Ingeniería Matemática
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

Apuntes Simulación Estocástica

MA4402 - Simulación Estocástica: Teoría y Laboratorio

DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

Profesor: Joaquín Fontbona 10 de agosto de 2023

El presente apunte contiene los contenidos del curso MA4402 Simulación Estocástica: Teoría y Laboratorio. Este curso es de carácter obligatorio para la carrera de Ingeniería Civil Matemática de la Universidad de Chile.

La versión presentada en este texto fue mayormente desarrollada entre los meses de agosto 2021 y agosto 2022, y toma como base, para las cinco primeras unidades, el curso tal como fue dictado el semestre de primavera 2021. Para la unidad 6, se utilizaron anotaciones del curso dictado en años anteriores por Roberto Cortez. La transcripción y edición de todo el texto fue realizada por Camilo Carvajal.

Indice

1. Repaso y preliminares

La presente sección está dedicada a aquellos elementos útiles de probabilidades y teoría de la medida y que serán utilizados a lo largo del curso. Primero se definirán objetos conocidos de probabilidades, para luego dar paso a un repaso de esperanza condicional.

Luego nos enfocaremos en convergencia en ley, que es una herramienta de suma importancia y que estará presente en las siguientes unidades. Este concepto, que suele ser visto en cursos de Probabilidades, está subyacente en varios algoritmos, además de ser la base del Teorema central del límite. Este último será también abordado al final de esta unidad.

1.1. Repaso Probabilidades

1.1.1. Ley y Esperanza

Consideraremos $(\Omega, \mathcal{F}, \mathbb{P})$ espacio de probabilidad (e.d.p.), (E, Σ) espacio medible y $X : \Omega \longrightarrow E$ variable aleatoria (función medible).

Definición 1.1.1 Ley de X
La *ley* de X es la medida de probabilidad:

$$\begin{aligned} \mu := \mathbb{P} \circ X^{-1} : \Sigma &\longrightarrow [0,1] \\ A \in \Sigma &\longmapsto \mu(A) = \mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A)) \end{aligned}$$

Corresponde a la medida inducida por $\mathbb{P}$ y X en Σ.

Notación.

- $\mu := Ley(X)$ o $X \sim \mu$.
- $\langle \mu, f \rangle = \int f(x) d\mu(x) = \int f(x)\mu(dx) \quad \forall f \in L^1(E, \Sigma, \mu)$.

Se usarán tanto $\langle \mu, f \rangle$ (que nos sugiere una noción de dualidad) como $\int f(x)d\mu(x) = \int f(x)\mu(dx)$, que es una versión más "probabilista" que aquella que suele usarse en cursos de teoría de la medida.

Definición 1.1.2 Esperanza
Si $Y : \Omega \longrightarrow \mathbb{R}$ es variable aleatoria e $Y \in L^1(\Omega, \mathcal{F}, \mathbb{P}), Y \geq 1$, $\mathbb{E}(Y)$ denota la integral de Lebesgue de Y con respecto a $\mathbb{P}$ y se define la esperanza de Y como sigue:

- $\mathbb{E}(Y) = \mathbb{P}(B)$ cuando $Y = \mathbf{1}_B$ con $B \in \mathcal{F}$.
- $\mathbb{E}(Y) = \sum_{i=1}^{n} b_i \mathbb{P}(B_i)$ cuando $Y = \sum_{i=1}^{n} b_i \mathbf{1}_{B_i}$ con $B_i \in \mathcal{F}$, i.e., cuando Y es una función simple.
- Para $Y \geq 0$, $\mathbb{E}(Y) = \lim_{n\to\infty} \mathbb{E}(Y_n)$ con $(Y_n)_{n\in\mathbb{N}}$ sucesión creciente de funciones simples tal que $Y_n \nearrow_{n\to\infty} Y$.
- $\mathbb{E}(Y) = \mathbb{E}(Y_+) - \mathbb{E}(Y_-)$ para $Y \in L^1$.

Proposición 1.1.1
Sea $X : \Omega \longrightarrow E$ v.a., $f : (E, \Sigma) \longrightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ medida ≥ 0 $f \in L^1(E, \Sigma, \mu) =_\mu Ley(X)$. Entonces

$$\mathbb{E}(f(X)) = \langle \mu, f \rangle, \quad \forall f \in L^1(\mu).$$

Más aún, $f \in L^1(\mu)$ si, y sólo si $f(X) \in L^1(\Omega, \mathcal{F}, \mathbb{P})$.

Demostración. Ejercicio
Indicación: demostrar primero para indicatrices de conjuntos medibles, luego para funciones simples, positivas y finalmente concluir el caso general.

Observación 1.1.1.

1. Si X es v.a. "discreta", tenemos que

$$\mu = Ley(X) = \sum_x p_x \delta_x$$

$$\mathbb{E}(f(X)) = \int f(x)\mu(dx) = \sum_x f(x)p_x \,.$$

   En lo anterior, δ_x son masas de Dirac, $\sum_x p_x = 1$ y las sumas son finitas o numerables.

2. Si X es v.a. absolutamente "continua", esto es, posee densidad f_X, entonces

$$\mu(dx) = f_X(x)dx \,,$$

$$\mathbb{E}(\varphi(X)) = \int_{\mathbb{R}} \varphi(x)\mu(dx) = \int \varphi(x) f_X(x)dx.$$

1.1.2. Esperanza Condicional

Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad completo y $G \subset \mathcal{F}$ una sub-σ-álgebra. Consideramos las siguientes interpretaciones:

- $\omega \in \Omega$ serán los "estados posibles de la naturaleza"
- $\mathcal{F}$ conjuntos cuya ocurrencia somos capaces de distinguir: dado $\omega \in \Omega$ y $B \in \mathcal{F}$, podemos responder si $\omega \in B$ (eventos), y podemos "medir", i.e., calcular $\mathbb{P}(B)$
  $\therefore$ $\mathcal{F}$ representa a qué información tenemos acceso.
- $\mathcal{G}$, al ser una sub-σ-álgebra, posee menos información (reconoce menos eventos).

Un ejemplo simple es pensar en un termómetro, con el cual podremos responder a preguntas acerca del clima. De cierto modo esta herramienta nos está definiendo una σ-álgebra, pues podemos "medir" eventos como si la temperatura está o no dentro de cierto rango. Por otro lado, si tuviésemos más herramientas, por ejemplo un pluviómetro además de un termómetro, entonces la σ-álgebra correspondiente sería más grande pues nos permite acceder a más información (y la σ-álgebra del termómetro sólo es una sub-σ-álgebra de esta).

Definición 1.1.3 Esperanza condicional, caso L^2
Sea $X \in L^2(\Omega, \mathcal{F}, \mathbb{P}), \mathcal{G} \subset \mathcal{F}$ una sub-σ-álgebra. Se define la esperanza condicional de X dado $\mathcal{G}$, denotada $\mathbb{E}(X|\mathcal{G})$, como la **proyección ortogonal** desde $L^2(\Omega, \mathcal{F}, \mathbb{P})$ de X en el subespacio vectorial $L^2(X, \mathcal{G}, \mathbb{P})$.

Observación 1.1.2.

- $L^2(X, \mathcal{G}, \mathbb{P})$ es cerrado en $L^2(\Omega, \mathcal{F}, \mathbb{P})$.
  En efecto $Z_n \xrightarrow[n\to\infty]{L^1} Z$ con $Z_n \in L^2(X, \mathcal{G}, \mathbb{P})$ implica que existe una subsucesión que converge a Z casi seguramente. Por lo tanto $Z \in \mathcal{G}$
- $\mathbb{E}(X|\mathcal{G})$ es $\mathcal{G}$-medible.
- $\mathbb{E}(\cdot|\mathcal{G}) : \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P}) \mapsto L^2(\Omega, \mathcal{G}, \mathbb{P})$ es bilineal y continua.

Observación 1.1.3 (Propiedad fundamental). $\mathbb{E}(X|\mathcal{G})$ queda caracterizada como la única variable aleatoria tal que:

1. $\mathbb{E}(X|\mathcal{G}) \in L^2(\Omega, \mathcal{G}, \mathbb{P})$
2. $\mathbb{E}(\mathbb{E}(X|\mathcal{G})|Z) = \mathbb{E}(X|Z) \quad \forall Z \in L^2(\Omega, \mathcal{G}, \mathbb{P})$

   En particular tenemos $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(X)$ y si además $X \in L^2(\Omega, \mathcal{G}, \mathbb{P})$, $E(X|\mathcal{G}) = X$

La siguiente propiedad es un ejercicio fácil.

Propiedad 1.1.1

1. $Z = \mathbb{E}(X|\mathcal{G})$ minimiza $Z \in L^2(\Omega, \mathcal{G}, \mathbb{P}) \mapsto \mathbb{E}((Z - X)^2)$.
2. Como consecuencia de lo anterior, tenemos la siguiente **interpretación estadística**: La v.a. $\mathcal{G}$-medible $\mathbb{E}(X|\mathcal{G})$ es el **mejor estimador de** X en el sentido de tener **menor error cuadrático medio** (en inglés *MSE*) usando la **información accesible** para la σ-álgebra $\mathcal{G}$.
3. Si $\mathcal{G}$ es la tribu trivial ($G = \{\emptyset, \Omega\}$), toda función $\mathcal{G}$-medible es constante. Luego $\mathbb{E}(X|\mathcal{G})$ es constante tal que $\mathbb{E}(E(X|\mathcal{G})) = \mathbb{E}(X) \implies \mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X)$.
   Dicho de otro modo, la mejor estimación es "trivial" y no usa información.

Lema 1.1.1
Sea $Y : \Omega \mapsto E$ v.a. y $Z : \Omega \mapsto \mathbb{R}$ v.a. medible con respecto a $\mathcal{G} = \sigma(Y) := \{Y^{-1}(A) : A \in \Sigma\} \subset \mathcal{F}$ (con Σ σ-álgebra de E), entonces existe $h : E \mapsto \mathbb{R}$ medible tal que $Z = h(Y)$.

Observación 1.1.4.

- En particular para $Z = \mathbb{E}(X|\mathcal{G})$, $\mathcal{G} = \sigma(Y)$ escribimos $\mathbb{E}(X|Y = y) := h(y)$, de modo que
$$\mathbb{E}(X|Y) = \mathbb{E}(X|\sigma(Y)) = h(Y) = \mathbb{E}(X|Y = y)|_{y=Y}\,.$$
- Si $Y = (Y_1, \ldots, Y_d) \in \mathbb{R}^d$, $\mathbb{E}(X|\sigma(Y_1, \ldots, Y_d)$ se denota $\mathbb{E}(X|Y_1, \ldots, Y_d)$. Por lo anterior, es una función de $(Y_1, \ldots, Y_d)$.

Demostración. (del Lema)

- Primero asumimos que $Z = \mathbf{1}_B$ con $B \in \sigma(Y)$, es decir $B = Y^{-1}(A)$ para cieto $A \in \Sigma$.
  Entonces $\mathbf{1}_B = \mathbf{1}_{Y^{-1}(A)} = \mathbf{1}_A(Y)$, luego $Z = h(Y)$ con $h(y) = \mathbf{1}_A(y)$.

- Ahora tomemos $Z = \sum_{i=1}^{n} b_i \mathbf{1}_{B_i}$ con $B_i = Y^{-1}(A_i)$, $A_i \in \Sigma$, $\forall i \in \{1, \dots, n\}$.
  Así, $Z = \sum_{i=1}^{n} b_i \mathbf{1}_{A_i}(Y)$, luego $Z = h(Y)$ con $h(y) = \sum_{i=1}^{n} b_i \mathbf{1}_{A_i}(y)$.

- Sea $Z \geq 0$, entonces existe una sucesión $(Z_k)_k$, todos $\mathcal{G} - medibles$ y $h^k : E \mapsto \mathbb{R}$ medibles tal que $h^k(Y) = Z_k \overset{\nearrow}{_{k\to\infty}} Z$ (puntualmente) c.s. . Entonces podemos definir:

$$h(y) = \begin{cases} \limsup_{k\to\infty} h^k(y) & \text{si } y \in Y(\Omega) \\ 0 & \text{si } y \notin Y(\Omega) \end{cases}.$$

  Luego dado que $Z_K \overset{\nearrow}{_{k\to\infty}} Z$, queda que $h(Y) = \limsup_{k\to\infty} Z_k$.

- El caso general se deduce de lo anterior y queda propuesto.

□

Teorema 1.1.1 Esperanza condicional, caso general L^1
Sean $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ v.a. y $\mathcal{G} \subset \mathcal{F}$ sub-σ-álgebra. Entonces existe una única variable aleatoria $Z \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ tal que:

- $Z \in L^1(\Omega, \mathcal{G}, \mathbb{P})$
- $\mathbb{E}(XH) = \mathbb{E}(ZH) \quad \forall H \in L^\infty(\Omega, \mathcal{G}, \mathbb{P})$ *(propiedad fundamental)*

Notación. Denotamos Z como $\mathbb{E}(X|\mathcal{G})$ y la llamamos **Esperanza condicional de X dado $\mathcal{G}$**

Observación 1.1.5.

- $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(X)$ (con $H = 1$)
- La propiedad fundamental equivale a $\mathbb{E}(X\mathbf{1}_A) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})\mathbf{1}_A) \forall A \in \mathcal{G}$
  Esto se demuestra usando aproximación por funciones simples y T.C.M. (Ejercicio)

Demostración. (del Teorema)
Ejercicio Para la existencia cuando $X \geq 0$ considerar $X_n := \min(X, n) \forall n \in \mathbb{N}$ y ver que $\mathbb{E}(X_{n+1}|\mathcal{G}) \geq \mathbb{E}(X_n|\mathcal{G})$ c.s. . Por T.C.D. se verifica que $X_n X$ y entonces tomando $Z = \lim_{n\to\infty} \nearrow \mathbb{E}(X_n|\mathcal{G})$ se prueba que $\mathbb{E}(XH) = \mathbb{E}(\mathbb{E}(ZH)) \forall H \in \mathcal{L}^\infty(\mathcal{G})$.

Para unicidad primero ver que si tomamos Z, Z' tal que satisfacen la propiedad fundamental entonces $\mathbb{E}((Z - Z')_{\{Z<Z'\}}) = 0$. Notar que lo anterior es simétrico y usarlo para concluir que $Z = Z'$ c.s. .

Propiedad 1.1.2

(i) $\mathbb{E}(\cdot|\mathcal{G}) : L^1(\Omega, \mathcal{F}, \mathbb{P}) \mapsto L^1(\Omega, \mathcal{G}, \mathbb{P})$ es una aplicación lineal continua

(ii) Si $X \in L^1(\Omega, \mathcal{G}, \mathbb{P})$ entonces $\mathbb{E}(X|\mathcal{G}) = X$

(iii) Si $F \in L^\infty(\Omega, \mathcal{G}, \mathbb{P})$ entonces $\mathbb{E}(XF|\mathcal{G}) = F\mathbb{E}(X|\mathcal{G})$

(iv) Si $\mathcal{H} \subseteq \mathcal{G} \subseteq \mathcal{F}$ σ-álgebras entonces $\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H}) = \mathbb{E}(\mathbb{E}(X|\mathcal{H})|\mathcal{G}) = \mathbb{E}(X|\mathcal{H})$

Demostración.

(i) Ejercicio

(ii) Ejercicio

(iii) Sean F, $H \in L^\infty(\mathcal{G})$, $\mathbb{E}((XF)H) = \mathbb{E}(X(FH)) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})FH)$
Como $\mathbb{E}(X|\mathcal{G})F \in L^1(\Omega, \mathcal{G}, \mathbb{P})$, por (ii), y tomando $H = 1$, $\mathbb{E}(XF) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})F) = \mathbb{E}(X|\mathcal{G})F$.

(iv) La segunda igualdad es directa pues $\mathbb{E}(X|\mathcal{H})$ es en particular $\mathcal{G}$-medible. Para la primera tomemos $H \in L^\infty(\mathcal{H})$, como $H \in L^\infty$ tenemos $\mathbb{E}(XH) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})H) = \mathbb{E}(\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H})H)$ pues $\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H})$ es $\mathcal{H}$-medible, entonces $\mathbb{E}(X|\mathcal{H}) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H})$, pero como tenemos que $\mathbb{E}(X|\mathcal{H}) = \mathbb{E}(\mathbb{E}(X|\mathcal{H})|\mathcal{G})$ (segunda igualdad), entonces concluimos que $\mathbb{E}(\mathbb{E}(X|\mathcal{H})|\mathcal{G}) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H})$.

□

Ejemplo 1.1.1 Aterrizando el concepto

- Sean $(B_n)_{n\in\mathbb{N}} \subset \mathcal{F}$ partición de Ω y $\mathcal{G} := \sigma((B_n)_n)$
  Se puede probar que $\mathcal{G} = \{\cup_{j\in J} B_j : J \subseteq \mathbb{N} \text{ numerable o finito }\} \cup \{\emptyset\}$ (Ejercicio).
  Sea $X \in L^1$, la esperanza condicional está dada por:

$$\mathbb{E}(X|\mathcal{G}) = \sum_{n\in\mathbb{N}} \mathbf{1}_{B_n} \mathbb{E}(X|B_n)\,.$$

  Demostración. $\sum_{n\in\mathbb{N}} \mathbf{1}_{B_n}\mathbb{E}(X|B_n)$ es $\mathcal{G}$-medible y está en L^1. Por otro lado, $\forall A \in \mathcal{G}$, $\exists (B_n)_{n\in\mathbb{N}}$ $A = \dot\cup_{j\in J} B_j$, luego se tiene

$$\mathbb{E}((\sum_{n\in\mathbb{N}} \mathbf{1}_{B_n}\mathbb{E}(X|B_n))\mathbf{1}_A) = \mathbb{E}(\sum_{j\in J} \mathbf{1}_{B_j}\mathbb{E}(X|B_j)) = \mathbb{E}(X\mathbf{1}_A)\,.$$

- Sean $(X, Y) \in \mathbb{R}^2$ par aleatorio continuo con densidad $f_{(X,Y)}$. La densidad condicional de $X|Y = y$ se define como $f_{X|Y}(x|y) = \dfrac{f_{(X,Y)}(x,y)}{f_Y(y)}\mathbf{1}_{\{f_Y(y)>0\}}$.
  Ejercicio: demostrar que

$$\mathbb{E}(X|Y)(\omega) = [\int x f_{X|Y}(x|y)dx]|_{y=Y(\omega)}\,.$$

1.2. Convergencia en Ley de variables aleatorias

La convergencia en ley es una herramienta que aparece varias veces a lo largo de este curso. A continuación se fijarán las bases teóricas de esta convergencia en un contexto más general del que se suele ver en cursos de probabilidades. Parte de esta subsección está basada en el libro de Billingsley [1].

Notación.

- (E, d) espacio métrico, $\mathcal{B}$ tribu boreliana
- $\mathcal{M}(E)$ medidas finitas ≥ 0 sobre E
- $\mathcal{P}(E)$ medidas de probabilidad, $\mathcal{M}_s(E)$ medidas con sigma finitas
- $\mathcal{C}_b(E)$ funciones continuas acotadas
- $BL(E)$ funciones Lipschitz acotadas
- Integral de f con respecto a $\mu \in \mathcal{M}(E)$ (con f medible y acotada):

$$\langle \mu, f \rangle = \int f(x)\mu(dx)$$

1.2.1. Definición de convergencia débil y en ley

Observación 1.2.1. $\nu \in \mathcal{M}(E)$ queda caracterizada por $\langle \nu, f \rangle$, $f \in \mathcal{C}_b(E)$

Demostración. Sea $F \subset E$ cerrado y $\epsilon > 0$. Consideramos $f_\epsilon(x) = \max(0, 1 - \frac{d(x,F)}{\epsilon})$ con $d(x, F) = \inf_{y \in F}(x, y)$. Notemos que esta función es $\frac{1}{\epsilon}$-Lipschitz y que $\mathbf{1}_F \leq f_\epsilon \leq \mathbf{1}_{F^\epsilon} \underset{\epsilon \to \infty}{\searrow} \mathbf{1}_F$, donde $F^\epsilon = \{x \in E : d(x, F) < \epsilon\}$. Entonces por teorema de convergencia dominada,

$$\langle \nu, f_\epsilon \rangle \underset{\epsilon \to \infty}{\searrow} \nu(F)\,.$$

Sea $\mathcal{H} = \{A \subset E \text{ tal que } \mu(A) = \sup_{F \subset A \text{cerrado}} \mu(F) = \mu(A) = \inf_{F \subset A \text{ abierto}} \mu(F)\}$. $\mathcal{H}$ es σ-álgebra y $\{F \subset A : F \text{ cerrado}\} \subset \mathcal{H}$. Entonces $\mathcal{B} = \sigma(cerrados) \subset \mathcal{H}$ y por lo tanto $\forall A \in (B)(E), A = \sup_{F cerrado \subset A} \mu(F), \Rightarrow \mu$ queda caracterizada por los cerrados y entonces por $\langle \mu, f \rangle, f \in \mathcal{C}_b(E)$. □

Definición 1.2.1 Convergencia Débil
Sean $\mu \in \mathcal{M}(E)$ y $(\mu_n)_{n \in \mathbb{N}} \subset \mathcal{M}(E)$ medidas finitas mayores o iguales a 0. Decimos que (μ_n) converge débilmente a μ si

$$\langle \mu_n, f \rangle \underset{n \to \infty}{\longrightarrow} \langle \mu, f \rangle, \quad \forall f \in \mathcal{C}_b(E),.$$

Esto se denota $\mu_n \underset{n \to \infty}{\Longrightarrow} \mu$.

Observación 1.2.2. $\mathcal{P}(E) \subset \mathcal{M}(E) \subset \mathcal{M}_s(E) \subset \mathcal{C}_b(E)^*$ y $\mu_n \underset{n \to \infty}{\Longrightarrow} \mu$ equivale a $\mu_n \overset{*}{\underset{n \to \infty}{\rightharpoonup}} \mu$

La inclusión $\mathcal{M}_s(E) \subset \mathcal{C}_b(E)^*$ se demuestra como sigue: para cada $\nu \in \mathcal{M}_s(E)$, la aplicación $f \in \mathcal{C}_b(E) \mapsto \langle \nu, f \rangle := \langle \nu_+, f \rangle - \langle \nu_-, f \rangle$ es lineal. Además es continua: $|\langle \nu, f \rangle| \leq (\langle \nu_+, 1 \rangle + \langle \nu_-, 1 \rangle)\|f\|_{unif}$. Entonces $\nu \in \mathcal{C}_b(E)^*$.

Ejemplo 1.2.1
Consideremos $E = \mathbb{R}$. Tenemos:

(i) $\mu_n = \delta_{\frac{1}{n}} \overset{\Longrightarrow}{_{n\to\infty}} \delta_0$

(ii) $\mu_n(dx) = \dfrac{n}{2}\mathbf{1}_{[-\frac{1}{n},\frac{1}{n}]}(x)dx \overset{\Longrightarrow}{_{n\to\infty}} \delta_0$

(iii) $\mu_n = \dfrac{1}{n}\displaystyle\sum_{n=0}^{n-1} \delta_{\frac{k}{n}} \overset{\Longrightarrow}{_{n\to\infty}} \mu$,
donde μ es la medida de Lebesgue en $[0,1]$

Demostración. Sea $f \in C_b(E)$,

(i) $\langle \mu_n, f\rangle = f(\frac{1}{n}) \overset{\longrightarrow}{_{n\to\infty}} f(0) = \langle \delta_0, f\rangle$

(ii) Observemos que $\displaystyle\int f(x)\mu_n(dx) = \frac{n}{2}\int_{-\frac{1}{n}}^{\frac{1}{n}} f(x)dt = \frac{n}{2}\cdot\frac{2}{n}\cdot f(\xi_n)$,
donde en la última igualdad usamos el teorema del valor medio para integrales y $\xi_n \in [\frac{-1}{n}, \frac{1}{n}]$. Como f es continua, $\frac{n}{2}\cdot\frac{2}{n}\cdot f(\xi_n) \overset{\longrightarrow}{_{n\to\infty}} f(0) = \langle \delta_0, f\rangle$

(iii) Tenemos que $\displaystyle\int f(x)\mu_n(dx) = \frac{1}{n}\sum_{k=0}^{n} f\left(\frac{k}{n}\right)$. El lado derecho es una suma de Riemann con paso $\frac{1}{n}$. Como f es continua entonces es Riemann integrable y luego $\displaystyle\frac{1}{n}\sum_{k=0}^{n} f\left(\frac{k}{n}\right) \overset{\longrightarrow}{_{n\to\infty}} \int_0^1 f(x)dx$.

□

Definición 1.2.2 Convergencia en Ley
Sean $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n), (\Omega, \mathcal{F}, \mathbb{P})$ espacios de probabilidad, Sean $X_n : \Omega_n \longrightarrow E, n \in \mathbb{N}$, y $X : \Omega \longrightarrow E$ v.a., decimos que X_n converge en ley o en distribución a X si

$$\mu_n := Ley(X_n) \overset{\Longrightarrow}{_{n\to\infty}} \mu := Ley(X) \text{ en } \mathcal{P}(E)\,.$$

Equivalentemente, $\forall f \in \mathcal{C}_b(E)$

$$\langle \mu_n, f\rangle \overset{\longrightarrow}{_{n\to\infty}} \langle \mu, f, \rangle\ ,$$

o

$$\mathbb{E}_n(f(X_n)) \overset{\longrightarrow}{_{n\to\infty}} \mathbb{E}(f(X))\,.$$

Notación. La convergencia en ley se denota del siguiente modo:

$$X_n \overset{ley}{\underset{n\to\infty}{\longrightarrow}} X\,.$$

Equivalentemente podemos denotarla $X_n \overset{\mathcal{L}}{\underset{n\to\infty}{\longrightarrow}} X$ o $X_n \overset{d}{\underset{n\to\infty}{\longrightarrow}} X$.

Ejemplo 1.2.2
Gracias al Ejemplo 1.2.1 tenemos lo siguiente:

- Sean $X_n \sim \frac{1}{n}\displaystyle\sum_{k=0}^{n-1} \delta_{\frac{k}{n}}$ y $X \sim \mathbb{U}[[0,1]$, entonces $X_n \overset{ley}{\underset{n\to\infty}{\longrightarrow}} X$.

- Sean $X_n \sim \mathbb{U}([-\frac{1}{n}, \frac{1}{n}])$ y $X \sim \delta_0(X \equiv 0)$ entonces $X_n \overset{ley}{\underset{n\to\infty}{\longrightarrow}} X$.

Observación 1.2.3.

- $X_n \overset{ley}{\underset{n\to\infty}{\longrightarrow}} X \not\Longrightarrow \mathbb{E}_n(f(X_n)) \underset{n\to\infty}{\longrightarrow} \mathbb{E}(f(X))$, para toda f medible acotada.
- $X_n \overset{ley}{\underset{n\to\infty}{\longrightarrow}} X \not\Longrightarrow \mathbb{P}(X_n \leq x) \underset{n\to\infty}{\longrightarrow} \mathbb{P}(X \leq x) \forall x \in \mathbb{R}$.

Teorema 1.2.1 Portmanteau
Sean $\mu, \mu_n \in \mathcal{P}(E)$. Son equivalentes:

(i) $\mu_n \underset{n\to\infty}{\Longrightarrow} \mu$

(ii) $\langle \mu_n, f\rangle \underset{n\to\infty}{\longrightarrow} \langle \mu, f\rangle, \quad \forall f$ *acotada, uniformemente continua*

(iii) $\langle \mu_n, f\rangle \underset{n\to\infty}{\longrightarrow} \langle \mu, f\rangle, \quad \forall f \in BL(E)$ *(Lipschitz acotada)*

(iv) $\limsup_{n\to\infty} \mu_n(F) \leq \mu(F) \quad \forall F$ *cerrado*

(v) $\liminf_{n\to\infty} \mu_n(G) \geq \mu(G) \quad \forall G$ *abierto*

(vi) $\lim_{n\to\infty} \mu_n(A) = \mu(A) \quad \forall A \in \mathcal{B}$ *tal que* $\mu(\partial A) = 0$ *(frontera de medida nula)*

(vii) $\langle \mu_n, f\rangle \underset{n\to\infty}{\longrightarrow} \langle \mu, f\rangle, \quad \forall f$ *acotada, continua* $\mu(dx)$*-c.s.*

Demostración. (vi)⇒(i)⇒(ii)⇒(iii) son directas.

(iii)⇒(iv) tomamos $f_\epsilon(x) = \max(0, 1 - \frac{d(x,F)}{\epsilon})$ y notemos que $\mathbf{1}_F \leq f_\epsilon \leq \mathbf{1}_{\bar{F^\epsilon}}$. Luego tenemos

$$\limsup_n \mu_n(F) \leq \limsup_n \langle \mu_n(F), f_\epsilon\rangle = \langle \mu, f\rangle .$$

A su vez $\mathbf{1}_{\bar{F^\epsilon}} \underset{\epsilon\to 0}{\searrow} \mathbf{1}_F$, de donde se concluye que

$$\limsup_n \mu_n(F) \leq \langle \mu, \mathbf{1}_{\bar{F^\epsilon}}\rangle \underset{\epsilon\to 0}{\searrow} \mu(F) ,$$

Donde en la última convergencia usamos T.C.D.

(iv)⇒(v) Basta tomar $F = G^c$.

(v)⇒(vi) $A^o \subset A \subset \bar{A}$ y como $\partial A = 0$, $\mu(\bar{A}) = \mu(A) = \mu(A^o)$. Ahora aplicamos (iv), $\limsup \leq \liminf$ y (v) para obtener:

$$\mu(\bar{A}) \geq \overline{\lim_n} \mu_n(\bar{A}) \geq \overline{\lim_n} \mu_n(A) \geq \underline{\lim_n} \mu_n(A) \geq \underline{\lim_n} \mu_n(A^0) \geq \mu(A^o) ,$$

donde hemos usando (v) en la primera y la última desigualdad. Concluimos que $\exists \lim_{n\to\infty} \mu_n(A) = \mu(A)$.

(vi)⇒(vii) Sin perder generalidad podemos suponer que $f : E \longrightarrow [0, 1]$. De otro modo basta cambiar f por $\frac{f - \inf f}{\sup f - \inf f}$.
Notar que usando Fubini, para todo $g \geq 0$ medible y $\nu \in \mathcal{P}(E)$ tenemos

$$\langle \nu, g\rangle = \int (\int_0^\infty \mathbf{1}_{\{(t,x): t<g(x)\}} dt)\nu(dx) = \int_0^\infty \nu(\{g > t\})dt .$$

Por otro lado, sea C_f los puntos de continuidad de f, se puede probar que $C_f \cap \partial\{f > t\} \subseteq \{f = t\}$ (Ejercicio).
Además, se tiene que $\mu(\{f = t\}) = 0$ salvo para un conjunto numerable de t's. En efecto, como μ es medida finita, $\mu(\partial\{f > t\}) = \mu(C_f \cap \partial\{f > t\}) = 0 \quad dt$-c.t.p..
Entonces

$$\mu_n(f > t) \xrightarrow[n\to\infty]{} \mu(f > t) \quad dt\text{-c.t.p.}.$$

Por último, por T.C.D. concluimos que $\langle \mu_n, f\rangle = \int_0^1 \mu_n(f > t)dt \xrightarrow[n\to\infty]{} \int_0^1 \mu(f > t)dt = \langle \mu, f\rangle$.

□

Teorema 1.2.2 Teorema del mapeo

Si X_n y X son v.a., en (E, d) tal que $X_n \xrightarrow[n\to\infty]{ley} X$ y $\Phi : (E, d) \longrightarrow (E', d')$ es función continua (no necesariamente acotada) entonces

$$\Phi(X_n) \xrightarrow[n\to\infty]{ley} \Phi(X)$$

Demostración. Ejercicio

Propiedad 1.2.1 (Consecuencias de Teo. del mapeo)
Sea $(X_n, Y_n) \in \mathbb{R}^2 d, Z_n \in \mathbb{R}$ v.a.,

- $(X_n, Y_n) \xrightarrow[n\to\infty]{ley} (X, Y) \Longrightarrow X_n + Y_n \xrightarrow[n\to\infty]{ley} X + Y$
- $(X_n, Z_n) \xrightarrow[n\to\infty]{ley} (X, Z) \Longrightarrow X_n \cdot Z_n \xrightarrow[n\to\infty]{ley} XZ$
- $X_n \xrightarrow[n\to\infty]{ley} X \Longrightarrow (X_n^{i_1}, \ldots, X_n^{i_k}) \xrightarrow[n\to\infty]{ley} (X^{i_1}, \ldots, X^{i_k})$
  (pues la proyección es una función continua)

Demostración. Ejercicio

Proposición 1.2.1
Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad. Entonces:

$$\left.\begin{array}{c} X_n \xrightarrow[n\to\infty]{L^p} X \\ o \\ X_n \xrightarrow[n\to\infty]{c.s.} X \end{array}\right\} \Longrightarrow X_n \xrightarrow[n\to\infty]{\mathbb{P}} X \Longrightarrow X_n \xrightarrow[n\to\infty]{ley} X\,.$$

Es decir, la convergencia en ley es más débil que las otras convergencias usuales de v.a. definidas en un mismo espacio de probabilidad.

Demostración. Ejercicio
Indicación: usar Portmanteau con la caracterización (iii) ($f \in BL(E)$).

Observación 1.2.4. En general $X_n \overset{ley}{\underset{n\to\infty}{\longrightarrow}} X \not\Longrightarrow X_n \overset{\mathbb{P}}{\underset{n\to\infty}{\longrightarrow}} X$, sin embargo

$$X_n \overset{ley}{\underset{n\to\infty}{\longrightarrow}} A \text{ determinista} \Longrightarrow X_n \overset{\mathbb{P}}{\underset{n\to\infty}{\longrightarrow}} A \text{ determinista}$$

Demostración. Ejercicio

Teorema 1.2.3 del mapeo generalizado
Si X_n y X son v.a., en (E,d) tal que $X_n \overset{ley}{\underset{n\to\infty}{\longrightarrow}} X$ y $\Phi : (E,d) \longrightarrow (E',d')$ es función continua $\mu(dx)$-c.s. en $X \in E$ con $\mu = Ley(X)$ entonces $\Phi(X_n) \overset{ley}{\underset{n\to\infty}{\longrightarrow}} \Phi(X)$.

Demostración. Ejercicio (usar Portmanteau (vii))

Proposición 1.2.2
Sean $X, X_n, n \in \mathbb{N}$ v.a., reales, son equivalentes:

$$X_n \overset{ley}{\underset{n\to\infty}{\longrightarrow}} X \quad y \quad F_{X_n}(x) \underset{n\to\infty}{\longrightarrow} F_X(x)\,,$$

para todo x punto de continuidad de F_X.

Demostración. $\Rightarrow$ Ejercicio. Para $\Leftarrow$ ver Billingsley [1].

1.2.2. Una métrica D para la convergencia débil en M(E)

Definición 1.2.3 Norma Lipschitz
Sea $f \in BL(E)$ y $\mu \in \mathcal{M}_s(E)$, definimos la norma Lipschitz como:

$$\|f\|_{BL(E)} := \sup_x |f(x)| + \sup_{x\neq y} \frac{|f(x)-f(y)|}{d(x,y)}\,.$$

Por otro lado definimos:

$$\|\mu\|_{BL(E)^*} := \sup_{f\in BL(E), \|f\|_{BL}\leq 1} |\langle \mu, f\rangle|\,.$$

Definición 1.2.4 Distancia Lipschitz dual
Sean $\mu, \nu \in \mathcal{M}(E)$. Definimos la distancia Lipschitz dual como:

$$D(\mu,\nu) := \|\mu-\nu\|_{BL(E)^*}\,.$$

Teorema 1.2.4
Si E es separable, D metriza la convergencia débil en $\mathcal{M}(E)$, esto es, para $(\mu_n) \subseteq \mathcal{M}(E), \mu \in \mathcal{M}(E)$,

$$\mu_n \underset{n\to\infty}{\Longrightarrow} \mu \quad ssi \quad D(\mu_n,\mu) \underset{n\to\infty}{\longrightarrow} 0$$

Además, los e.m. $(\mathcal{M}(E),D)$ y $(\mathcal{P}(E),D)$ son separables y más aún, son polacos (separables completos) si E lo es.

Demostración. Ver Billingsley [1]

Observación 1.2.5.

- No es cierto en general al cambiar $\mathcal{M}(E)$ por $\mathcal{C}_b(E)^*$. (Topología débil $*$ no siempre es metrizable).
- D es estrictamente más débil que la distancia de variación total, dada por:

$$\|\mu-\nu\|_{TV} := |\mu-\nu| = \sup_{f\in\mathcal{C}_b(E),\|f\|_{unif}\leq 1} |\langle \mu-\nu, f\rangle| = \|\mu-\nu\|_{\mathcal{C}_b(E)^*}\,.$$

  En efecto, vimos que

$$\mu_n = \frac{1}{n}\sum_{k=0}^{n-1}\delta_{\frac{k}{n}} \Longrightarrow \mu = \mathbb{U}_{[0,1]}\,,$$

  pero $\|\mu_n-\mu\|_{TV}=1 \quad \forall n\in\mathbb{N}$ pues $\mu_n \perp \mu$ (i.e., son singulares una con respecto a la otra). Entonces, la distancia variación total es demasiado rígida. Por ende el uso de funciones Lipschitz acotadas.

1.2.3. Compacidad en $(M(E), D)$: Tensión

Definición 1.2.5 Tensión (*Tightness*)
Una familia $M\subset\mathcal{M}(E)$ se dice **tensa** si:

(i) $\sup_{\mu\in M}\mu(E)<\infty$

(ii) $\forall\epsilon>0, \exists K_\epsilon\subset E$ compacto tal que $\sup_{\mu\in M}\mu(K_\epsilon^c)\leq\epsilon$

Dicho de otro modo, casi toda la masa de todas las $\mu\in M$ está en un mismo compacto.

Ejemplo 1.2.3

(a) Sea $E=\mathbb{R}$, $M=(\delta_n)_{n\in\mathbb{N}}$ no es tensa.

(b) Sea $E=\mathbb{R}$, $M=\}\mu_a\}_{a\in[0,R]}$ es tensa, con $\mu_a=\mathbb{U}_{[-a,a]}$.

(c) Sea $E=\mathbb{R}^d$, $M=\{\mu\}$ con μ medida finita es tensa (μ es regular interior por compactos).
(De hecho, se prueba en Billingsley [1] que $M=\{\mu\}$ es tensa si (E,d) es polaco).

(d) Sea $E=\mathbb{R}$, $\{\mu_\sigma\}_{\sigma\in[0,L]}$ con $\mu_\sigma=\mathcal{N}(0,\sigma^2)$ es tensa

De manera más general, $\{\mu_\lambda\}_{\lambda\in\Lambda}\subset\mathcal{P}(\mathbb{R})$ tal que $\exists p>0, \sup_{\lambda\in\Lambda}\int|x|^p\mu_\lambda(dx)=c<\infty$ es tensa.
Intuición: si los momentos están acotados, no puede haber masa muy lejos.

(e) Sea E un espacio arbitrario, si $M_1, M_2, \ldots, M_m$ son m familias tensas entonces $M=\cup_{i=1}^m M_i$ es una familia tensa.

En particular, las familias finitas en $\mathcal{M}(E)$ con (E,d) polaco, son tensas.

Demostración. Debemos probar los puntos (i) y (ii) de la definición 1.2.5 (Tensión), sin embargo el punto (i) es directo cuando las medidas son de probabilidad. Por ende en general probaremos sólo el punto (ii).

(a) Siempre habrá infinitas medidas fuera del compacto que nos demos, i.e., $\forall K$ compacto, existen infinitos $n \in \mathbb{N}$ tal que $n \in K^c$, entonces $\sup_n \delta_n(K^c) = 1$. Por ende $(\delta_n)_{n\in\mathbb{N}}$ no es tensa.

(b) Sea $\epsilon > 0$, $K_\epsilon = [-R, R]$ cumple $\mu_a(K_E^c) = 0 \leq \epsilon, \forall a \in [0, R]$.

(d) Sea $\epsilon > 0$, tenemos que:

$$\begin{aligned}\mu_\lambda([-R,R]^c) &= \int \mathbf{1}_{\{|x|^p > R^p\}} \mu_\lambda(dx) \\ &\leq \frac{1}{R^p} \int |x|^p \mu_\lambda(dx) \\ &\leq \frac{c}{R^p} < \epsilon \quad \forall \lambda \in \Lambda \quad \text{si} \quad R > \sqrt[p]{\tfrac{c}{\epsilon}}\,,\end{aligned}$$

Donde usamos que $\mathbf{1}_{\{\frac{|x|^p}{R^p} > 1\}} \leq \frac{|x|^p}{R^p}$.

(e) Sea $\epsilon > 0$ y K_ϵ^i tal que $\sup_{\mu \in M^i} \mu((K_\epsilon^i)^c) \leq \epsilon$.
Definimos K_ϵ como $K_\epsilon = K_\epsilon^1 \cup \cdots \cup K_\epsilon^m$, que es compacto.
Entonces, $\forall \mu \in M$ tenemos $\mu(K_\epsilon^c) \leq \mu((K_\epsilon^i)^c) \leq \epsilon$ (con i tal que $\mu \in M^i$).

□

Teorema 1.2.5 Prokhorov
Sea $M \subseteq \mathcal{M}(E)$, con (E, d) separable.

(i) M tensa $\Longrightarrow$ M relativamente compacta

(ii) Si además E es completo, la recíproca también es cierta.

Demostración. Ver Billingsley [1]

Ejemplo 1.2.4
Sean $(X_n)_{n\in\mathbb{N}}$ v.a., en $\mathbb{R}$ tal que $\sup_n \mathbb{E}(|X_n|^2) < \infty$ (i.e., sus leyes tienen momento de orden 2 acotado uniformemente). Entonces existe subsucesión $n_k \nearrow \infty$ y X v.a., en E tal que $X_{n_k} \overset{ley}{\underset{k\to\infty}{\longrightarrow}} X$.

Demostración. En efecto, sea M como sigue

$$M := (\mu_n := Ley(X_n))_{n\in\mathbb{N}} \subseteq \mathcal{P}(\mathbb{R})\,.$$

Por ejemplo 1.2.3 tenemos que M es tensa. Entonces por Teorema 1.2.5 (Prokhorov), M es relativamente compacta (secuencialmente). Entonces existen μ en $\mathcal{P}(\mathbb{R})$ y una subsucesión creciente a infinito n_k tal que $\mu_{n_k} \underset{k\to\infty}{\Longrightarrow} \mu$
Sea $(\Omega, \mathcal{F}, \mathbb{P}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu)$ y consideremos $X : \Omega \mapsto \mathbb{R}$ la identidad ($X = Id$),
entonces $Ley(X) = \mu$, y $X_{n_k} \overset{ley}{\underset{k\to\infty}{\longrightarrow}} X$ □

1.2.4. Convergencia débil y función característica

Ahora veamos una aplicación: función característica.

Definición 1.2.6 Función Característica
Sea X v.a. en $\mathbb{R}^d$ y $\mu = Ley(X)$. Se define la función característica de X como

$$\varphi_\mu(\xi) := \mathbb{E}(e^{i\langle \xi, X\rangle}) = \int \cos\langle \xi, x\rangle \mu(dx) + i \int \text{sen}\langle \xi, x\rangle \mu(dx) \in \mathbb{C}, \quad \xi \in \mathbb{R}^d.$$

También se denota φ_X. Notar que siempre está bien definida para cualquier v.a. X y para todo $\xi \in \mathbb{R}^d$.

Propiedad 1.2.2

(i) $|\varphi_\mu(\xi)| \leq 1, \forall \xi \in \mathbb{R}^d$ y $\varphi_\mu(0) = 1$.

(ii) φ_μ es uniformemente continua.

(iii) La función característica φ_μ caracteriza las medidas finitas μ:

$$\varphi_\mu = \varphi_\nu \iff \mu = \nu.$$

Demostración.

(i) Fácil.

(ii) Sean $\xi, \zeta \in \mathbb{R}^d$, notemos que

$$\begin{aligned} |\varphi_\mu(\xi) - \varphi_\mu(\zeta)| &= |\mathbb{E}(e^{i\langle \xi, X\rangle} - e^{i\langle \zeta, X\rangle})| \\ &= |\mathbb{E}([e^{i\langle \xi-\zeta, X\rangle} - 1] \cdot e^{i\langle \zeta, X\rangle})| \\ &\leq \mathbb{E}(|e^{i\langle \xi-\zeta, X\rangle} - 1|) \cdot 1\,. \end{aligned}$$

Llamemos $u = \xi - \zeta$. Por T.C.D. $\mathbb{E}(|e^{i\langle \xi-\zeta, X\rangle} - 1|) \xrightarrow[|u|\to 0]{} 0$

(iii) Para $d = 1$ el esquema de demostración es el siguiente:
(basta probar que $\langle \mu, f\rangle = \langle \nu, f\rangle \quad \forall f \in \mathcal{C}_0(\mathbb{R})$)

- Usando Teorema de Stone-Weirstrass, se prueba para cada $L > 0$ que las combinaciones lineales de f_n son densas en $\mathcal{C}_0([-L, L])$ para la convergencia uniforme.
- Por consiguiente, también son densas en las funciones f_L L-periódicas en $\mathbb{R}$, dadas por las traslaciones de funciones $f \in \mathcal{C}_0([-L, L])$ (con respecto a la convergencia uniforme en $\mathbb{R}$)
- Para esas funciones f_L, la hipótesis $\varphi_\mu = \varphi_\nu$ implica que $\langle \mu, f_L\rangle = \langle \mu, f_L\rangle$ (probar esto queda de Ejercicio)

$$\begin{aligned} \therefore |\langle \mu, f\rangle - \langle \mu, f\rangle| &\leq |\langle \mu, f_L\rangle - \langle \mu, f_L\rangle| + |\langle \mu, f - f_L\rangle - \langle \mu, f - f_L\rangle| \\ &\leq 2 \cdot \|f\|_{Unif}(\mu([-L, L]^c) + \mu([-L, L]^c)) \\ &\leq \epsilon, \quad \text{para } L \text{ suficientemente grande}\,. \end{aligned}$$

El caso $d \geq 2$ se deja de Ejercicio

□

Teorema 1.2.6 Lévy
Sea $(\mu_n)_{n\in\mathbb{N}} \subset \mathcal{P}(\mathbb{R}^d)$,

(i) Supongamos $\exists \mu \in \mathcal{P}(\mathbb{R}^d)$ *tal que* $\mu_n \underset{n\to\infty}{\Longrightarrow} \mu$, *luego*

$$\varphi_{\mu_n}(\xi) \underset{n\to\infty}{\longrightarrow} \varphi_\mu(\xi), \forall \xi \in \mathbb{R}\,.$$

(ii) Supongamos que $\varphi_{\mu_n}(\xi) \underset{n\to\infty}{\longrightarrow} \varphi(\xi), \forall \xi \in \mathbb{R}^d$ *con* φ *continua en* 0. *Entonces,*

$$\exists \mu \in \mathcal{P}(\mathbb{R}^d) \text{ tal que } \mu_n \underset{n\to\infty}{\Longrightarrow} \mu \text{ y } \varphi = \varphi_\mu\,.$$

Este es un teorema basado en lo visto anteriormente. Para su demostración se usará lo siguiente:

Lema 1.2.1
Si $\varphi_{\mu_n}(\xi) \underset{n\to\infty}{\longrightarrow} \varphi(\xi)$, $\forall \xi$, *con* φ *continua en* 0, *entonces* $(\mu_n)_{n\in\mathbb{N}}$ *es tensa.*

Demostración.

