\!\!\!\perp \mathcal{F}_t.$$

En palabras, el movimiento dado por $B_{s+t} - B_t$ es independiente a la sub σ -álgebra \mathcal{F}_t .

Los siguientes resultados son recuerdos de algunas propiedades de las **esperanzas condicionales** que serán de utilidad más adelante.

Proposición 19 Si X es una v.a. en el espacio $(\Omega, \mathcal{F}, \mathbb{P})$ y $\mathcal{G} \subseteq \mathcal{F}$ es una sub σ -álgebra, la **esperanza condicional** $Y = \mathbb{E}(X|\mathcal{G})$ es la única v.a. (\mathbb{P} -c.s.) \mathcal{G} -medible que cumple que

$$\mathbb{E}(\mathbb{1}_A X) = \mathbb{E}(\mathbb{1}_A Y)$$

Se cumple además:

- Si X es \mathcal{G} -medible $\Rightarrow \mathbb{E}(X|\mathcal{G}) = X$ c.s.
- Si $X \perp\!\!\!\perp \mathcal{G} \Rightarrow \mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X)$ c.s.
- Si Z es \mathcal{G} -medible $\Rightarrow \mathbb{E}(ZX|\mathcal{G}) = Z\mathbb{E}(X|\mathcal{G})$ c.s.
- Si $\mathcal{H} \subseteq \mathcal{G}$ es sub σ -álgebra $\Rightarrow \mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H}) = \mathbb{E}(X|\mathcal{H})$ c.s.

Dado un espacio filtrado $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$, se define una **martingala** como un proceso adaptado $(X_t)_{t \geq 0}$ tal que $X_t \in L^1 \forall t$ y $\forall, 0 \leq s \leq t$:

$$\mathbb{E}(X_t|\mathcal{F}_s) = X_s \quad \mathbb{P} - \text{c.s.}$$

La definición de martingala formaliza el concepto matemático de “juego justo” al otorgarle el mismo valor esperado a la variable a cada momento. Si interpretamos a X_t como el dinero ganado en un juego de apuestas sucesivas al momento t , entonces la definición anterior nos dice que lo que espero tener al momento t , si ya estoy en el momento s ($t \geq s$), es exactamente el dinero que tengo en el momento s sin importar lo ocurrido entremedio o anteriormente.

¹³Omitiremos (y asumiremos) que hablamos de un espacio de probabilidades y aclararemos cuando este no sea el caso.

Proposición 20 Dado un M.B., $(B_t)_{t \geq 0}$ en $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$.

- i. $(B_t)_{t \geq 0}$ es una mg.¹⁴.
- ii. $(B_t^2 - t)_{t \geq 0}$ es una mg.
- iii. $\forall \alpha > 0$, $\exp(\alpha B_t - \alpha^2 t/2)$ es una mg.

Demostración: i. Notemos que para cada $t \geq 0$, la sub σ -álgebra \mathcal{F}_t contiene la información del proceso hasta el tiempo t , esto incluye la trayectoria recorrida por B_t , lo que la hace \mathcal{F}_t -medible. Ahora, aplicando la propiedad de incrementos normales (definición 14, punto iii.) tenemos que para $0 \leq s \leq t$:

$$(B_t - B_s)|\mathcal{F}_s \sim \mathcal{N}(0, t - s),$$

por lo tanto

$$\begin{aligned} \mathbb{E}(B_t - B_s|\mathcal{F}_s) &= 0 \\ \Rightarrow \mathbb{E}(B_t|\mathcal{F}_s) &= \mathbb{E}(B_s|\mathcal{F}_s). \end{aligned}$$

Luego, como B_s es \mathcal{F}_s -medible, ocupando el primer punto de la proposición 19, obtenemos que

$$\mathbb{E}(B_s|\mathcal{F}_s) = B_s,$$

es decir:

$$\mathbb{E}(B_t|\mathcal{F}_s) = B_s.$$

Para demostrar ii. notemos que, para $0 \leq s \leq t$:

$$B_t^2 = (B_t - B_s)^2 + 2B_s(B_t - B_s) + B_s^2,$$

por lo tanto tenemos la siguiente igualdad

$$\mathbb{E}(B_t^2|\mathcal{F}_s) = \mathbb{E}((B_t - B_s)^2|\mathcal{F}_s) + 2\mathbb{E}(B_s(B_t - B_s)|\mathcal{F}_s) + \mathbb{E}(B_s^2|\mathcal{F}_s).$$

Primero, veamos que tanto B_s como B_s^2 son \mathcal{F}_s -medibles, ocupando el tercer y primer punto de la proposición 19, podemos reescribir la igualdad anterior como

$$\mathbb{E}(B_t^2|\mathcal{F}_s) = \mathbb{E}((B_t - B_s)^2|\mathcal{F}_s) + 2B_s\mathbb{E}((B_t - B_s)|\mathcal{F}_s) + B_s^2.$$

