Métodos Monte Carlo

José Ignacio Illana*

Departamento de Física Teórica y del Cosmos
Universidad de Granada

Enero de 2013

*Email: jillana@ugr.es

Índice

Tema 1

Introducción

1.1 Qué es un Monte Carlo

El término *Monte Carlo* se aplica a un conjunto de métodos matemáticos que se empezaron a usar en los 1940s para el desarrollo de armas nucleares en Los Alamos, favorecidos por la aparición de los ordenadores digitales modernos. Consisten en resolver un problema mediante la invención de juegos de azar cuyo comportamiento *simula* algún fenómeno real gobernado por una distribución de probabilidad (e.g. un proceso físico) o sirve para realizar un *cálculo* (e.g. evaluar una integral).

Más técnicamente, un Monte Carlo es un *proceso estocástico* numérico, es decir, una secuencia de estados cuya evolución viene determinada por sucesos aleatorios. Recordemos que un *suceso aleatorio* es un conjunto de resultados que se producen con cierta probabilidad. Veamos un par de ejemplos ilustrativos.

Ejemplo 1: Gotas de lluvia para estimar π

Consideremos un círculo de radio unidad circunscrito por un cuadrado. Suponiendo una *lluvia uniforme* sobre el cuadrado, podemos hallar el valor de π a partir de la probabilidad de que las gotas caigan dentro del círculo (figura 1.1):

$$\mathrm{P} = \frac{\text{área del círculo}}{\text{área del cuadrado}} = \frac{\displaystyle\int_{-1}^{1} \mathrm{d}x \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \mathrm{d}y}{\displaystyle\int_{-1}^{1} \mathrm{d}x \int_{-1}^{1} \mathrm{d}y} = \frac{2\displaystyle\int_{-1}^{1} \mathrm{d}x \sqrt{1-x^2}}{2 \cdot 2} = \frac{\pi}{4}. \tag{1.1}$$

Es decir, $\pi = 4\mathrm{P}$. Nótese que: (i) podemos simular fácilmente este experimento generando *aleatoriamente* con un ordenador puntos de coordenadas cartesianas (x, y); (ii) podemos mejorar nuestra estimación de π aumentando el número de puntos generados (ejercicio 1); (iii) tenemos un método para hallar la integral que aparece en la ecuación (1.1). Ciertamente el valor de π puede encontrarse de forma más rápida y precisa mediante otros métodos, pero veremos que el método Monte Carlo es el más eficiente para hallar integrales multidimensionales.

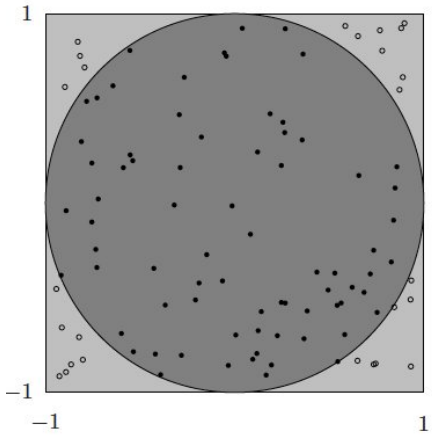


Figura 1.1: Experimento de las gotas de lluvia para estimar π. Extraído de [4].

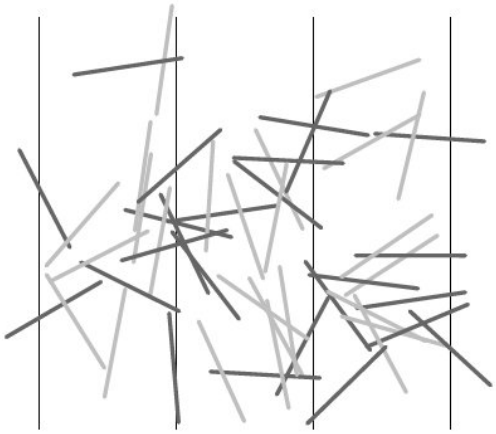


Figura 1.2: Experimento de las agujas de Buffon para estimar π. Extraído de [4].