- Para cada $i = 1, \ldots, d$, sean $(\mu_n^i)_n \subseteq \mathcal{P}(\mathbb{R})$ las leyes marginales de $(\mu_n)_{n\in\mathbb{N}}$. Es decir, si $\mu_n = Ley(X_n^1, \ldots, X_n^d)$, $\mu_n^i = Ley(X_n^i)$
- Basta probar que cada familia $(\mu_n^i)_{n\in\mathbb{N}}$ es tensa. Esto puesto que si son tensas, $\forall \epsilon > 0, \exists K_\epsilon^i \subseteq \mathbb{R}$ compacto tal que $\forall n \in \mathbb{N}$, $\mu_n^i((K_\epsilon^i)^c) \leq \epsilon$. Entonces tomando $K_\epsilon := K\epsilon^1 \times \cdots \times K_\epsilon^d$, que es un compacto en $\mathbb{R}^d$ tenemos que

  $$\mu_n(K_\epsilon^c) \leq \sum_{i=1}^{d} \mu_n^i((K_\epsilon^i)^c) \leq d\epsilon\,.$$

  $\therefore$ $(\mu_n)_{n\in\mathbb{N}}$ es tensa.
  Además, por continuidad en 0,

  $$\varphi_{\mu_n^i}(t) = \varphi_{\mu_n}(0, \ldots, 0, t, 0, \ldots, 0) \underset{n\to\infty}{\longrightarrow} \varphi(0, \ldots, 0, t, 0, \ldots, 0)\,,$$

  Con $t \in \mathbb{R}$ en la posición i-ésima. Luego basta suponer que $d = 1$ y probar el resultado en este caso.
- Sea $\delta > 0$, entonces para todo $\nu \in \mathcal{P}(\mathbb{R})$

  $$\begin{aligned}\frac{1}{\delta}\int_0^\delta [\int_{\mathbb{R}} 1 - cos(tx)\nu(dx)]dt &= \int_{\mathbb{R}} (1 - \frac{sen(x\delta)}{x\delta})\nu(dx) \\ &\geq \int_{|x\delta|>1} (1 - \frac{sen(x\delta)}{x\delta})\nu(dx) \\ &\geq c\nu(\{x : |x| > \frac{1}{\delta}\})\,.\end{aligned}$$

Donde usamos Teorema de Fubini. Luego queda que

$$\mu_n([-\frac{1}{\delta},\frac{1}{\delta}]^c) \leq \frac{c}{\delta}\int_0^\delta 1 - Re(\varphi_{\mu_n}(t))dt\,.$$

Y entonces

$$\limsup_{n\in\mathbb{N}} \mu_n([-\frac{1}{\delta},\frac{1}{\delta}]^c) \leq \frac{c}{\delta}\int_0^\delta 1 - Re(\varphi_{\mu_n}(t))dt \underset{\delta\to 0}{\longrightarrow} 0\,,$$

donde nuevamente hemos usado la continuidad de φ en 0.
De este modo, $\forall\epsilon > 0$, $\exists L = \frac{1}{\delta} > 0$ tal que

$$\limsup_{n\in\mathbb{N}} \mu_n([-L,L]^c) \leq \frac{\epsilon}{2}\,,$$

y entonces existe $n_0 \in \mathbb{N}$ tal que $\sup_{n\geq n_0} \mu_n([-L,L]^c) \leq \epsilon$

- Como cada $(\mu_0),\ldots,(\mu_{n_0-1})$ es tensa, por ejemplo 1.2.3, $\{(\mu_0),\ldots,(\mu_{n_0-1})\}$ también es una familia tensa.
  Luego, $\exists K_0 \leq \mathbb{R}$ compacto tal que $\mu_k(K_0^c) \leq \epsilon$, $\forall k = 0,\ldots,n_0-1$.
  Tomando $K_\epsilon := K_0 \cup [-1,1]$ obtenemos finalmente $\sup_{n\in\mathbb{N}} \mu_n(K_\epsilon^c) \leq \epsilon$.

□

Demostración. del Teorema 1.2.6 (Lévy)

(i) Directo, pues $x \mapsto cos(\langle\xi,x\rangle)$, $\mapsto sen(\langle\xi,x\rangle)$ son continuas acotadas.

(ii) Acá usaremos el Lema 1.2.1
En efecto, por Teorema 1.2.5 (Prokhorov), $\exists\mu \in \mathcal{P}(\mathbb{R}^d)$ y $(\mu_{n_k})_{k\in\mathbb{N}}$ subsucesión tal que $\mu_{n_k} \underset{k\to\infty}{\Longrightarrow} \mu$, y gracias a (i) tenemos que $\varphi_{\mu_{n_k}}(\xi) \underset{k\to\infty}{\longrightarrow} \varphi_\mu(\xi)$, y por hipótesis, $\varphi_{\mu_{n_k}}(\xi) \underset{k\to\infty}{\longrightarrow} \varphi(\xi)$.
Por lo tanto $\varphi = \varphi_\mu$ es una función característica.

Veamos entonces que se tiene $\mu_n \underset{n\to\infty}{\Longrightarrow} \mu$. Si suponemos lo contrario, existe una subsucesión $(\mu_{n_j})_{j\in\mathbb{N}}$ y un $\epsilon > 0$ tal que $D(\mu_{n_j},\mu) > \epsilon$.
Sin embargo, la familia $(\mu_{n_j})_{j\in\mathbb{N}}$ es también tensa y entonces existe $\nu \in \mathcal{P}(\mathbb{R}^d)$ y $(\mu_{n_{j_k}})_{k\in\mathbb{N}}$ subsucesión de la subsucesión que cumple $\mu_{n_{j_k}} \underset{k\to\infty}{\Longrightarrow} \nu$.
Luego por usando el argumento anterior, $\varphi_\nu = \varphi = \varphi_\mu$, es decir, $\mu = \nu$. Esto es una contradicción puesto que $D(\nu,\mu) \geq \epsilon > 0$.

□

1.3. Teorema central del límite en varias variables

Definición 1.3.1 Distribución Gaussiana multivariada
Decimos que Z es vector Gaussiano de media μ y varianza Γ si toda combinación de sus coordenadas es v.a. Gaussiana y $\mathbb{E}(Z) = \mu$, $Cov(Z) = \Gamma$. Esto se denota $Z \sim \mathcal{N}(\mu,\Gamma)$.

Observación 1.3.1. Si $Z \sim \mathcal{N}(\mu,\Gamma)$ entonces tenemos $\varphi_Z(\xi) = \mathbb{E}(e^{i\langle\xi,Z\rangle}) = e^{i\langle\xi,Z\rangle - \frac{\xi^T\Gamma\xi}{Z}}$.
Cuando el espacio es $\mathbb{R}$, esto se convierte en $\varphi_Z(t) = e^{ita-\frac{t^2\sigma^2}{Z}}$, con $a = \mathbb{E}(Z)$ y $\sigma = (Z)$.

Teorema 1.3.1 T.C.L. multivariado
Sean $X_1, \dots, X_n, \dots$ variables aleatorias independientes idénticamente distribuidas (i.i.d.) en L^2 ($\mathbb{E}(|X_n|^2) < \infty$). En $\mathbb{R}^d$, sean $\mu := \mathbb{E}(X_n)$ la media de las v.a. y $\Gamma = Cov(X_n)$ matriz de varianza-covarianza (i.e., $\Gamma = (Cov(X_n^i, X_n^j))_{ij}^d$).

Tomemos $\bar{X}_n := \frac{1}{n}\sum_{k=1}^{n} X_k$, entonces

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow[n\to\infty]{ley} \mathcal{N}(0, \Gamma)\,.$$

Para demostrarlo usaremos el siguiente lema:

Lema 1.3.1
Sea W v.a. real tal que $\mathbb{E}(W^2) < \infty$, entonces

$$\varphi_W(t) = 1 + it\mathbb{E}(W) - \frac{t^2}{2}\mathbb{E}(W^2) + o(t^2)\,.$$

Demostración. Notemos usando Taylor que

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

donde $R_2(ix) = \int_0^{ix} e^{it}(ix - t)^2 dt$.
Se puede demostrar (Ejercicio) que $|R_2(ix)| \le \frac{|x|^3}{6}$ y que $|R_2(ix)| \le 2|x|^2$. Entonces tenemos

$$\varphi_W(t) = \mathbb{E}(e^{itW}) = 1 + it\mathbb{E}(W) - \frac{t^2}{2}\mathbb{E}(W^2) + \mathbb{E}(R_2(itW))\,.$$

Basta ver que $\frac{1}{t^2}\mathbb{E}(R_2(itW)) \xrightarrow[t\to 0]{} 0$. Para esto usamos lo siguiente:

$$\mathbb{E}(|R_2(itW)|) \le \mathbb{E}(\min\{\frac{t^3|W|^3}{6}, 2t^2|W|^2\}) < \infty\,,$$

y como $\frac{1}{t^2}\mathbb{E}(R_2(itW)) \le \mathbb{E}(\min\{\frac{t|W|^3}{6}, 2|W|^2\})$, podemos concluir usando T.C.D. . □

Demostración de TCL multivariado 1.3.1. Denotemos $Z_n = \sqrt{n}(\bar{X}_n - \mu)$ y sea $Z \sim \mathcal{N}(0, \Gamma)$.

- Por Teorema 1.2.6 (Lévy), basta probar que $\varphi_{Z_n}(\xi) \xrightarrow[n\to\infty]{} \varphi_Z(\xi), \quad \forall \xi \in \mathbb{R}^d$.
- Primero notemos que podemos escribir $Z_n = \frac{\sum_{k=1}^n (\bar{X}_k - \mu)}{\sqrt{n}}$. Entonces

$$\varphi_{Z_n}(\xi) = \mathbb{E}(e^{\frac{i}{\sqrt{n}}\langle \xi, \sum_{k=1}^n (X_k - \mu)\rangle}) = \mathbb{E}(\Pi_{k=1}^n e^{\frac{i}{\sqrt{n}} W_k})\,,$$

donde $W_k := \langle \xi, X_k - \mu\rangle$. Luego usando la independencia de las W_k deducimos que

$$\mathbb{E}(\Pi_{k=1}^n e^{\frac{i}{\sqrt{n}} W_k}) = (\varphi_W\left(\frac{1}{\sqrt{n}}\right))^n\,.$$

- Ahora aplicaremos el Lema 1.3.1 a W_k. En efecto tomando $t = \frac{1}{\sqrt{n}}$, y usando que $Var(W) = \xi^T \Gamma \xi$, y que $\mathbb{E}(W) = 0$, nos queda que

$$\begin{aligned}\left(\varphi_W\left(\frac{1}{\sqrt{n}}\right)\right)^n &= \left(1 + 0 - \frac{1}{2n}\xi^T\Gamma\xi + o\left(\frac{1}{n}\right)\right)^n \\ &= \left(1 + \frac{1}{n}\left(\frac{-\xi^T\Gamma\xi}{2} + \frac{1}{\frac{1}{n}}o\left(\frac{1}{n}\right)\right)\right)^n \\ &\underset{n\to\infty}{\longrightarrow} e^{-\frac{\xi^T\Gamma\xi}{2}} = \varphi_Z(\xi)\,.\end{aligned}$$

  La última convergencia se tiene ya que $\frac{-\xi^T\Gamma\xi}{2} + \frac{1}{\frac{1}{n}}o(\frac{1}{n}))$ está contenido en una bola de centro $\frac{-\xi^T\Gamma\xi}{2}$ y radio ϵ. Luego para N suficientemente grande, se está suficientemente cerca de $\frac{-\xi^T\Gamma\xi}{2}$.

□

2. Métodos de Monte Carlo

Este capítulo está basado en los libros *Monte -Carlo methods in financial engineering*, capítulo 1, P. Glasserman [2] y *Processus de Markov et applications. Algorithmes, Réseaux, Génome et Finance*, capítulo 1, É. Pardoux[3].

2.0.1. Introducción

Monte Carlo es un nombre común para varios métodos, cuyo elemento en común es el uso de la aleatoriedad, con la cual típicamente se usarán para

- calcular cantidades/integrales
- simular objetos matemáticos

Ambos deterministas, en base a números/objetos aleatorios.

Los métodos de tipo Monte Carlo aparecen en variadas aplicaciones que incluyen: resolución de EDP, física, biología matemática, economía, ingeniería, finanzas, aprendizaje de máquinas, optimización, entre otras.

Origen: 1940-1950 (Fermi, Ulam, Metropolis, Von Neumann) Relacionada a los avances en física nuclear: simulación computacional aleatoria de soluciones de la ecuación de fisión.

2.1. Descripción de M.C.

Consideremos la siguiente integral:

$$I := \int_\Omega f(x)m(dx)\,,$$

con $\Omega \subset \mathbb{R}^d, m \in \mathcal{P}(\Omega)$ y $f \in L^1(m)$.

Estas integrales son **esperanzas**: $I = \mathbb{E}(f(X))$ si $X \sim m$. En particular si m tiene densidad g,

$$\mathbb{E}(f(X)) = \int_\Omega f(x)g(x)dx\,.$$

Luego *simulando* $X_1, X_2, \ldots, X_n, \ldots$ *replicas* i.i.d $\sim m$ podemos aproximar I, por **ley de grandes números**

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \overset{c.s.}{\underset{n\to\infty}{\longrightarrow}} I\,,$$

y también en L^p si $f(X_1) \in L^p, p \in [1,\infty)$.

¿Como se puede llevar a cabo esta simulación?

Los computadores pueden producir secuencias pseudoaleatorias de números $U_1, U_2, \ldots, U_n, \ldots$ que "parecen" i.i.d. $\sim \mathbb{U}([0,1])$. En realidad , son secuencias deterministas a valores en una grilla discreta muy fina, generadas con un sistema dinámico discreto con un ciclo larguísimo a partir de una *semilla* (seed) generada "aleatoriamente".

A partir de $U_1, U_2, \ldots, U_n, \ldots$ podemos generar (en teoría y en la práctica) v.a., $X_1, X_2, \ldots, X_n, \ldots$ i.i.d. en R^d de ley m cualquiera.

¿Qué tan buena es la aproximación (aleatoria) $\frac{1}{n}\sum_{k=1}^{n} f(X_k) \approx I = \mathbb{E}(Y_1)$?

Sabemos que si $Y_k := f(X_k) \in L^2$, tenemos la siguiente **cota para el error cuadrático medio**:

$$\mathbb{E}[|\frac{1}{n}\sum_{k=1}^{n} f(Y_k) - \mathbb{E}(Y_k)|^2] \leq \frac{Var(Y_1)}{n}\,.$$

En efecto,

$$\begin{aligned}\mathbb{E}(|\frac{1}{n}\sum_{k=1}^{n}(Y_k - I)|^2) &= \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{E}((Y_k - Y)^2) + \frac{1}{n^2}\sum_{j\neq k}\frac{\mathbb{E}((Y_k - I)(Y_j - I))}{\mathbb{E}(Y_k - I)\mathbb{E}(Y_j - I)} \\ &= \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{E}((Y_k - Y)^2) = \frac{Var(Y_n)}{n}\end{aligned}$$

Además por T.C.L. (1.3.1), obtenemos **intervalos de confianza asintóticos**: Sean $Y_1, Y_2, \ldots$ i.i.d. $\in L^2$ e $I = \mu = \mathbb{E}(Y_1)$:

$$\sqrt{n}(\bar{Y_n} - \mu) \overset{ley}{\underset{n\to\infty}{\longrightarrow}} \mathcal{N}(0, \sigma^2)\,,$$

con $\sigma^2 = Var(Y_1)$. Y entonces para n grande,

$$\mathbb{P}(|\bar{Y_n} - \mu| \leq \frac{\sigma}{\sqrt{n}} Z_{\frac{\alpha}{2}}) \approx 1 - \alpha\,,$$

con $\alpha \in (0,1)$ y $Z_{\frac{\alpha}{2}}$ tal que $\mathbb{P}(\mathcal{N}(0,1) > Z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$, donde hemos usado la caracterización *(vi)* del Teorema de Portmanteau (1.2.1).

Entonces **para lograr una precisión ϵ con probabilidad $\geq 1 - \alpha$**, i.e.,

$$\mathbb{P}(|\hat{Y}_n - \mu| \leq \epsilon) \geq 1 - \alpha\,,$$

tenemos que tomar n tal que:

$$\frac{\sigma Z_{\frac{\alpha}{2}}}{\sqrt{n}} \leq \epsilon \quad \Longleftrightarrow \quad n \geq \frac{\sigma^2 Z^2_{\frac{\alpha}{2}}}{\epsilon^2}\,.$$

Observación 2.1.1.

- El n requerido para lograr una precisión ϵ dada no depende de la dimensión d, contrariamente a métodos deterministas, donde aproximar $\int_{[0,1} f(x)dx$ con precisión ϵ para f Lipschitz requiere ϵ^{-d} evaluaciones, lo cual es impracticable para d no pequeño (> 3).
- El n requerido para una precisión dada será mejor (más pequeño) si σ^2 es más pequeño. Esto es importante pues si disponemos de $Y_1, \ldots, Y_n$ e $Y'_1, \ldots, Y'_n$ i.i.d. tal que $\mathbb{E}(Y_1) = \mathbb{E}(Y'_1) = \mu$ y $\sigma^2 = Var(Y_1) < \sigma^{2\prime} = Var(Y'_1)$ entonces es preferible hacer M.C. con $Y_1, \ldots, Y_n, \ldots$ para aproximar μ, siempre y cuando el costo de simular cada réplica de Y_n y de Y'_n sea similar.

- En general ¿cómo comparar entre dos M.M.C?
  Consideremos:

  - MC(1) v.a. i.i.d. $Y_1, \dots, Y_n, \dots$ con costo C por réplica
  - MC(2) v.a. i.i.d. $Y'_1, \dots, Y'_n, \dots$ con costo C' por réplica

  Para obtener una precisión $\epsilon > 0$ con probabilidad $\geq 1-\alpha$ necesitamos alguno de los siguientes:

$$n \approx \frac{\sigma^2 Z^2_{\frac{\alpha}{2}}}{\epsilon^2} \text{ con MC(1) a costo total } Cn$$

$$n' \approx \frac{\sigma^{2\prime} Z^2_{\frac{\alpha}{2}}}{\epsilon^2} \text{ con MC(2) a costo total } C'n'$$

  Luego MC(1) es preferible a MC(2) ssi:

$$Cn < C'n' \iff CVar(Y_1) < C'Var(Y'_1)\,.$$

- En general **no necesariamente conocemos** $\sigma^2 = Var(Y_1)$, que es necesario para los análisis anteriores.
  Sin embargo podemos estimarlo mediante una simulación *piloto* (más pequeña y previa a calcular (I)) usando el estimador (insesgado):

$$\sigma^2 \approx \frac{1}{q-1}\sum_{k=1}^{q}(\bar{Y}_q - Y_k)^2\,,$$

  con $q \approx 10^2$ o 10^3.

2.2. Simulación de variables aleatorias reales

Asumimos que podemos simular $(U_n)_{n\in\mathbb{N}}$ i.i.d. $\sim \mathbb{U}([0,1])$. Veremos como simular a partir de ellas realizaciones $(X_m)_{m\in\mathbb{N}}$ de v.a. de otras leyes. En teoría, basta una v.a. $\mathbb{U}([0,1])$ para simular variables aleatorias en cualquier espacio (E,d) polaco. Más aún se tiene:

Teorema 2.2.1 de Representación de Skorokhod

Sean $X_n, n \in \mathbb{N}$, X v.a. en (E,d) polaco, tal que $X_n \xrightarrow[n\to\infty]{ley} X$. Entonces $\exists Y_n : [0,1] \to E, n \in \mathbb{N}$, $Y : [0,1] \to E$ medibles tal que si $U \sim \mathbb{U}([0,1])$,

$$Y_n(U) \overset{ley}{=} X_n,\ Y(U) \overset{ley}{=} X \text{ y } Y_n(U) \xrightarrow[n\to\infty]{c.s.} Y(U)\,.$$

Demostración. En Billingsley [1]

Observación 2.2.1. Lamentablemente las funciones Y_n, Y no son construibles.
En $E = \mathbb{R}$ si se pueden definir como $Y_n = F^-_{X_n}, Y = F^-_X$, donde F^- se será enunciado más adelante en definición 2.2.4.

A continuación algunas **variables clásicas** que si se pueden simular con $\mathbb{U}([0,1])$.

2.2.1. Bernoulli, Binomial y Geométrica

Definición 2.2.1 Variable Bernoulli
Sea $p \in (0,1)$ una v.a. Bernoulli X puede realizarse como:

$$X := \mathbf{1}_{[0,p]}(U) \quad \text{con} \quad U \sim \mathbb{U}([0,1])\,.$$

Esto se denota $X \sim Ber(p)$.

Definición 2.2.2 Variable Binomial
Sea $p \in (0,1), N \in \mathbb{N}$, una variable binomial X está dada por:

$$X := \mathbf{1}_{[0,p]}(U_1) + \mathbf{1}_{[0,p]}(U_2) + \cdots + \mathbf{1}_{[0,p]}(U_N)$$

Lo denotamos $X \sim Bin(p\,; N)$

Definición 2.2.3 Variable Geométrica
Sea $p \in (0,1)$, decimos que X es una variable geométrica si está dada por:

$$X = \inf\{k \geq 1 : U_k \leq p\}\,.$$

Esto se denota $X \sim Geo(p)$

Observación 2.2.2 (Costos de simulación por réplica). Para las variables descritas anteriormente, los costos de simulación por réplica (denotado $C(\cdot)$) están dados por:

- **Bernoulli**
$$C(Ber(p)) = C(U) + evaluación$$
- **Binomial**
$$C(Bin(p,N)) = NC(Ber(p))$$
- **Geométrica**
$$C(Geo(p)) = Geo(p) \cdot C(Ber(p))$$

El costo $C(Geo(p))$ es aleatorio, y lo aproximamos usando la esperanza:

$$Geo(p) \cdot C(Ber(p)) \approx \mathbb{E}(Geo(p))C(Ber(p)) = \frac{1}{p}C(Ber(p))\,.$$

Además, para estimar el costo de simulación por réplica se puede proceder como sigue:

1. Simular $N(\approx 100, 1000)$ réplicas.
2. Contar tiempo T_N requerido.
3. Costo/réplica $\approx \dfrac{T_N}{N}$.

2.2.2. Variables reales generales

Utilizaremos el **método de la inversa generalizada**.

Definición 2.2.4 Inversa generalizada
Sea X v.a. real, F_X su función distribución. Definimos la inversa generalizada de $F_X : [0,1] \mapsto \mathbb{R}$,

$$F^-_X(t) := \inf\{x \in \mathbb{R} : F_X(x) > t\} \text{ con } t \in [0,1]\,.$$

Proposición 2.2.1
Sea X variable aleatoria real y F^-_X su inversa generalizada. Si $U \sim \mathbb{U}([0,1])$, entonces tenemos:

$$F^-_X(U) \sim Ley(X)\,.$$

Esta es la base del método de la inversa generalizada. Podemos entonces simular la ley de X usando uniformes.

Observación 2.2.3.

- F^-_X es creciente continua a la derecha
- Si F_X es creciente estricta $F^-_X = F^{-1}_X$
  Es por esto que se llama inversa generalizada, pues cuando la inversa existe entonces coincide con ella.
  Bajo la condición crecimiento estricto la Proposición 2.2.1 es directa puesto que:

$$\mathbb{P}(F^{-1}_X(U) \leq x) = \mathbb{P}(U \leq F_X(x)) = F_X(x)\,.$$

- En general, F^-_X es *inversa por la derecha*: $F_X \circ F^-_X = Id$

Demostración. Ejercicio

Demostración de Proposición 2.2.1. Sea X variable aleatoria real y F^-_X su inversa generalizada.

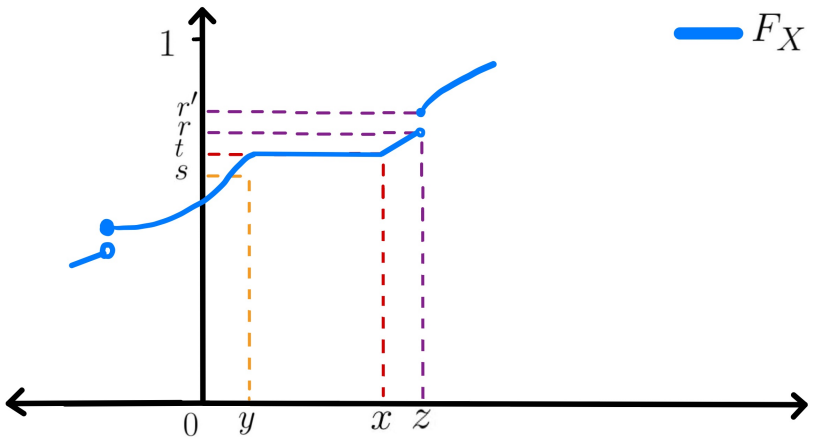
$$\begin{aligned} F^-_X \geq x &\iff \inf\{y \in \mathbb{R} : F_X(y) > U\} \geq x \\ &\iff \forall y \in \mathbb{R}, F_X(y) > U \implies y \geq x \\ &\iff \forall y < x, F_X(y) \leq U \\ &\iff F_X(x^-) \leq U \end{aligned}$$

Luego tenemos

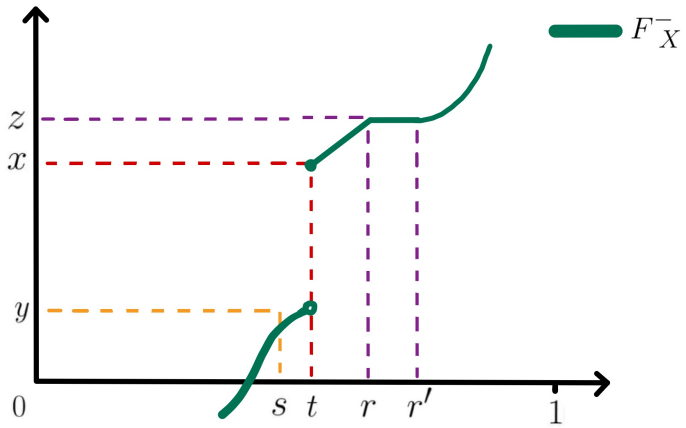
$$\begin{aligned} \mathbb{P}(F^-_X(U) \geq x) &= \mathbb{P}(F_X(x^-) \geq U) \\ &= \mathbb{P}(F_X(x^-) < U) \\ &= 1 - F_X(x^-) \\ &= 1 - \mathbb{P}(X < x) \\ &= \mathbb{P}(X \geq x) \end{aligned}$$

□

En la figura 1 mostramos un ejemplo de inversa generalizada para una función con discontinuidades. En ella tenemos que $y = F^-_X(s)$, $x = F^-_X(t)$ y $z = F^-_X(r) = F^-_X(r')$.



(a) Función F_X



(b) Función F^-_X

Figura 1: Ejemplo de inversa generalizada

Veamos como se aplica el método de la inversa generalizada a las variables aleatorias **exponencial** y **Poisson**.

Definición 2.2.5 Variable exponencial
Dado $\lambda > 0$ la v.a. X se dice exponencial si su densidad está dada por:

$$f_X(x) = \lambda e^{-\lambda x}, \forall x \geq 0\,,$$

lo cual se denota $X \sim exp(\lambda)$.

Sea la v.a. $X \sim exp(\lambda)$ con $\lambda > 0$. Notemos que su función distribución está dada por $F_X(x) = 1 - e^{-\lambda x}$. Luego por Proposición 2.2.1,

$$F^-_X(z) = \frac{-\ln(1-z)}{\lambda}\,.$$

Además, como tenemos $1 - U \overset{ley}{=} U$, se sigue el siguiente corolario:

Corolario 2.2.1 Simulación de una exponencial con inversa generalizada
Sea la v.a. $X \sim exp(\lambda)$ con $\lambda > 0$, entonces

$$X = \frac{-\ln(1-U)}{\lambda} \sim exp(\lambda)\,.$$

Definición 2.2.6 Distribución Poisson
Dado $\lambda > 0$ decimos que la v.a. N posee distribución de Poisson de parámetro λ si su función de distribución está dada por:

$$\mathbb{P}(N = n) = \frac{e^{-\lambda}\lambda^n}{n!}\,.$$

Observación 2.2.4. Recordemos que la distribución de Poisson aproxima el número de éxitos cuando repetimos muchas veces un experimento con probabilidad baja de éxito cada vez. Equivale al máximo $n \in \mathbb{N}$ tal que la suma de n exponenciales no sume más que λ.

Usando esto y el Corolario 2.2.1 tenemos lo siguiente:

Corolario 2.2.2 Simulación de una Poisson con inversa generalizada
Sean $(U_i)_{i\in\mathbb{N}}$ i.i.d. y sea $\lambda > 0$, entonces

$$N = \sup\{n \in \mathbb{N} : -\sum_{i=1}^{n} \ln(U_i) < \lambda\}$$

tiene ley de Poisson de parámetro λ.

2.2.3. Gaussianas

Consideremos dos distribuciones Gaussianas independientes (ver definición 1.3.1). Para simularlas usaremos la transformada de Box-Muller:

Proposición 2.2.2 Método de Box-Muller
Sean $U, V \sim \mathbb{U}([0,1])$, definimos:

$$R := \sqrt{-2\ln U}\,,$$

y

$$\hat{\theta} := (cos(2\pi V), sen(2\pi V))\,.$$

Luego tenemos

$$X = (X_1, X_2) := R\hat{\theta} \sim \mathcal{N}(0, I_2)\,.$$

O sea, $X_1 \perp\!\!\!\perp X_2$ y $X_1, X_2 \sim \mathcal{N}(0,1)$.

Demostración. Ejercicio. Indicación: usar teorema de cambio de variables en $\mathbb{R}^2$.

2.2.4. Variables aleatorias discretas cualquiera

Queremos simular X v.a. discreta que toma valores $x_1 < x_2 < x_3 < \cdots < x_n < \dots$ $n \in \mathbb{N}$.

Proposición 2.2.3
Sea X v.a. discreta y $p_n = \mathbb{P}(X = x_n)$ su función de probabilidad. Definimos una partición de $[0,1] : 0 = a_0 < a_1 < \cdots < a_n \leq 1$ mediante $a_n = \sum_{k \leq n} p_k$. Entonces podemos simular X usando

$$Y := \sum_{n \in \mathbb{N}} x_n \mathbf{1}_{a_{n-1}, a_n}(U) \overset{ley}{=} X\,.$$

Demostración. Basta ver que $Y = F^{-}_{X}(U)$. Ejercicio

2.2.5. Caso inversa no explícita

Sea X variable aleatoria real con densidad f_X estrictamente positiva. Entonces la función de distribución F_X es derivable y estrictamente creciente. Si **no conocemos** F_X o no la podemos invertir, usaremos el siguiente método:

Definición 2.2.7 Aproximación de Newton-Raphson
Dado $U \in [0,1]$ queremos encontrar el único $\bar{x}$ tal que $F_X(\bar{x}) = U$ (de este modo $\bar{x} = F^{-}_{X}(U)$). Es decir, buscamos la única raíz de $G(x) = F_X(x) - U$.
Luego el método de Newton-Raphson consiste en tomar:

$$x_n := x_{n-1} - \frac{G(x_{n-1})}{G'(x_{n-1})} = x_{n-1} - \frac{(F_X(x_{n-1}) - U)}{f_X(x_{n-1})} \underset{n \to \infty}{\longrightarrow} \bar{x}\,,$$

e iterar hasta que $(F_X(x_k) - U) \leq \delta$ para δ pequeño dado.

2.3. Método general

2.3.1. Aceptación-rechazo

Supongamos que queremos simular v.a. X en E espacio medible, con densidad conocida f con respecto a una medida $\lambda(dx)$ en E, y que existe Y v.a. en E con densidad g con respecto a λ tal que:

- sabemos simular fácilmente Y
- $\exists K > 0$ tal que $f(x) \leq K g(x)$ $\forall x$ tal que $g(x) > 0$

Veremos a continuación que podemos simular X usando una sucesión i.i.d. $(Y_n, U_n)_{n \geq 1}$ con $Y_n \overset{ley}{=} Y$, $U_n \sim \mathbb{U}([0,1])$ y $Y_n \perp\!\!\!\perp U_n$. La segunda condición nos dice que no es necesario que g "domine" f, sin embargo lo haría si amplificamos por una constante K.

Vamos a necesitar la función siguiente:

$$\alpha(x) := \begin{cases} \dfrac{f(x)}{K g(x)} & \text{si } g(x) > 0 \\ 0 & \text{si } g(x) \leq 0\,. \end{cases}$$

Notemos que $\alpha(x) \in [0,1]$. Además $\int f\lambda(dx) \leq K \int g\lambda(dx)$, luego $K \geq 1$.

Proposición 2.3.1 Método de aceptación-rechazo
Sea E un espacio medible cualquiera. Consideramos X v.a. en E con densidad f e Y v.a. en E con densidad g de modo que $\exists K > 0$ tal que

$$f(x) \leq Kg(x) \quad \forall x \in \{z \in E \,|\, g(z) > 0\}$$

Sea $N = \inf\{n \geq 1 : U_n \leq \alpha(Y_n)\}$, entonces tenemos:

(i) $N \sim Geom(\frac{1}{K})$

(ii) $Y_N \sim Ley(X)$

(iii) $Y_N \perp\!\!\!\perp N$

Demostración.

$$\begin{aligned}
\mathbb{P}(Y_N \in A, N = m) &= \mathbb{P}(Y_m \in A, U_m \leq \alpha(Y_m), U_k > \alpha(Y_k), k = 1, \ldots, m-1) \\
&= \mathbb{P}(Y_m \in A, U_m \leq \alpha(Y_m))(1-p)^{m-1} \\
&= \int_A \left(\int_0^{\alpha(y)} dt \right) g(y)\lambda(dy)(1-p)^{m-1} \\
&= \frac{1}{K} \int_A f(y)\lambda(dy)(1-p)^{m-1}\,,
\end{aligned}$$

donde $p = \mathbb{P}(U_1 \leq \alpha(Y_1))$. Luego tomando $A = E$ obtenemos

$$\begin{aligned}
\mathbb{P}(N = m) &= \mathbb{P}(U_m \leq \alpha(Y_m))(1-p)^{m-1} \\
&= p(1-p)^{m-1} \\
&= \frac{1}{K}(1-p)^{m-1}\,.
\end{aligned}$$

Luego $p = \mathbb{P}(U_m \leq \alpha(Y_m)) = \frac{1}{K}$, y $\therefore \quad N \sim Geom(p) = \frac{1}{K}$.
Así,

$$\begin{aligned}
\mathbb{P}(Y_N \in A, N = m) &= \int_A f(y)\lambda(dy)p(1-p)^{m-1} \\
&= \mathbb{P}(X \in A)p(1-p)^{m-1} \\
&= \mathbb{P}(X \in A)\mathbb{P}(N = m)\,.
\end{aligned}$$

Sumando sobre $m \geq 1$ nos queda que

$$\mathbb{P}(Y_N \in A) = \mathbb{P}(N = m)$$

Por ende $Y_n \overset{ley}{=} x$.
Finalmente,

$$\mathbb{P}(Y_N \in A, N = m) = \mathbb{P}(Y_N \in A)\mathbb{P}(N = m)$$

$\therefore \quad Y_N \perp\!\!\!\perp N$ □

Observación 2.3.1.

- Para obtener una realización de una v.a. de $Ley(X)$ se requiere simular una cantidad N finita, pero aleatoria $\sim Geom(\frac{1}{K})$ (desconocida a priori) de pares $(Y_1, U_1), \ldots, (Y_n, U_n), \ldots$.
  De ahí el nombre del método aceptación rechazo, pues si se cumple la condición aceptamos y si no rechazamos.
- $N \sim Geom(\frac{1}{k}) \implies \mathbb{E}(N) = K$. Luego, mientras más pequeño sea $K \geq 1$ tal que $f(x) \leq Kg(x)$, mejor, pues aceptamos el valor Y_m más pronto.
- K pequeño significa que g está más ajustada a f. Idealmente nos gustaría $K = 1$, pero en ese caso tendríamos $0 \leq g(x) - f(x) \implies 0 \leq \int |g - f| \lambda(dx) = 1 - 1 = 0 \implies g \equiv f$ o sea, $X \overset{ley}{=} Y$. Si tuviésemos esto entonces "ya sabríamos simular X" (Y es una v.a. que sabemos simular fácilmente).

2.3.2. Simulación condicional a subconjunto

Como aplicación tenemos el siguiente corolario, donde **no necesitamos simular las uniformes**:

Corolario 2.3.1 Simulación de una v.a. Y "condicional a que $Y \in A$"
Sean $(Y_n)_{n\geq 1}$ réplicas i.i.d. (simulables) de una v.a. Y en E y $A \subset E$. Sea $N := \inf\{n \geq 1 : Y_n \in A\}$, entonces

$$\hat{Y} := Y_N \sim Ley(Y|Y \in A) \quad y \quad Y_N \perp\!\!\!\perp N\,.$$

Demostración. $X \sim Ley(Y|Y \in A)$ tiene densidad f con respecto a $= Ley(Y)$, dada por

$$f(x) = K\mathbf{1}_A(x) = \mathbb{P}(Y \in A)^{-1}\mathbf{1}_A(x) \leq Kg(x)\,,$$

con $g \equiv 1$ la densidad Y con respecto a λ. Por el método aceptación-rechazo, $Y_N \sim Ley(Y|Y \in A)$, para $N = \inf\{n \geq 1 : U_n \leq \alpha(Y_n)\}$, donde $(U_n)_{n\in\mathbb{N}} \perp\!\!\!\perp (Y_n)_{n\geq 1}$ son i.i.d. y distribuyen $\mathbb{U}([0,1])$. Pero $U_n \leq \alpha(Y_n) = \mathbf{1}_A(Y_n) \iff Y_n \in A$ □

2.4. Variables condicionadas a estar en un intervalo

Proposición 2.4.1
Consideramos X una v.a. real cualquiera y $A = [a, b]$. Sea $V \sim \mathbb{U}([F_X(a), F_X(b)])$. Notemos que

$$V \overset{ley}{=} F_X(a) + U(F_X(b) - F_X(a))$$

con $U \sim \mathbb{U}([0,1])$. Entonces

$$F^-_X(V) \sim Ley(x \,|\, x \in [a,b])\,.$$

Demostración. Ejercicio

Indicación: demostrar y utilizar: $F_{\{x|x\in[a,b]\}}(x) = \dfrac{F_X(x) - F_X(a)}{F_X(b) - F_X(a)}$.

2.5. Técnicas de reducción de varianza en M.M.C

Para una presentación sucinta de estos temas ver el capítulo 1 del libro *Processus de Markov et applications. Algorithmes, Réseaux, Génome et Finance*, capítulo 1, É. Pardoux [3] y con el algo más de profundidad *Monte -Carlo methods in financial engineering* de P. Glasserman, capítulo 1. Para más detalles ver Asmussen "Stochastic Simulation" [4].

Idea: si disponemos de $Y_1, \ldots, Y_n, \ldots$ i.i.d. para aproximar $I = \mathbb{E}(Y)$, el error cometido será menor si $\sigma^2 = Var(Y)$ es menor (para igual $\mathbb{E}(Y)$).

- $\mathbb{E}(|\bar{Y}_n - I|^2) = \dfrac{Var(Y)}{n}$
- con un nivel de confianza $\alpha \in (0,1)$ dado, $\mathbb{P}(|\bar{Y}_n - I| < \epsilon) \gtrsim 1 - \alpha$ si $n \geq \dfrac{\sigma^2 Z^2_{\frac{\alpha}{2}}}{\epsilon^2}$

Luego si podemos simular $Y'_1, \ldots, Y'_n$ tal que $\mathbb{E}(Y') = I$ y $Var(Y') < Var(Y)$ (a costo comparable), necesitaremos menos "n"s y seremos más eficientes.

Los métodos de reducción de varianza que veremos son los siguientes:

1. Variable de control (2.5.1)
2. Variables antitéticas (2.5.2)
3. Muestreo preferencial (2.5.3)
4. Muestreo estratificado (2.5.4)

Notar de todos modos que para comparar dos métodos de calcular $I = \mathbb{E}(Y)$, de todas maneras hay que considerar el costo "por réplica".

2.5.1. Variable de Control

Supongamos que podemos simular $Y = f(X)$ y que conocemos $\mathbb{E}(h(X))$ explícitamente para cierta función real h, que llamaremos **variable de control**.

Disponemos entonces de muchos "estimadores" de I "posibles": para cada $c \in \mathbb{R}$,

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) + c\,[\frac{1}{n}\sum_{k=1}^{n} h(X_k) - \mathbb{E}(h(X))] = \frac{1}{n}\sum_{k=1}^{n}[f(X_k) + c(h(X_k) - \mathbb{E}(h(X))] = \frac{1}{n}\sum_{k=1}^{n}\varphi_c(X_k)\,,$$

con $\mathbb{E}(\varphi_c(X_1)) = I$. ¿Cual es su varianza?

$$\begin{aligned} Var(\varphi_c(X_1)) &= Var(f(X_1) + ch(X_1)) \\ &= Var(f(X_1)) + c^2 Var(h(X_1)) + 2\ c\ Cov(h(X_1), f(X_1))\,. \end{aligned}$$

De lo anterior se desprende que la varianza disminuye cuando se cumple:

$$c^2 Var(h(X)) + 2cCov(h(X), f(X)) < 0$$

Además podemos optimizar en $c \in \mathbb{R}$. Queda:

$$c^* = \frac{-Cov(f(X_1), h(X_1))}{Var(h(X_1))}\,.$$

Substituyendo arriba:

$$Var(\varphi_{c^*}(X_1)) = \sigma^2 - \frac{Cov(f(X_1, h(X_1)))^2}{Var(h(X_1))}\,.$$

Luego $Var(\varphi_{c^*}(X_1)) < \sigma^2$ si $Cov(f(X_1, h(X_1)) \neq 0$, i.e., reducimos la varianza del estimador.

En general, quizás no conocemos $Cov(f(X_1, h(X_1))$ ni $Var(h(X_1))$. Pero los podemos estimar con q (pequeño) simulaciones "piloto". Primero, usando la Ley de grandes números tenemos que,

$$\frac{\sum_{k=1}^{q}(Y_k - \bar{Y}_q)(Z_k - \mathbb{E}(Z_1))}{q-1} \underset{q\to\infty}{\longrightarrow} Cov(Y_1, Z_1)\,,$$

entonces

$$Cov(f(X_1, h(X_1))) \approx \frac{\sum_{k=1}^{q}(Y_k - \bar{Y}_q)(Z_k - \mathbb{E}(Z_1))}{q-1}\,,$$

con $Y_k = f(X_k), Z_k = h(X_k)$ y

$$Var(h(X_1)) \approx \frac{\sum_{k=1}^{q}(Z_k - \mathbb{E}(Z_k))^2}{q-1}\,,$$

luego,

$$\hat{c}^* = \frac{\sum_{k=1}^{q}(Y_k - \bar{Y}_q)(Z_k - \mathbb{E}(Z_1))}{\sum_{k=1}^{q}(Z_k - \mathbb{E}(Z_k))^2}\,,$$

y aproximaremos entonces I mediante el siguiente promedio empírico:

$$\frac{1}{n}\sum_{k=1}^{q}[f(X_k) - \hat{c}^*(h(X_k)) - \mathbb{E}(h(X_1))] = \frac{1}{n}\sum_{k=1}^{q}\hat{\varphi}(X_k)\,.$$

Ejemplo 2.5.1 Ejemplo "juguete" de aplicación de variable de control
Queremos calcular $I = \int_0^1 e^x dx$ (ya sabemos que esto vale $e - 1$).

Método "usual": $I = \mathbb{E}(f(X)) = \mathbb{E}(e^X), X \sim \mathbb{U}([0,1])$ donde $Var = Var(e^X) \approx 0,242$

Por otro lado, considerando $h(X_1) = \int_0^1 (x+1)dx$ como variable de control:

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

Tomando $c = -1, h(X) = x + 1$. En este caso se puede verificar que:

$$Var(e^X - 1 - x) \approx 0.00437\,.$$

Esto es 5 veces menos que $Var(e^X)$.

2.5.2. Variables antitéticas (de a pares)

Sea $I = \mathbb{E}(Y)$ con $Y = f(X)$ y X v.a. en $\mathbb{R}$. Consideremos el M.M.C usual usando $(X_n)_{n\geq 1}$ i.i.d. $\sim Ley(x)$, con $Var(f(x)) = \sigma^2$. Para un número par $2n$ de réplicas $Y_k = f(X_k)$, $\bar{Y}_{2n} = \frac{1}{2n}(Y_1 + \cdots + Y_{2n}) \underset{n\to\infty}{\longrightarrow} I$ con $ECM_{2m} = \mathbb{E}(|\bar{Y}_{2n} - I|^2) = \dfrac{\sigma^2}{2n}$ (error cuadrático medio).

Supongamos que simultáneamente podemos simular $(X'_n)_{n\geq 1} \sim Ley(X)$ de modo que X_n y X'_n están **correlacionadas**, y la sucesión de pares $(X_n, X'_n)_{n\geq 1}$ son i.i.d. (independiente de los anteriores). Entonces también:

$$\frac{1}{2}(\bar{Y}_n + \bar{Y}'_n) = \frac{1}{2n}[(Y_1 + Y'_1) + \cdots + (Y_n + Y'_n)] \underset{n\to\infty}{\longrightarrow} I\,.$$

¿Con qué ECM'_{2m}?
$\dfrac{1}{2}(\bar{Y}_n + \bar{Y}'_n) = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} \frac{(Y_k + Y'_k)}{2}$ con $\frac{Y_k+Y'_k}{2} = \frac{f(X_k)+f(X'_k)}{2}$, $k \geq 1$ i.i.d. de media $\mathbb{E}(f(X)) = I$.

$$\begin{aligned} ECM_2 n' &= \mathbb{E}\left(\left|\frac{1}{2}(\bar{Y}_n + \bar{Y}'_n - I)\right|^2\right) \\ &= \frac{1}{n} Var\left(\frac{Y_1 + Y'_1}{2}\right) \\ &= \frac{1}{4n}[2Var(Y_1) + 2Cov(Y_1, Y'_1)] \\ &= \frac{1}{2n}(\sigma^2 + Cov(f(X_1), f(X'_1))) \end{aligned}$$

Luego si $Cov(f(X_1), f(X'_1)) < 0$,

$$ECM'_{2n} < \frac{\sigma^2}{2n} = ECM_{2n}\,.$$

Aplicación usual: $X_n = U_n \sim \mathbb{U}([0,1])$, $X'_n = 1 - U_n \sim \mathbb{U}([0,1])$ (la realización U_n es la misma para X y X'), y las $2n$ variables aleatorias $U_1, \ldots, U_n$, $1 - U_1, \ldots 1 - U_n$ requieren solo n simulaciones.

¿Cuando se tiene $Cov(f(U_1), f(1 - U_1)) \leq 0$? Respondemos esto con el siguiente lema.

Lema 2.5.1
Sea $f : [0,1] \to \mathbb{R}$ función monótona tal que $f \in L^2([0,1], dx)$. Entonces $Cov(f(U), f(1-U)) \leq 0$.

Demostración. Sea $\tilde{U} \sim \mathbb{U}([0,1])$, $\tilde{U} \perp\!\!\!\perp U$. Para f creciente o decreciente siempre tenemos

$$(f(U) - f(\tilde{U}))(f(1-U) - f(1-\tilde{U})) \leq 0\,.$$

Tomando la esperanza se tiene que

$$2(\mathbb{E}(f(U)f(1-U)) - \mathbb{E}(f(U))\mathbb{E}(f(1-\tilde{U}))) \leq 0$$

$$\iff \mathbb{E}(f(U)f(1-U)) - \mathbb{E}(f(U))^2 \leq 0\,.$$

□

Receta general: Luego si tenemos $f : [0,1] \to \mathbb{R}$ monótona (por ejemplo $f = F^-_Z$ una función de distribución inversa) usando n réplicas $\mathbb{U}([0,1])$, el estimador:

$$\frac{1}{n}\sum_{k=1}^{n}[\frac{f(U_k)+f(1-U_k)}{2}]$$

aproxima $I = \mathbb{E}(f(U))$ con varianza menor a $\sigma^2/2n$, es decir, menos de la mitad de la varianza σ^2/n del estimador usual:

$$\frac{1}{n}\sum_{k=1}^{n} f(U_k)$$

construido con las mismas n réplicas.