Ahora, por la propiedad de incrementos independientes de los M.B., sabemos que $B_t - B_s \perp\!\!\!\perp \mathcal{F}_s$, por lo tanto también $(B_t - B_s)^2 \perp\!\!\!\perp \mathcal{F}_s$. Con esto podemos aplicar el segundo punto de la proposición 19:

$$\begin{aligned} \mathbb{E}(B_t^2|\mathcal{F}_s) &= \mathbb{E}((B_t - B_s)^2|\mathcal{F}_s) + 2B_s\mathbb{E}((B_t - B_s)|\mathcal{F}_s) + B_s^2 \\ &= \mathbb{E}((B_t - B_s)^2) + 2B_s\mathbb{E}(B_t - B_s) + B_s^2. \end{aligned}$$

Recordando que $B_t - B_s \sim \mathcal{N}(0, t - s)$, obtenemos que $\mathbb{E}(B_t - B_s) = 0$ y $\mathbb{E}((B_t - B_s)^2) = \text{Var}(B_t - B_s) = t - s$. Así,

$$\begin{aligned} \mathbb{E}(B_t^2|\mathcal{F}_s) &= t - s + B_s^2. \\ \Rightarrow \mathbb{E}(B_t^2 - t|\mathcal{F}_s) &= B_s^2 - s, \end{aligned}$$

probando el punto ii.

Ocupando argumentos similares que en la parte anterior, tenemos que:

$$\mathbb{E}(e^{\alpha B_t}|\mathcal{F}_s) = \mathbb{E}(e^{\alpha(B_t - B_s)}e^{\alpha B_s}|\mathcal{F}_s),$$

ocupamos que $e^{\alpha B_s}$ es \mathcal{F}_s -medible, $e^{\alpha(B_t - B_s)} \perp\!\!\!\perp \mathcal{F}_s$ y que $(B_t - B_s) \sim \mathcal{N}(0, t - s)$ para afirmar que

$$\begin{aligned} \mathbb{E}(e^{\alpha B_t}|\mathcal{F}_s) &= \mathbb{E}(e^{\alpha(B_t - B_s)}e^{\alpha B_s}|\mathcal{F}_s) = e^{\alpha B_s}\mathbb{E}(e^{\alpha(B_t - B_s)}|\mathcal{F}_s) \\ &= e^{\alpha B_s}\mathbb{E}(e^{\alpha(B_t - B_s)}) = e^{\alpha B_s}e^{\alpha^2(t-s)/2}. \end{aligned}$$

Es decir;

$$\mathbb{E}(\exp(\alpha B_t - \alpha^2 t/2)|\mathcal{F}_s) = \exp(\alpha B_s - \alpha^2 s/2).$$

■

¹⁴A.k.a.: Martingala.

9.1 Tiempos de parada (t.d.p.)

La noción detrás de un **Tiempo de Parada**, es de una variable aleatoria que “acumule” información temporal del proceso hasta cierto tiempo dado, sin el uso de la información futura.

Definición 16 Una variable T es llamada **tiempo de parada**, si es una variable aleatoria a valores en $[0, +\infty]$, tal que $\{T \leq t\} \in \mathcal{F}_t, \forall t \geq 0$.

Ejemplos:

- Si definimos T_a como la primera vez que el M.B. (B_t) alcanza el punto $a \in \mathbb{R}$, es decir,

$$T_a := \inf\{t \geq 0 \mid B_t = a\} \quad (9.1.1)$$

Entonces T_a es t.d.p.

- Definamos S_0 como la última vez antes de $t = 1$ en que el movimiento browniano, B_t , pasa por el 0. Es decir,

$$S_0 = \sup\{t \in [0, 1] \mid B_t = 0\}.$$

Notemos que S_0 **no es t.d.p.**, dado que conocer el supremo implica saber que el proceso B_t no vuelve a cero en todo el intervalo $(S_0, 1]$ “usando información futura”.

Dada una mg. $(X_t)_{t \geq 0}$, ¿será posible obtener un beneficio (en esperanza) al detener el proceso en un t.d.p. T ? Bajo ciertas condiciones de T , esto no sería posible.

Dado un t.d.p. T , se define la σ -álgebra \mathcal{F}_T como

$$\mathcal{F}_T := \{A \in \mathcal{F} \mid \forall t \geq 0, A \cap \{T \leq t\} \in \mathcal{F}_t\}$$

Teorema 25 (Muestreo opcional de Doob) Sea $(X_t)_{t \geq 0}$ mg, sean T, S , t.d.p. tal que $S \leq T \leq K$ (\mathbb{P} -c.s.), con K una constante positiva. Entonces

$$\mathbb{E}(X_T | \mathcal{F}_S) = X_S, \quad \mathbb{P} - \text{c.s.}$$

En particular tenemos que $\mathbb{E}(X_T) = \mathbb{E}(X_0)$.