Ejemplo 2: Las agujas de Buffon para estimar π

En 1773 Georges Louis Leclerc, conde de Buffon, propuso el siguiente problema, que él mismo resolvió en 1777. Consideremos un papel horizontal, en el que se han trazado rectas paralelas separadas una distancia d, sobre el que se dejan caer *aleatoriamente* agujas de longitud $L < d$ (figura 1.2). ¿Cuál es la probabilidad que las agujas crucen una cualquiera de las rectas? La respuesta es (ejercicio 2):

$$\mathrm{P} = \frac{2L}{\pi d}. \tag{1.2}$$

Esto nos permite estimar el valor de π, aunque de nuevo la convergencia es muy lenta.

En los dos ejemplos hemos destacado la palabra *aleatorio*. Sin embargo, un ordenador es determinista, pues solamente es capaz de generar una secuencia programada de números *pseudoaleatorios*. En el tema 2 estudiaremos este asunto, aprenderemos a muestrear distribuciones de probabilidad arbitrarias e introduciremos el concepto de camino aleatorio (cadena de Markov) [1, 2]. También presentaremos las técnicas de integración Monte Carlo más habituales, así como diversas formas de reducir los errores (control de la varianza) [1, 2, 3]. En el tema 3 veremos algunas aplicaciones de los métodos Monte

Carlo. Pero primero, en este mismo tema, repasaremos algunos conceptos de Probabilidad y Estadística que necesitaremos durante el curso.

1.2 Repaso de Probabilidad y Estadística

1.2.1 Sucesos aleatorios y definiciones

Un *experimento aleatorio* es aquél cuyo resultado no puede determinarse por adelantado. El ejemplo más sencillo es el lanzamiento de una moneda a cara ($\odot$) o cruz ($\oplus$).

El *espacio muestral* Ω es el conjunto de todos los posibles resultados del experimento. Si el experimento es 'lanzar la moneda 3 veces' el espacio muestral es:

$$\Omega = \{\odot\odot\odot,\ \odot\odot\oplus,\ \odot\oplus\odot,\ \oplus\odot\odot,\ \odot\oplus\oplus,\ \oplus\odot\oplus,\ \oplus\oplus\odot,\ \oplus\oplus\oplus\}. \tag{1.3}$$

Se llama *suceso* A a un subconjunto de Ω. Así, el suceso 'obtener dos caras' es:

$$A = \{\odot\odot\oplus,\ \odot\oplus\odot,\ \oplus\odot\odot\}. \tag{1.4}$$

Se dice que A sucede si el resultado del experimento es uno de los elementos de A. Los sucesos A y B son *sucesos disjuntos* si $A \cap B = \varnothing$.

La *probabilidad* es una regla que asigna un número $0 \leq \mathrm{P}(A) \leq 1$ a cada suceso A, con

$$\mathrm{P}(\Omega) = 1 \tag{1.5}$$

tal que para cualquier secuencia de sucesos disjuntos A_i se cumple la *regla de suma*:

$$\mathrm{P}(\bigcup_i A_i) = \sum_i \mathrm{P}(A_i). \tag{1.6}$$

Si la moneda no está trucada, $\mathrm{P}(A) = |A|/|\Omega| = 3/8$, donde $|A| = 3$ es el número de resultados en A y $|\Omega| = 8$.

La *probabilidad condicionada* de que ocurra B si ha ocurrido A, que denotaremos $\mathrm{P}(B|A)$, viene dada por

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(AB)}{\mathrm{P}(A)} \tag{1.7}$$

donde $\mathrm{P}(AB) \equiv \mathrm{P}(A \cap B)$ es la *probabilidad conjunta* de que ocurran A y B. En efecto, como $B \subset A$, el suceso B ocurre si $A \cap B$ ocurre y lo hará con probabilidad relativa $\mathrm{P}(A \cap B)/\mathrm{P}(A)$. Así la probabilidad condicionada de que al lanzar la moneda 3 veces obtengamos 'la primera vez cara' (suceso B)

$$B = \{\odot\odot\odot,\ \odot\odot\oplus,\ \odot\oplus\odot,\ \odot\oplus\oplus\}$$

si el 'número total de caras es dos' (suceso A), es:

$$\mathrm{P}(B|A) = \frac{(2/8)}{(3/8)} = \frac{2}{3}. \tag{1.8}$$

Despejando (1.7) deducimos la *regla del producto* de probabilidades:

$$\begin{aligned} \mathrm{P}(A_1 \cdots A_n) &= \mathrm{P}(A_1 \cdots A_{n-1})\mathrm{P}(A_n|A_1 \cdots A_{n-1}) \\ &= \mathrm{P}(A_1 \cdots A_{n-2})\mathrm{P}(A_{n-1}|A_1 \cdots A_{n-2})\mathrm{P}(A_n|A_1 \cdots A_{n-1}) \\ &= \mathrm{P}(A_1)\mathrm{P}(A_2|A_1)\mathrm{P}(A_3|A_1A_2) \cdots \mathrm{P}(A_n|A_1 \cdots A_{n-1}). \end{aligned} \tag{1.9}$$