2.5.3. Muestreo preferencial (*importance sampling*)

Queremos calcular $I = \mathbb{E}(f(X)) = \int f(x)\mu(dx)$ con $X \sim \mu$ y supongamos que podemos simular $Z \sim \nu$ simulable tal que $\nu \approx \mu$. Es decir $\exists L : \mathbb{R}^d \to \mathbb{R}_+$ densidad de Radon-Nikodym tal que $\nu(dx) = L(x)\mu(dx)$ con $L > 0$ $\mu - c.s.$,
Entonces

$$I = \int \frac{f(x)}{L(x)}\nu(dx)\,,$$

y podemos calcular I también como

$$I = \mathbb{E}\left(\frac{f(Z)}{L(Z)}\right).$$

¿Podemos encontrar ν así, de manera que además $Var(\frac{f(Z)}{L(Z)}) < Var(f(X))$?

Observación 2.5.1.

$$\begin{aligned} Var\left(\left(\frac{f(Z)}{L(Z)}\right)\right) &= \mathbb{E}\left(\frac{f(Z)^2}{L(Z)^2}\right) - I^2 \\ &= \int \frac{f(z)^2}{L(z)^2}\mu(dz) - [\int f(x)\mu(dx)]^2 \\ &= \int \frac{|f(z)|}{L(z)}\frac{|f(z)|}{L(z)}\mu(dz) - [\int f(x)\mu(dx)]^2 \\ &= \int |f(z)|\frac{|f(z)|}{L(z)}\mu(dz) - [\int f(x)\mu(dx)]^2\,. \end{aligned}$$

Si $f \geq 0$, tomando $L(z) = \dfrac{f(Z)}{I}$, o sea $\nu(dz) = \dfrac{f(Z)}{I}\mu(dz)$ y obtendríamos

$$Var\left(\frac{f(Z)}{L(Z)}\right) = 0\,,$$

pues $\frac{f(Z)}{L(Z)} = I$.

Lo anterior no se puede usar en la práctica porque **no conocíamos** I. Pero la fórmula $\nu(dz) = \frac{f(Z)}{I}\mu(dz)$ nos sugiere que sirve samplear de una ley parecida a μ (y equivalente) pero que da más peso a puntos z tales que $f(z)$ es más grande, o sea que son **más importantes** en el cálculo de $\int f(x)\mu(dx)$.

Heurística:

- Buscar una medida $\tilde{\nu}$ "parecida a $|f(x)|\mu(dx)$"
- Esta medida debe cumplir que $\nu(dx) := \dfrac{\tilde{\nu}(dx)}{\int \tilde{\nu}(dx)}$, sea "**fácilmente simulable**" donde $\int \tilde{\nu}(dx)$ es una constante de normalización calculable.

Ejemplo 2.5.2 Aplicación "artificial" de muestreo preferencial
Queremos aproximar

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

con $\mu(dx) = dx$ uniforme en $[0,1]$, $f(x) = \cos(\frac{\pi x}{2})$.

Elegimos $\nu(dz) = \frac{z}{2}(1-z^2)dz = L(z)dz$ y aproximamos

$$I \approx \frac{1}{n}\sum_{k=1}^{n} \cos\left(\frac{\pi z_k}{2}\right)/L(z_k), \quad Z_k \text{ i.i.d. } \sim \nu\,.$$

En la figura 2 se grafican ambas funciones. Se puede mostrar que la varianza se reduce en factor mayor que 10.

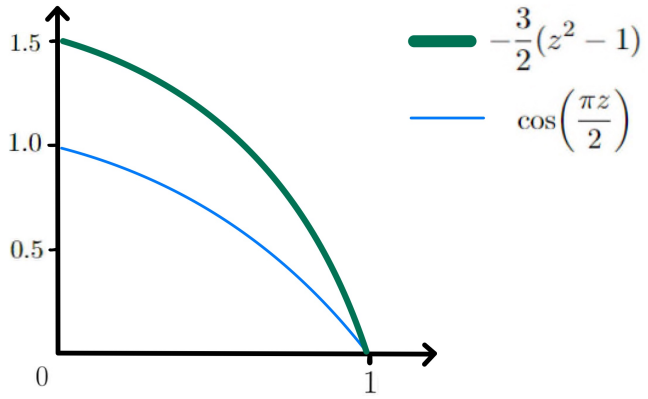


Figura 2: Ejemplo de muestreo preferencial

2.5.4. Muestreo estratificado (*stratified sampling*)

Queremos calucular $I = \mathbb{E}(Y), Y$ v.a. $\in \mathbb{R}$ con $\sigma^2 = Var(Y)$.
Supongamos que tenemos una partición $\mathcal{D} = \{D_1, \ldots, D_k\}$ ("estratos") de $\mathbb{R}$, conocemos $p_i = \mathbb{P}(Y \in D_i)$, y sabemos simular $Y^j \sim Ley(Y|Y \in D_j)$ para cada $j = 1, \ldots, k$.

Ejemplo 2.5.3
En $\mathbb{R}$, D_j de la forma $[a_j, b_j)$, vimos como simular Y^j usando F_Y^- y una v.a. $U[f_Y(a_j), f_X(b_j)]$.

Disponemos entonces de:

$$n = n_1 + \cdots + n_k \text{ v.a. independientes } : \begin{cases} Y_1^1, \ldots, Y_{n_1}^1 \text{ i.i.d.} & \overset{ley}{=} Y^1 \\ \ldots & \\ Y_1^k, \ldots, Y_{n_k}^k \text{ i.i.d.} & \overset{ley}{=} Y^k \end{cases}$$

¿Podemos estimar I con $Var < \dfrac{\sigma^2}{n} = Var\left(\dfrac{1}{n}\sum_{j=1}^{k} Y_j\right)$ con $(Y_n) \overset{ley}{=} Y$ i.i.d. ? Para responder a esto notemos que

$$\begin{aligned} \mathbb{E}(Y) &= \mathbb{E}(\mathbb{E}(Y|K)) \text{ con } K = j \text{ ssi } Y \in D_j \\ &= \mathbb{E}(\sum_{j=1}^{k} \mathbb{E}(Y|K=j)\mathbf{1}_{k=j}) \\ &= \sum_{j=1}^{k} p_j \mathbb{E}(Y^j)\,. \end{aligned}$$

Definimos

$$\hat{Y}_n = \sum_{j=1}^{k} p_j \frac{1}{n_j} \sum_{i=1}^{n_j} Y_i^j\,.$$

Observación 2.5.2.

- $\mathbb{E}(\hat{Y}_n) = I$, es decir, $\hat{Y}_n$ que es un estimador insesgado de I.
- $\hat{Y}_n \to I$ c.s. cuando $n_1, \ldots, n_k \to \infty$ con $n = n_1 + \cdots + n_k$, por Ley de Grandes Números.

¿Con qué ECM (varianza)?
Usando la independencia de las v.a. Y_k^j, tenemos:

$$Var(\hat{Y}_n) = \frac{1}{n}\sum_{j=1}^{k} \frac{p_j^2}{n_j}\sigma_j^2\,.$$

Veamos que esto es menor que σ^2/n . Usaremos la

Proposición 2.5.1 Fórmula de la "varianza total"
Sea Y v.a. real y K v.a. cualquiera, luego

$$Var(Y) = \mathbb{E}(Var(Y|K)) + Var(\mathbb{E}(Y|K))\,.$$

Donde $\mathbb{E}(Var(Y|K)) = \mathbb{E}((Y - \mathbb{E}(Y|K))^2|K)$.

Demostración. Ejercicio

Entonces aplicando lo anterior a la v.a. K definida como $K = j$ cuando $Y \in D_j$, obtenemos:

$$\sigma^2 = \sum_{j=1}^{k} p_j \sigma_j^2 + \sum_{j=1}^{k} p_j (\mathbb{E}(Y^j) - \mathbb{E}(Y))^2 \geq \sum_{j=1}^{k} p_j \sigma_j^2 .$$

Aquí, por un lado $\sum_{j=1}^{k} p_j \sigma_j^2$ corresponde a la esperanza de la varianza condicional, mientras que el término $\sum_{j=1}^{k} p_j (\mathbb{E}(Y^j) - \mathbb{E}(Y))^2$ es la varianza de la esperanza condicional (ver Figura 3).

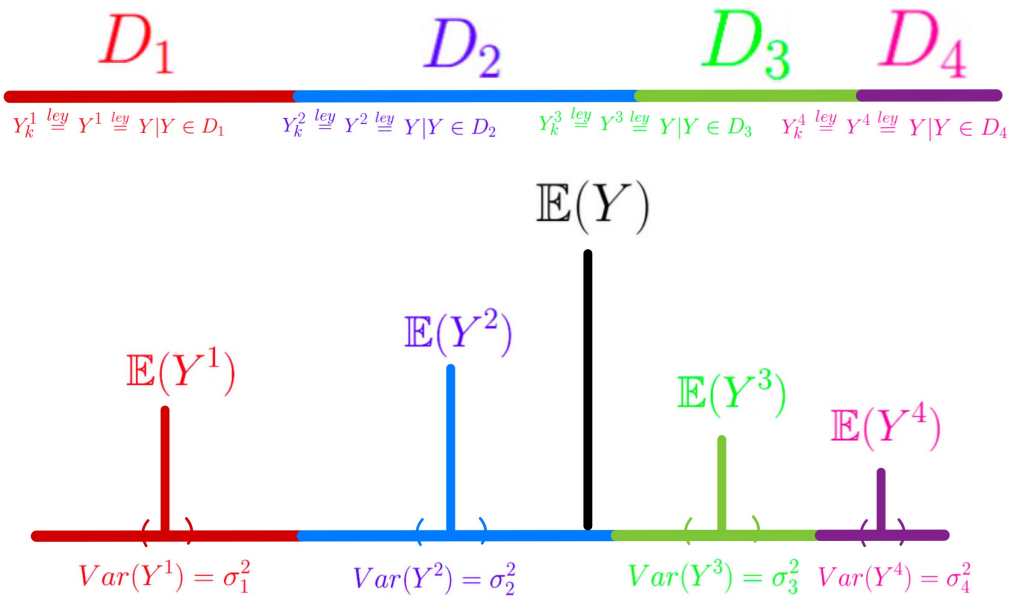


Figura 3: Ejemplo gráfico de muestreo estratificado

Notemos también que $\sum_{j=1}^{k} p_j \sigma_j^2 = Var(\hat{Y}_n)$ si elegimos $n_j \approx p_j n$. Vemos entonces que, en este caso, si $\exists j$ tal que $\mathbb{E}(Y^j) \neq \mathbb{E}(Y)$ tendremos que $Var(\bar{Y}_n) > Var(\hat{Y}_n)$, donde $\hat{Y}$ es el estimador usual. Dicho de otro modo, basta con que haya algún estrato para el cual la esperanza no sea igual a la esperanza de Y, para que se reduzca la varianza con esta elección de n_j's.

Más aún, podemos elegir los n_j's de manera aproximadamente óptima, resolviendo:

$$\min_{n_1,\ldots,n_k \geq 0} \quad \sum_{j=1}^{k} \sigma_j^2 \frac{p_j^2}{n_j}$$
$$\text{s.a.} \quad n_1 + \cdots + n_k = n$$

suponiendo los n_j reales y aplicando KKT. Obtenemos así como solución entera:

$$n_j \approx n \left(\frac{p_j \sigma_j}{\sum_{l=1}^{k} p_l \sigma_l} \right).$$

Los σ_j a su vez, se pueden estimar con simulaciones piloto.

3. Cadenas de Markov (CM)

Este capítulo contiene definiciones y resultados que ya han sido abordados en el curso Cadenas de Markov. No obstante, nos interesaremos también en simular cadenas de Markov de manera eficiente. También se propondrán algoritmos para simular la distribución invariante de una cadena.

Para revisar el tema con mayor profundidad, se puede consultar "Markov Chains" de J. Norris [5] o "Processus de Markov et applications" de E. Pardoux [3].

3.1. Recuerdo

En esta sección recordaremos las bases de Cadenas de Markov.

3.1.1. Definición

Definición 3.1.1 Cadena de Markov
Sea E un conjunto numerable, $P = (P_{xy})_{x,y\in E}$ matriz estocástica y $\lambda = (\lambda_x)_{x\in E}$ vector de probabilidad inicial.

La $(X_n)_{n\in N}$ sucesión de variables aleatorias con $X_n : (\Omega, \mathcal{F}, \mathbb{P}) \to E$ se dice **Cadena de Markov** C.M.(λ, P) (homogénea) si:

- $\forall n \in N, \forall x_0, \ldots, x_{n+1} \in E$,

$$\mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_0 = x_0) = P_{x_n, x_n+1} .$$

- $\forall x_0 \in E$, $\mathbb{P}(X_0 = x_0) = \lambda_0$.

Propiedad 3.1.1
Algunas consecuencias:

- $\mathbb{P}(X_{n+1} = y | X_n = x) = P_{xy}$.
- X es $CM(\lambda, P)$ si y solo si

$$\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) = \lambda_{x_0} P_{x_0,x_1} \ldots P_{x_{n-1}x_n} .$$

3.1.2. Definiciones y propiedades importantes

Notación. Denotaremos

- $\mathbb{P}_\mu(\cdot)$ a la ley de $(X_n)_{n\in N}$ cuando $X_0 \sim \lambda = \mu$.
- $\mathbb{P}_x(\cdot) = \mathbb{P}_{\delta_x}(\cdot)$, con δ_x masa de Dirac en $x \in E$.

Definición 3.1.2 x "pasa" a y
Sea $X = (X_n)_{n\in N}$, decimos que x "pasa" a y si

$$\mathbb{P}_x(\exists n \geq 0 \text{ tal que } X_n = y) > 0 ,$$

y se denota $x \longrightarrow y$.

Observación 3.1.1. $x \longleftrightarrow y$ es relación de equivalencia (donde $x \longleftrightarrow y$ si $x \longrightarrow y$ y $y \longrightarrow x$).

Definición 3.1.3 Cadena irreducible
Una cadena $X = CM(\lambda, P)$ se dice irreducible si E es la única clase de equivalencia para $\longleftrightarrow$.

Propiedad 3.1.2 (de Semigrupo)
Sea $X \sim CM(\lambda, P)$, entonces

$$\mathbb{P}_X(X_n = y) = (P^n)_{xy}\,.$$

Definición 3.1.4 Filtración
Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad. Sea $(\mathcal{F}_i)_{i\in\mathbb{N}}$ una familia de sub-σ-álgebras de $\mathcal{F}$.
Si $\forall i, j \in \mathbb{N}, i < j$ tenemos

$$\mathcal{F}_i \subseteq \mathcal{F}_j\,,$$

entonces $(\mathcal{F}_i)_{i\in\mathbb{N}}$ se dice una filtración.

Observación 3.1.2. Sea $(X_n)_{n\in\mathbb{N}}$ C.M. familia $(\mathcal{F}_i)_{i\in\mathbb{N}}$, dada por $\mathcal{F}_n = \sigma(X_0, \ldots, X_n)\,\forall n \in \mathbb{N}$, es una filtración.

Teorema 3.1.1 Propiedad de Markov
Sea $X \sim CM(\lambda, P)$ entonces $\forall F : E^{\mathbb{N}} \to \mathbb{R}$ medible y acotada se tiene

$$\mathbb{E}(F(X_{n+1}, X_{n+2}, \ldots)|\mathcal{F}_0, \ldots, \mathcal{F}_n) = \mathbb{E}_{X_n}(F(X_1, X_2, \ldots)) \quad c.s.\,,$$

donde $\mathcal{F}_n = \sigma(X_0, \ldots, X_n)$.

Definición 3.1.5 Tiempo de parada
Sea $(\mathcal{F}_i)_{i\in\mathbb{N}}$ una filtración y sea $\tau : \Omega \mapsto \mathbb{N} \cup \{\infty\}$ v.a.. τ se dice tiempo de parada si

$$\forall m \in \mathbb{N}, \quad \{\tau \leq m\} \in \mathcal{F}_m\,.$$

A la σ-álgebra $\mathcal{F}_\tau := \{A \in \mathcal{F} : A \cap \{\tau \leq m\} \forall m \in \mathbb{N}\}$ se le dice su tribu asociada.

Teorema 3.1.2 Propiedad de Markov Fuerte
Sea $X \sim CM(\lambda, P)$, entonces

$$\mathbb{E}(F(X_{\tau+1}, X_{\tau+2}, \ldots)|\mathcal{F}_\tau) = \mathbb{E}_{X_\tau}(F(X_1, \ldots))\ \textit{c.s. en}\ \{\tau < \infty\}\,.$$

Observación 3.1.3. El tiempo de retorno a $x \in E$ está dado por $\tau_X = \inf\{n \geq 1 : X_n = x\}$, y es un tiempo de parada.

Definición 3.1.6 Recurrencia y Transiencia
$x \in E$ se dice recurrente si:

$$\mathbb{P}_x(\tau_x < \infty) = 1\,.$$

En caso contrario se dice transiente.

Observación 3.1.4. Recurrencia es una propiedad de clase.

Proposición 3.1.1

- *Una $CM(\lambda, P)$ irreducible se dice recurrente/transiente si $x \in E$ lo es.*
- *Toda $CM(\lambda, P)$ irreducible en E finito es recurrente.*

Teorema 3.1.3
Sea $X \sim CM(\lambda, P)$ y denotemos

$$N_x := \sum_{n\in\mathbb{N}} \mathbf{1}_{\{X_n=x\}}\,,$$

*que corresponde al **número de visitas** de la cadena al estado $x \in E$.*
Entonces son equivalentes:

- *$x \in E$ recurrente.*
- *$N_x = \infty\ \mathbb{P}_X - c.s.$.*
- *$\sum_{n\in\mathbb{N}} (P^n)_{xx} = \mathbb{E}_x(N_x) = \infty$.*

Definición 3.1.7 Medida y Probabilidad Invariante
Sea $X \sim CM(\lambda, P)$, $\gamma = (\gamma_x)_{x\in E}$ con $\gamma_x \geq 0$ se dice **medida invariante** (con respecto a P) si

$$\gamma P = \gamma\,,$$

esto es:

$$\sum_{x\in E} \gamma_x P_{xy} = \gamma_y\ \forall y \in E \text{ y } \gamma \neq 0\,.$$

Una **probabilidad invariante** π es una medida invariante finita tal que $\sum_{x\in E} \pi_x = 1$.

Proposición 3.1.2
Sea π probabilidad invariante, si $X \sim CM(\lambda, P)$:

$$(X_{n+m})_{n\in\mathbb{N}} \sim CM(\pi, P)\ \forall n \in \mathbb{N}\,.$$

En particular $Ley(X_n) = \pi\ \forall n \in \mathbb{N}$.

Teorema 3.1.4
Sea X CM irreducible y recurrente, entonces existe una única γ medida invariante (salvo constante multiplicativa) tal que $\gamma_y > 0 \quad \forall y \in E$.
Más aún para cada x,

$$\gamma_y^{(x)} := \mathbb{E}_x\left(\sum_{n=1}^{\tau_x} \mathbf{1}_{\{X_n=y\}}\right), \quad y \in E\,,$$

es la única medida invariante tal que $\gamma_x^{(x)} = 1$.

Observación 3.1.5. Notemos que $\gamma_y^{(x)}$ es el número esperado de visitas a y entre dos visitas consecutivas a x .

Definición 3.1.8 Recurrencia Positiva o Nula
Denotamos $m_x = \mathbb{E}_x(\tau_x)$ como el tiempo de retorno esperado a x .
Un estado x se dice recurrente positivo si $m_x < \infty$ y recurrente nulo si $m_x = \infty$.

Observación 3.1.6. Recurrencia positiva y nula son propiedades de clase y

$$\mathbb{E}_x(\tau_x) < \infty \iff \mathbb{E}_x(\tau_y) < \infty \ \forall y \longleftrightarrow x\,.$$

Teorema 3.1.5
Sea X una CM irreducible, las siguientes son equivalentes:

- *$x \in E$ es recurrente positivo.*
- *Todo $x \in E$ es recurrente positivo.*
- *$\exists! \pi$ medida invariante para P estrictamente positiva, y está dada por:*

$$\pi = (\pi_x = m_x^{-1})_{x \in E} > 0\,.$$

Observación 3.1.7. $X \sim CM(\lambda, P)$ irreducible en E finito $\implies$ X recurrente positiva.
En la práctica, para efectos de simulaciones siempre nos restringiremos a espacios finitos. En los casos en que la cadena es recurrente positiva, esta suposición es suficientemente representativa.

Notación. $\mathbb{P}_\mu(X_n = y) = \mu P^n$

Definición 3.1.9 Aperiodicidad
X se dice aperiódica si $\forall x \in E, \exists n_0 \in \mathbb{N}$ tal que

$$(P^n)_{xx} > 0 \quad \forall n \geq n_0\,.$$

Teorema 3.1.6
Sea $X = (X_n)_{n \in \mathbb{N}}$ CM irreducible, entonces:

(a) Si X es transiente o recurrente nula, $\forall \mu, \forall y \in E$,

$$\mathbb{P}_\mu(X_n = y) \underset{n \to \infty}{\longrightarrow} 0\,.$$

(b) Si X es recurrente positivo, $\forall \mu, \forall y \in E$,

$$\frac{1}{n} \sum_{k=0}^{n} \mathbb{P}_\mu(X_k = y) \underset{n \to \infty}{\longrightarrow} \pi_y > 0\,,$$

con π distribución invariante.

(c) Si X es recurrente positiva y aperiódica, $\forall \mu, \forall y \in E$,

$$\mathbb{P}_\mu(X_n = y) \underset{n \to \infty}{\longrightarrow} \pi_y\,.$$

3.2. Simulación de cadenas de Markov

El siguiente resultado es útil para entender la construcción de cadenas de Markov. En particular será importante para su simulación.

Proposición 3.2.1
Sean X_0 v.a. en E, $(Z_n)_{n\in\mathbb{N}}$ v.a. i.i.d. a valores en un espacio medible (F,Σ) independientes de X_0. Sea $\Phi : E\times F\to E$ medible, entonces si definimos recursivamente:

$$X_{n+1} := \Phi(X_n, Z_{n+1}) \quad \forall n\in\mathbb{N}\,,$$

luego $(X_n)_{n\in N}$ es una cadena de Markov $CM(\nu,Q)$ con $\nu = Ley(X_0)$, y $Q_{xy} = \mathbb{P}(\Phi(x,Z_1)=y)$.

Demostración. Por construcción tenemos que $\nu = Ley(X_0)$.
Veamos que $(X_n)_{n\in\mathbb{N}}$ es cadena de Markov. Notemos que reemplazando, la probabilidad del cilindro $\mathbb{P}(X_{n+1}=x_{n+1}, X_n = x_n,\dots,X_0=x_0)$ puede escribirse como

$$\mathbb{P}(\Phi(x_n,Z_{n+1})=x_{n+1}, \Phi(x_{n-1},Z_n)=x_n,\dots,\Phi(x_0,Z_1)=x_1, X_0=x_0)\,.$$

Pero usando la independencia de los Z_n esto igual a

$$\mathbb{P}(\Phi(x_n,Z_{n+1})=x_{n+1})\mathbb{P}(X_n=x_n,\dots,X_0=x_0)\,,$$

que a su vez por definición corresponde a

$$Q_{x_nx_{n+1}}\mathbb{P}(X_n=x_n,\dots,X_0=x_0)\,.$$

Por otro lado, sumando sobre todos los estados posibles $x_0,\dots,x_n$ queda que

$$\mathbb{P}(X_{n+1}=x_{n+1},X_n=x_n) = \mathbb{P}(\Phi(X_n,Z_{n+1})=x_{n+1})\mathbb{P}(X_n=x_n) = Q_{x_n,x_{n+1}}\mathbb{P}(X_n=x_n)\,.$$

Despejando $Q_{x_nx_{n+1}}$ en ambas ecuaciones se obtiene la propiedad buscada. □

A partir de lo anterior deducimos como simular cadenas de markov fácilmente en un computador. En efecto tenemos el siguiente corolario:

Corolario 3.2.1 Simulación de cadenas de Markov
Sean $(U_n)_{n\in\mathbb{N}}$ i.i.d. $\sim \mathbb{U}([0,1])$, $\lambda = (\lambda_x)_{x\in E}$ vector de probabilidad y $P = (P_{xy})_{x,y\in E}$ matriz estocástica dados, con $E=\{y_0,y_1,\dots,y_n,\dots\}$. Sean

$$\Phi_0 : [0,1]\to E\,,\ y_n = \Phi_0(u) \quad si \quad u\in[\sum_{k=0}^{n-1}\lambda_{y_k}, \sum_{k=0}^{n}\lambda_{y_k}]$$

y

$$\Phi : E\times[0,1]\to E\,,\ y_n=\Phi(x,u) \quad si\in[\sum_{k=0}^{n-1}P_{xy_k}, \sum_{k=0}^{n}P_{xy_k}]\,.$$

Entonces el proceso X_0 definido por

$$X_0 := \Phi_0(U_0),\ X_{n+1} := \Phi(X_n,U_{n+1}), n\in\mathbb{N}\,,$$

es una cadena de Markov de parámetros λ, P.

Demostración. En la Proposición 3.2.1 tomar $Z_n = U_n$, $X_0=\Phi_0(U_0)$, y notar que $\mathbb{P}(X_0)=\lambda_y$ y $\mathbb{P}(\Phi(x,Z_1)=y)=P_{xy}$. □

Observación 3.2.1. Siempre podemos asumir que $E = \{y_0, \dots, y_n, \dots\} = \{0, 1, \dots, n, \dots\} = \mathbb{N}$. Luego, para simular $X_0 \sim \lambda$ y cada transición $Ley(X_{n+1}|X_n = x) = P_{x_0}$, podemos usar simulaciones de variables discretas en $\mathbb{N}$ como las vistas en sección 2.2.4.

3.3. Ley de grandes números para cadenas de Markov

Teorema 3.3.1 Ley de grandes números para CM o Teorema ergódico
Sea $X = (X_n)_{n\in N}$ cadena de markov recurrente positiva e irreducible y π su distribución invariante. Entonces $\forall \mu$ distribución inicial, $\forall f : E \to \mathbb{R}$ acotada,

$$\frac{1}{n}\sum_{k=0}^{n} f(X_k) \underset{n\to\infty}{\longrightarrow} \langle \pi, f\rangle = \sum_{y\in E} \pi_y f(y),\ \mathbb{P}_\mu - c.s..$$

En particular, $\forall y \in E$:

$$\frac{1}{n}\sum_{k=0}^{n} \mathbf{1}_{\{X_k=y\}} \underset{n\to\infty}{\longrightarrow} \pi_y = \mathbb{P}_\pi(X_0 = y)\ \mathbb{P}_\mu - c.s..$$

Como consecuencia ***podemos aproximar integrales con respecto a π usando una sola trayectoria de X.***

Demostración. Seguiremos la demostración de "Processus de Markov et applications" de E. Pardoux [3].

Estudiaremos primero el límite $\dfrac{N_x(n)}{n}$ cuando $n \to \infty$, donde

$$N_x(n) := \sum_{1\le k\le n} \mathbf{1}_{\{X_k=x\}}$$

es el número esperado de visitas al estado x antes de n.

Denotamos $S_x^0, S_x^1, \dots, S_x^k, \dots$ los largos de las sucesivas excursiones $\mathcal{E}_0, \mathcal{E}_1, \dots, \mathcal{E}_k, \dots$ de la cadena recurrente X partiendo y terminando en x. Tenemos que

$$S_x^0 + S_x^1 + \cdots + S_x^{N_x(n)-1} \le n < S_x^0 + S_x^1 + \cdots + S_x^{N_x(n)-1} + S_x^{N_x(n)},$$

con lo cual tenemos que

$$\frac{S_x^0 + S_x^1 + \cdots + S_x^{N_x(n)-1}}{N_x(n)} \le \frac{n}{N_x(n)} < \frac{S_x^0 + S_x^1 + \cdots + S_x^{N_x(n)-1} + S_x^{N_x(n)}}{N_x(n)}.$$

Usando la propiedad de Markov fuerte se demuestra que las excursiones $(\mathcal{E}_k)_{k\in\mathbb{N}}$ son i.i.d. y por ende también lo son las v.a. $(S_x^k)_{k\in\mathbb{N}}$. Además, cada S_x^k tiene la misa ley que T_x bajo $\mathbb{P}_x$, donde $T_x = \inf\{n > 0 : X_n = x\}$. Deducimos que

$$\frac{S_x^0 + S_x^1 + \cdots + S_x^{N_x(n)-1} + S_x^{N_x(n)}}{N_x(n)} \underset{n\to\infty}{\longrightarrow} \mathbb{E}_x(T_x) = m_x \quad \mathbb{P}_x - c.s..$$

Por Teorema 3.1.3, $N_x(n) \to +\infty \;\; \mathbb{P}_x - c.s.$ cuando $n \to \infty$, y entonces obtenemos

$$\frac{N_x(n)}{n} \underset{n\to\infty}{\longrightarrow} \frac{1}{m_x} \quad \mathbb{P}_x - c.s..$$

Sea ahora $F \subset E$ y denotemos $\bar{f} = \sum_{x\in E} \pi_x f(x)$. Luego, con $c > 0$ una cota para $|f|$, se tiene

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

En la 3era línea cambiamos $\sum_{x\notin F}\frac{N_x(n)}{n} = 1 - \sum_{x\in F}\frac{N_x(n)}{n} = \sum_{x\in E}\pi_x - \sum_{x\in F}\frac{N_x(n)}{n}$. Elegimos ahora F tal que $\sum_{x\notin F}\pi_x \leq \frac{\epsilon}{4c}$ y $N(\omega)$ tal que $\forall n \geq N(\omega)$

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

Entonces, $\forall n \geq N(\omega)$ tenemos

$$\left|\frac{1}{n}\sum_{k=1}^{n} f(X_k) - \bar{f}\right| \leq \epsilon$$

concluyendo la convergencia. □

3.3.1. Estimación de matriz de transición

Corolario 3.3.1 Estimación de probabilidades de transición
Sea $(X_n)_{n\in N} CM(\lambda, P)$ recurrente positiva. $\forall \mu$, $\forall x, y \in E$ tal que $P_{xy} > 0$,

$$\frac{1}{n}\sum_{k=0}^{n} \mathbf{1}_{\{X_{k+1}=y, X_k=x\}} \overset{\mathbb{P}_\mu - c.s.}{\underset{n\to\infty}{\longrightarrow}} \pi_x P_{xy}$$

y además,

$$\frac{\sum_{k=0}^{n} \mathbf{1}_{\{X_{k+1}=y, X_k=x\}}}{\sum_{k=0}^{n} \mathbf{1}_{\{X_k=x\}}} \overset{\mathbb{P}_\mu - c.s.}{\underset{n\to\infty}{\longrightarrow}} P_{xy}\,.$$

Demostración. Ejercicio:
$(\hat{X}_n)_{n\in\mathbb{N}} := (X_n, X_{n+1})_{n\in\mathbb{N}}$ es $CM(\hat{\mu}, \hat{P})$ en $\hat{E} := \{(x,y) \in E \times E : P_{xy} > 0\}$ con distribución inicial $\hat{\mu}_{(x,y)} := \mu_x P_{xy}$ y $\hat{P}_{(z,x|(\omega,y))} := \begin{cases} P_{xy} & \text{si } x = \omega\,, \\ 0 & \text{si no.} \end{cases}$

Además es irreducible y con distribución invariante $\hat{\pi}_{(x,y)} = \pi_x P_{xy}$.
Por la parte anterior, aplicada a $\hat{X} \sim CM(\hat{\mu}, \hat{P})$,

$$\frac{1}{n}\sum_{k=0}^{n} \mathbf{1}_{\{X_{k+1}=y, X_k=x\}} \overset{\mathbb{P}_\mu - c.s.}{\underset{n\to\infty}{\longrightarrow}} \pi_x P_{xy}\,,$$

y tomar cociente (dividir por la aproximación de π_x). □

3.4. Distancia de Variación total y coupling

Nos interesa cuantificar la convergencia de cadenas de Markov y en particular encontrar condiciones para convergencia "rápida" (geométrica) al equilibrio. Para ello estudiaremos una noción apropiada de distancia entre probabilidades en E.

Definición 3.4.1 Distancia de variación total
Dadas μ, $\nu \in \mathcal{E}$, con E numerable, su distancia de variación total es

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

o sea, la distancia en $l^1(E) = L^1(E, \#)$ con $\#$ la medida de conteo.
En general, en un espacio medible (E, Σ) cualquiera, la distancia en variación total está dada por:

$$\|\mu - \nu\|_1 = \int_E |\frac{d\mu}{d\lambda}(x) - \frac{d\nu}{d\lambda}|\lambda(dx)\,,$$

donde λ es cualquier medida σ-finita tal que $\mu, \nu \ll \lambda$ (ejemplo: $\lambda = \mu + \nu$), y se puede verificar que en el caso de E discreto esta noción corresponde con la antes dada.

Definición 3.4.2 Coupling
Dadas μ, $\nu \in \mathcal{P}(E)$, un coupling (o acoplamiento) entre μ y ν es un par de variables aleatorias X e $Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, \beta)^2$ definido en algún espacio $(\Omega, \mathcal{F}, \mathbb{P})$ común tal que $X \sim \mu$, $Y \sim \nu$.

Observación 3.4.1. Como el espacio de probabilidad es común, podemos samplear X e Y simultáneamente.

Lema 3.4.1 Desigualdad de Coupling
$\forall \mu, \nu \in \mathcal{P}(E)$,

$$\|\mu - \nu\|_1 = 2 \inf_{(X,Y)\ \textit{coupling de}\ \mu\ \textit{y}\ \nu} \mathbb{P}(X \neq Y)\,.$$

Además, este ínfimo se alcanza. En particular $\forall (X, Y)$ *coupling de* μ *y* ν *se tiene:*

$$\|\mu - \nu\| \leq 2\mathbb{P}(X \neq Y)\,.$$

A esta última se le llama desigualdad de coupling.

Observación 3.4.2.

- Elegir X e Y de manera inteligente nos permite tener una buena cota para la distancia de variación total.
- Siempre existen coupling. Por ejemplo tomar $X \perp\!\!\!\perp Y$, sin embargo esta no es la mejor cota que podemos encontrar.

Demostración del Lema 3.4.1. Sea (X, Y) coupling de μ, ν, entonces

$$\begin{aligned}\mathbb{P}(X = Y) &= \sum_{x\in E} \mathbb{P}(X = Y = x) \\ &\leq \sum_{x\in E} \min\{\mu_x, \nu_x\}\,.\end{aligned}$$

$$\begin{aligned}\therefore \mathbb{P}(X \neq Y) =& \geq 1 - \sum_{x\in E} \min\{\mu_x, \nu_x\} \\ &= \sum_{x\in E} (\mu_x - \min\{\mu_x, \nu_x\}) \\ &= \sum_{x\in E} (\mu_x - \nu_x)_+.\end{aligned}$$

Luego, como $|x| = |x|_+ + |-x|_+$,

$$\|\mu_x - \nu_x\|_1 = \sum_{x\in E} |\mu_x - \nu_x| \leq 2\mathbb{P}(X \neq Y)\,.$$

Ahora construiremos explícitamente un coupling "óptimo" donde el ínfimo se alcanza.
Sea $\alpha = \sum_{x\in E} \min\{\mu_x, \nu_x\} \leq 1$ y $\xi \in \mathcal{P}(E)$ dada por

$$\xi_x = \alpha^{-1} \min\{\mu_x, \nu_x\}\,.$$

Sean ξ, U, V, W v.a. independientes con leyes $\xi \sim Ber(\alpha)$, $U \sim \xi$, $V \sim \left(\bar{\mu}_x := \dfrac{(\mu_x - \nu_x)_+}{1-\alpha}\right)_{x\in E}$, $W \sim \left(\bar{\nu}_x := \dfrac{(\nu_x - \mu_x)_+}{1-\alpha}\right)_{x\in E}$. Entonces definimos

$$X := \begin{cases} U & \text{si } \xi = 1 \\ V & \text{si } \xi = 0 \end{cases}, \quad Y := \begin{cases} U & \text{si } \xi = 1 \\ W & \text{si } \xi = 0 \end{cases}$$

Veamos que (X, Y) es un coupling de μ y ν. En efecto

$$\begin{aligned}\mathbb{P}(X = x) &= \mathbb{P}(X = x|\xi = 0)\mathbb{P}(\xi = 1) + \mathbb{P}(X = x|\xi = 0)\mathbb{P}(\xi = 0) \\ &= \alpha\mathbb{P}(U = x) + (1-\alpha)\mathbb{P}(V = x) \\ &= \alpha\left(\frac{\min\{\mu_x, \nu_x\}}{\alpha}\right) + (1-\alpha)\frac{(\mu_x - \nu_x)_+}{1-\alpha} \\ &= \min\{\mu_x, \nu_x\} + (\mu_x - \nu_x)_+ = \mu_x\,.\end{aligned}$$

Por ende tenemos que $X \sim \mu$. De manera completamente análoga podemos concluir que $Y \sim \nu$. Veamos ahora que el coupling es óptimo. Notemos que $\mathbb{P}(X \neq Y) = 1 - \alpha$. En efecto,

$$\begin{aligned}\mathbb{P}(X \neq Y) &= \mathbb{P}(X \neq Y|\xi = 0)\mathbb{P}(\xi = 0) + \mathbb{P}(X \neq Y|\xi = 1)\mathbb{P}(\xi = 1)\\ &= (1 - \mathbb{P}(V = W))(1 - \alpha)\\ &= \left(1 - \sum_{x\in E} \frac{(\mu_x - \nu_x)_+(\nu_x - \mu_x)_+}{(1-\alpha)^2}\right)(1 - \alpha)\\ &= 1 - \alpha\,,\end{aligned}$$

pues $(\mu_x - \nu_x)_+(\nu_x - \mu_x)_+ = 0$. Usando esto tenemos que

$$\begin{aligned}\|\mu - \nu\|_1 &= \sum_{x\in E} |\mu_x - \nu_x|\\ &= \sum_{x\in E} [\mu_x + \nu_x - 2\min\{\mu_x, \nu_x\}]\\ &= 2 - 2\sum_{x\in E} \min\{\mu_x, \nu_x\}\\ &= 2 - 2\alpha = 2\mathbb{P}(X \neq Y)\,.\end{aligned}$$

$\therefore$ este coupling alcanza el ínfimo. □

Observación 3.4.3. La distancia en variación total se puede definir en (E, Σ) espacio medible cualquiera como

$$\|\mu - \nu\|_1 = \int_E |\frac{d\mu}{d\lambda}(x) - \frac{d\nu}{d\lambda}(x)|\lambda(dx)\,,$$

donde λ es cualquier medida σ-finita tal que $\mu, \nu << \lambda$ (por ejemplo $\lambda = \mu + \nu$).
Esta definición coincide con la del curso de Teoría de la Medida:

$$\|\mu - \nu\| = |\mu - \nu|(E)$$

donde $|\ \ |$ es la medida de variación total; y también con la definición 3.4.1 que hicimos para el caso E numerable, excepto que esta vale para E espacio medible cualquiera.

Proposición 3.4.1
Sea E numerable, se tiene:

$$\mu^n \underset{n\to\infty}{\Longrightarrow} \mu \quad \Longleftrightarrow \quad \|\mu^n - \mu\|_1 \underset{n\to\infty}{\longrightarrow} 0\,.$$

Demostración. Ejercicio Indicación: Notar que E es un espacio métrico con $d(x, y) = \mathbf{1}_{x\neq y}$, y toda función es Lipschitz. La distancia en variación total es un caso particular de una familia de métricas generales definidas sobre medidas de probabilidad, conocidas como distancia de transporte:

Proposición 3.4.2 Distancia Wasserstein
Sea (E, d) espacio métrico cualquiera, d induce una distancia W_d en $\mathcal{P}(E)$ mediante:

$$W_d(\mu, \nu) = \inf_{(X,Y)\ \textit{coupling de } \mu,\nu} \mathbb{E}(d(X, Y))\,.$$

Se le llama ***distancia de Wasserstein o de transporte****.*

3.5. Convergencia Geométrica

Definición 3.5.1 Condición de Doeblin
Decimos que se cumple la condición de Doeblin, denotada por (D), si $\exists n_0 \geq 1$, $\exists \beta > 0$, $\exists m \in \mathcal{P}(E)$ tal que

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

Observación 3.5.1.

- Si vemos $(P^{n_0})_{xy}, y \in E$ como medida de probabilidad, (D) dice que esta está minorada por una medida de probabilidad m_y que no depende de x.
- Cuando se simula una cadena de Markov, si queremos simular una transición entre x e y, la condición nos dice que a veces podremos hacerlo sampleando una variable aleatoria con ley m_y, olvidándonos de que x partimos. Esta capacidad de "olvidar el pasado" nos va a dar la convergencia al equilibrio.

Propiedad 3.5.1
Consideremos la condición de Doeblin (D), entonces:

$$\begin{aligned}(D) &\iff (\exists y \in E)(\exists n_0 \geq 1) \text{ tal que } \inf_{x\in E}(P^{n_0})_{xy} > 0\\ &\iff (\exists n_0 \geq 1) \text{ tal que } \sum_{y\in E}\inf_{x\in E}(P^{n_0})_{xy} > 0\,.\end{aligned}$$

Demostración. Ejercicio

Observación 3.5.2.

- $\beta \in [0,1]$ (sumando con respecto a y en (D)) y $\beta < 1$ a menos que $(P^{n_0})_{xy} = m_y \ \forall x, y \in E$.
- (D) es raramente satisfecha cuando $|E| = +\infty$. Típicamente sucederá que $\forall n \in \mathbb{N}$, $\forall y \in E$ $\inf_{x\in E}(P^n)_{xy} = 0$.
- P irreducible y aperiódica y $|E| < \infty$ implica que (D) se cumple.
  Ejercicio

Teorema 3.5.1
Supongamos P irreducible y que cumple (D). Entonces P es aperiódica y recurrente positiva, y si π denota su distribución invariante, se tiene

$$\|\mu^{P^n} - \pi\|_1 \leq 2(1-\beta)^{\lfloor \frac{n}{n_0} \rfloor},\ \forall \mu \in \mathcal{P}(E), \forall n \in \mathbb{N}\,,$$

donde $\beta \in (0,1)$ y $n_0 \geq 1$ vienen de (D). Es decir, la convergencia al equilibrio tiene lugar a tasa geométrica.

Demostración.

1. Probemos que $\forall \mu, \nu \in \mathcal{P}(E)$, $\forall n \in \mathbb{N}$

$$\|\mu P^n - \pi\|_1 \leq 2(1-\beta)^{\lfloor \frac{n}{n_0} \rfloor}\,.$$

Basta construir para cada n un coupling (X_n, Y_n) de μP^n y νP^n tal que

$$\mathbb{P}(X_n \neq Y_n) \leq (1-\beta)^{\lfloor \frac{n}{n_0} \rfloor}$$

y tomar después $\nu = \pi$, de modo que $\nu P^n = \pi$ para todo n. Escribimos: $n = kn_0 + j$, con $j < n_0$.

- Notar que $Q_{xy} := \dfrac{(P^{n_0})_{xy} - \beta_{m_y}}{1-\beta}$ es matriz estocástica por (D).
  Sea $f : E \times [0,1] \mapsto E$ función de transición asociada:

$$\mathbb{P}(f(x,U) = y) = Q_{xy} \text{ si } U \sim \mathbb{U}([0,1])\,.$$

- Sean $X_0 \sim \mu$, $Y_0 \sim \nu$, $\xi_l \sim Ber(\beta)$, $U_l \sim \mathbb{U}([0,1])$, $W_l \sim m$, $l = 1, \ldots, k$ independientes. Definimos recursivamente para $l = 0, \ldots, k-1$

$$X_{(l+1)n_0} = \begin{cases} W_{l+1} & \text{si } \xi_{l+1} = 1 \\ f(X_{l_{n_0}}, U_{l+1}) & \text{si no} \end{cases}$$

$$Y_{(l+1)n_0} = \begin{cases} W_{l+1} & \text{si } \xi_{l+1} = 1 \\ f(Y_{l_{n_0}}, U_{l+1}) & \text{si no}\,. \end{cases}$$

  Se puede probar que

$$(P^{n_0})_{xy} = \begin{cases} \mathbb{P}(X_{(l+1)_{n_0}} = y | X_{l_{n_0}} = x) \\ \mathbb{P}(Y_{(l+1)_{n_0}} = y | Y_{l_{n_0}} = x) \end{cases}$$

  y

$$(X_{l_{n_0}})_{l=0}^k \sim C.M.(\mu, P^{n_0})$$
$$(Y_{l_{n_0}})_{l=0}^k \sim C.M.(\nu, P^{n_0})$$

  pues ambas son construcciones recursivas con "innovaciones" independientes.
- Sea ahora $\hat{f} : E \times [0,1] \mapsto E$ función de transición asociada a P^j y $\hat{U} \sim \mathbb{U}([0,1])$, independiente a todo lo anterior. Entonces (X_n, Y_n) con

$$X_n := \hat{f}(X_{k_{n_0}}, \hat{U})$$

$$Y_n := \hat{f}(Y_{k_{n_0}}, \hat{U})$$

  es un coupling de μP^n y νP^n, y se tiene: $\{X_n \neq Y_n\} \subseteq \bigcap_{l=0}^{k} \{X_{l_{n_0}} \neq Y_{l_{n_0}}\} \subseteq \bigcap_{l=0}^{k} \{\xi_l = 0\}$,

$$\therefore \mathbb{P}(X_n \neq Y_n) \leq (1-\beta)^k = (1-\beta)^{\lfloor \frac{n}{n_0} \rfloor}$$

$$\text{y } \|\mu P^n - \nu P^n\|_1 \leq 2(1-\beta)^{\lfloor \frac{n}{n_0} \rfloor}\,,$$

  gracias al Lema 3.4.1, que es lo que queríamos.

2. Veamos que μP^n converge a algo en $l^1(E)$ (variación total)

Tomando $\nu = \mu P^m$, $m \in \mathbb{N}$ fijo, obtenemos

$$\begin{aligned} &\|\mu P^n - \mu P^{n+m}\|_1 \leq 2(1-\beta)^{\frac{n}{n_0}+1} \, \forall m \in \mathbb{N} \\ &\Longrightarrow \{\mu P^n\}_{n\in\mathbb{N}} \text{ suc. de Cauchy en } l^1(E) \\ &\Longrightarrow \exists \pi \in \mathcal{P}(E) \text{ tal que } \mu P^n \underset{n\to\infty}{\longrightarrow} \pi \,. \end{aligned}$$

3. Notar que $\|\mu P - \nu P\|_1 \leq \|\mu - \nu\|_1$, o sea $\nu \mapsto \nu P$ es Lipschitz con respecto a $\| \, \|_1$, por lo que tomando $\nu = \mu P$ tenemos

$$\|\mu P^n - \mu P^{n+1}\|_1 = \|\mu P^n - (\mu P^n)P\|_1 \underset{n\to\infty}{\longrightarrow} \|\pi - \pi P\|_1 \,.$$

Por otro lado, como μP^n es sucesión de Cauchy, $\|\mu P^n - (\mu P^n)P\|_1 \underset{n\to\infty}{\longrightarrow} = 0$, con lo cual $\|\pi - \pi P\|_1 = 0$. Sigue que π es medida de probabilidad invariante y P es recurrente positiva. Además, $(P^n)_{xx} \underset{n\to\infty}{\longrightarrow} \pi_x > 0$, luego P es aperiódica. Finalmente tomando $\nu = \pi$

$$\Longrightarrow \|\mu P^n - \pi\|_1 \leq 2(1-\beta)^{\lfloor \frac{n}{n_0} \rfloor} \, \forall n \in \mathbb{N} \,.$$

□

3.6. Teorema central del límite para CM

Definición 3.6.1 Uniforme ergodicidad
Una $CM(\lambda, P)$ (o su matriz de transición) se dice uniformemente ergódica si es irreducible, recurrente positiva y $\exists m > 0$, $\rho \in (0,1)$ tal que $\sum_{y\in\mathbb{E}} |(P^n)_{xy} - \pi_y| \leq M\rho^n$, $\forall n \in \mathbb{N}$, $\forall x \in E$.