La condición $T \leq K$, suele ser muy restrictiva. Típicamente, se aplica el teorema anterior a $T \wedge n$, para luego hacer $n \rightarrow +\infty$. Esta técnica se denomina *localización*.

Proposición 21 Sea $a \in \mathbb{R}$, sea T_a como en la expresión 9.1.1. Entonces $T_a < \infty$ ($\mathbb{P} - \text{c.s.}$) y su ley viene dada por su transformación de Laplace

$$\mathbb{E}(e^{-\lambda T_a}) = e^{-\sqrt{2\lambda}|a|}, \quad \forall \lambda > 0. \quad (9.1.2)$$

O, equivalentemente, su densidad está dada por

$$f_{T_a}(x) = \frac{|a|}{\sqrt{2\pi x^3}} e^{-a^2/2x}, \quad \forall x > 0$$

Demostración: Supongamos $a > 0$, para $\eta > 0$, definamos la mg:

$$X_t = e^{\eta B_t - \eta^2 t/2},$$

donde B_t es M.B. estándar.

Para cada $n \in \mathbb{N}$, el t.d.p. $T_a \wedge n$ es acotado. Ocupando el Teorema 25 tenemos que

$$\mathbb{E}(X_{T_a \wedge n}) = \mathbb{E}(X_0) = 1.$$

Además notamos que $X_{T_a \wedge n}$ es acotada (como sucesión en n). En efecto, como $B_{T_a \wedge n} \leq a$:

$$X_{T_a \wedge n} = e^{\eta B_{T_a \wedge n} - \eta^2 (T_a \wedge n)/2} \leq e^{\eta a}.$$

Dado el evento $\{T_a < \infty\}$, entonces $X_{T_a \wedge n} \rightarrow X_{T_a}$ cuando $n \rightarrow \infty$. De lo contrario, dado $\{T_a = \infty\}$, para todo $n \in \mathbb{N}$ tenemos que $X_{T_a \wedge n} = X_n$ y

$$X_n = e^{\eta B_n - \eta^2 n/2} \rightarrow 0$$

cuando $n \rightarrow \infty$, puesto que $B_n < a$. Juntando las expresiones anteriores y ocupando el **T.C.D**, obtenemos que:

$$\begin{aligned} 1 &= \lim_{n \rightarrow \infty} \mathbb{E}(X_{T_a \wedge n}) = \mathbb{E}\left(\lim_{n \rightarrow \infty} X_{T_a \wedge n}\right) \\ &= \mathbb{E}\left(\lim_{n \rightarrow \infty} \mathbb{1}_{\{T_a < \infty\}} X_{T_a \wedge n} + \lim_{n \rightarrow \infty} \mathbb{1}_{\{T_a = \infty\}} X_{T_a \wedge n}\right) \\ &= \mathbb{E}\left(\mathbb{1}_{\{T_a < \infty\}} X_{T_a} + 0\right) \\ &= \mathbb{E}\left(\mathbb{1}_{\{T_a < \infty\}} e^{\eta a - \eta^2 T_a/2}\right) \end{aligned}$$

Es decir, $\mathbb{E}\left(\mathbb{1}_{\{T_a < \infty\}} e^{-\eta^2 T_a/2}\right) = e^{-\eta a}$. Tomando $\eta \rightarrow 0^+$ tenemos $\mathbb{P}(T_a < \infty) = 1$ y con esto $\mathbb{E}\left(e^{-\eta^2 T_a/2}\right) = e^{-\eta a}$. La expresión 9.1.2 se obtiene tomando $\eta = \sqrt{2\lambda}$.

El caso $a < 0$ se deduce sabiendo que $-B_t$ también es m.b.

■

Teorema 26 (Desigualdad de Doob) Si $(X_t)_{t \in [0, T]}$ es mg. continua¹⁵, entonces:

$$\mathbb{E}\left(\sup_{0 \leq t \leq T} X_t\right)^2 \leq 4 \mathbb{E}(X_T^2)$$

Demostración: Un caso más general se puede encontrar en el Teorema 3.8 de [4]. ■

¹⁵De trayectoria continua.

References

- [1] BILLINGSLEY, P., *Convergence of Probabilistic Measure*, second edition, Wiley series in probability and statistic, 1999.
- [2] UNIVERSITY OF WATERLOO , Notas de curso "STAT 901: Probability" formato PDF, http://sas.uwaterloo.ca/~dlmcleis/s901/s901_2005.pdf, 2005.
- [3] PARDOUX, E., *Markov Processes and Applications*, Wiley series in probability and statistic, 2008.
- [4] KARATZAS, I.; SHREVE, S., *Brownian Motion and Stochastic Calculus*, Second Edition, Editorial Springer, 1991.
- [5] LAMPEYRE, B.; LAMBERTON, D. *Introduction to Stochastic Calculus Applied to Finance*, Second Edition, CHAPMANHALL/CRC, 2008.