Se dice que A y B son *sucesos independientes* si el hecho de que ocurra A no cambia la probabilidad de que ocurra B, es decir si $\mathrm{P}(B|A) = \mathrm{P}(B)$. Aplicando (1.7) vemos que

$$A \text{ y } B \text{ independientes} \Leftrightarrow \mathrm{P}(AB) = \mathrm{P}(A)\mathrm{P}(B). \tag{1.10}$$

Por ejemplo, consideremos ahora el experimento 'lanzar la moneda n veces'. Sean A_i los sucesos en los que sale cara en el i-ésimo lanzamiento ($i = 1, \ldots, n$). Los sucesos A_i son independientes. Sea p la probabilidad de que salga cara en cada lanzamiento ($p = 1/2$ si la moneda no está trucada). Entonces la probabilidad de que en los n lanzamientos los primeros k sean caras y los siguientes $n - k$ sean cruces es:

$$\begin{aligned} \mathrm{P}(A_1 \cdots A_k \bar{A}_{k+1} \cdots \bar{A}_n) &= \mathrm{P}(A_1) \cdots \mathrm{P}(A_k)\mathrm{P}(\bar{A}_{k+1}) \cdots \mathrm{P}(\bar{A}_n) \\ &= p^k(1-p)^{n-k}. \end{aligned} \tag{1.11}$$

1.2.2 Variables aleatorias y distribuciones de probabilidad

La descripción completa de Ω y P de un experimento aleatorio no suele resultar conveniente ni necesaria. Generalmente basta con asociar uno o varios *números* X (*variable(s) aleatoria(s)*) a los resultados del experimento. Por ejemplo, en el experimento anterior, estamos más interesados en conocer la probabilidad de que salga k veces cara, $\mathrm{P}(X = k)$, que en conocer la probabilidad de obtener cierta secuencia de resultados como $\mathrm{P}(A_1 \cdots A_k \bar{A}_{k+1} \cdots \bar{A}_n)$.

Se dice que la variable aleatoria X tiene una *distribución discreta* si solamente para un conjunto numerable de valores x_i se tiene

$$\mathrm{P}(X = x_i) > 0, \quad \text{con} \quad \sum_i \mathrm{P}(X = x_i) = 1. \tag{1.12}$$

Llamamos *función distribución de probabilidad* (pdf) a $f(x) = \mathrm{P}(X = x)$. En el ejemplo, la variable aleatoria X toma valores discretos $k \in \{0, 1, \ldots, n\}$. Su pdf viene dada por la fórmula binomial

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n. \tag{1.13}$$

Se dice que la variable aleatoria X tiene una *distribución continua* si existe una función f con integral total unidad, tal que para todo $x_1 \leq x_2$

$$\mathrm{P}(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} \mathrm{d}x \; f(x), \quad x \in \mathbb{R}. \tag{1.14}$$

Esta función f es la *densidad de probabilidad*, que también llamaremos pdf. Queda determinada especificando la *función de distribución acumulada* (cdf) definida por

$$F(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} \mathrm{d}x \; f(x) \quad \Rightarrow \quad f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}. \tag{1.15}$$

Está claro que $F(x)$ es una función *no decreciente* en x, $F(-\infty) = 0$ y $F(\infty) = 1$. En el caso de que $F(x)$ no sea continua en algún punto x_i podemos usar la *delta de Dirac* para describir f en esos puntos (la pdf no es por tanto en este caso una función acotada):

$$f(x) = \sum_i \delta(x - x_i) p_i. \tag{1.16}$$

Muchas veces el experimento aleatorio (también llamado proceso estocástico) viene descrito por más de una variable aleatoria $\boldsymbol{X} = (X_1, \ldots, X_n)$, cuyos posibles valores constituyen el espacio de parámetros del proceso, ya sea continuo o discreto. La *probabilidad conjunta*, en el caso discreto, viene dada por la pdf conjunta:

$$f(x_1, \ldots, x_n) = \mathrm{P}(X_1 = x_1, \ldots, X_n = x_n). \tag{1.17}$$

En el caso continuo, la cdf conjunta es:

$$F(x_1, \ldots, x_n) = \mathrm{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \int_{-\infty}^{x_1} \mathrm{d}x_1 \cdots \int_{-\infty}^{x_n} \mathrm{d}x_n \, f(x_1, \ldots, x_n) \tag{1.18}$$

y la pdf conjunta:

$$f(x_1, \ldots, x_n) = \frac{\partial^n F(x_1, \ldots, x_n)}{\partial x_1 \cdots \partial x_n}. \tag{1.19}$$