Teorema 3.6.1 T.C.L para cadenas de Markov
Sea $(X_n)_{n\in\mathbb{N}}$ irreducible, aperiódica y uniformemente ergódica. Sea π su ley invariante y $f \in L^2(E,\pi)$, es decir, $\sum_{x\in E} f^2(x)\pi_x < \infty$. Entonces

$$\sqrt{n}\left(\frac{\sum_{k=0}^n f(X_k) - \langle \pi, f\rangle)}{n}\right) \underset{n\to\infty}{\overset{ley}{\longrightarrow}} \mathcal{N}(0, \sigma_f^2) \,,$$

con $\sigma_f^2 = \langle \pi, (Qf)^2\rangle - \langle \pi, (PQf)^2\rangle$ donde $(Qf)_x = \sum_{n=0}^{\infty} \mathbb{E}_x(f(X_n))$, $x \in E$.

Notar que $\sum_{n=0}^{\infty}$ converge gracias a la uniforme ergodicidad.

Demostración. En Pardoux [3].

Corolario 3.6.1
Sea $(X_n)_{n\in N}$ una cadena de Markov irreducible que satisface (D), en particular si E es finito, satisface el TCL (3.6.1).

3.7. Simulación exacta de una ley invariante

Consideremos $(X_n)_{n\in\mathbb{N}} \sim CM(\lambda, P)$ irreducible, recurrente positiva y aperiódica. Sabemos que:

$$X_n \overset{ley}{\underset{n\to\infty}{\longrightarrow}} X_\infty \sim \pi$$

Sin embargo esto sigue siendo una aproximación. Sin embargo nos gustaría simular una X_∞ que distribuya π de manera **exacta**, no aproximada. Se darán dos condiciones y algoritmos para aquello.

3.7.1. Algoritmo simulación perfecta

Supongamos (D) con $n_0 = 1$ y sea $\beta = \sum_{y\in E} \inf_{x\in E} P_{xy} \in (0,1)$ y consideremos la siguiente elección de m:

- $m \in \mathcal{P}(E)$, con $m_y := \dfrac{\inf_{x\in E} P_{xy}}{\beta}$, $y \in E$.
- $\left(Q_{xy} = \dfrac{P_{xy} - \beta m_y}{1-\beta}\right)_{x,y\in E}$ matriz estocástica.

Notemos que si simulamos $\xi \sim Ber(\beta)$, $Z \sim m$, $Y \sim Q_{X_0}$ independientes, entonces X dada por

$$X := \begin{cases} Z & \text{si } \xi = 1 \\ Y & \text{si } \xi = 0 \end{cases}$$

tiene ley P_{x_0}. En efecto:

$$\begin{aligned} \mathbb{P}(X = z) &= \mathbb{P}(X = z|\xi = 1)\mathbb{P}(\xi = 1) + \mathbb{P}(X = z|\xi = 0)\mathbb{P}(\xi = 0) \\ &= \mathbb{P}(Z = z)\beta + \mathbb{P}(Y = z)(1-\beta) \\ &= \beta m_Z + (1-\beta)Q_{xz} = P_{xz}\,, \end{aligned}$$

i.e., obtenemos P_{xz} por construcción.

Entonces construimos una función de transición $f : E \times [0,1] \to E$ tal que si $U \sim \mathbb{U}([0,1])$,

$$\mathbb{P}(f(x,U) = y|U \le \beta) = m_y$$

$$\mathbb{P}(f(x,U) = y|U > \beta) = Q_{xy}\,.$$

Basta tomar:

- $f_1 : E \times [0,1] \to E$ función de transición asociada a $Q : \mathbb{P}(f_1(x,U) = y) = Q_{xy}$.
- $f_0 : [0,1] \to E$ función para simular $m : \mathbb{P}(f_0(U) = y) = m_y$.

Y definimos:

$$f(x,u) = f_0\left(\frac{u}{\beta}\right)\mathbf{1}_{u\le\beta} + f_1\left(x, \frac{u-\beta}{1-\beta}\right)\mathbf{1}_{u>\beta}\,.$$

Esta función cumple lo requerido (Ejercicio) y además

$$\forall x \in E, \quad \mathbb{P}(f(x,U) = y) = P_{xy}\,.$$

Teorema 3.7.1 Algoritmo de simulación perfecta

(i) *Sean* $U_0, U_{-1}, U_{-2}, \ldots,$ *i.i.d.* $\sim \mathbb{U}([0,1])$ *y* $\tau = \max\{k \leq 0 : U_k < \beta\}$ *.*

(ii) *Sea* $X_\tau := f(\bar{x}, U_\tau)$ *con* $\bar{x}$ *fijo cualquiera.*

(iii) *Para* $k = \tau+1, \tau+2, \ldots, -1, 0$ *sean* $X_k = f(X_{k-1}, U_k)$ *donde las* U_k *son las mismas uniformes de antes.*

Entonces $-(\tau-1) \sim Geo(\beta)$*,* $X_\tau \sim m$*,* $X_\tau \perp\!\!\!\perp -(\tau-1)$ *y*

$$X_0 \sim \pi\,,$$

donde π *es invariante para* P *.*

Demostración. Sean $k \in \mathbb{N}$, $x_0, x_1, \ldots, x_k \in E$,

$$\begin{aligned}
&\mathbb{P}(\tau = -k, X_{-k} = x_k, X_{-(k-1)} = x_{k-1}, \ldots, X_{-1} = x_{-1}, X_0 = x_0) \\
&= \mathbb{P}[f(\bar{x}, U_{-k}) = x_k, U_{-k} < \beta, f(x_k, U_{-(k-1)}) = x_{k-1}, U_{-(k-1)} \geq \beta, \ldots, f(x_1, U_0) = x_0, U_0 \geq \beta] \\
&= m_{x_k}\beta \cdot Q_{x_k x_{k-1}}(1-\beta)\ldots Q_{x_1 x_0}(1-\beta)\,,
\end{aligned}$$

donde usamos que las uniformes son independientes. Por otro lado, sumando, para $x_0, \ldots, x_k$ se verifica que

$$\mathbb{P}(-(\tau+1) = j, X\tau = x) = m_x\beta(1-\beta)^{j-1}\,,$$

que es la ley geométrica buscada. Veamos ahora que $\mathbb{P}(X_0 = x)$ es invariante.
Denotando $\mu_x = \mathbb{P}(X_0 = x)$, tenemos entonces que

$$\begin{aligned}
\mu_x &= \sum_{k=0}^{\infty} \sum_{x_1,\ldots,x_k \in E} \mathbb{P}(\tau = -k, X_{-k} = x_k, \ldots, X_{-1} = x_1, X_0 = x) \\
&= \sum_{k=0}^{\infty} (mQ^k)_x \beta(1-\beta)^k \\
&= \beta \sum_{k=0}^{\infty} (mQ^{k+1})_x (1-\beta)^{k+1} + \beta m_x \\
&= \sum_{y \in E} \beta \sum_{k=0}^{\infty} (mQ^k)_y (1-\beta)^k [(1-\beta)Q_{yx} + \beta m_x] \\
&= \sum_{y \in E} \mathbb{P}(x_0 = y) P_{yx} = \sum_{y \in E} \mu_y P_{yx}\,,
\end{aligned}$$

gracias a que $P_{xy} = (1-\beta)Q_{yx} + \beta m_x$ y que $\mathbb{P}(x_0 = y) = \beta \sum_{k=0}^{\infty} (mQ^k)_y (1-\beta)^k$.
Como $\mu_x = \sum \mu_y P_{yx}$ tenemos que μ_x es medida invariante. Como la medida invariante es única concluimos que $\mu_x = \pi$. □

3.7.2. Coupling from the past

Sea P finita irreducible (y entonces recurrente positiva) en un espacio $E = \{1, \dots, N\}$. Sea $f : E \times [0,1] \mapsto E$ la función de transición asociada. Sean $(U_n : n \in \mathbb{Z}, y \in E) \sim \mathbb{U}([0,1])$ i.i.d. que simulamos una sola vez. Entonces, tenemos funciones aleatorias independientes

$$\begin{aligned} \Phi_n^y : E &\longmapsto E \\ x &\longmapsto \Phi_n(x) = f(x, U_n)\,, \end{aligned}$$

con $n \in \mathbb{Z}, y \in E$. Las usaremos para construir, para cada $n \in \mathbb{N}$, N cadenas de Markov $(X_{n,m}^y)_{m \geq n}$, $y = 1, \dots, N$, partiendo en tiempo n de los N estados distintos, pero **coalescentes**. Esto es:

- Para cada $y = 1, \dots, N$, $X_{n,m}^y$ parte de y y evoluciona usando el flujo $\Phi_m \circ \cdots \circ \Phi_n : E \longmapsto E$, es decir $X_{n,m}^y = \Phi_m \circ \dots \Phi_n(y)$.
- Si $\exists m > n$ tal que $X_{n,m}^y = X_{n,m}^z = x$, en ese momento coalescen, pues ambas evolucionan desde ahí como $\Phi_{m+k} \circ \cdots \circ \Phi_m(x)$, i.e., siguen igual desde aquel punto.
- Para cada $n \in -\mathbb{N}$, sea $S_n := \inf\{m > n : X_{n,m}^1 = \dots X_{n,m}^N\}$ el primer instante en que coalescen todas las cadenas iniciadas a tiempo n. Notar que S_n puede ser infinito, pues puede ser que en ningún momento se junten todas. En la figura 4 visualizamos como van coalesciendo las cadenas.
- Sea $\tau_0 := \sup\{n \leq 0 : S_n \leq 0\}$ el último $n \leq 0$ tal que las cadenas iniciadas en n coalescen antes de $m = 0$. τ_0 representa aquel tiempo "menos en el pasado" tal que al lanzar las cadenas en ese momento hayan coalescido en 0.
- En general, se necesitan condiciones para asegurar que $\tau_0 > -\infty$.

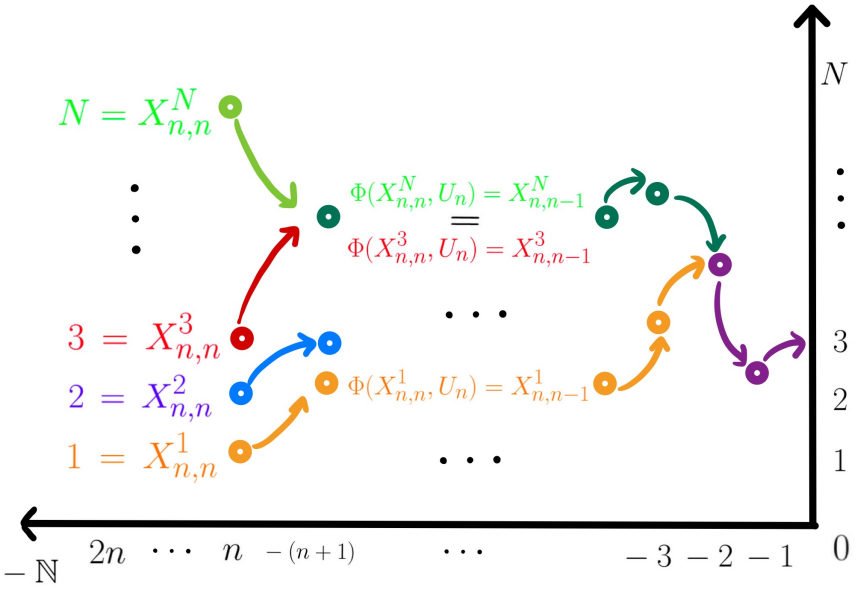


Figura 4: Ejemplo de $(X_{n,m}^y)_{m \geq n}$ que coalescen.

Teorema 3.7.2
Si $\tau_o > -\infty$ c.s., se tiene

$$X^y_{\tau_0,0} \sim \pi \quad \forall y \in E\,.$$

Demostración. $\forall k \in -\mathbb{N}$, $x, y \in E$

$$\mathbb{P}(X^y_{\tau_0,0} = x, \tau_0 > k) = \mathbb{P}(X^y_{k,0} = x, \tau_0 > k)\,.$$

Luego $\mathbb{P}(X^y_{\tau_0,0} = x) = \lim_{k\to-\infty} \mathbb{P}(X^y_{k,0} = x)$. Por otro lado

$$\begin{aligned}|\mathbb{P}(X^y_{k,0} = x) - \pi_x| &= |\mathbb{P}(X_0 = x|X_k = y) - \sum_z \pi_z \mathbb{P}(X_0 = x|X_n = z)| \\ &\le \sum_z \pi_z |\mathbb{P}(X_0 = x|X_k = y) - \mathbb{P}(X_0 = x|X_k = z)|\,.\end{aligned}$$

En general, si (V, W) es un coupling de dos leyes μ y ν,

$$\begin{aligned}\mu_y = \mathbb{P}(V = y) &= \mathbb{P}(V = y, W = y) + \mathbb{P}(V = y, W \neq y) \\ &\le \mathbb{P}(W = y) + \mathbb{P}(V \neq W) \\ &= \nu_y + \mathbb{P}(V \neq W)\,. \\ \therefore |\mu_y - \nu_y| &\le \mathbb{P}(V \neq W) \\ \therefore |\mathbb{P}(X^y_{k,0} = x) - \pi_x| &\le \sum_z \pi_z \mathbb{P}(X^y_{k,0} \neq X^z_{k,0}) \\ &\le \mathbb{P}(\tau_0 < k) \underset{k\to-\infty}{\longrightarrow} 0\,.\end{aligned}$$

Entonces,

$$\mathbb{P}(X^y_{\tau,0} = x) = \underset{k\to-\infty}{\longrightarrow} \mathbb{P}(X^y_{k,0} = x) = \pi_x\,.$$

□

3.7.3. Criterio Foster-Lyapunov para convergencia geométrica

Teorema 3.7.3 de Harris
Sea $(X_n)_{n\in\mathbb{N}}$ cadena de Markov en E con matriz P irreducible tal que

- *$\exists K \subseteq E$, $\exists \beta > 0$, $m \in \mathcal{P}(E)$, $n_0 \in \mathbb{N}$ tal que*

  $$(P^{n_0})_{xy} \ge \beta_{my} \quad \forall x \in K, \forall y \in E\,.$$

  Esto es, una condición de tipo Doeblin (D) en K.

- *$\exists V : E \longmapsto [1, \infty)$, $\rho \in (0, 1)$, $c > 0$ tal que*

  $$PV(x) \le \rho V(x) + c\mathbf{1}_K(x) \quad \forall x \in E\,.$$

  (Esta condición de "tipo Lyapunov" fuera de K nos dice que V tiende a decrecer en promedio.)

Entonces, $(X_n)_{n\in\mathbb{N}}$ *es recurrente positiva, y* $\exists\theta\in(0,1)$*,* $M>0$ *tal que* $\forall x\in E$

$$\|P^n_{x_0}-\pi\|_1\leq M\theta^n\,,$$

i.e., es uniformemente ergódica.

Probaremos sólo un resultado intermedio:

Lema 3.7.1
Sea $\tau_K=\inf\{n\geq 1: X_n\in K\}$ *tiempo de parada. Entonces* $\forall x\notin K$*,*

$$\mathbb{E}(\rho^{-\tau_K})<\infty\,.$$

Observación 3.7.1. Luego τ_K tiene un momento exponencial ($\rho^{-1}>1$), y $\forall n$

$$\mathbb{P}_x(\tau_K>n)=\mathbb{E}_X(\mathbf{1}_{\tau_k>n})=\mathbb{E}_X(\mathbf{1}_{\rho^{-\tau_X}>\rho^{-n}})\leq\rho^n\mathbb{E}_X(\rho^{-\tau_k})\,,$$

i.e., τ_k tiene "cola geométrica" (y entonces toma valores grandes con baja probabilidad).

Demostración. Probaremos que $Y_n:=\rho^{-\min\{n,\tau_K\}}V(X_{\min\{n,\tau_K\}})$ es una sobre-martingala en la filtración $\mathcal{F}_n:=\sigma(X_0,\ldots,X_n)$, esto es. $\mathbb{E}_x(Y_{n+1}|\mathcal{F}_n)\leq Y_n$. Como las esperanzas decrecen, $\mathbb{E}(Y_n)\leq E_x(Y_0)$, y dado que $V(\cdot)\geq 1$, se obtendrá entonces que

$$\mathbb{E}_x(\rho^{\min\{n,\tau_K\}})\leq\mathbb{E}_X(\rho^{\min\{n,\tau_K\}V(X_{\min\{n,\tau_K\}})})\leq V(x)<\infty\,.$$

Luego, tomando $n\to\infty$, por T.C.M. se concluye que $\mathbb{E}_x(\rho^{-\tau_K})<\infty$. Estudiemos entonces

$$\mathbb{E}_x(Y_{n+1}|\mathcal{F}_n)=\mathbb{E}_x(\rho^{-(n+1)}V(X_{n+1})\mathbf{1}_{\tau_K>n}|\mathcal{F}_n)+\mathbb{E}_x(\rho^{-\tau_K}V(X_{\tau_K})\mathbf{1}_{\tau_K\leq n}|\mathcal{F}_n)\,.$$

El primer término queda

$$\begin{aligned}\mathbb{E}_x(\rho^{-(n+1)}V(X_{n+1})\mathbf{1}_{\tau_K>n}|\mathcal{F}_n)&=\rho^{-(n+1)}\mathbb{E}(V(X_{n+1})|\mathcal{F}_n)\mathbf{1}_{\tau_K>n}\\&=\rho^{-(n+1)}PV(X_n)\mathbf{1}_{\tau_K>n}\\&\leq\rho^{-n}V(X_n)\mathbf{1}_{\tau_K>n},\end{aligned}$$

donde usamos la propiedad de Markov y que $X_n\notin K$ en $\{\tau_K>n\}$ (entonces $c\mathbf{1}_K(x)=0$). Por otro lado en el segundo término podemos sacar la $\mathbb{E}(\cdot|\mathcal{F}_n)$ pues $\mathcal{F}(X_{\tau_K})\mathbf{1}_{\tau_K\leq n}$ es $\mathcal{F}_n$-medible. Luego

$$\begin{aligned}\mathbb{E}_x(Y_{n+1}|\mathcal{F}_n)&\leq\rho^{-n}V(X_n)\mathbf{1}_{\tau_K>n}+\rho^{-\tau_K}V(X_{\tau_K})\mathbf{1}_{\tau_K\leq n}\\&=\rho^{\min\{n,\tau_K\}}V(X_{\min\{n,\tau_K\}})=Y_n\,.\end{aligned}$$

□

4. Algoritmos estocásticos basados en CM

A partir de la unidad anterior podemos obtener una familia de algoritmos que se basa en el uso de cadenas de Markov. Estos se llaman **Markov Chain Monte Carlo** y tienen variadas aplicaciones, incluyendo optimización global de funciones no convexas.

Algunos de los casos de uso incluyen situaciones en las que uno no necesariamente desea llegar a un óptimo global sino más bien un buen mínimo local. Pese a esto, en algunos casos existen garantías de convergencia y son competitivos con métodos deterministas.

Referencias en Pardoux [3] y Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E. [6].

4.1. Cadenas de Markov reversibles

Proposición 4.1.1
Sea $(X_n)_{n\in N} \sim CM(\lambda, P)$, $n \in \mathbb{N}$ *entonces* $(\hat{X}_n)_{n=0}^N := (X_{N-n})_{n=0}^N$ *es cadena de Markov* ***no homogénea*** *con*

$$\mathbb{P}(\hat{X}_{n+1} = y | \hat{X}_n = x) = \frac{(\mu P^{N-n-1})_y}{(\mu P^N)_x} P_{yx}\,.$$

En particular si $\mu = \pi$ *con* π *distribución invariante de* P*, donde* P *es irreducible, entonces* $(\hat{X}_n)_{n=0}^N$ *es* $CM(\pi, \hat{P})$ *homogénea con matriz de transición:*

$$\hat{P}_{xy} = \frac{\pi_y}{\pi_x} P_{yx}\,.$$

Demostración.

$$\begin{aligned}
\mathbb{P}(\hat{X}_{n+1} = y | \hat{X}_n = X_n, \dots, \hat{X}_0 = x_0) &= \frac{\mathbb{P}_\mu(X_N = x_0, \dots, X_{N-n} = x_n, X_{N-n-1} = y)}{\mathbb{P}_\mu(X_N = x_0, \dots, X_{N-n} = x_n)} \\
&= \frac{(\mu P^{N-n-1})_y P_{y,x_n} P_{x_n,x_{n-1}}, \dots, P_{x_0,x_0}}{(\mu P^{N-n})_{x_n} P_{y,x_n} P_{x_n,x_{n-1}}, \dots, P_{x_1,x_0}} \\
&= \frac{(\mu P^{N-n-1})_y P_{y,x_n}}{(\mu P^{N-n})_{x_n}} = \frac{\mathbb{P}_\mu(X_{N-n-1} = y, X_{N-n} = x_n)}{\mathbb{P}_\mu(X_{N-n} = x_n)} \\
&= \mathbb{P}(\hat{X}_{n+1} = y | \hat{X}_n = x_n)\,.
\end{aligned}$$

□

Definición 4.1.1 Reversibilidad
Sea $(X_n)_{n\in N}$ cadena de Markov irreducible recurrente positiva en equilibrio. Decimos que es reversible si $\forall n \in \mathbb{N}$

$$Ley((X_n)_{n=0}^N) = Ley((X_{N-n})_{n=0}^N) \quad (= Ley((\hat{X}_n)_{n=0}^N))\,.$$

Proposición 4.1.2 Condición de balance detallado
Sea $(X_n)_{n\in N}$ *cadena de Markov irreducible recurrente positiva en equilibrio.* X *es* ***reversible*** *si y sólo si* (π, P) *cumplen la condición de balance detallado:*

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

Demostración.
($\Rightarrow$) Reversible $\implies \mathbb{P}_\pi(X_0 = x, X_1 = y)(= \pi_x P_{xy}) = P_\pi(X_1 = x, X_0 = y)(= \pi_y P_{yx})$
($\Leftarrow$) Balance detallado $\implies \hat{P}_{xy} := \frac{\pi_y}{\pi_x} P_{yx} = P_{xy}$
$\therefore (\hat{X}_n)_{n=0}^N \sim CM(\pi, P)$ gracias a la proposición 4.1.1. □

Observación 4.1.1.

- Notación: si (π, P) están en balance detallado, decimos también que π es reversible con respecto a P y vice-versa.
- Si tenemos P matriz estocástica irreducible y $\pi \in \mathcal{P}(E)$ es reversible (i.e., en balance detallado) con respecto a P, entonces π es invariante para P. En efecto:

$$\pi_x P_{xy} = \pi_y P_{yx} \implies (\pi P)_y = \pi_y(\text{ sumando para } x \in E)\,.$$

  La recíproca no es cierta en general.
- π es reversible con respecto a P si y solo si

$$\mathbb{P}_\pi(X_{n+1} = y, X_n = x) = \mathbb{P}_\pi(X_{n+1} = x, X_n = y) \quad \forall x, y \in E\,.$$

- π es invariante con respecto a P si y solo si

$$\forall y \in E, \quad \mathbb{P}_\pi(X_{n+1} = y, X_n \neq y) = \mathbb{P}_\pi(X_{n+1} \neq y, X_n \neq y)\,.$$

Ejemplo 4.1.1 Grafo no-orientado finito
Sea G grafo no-orientado finito. Sea $(X_n)_{n\in\mathbb{N}}$ un paseo aleatorio simple, es decir:

$$P_{xy} = \mathbb{P}(X_{n+1}|X_n = x) := \begin{cases} \frac{1}{deg_x} & \text{si } y \sim x \\ 0 & \text{si no,} \end{cases}$$

con $deg_x = |\{y|y \sim x\}|$ (i.e., el grado de cada vértice). Entonces

$$deg_x \cdot P_{xy} = 1 = deg_y \cdot P_{yx}, \forall x, y$$

$$\therefore \pi = (\pi_x)_{x\in E} = \left(\frac{deg_x}{\sum_{y\in E} deg_y}\right)_{x\in E} \text{ está en balance detallado con } P\,.$$

$$\therefore \pi \text{ es invariante.}$$

4.2. Markov Chain Monte Carlo

4.2.1. Idea general

Sea $\pi \in \mathcal{P}(E), \pi > 0$
Pregunta: ¿existe P matriz estocástica (irreducible) tal que $\pi P = \pi$? (π invariante con respecto a P).

La utilidad de esto sería que si queremos **simular** aproximadamente una variable aleatoria $x_\infty \sim \pi$, basta encontrar P tal que $\pi P = \pi$ y simular $(X_n)_{n\in\mathbb{N}} \sim CM(\lambda, P)$ por tiempo suficiente.

Es más fácil buscar P matriz estocástica tal que $\pi_x P_{xy} = \pi_y P_{yx}$ (reversible) , $\forall x.y \in E$.

Objetivo: dado $\pi \in \mathcal{P}(E), \pi > 0$, queremos construir P irreducible tal que (π, P) estén en balance detallado y tal que $(X_n)_{n\in\mathbb{N}} \sim CM(\mu, P)$ es fácilmente simulable.

4.2.2. Los métodos MCMC

Partimos con $R = (R_{xy})_{xy\in E}$ matriz de transición irreducible "cualquiera" tal que $\forall x, y$, $R_{xy} > 0 \implies R_{yx} > 0$ y además, cuyas transiciones (de $CM(\pi, R)$) sean fáciles de simular.
Luego definimos

$$P_{xy} = \begin{cases} \min(R_{xy}, (\frac{\pi_y}{\pi_x})R_{yx}) & \text{si } x \neq y \\ 1 - \sum_{z\neq x} P_{xz} & \text{si } x = y \end{cases}$$

Proposición 4.2.1
$P = (P_{xy})$ es matriz estocástica irreducible y (π, P) están en balance detallado.

Demostración.

- Veamos que es matriz estocástica: $\sum_{y\neq x} p_{xy} \leq \sum_{y\neq x} R_{xy} \leq 1 \implies P_{xx} \in [0, 1]$ y $\sum_z P_{xz} = 1$.
- Para la irreducibilidad notemos que $\forall x, y \in E$ existen $n, x_1, \ldots, x_n \in E$ tal que

  $$R_{xx_1}, R_{x_2,x_3}, \ldots, R_{x_n y} > 0\,,$$

  luego $R_{yx_n}, R_{x_n,x_{n-1}}, \ldots, R_{x_1 y} > 0$, con lo cual $P_{xx_1}, \ldots, P_{x_n y} > 0$.
- El caso $x = x$ es directo. Para $x \neq y$, $\pi_x P_{xy} = \pi_x R_{xy} \wedge \pi_y R_{yx} = \pi_y P_{yx}$.
  Entonces se tiene la condición de balance detallado.

□

¿Cómo escoger R?
Elegimos un grafo no orientado G con conjunto de vértices E y R tal que $\forall x, y$, $R_{xy} > 0$ si y sólo si $x \sim y$ (vecino) en G.
Dos elecciones "clásicas" son:

- **Gibbs sampler** (muestreo de Gibbs)

  $$R_{xy} = \begin{cases} \pi_y(\sum_{z\sim x} \pi_z)^{-1} & \text{si } x \sim y \\ 0 & \text{si no} \end{cases}$$

- **Algoritmo metrópolis** (paseo aleatorio simple en G)

  $$R_{xy} = \begin{cases} \dfrac{1}{deg_x} & \text{si } x \sim y \text{ con } deg_x = |\{y : y \sim x\}| \\ 0 & \text{si no} \end{cases}$$

4.2.2.1. Metropolis-Hasting

¿Cómo simular $(X_n)_{n\in\mathbb{N}} \sim CM(\mu, P)$?
Sean $(V_n)_{n\in\mathbb{N}} \sim$ i.i.d. $\mathbb{U}([0,1])$ y $f : [0,1] \times E \to E$ función de transición asociada a R.
Sean $(U_n)_{n\geq 1} \sim$ i.i.d. $\mathbb{U}([0,1])$ independientes de las $(V_n)_{n\in\mathbb{N}}$. Simulamos $X_0 = Y_0 \sim \mu$ usando V_0.
Luego, recursivamente definimos:

- Dado $X_n = x$, simulamos

$$Y_{n+1} := f(V_{n+1}, x) = y, \text{ es decir, una transición según } R\,.$$

- Definimos

$$X_{n+1} = \begin{cases} Y_{n+1} & \text{si} \quad U_{n+1} \leq \dfrac{\pi_y R_{yx}}{\pi_x R_{xy}} \\ X_n & \text{si} \quad U_{n+1} > \dfrac{\pi_y R_{yx}}{\pi_x R_{xy}} \end{cases}$$

Proposición 4.2.2

$$(X_n)_{n\in\mathbb{N}} \sim CM(\mu, P)$$

Demostración. $\forall x \neq y$ tenemos:

$$\begin{aligned} \mathbb{P}(\hat{X}_{n+1} = y, \hat{X}_n = y) &= \mathbb{P}\left(f(V_{n+1}, x) = y, X_n = x, U_{n+1} \leq \frac{\pi_y R_{yx}}{\pi_x R_{xy}} \right) \\ &= R_{xy}\mathbb{P}\left(U_{n+1} \leq \frac{\pi_y R_{yx}}{\pi_x R_{xy}} \right) \mathbb{P}(X_n = x) \\ &= R_{xy} \min\left(\frac{\pi_y R_{yx}}{\pi_x R_{xy}}, 1 \right) \mathbb{P}(X_n = x)\,, \end{aligned}$$

$\implies \mathbb{P}(X_{n+1} = y | X_n = x) = \min\{\frac{\pi_y}{\pi_x} R_{yx}, R_{xy}\} = P_{xy}\,,$
y $\mathbb{P}(\hat{X}_{n+1} = y | \hat{X}_n = x) = 1 - \sum_{y\neq x} \mathbb{P}(X_{n+1} = y | X_n = x) = 1 - \sum_{y\neq x} P_{xy} = P_{xx}\,.$ □

Observación 4.2.1. La construcción sólo requiere conocer $\lambda = \alpha\pi$ con $\alpha > 0$ una constante (depende sólo de $\frac{\pi_x}{\pi_y}, x$ e y). Esto es muy importante en la práctica pues muchas veces se conoce sólo la medida λ en E, y calcular la constante de normalización puede ser inviable numéricamente si E es grande.

Observación 4.2.2. El grafo G debe escogerse idealmente de forma que

- No haya estados "muy aislados", de modo que una $CM(\mu, R)$ lo "recorre bien" y $(X_n)_{n\in\mathbb{N}} \sim CM(\mu, P)$ "alcanza rápido" el equilibrio.
- Las transiciones desde cada x sean fáciles de simular (lo que es más difícil si x tiene demasiados vecinos).

Notar que estas dos propiedades apuntan en sentidos contrarios.

Observación 4.2.3. En el caso Gibbs, para calcular $(R_{xy})_{xy\in E}$, hay que calcular sumas $\sum_{z\sim x}\pi_z$, $x\in E$.
Atención: si x tiene muchos vecinos, calcular estas sumas puede ser impracticable. Entonces es mejor usar Metropolis.

4.3. Aplicación de MCMC: simulated annealing

Simmulated annealing ("recocido simulado") tiene como objetivo **minimizar** una función o "energía" $U : E \to \mathbb{R}$ (donde E es grande) con un algoritmo estocástico.

Consideramos un parámetro $\beta > 0$ que denominamos "temperatura inversa" (i.e., $T = \beta^{-1}$ representa la temperatura).

Definición 4.3.1 Medida de Gibbs
Definimos $\pi^\beta \in \mathcal{P}(E)$ mediante:

$$\pi_x^\beta = \frac{\exp^{-\beta U(x)}}{Z_\beta},$$

con

$$Z_\beta = \sum_{y\in E} \exp^{-\beta U(y)} \quad \text{constate de normalización}.$$

A π^β se le llama medida de Gibbs.

Observación 4.3.1.

- π^β da más probabilidad a los x con menor $U(x)$.
- Cuando $\beta \searrow 0$ ($T \nearrow \infty$): $\pi^\beta \underset{\beta\to\infty}{\longrightarrow} Unif(E)$, i.e., es indiferente de la energia U.
- ¿Que pasa cuando $\beta \nearrow \infty$ ($T \searrow 0$)?
  Sean $U_* = \min U$, $A_U = \arg\min U$,

$$\begin{aligned}\pi_x^\beta &= \frac{e^{-\beta U(x)}}{\sum_{y_U} e^{-\beta U(y)} + \sum_{y\in A_U^C} e^{-\beta U(y)}} \\ &= \frac{e^{-\beta U(x)}}{\#A_U e^{-\beta U_*} + \sum_{y\in A_U^C} e^{-\beta U(y)}} \\ &= \frac{e^{-\beta(U(x)-U_*)}}{\#A_U + \sum_{y\in A_U^C} e^{-\beta(U(y)-U_*)}}) \underset{\beta\to\infty}{\longrightarrow} \begin{cases}\frac{1}{\#A_U} & \text{si } x\in A_U \\ 0 & \text{si } x\notin A_U\end{cases}\end{aligned}$$

  Entonces si $\beta \nearrow \infty$, $\pi^\beta \underset{\beta\to\infty}{\longrightarrow} \pi^\infty = Unif(A_U)$.
- Para cada $\beta > 0$ sabemos simular $(X_n^\beta)_{n\in\mathbb{N}}$ (reversible) que converge en ley a $\pi^\beta \propto e^{-\beta U}$ cuando $n \to \infty$.

La idea del método **Simmulated annealing** consiste en tomar $\beta_n \searrow \infty$ y simular una cadena de markov no homogénea $(X_n)_{n\in N}$"="$(X_n^{\beta_n})_{n\in\mathbb{N}}$ (usando Metropolis-Hastings (4.2.2.1.)), es decir, en cada tiempo n, simular transición tal que

$$\mathbb{P}(X_{n+1} = y | X_n = x) = \mathbb{P}(X_{n+1}^{\beta_{n+1}} = y | X_n^{\beta_{n+1}} = x)\,.$$

Antes de estudiar como hacerlo, veamos qué hace $(X_n^\beta)_{n\in\mathbb{N}}$ cadena simulada con Metropolis-Hastings con $\beta > 0$ fijo.

- Contamos con G grafo regular no orientado con conjunto de vértices E.
- $\pi^\beta \propto e^{-\beta U(x)}, \forall x \in E$.
- En el tiempo n, dado que $X_n = x$, simulamos $y = Y_{n+1} \sim Unif\{z : z \sim x\}$, transición desde x para un paseo aleatorio simple $(R_{xy} = (deg(G))^{-1}, \forall x \sim y)$.
- Sampleamos $U_{n+1} \sim \mathbb{U}([0,1]) \perp\!\!\!\perp$ de todo y

$$X_{n+1}^\beta = \begin{cases} Y_{n+1} = y & \text{si } U_{n+1} \leq \min\left\{\frac{\pi_y^\beta}{\pi_x^\beta}, 1\right\}, \\ X_n^\beta & \text{si } U_{n+1} > \min\left\{\frac{\pi_y^\beta}{\pi_x^\beta}, 1\right\} \end{cases}.$$

Notemos que $\min\left\{\frac{\pi_y^\beta}{\pi_x^\beta}, 1\right\} = \min\{e^{-\beta(U(y)-U(x))}, 1\}$. Luego:

- Si $U(y) < U(x)$ la transición siempre se realiza pues el mínimo es 1 y $U_{n+1} \leq 1$ siempre. Esto quiere decir que si la energía del estado propuesto y es menor que la del estado actual x, la transición se realiza de todos modos.
- Si $U(y) \geq U(x)$ el mínimo es menor que 1, luego se realiza la transición si $U_{n+1} \leq e^{-\beta(U(y)-U(x))}$, y en caso contrario $(>)$ no. Así, si la propuesta es transitar a un estado de mayor energía, esto puede llevarse acabo a veces, pues puede permitir salir de un mínimo local, distinto del mínimo global buscado.

Entonces, si $\beta \gg 1$ $(0 < T \ll 1)$ el valor de la energía es menos propenso a "subir" en cada paso, y **tiende a ir hacia un mínimo** (que puede no ser global). Por otro lado si $0 < \beta \ll 1$ $(T \gg 1)$, la energía fluctúa de manera más aleatoria, pudiendo subir o bajar en cada paso, lo que permite **escapar de mínimos locales**.

Así, para n pequeño (β_n pequeño y T_n alta), X_n tiende a pasearse por todo E sin tomar muy en cuenta el "paisaje de energía" dado por U (desorden), mientras que, para n grande (β_n grande y T_n chico) X_n se mueve muy poco, sólo alrededor de algún mínimo (local) hasta que se "congela" ahí.

El siguiente resultado indica una elección teórica para β_n que asegura convergencia a mínimos globales:

Teorema 4.3.1
Sea

$$\triangle > Osc(U) := \max_{x\in E} U(x) - \min_{x\in E} U(x)\,.$$

Entonces, si $\beta_n := \dfrac{1}{\triangle}\log(1-n)$

$$\Longrightarrow\ Ley(X_n) \underset{n\to\infty}{\longrightarrow} \pi^{\infty} = Unif(A_U)\ \textit{minimizante}\,.$$

Más aún, se prueba que $X_n \overset{\mathbb{P}}{\underset{n\to\infty}{\longrightarrow}} A_U$ *(mínimo global).*

Demostración. Ver Pardoux cap 10 [3] para una versión para C.M. en tiempo continuo.

Observación 4.3.2.

- En la práctica $\ln(1+n)$ es demasiado lento para poder observar una evolución descendiente de la energía.
- Usualmente se suele escoger:
  - β_n polinomial en n (por ejemplo n^2)
  - β_n exponencial en n,

  pero no hay garantías de convergencia a A_U en esos casos.
- En cada problema hay que jugar con distintas sucesiones $(\beta_n)_{n\in\mathbb{N}}$ tal que $\beta_n \nearrow \infty$ y puntos de inicio X_0, y escoger finalmente el mejor mínimo encontrado.
- El algoritmo da una heurística para encontrar buenos mínimos locales más eficiente que optimización discreta.

Idea física:
"Annealing" = recocido, viene de la metalurgia y se refiere a un procedimiento para hacer más dúctil (menos duro) una aleación, calentando(a) por sobre el nivel de cristalización y dejándola enfriar lentamente para alcanzar un estado de energía potencial basal material homogéneo, pocas "dislocaciones".
Si el enfriamiento es muy rápido queda un material homogéneo pero duro:
"Quenching"=templado.

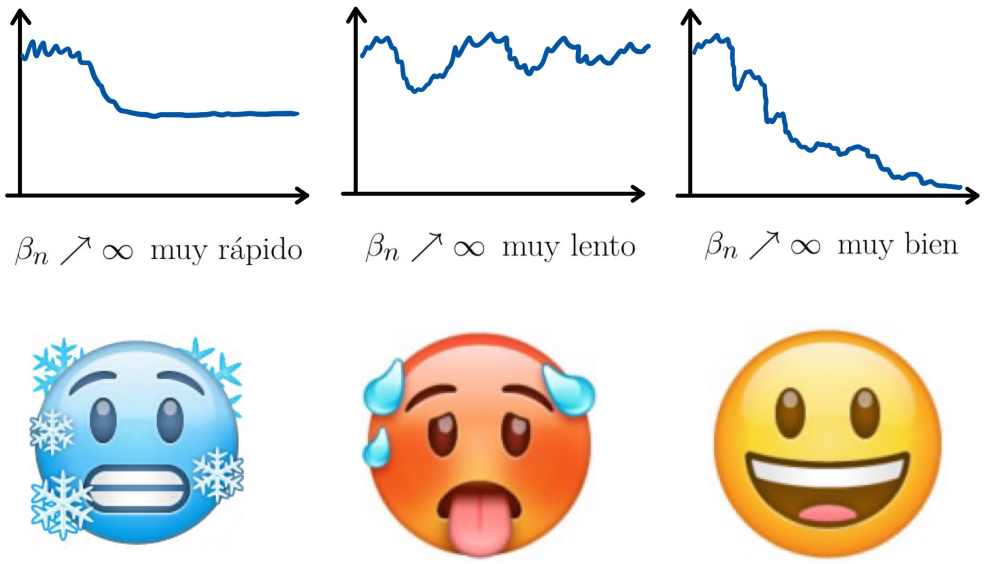


Figura 5: Idea del efecto de las sucesiones β_n en la minimización de la energía U.

4.4. MCMC y estadística Bayesiana

4.4.1. Recuerdo de estadística Bayesiana

Idea: modelamos simultaneamente y probabilisticamente:

- Observaciones x de un fenómeno o dato aleatorio, cuya ley "$p(x|\theta)$" depende de un parámetro $\theta \in \Theta$.
- La incertidumbre, desconocimiento o conocimiento parcial de θ, lo que representamos asumiendo que θ es aleatorio y sólo conocemos su distribución "$p(\theta)$" "a priori".
- En lo anterior, la observación x está en un conjunto $\mathfrak{X}$ que puede ser subconjunto de $\mathbb{R}^d$, un conjunto finito o numerable, etc.
  En general $x \mapsto p(x|\theta)$ (ley de x dado θ) es
  - o bien una función de masa discreta:
  $$\sum_{x \in \mathfrak{X}} p(x|\theta) = 1$$
  - o bien una densidad de probabilidad:
  $$\int_{\mathfrak{X}} p(x|\theta) dx = 1\,.$$
- x puede ser una "muestra" $x = (x_1, \ldots, x_n)$.

- Θ puede ser subconjunto de $\mathbb{R}^k$, conjunto finito o numerable y $p(\theta)$ denota, según corresponda:
  - una función de masa discreta
  - o bien una densidad de probabilidad,

  con respecto a θ, y se le llama **ley a priori**.
- $p(x, \theta) := p(x|\theta)p(\theta)$ es ley conjunta en $\mathfrak{X} \times \Theta$ del parámetro y una observación.
  Notar que
  $$\int_\Theta \int_\mathfrak{X} p(x, \theta) dx d\theta = \int_\Theta \int_\mathfrak{X} p(x|\theta) dx p(\theta) d\theta = 1 \,.$$
- Dado $\theta \in \Theta$, $x \mapsto p(x|\theta)$ se llama ley de x dado θ, o bien la densidad de x condicional a θ.
- Dado $X \in \mathfrak{X}$ observación, $\theta \mapsto L(\theta) = L_X(\theta) = p(X|\theta)$ se llama **función de verosimilitud** (likelihood). Notar que $\int L(\theta) d\theta \neq 1$ en general.

Objetivo de la inferencia Bayesiana: "Re-estimar" el parámetro θ (modificar o actualizar la ley que describe lo que sabemos de θ) usando la información que nos da el observar x.

Teorema 4.4.1 Bayes
La ley de θ dado x, también llamada ley posterior de θ, está dada por:

$$\begin{aligned} p(\theta|x) &:= \frac{p(x, \theta)}{p(x)} \\ &= \frac{p(x|\theta)p(\theta)}{p(x)} \,, \end{aligned}$$

con $p(x) = \displaystyle\int_\Theta p(x, \theta) d\theta$ *ley marginal (no condicional).*

Luego $p(\theta|x) \propto p(x|\theta)p(\theta)$. De esta forma, $p(x)$ aparece sólo como una constante de normalización, dependiente de la observación x, cuyo cálculo requiere en general sumar o integrar sobre todo el espacio. Por ello, **calcular la constante de normalización $p(x)$ puede ser muy costoso, y hay evitar tener que hacerlo**. Es por este motivo que los métodos MCMC son muy útiles en este contexto, como veremos un poco más adelante.

4.4.2. Aplicaciones de estadística Bayesiana

- Observamos $x_!, \ldots, x^n$ i.i.d. $\sim p(x|\theta)dx$ con θ fijo.
  Sean $\mathcal{D} = \{x_1, \ldots, x_n\}$ "datos" y su densidad conjunta dada por:
  $$p(\mathcal{D}|\theta) := p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^n p(x_i|\theta) \,.$$
  Además, consideramos su función de verosimilitud:
  $$L_\mathcal{D}(\theta) = L_{x_1,\ldots,x_n}(\theta) = \prod_{i=1}^n L_{x_i}(\theta)$$

- Luego la ley a posteriori de θ dados $\mathcal{D}$ es la ley en Θ:

$$\mathcal{D} = p(\theta|x_1,\ldots,x_n) = \frac{\prod_{i=1}^n p(x_i|\theta)p(\theta)}{p(\mathcal{D})} \propto \prod_{i=1}^n p(x_i|\theta)p(\theta)$$

  donde la última expresión es evaluable. $p(\mathcal{D})$ es la densidad marginal de los datos y requiere integrar:

$$\int_\Theta p(\mathcal{D}|\theta)p(\theta)d\theta = \int_\Theta \prod_{i=1}^n p(x_i|\theta)p(\theta)d\theta\,.$$

- En base a $p(\theta|\mathcal{D}) = p(\theta|x_1,\ldots,x_n)$, finalmente se construye un "**posterior predictivo**", i.e., un valor $\hat{\theta}_n \in \Theta$ ("estimador de θ") que mejor explica los datos $x_1,\ldots,x_n$.

4.4.2.1. Ejemplos de estimadores Bayesianos

Los siguiente son estimadores Bayesianos clásicos (hay muchos otros):

- **Media a posteriori**
  Considera como estimador predictivo la media con respecto a la ley a posteriori de θ.

$$\hat{\theta}_n = \int_\Theta p(\theta|\mathcal{D})d\theta\,.$$

- **Máximo a posteriori**
  Es un parámetro cuyo valor maximiza la probabilidad a posteriori de observar la data $\mathcal{D}$

$$\begin{aligned}\hat{\theta}_n &:= \arg\max_\theta p(\theta|\mathcal{D}) \\ &= \arg\max_\theta [\sum_{i=1}^n \log p(x_i|\theta) + \log p(\theta)]\,.\end{aligned}$$

  En la práctica, la maximización se lleva a cabo como en la última expresión.

Observación 4.4.1.

- También es posible construir regiones o **regiones o intervalos de confianza** para θ, es decir (para θ real), un intervalo I tal que $\mathbb{P}(\theta \in I|X_1,\ldots,X_n) \approx 95\,\%$ (por ejemplo).
- Se requiere "conocer" $p(\theta|\mathcal{D})$ para optimizar en θ o bien para integrar $p(\theta|\mathcal{D})$ con respecto a θ. En algunos (pocos) casos, $p(\theta|\mathcal{D})$ tiene forma analítica explícita.

4.4.3. Uso de MCMC

En general $p(\theta|x_1,\ldots,x_n) = \prod_{i=1}^n p(x_i|\theta)p(\theta)$ no tiene forma cerrada.
Además para tener su valor numérico, se requiere integrar $\int \prod_{i=1}^n p(x_i|\theta)p(\theta)d\theta = p(x_1,\ldots,x_n)$, lo cual puede ser muy costoso.

Idea: samplear de

$$\begin{aligned}\Pi(\theta) &:= p(\theta|x_1,\ldots,x_n)\\ &= \frac{\prod_{i=1}^n p(x_i|\theta)p(\theta)}{p(\mathcal{D})}\end{aligned}$$

usando MCMC en Θ, pues con este método **no se requiere conocer ni calcular** $p(\mathcal{D})$**, y basta poder evaluar rápidamente** $\prod_{i=1}^n p(x_i|\theta)p(\theta)$. La cadena de Markov construida con MCMC vive en Θ y tiene distribución invariante

$$\Pi(\theta) \propto \prod_{i=1}^n p(x_i|\theta)p(\theta)\,.$$

Cabe notar que:

- el método puede aplicarse tanto para Θ discreto como $\Theta = \mathbb{R}^k$ (usando "paseo aleatorio en $\mathbb{R}^k$ como cadena Markov de base").
- la simulación es costosa en general, sobretodo si el espacio de parámetros tiene dimensión muy grande (se requiere correr muchas veces la cadena de MCMC, por mucho tiempo).

5. Algoritmos estocásticos en aprendizaje de máquinas

Esta unidad estará dedicada a la exploración de algoritmos tipo **gradiente estocástico**. Si bien los conceptos han sido introducidos hace tiempo, su uso ha subido significativamente en años recientes gracias a su uso en iteligencia artificial y aprendizaje de máquinas. Más precisamente, estos métodos son utilizados para el entrenamiento de redes neuronales profundas, cuyo caso de uso será también introducido en esta sección.

5.1. Introducción

Observamos $(x_1, y_1), \ldots, (x_n, y_n), \cdots \in \mathbb{R}^d \times \mathbb{R}$ muchos datos (potencialmente infinitos) de dimensión (posiblemente) grande.

Ejemplo 5.1.1 Clasificación de imágenes
Tenemos un conjunto de imágenes, cada una con una etiqueta. Por ejemplo, la foto de un gato, como se grafica en figura 6[1], puede representarse como un vector al "aplanar" la matriz correspondiente a la imagen. Las distintas k clases posibles pueden enumerarse de modo que al concepto *gato* se le asigna un número $y \in \{1, \ldots, k\} \subset \mathbb{N} \subset \mathbb{R}$.

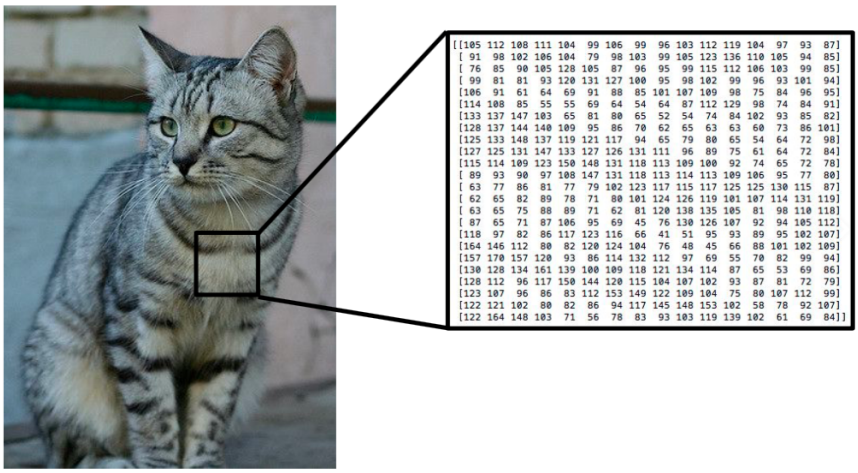


Figura 6: Imagen con etiqueta "gato".

De este modo, dado un elemento i de nuestros n datos, $x_i \in \mathbb{R}^d$ es un vector que representa la imagen, mientras que $y \in \{1, \ldots, k\}$ corresponde a una etiqueta.

Objetivo: aprender de los datos, i.e., aprender a predecir, al ver un x nuevo, la etiqueta y correspondiente.

Ejemplo 5.1.2 Redes Neuronales
Se busca predecir y mediante

$$\hat{y}(x, \theta) = \sum_{j=1}^{N} \sigma_*(x_j, \theta_j)\,.$$

[1] Fuente: https://dongminlee.tistory.com/18

donde:

- N es el número de neuronas.
- $\theta = (\theta_1, \dots, \theta_N) \in (\mathbb{R}^D)^N$ son los parámetros ("pesos").
  En general, para $j = 1, \dots, N$ tomamos $\theta_j = (a_j, b_j, \omega_j) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^D$ y σ_* de la forma

$$\sigma_*(x, \theta_j) = a_i \sigma(\langle x, \omega_i \rangle + b_i)\,.$$

¿Como aprende?

- Consideramos $l : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ función de pérdida.
  Ej: $l(y, \hat{y}) = (y - \hat{y})^2$
- Idealmente, buscamos

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{M} l(y_i, \hat{y}(x_i, \theta))\,,$$

  con M grande.
- Primera idea: aplicar un algoritmo de optimización para "entrenar la red" con los datos conocidos. Es decir, encontrar $\hat{\theta}$ óptimo (o casi).

Problemas:

- Función objetivo costosa de evaluar, debido a una o varias de las razones siguientes:
  - M es grande
  - d es grande, donde d es la dimensión de x_i.
  - N es grande
- Requiere tiempo y memoria computacionales (se requiere usar toda la información en todos los pasos).
- Sobreajuste (overfitting): Si entrenamos "perfectamente" usando todos los $(x_1, y_1), \dots, (x_n, y_n), \dots$ (mínimo global) sobreajustamos y perdemos la "capacidad de generalización".

Segunda idea: Supongamos $(x_i, y_i) \sim$ i.i.d. de ley μ no conocida.
En vez de buscar

$$\begin{aligned}\hat{\theta} &= \arg\min_{\theta} \sum_{i=1}^{M} l(y_i, \hat{y}(x_i, \theta)) \\ &= \arg\min_{\theta} \frac{1}{M} \sum_{i=1}^{M} l(y_i, \hat{y}(x_i, \theta),\end{aligned}$$

buscamos

$$\hat{\theta} = \arg\min_{\theta} \mathbb{E}(l(y, \hat{y}(X, \theta)))$$

con $(X, Y) \sim \mu$.

¿Tiene sentido?, ¿de qué sirve si no conocemos μ? ¿qué cambia?

5.2. Algoritmo de gradiente estocástico

El descenso de gradiente estocástico (o *S.G.D.* por sus siglas en inglés) es uno de los pilares del desarrollo reciente del aprendizaje de máquinas y de la inteligencia artificial.

Bibliografía: *Stochastic Approximation* (Robbins & Monro) [7]..... Referencia más actual: *Online Learning and Stochastic Approximations* [8].

5.2.1. El algoritmo

Consideramos el problema

$$\min_{\theta} \mathbb{E}(f(\theta, X)),$$

con:

- $f : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}$
- $X \sim \mu$
- $(\gamma_t)_{t\in\mathbb{N}}$ pasos o tasa de aprendizaje (*learning rate*)

Asumiremos en lo que sigue que f es tal que $F(\theta) := \mathbb{E}(f(\theta, X))$ está bien definida, y que se cumple la relación $\nabla F(\theta) = \mathbb{E}(\nabla_\theta f(\theta, X))$ para todo θ.

En vez de usar un algoritmo gradiente usual $\theta^{t+1} = \theta^t - \gamma_{t+1}\nabla F(\theta^t) = \theta^t - \gamma_{t+1}\mathbb{E}(\nabla_\theta f(\theta^t, X))$, con $(\gamma_t)_{t\in\mathbb{N}}$ sucesión en $\mathbb{R}_+$ pasos, la idea será **considerar observaciones** $x_1, x_2, \ldots, x_t$ i.i.d. $\sim \mu$ y el algoritmo definido por las iteraciones:

$$\theta^{t+1} := \theta^t - \gamma_{t+1}\nabla_\theta f(\theta^t, x_{t+1}).$$

Observación 5.2.1.

- El término $\nabla_\theta f(\theta^t, x_{t+1})$ se puede ver como un gradiente exacto perturbado por cierto ruido aleatorio, más precisamente,

  $$\nabla_\theta f(\theta^t, x_{t+1}) = \mathbb{E}(\nabla_\theta f(\theta^t, X)) + \Delta_{t+1} = \nabla F(\theta^t) + \Delta_{t+1},$$

  con $\Delta_{t+1} = \nabla_\theta f(\theta^t, x_{t+1}) - \mathbb{E}(\nabla_\theta f(\theta^t, X))$ v.a. centrada.
- En cada paso necesitamos evaluar una sola vez $\nabla_\theta f$ (en un solo dato nuevo).
- Podemos usar los datos a medida que llegan.
- **Caso particular importante (literatura de optimización)**

  Se quiere optimizar $\tilde{F}(\theta) := \sum_{i=1}^{M} \tilde{f}_i(\theta)$, con $\tilde{f}_i : \mathbb{R}^d \to \mathbb{R}$ (M funciones), lo cual es equivalente a optimizar $F(\theta) = \frac{1}{M}\sum_{i=1}^{n} \tilde{f}_i(\theta) = \mathbb{E}(f(I, \theta))$ con $I = X \sim \frac{1}{M}\sum_{i=1}^{M} \delta_i$ (esto es, $I \in \{1, \ldots, M\}$ es un índice aleatorio elegido de manera uniforme), y $f(i, \theta) = \tilde{f}_i(\theta)$. Entonces,

  $$\theta^{t+1} := \theta^t - \gamma_t \nabla_\theta f(\theta^t, I_{t+1}) = \theta^t - \gamma_t \nabla_\theta \tilde{f}_{I_{t+1}}(\theta),$$

  con $I_1, I_2, \ldots$ i.i.d. $\sim \mathbb{U}(\{1, \ldots, M\})$.

El siguiente es un enunciado clásico sobre este tipo de algoritmo, que damos inicialmente sin todos los detalles:

Teorema 5.2.1 Sigmund, Robbins, (1951)
Bajo hipótesis razonables sobre f (regularidad en θ, integrabilidad de $\nabla_\theta f_+$, cotas), si $\gamma_t \overset{\searrow}{_{t\to\infty}} 0$ suficientemente lento la sucesión:

$$\theta^t \overset{c.s.}{\underset{t\to\infty}{\longrightarrow}} \text{ Conjunto de puntos críticos de } F(\theta) = \mathbb{E}(f(\theta, X)).$$

Más aún, si f es convexa, estrictamente en θ (más algunas hipótesis adicionales), entonces

$$\theta^t \overset{c.s.}{\underset{t\to\infty}{\longrightarrow}} \arg\min_\theta \mathbb{E}(f(\theta, X)) \,.$$

Observación 5.2.2. Si bien este tipo de resultados fueron obtenidos por primera vez hace cerca de 70 años, la utilización del descenso de gradiente estocástico (o *S.G.D.* ha sido fuertemente reimpulsada con el desarrollo reciente del aprendizaje de máquinas y de la inteligencia artificial pues, entre otros motivos, los siguientes:

- Permite "entrenar" (estimar) parámetros con datos, a medida que estos van llegando.
- Permite "actualizar" la estimación en tiempo real (a medida que llegan nuevos datos).
- Es "escalable": el costo es temporal (uso de memoria constante) y proporcional al tamaño del conjunto de datos.
- Si bien no está garantizada la convergencia a un óptimo global, muchas veces permite encontrar un mínimo local $\hat{\theta} = \theta^t$ que "generaliza bien" (como estimador) en el sentido siguiente: si t es suficientemente grande, $X_1, \ldots, X_t$ es el conjunto de entrenamiento, y $X_{t+i}, i = 0, \ldots, N$ son datos nuevos, entonces

$$\frac{1}{N}\sum_{i=1}^{N} f(\theta^t, X_{t+i}) \text{ es cercano a } \mathbb{E}(f(\theta^t, X)) \text{ y } \mathbb{E}(f(\theta^t, X)) \text{ es cercano a } \min_\theta \mathbb{E}(f(\theta, X)).$$

Observación 5.2.3. Desde el punto de vista práctico:

- Típicamente, se buscan mínimos locales "buenos", de valores cercanos al global, pero no necesariamente un mínimo global.
- Se sugiere correr el algoritmo desde muchos θ_0 iniciales y guardar el mejor valor obtenido.
- Se sugiere probar además distintas elecciones de paso γ_t y guardar la que de mejores resultados. En la práctica se suele usar γ_t =constante o constante por tramos (t no tiende a infinito en la vida real)
- Hay muchas variantes para acelerar "convergencia" o bien para explorar mejor el espacio. Un ejemplo de eso último es agregar "ruido" o aleatoriedad adicional, para evitar quedar atrapado en mínimos locales malos.

Notación. En lo que sigue, denotamos $\partial f(\theta, x) = \nabla_\theta \,.$

5.2.2. Convergencia en el caso convexo

Para probar la convergencia en el caso convexo necesitaremos herramientas de cálculo estocástico.

Definición 5.2.1 Martingala
Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad dotado de una filtración $(\mathcal{F}_t)_{t\in\mathbb{N}}$ (i.e., σ-álgebras tal que $\forall t \in \mathbb{N}$, $\mathcal{F}_t \subset \mathcal{F}_{t+1} \subset \mathcal{F}$). Una familia de variables aleatorias se dice martingala si

$$X_t \in L^1(\mathcal{F}_t) \forall t \in \mathbb{N}$$

y

$$\forall t \in \mathbb{N} \quad \mathbb{E}(X_{t+1}|\mathcal{F}_t) = X_t \text{ c.s.}$$

Notación. Si en la última ecuación tenemos $\geq$ en vez de igualdad entonces la familia se dice **sub-martingala**. Si tenemos $\leq$ entonces la llamamos **sobre-martingala**.

En particular usaremos el siguiente teorema del curso de cálculo estocástico, que asumiremos sin demostración.

Teorema 5.2.2 Convergencia sobre-martingalas
Sea $(X_t)_{t\in\mathbb{N}}$ una sobre-martingala con respecto a $(\mathcal{F}_t)_{t\in\mathbb{N}}$ tal que $\sup_{t\in\mathbb{N}} \mathbb{E}(X_t^-) < \infty$. Entonces $\exists X_\infty \in L^1$ tal que

$$X_t \xrightarrow[t\to\infty]{c.s.} X_\infty \,.$$

Teorema 5.2.3 Descenso de gradiente estocástico caso convexo
Supongamos que los datos .... son v.a. i.i.d. de ley μ definas en un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$. Adems, supongamos que:

i) $\forall \theta \,, f(\theta, \cdot) \in L^1(\mu)$. *Además* $f(\cdot, x) \in \mathcal{C}^1 \mu(dx) - c.s.,$ *y* $\partial_\theta f(\theta, \cdot) \in L^1(\mu)$

ii) F *tiene un único mínimo* θ^*

iii) $\forall \epsilon > 0, \inf_{\theta : \|\theta - \theta^*\|^2 > \epsilon} (\theta - \theta^*) \nabla F(\theta) > 0$
i.e., lejos de θ^*, F *decrece estrictamente, uniformemente hacia* $F(\theta^*)$

iv) $\exists A, B \geq 0$ *tal que* $\mathbb{E}(\|\partial f(\theta, X)\|^2) \leq A + B\|\theta - \theta^*\|^2, \forall \theta$

v) $\sum_{t\in\mathbb{N}} \gamma_t = \infty$, $\sum_{t\in\mathbb{N}} \gamma_t^2 < \infty$ *(por ejemplo:* $\gamma_t = \frac{1}{t}$*).*

Entonces

$$\theta_t \xrightarrow[t\to\infty]{c.s.} \theta^* \,.$$

Observación 5.2.4.

- (i) $\implies$ $F \in \mathcal{C}^1(\mathbb{R}^d)$ con $\nabla F = \mathbb{E}(\partial_\theta f(\theta, x))$.
- Si bien no pedimos explícitamente que la función F sea convexa, haciendo un Taylor en θ se puede verificar que las hipótesis ii) y iii) se cumplen, si F es de clase $\mathcal{C}^2$ y estrictamente convexa.
- (iii) impide que el gradiente se "aplane" lejos del mínimo global.

- (iv) se cumple si por ejemplo $f(\cdot, x)$ es de clase $\mathcal{C}^2$ con $\mathbb{E}(\|Hess_\theta f(\theta, X)\|) < \infty$

Demostración de Teorema 5.2.3 descenso de gradiente estocástico, caso convexo. Tomemos la función siguiente como "función de de Lyapunov":

$$h(\theta) = \|\theta - \theta^*\|^2 ,$$

y denotemos $h_t := h(\theta_t)$. Queremos probar que h_t converge a 0 c.s., con lo cual tendremos que $\theta_t \overset{c.s.}{\underset{t\to\infty}{\longrightarrow}} \theta^*$. En efecto,

$$\theta_{t+1} - \theta^* = \theta_t - \theta^* - \gamma_t \partial f(\theta_t, x_{t+1}) .$$

Luego aplicando $\| \ \|^2$,

$$\|\theta_{t+1} - \theta^*\|^2 = \|\theta_t - \theta^*\|^2 - 2\gamma_t(\theta_t - \theta^*)\partial f(\theta_t, x_{t+1}) + \gamma_t^2 \|\partial f(\theta_t, x_{t+1})\|^2 ,$$

entonces

$$h_{t+1} - h_t = -2\gamma_t(\theta_t - \theta^*)\partial f(\theta_t, x_{t+1}) + \gamma_t^2 \|\partial f(\theta_t, x_{t+1})\|^2 .$$

Ahora haremos aparecer una sobre-martingala. Tomemos como filtración la tribu generada por las observaciones, i.e.,$(\mathcal{F}_t = \sigma(x_0, \dots, x_t))_{t\in\mathbb{N}}$, asumiendo además que la condición inicial θ_0 es determinista.

Por inducción se ve fácilmente que $\theta_t \in L^2(\Omega, \mathcal{F}, \mathbb{P}) \, \forall t \in \mathbb{N}$, gracias a la condición iv). Tomando esperanza condicional respecto a $\mathcal{F}_t$:

$$\mathbb{E}(h_{t+1} - h_t|\mathcal{F}_t) = -2\gamma_t(\theta_t - \theta^*)\mathbb{E}(\partial f(\theta, X))\big|_{\theta=\theta_t} + \gamma_t^2 \mathbb{E}(\|\partial f(\theta, X)\|^2)\big|_{\theta=\theta_t} ,$$

donde usamos que $x_{t+1} \perp\!\!\!\perp \mathcal{F}_t$, con lo cual $\mathbb{E}(G(\theta_t, x_{t+1})|\mathcal{F}_t)) = \mathbb{E}(G(\theta, X))|_{\theta=\theta_t}$ para toda G apropiada. Además, por (iv), $\exists A, B \geq 0$ tal que $\mathbb{E}(\|\partial f(\theta, X)\|^2)\big|_{\theta=\theta_t} \leq A + B\|\theta_t - \theta^*\|^2$, por ende

$$\mathbb{E}(h_{t+1} - h_t|\mathcal{F}_t) \leq -2\gamma_t(\theta - \theta^*)\mathbb{E}(\partial f(\theta, X))\big|_{\theta=\theta_t} + A\gamma_t^2 + B\gamma_t^2 h_t .$$

Como h_t es $\mathcal{F}_t$-medible, puede entrar en la esperanza condicional del lado izquierdo, con lo cual

$$\mathbb{E}(h_{t+1} - (1 + \gamma_t^2 B)h_t|\mathcal{F}_t) \ \leq \ -2\gamma_t(\theta_t - \theta^*)\nabla F(\theta_t) + \gamma_t^2 A.$$

Definimos $\mu_t := \prod_{k=1}^{t-1} \frac{1}{1+\gamma_k^2 B}$ y $h'_t := \mu_t h_t$. Multiplicando por μ_{t+1} queda que

$$\mathbb{E}(h'_{t+1} - h'_t|\mathcal{F}_t) \leq -2\gamma_t(\theta_t - \theta^*)\nabla F(\theta_t) + \gamma_t^2 A\mu_{t+1} \ \leq \gamma_t^2 A\mu_t ,$$

donde usamos (iii) y el hecho que $\mu_{t+1} \leq \mu_t$. Notar que $\mu_t \overset{}{\underset{t\to\infty}{\longrightarrow}} \mu_\infty \in (0, \infty)$, como puede verse tomando logaritmo y usando que $\sum \gamma_t^2 < \infty$. De la desigualdad anterior deducimos que

$$X_t := h'_t - \sum_{k=0}^{t-1} A\gamma_t^2 \mu_t$$

es una sobre-martingala con respecto a $(\mathcal{F}_t)_{t\in\mathbb{N}}$, es decir $\mathbb{E}(X_{t+1}|\mathcal{F}_t) \leq X_t \, \forall t \in \mathbb{N}$. Para usar el Teorema 5.2.2 de convergencia de sobre-martingalas, veamos que es acotada por debajo. En efecto,

como $h'_t \geq 0$, tenemos que

$$X_t \geq 0 - A\sum_{k=0}^{\infty} \gamma_t^2 \cdot (\sup_{j\in\mathbb{N}} \mu_j) > -\infty.$$

Así, por el Teorema de convergencia de sobre-martingalas, $\exists X_\infty \in L^1$ tal que $X_t \overset{c.s.}{\underset{t\to\infty}{\longrightarrow}} X_\infty$. Dado que

$$A\sum_{t=1}^{N} \gamma_t^2 \mu_t \underset{N\to\infty}{\longrightarrow} A\sum_{t=1}^{\infty} \gamma_t^2 \mu_t < \infty,$$

se sigue que $h'_t \overset{c.s.}{\underset{t\to\infty}{\longrightarrow}} h'_\infty$, para cierta v.a. $h'_\infty \in L^1$, y puesto que $\mu_t \underset{t\to\infty}{\longrightarrow} \mu_0 \in (0,\infty)$, deducimos que

$$h_t \overset{c.s.}{\underset{t\to\infty}{\longrightarrow}} h_\infty$$

para cierta $h_\infty \in L^1$. Para concluir el teorema, veamos que $h_\infty = 0\, c.s.$. Volviendo a la desigualada satisfecha por $\mathbb{E}(h'_{t+1} - h'_t|\mathcal{F}_t)$, vemos que

$$0 \leq 2\mu_t\gamma_t(\theta_t - \theta^*)\nabla F(\theta_t) \leq \gamma_t^2 A\mu_t + \mathbb{E}(h'_t - h'_{t+1}|\mathcal{F}_t)\,,$$

de donde

$$\begin{aligned} 2\mathbb{E}(\sum_{t=0}^{N} \gamma_t\mu_t(\theta_t - \theta^*)\nabla F(\theta_t)) &\leq \sum_{t=1}^{N} \gamma_t^2 A\mu_t + \mathbb{E}(h_0) - \mathbb{E}(h_N) \\ &\leq A\sum_{t=1}^{\infty} \gamma_t^2\mu_t + \mathbb{E}(h_0) < +\infty \end{aligned}$$

$$\therefore \quad \sum_{t=0}^{\infty} \gamma_t\mu_t(\theta_t - \theta^*)\nabla F(\theta_t) < \infty \quad c.s..$$

Puesto que $\mu_t \underset{t\to\infty}{\longrightarrow} \mu_\infty \in (0,\infty)$, μ_t está acotada por debajo lejos de 0, por lo que

$$\sum_t \gamma_t(\theta_t - \theta^*)\nabla F(\theta_t) < \infty\, c.s..$$

Usando (v), $\sum_t \gamma_t = \infty$, con lo cual necesariamente se cumple que

$$\liminf_{t\to\infty} (\theta_t - \theta^*)\nabla F(\theta_t) = 0\, c.s..$$

Para concluir, consideremos $\Omega'_\epsilon = \{h_\infty > \sqrt{\epsilon}\}$ y supongamos que $\mathbb{P}(\Omega'_\epsilon) > 0$. Usando (iii) tenemos que $\mathbb{P}(\liminf_{t\to\infty}(\theta_t - \theta^*)\nabla F(\theta_t) > 0) \geq \mathbb{P}(\Omega'_{\epsilon/2}) > 0$, que es una contradicción.

$$\therefore \quad \mathbb{P}(\Omega'_\epsilon) = 0 \quad \forall\epsilon > 0, \quad \Longrightarrow \quad h_\infty = 0 \quad c.s..$$

□

5.2.3. Convergencia en caso no-convexo

Teorema 5.2.4 Extensión de Descenso de gradiente estocástico (caso no-convexo) *Supongamos:*

i) $f(\theta,\cdot)\in L^1(\mu)$, $f(\cdot,x)\in\mathcal{C}^3, \mu(dx)-c.s.$, $\partial^k f(\theta,\cdot)\in L^1(\mu), k=1,2,3$ *(luego* $F\in\mathcal{C}^3$*)*

ii) F *es acotada por debajo*

iii) $\exists A_k, B_k\geq 0$ *tal que* $\mathbb{E}(\|\partial f(\theta,X\|^k)\leq A_k+B_k\|\theta\|^2, \quad k=1,2,3$

iv) $\forall M>0, \inf_{\theta:\|\theta\|^2>M}\theta\nabla F(\theta)>0$

v) $\sum_{t\in\mathbb{N}}\gamma_t=\infty$, $\sum_{t\in\mathbb{N}}\gamma_t^2<\infty$

Entonces tenemos lo siguiente (casi seguramente):

a) $(\theta_t)_{t\in\mathbb{N}}\subset$ *un compacto* $\mathbb{R}^d$

b) $\exists F_\infty\in L^1$ *tal que* $F(\theta_t)\overset{c.s.}{\underset{t\to\infty}{\longrightarrow}}0$

c) $\nabla F(\theta_t)\overset{c.s.}{\underset{t\to\infty}{\longrightarrow}}0$

Observación 5.2.5.

- El resultado no implica convergencia c.s. de $(\theta_t)_{t\in\mathbb{N}}$.
- (a) $\implies$ toda subsucesión tiene una subsucesión convergente c.s. $\theta_{t_k}\underset{k\to\infty}{\longrightarrow}\tilde{\theta}$
  (b), (c) $\implies$ por continuidad de F y ∇F,

$$F(\theta_{t_k})\underset{k\to\infty}{\longrightarrow}F(\tilde{G}) \quad\wedge\quad \nabla F(\theta_{t_k})\overset{c.s.}{\underset{k\to\infty}{\longrightarrow}}0=\nabla F(\tilde{\theta})\,.$$

  $\therefore$ si $Crit:\{\theta:\nabla F(\theta)=0\}$, $\forall\epsilon>0$, $\mathbb{P}(dist(\theta_t,Crit)>\epsilon)\underset{t\to\infty}{\longrightarrow}0$, i.e., la probabilidad de estar a distancia positiva de puntos críticos tiende a 0.

Esquema de Demostración de Teorema 5.2.4, gradiente estocástico, caso no-convexo. Para más detalles ver Bottou et al. [8].

Usaremos nuevamente el argumento con sobre-martingalas usando una función de Lyapunov ($h(\theta)$). Veremos un confinamiento, o sea que el algoritmo nos llevará a estar en un compacto. Tomamos

$$h(\theta)=(\|\theta\|^2-M)_+^2=\phi(\|\theta\|^2)\,,$$

donde $\phi(r)=(r-M)_+^2$, y tomamos $h_t:=h(\theta_t)\,\forall t\in\mathbb{N}$. Usando la hipótesis (iii) y un Taylor de orden 1 se demuestra que

$$\mathbb{E}(h_{t+1}-h_t|\mathcal{F}_t)\leq -2\gamma_t\theta_t\nabla F(\theta_t)\phi'(\|\theta\|^2)+\gamma_t^2(A+Bh_t)\,,$$

con $\phi'(r) = \begin{cases} 0 & r \leq M \\ 2(r-M) & r > M \end{cases}$.
Luego por Teorema 5.2.2, $h_t \underset{t\to\infty}{\longrightarrow} h_\infty \in L^1$. Entonces

$$\sum_{t=0}^{\infty} \gamma_t \theta_t \nabla F(\theta_t) \phi'(\|\theta_t\|^2) < \infty \quad c.s.\,,$$

y como por (v) tenemos $\sum \gamma_t = \infty$, necesariamente $\liminf_{t\to\infty} \theta_t \nabla F(\theta_t)\phi'(\|\theta_t\|^2) = 0$. Además, por (iv) $\liminf_{t\to\infty} \phi'(\|\theta_t\|^2) = 0 \quad \Longrightarrow \quad h_\infty = 0\, c.s. \quad , \therefore \|\theta_t\| \leq 2M \quad c.s.$ para todo t suficientemente grande (aleatorio).

Para demostrar la convergencia de $F(\theta_t)$ usamos un Taylor de orden 1 para obtener

$$F(\theta_{t+1}) - F(\theta_t) \leq -\gamma_T \partial f(\theta_t, X_{t+1}) \nabla F(\theta_t) + \frac{1}{2}\gamma_t^2 \|\partial f(\theta_t, X_{t+1})\|^2 \partial^2 F(\tilde{\theta}_t)\,,$$

luego obtenemos una sobre-martingala al tomar esperanza condicional con respecto a $\mathcal{F}_t$. Luego definiendo $h_t := F(\theta_t) \quad (\geq cte)$ se demuestra que

$$\mathbb{E}(h_{t+1} - h_t | \mathcal{F}_t) \leq -\gamma_t \|\nabla F(\theta_t)\|^2 + \gamma_t^2 C\,,$$

y nuevamente tenemos que existe $F_\infty \in L^1$ tal que $F(\theta_t) = h_t \underset{t\to\infty}{\longrightarrow} F_\infty$.

Para la convergencia de $\nabla F(\theta_t)$ a 0 notemos que

$$\mathbb{E}(h_{t+1} - h_t | \mathcal{F}_t) \leq -\gamma_t \|\nabla F(\theta_t)\|^2 + \gamma_t^2 C \quad \Longrightarrow \quad \sum_{t=1}^{\infty} \gamma_t \|\nabla F(\theta_t)\|^2 < \infty\,,$$

con lo cual $\liminf_{t\to\infty} \|\nabla F(\theta_t)\|^2 = 0$. Para concluir debemos mostrar que $h_t := \|\nabla F(\theta_t)\|^2$ es convergente. En efecto haciendo un taylor de orden 2 para $\|\nabla F\|^2$ tenemos

$$\|\nabla F(\theta_{t+1})\|^2 - \|\nabla F(\theta_t)\|^2 \leq -2\gamma_t \partial f(\theta, X_{t+1}) Hess(F(\theta_t)) \nabla F(\theta_t) + \gamma_t^2 \|\partial^2 f(\theta, X_{t+1})\| M'''\,,$$

luego

$$\mathbb{E}(h_{t+1} - h_t | \mathcal{F}_t) \leq 2\gamma_t \|\nabla F(\theta_t)\|^2 + \gamma_t^2 M'''\,.$$

Basta notar que tanto $2\gamma_t \|\nabla F(\theta_t)\|^2$ como $\gamma_t^2 M'''$ son convergentes, por ende podemos usar nuevamente el Teorema 5.2.2 para encontrar un h_∞, que necesariamente debe ser 0 por lo anterior, i.e.,

$$h_t = \|\nabla F(\theta_t)\|^2 \underset{t\to\infty}{\longrightarrow} 0\, c.s.$$

□

Observación 5.2.6 (**Posibilidades a tener en mente**). Al correr el algoritmo podemos encontrarnos en los siguientes casos, graficado en la figura 7.

- Mínimo local: Ok, eligiendo el mejor entre muchas corridas.
- Máximo local (u otra cosa): Malo, pero fácil de descartar.

- Mínimo global: No necesariamente deseable pues se corre el riesgo de sobreajustar. Además es “inalcanzable”.
- Plateau asintótico: Muy malo. No deberia ocurrir bajo la hipótesis (iv).

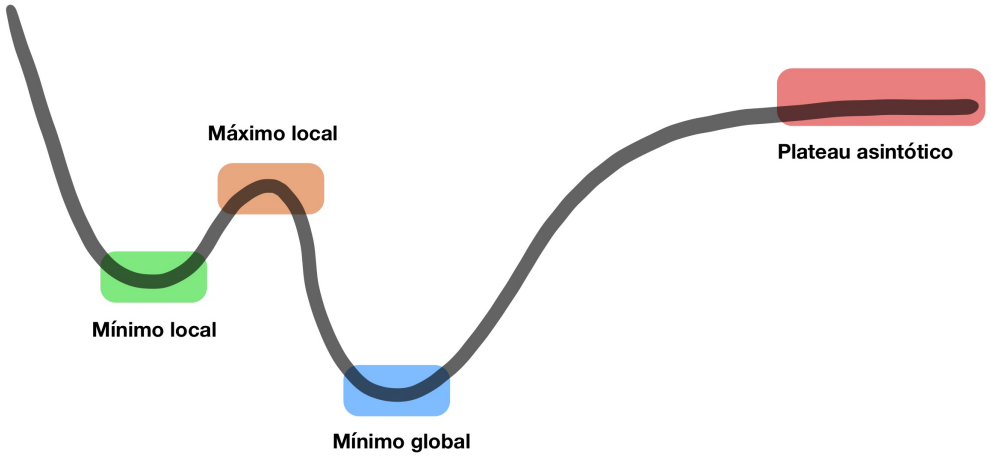


Figura 7: Distintos resultados al aplicar S.G.D.

5.2.4. Tasa de convergencia

Sobre la tasa de convergencia tenemos las siguientes comparaciones entre descenso de gradiente y descenso de gradiente estocástico:

Descenso Gradiente (algoritmo clásico determinista)

- F convexa: $F(\theta_t) - F(\theta^*) = O(\frac{1}{\sqrt{t}})$
- Convexa con ∇F Lipschitz: $F(\theta_t) - \mathcal{F}(\theta^* = O(\frac{1}{t})$
- Fuertemente convexa + ∇F Lipschitz: $F(\theta_t) - F(\theta^*) = O(e^t), \quad e \in (0,1)$

Descenso de gradiente estocástico con pasos $\gamma_t \sim \frac{1}{t}$

- F convexa: $F(\theta_t) - F(\theta^*) = O(\frac{1}{\sqrt{t}})$
- Convexa con ∇F Lipschitz: $F(\theta_t) - \mathcal{F}(\theta^* = O(\frac{1}{\sqrt{t}})$
- Fuertemente convexa + ∇F Lipschitz: $F(\theta_t) - F(\theta^*) = O(\frac{1}{t})$

5.3. Variantes de Gradiente Estocástico

5.3.1. Gradiente estocástico con Mini-Batch

Es una variante muy utilizada en la práctica. Viene del anglicismo Batch que significa "lote".

Idea: en general tenemos

$$\theta_{t+1} = \theta_t - \gamma_t \partial_\theta f(\theta_t, X_{t+1}) = \theta_t - \gamma_t(\nabla F(\theta_t) + \partial_\theta f(\theta_t, X_{t+1}) - \nabla F(\theta_t)\,,$$

donde $\Delta_t = \partial_\theta f(\theta_t, X_{t+1}) - \nabla F(\theta_t)$ es aleatorio y lo interpretamos como ruido, con $\mathbb{E}(\Delta_t) = 0$. Dicho de otro modo, tenemos un "gradiente perturbado".
Notemos que $\partial_\theta f(\theta_t, X_{t+1})$ es un **estimador insesgado** de $\nabla F(\theta)$. Si tomamos esperanza nos da

$$\mathbb{E}(\partial_\theta f(\theta_t, X_{t+1})) = \nabla F(\theta_t)\,.$$

Sin embargo también lo es la siguiente expresión, para m fijo:

$$\frac{1}{m}\sum_{i=1}^{m} \partial_\theta f(\theta_t, X_{t+i})\,.$$

Más aún, con lo anterior estaremos reduciendo varianza.

Descenso de gradiente con mini-batch:
Sea m fijo (tamaño del mini-batch), y $x_1, \ldots, x_t, \ldots$ i.i.d. $\sim \mu$, entonces tomamos:

$$\theta_{t+1} = \theta_t - \gamma_t \cdot \frac{1}{m}\sum_{k=1}^{m} \partial_\theta f(\theta_t, x_{t+i})$$

con

$$Var(\frac{1}{m}\sum_{k=1}^{m} \partial_\theta f(\theta, x_{t+i})) = \frac{Var(\partial_\theta f(\theta, x))}{m} = \frac{\sigma^2}{m}$$

Posee entonces las siguientes ventajas:

- El estimador del gradiente posee mucha menos varianza.
- Lo anterior implica que hay fluctuaciones más pequeñas.
- Aprovecha "vectorización en computadores" (GPU, paralelización, ...).

Observación 5.3.1.

- En el caso conexo fuerte con $\gamma_t = \gamma < \frac{2}{\mu}$ se puede probar que

$$\mathbb{E}(\|\theta_t - \theta^*\|^2) \leq (1 - 2\alpha\mu)^t \|\theta_0 - \theta^*\|^2 + \frac{\gamma\sigma^2}{m2\mu}\,,$$

  donde

$$F(\theta) - F(\eta) \geq \nabla F(\theta)(\theta - \eta) + \frac{\mu}{2}\|\theta - \eta\|^2\,.$$

- m y η_t se deben ajustar conjuntamente.

- "mini-batch": entre el lote completo de datos (Descenso de gradiente usual) y S.G.D. "en línea".
- En la práctica, S.G.D. se usa casi siempre con mini-batch, de tamaño fijo m para aprovechar capacidad de cálculo "paralelo". El costo de un paso con mini-batch es menor a m por el costo de un paso con un dato.
- Tener varianza pequeña no siempre es deseable al comienxo: ayuda a no quedar bloqueado en mínimos locales, por ende tomamos $\gamma_t \searrow$.
- Tampoco es necesario "esperar" a que las fluctuaciones "lleguen a 0" (con $\gamma_t \to 0$).
  Se puede parar fijando un criterio $|F(\theta_{t_k}) - F(\theta_{t_{k+1}})| < \epsilon$.

5.3.2. Más variantes de Gradiente Estocástico

Existen variantes de descenso de gradiente estocástico que explotan la dinámica determinista subyacente para una convergencia más rápida. Si bien en algunos casos tendremos convergencia, muchas veces estos métodos son heurísticas. Los algoritmos mencionados en esta sección están disponibles en las principales bibliotecas de *python* como pytorch, tensorflow y Apache MXNet.

5.3.2.1. Momentum

El método momentum utiliza el gradiente de la etapa anterior mediante la siguiente recurrencia:

$$\theta_{t+1} = \theta_t - \gamma_t m_{t+1}\,,$$

con

$$m_{t+1} = \beta m_t + (1-\beta)\nabla_\theta f(\theta_t, X_{t+1}), \quad m_0 = 0 \quad .$$

Esto mantiene algo de las direcciones de descenso anteriormente usadas ("olvido exponencial").

5.3.2.2. AdaGrad

AdaGrad consiste en dividir el paso por un factor proporcional a la norma de los gradientes acumulados hasta entonces.

$$\theta_{t+1} = \theta_t - \frac{\gamma_t}{\sqrt{v_t + \epsilon}}\nabla_\theta f(\theta, X_{t+1})\,,$$

para ϵ fijo y con

$$v_t = \sum_{k=1}^{t-1} \|\nabla_\theta f(\theta, X_k)\|^2\,, \quad v_1 = 0\,.$$

Este método va haciendo que los pasos sean más chicos si hemos dado pasos grandes, evitando salir de buenos mínimos locales. Por otro lado, Adam (*Adaptive momentum estimation*) es una variación de AdaGrad que considera una ventana de gradientes (con una recursión de mayor complejidad que no discutiremos). En este caso, al llegar a una zona plana este factor aumentará pues los últimos gradientes serán pequeños, de modo que aumenta la tasa y se podría salir hacia mejores valores.

5.3.2.3. Variante Estocástica (Stochastic Gradient Langevin Dynamics)

La siguiente variante "agrega estocasticidad" para explorar de mejor manera el espacio y evitar mínimos locales. Tomamos

$$\theta_{t+1} = \theta_t - \gamma_t \nabla_\theta f(\theta_t, X_{t+1}) + \beta_t \mathcal{N}_t(0, I_d)$$

donde las $\mathcal{N}_t(0, I_d)$ son independientes y representan ruido exógeno. Podemos tomar β_t tendiendo a cero (por ejemplo $\beta_t = \sqrt{2\gamma t}$) o bien $\beta_t = \beta$ constante. Con este último, siempre se estará agregando ruido, lo cual puede ser útil para recorrer más el espacio de soluciones.

5.3.2.4. Otras variantes

Existen más variantes, muchas de las cuales son variantes del método de gradiente determinista adaptadas a gradiente estocástico. Algunas de ellas son:

1. RMSprop
2. NAG (*Nesterov Accelerated Gradient*)
3. AdaDelta
4. Adamax

Hay teoría que indica cuando podría ser mejor uno u otro. En la práctica se recomienda probar con varios para encontrar la mejor alternativa.

5.4. Introducción a las Redes Neuronales

Recordemos el Ejemplo 5.1.2. Queremos aproximar las etiquetas y usando

$$\hat{y}(x, \theta) = \sum_{j=1}^{N} a_j \sigma(w_j^T x + b_j)$$

Cada uno de los elementos $\sigma(w_j^T x + b_j)$ define una función, que se le llama **perceptrón**.

El perceptrón tiene una motivación biológica: se inspira de las neuronas biológicas. En una red neuronal combinaremos la acción de varios perceptrones, de modo que la transmisión de información entre un perceptrón y otro simula la sinapsis entre neuronas del cerebro. En este caso, la señal transmitida será numérica, y el que un perceptrón esté o no activado se modelará usando la función de activación σ, similarmente a como se activaría una neurona. El ejemplo más clásico es una función de activación de Heaviside, que retorna 0 cuando el input es ≤ 0 y 1 en caso contrario.

A la función del ejemplo la llamamos una red de perceptrón de una capa. Analizaremos las capacidades aproximativas de esta arquitectura, que mostramos de manera visual en la figura 8.

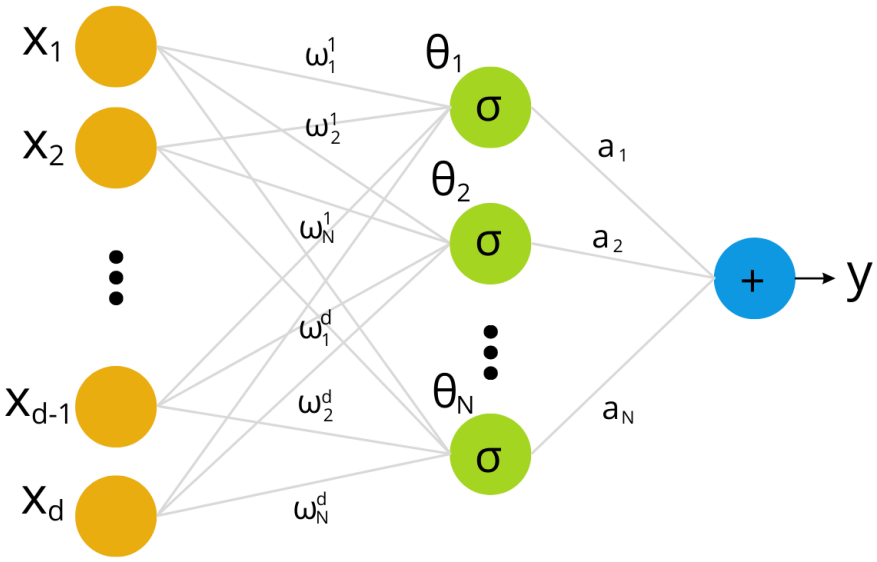


Figura 8: Visualización de una red neuronal de una capa

5.4.1. Teoremas de aproximación universal

La siguiente subsección está basada en el trabajo de Cybenko (1989) [9]. Para la demostración usaremos elementos de Análisis Funcional, como lo son el Teorema de Hahn-Banach y el Teorema de representación de Riesz. Por otro lado, una versión del mismo resultado demostrado en el mismo periodo por Hornik et al. usa el Teorema de Stone-Weirstrass [10].

Definición 5.4.1 Función sigmoide
Decimos que una función σ es sigmoide si cumple

$$\sigma(t) = \begin{cases} 1 & \text{cuando } t \to \infty \\ 0 & \text{cuando } t \to -\infty \end{cases}$$

Ejemplo 5.4.1 Función logística
A la función $\sigma : \mathbb{R} \mapsto (0, 1)$

$$\sigma(t) := \frac{1}{1 + e^{-t}}$$

se le llama función logística. Es el ejemplo más común de una función sigmoide continua.
La activación de Heavisde es también una función sigmoide.

Asumimos el siguiente resultado sin demostración. Está basado en el uso del teorema de convergencia dominada.

Lema 5.4.1
Sea σ una función sigmoide, medible y acotada, entonces cumple la siguiente propiedad:

$$\int_{[0,1]^n} \sigma(y^T x + \theta) d\mu(x) = 0 \quad \forall y \in \mathbb{R}^n, \forall \theta \in \mathbb{R} \quad \Longrightarrow \quad \mu \equiv 0$$

En particular esto es cierto para cualquier función sigmoide continua.

Teorema 5.4.1 Teorema de aproximación universal para activación sigmoide
Sea σ una función sigmoide continua. Entonces el conjunto de funciones de la forma

$$G(x) = \sum_{j=1}^{N} a_j \sigma(w_j^T x + b_j)$$

es denso en $\mathcal{C}([0,1]^n)$, con $j \in \mathbb{R}$, $a_j \in \mathbb{R} \quad \forall j = 1, \ldots, N$ y $w_j \in \mathbb{R}^n \quad \forall j = 1, \ldots, N$. Dicho de otro modo, para todo $f \in \mathcal{C}([0,1]^n)$ y $\epsilon > 0$, $\exists G(x)$ de la forma anterior que cumple

$$|G(x) - f(x)| < \epsilon \quad \forall x \in [0,1]^n \, .$$

Demostración. Sea $S \subset \mathcal{C}([0,1]^n)$ el conjunto de las funciones de la forma $\sum_{i=j}^{N} a_j \sigma(w_j^T x + b_j)$, con $b_j \in \mathbb{R}$, $a_i \in \mathbb{R} \quad \forall j = 1, \ldots, N$ y $w_j \in \mathbb{R}^n \quad \forall j = 1, \ldots, N$, y notemos que es un subespacio lineal de $\mathcal{C}([0,1]^n)$. Veamos que $\bar{S}$ (la cerradura de S) es $\mathcal{C}([0,1]^n)$. Por contradicción, si asumimos lo contrario entonces $\bar{S}$ es un subespacio cerrado propio de $\mathcal{C}([0,1]^n)$. Luego por Hahn-Banach existe un funcional lineal L en $\mathcal{C}([0,1]^n)$, no nulo, pero que se anula en $\bar{S}$ (y por ende en S).

Por el teorema de representación de Riesz, el funcional L debe ser de la forma

$$L(h) = \int_{[0,1]^n} h(x) d\mu(x)$$

para μ medida en $[0,1]^n$ y $\forall h \in \mathcal{C}([0,1]^n)$. Luego en particular se cumple que

$$\int_{[0,1]^n} \sigma(y^T x + \theta) d\mu(x) = 0 \quad \forall y \in \mathbb{R}^n, \forall \theta \in \mathbb{R}$$

. Por lema 5.4.1, $\mu \equiv 0$, con lo cual el funcional lineal L es nulo. Como esto es una contradicción, $\bar{S} = \mathcal{C}([0,1]^n)$, i.e., S es denso en $\mathcal{C}([0,1]^n)$. □

Observación 5.4.1. A las funciones que cumplen la propiedad del lema 5.4.1 se les llama funciones discriminatorias. El resultado vale entonces para cualquier función discriminatoria continua. Además, se puede generalizar a funciones sigmoide medibles y funciones sigmoide arbitrarias agregando ciertas restricciones adicionales.