Consideremos ahora, por simplificar, dos variables aleatorias X e Y (ambas discretas o continuas). A partir de (1.7) definimos la *pdf condicionada* de obtener Y dado X como:

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}, \tag{1.20}$$

donde se ha introducido la *pdf marginal* de obtener X (probabilidad conjunta de X para cualquier Y):

$$f_X(x) = \sum_i f(x, y_i) \quad \text{o bien} \quad \int_{-\infty}^{\infty} \mathrm{d}y \, f(x,y), \tag{1.21}$$

según sean discretas o continuas, respectivamente. Si $f_{Y|X}(y|x) = f_Y(y)$ entonces

$$X \text{ e } Y \textit{ independientes} \Leftrightarrow f(x,y) = f_X(x) f_Y(y). \tag{1.22}$$

1.2.3 Esperanza, varianza y covarianza de variables aleatorias

La *esperanza* de una variable aleatoria X es el *valor medio* o *esperado* de su distribución:

$$\mu_X = \mathrm{E}[X] = \begin{cases} \sum_i x_i f(x_i) & \text{(caso discreto)}, \\ \int_{-\infty}^{\infty} \mathrm{d}x \; x f(x) & \text{(caso continuo)}. \end{cases} \tag{1.23}$$

Una función cualquiera $g(X)$ es también una variable aleatoria (veremos en §1.2.5 cómo hallar su pdf). Su esperanza es:

$$\mathrm{E}[g(X)] = \begin{cases} \sum_i g(x_i) f(x_i) & \text{(caso discreto)}, \\ \int_{-\infty}^{\infty} \mathrm{d}x \; g(x) f(x) & \text{(caso continuo)}. \end{cases} \tag{1.24}$$

Si tenemos una función de varias variables aleatorias entonces

$$\mathrm{E}[g(X_1,\ldots,X_n)] = \int \mathrm{d}x_1 \cdots \int \mathrm{d}x_n \; g(x_1,\ldots,x_n) f(x_1,\ldots,x_n). \tag{1.25}$$

Así tenemos que, en general,

$$\mathrm{E}[a + b_1 X_1 + \cdots + b_n X_n] = a + b_1\mu_1 + \cdots + b_n\mu_n \tag{1.26}$$

siendo $\mu_i = \mathrm{E}[X_i]$ y a, b_i constantes. Y, solamente *si son independientes*,

$$\mathrm{E}[X_1 X_2 \cdots X_n] = \mu_1\mu_2\cdots\mu_n. \tag{1.27}$$

La *esperanza condicionada* de Y dado X es:

$$\mathrm{E}[Y|X] = \int_{-\infty}^{\infty} \mathrm{d}y \; y f_{Y|X}(y|x) = \frac{\displaystyle\int_{-\infty}^{\infty} \mathrm{d}y \; y f(x,y)}{f_X(x)} \tag{1.28}$$

que es también una variable aleatoria cuya esperanza es:

$$\mathrm{E}[\mathrm{E}[Y|X]] = \int_{-\infty}^{\infty} \mathrm{d}x \; \mathrm{E}[Y|X] f_X(x) = \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{d}y \; y f(x,y) = \mathrm{E}[Y], \tag{1.29}$$

como era de esperar.