Más aún, existen variados resultados que proponen cotas para el tamaño de la red neuronal dependiendo de la función a aproximar. La versión del Teorema de Aproximación Universal que hemos visto corresponde al caso de ancho arbitrario, donde ancho se refiere a la cantidad de perceptrones de la capa. Análogamente existen versiones de ancho acotado y profundidad arbitraria, donde profundidad se refiere a la cantidad de capas.

El siguiente resultado nos muestra que el Teorema de aproximación universal nos sirve para implementar un clasificador basado en separaciones del espacio con una red de una sola capa. Primero sea $P_1, \ldots, P_k$, $k \in \mathbb{N}$ una partición de $[0,1]^n$, definiremos una función de decisión como un $f : [0,1]^n \mapsto \{1, \ldots, k\}$ que cumple

$$f(x) = j \quad \Longleftrightarrow \quad x \in P_j \, .$$

Teorema 5.4.2
Sea σ una función sigmoidal continua. Sea f una función de decisión para alguna partición μ-medible y finita, donde μ es la medida de Lebesgue en $[0,1]^n$. Para cualquier $\epsilon > 0$ existe una suma finita de la forma

$$G(x) = \sum_{j=1}^{N} a_j \sigma(w_j^T x + b_j) \, ,$$

y un conjunto $D \subset [0,1]^n$ tal que $\mu(D) \geq 1 - \epsilon$ y que cumple

$$|G(x) - f(x)| < \epsilon \quad \textit{para } x \in D \, .$$

Demostración. Recordemos el Teorema de Lusin, que nos dice que una función finita μ-c.s. es medible si y solo si es una función continua en casi todo su dominio. Luego existe una función continua h y un conjunto D tal que $\mu(D) \geq 1 - \epsilon$ de modo que

$$h(x) = f(x) \quad \forall x \in D \, .$$

Como h es continua, por Teorema 5.4.1, existe una función G de la forma

$$G(x) = \sum_{i=j}^{N} a_j \sigma(w_j^T x + b_j) \, ,$$

con $b_j \in \mathbb{R}$, $a_i \in \mathbb{R} \quad \forall j = 1, \ldots, N$ y $w_j \in \mathbb{R}^n \quad \forall j = 1, \ldots, N$ que satisface

$$|G(x) - h(x)| < \epsilon \quad \forall x \in [0,1]^n \, .$$

Entonces para $x \in D$ tendremos $|G(x) - f(x)| < \epsilon \, .$ □

5.4.2. Entrenamiento de una red neuronal con Gradiente Estocástico

Consideremos el caso en el que tengamos una sucesión de n puntos de datos $(x_1, y_1), \ldots, (x_n, y_n) \subset \mathbb{R}^d \times \mathbb{R}$. Sea además, $l : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ una función de error, y denotemos:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{n} l(y_i, \hat{y}_i(x_i, \theta)) \, .$$

Usaremos el Algoritmo de Descenso de Gradiente Estocástico para minimizar esta función de pérdida. Sin embargo notemos que esto no es tarea fácil, pues involucra el cómputo del gradiente de $\mathcal{L}$, considerando que la dimensionalidad de θ corresponde al número de parámetros totales de la red neuronal (potencialmente muy grande). Estudiemos como se realiza esto en la práctica.

5.4.2.1. Cálculo de gradiente con *Back-propagation*

Back-Propagation es un algoritmo para calcular derivadas en contextos generales, aunque es principalmente usada en el contexto de redes neuronales. Es particularmente útil cuando tenemos cálculos sucesivos, lo cual es el caso de las redes neuronales, más aún si constan de varias capas de profundidad. Veamos un ejemplo para inspirar su uso.

Sean θ el conjunto de parámetros de una red neuronal que queremos ajustar. Sabemos que para Gradiente Estocástico tendremos que saber computar el gradiente en un punto para cualquier parámetro en θ. Consideremos una pérdida cuadrática y una aproximación de un dato y_i de la forma

$$\hat{y}_i = \sum_{j=1}^{N} a_j \sigma(\omega_j^T x_i + b_j)$$

Tratemos de calcular la derivada de $l_i = (y_i - \hat{y}_i)^2$ con respecto a a_k, con $k \in \{1, \dots, N\}$:

$$\frac{\partial l_i}{\partial a_k} = \frac{\partial (y_i - \hat{y}_i)^2}{\partial a_k} = 2(y_i - \hat{y}_i)\frac{\partial (y_i - \hat{y}_i)}{\partial a_k}\,.$$

Notemos que hasta este punto, sólo usamos la regla de la cadena. Ahora como y_i es un punto fijo de dato que no depende de a_k, nos queda:

$$\frac{\partial l_i}{\partial a_k} = 2(y_i - \hat{y}_i)\frac{\partial (y_i - \hat{y}_i)}{\partial a_k} = -2(y_i - \hat{y}_i)\frac{\partial \hat{y}_i}{\partial a_k} = -2(y_i - \hat{y}_i)\frac{\partial [\sum_{j=1}^{N} a_j \sigma(\omega_j^T x_i + b_j)]}{\partial a_k} = -2(y_i - \hat{y}_i)\sigma(\omega_k^T x_i + b_k)$$

Denotaremos ahora las variables auxiliares siguientes. El objetivo será simplificar lo anterior:

- $z_i = (y_i - \hat{y}_i)$
- $u_i^j = a_j \sigma(\omega_j^T x_i + b_j)$
- $v_i^j = \omega_j^T x_i + b_j$

Podemos reescribir el cálculo de la última linea usando las variables auxiliares:

$$\frac{\partial l_i}{\partial a_k} = 2(y_i - \hat{y}_i)\frac{\partial [\sum_{j=1}^{N} a_j \sigma(\omega_j^T x_i + b_j)]}{\partial a_k} = 2z_i \frac{\partial [\sum_{j=1}^{N} u_i^j]}{\partial a_k} = 2z_i \frac{\partial u_i^k}{\partial a_k}\,.$$

En este punto podemos notar que

$$\frac{\partial u_i^k}{\partial a_k} = \frac{\partial [a_k \sigma(v_i^j)]}{\partial a_k} = \sigma(v_i^j) \quad \therefore \quad \frac{\partial l_i}{\partial a_k} = 2z_i \sigma(v_i^j)$$

Si bien agregar las variables auxiliares nos hace pesada la notación, la idea es guardar la información paso por paso a medida que los valores "avanzan" en la red neuronal hasta llegar a la predicción. Esto a la larga será útil, pues se nos facilitan los cálculos, como vimos en derivar $\frac{\partial l_i}{\partial a_k}$. Por otro lado notemos que podemos llegar a l_i de manera acumulativa desde los parámetros usando las variables auxiliares.

- $l_i = z_i^2$
- $z_i = (y_i - \sum_{j=1}^{N} u_i^j)$

- $u_i^j = a_j \sigma(v_i^j)$
- $v_i^j = \omega_j^T x_i + b_j$

Luego usando la regla de la cadena podemos deducir que

$$\frac{\partial l_i}{\partial a_k} = \frac{\partial l_i}{\partial z_i}\frac{\partial z_i}{\partial u_i^j}\frac{\partial u_i^j}{\partial a_k} = 2z_i \cdot 1 \cdot \sigma(v_i^j)$$

i.e., hicimos el mismo cálculo anterior pero basado en cálculos simples de derivadas usando las variables auxiliares. Veamos otro ejemplo. Calcularemos $\frac{\partial l_i}{\partial b_k}$ usando este principio:

$$\frac{\partial l_i}{\partial b_k} = \frac{\partial l_i}{\partial z_i}\frac{\partial z_i}{\partial u_i^j}\frac{\partial u_i^j}{\partial v_i^j}\frac{\partial v_i^j}{\partial b_k} = 2z_i \cdot 1 \cdot a_k \sigma'(v_i^j) \cdot 1 = 2(y_i - \sum_{j=1}^{N} u_i^j) a_k \sigma'(\omega_j^T x_i + b_j)$$

donde σ' es la derivada de la función de activación σ (unidimensional).

Notemos la facilidad con la que llegamos a la expresión final, y más aún, notemos que muchas de las derivadas parciales que usamos también se usaron para computar $\frac{\partial l_i}{\partial a_k}$. Se puede hacer un cálculo análogo para llegar a $\frac{\partial l_i}{\partial \omega_k}$, con lo cual tendríamos calculado el gradiente de l_i.

El algoritmo Back-Propagation consiste en el uso exhaustivo de la **regla de la cadena**. Calculamos derivadas parciales paso por paso y las guardamos para usarlas en cada parámetro. De este modo que sólo computaremos derivadas simples y guardarlos evitará hacer cálculos redundantes. En nuestro caso no era imposible calcular el gradiente, pero a medida que las redes neuronales sean más profundas, más profundas serán las dependencias entre variables y por ende más cálculos se ahorrarán si usamos regla de la cadena, además de simplificar el computo.

Las bibliotecas que implementan redes neuronales, como *pytorch* implementan este algoritmo implícitamente, de modo que al definir las operaciones entre variables, se guardan las derivadas parciales respectivas. Durante el entrenamiento, una vez que la información pasa a través de la red neuronal para computar una predicción, se computará un error (*forward*). Enseguida, se computa el gradiente usando las operaciones que estaban guardadas, haciendo fluir la información en reversa hasta tener el valor de cada derivada parcial (*backward*). Esto nos da una manera eficiente de usar descenso de gradiente estocástico.

5.4.2.2. Problemas y prácticas usuales

Hasta ahora hemos observado las capacidades teóricas de las redes neuronales de una sola capa. Si bien poder aproximar funciones continuas parece una capacidad muy valiosa, no siempre estaremos intentando aproximar funciones con buenas propiedades. Es más, muchas veces no tendremos ninguna garantía de que las funciones que estamos tratando de aproximar existan. Tener un cierto grado de profundidad nos ayuda a aumentar el poder de expresividad de los valores de salida. A su vez podremos disminuir el número de perceptrones en cada capa.

Lamentablemente, un problema que se encuentra al entrenar redes neuronales profundas es el **desvanecimiento del gradiente**. Esto sucede cuando el valor de las derivadas parciales es pequeño, evitando que puedan actualizarse valores de capas iniciales usando descenso de gradiente.

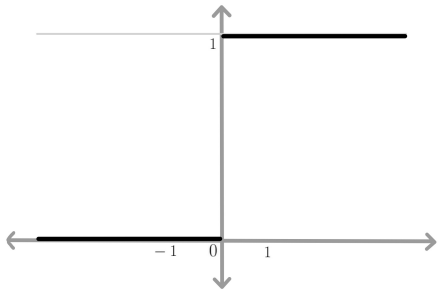
Dentro de las soluciones que se han explorado para evitar este problema está el uso de la función de activación *ReLU* (*rectified linear unit*), que está dada por:

$$ReLU(t) = \begin{cases} t & \text{cuando } t \geq 0 \\ 0 & \text{cuando } t < 0 \end{cases}$$

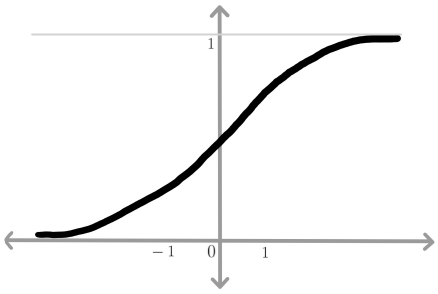
Si bien esta función no es sigmoide y es no-acotada, es mucho más fácil de computar y es invariante a la escala. El problema de que no sea derivable en 0 se arregla simplemente asignando 1 o 0 como valor de la derivada en aquel punto. Por último notar que podemos lograr funciones que si son sigmoides al acoplar dos funciones *ReLU* o más. Un ejemplo de esto es la función σ siguiente:

$$\sigma(t) = \frac{1}{2}(ReLU(t+1) - 1 - (ReLU(t-1) - 1)),$$

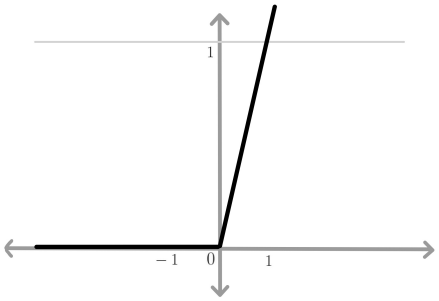
que es una función sigmoide continua, y por ende cumple con las hipótesis del Teorema 5.4.1. En la figura 9 se muestran algunas de las funciones de activación que se han mencionado en esta subsección.



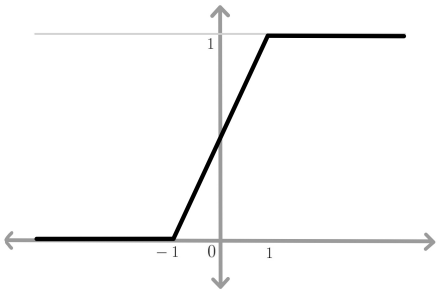
(a) Activación de Heaviside



(b) Logística



(c) $ReLU(t)$



(d) $\frac{1}{2}(ReLU(t+1) - 1 - (ReLU(t-1) - 1))$

Figura 9: Ejemplos de funciones de activación

6. Movimiento Browniano y difusiones

Este capítulo tiene como objetivo introducir conceptos que se abordan de manera más exhaustiva en el curso de cálculo de estocástico. Más precisamente, veremos el movimiento Browniano, objeto central para procesos estocásticos y que tiene variadas aplicaciones.

Nos interesará la simulación del movimiento Browniano y con ellos procesos conducidos por él, que son a su vez representaciones microscópicas de lo que describen ciertas ecuaciones derivadas parciales. Con esto obtendremos métodos estocásticos para la resolución de ciertas EDPs, los cuales son competitivos con métodos deterministas en dimensión alta.

6.1. Movimiento Browniano

Recordemos la definición 1.3.1. Un vector se dice Gaussiano si toda combinación de sus coordenadas es v.a. Gaussiana. Además llamamos media de X al vector

$$\mu = \mathbb{E}(X) = (\mathbb{E}(X^i))_{i=1}^n$$

y matriz de varianza-covarianza a

$$\Sigma = \mathbb{E}((X-\mu)(X-\mu)^T) = (Cov(X^i, X^j))_{i,j=1,1}^n .$$

Recordemos además que podemos simular una variable $X \sim \mathcal{N}(\mu, \Sigma)$ usando distribuciones normales estándar usando que

$$X \overset{ley}{=} \mu + \Gamma \mathcal{N}(0, I_d)$$

donde Γ cumple $\Gamma\Gamma^T = \Sigma$.

Proposición 6.1.1
Sea X vector Gaussiano. Tenemos las siguientes propiedades.

- *Sea $J \subseteq \{1, \dots, n\}$, el subvector $\tilde{X} = (X^j)_{j\in J}$ también es vector Gaussiano.*
- *Por definición, $\langle \lambda, X\rangle$ es un vector Gaussiano:*

$$\langle \lambda, X\rangle \sim \mathcal{N}\Big(\langle \lambda, \mu\rangle, \frac{\lambda^T \Sigma \lambda}{2}\Big)$$

  y su función característica está dada por

$$\mathbb{E}\Big(e^{i\langle \lambda, X\rangle}\Big) = e^{i\langle \mu, \lambda\rangle - \frac{\lambda^T\Sigma\lambda}{2}} .$$

- *Un subvector $\tilde{X} = (X^j)_{j\in J}$ tiene coordenadas independientes ssi $Cov(X^i, X^j) = 0 \quad \forall i, j \in J$, $j \neq i$*
- *Si Σ no es degenerada, X tiene densidad*

$$\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} exp\Big(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\Big) \quad \text{para } x \in \mathbb{R}^n$$

Definición 6.1.1 Proceso Gaussiano
Sea $\mathbb{T}$ un conjunto de índices. Un proceso Gaussiano $(X_t)_{t\in\mathbb{T}} \subseteq \mathbb{R}^{\mathbb{T}}$ de función de medias $m = (m_t)_{t\in\mathbb{T}} \in \mathbb{R}^P i$ y función de varianza-covarianzas $K = (K_{t,s})_{t,s\in\mathbb{T}}$ (que es simétrica ≥ 0) es una familia $X_t : (\Omega, \mathcal{F}, \mathbb{P}) \mapsto (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ de variables aleatorias tal que $\forall S \subseteq \mathbb{T}$ finito $(X_t)_{t\in S}$ es vector Gaussiano con media $(m_t)_{t\in S}$, y matriz de covarianza $(K_{t,s})_{t,s\in S}$.

Teorema 6.1.1 Existencia del proceso Gaussiano
Dados m y K como en la definición 6.1.1, $\exists (X_t)_{t\in\mathbb{T}}$ proceso Gaussiano con media m y varianza-covarianza K. Además su ley en $\mathbb{R}^{\mathbb{T}}$ es única.

La demostración consiste en utilizar el teorema de consistencia de Kolmogorov.

Definición 6.1.2 Movimiento Browniano
Un proceso $(B_t : t \geq 0)$ definido en un espacio de probabilidad completo $(\Omega, \mathcal{F}, \mathbb{P})$ es un movimiento Browniano real si es un proceso Gaussiano con

- $\mathbb{E}(B_t) = 0 \quad \forall t \geq 0$
- $\mathbb{E}(B_t B_s) = Cov(B_t, B_s) = t \vee s$
- La aplicación $t \mapsto B_t(\omega)$ es continua $\mathbb{P}(d\omega)$ c.s. .

La siguiente es una definición equivalente:

Definición 6.1.3 Movimiento Browniano, definición equivalente
Un movimiento browniano es un proceso estocástico (es decir, una colección de variables aleatorias indexadas por $t \geq 0$) denotado $(B_t)_{T\geq 0}$, que satisface:

i) $B_0 = 0$ casi seguramente.

ii) Tiene **incrementos independientes**: $\forall t \geq 0$,

$$(B_{t+s} - B_t)_{s\geq 0}, \text{ es independiente de } (B_s)_{0\leq s\leq t}\,.$$

iii) Tiene **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 $t \mapsto B_t$ es continua en t.

La ley del proceso $(B_t^{(a)})_{t\geq 0}$ converge, cuando $a \to \infty$, a la ley del movimiento browniano $(B_t)_{t\geq 0}$, en un sentido adecuado (esto se llama *teorema de Donsker*). Por otra parte, $(B_t)_{t\geq 0}$ también se conoce como **Proceso de Wiener**. El nombre *browniano* proviene del botánico Robert Brown, quien observó el movimiento errático de partículas de polen en el agua. Posteriormente, Einstein explicó este movimiento como resultado de muchas pequeñas colisiones del polen con las moléculas de agua circundantes.

Proposición 6.1.2
$\forall\ 0 = t_0 < t_1 < \ldots < t_n$ *, la colección*

$$\left(\frac{B_{t_i} - B_{t_{i-1}}}{\sqrt{t_i - t_{i-1}}}\right)_{i=1,\ldots,n} \textit{ es i.i.d } \sim \mathcal{N}(0,1)\,.$$

Demostración. Directa de la definición.

Esto da lugar a un método sencillo para generar un movimiento browniano discretizado: para un horizonte $T > 0$ y $n \in \mathbb{N}$, sea $\Delta t = \frac{T}{n}$, y $t_i = i\Delta t$. Dadas $Z_1, \ldots, Z_n$ i.i.d. $\mathcal{N}(0,1)$, entonces el proceso $(Y_t)_{t\in[0,T]}$:

$$Y_{t_i} := \sqrt{\Delta t} \sum_{j=1}^{i} Z_j$$

interpolado linealmente entre t_i's, es una aproximación de $(B_t)_{t\in[0,T]}$ (de hecho, tiene exactamente la misma ley en la malla $0 = t_0 < t_1 < \ldots < t_n = T$).

Proposición 6.1.3
Se tienen las siguientes proposiciones:

1. $\forall t_0 > 0$, *el proceso* $X_t := B_{t+t_0} - B_{t_0}$ *es un movimiento browniano.*
2. $\forall c > 1$, *el proceso* $X_t := \frac{1}{\sqrt{c}} B_{ct}$ *es un movimiento browniano.*
3. *El proceso* $X_t := B_1 - B_{1-t}$ *es un movimiento browniano en* $[0,1]$.
4. *El proceso* $X_t := tB_{\frac{1}{t}}$ *es un movimiento browniano.* ($X_0 = 0$)
5. *El proceso* $X_t := -B_t$ *es un movimiento browniano.*

Demostración. Probemos 2, el resto queda propuesto como Ejercicio.
Debemos probar que $X_t = \frac{1}{\sqrt{c}} B_{ct}$ cumple con las condiciones de la definición (6.1.3). Es claro que las condiciones I,II y IV son directas del hecho que $(B_t)_{t\geq 0}$ es un movimiento browniano. Además:

$$X_{t+s} - X_t = \underbrace{\frac{1}{\sqrt{c}} \left(B_{ct+cs} - B_{ct}\right)}_{\mathcal{N}(0,cs)} \sim \mathcal{N}(0,s)\,.$$

□

A pesar de que son continuas, las **trayectorias** de un movimiento browniano son bastante irregulares. La trayectoria de un movimiento browniano es la función aleatoria:

$$t \mapsto B_t$$

Teorema 6.1.2

$$\mathbb{P}(\exists t \geq 0 \text{ tal que } B_t \text{ es derivable en } t) = 0\,.$$

La demostración de dicho teorema escapa a los contenidos del curso.

El comportamiento de las **oscilaciones** del movimiento browniano cerca de $t = 0$ y $t = \infty$ queda descrito por el siguiente resultado:

Teorema 6.1.3 Ley del Logaritmo Iterado
$\mathbb{P}$- *casi seguramente se cumple:*

1. $\limsup_{t\searrow 0} \frac{B_t}{\sqrt{2t\log\log 1/t}} = 1$
2. $\liminf_{t\searrow 0} \frac{B_t}{\sqrt{2t\log\log 1/t}} = -1$
3. $\limsup_{t\to\infty} \frac{B_t}{\sqrt{2t\log\log t}} = 1$
4. $\liminf_{t\to\infty} \frac{B_t}{\sqrt{2t\log\log t}} = -1$

La demostración de este resultado utiliza herramientas fuera del alcance del curso. Gráficamente 1 y 2 significa que la trayectoria de $(B_t)_{t\geq 0}$ cruza o pasa cerca del gráfico de la función $f(t) = \sqrt{2t\log\log(t)}$ infinitas veces cuando $t \to \infty$, y lo mismo para $-f(t)$.

Proposición 6.1.4

$$\mathrm{Cov}(B_t, B_s) = \min(t, s)\,.$$

Demostración. Para $0 \leq s \leq t$ tenemos:

$$\begin{aligned}
\mathrm{cov}(B_t, B_s) &= \mathbb{E}(B_t B_s) - \underbrace{(\mathbb{E}(B_t))(\mathbb{E}(B_s))}_{=0} \\
&= \mathbb{E}\underbrace{((B_t - B_s)B_s)}_{\text{indep.}} + \mathbb{E}(B_s) \\
&= \underbrace{(\mathbb{E}(B_t - B_s))}_{=0} + \underbrace{Var(Bs)}_{s} \\
&= s = \min(s, t)
\end{aligned}$$

□

6.2. Martingalas

Para definir lo que es una martingala y otros conceptos relacionados se requiere del siguiente concepto:

Definición 6.2.1 Filtración
Dado un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$, una **filtración** es una colección $\mathbb{F} = (\mathcal{F}_t)_{t\geq 0}$ de sub σ-álgebras de $\mathcal{F}$ que es **creciente**, es decir, $\mathcal{F}_t \subseteq \mathcal{F}_s$, $\forall s \leq s$.

$\mathcal{F}_t$ representa la **información** disponible en momento t, es decir, la colección de eventos de $\mathcal{F}$ que pueden definirse a partid e la información hasta t.

Dado un proceso estocástico $(X_t)_{t\geq 0}$, definimos su **filtración natural** como

$$\mathcal{F}_t^X := \sigma\left(\{X_s : s \leq t\}\right)$$

Decimos que un proceso $(X_t)_{t\geq 0}$ es **adaptado** a una filtración $(\mathcal{F}_t)_{t\geq 0}$ si X_t es $\mathcal{F}_t$ medible $\forall t$. En general se asume que los procesos con los que se trabaja son adaptados. Evidentemente, $(X_t)_{t\geq 0}$ es $(\mathcal{F}_t^X)_{t\geq 0}$ - adaptado.

Cuando trabajamos con un proceso (por ejemplo un movimiento browniano y no se menciona la filtración, esta implícito que se usa la filtración natural.

Para comprender el concepto de martingala es provechoso reflexionar sobre el siguiente ejemplo: si X_t representa la riqueza de una persona en el instante t, la cual evoluciona de acuerdo a reglas aleatorias de un cierto *juego* (ejemplo: apuestas, fluctuaciones de la bolsa), entendemos que el juego es **justo** si la riqueza futura no crece ni decrece, en esperanza.

Definición 6.2.2 Martingala
Dado un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ con filtración $\mathbb{F} = (\mathcal{F}_t)_{t\geq 0}$, una **martingala** es un proceso $(X_t)_{t\geq 0}$ adaptado tal que:

- $\forall t \geq 0$, $X_t \in L^1$, es decir, $\mathbb{E}|X_t| < \infty$.
- $\forall s \leq t$,
$$\mathbb{E}\left(X_t \mid \mathcal{F}_s\right) = X_s, \quad \text{casi seguramente}\,.$$

Notemos que entonces $\mathbb{E}(X_t) = \mathbb{E}(X_0)$ que es constante para todo t.

Cuando hay una filtración subyacente, se hace necesario re-definir el concepto de movimiento browniano:

Definición 6.2.3 Movimiento Browniano para un espacio filtrado
Dado un espacio de probabilidad filtrado $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, un **movimiento browniano** es un proceso $(B_t)_{t\geq 0}$ adaptado que satisface:

i) $B_0 = 0$, casi seguramente.

ii) $\forall t \geq 0$, $(B_{t+s} - B_t)_{s\geq 0} \perp\!\!\!\perp \mathcal{F}_t$.

iii) $B_{t+s} - B_t \sim \mathcal{N}(0, s)$, $\forall t, s \geq 0$.

iv) Posee trayectorias continuas.

Si no se especifica la filtración, se asuma la filtración natural (es decir, la definición original).

Proposición 6.2.1
Dado $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ y un movimiento browniano $(B_t)_{t\geq 0}$, entonces:

1. *$(B_t)_{t\geq 0}$ es una martingala.*
2. *$\left(B_t^2 - t\right)_{t\geq 0}$ es una martingala.*
3. *$\exp\left(\sigma B_t - \frac{\sigma^2}{2}t\right)$ es una martingala , $\forall \sigma > 0$.*

Demostración. Comenzamos por 1:

$$\begin{aligned}\mathbb{E}(B_t \mid \mathcal{F}_s) &= \mathbb{E}(B_t - B_s \mid \mathcal{F}_s) + \mathbb{E}(B_s \mid \mathcal{F}_s) \\ &= \mathbb{E}(B_t - B_s) + B_s \\ &= B_s\,.\end{aligned}$$

Por otra parte, para 2:

$$\begin{aligned}\mathbb{E}(B_t^2 \mid \mathcal{F}_s) &= \mathbb{E}\left((B_t - B_s) \mid \mathcal{F}_s\right) + \mathbb{E}(B_s \mid \mathcal{F}_s) - 2\mathbb{E}\left((B_t - B_s)B_s \mid \mathcal{F}_s\right) \\ &= \mathbb{E}\left((B_t - B_s)^2\right) + B_s^2 - 2\mathbb{E}\left((B_t - B_s) \mid \mathcal{F}_s\right) \\ &= t - s + B_s^2\,.\end{aligned}$$

Finalmente para 3, recordemos que si $X \sim \mathcal{N}(\mu, \tau^2)$, entonces

$$\mathbb{E}(e^{\lambda X}) = \exp\{\lambda\mu + \frac{\lambda^2\tau^2}{2}\}\,.$$

Con esto

$$\begin{aligned}\mathbb{E}\left(e^{\sigma B_t} \mid \mathcal{F}_s\right) &= \mathbb{E}\left(e^{\sigma(B_t - B_s)}e^{\sigma_s} \mid \mathcal{F}_s\right) \\ &= e^{\sigma B_s}\mathbb{E}\left(e^{\sigma(B_t - B_s)} \mid \mathcal{F}_s\right) \\ &= e^{\sigma B_s}\mathbb{E}\left(e^{\sigma(B_t - B_s)}\right) \\ &= e^{\sigma B_s}e^{\sigma^2(t-s)/2}\,.\end{aligned}$$

□

6.3. Tiempos de Parada

Buscamos definir tiempo aleatorios que **no utilicen información futura**, solamente lo que a ocurrido hasta el momento.

Definición 6.3.1 Tiempo de parada
Dado un espacio de probabilidad filtrado $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, un **tiempo de parada** es una variable aleatoria $T \in [0, \infty]$ tal que

$$\{T \leq t\} \in \mathcal{F}_t, \quad \forall t\,.$$

Ejemplo 6.3.1
Para ilustrar la definición anterior, se estudian los siguientes ejemplos:

- Si se define $T_b := \inf\{t \geq 0 : B_t = b\}$ como *la primera vez que B llega a b*, se tiene entonces que T_b es un tiempo de parada.
- Si se define $S_0 := \{t \in [0, 1] : B_t = 0\}$ como *la última vez antes de $t = 1$ que B toca a 0*. Entonces S_0 no es tiempo de parada.

Demostración. Veamos que T_b es un tiempo de parada:

$$\begin{aligned}\{T_b \leq t\} &= \{\exists s \in [0,1] : B_s = b\} \\ &= \{\forall n \in \mathbb{N}, \exists r \in [0,1] \cap \mathbb{Q}, |B_r - v| \leq \frac{1}{n}\} \\ &= \bigcap_{n\in\mathbb{N}} \bigcup_{r\in[0,1]\cap\mathbb{Q}} \{|B_r - b| \leq \frac{1}{n}\} \in \mathcal{F}_t\,.\end{aligned}$$

Lo que comprueba dicha afirmación. □

Surge la pregunta, cuando $(X_t)_{t\geq 0}$ es una martingala (es decir, un juego justo). Si uno detiene el juego en un tiempo aleatorio con la información disponible hasta el momento, ¿será posible obtener un beneficio de aquello?

Bajo cierta condición en el tiempo de parada, la respuesta será **no**. Para especificar lo anterior, necesitamos definir la σ-álgebra correspondiente a la información disponible hasta cierto tiempo de parada.

Definición 6.3.2 σ-álgebra asociada a un tiempo de parada
Dado $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ y un tiempo de parada T, la σ-álgebra asociada a T es:

$$\mathcal{F}_T := \{A \in \mathcal{F} : \forall t \geq 0,\ A \cap \{T \leq t\} \in \mathcal{F}_t\}$$

Ejercicio: Probar que $\mathcal{F}_T$ es σ-álgebra.

Teorema 6.3.1 Muestreo opcional de Doob
Dados $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ y una martingala $(X_t)_{t\geq 0}$, sean T, S tiempos de parada tales que $S \leq T \leq K$ con $K > 0$ una constante. Entonces

$$\mathbb{E}\left[X_t \mid \mathcal{F}_s\right] = X_s, \quad c.s..$$

La demostración de este teorema utiliza herramientas fuera del alcance del curso.

El resultado anterior es útil para algunos cálculos sobre tiempos de parada. La condición de T acotado suele ser demasiado restrictiva. Para T tiempo de parada no acotado, se trabaja con $\min(T, n)$ y luego se hace $n \to \infty$. Esta técnica se denomina **localización**. Veamos un ejemplo:

Proposición 6.3.1
Para $a \in \mathbb{R}$, sea $T_a = \inf\{t \geq 0 : B_t = a\}$. Entonces $T_a < \infty$ casi seguramente y su ley viene dada por su transformada de Laplace.

$$\mathbb{E}\left(e^{-\lambda T_a}\right) = e^{-\sqrt{2\lambda}|a|}\,,$$

o equivalente, su densidad es

$$f_{T_a} = \frac{|a|}{\sqrt{2\pi x^3}} \exp\Big(-a^2/2x\Big), \quad \forall x > 0\,.$$

Demostración. Sea $a > 0$. Sea además la martingala $X_t = \exp\Big(\sigma B_t - \frac{\sigma^2}{2}t\Big)$ y usemos el teorema al tiempo de parada acotado $\min(T_a, n)$, para $n \in \mathbb{N}$. Luego $\mathbb{E}(X_{\min(T_a,n)}) = \mathbb{E}(X_0) = 1$. Queremos hacer $n \to \infty$. Notemos que:

- $X_{\min(T_a,n)} = \exp\left(\sigma B_{\min(T_a,n)} - \frac{\sigma^2}{2}\min(T_a,n)\right) \leq e^{\sigma a}$, es decir, la sucesión $(X_{\min(T_a,n)})_{n\in\mathbb{N}}$ está dominada por $e^{\sigma a}$.
- En $\{T_a < \infty\}$, $X_{\min(T_a,n)} \underset{n}{\to} X_{T_a}$
- En $\{T_a = \infty\}$, $X_{\min(T_a,n)} = X_n = \exp\left(\sigma B_n - \frac{\sigma^2 n}{2}\right) \underset{n}{\to} 0$.

Luego, usando el teorema de convergencia dominada:

$$\begin{aligned}
1 &= \lim_n \mathbb{E}(X_{\min(T_a,n)}) \\
&\overset{\text{TCD}}{=} \mathbb{E}\left(\lim_n X_{\min(T_a,n)}\right) \\
&= \mathbb{E}\left(\lim_n 1_{\{T_a<\infty\}} X_{\min(T_a,n)} + \lim_n 1_{\{T_a=\infty\}} X_{\min(T_a,n)}\right) \\
&= \mathbb{E}\left(1_{\{T_a<\infty\}} X_{T_a}\right) = \mathbb{E}\left(1_{\{T_a<\infty\}} \exp\left(\sigma B_{T_a} - \frac{\sigma^2 T_a}{2}\right)\right).
\end{aligned}$$

Es decir,

$$\mathbb{E}\left[1_{\{T_a<\infty\}} \exp\left(-\sigma^2 T_a/2\right)\right] = e^{-\sigma a}$$

Haciendo $\sigma \to 0$, se obtiene $\mathbb{P}(T_a < \infty) = 1$. Tomando $\sigma = \sqrt{2\lambda}$, se llega a lo deseado. El caso $a < 0$ se deduce directamente pues $-B_t$ es también un movimiento browniano. □

La siguiente desigualdad también es útil:

Teorema 6.3.2 Desigualdad de Doob
Si $(X_t)_{t\in[0,T]}$ es una martingala continua, entonces

$$\mathbb{E}\left[\sup_{0\leq t\leq T} |X_t|^2\right] \leq 4\mathbb{E}\left[|X_T|^2\right].$$

6.4. Integral Estocástica y Cálculo de Itô

Como motivación consideremos $(X_t)_{t\geq 0}$ un proceso con trayectorias continuas y derivables. Se puede definir la **integral** de f con respecto a X como

$$\int_0^t f(s)dX_s := \int_0^t f(s)\frac{dX_s}{ds}ds\,.$$

Lamentablemente, no podemos hacer los mismo con un movimiento browniano $(B_t)_{t\geq 0}$, pues $\frac{dB_t}{dt}$ **no existe**. Luego, si queremos definir

$$\int_0^t f(s)dB_s$$

tendremos que seguir un enfoque distinto. Integrales de este tipo son muy útiles; por ejemplo, para definir ecuaciones diferenciales estocásticas.

Permitiremos que el integrando $f(s)$ también sea un proceso, i.e., definiremos

$$\int_0^t X_s dB_s \, .$$

6.4.1. Construcción

Sea $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ un espacio filtrado, sea $(B_t)_t t \geq 0$ un movimiento browniano con respecto a este espacio. En primer lugar, definiremos la integral con respecto a B para una cierta clase de procesos, llamados **procesos simples predecibles**:

Definición 6.4.1 Proceso Simple Predecible
Un proceso $(H_t)_{0\leq t\leq T}$ se dice simple predecible si se escribe de la forma

$$H_t = \sum_{i=1}^{k} A_i 1_{\{t_{i-1}, t_i\}}(t) \, .$$

para ciertos instantes $0 = t_0 < t_1 < \ldots, t_k = T$, $(A_i)_{i=1}^k$ son variables aleatorias $\mathcal{F}_{t_{i-1}}$ - medibles, y tales que H cumple $\mathbb{E}(\int_0^T H_s^2 ds) < \infty$.

Definición 6.4.2 Integral Estocástica
Dado un tal H, la **integral estocástica** de H con respecto a B es el proceso $(I(H)_t)_{t\in[0,T]}$. Definido como

$$I(H)_t = \sum_{i=1}^{k} A_i \left(B_{\min(t_i, n)} - B_{\min(t_{i-1}, t)} \right) =: \int_0^t H_s dB_s \, .$$

Es decir, para $t \in (t_j, t_{j+1}]$

$$I(H)_t = \sum_{i=1}^{j} A_i \left(B_{t_i} - B_{t_{i-1}} \right) + A_{j+1}(B_t - B_{t_j}) \, .$$

Notar que $I(H)_t$ es continuo en t (ejercicio). $I(H)_t$ corresponde a la trayectoria de B_t, donde cada tramo está ponderado por el A_i correspondiente.

La idea ahora es extender esta definición a una clase más general de procesos. Para ello, la siguiente proposición es fundamental.

Proposición 6.4.1
Sea $(H_t)_{t\in[0,T]}$ un proceso simple predecible. Entonces

1. *$(I(H)_t)_{t\in[0,T]}$ es una martingala continua.*
2. $\mathbb{E}\left(\left[\int_0^t H_s dB_s \right]^2 \right) = \mathbb{E}\left(\int_0^t H_s^2 ds \right)$
3. $\mathbb{E}\left(\sup_{0\leq t\leq T} \left| \int_0^t H_s dB_s \right|^2 \right) \leq 4\mathbb{E}\left(\int_0^T H_s^2 ds \right)$

Demostración. Para demostrar 1 se requiere que $\forall s < t$

$$\mathbb{E}\left(\int_0^t H_u dB_u \mid \mathcal{F}_s \right) = \int_0^s H_u dB_u \, .$$

Sin pérdida de generalidad, podemos asumir que $s = t_l < t_m < t$ para ciertos l, m (en caso contrario, t y s se pueden incluir en la partición $0 = t_0 < t_1, \ldots, t_k = T$ y el proceso H obtenido seguirá siendo simple predecible). Tenemos:

$$\mathbb{E}\left[\int_0^{t_m} H_u dB_u \mid \mathcal{F}_{t_l}\right] = \sum_{i=1}^{m} \mathbb{E}\left(A_i(B_{t_i} - B_{t_{i-1}}) \mid \mathcal{F}_{t_l}\right) .$$

Para $i \geq l+1$: $\mathcal{F}_{t_{i-1}} \supseteq \mathcal{F}_{t_l}$, luego

$$\begin{aligned}\mathbb{E}\left[A_i(B_{t_i} - B_{t_{i-1}}) \mid \mathcal{F}_{t_l}\right] &= \mathbb{E}\left[\mathbb{E}(A_i(B_{t_i} - B_{t_{i-1}}) \mid \mathcal{F}_{t_{i-1}} \mid \mathcal{F}_{t_l}\right] \\ &= \mathbb{E}\left[A_i \mathbb{E}(B_{t_i} - B_{t_{i-1}}) \mid \mathcal{F}_{t_l}\right] = 0 .\end{aligned}$$

Luego en $\sum_{i=1}^{m} \mathbb{E}\left(A_i(B_{t_i} - B_{t_{i-1}}) \mid \mathcal{F}_{t_l}\right)$ quedan sólo los términos $i \leq l$, y la esperanza condicional desaparece pues los términos dentro son $\mathcal{F}_{t_l}$-medibles. Luego

$$\mathbb{E}\left[\int_0^{t_m} H_u dB_u \mid \mathcal{F}_{t_l}\right] = \sum_{i=1}^{l} A_i(B_{t_i} - B_{t_{i-1}}) = \int_0^{t_l} H_u dB_u .$$

Para 2, sin pérdida de generalidad se asume nuevamente $t = t_m$ para cierto m. Se tiene así:

$$\begin{aligned}&\mathbb{E}\left(\left[\int_0^t H_s dB_s\right]^2\right) = \mathbb{E}\left(\left[\sum_{i=1}^{m} A_i(B_{t_i} - B_{t_{i-1}}\right]^2\right) \\ &= \sum_{i=1}^{m} \mathbb{E}\left[A_i^2 (B_{t_i} - B_{t_{i-1}})^2\right] + 2\sum_{i<j}^{m} \mathbb{E}\left[A_i A_j (B_{t_i} - B_{t_{i-1}})(B_{t_j} - B_{t_{j-1}})\right] .\end{aligned}$$

Por independencia: $\mathbb{E}[A_i^2 (B_{t_i} - B_{t_{i-1}})^2] = \mathbb{E}[A_i^2]\mathbb{E}\left[(B_{t_i} - B_{t_{i-1}})^2\right] = \mathbb{E}[A_i^2](t_i - t_{i-1})$.
Para el término cruzado:

$$\begin{aligned}\mathbb{E}(A_i A_j \Delta B_i \Delta B_j) &= \mathbb{E}\left[\mathbb{E}(A_i A_j \Delta B_i \Delta B_j) \mid \mathcal{F}_{t_{j-1}}\right] \\ &= \mathbb{E}\left[A_i A_j \Delta B_i \, \mathbb{E}(\Delta B_j)\right] = 0 .\end{aligned}$$

Por lo tanto:

$$\mathbb{E}\left(\left[\int_0^t H_s dB_s\right]^2\right) = \mathbb{E}\left[\sum_{i=1}^{m} A_i^2 \Delta t_i\right] = \mathbb{E}\left[\int_0^t H_s^2 ds\right] .$$

Finalemente, para 3 basta utilizar la desigualdad de Doob. □

Notación. La propiedad càglàd corresponde a tener continuidad a la derecha con límites izquierdos.

El objetivo ahora consiste en extender I. Para ello, sean

- $\mathcal{H} := \left\{\text{Procesos } (X_t)_{t\in[0,T]} \text{ adaptados càglàd, tales que } \mathbb{E}\left[\int_0^T X_t^2 dt < \infty\right]\right\}$, con la norma $|X|^2 = \mathbb{E}\left[\int_0^T X_t^2 dt\right]$.
- $\mathcal{M} := \left\{\text{Martingalas continuas } (M_t)_{t\in[0,T]} \text{ tales que } \mathbb{E}\left[\sup_{t\in[0,T]|M_t|^2}\right] < \infty\right\}$, con la norma $\|M\|^2 = \mathbb{E}\left[\sup_{t\in[0,T]} |M_t|^2\right]$.

- $\mathcal{S} := \{\text{Procesos simples predecibles}\} \subseteq \mathcal{H}$

Observación 6.4.1. $(\mathcal{H}, |\cdot|)$ y $(\mathcal{M}, \|\cdot\|)$ son espacios vectoriales normados completos con la convención usual de identificar X e Y como el mismo proceso si $X \equiv Y$ casi seguramente.

Gracias a la proposición anterior, tenemos que

$$\begin{aligned} I: \ & \mathcal{S} \to \mathcal{M} \\ & H \to I(H) = \int_0^{\cdot} H_s dB_s \end{aligned}$$

es un operador lineal y continuo:

$$\begin{aligned} \|I(H)\|^2 &= \mathbb{E}\left[\sup_{t\in[0,T]} \left|\int_0^t H_s dB_s\right|^2\right] \\ &\leq 4\mathbb{E}\left[\int_0^T H_s^2 ds\right] = 4|H|^2 . \end{aligned}$$

Luego puede extenderse de manera continua a un operador definido en la adherencia de $\mathcal{S}$.

Proposición 6.4.2
$\mathcal{S}$ es denso en $\mathcal{H}$

Utilizando la proposición anterior se ha demostrado el siguiente resultado:

Proposición 6.4.3
*Existe un operador que extiende $I : \mathcal{S} \to \mathcal{M}$ a todo $\mathcal{H}$, que también denotaremos I. Dicho operador se denomina **integral estocástica**:*

$$\forall x \in \mathcal{H}, \quad \int_0^t X_s dB_s := I(X)_t .$$

Se cumple:

1. $\forall x \in \mathcal{H},\ \forall t \in [0,T],\ \mathbb{E}\left(\left[\int_0^t X_s dB_s\right]^2\right) = \mathbb{E}\left[\int_0^t X_s^2 ds\right]$.
2. $\forall x \in \mathcal{H}$,
$$\mathbb{E}\left[\sum_{t\in[0,T]} \left|\int_0^t X_s dB_s\right|^2\right] \leq 4\mathbb{E}\left[\int_0^T X_s^2 ds\right] .$$
3. $\forall x \in \mathcal{H},\ \left(\int_0^t X_s dB_s\right)_{t\in[0,T]}$ *es una martingala continua.*

Demostración. La demostración de 2 viene simplemente del hecho que la extensión de I_4 preserva la norma. 3 es por construcción. Probemos 1:
Por definición de I y por densidad, $\exists X^n \in \mathcal{S}$ tal que $|X^n - X| \underset{n}{\to} 0$, y luego $\|I(X^n) - I(X)\| \underset{n}{\to} 0$.
Ahora para $t \in [0,T]$ fijo, tenemos:

$$\begin{aligned} \mathbb{E}\left[|I(X^n) - I(X)_t|^2\right] &\leq \mathbb{E}\left[\sup_{s\in[0,T]} |I(X^n)_s - I(X)_s|^2\right] \\ &= \|I(X^n) - I(X)\|^2 \underset{n}{\to} 0 . \end{aligned}$$

Es decir, $I(X^n)_t \underset{n}{\to} 0$ en $L^2(d\mathbb{P})$. Esto implica que $\|I(X^n)\|_{L^2(d\mathbb{P})}$ Pero como $X^n \in \mathcal{S}$, sabemos que

$$\|I(X^n)_t\|^2_{L^2(d\mathbb{P})} = \mathbb{E}\left(\left[\int_0^t X_s^n dB_s\right]^2\right) = \mathbb{E}\left[\int_0^t |X_s^n|^2 ds\right].$$

Similarmente

$$\mathbb{E}\left[\int_0^t |X_s^n - X_s|^2 ds\right] \leq \mathbb{E}\left[\int_0^T |X_s^n - X_s|^2 ds\right] = |X^n - X|^2 \underset{n}{\to} 0\,.$$

Luego $X^n \to X$ en $L^2(\Omega \times [0,t], d\mathbb{P} \otimes ds)$ lo cual implica

$$\|X^n\|_{L^2(d\mathbb{P}\otimes ds)} \to \|X\|_{L^2(d\mathbb{P}\otimes ds)} = \mathbb{E}\left[\int_0^t |X_s|^2 ds\right].$$

Con esto hemos probado:

$$\begin{aligned}\mathbb{E}\left(\left[\int_0^t X_s dB_s\right]\right) &= \|I(X)_t\|_{L^2(d\mathbb{P})} \\ &= \lim_n \|I(X^n)_t\|_{L^2(d\mathbb{P})} = \lim_n \mathbb{E}\left[\int_0^t |X_s|^2 ds\right] \\ &= \mathbb{E}\left[\int_0^t |X_s|^2 ds\right].\end{aligned}$$

□

Observación 6.4.2. En general, la integral $\int_0^t X_s dB_s$ **no es un limite trayectorial**, es decir, no se obtiene w por w como el limite de una secesión $\int_0^t X_s^n dB_s$, sin embargo, como acabamos de mostrar, se tiene la siguiente proposición.

Proposición 6.4.4
Dado $x \in \mathcal{H}$ y $X^n \in \mathcal{S}$ tal que $|X^n - X| \underset{n}{\to} 0$, $\forall t \in [0,T]$ se tiene

$$\int_0^t X_s^n dB_s \xrightarrow[n\to\infty]{L^2(\mathbb{P})} \int_0^t X_s dB_s\,.$$

A continuación se presenta otra propiedad útil.

Proposición 6.4.5
$\forall x \in \mathcal{H}$, $\forall \tau$ tiempo de parada,

$$\int_0^\tau X_s dB_s = \int_0^T 1_{\{s\leq\tau\}} X_s dB_s\,.$$

casi seguramente.

Para hacer cálculos más explícitos, necesitaremos el siguiente resultado de aproximación. Puede verse como una versión levemente restringida del hecho que $\mathcal{S}$ es denso en $\mathcal{H}$, aunque más explicita:

Proposición 6.4.6
Sea $x \in \mathcal{H}$ tal que $\mathbb{E}\{\sup_{t\in[0,T]} |X_t|^2\} < \infty$. Sea $(\pi^n)_{n\in\mathbb{N}}$ una colección de particiones de $[0,T]$, $\pi^n = \{0 = t_0^n < t_1^n < \ldots, < t_{k_n}^n = T\}$ tal que

$$\|\pi^n\| := \sup_{i=1,\ldots,k_n} \left|t_i^n - t_{i-1}^n\right| \xrightarrow[n\to\infty]{} 0\,.$$

Sea

$$X_t^n := \sum_{i=1}^{k_n} X_{t_{i-1}^n} 1_{(t_{i-1}^n, t_i^n]}(t) \in \mathcal{S}\,,$$

entonces $|X^n - X| \underset{n}{\to} 0$, y consecuentemente $\|I(X^n) - I(X)\| \underset{n}{\to} 0$.

Demostración. Fijemos $t \in [0,T]$, y $\forall n \in \mathbb{N}$, sea i_n tal que $t \in (t_{i_n-1}^n, t_{i_n}^n]$. Luego, $X_t^n = X_{t_{i_n-1}^n}$, y entonces

$$\begin{aligned}\lim_{n\to\infty} X_t^n &= \lim_{n\to\infty} X_{t_{i_n}-1}^n \\ &= \lim_{s\to t^-} X_s = X_t\,,\end{aligned}$$

donde se utiliza $t_{i_n-1}^n \to t^-$ y que X es continuo a la izquierda (càglàd). Además

$$\begin{aligned}|X_s^n - X_s| &\le 2|X_s^n|ds + 2|X_s|^2 \\ &\le 4 \sum_{s\in[0,T]} |X_s|^2 \in L^1(d\mathbb{P}\otimes ds)\,.\end{aligned}$$

Luego por el teorema de convergencia dominada:

$$|X^n - X|^2 = \mathbb{E}\left[\int_0^T |X_s^n - X_s|^2 ds\right] \xrightarrow[n\to\infty]{} 0\,.$$

□

Observación 6.4.3. La condición $\mathbb{E}\left[\sup_{t\in[0,T]} |X_t|^2\right] < \infty$ se cumple, por ejemplo, si X es un proceso acotado, o si X es una martingala continua con $\mathbb{E}[|X_T|^2] < \infty$ (como el movimiento browniano), por la desgualdad de Doob.

Ejemplo 6.4.1
Podemos usar lo anterior para comprobar que

$$\int_0^t B_s dB_s = \frac{1}{2}B_t^2 - \frac{1}{2}t\,.$$

Demostración. Sea $(\pi^n)_{n\in\mathbb{N}}$ secuencia de particiones de $[0,t]$ como en la proposición. (para $t > 0$ fijo), luego, por dicha proposición y por lo hecho previamente:

$$\int_0^t B_s dB_s = \lim_{n\to\infty} \int_0^t B_s^n ds \quad \text{en } L^2(d\mathbb{P})\,,$$

donde

$$\int_0^t B_s^n dBs = \sum_{i=1}^{k_n} B_{t_{i-1}^n}(B_{t_i^n} - B_{t_{i-1}^n}) = \frac{1}{2}\sum_{i=1}^{k_n}(B_{t_i^n}^2 - B_{t_{i-1}^n}^2) - \frac{1}{2}\sum_{i=1}^{k_n}(\Delta B_i)^2 \,.$$

Pero es fácil ver que $\sum_{i=1}^{k_n}(\Delta B_i^2)^2 \xrightarrow[n\to\infty]{} t$ en $L^2(d\mathbb{P})$:

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^{k_n}(\Delta B_i)^2 - t\right)^2\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{k_n}((\Delta B_i)^2 - \Delta t_i)\right)^2\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{k_n}\Big((\Delta B_i)^2 - \Delta t_i\Big)^2\right] + \mathbb{E}\left[\sum_{i\neq j}\Big((\Delta B_i)^2 - \Delta t_i\Big)\Big((\Delta B_j)^2 - \Delta t_j\Big)\right] \\ &= \sum_{i=1}^{k_n}\left(\mathbb{E}\left[(\Delta B_i)^4\right] + (\Delta t_i)^2 - 2\Delta t_i \mathbb{E}\left[(\Delta B_i)^2\right]\right) \,.\end{aligned}$$

Luego, $\mathbb{E}\left[\left(\sum_{i=1}^{k_n}(\Delta B_i)^2 - t\right)\right] = 2\sum_{i=1}^{k_n}(\Delta t_i)^2 \xrightarrow[n\to\infty]{} 0$, pues $\|\pi^n\| \underset{n}{\to} n$. Como todos estos limites son en $L^2(d\mathbb{P})$, se tiene:

$$\int_0^t B_s dB_s = \frac{1}{2}B_t^2 - \frac{1}{2}t \,.$$

□

Observación 6.4.4. Si $(A_t)_{t\geq 0}$ es un procesos con trayectorias derivables y $A_0 = 0$,

$$\int_0^t A_s dA_s = \int_0^t A_s \frac{dA_s}{ds} ds = \frac{1}{2}A_t^2 \,,$$

lo cual difiere del ejemplo 6.4.1. Esto se debe a que el proceso A tiene **variación cuadrática** 0 por tener trayectorias suaves, es decir,

$$\lim_n \sum_{i=1}^{k_n}(\Delta A_i)^2 = 0 \,.$$

La formula del ejemplo 6.4.1 es un caso particular de la **formula de Itô**.

6.4.2. Cálculo de Itô

Notemos que (6.4.1) puede escribirse como

$$f(B_t) = \int_0^t f'(B_s)dB_s + \frac{1}{2}\int_0^t f''(B_s)ds \,.$$

Para $f(x) = x^2$. Queremos deducir esto para funciones de clase $\mathcal{C}^2$ generales, y para procesos X más generales que el movimiento browniano B.

Definición 6.4.3 Proceso de Itô
Dado un espacio de probabilidad filtrado $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ y un movimiento browniano $(B_t)_{t\geq 0}$ definido sobre él. Decimos que un proceso $(X_t)_{t\in[0,T]}$ es un **proceso de Itô** si se escribe como:

$$X_t = X_0 + \int_0^t K_s ds + \int_0^t H_s dB_s \quad \forall t \in [0,T]$$

Donde:

- X_0 es $\mathcal{F}_0$-medible.
- $(K_t)_{t\in[0,T]}$, $(H_t)_{t\in[0,T]}$ son adaptados.
- $\int_0^T |K_s| ds < \infty$, $\int_0^T |H_s|^2 ds < \infty$ casi seguramente.

Observación 6.4.5. Puede probarse que la escritura de la Definición 6.4.3 es única, es decir, si reemplazamos K, H por $\tilde{K}, \tilde{H}$, entonces $K \equiv \tilde{K}$, $H \equiv \tilde{H}$ casi seguramente con respecto a $ds \otimes d\mathbb{P}$.

Definición 6.4.4 Variación Cuadrática
Dado un proceso de Itô $(X_t)_{t\geq 0}$ con descomposición de la Definición 6.4.3, su **variación cuadrática** se define como el proceso $\langle X \rangle_t$ dado por

$$\langle X \rangle_t := \int_0^t H_s^2 ds \,.$$

Observación 6.4.6. Puede probarse que

$$\langle X, Y \rangle_t = \lim_n \sum_{i=1}^{k_n} (\Delta X_i)^2, \quad \text{en } L^2(d\mathbb{P})$$

Donde $(\pi^n)_{n\geq 1}$ es una secuencia de particiones de $[0,T]$ como antes. Esto justifica el nombre *variación cuadrática*.

Teorema 6.4.1 Lema de Itô, Regla de Itô, Fórmula de Itô
Sea $(X_t)_{t\geq 0}$ un proceso de Itô con descomposición de la Definición 6.4.3, y $f : \mathbb{R} \to \mathbb{R}$ una función $\mathcal{C}^2$. Entonces

$$f(X_t) = f(X_0) + \int_0^t f'(X_s) dX_s + \frac{1}{2} \int_0^t f''(X_s) d\langle X \rangle_s \,.$$

Observación 6.4.7. La expresión del Teorema 6.4.1 suele escribirse en *forma diferencial*, la cual es más fácil de recordar:

$$df(X_t) = f'(X_t) dX_t + \frac{1}{2} f''(X_t) d\langle X \rangle_t \,.$$

Alternativamente, usando las definiciones de $\langle X \rangle_t$ y $\int_0^t f'(X_s) dX_s$, se tiene

$$\begin{aligned} df(X_t) &= f'(X_t) K_t dt + f'(X_t) H_t dB_t + \frac{1}{2} f''(X_t) H_t^2 dt \\ &= \left[f'(X_t) K_t + \frac{1}{2} f''(X_t) H_t^2 \right] dt + f'(X_t) H_t dB_t \,. \end{aligned}$$

Lo cual muestra que $f(X_t)$ es un proceso de Itô y nos entrega su descomposición.

Observación 6.4.8. La regla de Itô también se conoce como *regla de la cadena estocástica*, la cual se justifica al observar la forma diferencial de la Observación **??**. Notar que $\frac{1}{2}f''(X_t)d\langle X\rangle_t$ es una novedad en el sentido de que no aparece en cálculo clásico.

Ejemplo 6.4.2
Para $X = B$, $f(x) = x^2$, tenemos

$$dB_t^2 = 2B_t dB_t + \frac{1}{2}2d\langle B\rangle_t = 2B_t dB_t - dt\,.$$

Lo cual es exactamente (6.4.1) escrito en forma diferencial.

Ejemplo 6.4.3 Movimeinto Browniano Geométrico
Dados $\mu \in \mathbb{R}$ un *coeficiente de drift*, $\sigma > 0$ *volatilidad*. Decimos que $(S_t)_{t\geq 0}$ es un movimiento browniano *geométrico* si cumple:

$$dS_t = \mu dS_t dt + \sigma S_t dB_t\,.$$

Para resolver esto, sea $(X_t)_{t\geq 0}$ dado por

$$dX_t = (\mu - \frac{1}{2}\sigma^2)dt + \sigma dB_t\,.$$

y definimos $S_t = S_0 e^{X_t}$. Verificamos que S es un movimiento browniano geométrico: usando la regla de Itô,

$$\begin{aligned} dS_t &= S_0\left[e^{X_t}dX_t + \frac{1}{2}e^{X_t}d\langle X\rangle_t\right] \\ &= S_0\left[e^{X_t}(\mu - \frac{1}{2}\sigma^2)dt + e^{X_t}\sigma dB_t + \frac{1}{2}e^{X_t}\sigma^2 dt\right] \\ &= S_0 e^{X_t}\mu dt + S_0 e^{X_t}\sigma dB_t = \mu S_t dt + \sigma S_t dB_t\,. \end{aligned}$$

El movimiento browniano geométrico es un modelo simple de precios de acciones.

Definición 6.4.5 Covariación Cuadrática
Sean dos procesos de Itô X, Y con descomposición

$$\begin{aligned} dXt &= K_t dt + H_t dB_t \\ dY_t &= \tilde{K}_t dt + \tilde{H}_t dB_t\,. \end{aligned}$$

Su **covariación cuadrática** es el proceso $(\langle X, Y\rangle)_{t\geq 0}$,

$$\langle X, Y\rangle_t := \int_0^t H_s\tilde{H}_s ds\,.$$

Observación 6.4.9. Puede probarse que

$$\langle X, Y\rangle = \lim_n \sum_i (\Delta X_i)(\Delta Y_i)$$

La siguiente propiedad completa lo básico del calculo estocástico:

Proposición 6.4.7 Integración por Partes
Dados X, Y proceso de Itô con descomposición de la Definición 6.4.5,se tiene

$$X_t Y_t = X_0 Y_0 + \int_0^t X_s dY_s + \int_0^t Y dX_s + \langle X, Y\rangle_t$$

o en forma diferencial

$$d(X_t, Y_t) = X_t dY_t + Y_t dX_t + \langle X, Y\rangle_t$$

Demostración. Por la regla de Itô

$$\begin{aligned} X_t^2 &= X_0^2 + \int_0^t 2X_s dX_s + \int_0^t \frac{1}{2} 2 d\langle X\rangle_s \\ &= X_0^2 + 2\int_0^t X_s dX_s + \int_0^t H_s^2 ds\,. \end{aligned}$$

Análogamente

$$Y_t^2 = Y_0^2 + 2\int_0^t Y_s dY_s + \int_0^t \tilde{H}_s^2 ds\,.$$

La descomposición de $X + Y$ claramente es:

$$d(X+Y) = (K_t + \tilde{K}_t)dt + (H_t + \tilde{H}_t)dB_t\,.$$

Luego

$$\begin{aligned} (X_t + Y_t)^2 &= (X_0 + Y_0)^2 + \int_0^t 2(X_s + Y_s) d(X_s + Y_s) + \frac{1}{2}\int_0^t 2 d\langle X + Y\rangle_2 \\ &= X_0^2 + Y_0^2 + 2X_0 Y_0 + 2\int_0^t X_s dX_s + 2\int_0^t Y_s dY_s \\ &+ 2\int_0^t X_s dY_s + 2\int_0^t Y_s dX_s + \int_0^t H_s^2 ds + \int_0^t \tilde{H}_s^2 ds + 2\int_0^t H_s \tilde{H}_s ds\,. \end{aligned}$$

Haciendo $\frac{1}{2}((3) - (1) - (2))$ se tiene el resultado □

6.5. Ecuaciones Diferenciales Estocásticas

Dada una ecuación diferencial clásica, digamos

$$\frac{dX_t}{dt} = b(X_t)\,,$$

o en forma diferencial

$$dX_t = b(X_t)dt\,,$$

donde $b : \mathbb{R} \to \mathbb{R}$ es una función determinista, queremos añadir *ruido* a la evolución de X. Este ruido puede modelar distintos fenómenos: imperfecciones en la medición, variaciones bursátiles, interacciones con el medio, etc. Si el ruido es proporcional a una función de X, obtenemos la **ecuación diferencial estocástica** (EDE o bien *SDE* en inglés)

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t\,.$$

Aquí $(B_t)_{t\geq 0}$ es un movimiento browniano. Esto debe entenderse como una notación para la siguiente definición rigurosa:

Definición 6.5.1 Solución exacta de una EDE
Dado un espacio de probabilidad filtrado $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, con un movimiento browniano $(B_t)_{t\geq 0}$ asociado y una variable aleatoria $X_0 \in \mathcal{F}_0$, junto con funciones $b : \mathbb{R} \to \mathbb{R}$, $\sigma : \mathbb{R} \to \mathbb{R}$, decimos que $(X_t)_{t\geq 0}$ es una solución fuerte de la ecuación diferencial estocástica $dX_t = b(X_t)dt + \sigma(X_t)dB_t$, si

- $\forall t \geq 0$, $\int_0^t |b(X_s)|ds < \infty$ y $\int_0^t |\sigma(X_s)|^2 ds < \infty$ casi seguramente.
- $\forall t \geq 0$,
$$X_t = X_0 + \int_0^t b(X_s)ds + \int_0^t \sigma(X_s)dB_s$$
casi seguramente.

Observación 6.5.1. También puede incluirse el caso en que b y σ dependen de t, i.e., $b = b(t, x)$, $\sigma = \sigma(t, x)$.

Observación 6.5.2. Un X cumpliendo

$$X_t = X_0 + \int_0^t b(X_s)ds + \int_0^t \sigma(X_s)dB_s$$

se denomina **difusión**.

Ahora estudiemos su existencia y unicidad. Notemos que incluso en el caso determinista ($\sigma \equiv 0$), la ecuación 6.5.2 puede tener mal comportamiento (múltiples soluciones o soluciones que explotan) si la función b no es Lipschitz. (por ejemplo $\frac{dx}{dt} = x^2$). Esto resulta ser suficiente:

Teorema 6.5.1 Existencia y Unicidad de Soluciones para EDEs
Supongamos que $\exists K > 0$ *constante tal que*

- $|b(x) - b(y)| + |\sigma(x) - \sigma(y)| \leq K|x - y|$, $\forall x, y \in \mathbb{R}$.
- $|b(x)| + |\sigma(x)| \leq K(1 + |x|)$, $\forall x \in \mathbb{R}$.
- $\mathbb{E}[X_0^2] < \infty$.

Entonces, para todo T,

$$X_t = X_0 + \int_0^t b(X_s)ds + \int_0^t \sigma(X_s)dB_s$$

admite casi seguramente una única solución fuerte $(X_t)_{t\in[0,T]}$ *en* $[0, T]$, *la cual además cumple*

$$\mathbb{E}\left[\sup_{t\in[0,T]} |X_t|^2\right] < \infty\,.$$

Demostración. Usaremos un argumento de punto fijo. Para ello, sea

$$\xi = \{(X_t)_{t\in[0,T]} \text{ adaptado y continuo tal que } \|X\|_T < \infty\}\,,$$

donde $\|X\|_T^2 := \mathbb{E}\left[\sup_{t\in[0,T]} |X_t|^2\right]$. $(\xi, \|\cdot\|_T)$ es un espacio de Banach. $\forall \in \xi$, definimos $(\Phi(X)_t)_{t\in[0,T]}$ como

$$\Phi(X)_t = X_0 + \int_0^t b(X_s)ds \int_0^t \sigma(X_s)dB_s\,.$$

Es fácil ver que $\Phi(X) \in \xi$. Además

$$\begin{aligned}|\Phi(X)_t - \Phi(Y)_t|^2 &\leq \left|\int_0^t (b(X_s) - b(Y_s))ds + \int_0^t (\sigma(X_s) - \sigma(Y_s))dB_s\right|^2 \\ &\leq 2\sup_{t\in[0,T]}\left|\int_0^t (b(X_s) - b(Y_s))ds\right|^2 + 2\sup_{t\in[0,T]}\left|\int_0^t (\sigma(X_s) - \sigma(Y_s))dB_s\right|^2.\end{aligned}$$

Por otra parte

$$\begin{aligned}\|\Phi(X) - \Phi(Y)\|_T^2 &= \mathbb{E}\left[\sup_{t\in[0,T]} |\Phi(X)_t - \Phi(Y)_t|^2\right] \\ &\leq 2\mathbb{E}\left[\sup_{t\in[0,T]}\left(\int_0^t |b(X_s) - b(Y_s)|ds\right)^2\right] \\ &\quad + 2\mathbb{E}\left[\sup_{t\in[0,T]}\left|\int_0^t (\sigma(X_s) - \sigma(Y_s))dB_s\right|^2\right] \\ &\leq 2K^2T^2\mathbb{E}\left[\sup_{t\in[0,T]} |X_t - Y_t|^2\right] + 8K^2T\mathbb{E}\left[\sup_{t\in[0,T]} |X_t - Y_t|^2\right] \\ &= \left(2K^2T^2 + 8K^2T\right)\|X - Y\|_T^2\,.\end{aligned}$$

Luego, Φ es Lipschitz con constante $L(T) \leq \sqrt{2K^2T^2 + 8K^2T}$, luego, si T es suficientemente pequeño, de modo que $L(T) < 1$, Φ será contractante, y entonces, tendrá un único punto fijo en ξ. Para pasar a un T cualquiera, basta concatenar soluciones en intervalos

$$[0, \frac{T}{n}], [\frac{T}{n}, \frac{2T}{n}], \ldots, [\frac{(n-1)T}{n}, T]\,.$$

para n adecuado. □

Observación 6.5.3. El argumento anterior prueba la unicidad de soluciones en ξ. Sin embargo, puede probarse que cualquier selección de 6.5.2 debe pertenecer a ξ, lo cual significa que el argumento anterior efectivamente implica unicidad de soluciones.

6.5.0.1. Esquemas Numéricos

Dada una ecuación diferencial estocástica

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t\,,$$

queremos aproximar la solución por X_t^n que se pueda generar efectivamente en un computador.

El parámetro $n \in \mathbb{N}$ se tomará grande, y se espera que $X^n \underset{n}{\to} X$ en algún sentido.

El primer esquema es clásico **esquema de Euler** que se justifica por la siguiente heurística:

$$\begin{aligned} X_{t+h} &= X_t + \int_t^{t+h} b(X_s)ds + \int_t^{t+s} \sigma(X_s)dB_s \\ &\approx X_t + b(X_t)h + \sigma(X_t)(B_{t+h} - B_t) \end{aligned}$$

Es decir, reemplazamos $(X_s)_{s\in[t,t+h]}$ por su valor en el extremo izquierdo del intervalo $[t, t+h]$.

Algoritmo 6.5.1 Esquema de Euler
Dado: $T > 0$, $n \in \mathbb{N}$, $h = \frac{T}{n}$.
Definir $(X^n_{ih})_{i=0,\dots,n}$ *mediante* $X^n_0 = X_0$ *y*

$$X^n_{ih+h} = X^n_{ih} + b(X^n_{ih})h + \sigma(X^n_{ih})(B_{ih+h} - B_{ih})\,.$$

Refinemos un poco lo anterior: partiendo desde $dX_t = b(X_t)dt + \sigma(X_t)dB_t$,

$$X_{t+h} \approx X_t + b(X_t)h + \sigma(X_t)\Delta B\,.$$

Luego, para mejorar el esquema, buscamos mejorar ese ΔB, pues es de orden sólo $h^{\frac{1}{2}}$ y buscamos h. Por Itô, para $s \geq t$,

$$\begin{aligned} \sigma(X_s) &= \sigma(X_t) + \int_0^t \sigma'(X_u)dX_u + \frac{1}{2}\int_t^s \sigma''(X_u)d\langle X\rangle_u \\ &= \sigma(X_t) + \int_t^s \sigma'(X_u)[b(X_u)ds + \sigma(X_u)dB_u] + \frac{1}{2}\int_t^s \sigma''(X_u)\sigma^2(X_u)du\,. \end{aligned}$$

Reemplazaremos esto en $\int_t^{t+h}\sigma(X_s)dB_s$. Con la siguiente heurística:

$$\begin{aligned} du\ dB_s &\sim h \cdot h^{1/2} = h^{3/2} \\ du\ ds &\sim h \cdot h = h^2 \\ dB_u\ dB_s &\sim h^{1/2} \cdot h^{1/2} = h\,. \end{aligned}$$

Descartaremos las integrales $dudB_s$ y $duds$. Obtenemos:

$$\begin{aligned} X_{t+h} &= X_t \int_t^{t+h} b(X_s)ds + \int_t^{t+h} \sigma(X_s)dB_s \\ &\approx X_t + b(X_t)h + \sigma(X_t)(B_{t+h} - B_t) + \sigma'(X_t)\sigma(X_t)\int_t^{t+h}\int_t^s dB_u dB_s\,. \end{aligned}$$

Pero:

$$\begin{aligned}\int_t^{t+h}\int_t^s dB_u dB_s &= \int_t^{t+h}(B_s - B_t)dB_s \\ &= \int_t^{t+h} B_s dB_s - B_t\int_t^{t+h} dB_s \\ &= \frac{1}{2}B_{t+h}^2 + \frac{1}{2}(t+h) - \left[\frac{1}{2}B_t^2 - \frac{1}{2}t\right] - B_t(B_{t+h} - B_t) \\ &= \frac{1}{2}B_{t+h}^2 + \frac{1}{2}B_t^2 - \frac{1}{2}h - B_t B_{t+h} + B_t^2 \\ &= \frac{1}{2}(B_{t+h} - B_t)^2 - \frac{1}{2}h\,.\end{aligned}$$

Obtenemos

Algoritmo 6.5.2 Esquema de Milstein
Dado: $T > 0$, $n \in \mathbb{N}$, $h := \frac{T}{n}$.
Se define $(X_{ih}^n)_{i=0,\dots,n}$ *por medio de:* $X_0^n = 0$ *y*

$$X_{ih+h}^n = X_{ih}^n + b(X_{ih}^n)h + \sigma(X_{ih}^n)(B_{ih+h} - B_{ih}) + \frac{1}{2}\sigma(X_{ih}^n)\sigma'(X_{ih}^n)\left[(B_{ih+h} - B_{ih})^2 - h\right]$$

Se tienen las siguientes tasas de convergencia.

Teorema 6.5.2 Convergencia Esquema de Euler
Supongamos que b, σ *son Lipschitz de constante* L, *y que* $\exists m \geq 1$ *entero tal que* $\mathbb{E}[|X_0|^{2m}] < \infty$. *Entonces,* $\exists K = K(T, L, m, \mathbb{E}[|X_0|^{2m}])$ *constante, tal que la aproximación de Euler* (X_{ih}^n) *(con* $h = \frac{T}{n}$*) cumple* $\forall n$*:*

$$\mathbb{E}\left[\sup_{i=1,\dots,n}|X_{ih} - X_{ih}^n|^{2m}\right] \leq \frac{K}{n^{2m}}\,,$$

y

$$\forall \alpha \in \left[0, \frac{1}{2} - \frac{1}{2m}\right), \quad n^\alpha \cdot \sup_{i=1,\dots,n}|X_{ih} - X_{ih}^n| \xrightarrow[c.s.]{n} 0\,.$$

Similarmente

Teorema 6.5.3 Convergencia del Esquema de Milstein
Supongamos que $b, \sigma \in \mathcal{C}^2$ *con derivadas hasta orden 2 acotadas por una constante* $C > 0$, *y que* $\exists m \geq 1$ *entero, tal que,* $\mathbb{E}\left[|X_0|^{4m}\right] < \infty$. *Entonces* $\exists K = K(T, C, m, \mathbb{E}[|X_0|^{4m}])$, *tal que la aproximación de Milstein* $(X_{ih}^n)_{i=1,\dots,n}$ *cumple* $\forall n$*:*

$$\sup_{i=1,\dots,n}\mathbb{E}\left[|X_{ih} - X_{ih}^n|^{2m}\right] \leq \frac{K}{n^{2m}}\,,$$

y

$$\forall \alpha \in \left[0, \frac{1}{2} - \frac{1}{m}\right), \quad n^\alpha \cdot \sup_{i=1,\dots,n}|X_{ih} - X_{ih}^n| \xrightarrow[c.s.]{n} 0\,.$$

Las demostraciones correspondientes escapan al alcance del curso, sin embargo, los teoremas nos dicen en esencia:

- $$\mathbb{E}\left[\sup_{i=1,\dots,n}|X_{ih}-X_{ih}^n|^{2m}\right]\leq\frac{K}{n^{2m}}$$

  implica que el esquema de Euler es de orden $n^{\frac{1}{2}}$.

- $$\sup_{i=1,\dots,n}\mathbb{E}\left[|X_{ih}-X_{ih}^n|^{2m}\right]\leq\frac{K}{n^{2m}}$$

  implica que el esquema de Milstein es de orden n:

$$\mathbb{E}\left[X_{ih}-X_{ih}^n\right]\leq\left(\mathbb{E}\left[|X_{ih}-X_{ih}^n|^{2m}\right]\right)^{\frac{1}{2m}}\leq\begin{cases}\frac{K^{1/2m}}{n^{1/2}}, & \text{Euler}\\ \frac{K^{1/2m}}{n}, & \text{Milstein}\,.\end{cases}$$

6.5.1. Cálculo Estocástico y EDPs

Se puede establecer una profunda conexión entre ciertas EDPs en $\mathbb{R}^d$ y las difusiones. Para ello, se necesita primero definir el movimiento browniano en $\mathbb{R}^d$.

Definición 6.5.2
Dado un espacio de probabilidad filtrado $(\Omega,\mathcal{F},(\mathcal{F}_t)_{t\geq0},\mathbb{P})$, decimos que $B=)B_t^1,\dots,B_t^{\tilde{d}})_{t\geq0}^T$. Es un movimiento browniano d-dimensional si los $(B_t^i)_{t\geq0}$ son movimientos brownianos independientes (en $\mathbb{R}^d$).

Definición 6.5.3 Proceso de Itô d-simensional
Un proceso $X=(X_t^1,\dots,X_t^d)_{t\geq0}^T$ a valores en $\mathbb{R}^d$ se dice un **proceso de Itô d-dimensional** si $\forall i=1,\dots,d$

$$X_t^i=X_0^i+\int_0^t K_s^i ds+\sum_{j=1}^{\tilde{d}}H_s^{ij}dB_s^j\,,$$

donde $(K_t^i)_{t\geq0,i=1,\dots,d}$, $(H_t^{ij})_{t\geq0,\,j=1,\dots,\tilde{d}}^{i=1,\dots,d}$ son procesos adaptados cumpliendo $\forall T>0$:

$$\forall i,\ \int_0^T\left|K_s^i\right|ds<\infty,\text{c.s.},\quad\int_0^T\left|H_s^{ij}\right|^2ds<\infty,\quad\forall ij\,.$$

Matricialmente: si $K_t=(K_t^1,\dots,K_t^d)^T\in\mathbb{R}$, $H_t=(H_t^{ij})_{\substack{i=1,\dots,d\\ j=1,\dots,\tilde{d}}}\in\mathbb{R}^{d\times\tilde{d}}$, luego la ecuación de la Definición 6.5.3 equivale a $X_t=X_0+\int_0^t K_s ds+\int_0^t H_s dB_s$.

Ahora que trabajamos con varios brownianos i.i.d., debemos extender la definición de covariación cuadrática:

$$d\langle X^i,X^j\rangle_t:=\sum_{l=1}^{\tilde{d}}H_t^{il}H_t^{jl}dt\,,$$

es decir,

$$\langle X^i,X^j\rangle:=\sum_{l=1}^{\tilde{d}}\int_0^t H_s^{il}H_s^{jl}ds\,.$$

Por ejemplo: $X^i = B^i$, $X_j = B^j$ se tiene

$$H_s^{il} = \begin{cases} 1 & , \text{ si } l = i \\ 0 & \sim , \end{cases}$$

que opera de manera análoga para H_s^{jl}. Por otra parte

$$\langle B^i, B^j \rangle_t = \sum_{l=1}^{\tilde{d}} \int_0^t H_s^{il} H_s^{jl} ds = \delta_{ij} t \, .$$

Esto se justifica probando (propuesto) que para $i \neq j$

$$\sum_k (\Delta B_k^i)(\Delta B_k^j) \xrightarrow[|\pi| \to 0]{L^2(d\mathbb{P})} 0 \, ,$$

donde ΔB_k^i, ΔB_k^j son los incrementos en intervalo $[t_{k-1}, t_k]$ de la malla temporal.
Observación 6.5.4. La expresión

$$\langle X^i, X^j \rangle := \sum_{l=1}^{\tilde{d}} \int_0^t H_s^{il} H_s^{jl} ds$$

se obtiene a partir de la siguientes propiedades equivalentes:

- $\langle \cdot, \cdot \rangle$ es bilineal y simétrica.
- $\langle X, \int_0^{\cdot} \tilde{H}_s ds \rangle_t = 0$
- $\langle \int_0^{\cdot} H_s dB_s^i, \int_0^{\cdot} \tilde{H}_s dB_s^j \rangle = \delta_{ij} \int_0^t H_s \tilde{H}_s ds$

Ahora podemos enunciar (sin demostración):

Teorema 6.5.4 Regla de Itô en $\mathbb{R}^d$
Dado un tal $(X_t)_{t \geq 0}$ y una función $f : \mathbb{R}^d \to \mathbb{R}$ $\mathcal{C}^2$, se tiene $\forall t \geq 0$,

$$f(X_t) = f(X_0) + \sum_{i=1}^d \int_0^t \frac{\partial f}{\partial x^i}(X_s) dX_s^i + \frac{1}{2} \sum_{i,j=1}^d \int_0^t \frac{\partial^2 f}{\partial x^i \partial x^j}(X_s) d\langle X^i, X^j \rangle_s \, ,$$

donde $dX_s^i := K_s^i ds + \sum_{l=1}^{\tilde{d}} H_s^{il} dB_s^l$.

Desarrollando $d\langle X^i, X^j \rangle_s = \sum_{l=1}^{\tilde{d}} H_s^{il} H_s^{jl} ds$, y anotando $\mathrm{tr}(\cdot)$ la traza de una matriz cuadrada, podemos escribir la regla de Itô matricialmente:

$$f(X_t) = f(X_0) + \int_0^t \nabla f(X_s) \cdot dX_s + \frac{1}{2} \int_0^t \mathrm{tr} \left(D^2 f(X_s) H_s H_s^T \right) ds \, ,$$

siendo $D^2 f(x)$ la matriz hessiana del f en x.
Como caso particular se puede analizar: $dX_t^i = K_t dt + \sigma dB_t^i$ $(d = \tilde{d})$, es decir, $H_t^{ij} \equiv 0$, $i \neq j$,

$H_t^{ii} \equiv \sigma$. En tal caso $d\langle X^i, X^j\rangle_s = \sigma^2\delta_{ij}ds$, luego

$$\begin{aligned} f(X_t) &= f(X_0) + \int_0^t \nabla f(X_s)\cdot dX_s + \frac{1}{2}\sigma^2\int_0^t \Delta f(X_s)ds \\ &= f(X_0) + \int_0^t \left[\nabla f(X_s)\cdot K_s + \frac{1}{2}\sigma^2\Delta f(X_s)\right]ds + \sigma\int_0^t \nabla f(X_s)\cdot dB_s\,. \end{aligned}$$

Definición 6.5.4
Dado $x \in \mathbb{R}^{\tilde{d}}$, denotaremos $\mathbb{P}_x$, $\mathbb{E}_x$ a la probabilidad y esperanza de la medida bajo la cual el movimiento browniano $(B_t)_{t\geq 0}$ parte desde x, es decir,

$$\mathbb{P}_x(X_0 = x) = 1$$

6.5.2. Problema de Dirichlet

Sea $D \subseteq \mathbb{R}^d$ abierto, $f : \partial D \to \mathbb{R}$ función continua. Consideremos el **problema de Dirichlet**: encontrar $u : \overline{D} \to \mathbb{R}$ de clase $\mathcal{C}^2$ y continua en $\overline{D}$, tal que

$$\begin{cases} \Delta u = 0 & \text{en } D \\ u = f & \text{en } \partial D \end{cases}. \tag{D}$$

A continuación escribiremos una **representación probabilista** de la solución: sea

$$\begin{aligned} \tau_D = &\text{ tiempo de llegada de } B \text{ a } D^c \\ &= \inf\{t \geq 0 : B_t \in D^c\}\,, \end{aligned}$$

de manera heurística, la solución de (D) tiene la representación

$$u(x) = \mathbb{E}_x\left[f(B_{\tau_D})\right], \quad \forall x \in \overline{D}$$

De manera más rigurosa:

Teorema 6.5.5
Si f es acotada y $\forall x \in D$, $\mathbb{P}_x(\tau_D < \infty) = 1$, entonces cualquier solución acotada de

$$\begin{cases} \Delta u = 0 & \textit{en } D \\ u = f & \textit{en } \partial D \end{cases}. \tag{D}$$

tiene la representación

$$u(x) = \mathbb{E}_x\left[f(B_{\tau_D})\right], \quad \forall x \in \overline{D}\,.$$

Idea de Demostración. Sea u solución acotada de (D), por el lema de Itô:

$$u(B_t) = u(B_0) + \int_0^t \nabla e u(B_s)\cdot dB_s + \frac{1}{2}\int_0^t \Delta u(B_s)ds\,.$$

De manear heurística:

$$u(B_{\tau_D}) = u(B_0) + \int_0^{\tau_D} \nabla u(B_s)\cdot dB_s + \frac{1}{2}\int_0^{\tau_D} \Delta u(B_s)ds\,.$$

Luego

$$\mathbb{E}_x\left[f(B_{\tau_D})\right] = \mathbb{E}_x[u(B_0)] + \mathbb{E}_x\left[\int_0^t \nabla u(B_s)\cdot dB_s\right],$$

que finalmente implica $u(x) = \mathbb{E}_x[f(B_{\tau_D})]$. □

Si a priori no conoce la existencia de solución del problema de Dirichlet, se espera que $\mathbb{E}_x\left[f(B_{\tau_D})\right]$ sea la solución. Esto es cierto, bajo ciertas condiciones de **regularidad** de ∂D que aseguren que (D) entregue una función continua hasta ∂D. Específicamente tenemos la definición siguiente.

Definición 6.5.5 Punto Regular
Dado $D \subseteq \mathbb{R}^d$ abierto, sea

$$\sigma_D = \inf\{t > 0 : B_t \in D^c\}.$$

Decimos que un punto $x \in \partial D$ es regular si $\mathbb{P}_x(\sigma_D = 0) = 1$, es decir, con probabilidad 0, la trayectoria de B entra a D inmediatamente y se queda D por intervalo de tiempo.

Teorema 6.5.6
Dado $D \subseteq \mathbb{R}^d$ abierto, con $d \geq 2$, y $x \in \partial D$, son equivalentes:

1. *x es regular.*
2. *$\forall f : \partial D \to \mathbb{R}$ medible y acotada, continua en x, se tiene*

$$\lim_{\substack{y\to x\\ y\in D}} \mathbb{E}_y[f(B_{\tau_D})] = f(x).$$

Con esto se obtiene

Teorema 6.5.7
Sea $D \subseteq \mathbb{R}^d$ abierto con ∂D regular (es decir, $\forall x \in \partial D$, x es regular), sea $f : \partial D \to \mathbb{R}$ continua y acotada. Si $\forall \in D$, $\mathbb{P}_x(\tau_D < \infty) = 1$, entonces

$$u(x) := \mathbb{E}_x[f(B_{\tau_D})]$$

es la única solución del problema de Dirichlet.

Más generalmente

Teorema 6.5.8
Sea $D \subseteq \mathbb{R}^d$ abierto y acotado, sean $g : D \to \mathbb{R}$, $f : \partial D \to \mathbb{R}$ acotadas y continuas. Sea $u : \overline{D} \to \mathbb{R}$ $\mathcal{C}^2$ en D y continua en $\overline{D}$ tal que

$$\begin{cases} \Delta u = g & \text{en } D \\ u = f & \text{en } \partial D \end{cases},$$

entonces

$$u(x) = \mathbb{E}_x\left[f(B_{\tau_D}) + \frac{1}{2}\int_0^{\tau_D} g(B_s)ds\right].$$

Con esto se puede idear un algoritmo. En efecto, basta simular N movimientos brownianos independientes en $\mathbb{R}^d$, $B^{(1)}, \ldots, B^{(N)}$ cada uno partiendo desde $x \in D$ en incrementos temporales

de paso h fijo. Sea t^k el primer instante en que B^k sale de D. Se aproxima (D) como:

$$u(x) = \frac{1}{N}\sum_{k=1}^{N} f(B^k_{t^{(k)}})$$

Ejercicio: Establezca el algoritmo descrito anteriormente.

6.5.3. Ecuación de Feynman-Kac

Se busca encontrar una representación probabilista para las soluciones de cierta clase de EDPs parabólicas (es decir, con derivadas temporales; el problema de Dirichlet es una ecuación elíptica).

Necesitaremos un versión del lema de Itô que admite dependencia temporal de la función:

Teorema 6.5.9 Lema de Itô
Sea $X_t \in \mathbb{R}^d$ un proceso con descomposición:

$$dX_t = K_t dt + H_t dB_t\,,$$

donde $K_T \in \mathbb{R}^d$, $H_t \in \mathbb{R}^{d\times\tilde{d}}$, B_t movimiento browniano en $\mathbb{R}^{\tilde{d}}$. Sea $f : [0,\infty)\times\mathbb{R}^d \to \mathbb{R}$ de clase $\mathcal{C}^{1,2}$. Entonces

$$f(t,X_t) = f(0,X_0) + \int_0^t \partial_s f(s,X_s)ds + \int_0^t \nabla_x f(s,X_s)\cdot dX_s + \frac{1}{2}\int_0^t tr(D_x^2 f(s,X_s)H_s H_s^T)ds\,.$$

Especificaremos la EDP

Teorema 6.5.10 Fórmula de Feynman-Kac
Sean $b : \mathbb{R}^d \to \mathbb{R}$, $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times\tilde{d}}$ funciones dadas, y sea $a : \mathbb{R}^d \to \mathbb{R}^d$ dada por $a = \sigma\sigma^T$. Consideremos el operador:

$$\mathcal{L} := b(\cdot)\cdot\nabla + \frac{1}{2}tr(a(\cdot)D^2(\cdot))\,,$$

es decir $(\mathcal{L}f)(x) := b(x)\cdot\nabla f(x) + \frac{1}{2}tr(D^2 f(x)a(x))$.

Consideremos la siguiente EDP parabólica para una función $u(t,x)$, $t \geq 0$, $x \in \mathbb{R}^d$:

$$\begin{cases} \partial_t u + \mathcal{L}u = 0 & (t,x) \in [0,T]\times\mathbb{R}^d \\ u(T,x) = f(x) & x \in \mathbb{R}^d \end{cases} \tag{1}$$

Supongamos que admite solución $\mathcal{C}^{1,2}$ con derivadas con crecimiento polinomial, con b, σ Lipschitz y tal que f tiene crecimiento a lo más polinomial. Por otro lado, sea X_t^x solución de la siguiente EDE: (para $x \in \mathbb{R}^d$ fijo)

$$X_t^x = x + \int_0^t b(X_s^x)ds + \int_0^t \sigma(X_s^x)dB_s\,, \tag{2}$$

entonces

$$u(t,x) = \mathbb{E}[f(X_{T-t}^x)]$$

Demostración. Fijemos $0 < t < T$. Sea $(Y_r^{x,t})_{r\in[0,T]}$ solución de la EDE:

$$Y_r^{x,t} = x + \int_0^r b(Y_s^{x,t})ds + \int_t^r \sigma(Y_s^{x,t})dB_s$$

Probaremos primero que

$$u(t,x) = \mathbb{E}[f(Y_T^{x,t})]$$

Usamos Itô sobre $u(r, Y_r^{x,t})$, en $r = T$:

$$\begin{aligned} u(T, Y_T^{x,t}) &= u(t, Y_t^{x,t}) + \int_t^T \partial_s u(s, Y_s^{x,t})ds + \int_t^T \nabla u(s, Y_s^{x,t}) \cdot dY_s^{x,t} \\ &\quad + \frac{1}{2}\int_t^T \mathrm{tr}(D^2 u(s, Y_s^{x,t}))\sigma(Y_s^{x,t})\sigma(Y_s^{x,t})^T)ds \\ &= u(t,x) + \int_t^T \sigma(Y_s^{x,t})dB_s \\ &\quad + \int_t^T \left[\partial_s u(s, Y_s^{x,t}) + \nabla u(s, Y_s^{x,t}) \cdot b(Y_s^{x,t}) + \frac{1}{2}\mathrm{tr}(D^2 u(s, Y_s^{x,t}))a(Y_s^{x,t})\right] ds\,. \end{aligned}$$

Como u resuelve (1), se tiene que $[\ ] = 0$ y que $u(T, Y_T^{x,t}) = f(Y_T^{x,t})$. Tomando $\mathbb{E}[\cdot]$, la martingala se anula, luego

$$\mathbb{E}[f(Y_t^{x,t})] = u(t,x)\,.$$

Para concluir que $u(t,x) = \mathbb{E}[f(X_{T-t}^x)]$, basta notar que los procesos $(X_r^x)_{0\le r\le T-t}$ y $(Y_r^{x,t})_{t\le r\le T}$ tiene la misma ley, pues ambos resuelven la EDE (2) pero con movimientos brownianos distintos:

- X^x resuelve (2) para el movimiento browniano $(B_r)_{0\le r\le T-t}$.
- $Y^{x,t}$ resuelve (2) para el movimiento browniano $(B_{t+r} - B_t)_{0\le r\le T-t}$.

Luego, $X_{T-t}^x \overset{d}{=} Y_T^{x,t}$, así

$$u(x,t) = \mathbb{E}[f(Y_T^{x,t})] = \mathbb{E}[f(X_{T-t}^x)]\,.$$

□

En vista de lo anterior, se obtiene el algoritmo Monte Carlo (6.5.3) para aproximar la solución de (1).

Algoritmo 6.5.3 Solución de Monte Carlo para EDPs parabólicas de la forma (1).
Dado: *Fijar $T > 0$, $n \in \mathbb{N}$, $N \in \mathbb{N}$.*

1. *Definir una malla temporal $0 = t_0 <, \ldots, < t_n = T$.*
2. *Definir N procesos, como copias independientes de la aproximación de Euler de la EDE (2).*
$$\left(X_{t_i}^{x,n,k}\right)_{\substack{k=1,\ldots,N \\ i=0,\ldots,n}}\,.$$
3. *Aproximar (1) mediante:*
$$u^{n,N}(t_i,x) = \frac{1}{N}\sum_{k=1}^{N} f\left(X_{T-t_i}^{x,n,k}\right), \quad \forall i = 0,\ldots,n\,.$$

Observación 6.5.5. Puede probarse que la aproximación de Euler tiene error de orden $\frac{1}{n}$ (antes de usar Monte Carlo) en el sentido que

$$\mathbb{E}[f(X_t^x)] - \mathbb{E}[f(X_T^{x,n,1})] \sim \frac{1}{n}\,.$$

Referencias

[1] Dudley, R. M. (1971). P. Billingsley, Convergence of probability measures. Bulletin of the American Mathematical Society, 77(1), 25-27.

[2] Glasserman, P. (2013). Monte Carlo methods in financial engineering (Vol. 53). Springer Science & Business Media.

[3] Pardoux, É. (2006). Processus de Markov et applications. Algorithmes, Réseaux, Génome et Finance.

[4] Asmussen, S., & Glynn, P. W. (2007). Stochastic simulation: algorithms and analysis (Vol. 57). Springer Science & Business Media.

[5] Norris, J. R. (1998). Markov chains (No. 2). Cambridge university press.

[6] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. The journal of chemical physics, 21(6), 1087-1092.

[7] Robbins, H., & Monro, S. (1951). A Stochastic Approximation Method. The Annals of Mathematical Statistics, 22(3), 400–407.

[8] Bottou, L. (1998). On-line learning and stochastic approximations. In On-Line Learning in Neural Networks, 9–42.

[9]] G. Cybenko. (1989). Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems, 2(4), 303–314.

[10] K. Hornik, M. Stinchcombe, & H. White. (1989). Multilayer feedforward networks are universal approximators. Neural networks, 2(5), 359–366.

[11] Karatzas, I., & Shreve, S. (2012). Brownian motion and stochastic calculus (Vol. 113). Springer Science Business Media.

Anexo A. Laboratorios

A.1. Laboratorio 1

Monte Carlo y eficiencia de simulación

Preliminar

1. Programe el método *DiscreteQuantile(f,u)*, que recibe como parámetros una función de masa discreta f y un vector $u \in [0,1]^r$, y retorne el menor vector $n \in \mathbb{N}^r$ (coordenada a coordenada) tal que $\sum_{j=0}^{n_i} f(j) \geq u_i$.
2. Programe el método *DiscreteQuantileF(F,u)*, que recibe como parámetros una función de distribución F y un vector $u \in [0,1]^r$, y retorne el menor vector $n \in \mathbb{N}^r$ tal que $F(n_i) \geq u_i$.