La *varianza* de X mide la dispersión de la distribución:

$$\sigma_X^2 = \mathrm{var}(X) = \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2. \tag{1.30}$$

La raíz cuadrada de la varianza se llama *desviación estándar* σ. Además de expresar la dispersión de los resultados respecto al valor medio, la σ se usa para medir el nivel de confianza en las estimaciones estadísticas (véase §1.2.6). La varianza de una función $g(X)$ es:

$$\mathrm{var}(g(X)) = \mathrm{E}[g^2(X)] - (\mathrm{E}[g(X)])^2. \tag{1.31}$$

La *covarianza* de dos variables aleatorias X e Y se define como

$$\mathrm{cov}(X,Y) = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)]. \tag{1.32}$$

Es una medida de la cantidad de dependencia lineal entre las variables (véase (1.27)). Así, tenemos que

$$\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y) + 2\mathrm{cov}(X,Y), \tag{1.33}$$

y si X e Y son independientes entonces $\mathrm{cov}(X,Y) = 0$. Una versión normalizada es el *coeficiente de correlación*

$$\varrho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}. \tag{1.34}$$

Puede probarse (ejercicio 3) que $-1 \leq \varrho(X,Y) \leq 1$.

1.2.4 Distribuciones más habituales

A continuación listamos las pdf discretas y continuas más importantes, así como sus valores medios y varianzas. Cuando una variable aleatoria X se distribuya según f diremos que $X \sim f$.

Discretas	Uniforme	Binomial	Geométrica	Poisson
Notación	$\mathsf{DU}\{1\dots n\}$	$\mathsf{Bin}(n,p)$	$\mathsf{G}(p)$	$\mathsf{Poi}(\lambda)$
$f(k)$	$\frac{1}{n}$	$\binom{n}{k}p^k(1-p)^{n-k}$	$p(1-p)^{k-1}$	$\mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}$
$k \in$	$\{1,2,\dots,n\}$	$\{0,1,\dots,n\}$	$\mathbb{N}^*$	$\mathbb{N}$
parámetros	$n \in \{1,2,\dots\}$	$0 \le p \le 1,\ n \in \mathbb{N}^*$	$0 \le p \le 1$	$\lambda > 0$
$\mathrm{E}[X]$	$\frac{n+1}{2}$	np	$\frac{1}{p}$	λ
$\mathrm{var}(X)$	$\frac{n^2-1}{12}$	$np(1-p)$	$\frac{1-p}{p^2}$	λ

Continuas	Uniforme	Normal	Exponencial	Gamma
Notación	$\mathsf{U}[\alpha,\beta]$	$\mathsf{N}(\mu,\sigma^2)$	$\mathsf{Exp}(\lambda)$	$\mathsf{Gamma}(\alpha,\lambda)$
$f(x)$	$\frac{1}{\beta-\alpha}$	$\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-(x-\mu)^2/(2\sigma^2)}$	$\lambda\mathrm{e}^{-\lambda x}$	$\frac{\lambda^\alpha \mathrm{e}^{-\lambda x}x^{\alpha-1}}{\Gamma(\alpha)}$
$x \in$	$[\alpha,\beta]$	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}^+$
parámetros	$\alpha < \beta$	$\sigma > 0,\ \mu \in \mathbb{R}$	$\lambda > 0$	$\alpha,\lambda > 0$
$\mathrm{E}[X]$	$\frac{\alpha+\beta}{2}$	μ	$\frac{1}{\lambda}$	$\frac{\alpha}{\lambda}$
$\mathrm{var}(X)$	$\frac{(\beta-\alpha)^2}{12}$	σ^2	$\frac{1}{\lambda^2}$	$\frac{\alpha}{\lambda^2}$

La función $\Gamma(\alpha) = \displaystyle\int_0^\infty \mathrm{d}x\ \mathrm{e}^{-x}x^{\alpha-1}$. Si $n \in \mathbb{N}$, $\Gamma(n) = (n-1)!$ con $\Gamma(0) = 1$.

Ya hemos visto que la *distribución binomial* describe la probabilidad de acertar k veces en una secuencia de n experimentos independientes con dos resultados posibles (sí/no, cara/cruz, etc.) en cada uno de los cuales se acierta con probabilidad p. El caso $n = 1$ se llama *distribución de Bernoulli*, $\mathsf{Ber}(p)$.

La *distribución geométrica* describe la probabilidad de acertar a la de k intentos en una secuencia de experimentos independientes con dos resultados posibles en cada uno de los cuales se acierta con probabilidad p. Por ejemplo, la probabilidad de obtener cara después de k lanzamientos de una moneda viene dada por $\mathsf{G}(p)$, donde $p = 1/2$ si la moneda no está trucada.

La *distribución de Poisson* describe la probabilidad de que ocurra un suceso aleatorio k veces en un intervalo de tiempo fijo si sabemos que este suceso se repite en promedio un número de veces λ independientemente del tiempo transcurrido desde el suceso anterior.

Por ejemplo, si en promedio me llegan λ emails cada día, la probabilidad de que un día reciba k viene dada por $\mathsf{Poi}(\lambda)$.

La *distribución normal* o *gaussiana* juega un papel esencial. El teorema del límite central (véase §1.2.7) establece que el promedio de N variables aleatorias *independientes e idénticamente distribuidas* (iid) *cualesquiera* se distribuye según una normal $\mathsf{N}(\mu, \sigma^2/N)$ cuando N es grande.

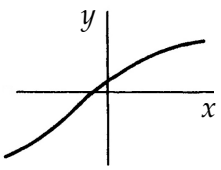
La *distribución exponencial* describe la probabilidad de que ocurra un suceso al cabo de un tiempo x, si la probabilidad de que ocurra en un intervalo de tiempo entre x y $x + \mathrm{d}x$ es proporcional al intervalo, con constante de proporcionalidad λ. Por ejemplo, nos dice cuál es la probabilidad de que ocurra un terremoto al cabo de x años si hay en promedio λ cada año. También modela el decrecimiento/crecimiento de una población cuyo ritmo disminuye/aumenta proporcionalmente al tiempo transcurrido. Por ejemplo, nos dice cuál es la probabilidad de que una partícula inestable viva un tiempo x si su vida media es $1/\lambda$. La distribución exponencial es la versión continua de la distribución geométrica.

La *distribución Gamma* describe la probabilidad de que ocurra un suceso aleatorio α veces al cabo de un tiempo x, si la probabilidad de que ocurra uno en un intervalo de tiempo entre x y $x + \mathrm{d}x$ es proporcional al intervalo, con constante de proporcionalidad λ. Por ejemplo, si sabemos que hay en promedio una inundación cada 6 años, la probabilidad de que haya 4 inundaciones en un tiempo x viene dada por $\mathsf{Gamma}(4, 6)$. La $\mathsf{Gamma}(N, \lambda)$ es asimismo la distribución de la suma $S_N = X_1 + \cdots + X_N$ de N variables aleatorias independientes donde cada $X_i \sim \mathsf{Exp}(\lambda)$ (véase §1.2.5). Otro caso particular es la *distribución* χ^2 (o *verosimilitud*) para n grados de libertad, que corresponde a $\chi^2(n) = \mathsf{Gamma}(\alpha = n/2, \lambda = 1/2)$ y está relacionada con la bondad de un ajuste y los intervalos de confianza en el contraste de hipótesis (véase §1.2.6).

1.2.5 Función de variables aleatorias y transformadas

Supongamos una variable aleatoria $X \sim f_X$. A veces nos conviene conocer la pdf f_Y de una función *monótona* $Y = y(X)$, que es también una variable aleatoria. Para ello hay que relacionar las cdf y distinguir dos casos:

– Si $Y = y(X)$ es una *función no decreciente* de x, es decir,



$$y(X) \leq y(x) \quad \text{sii} \quad X \leq x\,, \tag{1.35}$$

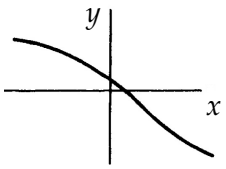
entonces en $y = y(x)$,

$$F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(y(X) \leq y(x)) = \mathrm{P}(X \leq x) = F_X(x)\,. \tag{1.36}$$

Derivando respecto a y,

$$f_Y(y) = \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}y} = f_X(x)\frac{\mathrm{d}x}{\mathrm{d}y}\,. \tag{1.37}$$

– Si $Y = y(X)$ es una *función no creciente* de x, es decir,



$$y(X) \le y(x) \quad \text{sii} \quad X \ge x\,, \tag{1.38}$$

entonces en $y = y(x)$,

$$F_Y(y) = \mathrm{P}(Y \le y) = \mathrm{P}(y(X) \le y(x)) = \mathrm{P}(X \ge x) = 1 - \mathrm{P}(X < x) = 1 - F_X(x)\,. \tag{1.39}$$

Derivando respecto a y,

$$f_Y(y) = \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} = -\frac{\mathrm{d}F_X(x)}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}y} = -f_X(x)\frac{\mathrm{d}x}{\mathrm{d}y}\,. \tag{1.40}$$