Problema 1

En *python* existen distintas librerías que permiten simular variables aleatorias. Entre ellas se destacan las siguientes:

- *Numpy*
- *Scipy*
- *Random*

1. Programe la función `uniforme`, que reciba un valor entero n y un método (*Numpy*, *Scipy* o *Random*), y retorne n simulaciones de una variable aleatoria uniforme en $[0,1]$.
2. Genere 10^6 uniformes para cada librería y grafique en un histograma cada muestra generada, utilizando la librería *seaborn*.
3. Genere 100 muestras de 1000 uniformes en $[0,1]$ y utilizando la librería *time* calcule los tiempos de ejecución que toman generar cada muestra para cada librería. Grafique los tiempos encontrados para cada método y calcule la media y varianza del tiempo de ejecución por muestra de cada método.
4. En base a los resultados anteriores. ¿Cual es el mejor método a utilizar? Argumente.
5. Genere funciones que permitan obtener una muestra para las siguientes variables a partir de uniformes:
   - $Bernoulli(p)$
   - $Binomial(p, N)$
   - $Geometrica(p)$

   Utilice estas funciones y la librería *random* para generar muestras de estas variables. Compare los tiempos de ejecución con los métodos para simular directamente estas variables disponibles en las librerías *numpy* y *scipy* (note que $Bernoulli(p) = Binomial(p, 1)$).

Las librerías anteriores generan números **pseudoaleatorios** que se asemeja bastante a lo que se necesita.

6. Averigüe y explique en qué consisten los métodos para generar números pseudoaleatorios uniformes en [0,1] disponibles en la versión que se usará de Python. Especifique: número de bits, período (de congruencias lineal utilizada o medida equivalente para el método que corresponda), posibilidad y manera de cambiar semilla. Utilizar aproximadamente media plana de desarrollo incluyendo tablas y/o figuras.

Problema 2

Tomando en cuenta que

$$I = \frac{\pi}{4} = \int_0^1 \sqrt{1-x^2}\ \mathrm{d}x = \int_0^1 \int_0^1 \mathbf{1}_{\{x^2+y^2\leq 1\}}\ \mathrm{d}x\mathrm{d}y,$$

se considerarán dos métodos de Monte Carlo para calcular numéricamente I:

- Utilizando la variable aleatoria $X = \sqrt{1-U^2}$, con U v. a. uniforme en $[0,1]$.
- Utilizando la variable aleatoria $Z = \mathbf{1}_{\{U_1^2+U_2^2\leq 1\}}$, con U_i v. a. uniforme en $[0,1]$ e independientes.

1. Calcule las varianzas $\mathrm{Var}(X)$ y $\mathrm{Var}(Z)$ de forma teórica y de forma simulada con diferentes cantidades de réplicas n. Grafique. Estime una cantidad de réplicas necesarias para X y Z con tal de obtener una aproximación de la varianza con un error del orden del $1\,\%$.

2. Calcule la cantidad de réplicas necesarias para X y Z con tal de aproximar I con un error máximo de $\mathrm{Err}_1 = 0.1$ y probabilidad $\mathrm{Pr}_1 = 90\,\%$. Haga el mismo ejercicio con $\mathrm{Err}_2 = 0.01$ y $\mathrm{Pr}_2 = 95\,\%$, $\mathrm{Err}_3 = 0.001$ y $\mathrm{Pr}_3 = 99\,\%$.

3. Aproxime las esperanzas $\mathbb{E}(X)$ y $\mathbb{E}(Z)$ de forma simulada con diferentes cantidades de réplicas n hasta llegar al n^* tal que se cumple $Err3$ y $Pr3$.

   - Grafique las aproximaciones en función de la cantidad de réplicas.
   - Grafique el tiempo utilizado en aproximar las esperanzas en función de la cantidad de réplicas.
   - Estime los costos de simular una réplica de X y una réplica de Z.

4. Considerando $Err3$ y $Pr3$ calcule un intervalo de confianza para I utilizando X y Z. Mida el tiempo total utilizado por cada método para obtener dicha precisión y compare los errores de estimación. Compare los costos totales para cada método ¿Cuál método es más eficiente?

5. Considerando Err_3 y Pr_3 calcule el costo teórico de estimar I utilizando X y Z, tomando como costo la cantidad de variables aleatorias uniformes necesarias ¿Cuál método es mas eficiente bajo este criterio? ¿Qué diferencia se observa entre comparar las eficiencias usando este criterio (número de uniformes) y el criterio anterior (costo total)? ¿Qué indica esa diferencia? ¿Cuál criterio debería preferirse en general?

Problema 3

1. Programe el método "NewtonRaphson", que recibe como parámetros una función de distribución F, su función de densidad f y un vector $u \in [0,1]^r$, y aplique el método de Newton-Raphson para calcular el vector $x \in R^r$ tal que $|F(x_i) - u_i| \leq error$, donde error es un parámetro de la clase inicializado con $error = 10^{-4}$.

   Decimos que X es una variable $Beta$ de parámetros θ_1, $\theta_2 > 0$ si su función de densidad f_X cumple

   $$f_X(x) = \frac{x^{\theta_1 - 1}(1-x)^{\theta_2 - 1}}{B(\theta_1, \theta_2)} \mathbf{1}_{[0,1]},$$

   donde $B(\theta_1, \theta_2) = \int_0^1 t^{\theta_1 - 1}(1-t)^{\theta_2 - 1} dt$. Para esta función puede usar el método de la biblioteca *Scipy*.

2. 2 - Grafique en una misma figura la función de densidad para:

   a) $\theta_1 = 2$, $\theta_2 = 5$,
   b) $\theta_1 = 2$, $\theta_2 = 2$,
   c) $\theta_1 = 1$, $\theta_2 = 3$,
   d) $\theta_1 = 0.5$, $\theta_2 = 0.5$.

   De ahora en adelante fijamos los parámetros $\theta_1 = 2$ y $\theta_2 = 5$.

3. Utilice el método "NewtonRaphson" para simular 10000 réplicas de $X \sim Beta(\theta_1, \theta_2)$. Grafique los resultados.

4. Programe el método de "AceptacionRechazo" que tome una función de densidad f definida en $[0,1]$, una cota apropiada K y dos vectores $u, v \in [0,1]^r$ y retornen réplicas de de una v.a. X de densidad f usando el método de aceptación rechazo usando v.a. auxiliares de densidad $g \sim Unif([0,1])$.

5. Encuentre una cota K para la variable $Beta(\theta_1, \theta_2)$, con θ_1, θ_2 como en el punto anterior. Implemente el método de aceptación-rechazo con f_X y K para simular réplicas de $Beta(\theta_1, \theta_2)$ usando 10000 uniformes. Grafique los resultados.

6. - Para $k = 3, \ldots, 5$, simule $n = 10^k$ réplicas de X para cada uno de los métodos implementados. Para cada k grafique los histogramas de las muestras obtenidas.

   - ¿En que medida coinciden los resultados ambos métodos y por qué?
   - Grafique el tiempo de ejecución en función de la cantidad de réplicas, estime el costo por réplica de cada uno de los métodos y ordene los métodos según su eficiencia.

7. Usando el método más eficiente, simule $n = 100000$ réplicas de X, calcule las medias y varianzas muestrales, y luego compare los resultados con los valores teóricos.

Problema 4

Considere $Y_{\lambda,s}$ variable aleatoria discreta con

$$\mathbb{P}(Y_{\lambda,t,s} = k) = \frac{e^{-\lambda}\lambda^k/k!}{\sum_{j=t}^{s} e^{-\lambda}\lambda^j/j!} \text{ para } k = t, \ldots, s\,.$$

Para las evaluaciones considere $\lambda = 5$.
Analizaremos dos métodos que reciban t y s, y simulen n réplicas de $Y_{\lambda,t,s}$.

1. Implemente el método de aceptacion-rechazo utilizando variables uniformes discretas.
2. Implemente el método de simulación condicional de una variable aleatoria apropiada.
3. Evalue la eficiencia teórica de ambos métodos.
4. Compare la precisión numérica de los dos métodos considerando sus histogramas y las eficiencias numéricas para los siguientes casos:
   - $t = 0$, $s = 10$.
   - $t = 10$, $s = 20$.

A.2. Laboratorio 2

Reducción de varianza y cadenas de Markov

Problema 1: Reducción de varianza

Considere la cantidad

$$\alpha = [e^{bZ}\mathbf{1}_{Z>0}],$$

donde Z es una variable normal estándar y $b \in$ es una constante. Supondremos para este problema que la normal estándar es la única variable eficientemente simulable. Se desea aproximar α mediante un algoritmo de Monte Carlo con baja varianza.

1. Proponga un método de muestreo preferencial.
2. Sabiendo que $\exp(bZ) = \exp(b^2/2)$, proponga un método de variable de control.
3. Mejore el método del ítem anterior usando una variable antitética.

Disponemos entonces de cuatro métodos de Monte Carlo para aproximar α: usando directamente la ecuación para α, y los tres métodos anteriores propuestos por usted. En lo que sigue, trabaje con $b = 2$.

4. Aproxime numéricamente la raíz de la varianza de la variable aleatoria que da lugar a cada uno de los cuatro métodos, para distintos tamaños de muestras, y grafique. Obtenga una estimación de la raíz de la varianza en cada caso.
5. Usando la estimación de la raíz de la varianza del punto previo y aproximando con el TCL, calcule el tamaño de muestra necesario para cada método de modo de que el error obtenido sea inferior a $\varepsilon = 0.02$ con probabilidad de 95 %. Comente.
6. Sea $N_{\max}$ el tamaño de muestra máximo entre los calculados en el ítem anterior para los tres métodos propuestos por usted (es decir, excluyendo el método que usa directamente la ecuación para α). Para distintos tamaños de muestra crecientes hasta $N_{\max}$, obtenga la estimación de α de cada uno de los cuatro métodos y grafique.
7. En base a lo obtenido en los puntos previos, ¿cuál método es el mejor y cuál el peor? Obtenga el valor exacto de α usando una herramienta adecuada (por ejemplo: www.wolframalpha.com), y compare con el valor entregado por los cuatro métodos, usando el mismo $N_{\max}$ para todos. Comente.

Problema 2: Simulación de cadenas de Markov y flujos Markovianos.

Sea E un conjunto finito que supondremos sin perder generalidad igual a $\{1, \ldots, N\}$, donde $N \in \mathbb{N}$ está fijo. Sea $f : E \times [0, 1] \to E$ una función, X_0 una v.a. con ley μ y $(U_n)_{n\geq 1}$ una colección

de v.a.'s i.i.d. uniformes en $[0,1]$, independientes de X_0. Para $n \geq 0$ se define por recurrencia la sucesión aleatoria

$$X_{n+1} = f(X_n, U_{n+1}).$$

Se sabe del curso que $(X_n)_{n\geq 1}$ es una cadena de Markov homogénea.

1. 1 - Sea P una matriz estocástica indexada por E. Si f es tal que $\mathbb{P}(f(x,U) = y) = P_{xy}$ para todo $x, y \in E$, diremos que es una *función de transición* asociada a la matriz P. Muestre que $f(x,u) := \inf\{y \in E : \sum_{z=1}^{y} P_{xz} \geq u\}$ cumple esa condición, y que la cadena $(X_n)_{n\in\mathbb{N}}$ así construida tiene entonces matriz de transición P.

   Note que dada **una** v.a. uniforme U en $[0,1]$, $\Phi := f(\cdot, U) : E \to E$ es una función aleatoria, que entrega transiciones de la cadena desde un estado x cualquiera, a algún estado $y = \Phi(x) = f(x,U)$.

2. Programe una función $Trans(x,u,P)$ que tiene como parámetros un valor u en $[0,1]$ y una matriz estocástica P indexada por $\{1,...N\}$, y entrega **para cada** x el valor correspondiente de la función de transición asociada a P con el parámetro u dado. Puede usar para ello una función ya programada en el Laboratorio 1 si lo desea.

   En base a lo anterior, construya también un método $CM(u,\mu,P)$ que tome un vector u de n uniformes y simule n pasos de la cadena de Markov homogénea con matriz de transición P y distribución inicial μ.

3. Usando las funciones antes construidas, y utilizando (solo) $n = 100$ v.a. uniformes, simule y grafique $n = 100$ pasos de $K = 10$ trayectorias de un paseo aleatorio en el conjunto $\{1,...,N\}$ para $N = 10$, donde cada trayectoria parte de un estado distinto. En los extremos $x = 1, 10$ el paseo aleatorio se queda en el estado actual con probabilidad $1-p$ y salta con probabilidad p. Realice esto en cada uno de los 3 casos siguientes: $p = 1/2$, $p = 1/3$, $p = 2/3$.

   Consideremos el planeta Eternium, que yace en el centro del universo. Gracias a su particular ubicación, el clima de un día depende únicamente del día anterior. Supongamos primeramente que los estados no-lluvioso y lluvioso obedecen:

   - 10 % de probabilidad de lluvia después de un día no-lluvioso, 90 % de probabilidad de volver a tener un día no-lluvioso
   - Equiprobabilidad de día lluvioso/no-luvioso después de un día lluvioso.

4.a Modele el problema utilizando como una cadena de Markov. Derive la medida invariante de manera teórica e interprete su significado.

4.b Simule una cantidad apropiada de pasos, grafique y compare con lo anterior.

   Gracias a un análisis más avanzado obtenemos un modelo con cuatro estados: nieve, lluvia, parcial y soleado. Además, sabemos que las transiciones entre ellos siguen pueden ser modeladas con la matriz

$$P = \begin{bmatrix} 0.3 & 0.35 & 0.25 & 0.1 \\ 0.05 & 0.45 & 0.3 & 0.2 \\ 0 & 0.15 & 0.55 & 0.3 \\ 0 & 0.1 & 0.4 & 0.5 \end{bmatrix}$$

4.c Simule el clima de este planeta por 10 años. Grafique y derive sus estadísticas climatológicas anuales.

Problema 3: Aplicación a un modelo de colas

Considere una cola a tiempo discreto a la que, en cada instante $n \in$ llega un cliente con probabilidad $p \in (0,1)$ y no llegan clientes con probabilidad $1-p$. Durante cada intervalo de tiempo en que hay al menos un cliente en la cola, un cliente es atendido y sale de la cola con probabilidad $q \in (0,1)$ y no se va ningún cliente con probabilidad $1-q$. Denote por X_n la cantidad de clientes en la cola en el instante n.

1. Escriba X_n como $X_n = F(X_{n-1}, Y_n, Z_n)$ explícitamente en términos de v.a.'s $(Y_n) \sim Bernoulli(p)$ y $(Z_n) \sim Bernoulli(q)$ independientes. Justifique que (X_n) es irreducible. Se puede probar que para $p > q$, la cadena (X_n) diverge c.s. Además, en el caso $p = q$ la medida $(1-p, 1, 1, \dots)$ es invariante, y para $p < q$ lo es la medida

$$\pi_0 = \frac{q-p}{q}, \qquad \pi_x = \left(\frac{p(1-q)}{q(1-p)}\right)^x \frac{q-p}{q(1-q)}, \quad x \geq 1.$$

   Explicite el rango de parámetros para los que (X_n) es recurrente positiva, recurrente nula o transiente.

2. En lo que sigue se considerará una versión "truncada" (y simulable) de la cadena. Para ello, se fijará N un número máximo de clientes que la cola permite, y se modifica la matriz de transición con la convención de que si ya hay N clientes, un nuevo cliente simplemente no se queda, pero la dinámica de los que ya están es la misma de antes. Simule y grafique trayectorias de (X_n) hasta $n = 1000$ para 3 pares de valores representativos de (p, q), fijando en cada caso valores convenientes de N que deberá determinar.

3. Para un par p, q tal que $p < q$ y un valor N fijos que usted determine, compare las siguientes 3 estrategias para estimar numéricamente la distribución invariante π^N asociada:

   - Simular K (grande) CM independientes en un tiempo T (grande) y obtener el histograma correspondiente a la *medida empírica* para las K trayectorias en ese tiempo.
   - Simular una CM por un tiempo T (grande) y obtener el histograma de las *medias ergódicas*

$$\frac{1}{T}\sum_{k=1}^{T} \mathbf{1}_{\{X_k = i\}}, \quad i \in \{0, ..., N\}.$$

   En todos los casos, puede fijar K y/o T en términos de la cantidad de uniformes a simular o según un cierto error. Compare los métodos en distancia en variación total a π, y haga un análisis completo de las distintas estrategias.

A.3. Laboratorio 3

Algoritmos estocásticos usando cadenas de Markov

Problema 1: Modelo de Ising en $\mathbb{Z}^2$

Queremos modelar el ferromagnetismo en una placa metálica plana idealizada. Específicamente para $N \in \mathbb{N}$ (grande), las moléculas de la placa se ubican en la grilla 2-dimensional $\Lambda = \Lambda_N = \{-N, \ldots, N\}^2 \subseteq \mathbb{Z}^2$, y cada molécula posee un momento magnético o *spin*, el cual puede estar orientado hacia arriba o hacia abajo. Por lo tanto, el conjunto de posibles configuraciones es $E_N = \{-1, 1\}^\Lambda$. Dado $x \in E_N$, y $m \in \Lambda$, denotamos $x(m) \in \{-1, 1\}$ el spin del sitio m en la configuración x. Trabajaremos en el espacio E de configuraciones con spin fijo hacia arriba en el borde:

$$E = \{x \in E_N : x(m) = 1, \forall m \in \partial\Lambda\}, \quad \text{donde } \partial\Lambda = \Lambda_N \setminus \Lambda_{N-1}.$$

En un material ferromagnético, los spins de sitios cercanos tienen tendencia a alinearse; es decir, spins iguales en sitios contiguos tienen asociada menor energía. Específicamente, la energía de una configuración $x \in E_N$ viene dada por

$$H(x) = \sum_{m \sim m'} [x(m) - x(m')]^2,$$

donde $m \sim m'$ denota que m y m' son vecinos en la grilla, es decir, que están a distancia 1 (cada par $m, m' \in \Lambda$ aparece una sola vez en la sumatoria). Supondremos que la probabilidad de que el sistema se encuentre en la configuración $x \in E$ está dada por

$$\pi_x = \frac{e^{-\beta H(x)}}{Z(\beta)},$$

para $\beta > 0$ dado ($1/\beta$ es la temperatura), y $Z(\beta) = \sum_{y \in E} e^{-\beta H(y)}$ es la constante de normalización. Observe que lo anterior hace improbables las configuraciones con mayor energía, lo cual es consecuente con lo que ocurre en los sistemas físicos. Se desea simular realizaciones de la distribución $\pi \in \mathcal{P}(E)$.

1. Calcule $|E|$. Argumente por qué es imposible en la práctica calcular $Z(\beta)$ explícitamente, incluso para N pequeño (por ejemplo, $N = 10$).

Debido a lo anterior, no es posible simular π directamente. Por este motivo, utilizaremos un algoritmo tipo *Markov chain Monte Carlo* (MCMC) para realizar simulaciones aproximadas de π, según lo descrito en cátedra. Para esto, siga los siguientes pasos.

2. Consideremos el grafo G sobre E en el cual xy es una arista de G si y sólo si x e y difieren en exactamente un sitio, es decir, si y sólo si

$$\exists m_0 \in \Lambda \setminus \partial\Lambda, \text{ tal que } x(m_0) = -y(m_0), \text{ y además } x(m) = y(m), \forall m \neq m_0.$$

Escriba la matriz estocástica R asociada a este grafo para el algoritmo de Metropolis y el de Gibbs. Obtenga una expresión explícita simple para $\frac{\pi_y R_{yx}}{\pi_x R_{xy}}$ en ambos casos ¿Cuál algoritmo es más conveniente usar y por qué?

3. Describa el algoritmo MCMC correspondiente a este caso usando pseudo-código. En base a él, programe un método `X=Ising(N,beta,nf)` que simule `nf` pasos de la cadena, grafique su estado cada cierta cantidad de pasos (mostrando la grilla Λ y asociando un color al spin -1 y otro a 1), y retorne el estado final `X`, en caso que `nf` sea finito. Escoja los spins iniciales independientes con ley $2 \cdot \text{Bernoulli}(p) - 1$, para algún $p \in (1/2, 1)$ (por ejemplo $p = 2/3$ ó $p = 3/4$). Haga esto para 2 valores de p distintivos. Comente la diferencia o similitud observadas entre ambos casos y la relevancia de dicha elección. Fije un valor para todo lo que viene.

4. En lo que sigue, fije $N \in \{50, \ldots, 200\}$ (lo más grande que se pueda, mientras la simulación sea fluida). Observe el comportamiento de la cadena en el tiempo largo para un β pequeño y otro grande (digamos, escoja un $\beta < 0.1$ y otro $\beta > 5$). Grafique y comente.

Lo observado en el punto anterior se conoce como *transición de fase:* existe un valor crítico $\beta_C > 0$ tal que para cualquier $\beta > \beta_C$ y N grande, π asigna probabilidad casi 1 a configuraciones con spin hacia arriba en la gran mayoría de los sitios (con la condición de borde que usamos), mientras que cuando $\beta < \beta_C$ se observa coexistencia de ambos spins. A continuación estudiaremos este fenómeno.

5. Fije `nf` grande (en el orden de los millones), de modo que el algoritmo tenga tiempo de acercarse a π. Fije una malla del intervalo $[0, 1]$ de distintos valores de β. Para cada β en la malla, obtenga el estado X de la cadena luego de `nf` pasos, y calcule el *spin medio* $s = \frac{1}{|\Lambda|} \sum_{m \in \Lambda} X(m)$. Grafique s en función de β, y estime visualmente el valor crítico β_C.

6. Repita lo anterior en un intervalo más pequeño centrado en su estimación de β_C, y con `nf` más grande aún, de modo de obtener una estimación más fina. Grafique. Averigüe el valor exacto de β_C y compárelo con su estimación.

Problema 2: Problema del vendedor viajero

Considere un conjunto $\{1, 2, ..., N\}$ de ciudades en el dominio plano $[0, 1]^2$. El problema del vendedor viajero consiste en encontrar un ciclo que recorra todas las ciudades una y solo una vez, partiendo y terminando en la primera, que minimice la distancia recorrida.

Sea $E = \{\sigma \in S_N : \sigma(1) = 1\}$ el conjunto de todas las posibles rutas que empiezan en 1. Es fácil verificar que $|E| = (N-1)!$, por lo que si consideramos el problema con 15 ciudades, la cantidad de rutas posibles es 87.178.291.200, por lo cual es claro que es prácticamente imposible recorrer todas las posibilidades. La idea es construir un algoritmo estocástico, denominado *recocido simulado (simulated annealing)*, para minimizar la función de distancia total recorrida

$$w(\sigma) = \sum_{i=1}^{N} d(\sigma(n), \sigma(n+1))$$

en donde $d : \{1, ..., N\}^2 \to \mathbb{R}_+$ es la distancia euclidiana usual entre dos ciudades, y se usa la convención $\sigma(N+1) = 1$ para $\sigma \in E$. Definimos el grafo G sobre E dado por la siguiente relación

de adyacencia: $\sigma \sim \tau$ si y sólo si τ se obtiene permutando exactamente 2 ciuidades de σ.

1. Programe una función que genere N ciudades uniformemente distribuidas en $[0,1]^2$, y que genere luego la matriz $D = (D_{ij})_{i,j=1}^N$ de distancias, donde D_{ij} es la distancia entre la ciudad i y la ciudad j.

2. Programe una función que dado un camino que recorre las N ciudades, en el orden dado por la permutación σ, grafique dicho camino.

3. Se define $osc_K(w) = \max\{w(\tau) - w(\sigma) : \sigma \sim \tau\}$. Dé una cota superior para $osc_K(w)$ para cada N, que no dependa de la posición de las ciudades.

4. Considere una sucesión de temperaturas inversas $\beta_n = \frac{1}{C}\ln(n+e)$ con $C > (N-1)osc_K(w)$ y una cadena de Markov (X_n) no homogénea tal que, en el tiempo n, su matriz de transición está dada por la matriz de la cadena $(X_k^{\beta_n})$, según el método visto en cátedra. Es decir: dado $X_n = \sigma$, se escoge un vecino $\tau \sim \sigma$ uniformemente al azar, y con probabilidad $e^{-\beta_n(w(\tau)-w(\sigma))} \wedge 1$ se define $X_{n+1} = \tau$; si no, se mantiene $X_{n+1} = \sigma$. Se puede probar que dicha cadena converge en probabilidad a una variable aleatoria distribuida uniformemente en el conjunto de mínimos globales de la función w. Implemente un método que simule esta cadena para un estado inicial que usted escoja.

5. Para $N = 20$ ciudades fijas, encuentre un mínimo global aproximado de la función w. Pruebe con sucesiones β_n de distintas formas, por ejemplo lineal, cuadrática, exponencial, etc. Grafique en cada caso la evolución de la función w evaluada en el estado de la cadena, durante el tiempo de ejecución del algoritmo (el que usted deberá determinar dependiendo de la sucesión β_n escogida). Grafique para algunos tiempos representativos los caminos respectivos.

A.4. Laboratorio 4.1

Descenso de gradiente estocástico y aplicaciones

Problema 1: Regresión Lineal

Dado un conjunto de datos $(x_i, y_i)_{i=1}^n \subset \mathbb{R}^m \times \mathbb{R}$, se propone la siguiente relación entre sus componentes

$$y = \theta^T x + \varepsilon\,,$$

en donde ε es una variable aleatoria con valor esperado 0 y distribución desconocida. El problema de regresión lineal consiste en encontrar el parámetro θ tal que el conjunto de datos satisfaga la ecuación anterior de manera que $\mathrm{Var}(\varepsilon)$ sea lo más pequeña posible. Esto conlleva al problema de optimización

$$\hat{\theta} = \arg\min_{\theta} \mathrm{Var}(\varepsilon) = \arg\min_{\theta} \mathbb{E}\left((y - \theta^T x)^2\right).$$

El objetivo de esta pregunta es aplicar el modelo anterior sobre el conjunto de datos **Diabetes** y estimar el mejor parámetro posible utilizando el método de descenso de gradiente estocástico.

1.a Cargue el conjunto de datos utilizando el siguiente código.

```
from sklearn.datasets import load_diabetes

df, target = load_diabetes(return_X_y=True,as_frame=True)
print(load_diabetes().DESCR)
print("Target variable statistcs:\n"+str(target.describe()))
df.head()
```

Observe la cantidad de variables y el tipo de datos que posee. ¿Que complicaciones pueden surgir de usar un modelo predictivo en un contexto real? Reflexione acerca de la naturaleza del dataset y la tarea que se quiere ejecutar.

Nos referimos a estandarizar cuando forzamos los datos a tener una distribución normal estándar. Para esto, reemplazamos cada punto de dato x_i por:

$$\tilde{x}_i = \frac{x_i - \mu}{\sigma}$$

En donde μ es la media de la columna y σ es su desviación estándar.

1.b Estandarice los datos (sin usar funciones de pre-procesamiento) para que el modelo a trabajar funcione. Extienda la base de datos (agregando una columna) para obtener un modelo de regresión lineal, esta vez representado por una función afín de la forma $y_i = \theta^T x_i + b + \varepsilon_i$, con $\theta \in \mathbb{R}^m$, $b \in \mathbb{R}$.

Nota: no estandarice la columna objetivo.

1.c Separe los datos en un conjunto de entrenamiento y otro de prueba según la proporción 80 % y 20 %, Justifique brevemente por qué esto es necesario. Le será útil la función **train_test_split** de la biblioteca **scikit-learn**.

En lo que sigue justifique sus respuestas graficando la función de costos cada cierta cantidad de iteraciones. Cuando se pida comparar diferentes implementaciones debe realizarlo en base al conjunto de datos de prueba y el error cuadrático medio incurrido con la estimación obtenida. Tanto la cantidad de iteraciones como los parámetros pueden ser escogidos libremente.

2. Implemente el algoritmo de descenso de gradiente estocástico para un modelo de regresión lineal, especificando cuál es la función de costos y su gradiente. Considere los siguientes casos

   a) *Learning rates* constantes.

   b) *Learning rates* variables (proponga al menos 2).

   Proponga al menos dos en cada caso. Compare los resultados obtenidos para cada elección ¿Cuál es la mejor elección?

3. Modifique el algoritmo anterior para trabajar con *mini-batch*. Pruebe el desempeño (nuevamente en términos del error cuadrático medio para el conjunto de prueba) del algoritmo para distintos tamaños de *mini-batch* y *learning rates* y justifique cual es el mejor ¿Existe alguna relación entre ambos parámetros?

4. Implemente las siguientes variantes de descenso de gradiente estocástico y compare el desempeño de estos con los algoritmos implementados en las partes anteriores

   a) **Momentum:** El método consiste en ir generando los pasos de descenso como

   $$m_i = \beta m_{i-1} + (1-\beta)\nabla_\theta f(\theta_i, x_i), \quad m_0 = 0,$$

   tal que

   $$\theta_{i+1} = \theta_i - \eta\, m_i,$$

   donde $\beta \in (0,1)$ (debe ser elegido, asi como también η).

   b) **Adagrad:** El *learning rate* es variable y se genera de la siguiente manera:

   $$\eta_i = \frac{\eta}{\sqrt{v_i + c}},$$

   donde $c > 0$, $\eta > 0$ y

   $$v_i = \sum_{j=1}^{i} \|\nabla_\theta f(\theta_j, x_j)\|^2, \quad v_0 = 0.$$

5. (Opcional) Realice una búsqueda de grilla (*gridsearch*) para escoger los mejores hiper-parámetros para Momentum y Adagrad. Dado un *batch size* m, reporte los errores del test set en un mapa de calor para valores de β y ν (respectivamente c y ν) de su elección. Repita para otro valor de m. Describa los resultados y concluya.

Problema 2: Redes Neuronales

El objetivo de este problema es explorar el uso de una biblioteca de aprendizaje profundo. Más precisamente, implementaremos modelos de redes neuronales para clasificar imágenes usando la biblioteca *PyTorch*. Para esto usaremos la base de datos *MNIST*, que consiste en dígitos escritos a mano. La tarea consiste en entrenar un modelo que identifique de manera automática el dígito en cuestión. Para importar el conjunto, se debe descargar el archivo *mnist.pkl.gz* y ejecutar el siguiente código.

```
from sklearn.datasets import load_diabetes

df, target = load_diabetes(return_X_y=True,as_frame=True)
print(load_diabetes().DESCR)
print("Target variable statistcs:\n"+str(target.describe()))
df.head()
```

Un **Tensor** en el contexto de aprendizaje de máquinas, corresponde a la generalización de una matriz a dimensiones más altas, similar al concepto de arreglo de bibliotecas como *numpy*. *PyTorch* utiliza tensores para las computaciones, con la característica especial de que poder operarlos de manera rápida usando GPUs.

```
import torch

x_train, y_train, x_valid, y_valid = map(torch.tensor, (x_train, y_train, x_valid, y_valid))
n, d = x_train.shape;
```

El primer objetivo será definir un modelo que conste de una sola capa lineal, lo cual equivale a una regresión logística si le aplicamos la función *softmax*. Definimos los parámetros de nuestra regresión como tensores, los cuales serán después optimizados. A modo de ejemplo, mostramos el resultado de darle un "mini-batch.[a] los parámetros de nuestro modelo no entrenado.

```
import math

pesos = torch.randn(784, 10) / math.sqrt(784)
pesos.requires_grad_()
sesgo = torch.zeros(10, requires_grad=True)
batch_size = 64
x_batch = x_train[0:batch_size]
preds = x_batch @ pesos + sesgo
```

PyTorch guarda la información de los pasos aplicados a tensores. Esto se usa posteriormente para computar los gradientes para cada parámetro, los cuales se usarán en Descenso de Gradiente. Este método se llama Diferenciación Automática.

Necesitamos definir una función de pérdida a minimizar. Esta sección se concentrará en justificar el uso de la función de pérdida entropía cruzada. Nos gustaría ajustar los parámetros de tal forma que la distribución de probabilidad dada por la red se asemeje lo más posible a la real distribución de

los datos. Como solo tenemos observaciones de estos datos, para lo anterior buscaremos el estimador de máxima verosimilitud. Por ende queremos maximizar

$$\prod_{i=1}^{N} p_\theta(y^{(i)}|x^{(i)})$$

donde $y^{(i)} \in \{c_1, \ldots, c_k\}$ (las clases posibles) y θ son los parámetros de la red, que a su vez computa $p_\theta(c_j|x)$ en un paso *forward* dado un punto de dato x. En otras palabras

$$\hat{\theta} = \arg\max_\theta \prod_{i=1}^{N} p_\theta(y^{(i)}|x^{(i)})$$

1.a Justifique que

$$\hat{\theta} = \arg\max_\theta \frac{1}{N} \sum_{i=1}^{N} log(P_\theta(y^{(i)}|x^{(i)}))$$

Sean p y q dos distribuciones de probabilidad discreta. La **entropía cruzada** está dada por

$$H(p,q) = \sum_x p(x) log(\frac{1}{q(x)}))\,.$$

Se define entonces la función de pérdida de entropía cruzada por

$$L(q,p) = \frac{1}{N} \sum_{i=1}^{N} H(p_i, q_i)$$

1.b Demuestre que

$$\hat{\theta} = \arg\min_\theta L(p_\theta, p)$$

donde p es la distribución de probabilidad empírica.

1.c Comente brevemente la relación entre la entropía cruzada y la teoría de la información.

La función de pérdida de entropía cruzada puede ser accedida en *PyTorch* mediante:

```
import torch.nn.functional as F

func_costo = F.cross_entropy
print(func_costo(x_valid @ pesos + sesgo,y_valid).item())
```

2.a Complete el siguiente código que implementa **Descenso de Gradiente Mini-batch** para ajustar los parámetros del modelo dada una tasa de aprendizaje.

```
def DescensoGradiente(pesos,sesgo,tasa):
        perdidas_epoch = []
        # rellenar acá
        cantidad_batches = None  # cambiar
        # ------------
        for i in range(cantidad_batches):
```

```
        # rellenar acá
        loss = None  # calcular pérdida
        # ------------
        loss.backward()
        with torch.no_grad():
            pesos -= pesos.grad * tasa
            sesgo -= sesgo.grad * tasa
            pesos.grad.zero_()
            sesgo.grad.zero_()
        perdidas_epoch.append(loss.item())
    return perdidas_epochnd(loss.item())
    return perdidas_epoch
```

2.b Implemente una función entrenar que tome un número de épocas y ejecute el método anterior en cada iteración, imprimiendo el valor de la función de pérdida. Guarde además estas pérdidas en una lista. (A una barrida completa al conjunto de entrenamiento se le llama época (o *epoch* en inglés).)

2.c Ejecute el bucle para K épocas y grafique lo encontrado. Utilice otras métricas para evaluar la calidad de la clasificación, ¿que clases son más y menos fáciles de reconocer con el algoritmo? Le puede ser útil el reporte de clasificación de scikit-learn y la siguiente función para obtener las predicciones:

```
logsoftmax = torch.nn.LogSoftmax(dim=0)

def predicciones(output):
    return torch.argmax(logsoftmax(output),axis=1)
```

La biblioteca *PyTorch* nos provee de variadas herramientas que pueden ser útiles para el entrenamiento. Una de ellas es el módulo torch.distributions. Esta nos provee de algunas variables aleatorias conocidas de las cuales podemos samplear. Por ejemplo, para samplear de una distribución Gaussiana de dimensión 10, centrada en cero y con diagonal $= I$ ejecutamos:

```
from torch.distributions import MultivariateNormal

normal = MultivariateNormal(torch.zeros(2), torch.eye(2))
normal.sample()
```

3 Usando esto, implemente la variante **Langevin Dynamics** del algoritmo de Gradiente Estocástico. Re-defina los parámetros del modelo y ajústelos usando esta variante. Grafique y comente. ¿Que desventajas puede tener usar este método para optimizar otros modelos?

Observación: le puede ser útil el método *.reshape()*, que funciona para tensores del mismo modo que para arreglos de numpy.

A continuación se implementaremos lo anterior pero con el procedimiento usual que se usa en *PyTorch*. Necesitamos definir nuestro modelo como una clase, que heredará atributos de la clase

torch.nn.Module. En *torch.nn* podemos encontrar numerosos "bloques"para armar modelos. En este caso, simplemente usamos una capa lineal.

Una parte esencial de nuestro modelo es el paso *forward*, que será aquel que se ejecute cuando llamemos a nuestro modelo en uno o varios puntos de datos. Para más detalles ver la documentación.

```
from torch import nn

class Reg_Logistica(nn.Module):
    def ___init___(self):
        super().___init___()
        self.lineal = nn.Linear(784, 10)

    def forward(self, xb):
        return self.lineal(xb)
```

Además, usamos el módulo torch.optim para entrenar las redes, lo cual hace más consciso el código de gradiente estocástico escrito anteriormente.

```
from torch import optim

modelo = Reg_Logistica()
opt_SGD = optim.SGD(modelo.parameters(), lr=learning_rate)
```

4.a En base a lo anterior, re-defina el método entrenar para que tome un optimizador del módulo *optim*, un modelo (sin entrenar) y un número de épocas. Ejecute el método con Descenso de Gradiente y grafique. Escoja dos métodos más del módulo *optim*, entrene modelos con ellos y compare.

4.b Re-defina el modelo usando una capa intermedia. Entrenelo usando alguna de los optimizadores. ¿Cómo se comparan los resultados de esta nueva red en ambos conjuntos con los modelos anteriores? Justifique.

5 (Opcional)Un tipo de red neuronal que funciona bien para el procesamiento de imágenes son las **redes convolucionales (*CNN*)**. Averigue a que corresponde un **Núcleo (*Kernel*)** en procesamiento digital de imágenes y resuma en un parrafo. Investigue y mencione alguna ventaja de usar redes convolucionales. A continuación complete el siguiente código usando redes convolucionales (en vez de capas lineales) y ReLu como no linealidad. Pruebe su red del mismo modo que antes y compare los resultados con las partes anteriores.

```
class CNN(nn.Module):
    def ___init___(self):
        super().___init___()
        # defina acá dos o más transformaciones convolucionales

    def forward(self, xb):
        xb = xb.view(-1, 1, 28, 28)
        # transforme xb componiendo las capas convolucionales con relu
        xb = F.avg_pool2d(xb, 4)
        return xb.view(-1, xb.size(1))
```

A.5. Laboratorio 4.2

Laboratorio 4.2: Integral estocástica y EDEs

Problema 1: Movimiento Browniano

1. Programe una función `BrownianTrajectories` que reciba como parámetros:

   - un vector x de N condiciones iniciales,
   - un tiempo final $T > 0$,
   - un entero $K > 0$,

   y que simule N copias independientes de un movimiento browniano $(B_t)_{t\in[0,T]}$ en la grilla $t = (t_0, t_1, \ldots, t_K)$, con $t_i = iT/K$, partiendo de las condiciones iniciales indicadas en el vector x. Debe retornar `[t, B]`, siendo `B` la matriz con todas las trayectorias simuladas.

2. Utilizando la función creada, genere y grafique $N = 50$ trayectorias brownianas partiendo de $x = 0$ hasta T muy grande ($T \geq 10^{10}$, por ejemplo), en una malla temporal suficientemente fina. Agregue los gráficos de las funciones L y $-L$, donde $L(t) := \sqrt{2t \log\log t}$. Observe y comente.

3. Fije un $T \in [1, 10]$ a gusto. Definimos el *valor absoluto* $|B|$ y el *máximo acumulado* M del browniano, como
$$|B|_t := |B_t| \quad \text{y} \quad S_t := \max_{s\leq t} B_t.$$

   a) Para $N = 3$ brownianos simulados, grafique las trayectorias de $|B|$ y S asociadas, comentando en qué se diferencian.

   b) Para N grande ($N \geq 10^5$, por ejemplo), grafique histogramas de $|B|_T$ y S_T. Observe, compare y comente. Busque en Google o en un libro el resultado matemático correspondiente a lo que ilustran estas simulaciones.

Problema 2: Resolución numérica de ecuaciones diferenciales estocásticas

Consideremos una ecuación diferencial estocástica genérica:

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t, \quad X_0 = x_0, \tag{A.1}$$

donde $(B_t)_{t\geq 0}$ es un movimiento browniano en $\mathbb{R}$, $x_0 \in \mathbb{R}$ está dado, y $b, \sigma : \mathbb{R} \mapsto \mathbb{R}$ son funciones conocidas. En esta parte del laboratorio deseamos resolver numéricamente esta ecuación, y evaluar el desempeño de los algoritmos utilizados.

1. Implemente una función `[t,B,E]=SDEEuler(N,K,T,x0,b,s)` que realice lo siguiente:

   - Genere una malla `t` del intervalo $[0, T]$ usando paso T/K.

- Genere N trayectorias brownianas independientes en dicha malla, retornando el resultado en `B`.
- Por cada trayectoria browniana, genere las aproximaciones de la solución de (A.1) mediante el esquema de Euler, retornando el resultado en `E`. Las variables `x0`, `b` y `s` corresponden a x_0, b y σ.

Utilice esta función para explorar la naturaleza de las soluciones en función de los coeficientes b y σ: realice pruebas y gráficos para $N = 1$ trayectoria y para variados b y σ que usted estime convenientes; por ejemplo, puede fijar b y considerar distintos σ que son múltiplos de una función fija, para ver los efectos del ruido en la ecuación. Deberá por tanto simular las trayectorias de los diferentes procesos usando la misma realización de un movimiento Browniano.

2. En el caso particular en que $b(x) = bx$ y $\sigma(x) = \sigma x$ para ciertas constantes b, σ, la ecuación (A.1) describe un *movimiento browniano geométrico*, y puede resolverse explícitamente:

$$X_t = x_0 \exp\left((b - \frac{\sigma^2}{2})t + \sigma B_t\right). \tag{A.2}$$

Para parámetros N, K, T, x_0, b y σ que usted estime convenientes, obtenga la aproximacione de Euler. Compare gráficamente contra la trayectoria de la solución exacta dada por (A.2) utilizando el mismo movimiento Browniano. Comente.

3. Siguiendo con el caso del movimiento browniano geométrico, en esta parte se desea cuantificar lo observado gráficamente en el punto anterior. Específicamente: si E_t^K denota la aproximacion de Euler de paso T/K al tiempo t , queremos estudiar

$$\mathbb{E}|X_T - E_T^K|, \tag{A.3}$$

y en particular ver a qué tasa converge a 0 en función de K. Se espera que lo anterior sea de orden $K^{-\alpha}$, para cierto $\alpha > 0$ por determinar experimentalmente. Para esto, fije $T = x_0 = b = \sigma = 1$ y un N a conveniencia (100 ó 1000 debería bastar), e implemente una función `TestSDEEuler()` que realice lo siguiente:

- Genere un vector creciente de distintos valores de K, desde K relativamente pequeño (~ 10 ó ~ 100) a un a K grande ($\sim 10^5$ o más). De preferencia, que sea un vector equiespaciado en escala logarítmica.
- Para cada K generado, obtenga N realizaciones de la aproximacion del esquema de Euler de (A.1). Utilice estas realizaciones para aproximar la esperanza (A.3) usando Monte Carlo, donde para cada trayectoria, X_T se calcula de manera exacta usando la expresión (A.2) con la misma realización de B_T que la del movimiento Browniano discretrizado utilizado en el esquema de Euler. Grafique en función de K.
- Mediante una regresión lineal entre $\log(K)$ y el logaritmo de las aproximaciones de (A.3) obtenidas en el punto anterior, obtenga una estimación de las tasa α del esquema y grafique los resultados de la regresión.

Comente los resultados obtenidos.

4. Muestre que $\mathbb{E}(X_T) = e^{bT}$. Usando un esquema de Euler y Monte Carlo, calcule ahora $\mathbb{E}(E_T^K)$ y haga un analisis similar al anterior para el denominado "error débill ": $|\mathbb{E}(X_T) - \mathbb{E}(E_T^K)|$ (en contraste con el error "fuerte" (A.3)). Comente.

Problema 3: Problema de Dirichlet en $\mathbb{R}^2$

En esta parte estudiaremos el *problema de Dirichlet en* $\mathbb{R}^2$:

$$\begin{cases} \Delta u(x) = 0, & x \in D \\ u(x) = f(x), & x \in \partial D, \end{cases} \tag{A.4}$$

donde $D \subseteq \mathbb{R}^2$ es un abierto acotado, y $f : \partial D \to \mathbb{R}$ es una función continua. Denotamos por $(B_t)_{t\geq 0}$ un movimiento Browniano en $\mathbb{R}^2$, es decir $(B_t)_{t\geq 0} = (B_t^1, B_t^2)_{t\geq 0}$ con B^1, B^2 movimientos Brownianos independientes en $\mathbb{R}$.

Se prueba (ver e.g. Karatzas-Shreve , "Brownian motion and stochastic calculus") que la solución de (A.4) tiene la representación probabilística

$$u(x) = \mathbb{E} f(B_{\tau^x}^x) \quad \forall x \in \bar{D}, \tag{A.5}$$

donde $(B_t^x)_{t\geq 0}$ es un movimiento Browniano en $\mathbb{R}^2$ partiendo de x, es decir $(B_t^x)_{t\geq 0} = (B_t + x)_{t\geq 0}$, y $\tau^x = \inf\{t \geq 0 : B_t^x \notin D\}$ es el tiempo de parada en que B_t^x sale de D.

1. Implemente una función que simule N trayectorias Brownianas independientes en $\mathbb{R}^2$ partiendo de $(0, 0)$, en una grilla temporal de paso h, durante K pasos. Defina una función que implemente una grilla espacial fina de $D = [-1, 1]^2$ y $D = B((0, 0), 1)$, de ancho $\epsilon > 0$. Usando (solo) N trayectorias Brownianas independientes partiendo de $(0, 0)$ implemente en base a lo anterior una función que simule para cada x en la grilla, N trayectorias Brownianas partiendo de x, durante K pasos de paso h (MUY IMPORTANTE: en total solo deben usarse N trayectorias Brownianas, no $N\times$(número de puntos en la grilla)).

2. Programe una función que, dada $N = 1$ trayectoria (discretizada) simulada $(B_{ih}^x)_{i=1}^K$ de movimiento Browniano de K pasos de paso h partiendo de x, retorne $B_{\tau^x \wedge (Kh)}^x$, es decir, el punto en la frontera por el cual el proceso salió de D, o bien su posición en el tiempo Kh si no salió hasta ese momento. Alternativamente, puede implementar una función, que dada dicha trayectoria, entregue la trayectoria "detenida" $(B_{(ih)\wedge\tau^x}^x)_{i=1}^K$ en el tiempo τ^x. Note que dado que el tiempo es discreto, en el tiempo τ^x el proceso se encontrará en realidad fuera de D, por lo que deberá escoger como punto de salida un punto en ∂D que interpole entre B_τ^x y $B_{\tau-h}^x$.

3. Combine lo antes implementado para simular por bloques de K pasos de paso h, N trayectorias Brownianas partiendo de cada punto x de la grilla espacial, hasta que todas las trayectorias hayan salido del D, encontrando para cada trayectoria $(B_{ih}^x)_{i\in\mathbb{N}}$ el punto $B_{\tau^x}^x \in \partial D$ respectivo. Haga esto para cada uno de los 2 dominios D especificados. Se recomienda agregar para cada trayectoria $(B_{ih}^x)_{i\in\mathbb{N}}$ una variable binaria que indique si el proceso ya salió del dominio, con el fin de optimizar el uso de la función programada en 2).

4. Busque ejemplos de funciones $f : \partial D \to \mathbb{R}$ para los cuales el problema de Dirichlet (A.5) tiene solución analítica conocida, en los dos dominios $D = [-1, 1]^2$ y $D = B((0, 0), 1)$ considerados. Calcule numéricamente $u(x)$ para cada x en la grilla espacial correspondiente, mediante un método de Monte Carlo usando $N > 10000$ trayectorias de movimientos Brownianos partiendo de $(0, 0)$ y lo antes desarrollado. Haga esto para distintos valores de h y ϵ, compare entre si los resultados obtenidos con distintos sets de parámetros, y con la solución analítica conocida. Para una elección apropiada de parámetros N, h y ϵ grafique la solución calculada con la representación probabilista y la solución exacta, en cada uno de los dominios.