Como las pdf son siempre no negativas, (1.37) y (1.40) nos dicen que la pdf f_Y de una función monótona $Y = y(X)$ con $X \sim f_X$ es

$$f_Y(y) = f_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| = f_X(x)\left|\frac{\mathrm{d}y}{\mathrm{d}x}\right|^{-1}\,. \tag{1.41}$$

Por ejemplo:

$$y = y(x) = \mathrm{e}^{-\lambda x} \quad \Rightarrow \quad x = -\frac{1}{\lambda}\ln y; \quad \left|\frac{\mathrm{d}y}{\mathrm{d}x}\right| = \lambda y \quad \Rightarrow \quad f_Y(y) = \frac{f_X(-(\ln y)/\lambda)}{\lambda y}$$

Así, si $X \sim \mathsf{Exp}(\lambda)$ entonces $Y = \mathrm{e}^{-\lambda X}$ es uniforme en el intervalo $[0,1]$, pues

$$f_X(x) = \lambda \mathrm{e}^{-\lambda x} \quad \Rightarrow \quad f_Y(y) = 1.$$

Para conocer la distribución de una función de varias variables aleatorias independientes se procede análogamente. Concentrémonos, por su importancia, en el caso de la suma de dos variables $S = X + Y$. Como son independientes $f(x,y) = f_X(x)f_Y(y)$ y la cdf de la suma será:

$$F_S(s) = \mathrm{P}(X + Y \le s) = \int_{-\infty}^{\infty}\mathrm{d}x \int_{-\infty}^{s-x}\mathrm{d}y\; f_X(x)f_Y(y) = \int_{-\infty}^{\infty}\mathrm{d}x\; f_X(x)F_Y(s-x)\,. \tag{1.42}$$

Así que la pdf de la suma $S = X + Y$ resulta ser la *convolución* de f_X y f_Y:

$$f_S(s) = \frac{\mathrm{d}F_S(s)}{\mathrm{d}s} = \int_{-\infty}^{\infty}\mathrm{d}x\; f_X(x)\frac{F_Y(s-x)}{\mathrm{d}s} = \int_{-\infty}^{\infty}\mathrm{d}x\; f_X(s)f_Y(s-x). \tag{1.43}$$

Por ejemplo, si X_1 y X_2 son dos variables aleatorias iid con pdf $f_X = \mathsf{U}[0,1]$ entonces la distribución de su suma $S_2 = X_1 + X_2$ es

$$f_{S_2}(s) = \int_{-\infty}^{\infty}\mathrm{d}x\; f_X(x)f_X(s-x) = \begin{cases} \displaystyle\int_0^s \mathrm{d}x = s & \text{si } 0 \le s \le 1 \\ \displaystyle\int_{s-1}^1 \mathrm{d}x = 2 - s & \text{si } 1 \le s \le 2\,. \end{cases} \tag{1.44}$$

Nótese que ¡la suma de variables uniformes no es uniforme! Esto no va contra la intuición si se piensa un poco (ejercicio 5). Podemos hallar recursivamente la distribución de la suma $S_N = X_1 + \cdots + X_N$ de N variables aleatorias iid,

$$f_{S_N}(s) = \int_{-\infty}^{\infty} \mathrm{d}x \; f_X(x) f_{S_{N-1}}(s-x). \tag{1.45}$$

Así (ejercicio 6), la pdf de la suma de tres variables uniformes $S_3 = X_1 + X_2 + X_3$ es:

$$f_{S_3}(s) = \begin{cases} \displaystyle\int_0^s \mathrm{d}x \; (s-x) = \frac{s^2}{2} & \text{si } 0 \le s \le 1 \\ \displaystyle\int_0^{s-1} \mathrm{d}x \; (2-s+x) + \int_{s-1}^1 \mathrm{d}x \; (s-x) = -\frac{3}{2} + 3s - s^2 & \text{si } 1 \le s \le 2 \\ \displaystyle\int_{s-2}^1 \mathrm{d}x \; (2-s+x) = \frac{(s-3)^2}{2} & \text{si } 2 \le s \le 3\,. \end{cases} \tag{1.46}$$

Aplicando (1.41) también podemos hallar la distribución del promedio $\overline{X}_N = S_N/N$:

$$f_{\overline{X}_N}(\overline{x}) = N f_{S_N}(N\overline{x}). \tag{1.47}$$

Por otro lado, muchos cálculos y manipulaciones se simplifican gracias al uso de *transformadas*, que son el valor esperado de una función de variables aleatorias, ya sean discretas o continuas. Veamos dos ejemplos importantes:

- La *función generadora de probabilidad* de una variable aleatoria *discreta* definida positiva, N:

$$G(z) = \mathrm{E}[z^N] = \sum_{k=0}^{\infty} z^k \, \mathrm{P}(N=k), \quad |z| \le 1 \tag{1.48}$$

- La *transformada de Laplace* de una variable aleatoria definida positiva, X:

$$L(s) = \mathrm{E}[\mathrm{e}^{-sX}] = \begin{cases} \displaystyle\sum_i \mathrm{e}^{-sx_i} f(x_i) & \text{(caso discreto)} \\ \displaystyle\int_0^{\infty} \mathrm{d}x \; \mathrm{e}^{-sx} f(x) & \text{(caso continuo)} \end{cases}, \quad s \ge 0 \tag{1.49}$$

Todas las transformadas poseen la *propiedad de unicidad*: dos distribuciones son las mismas si y sólo si sus respectivas transformadas son las mismas. Esto nos permite demostrar propiedades interesantes de las distribuciones. Por ejemplo:

- La suma de dos variables aleatorias discretas independientes $M \sim \mathsf{Poi}(\mu)$ y $N \sim \mathsf{Poi}(\nu)$ es también una distribución de Poisson $M + N \sim \mathsf{Poi}(\mu + \nu)$.

  En efecto. Hallemos la función generadora de probabilidad de M,

$$G(z) = \sum_{k=0}^{\infty} z^k \mathrm{e}^{-\mu} \frac{\mu^k}{k!} = \mathrm{e}^{-\mu} \sum_{k=0}^{\infty} \frac{(z\mu)^k}{k!} = \mathrm{e}^{-\mu} \mathrm{e}^{z\mu} = \mathrm{e}^{-\mu(1-z)}. \tag{1.50}$$

Como la función generadora de $M+N$ es

$$\mathrm{E}[z^{M+N}] = \mathrm{E}[z^M]\mathrm{E}[z^N] = \mathrm{e}^{-(\mu+\nu)(1-z)}, \tag{1.51}$$

vemos que la transformada de $M+N$ coinicide con la de una variable de Poisson de parámetros $\mu+\nu$. Aplicando la propiedad de unicidad, tenemos que $M+N \sim \mathsf{Poi}(\mu+\nu)$.

- La suma de N variables aleatorias continuas independientes $X_i \sim \mathsf{Exp}(\lambda)$, $S_N = X_1 + \cdots + X_N$, se distribuye según $S_N \sim \mathsf{Gamma}(N,\lambda)$.

  En efecto. Hallemos la transformada de Laplace de X,

$$\mathrm{E}[\mathrm{e}^{-sX}] = \int_0^\infty \mathrm{d}x\ \mathrm{e}^{-sx}\frac{\mathrm{e}^{-\lambda x}\lambda^\alpha x^{\alpha-1}}{\Gamma(\alpha)} = \left(\frac{\lambda}{\lambda+s}\right)^\alpha \tag{1.52}$$

  donde se ha cambiado la variable x por $(\lambda+s)x$ y sustituido la definición de $\Gamma(\alpha)$. Si hacemos ahora la transformada de Laplace de S_N,

$$\mathrm{E}[\mathrm{e}^{-sS_N}] = \mathrm{E}[\mathrm{e}^{-sX_1}\cdots\mathrm{e}^{-sX_N}] = \mathrm{E}[\mathrm{e}^{-sX_1}]\cdots\mathrm{E}[\mathrm{e}^{-sX_N}] = \left(\frac{\lambda}{\lambda+s}\right)^N \tag{1.53}$$

  vemos que es la misma que la de una distribución Gamma de parámetros N,λ. Por tanto, por la propiedad de unicidad, tenemos que $S_N \sim \mathsf{Gamma}(N,\lambda)$.

1.2.6 Tests, ajustes e intervalos de confianza

Supongamos que queremos ajustar una función a unos datos o bien queremos comprobar una predicción teórica. Para ello tenemos que comparar una serie de datos X_i con una hipótesis dada por los correspondientes valores esperados, $E_i = \mathrm{E}[X_i]$. Si los datos X_i *se distribuyen según una normal* (lo que ocurre e.g. si cada X_i es el promedio de un número grande de medidas, como veremos en §1.2.7) entonces la variable de prueba

$$\chi^2 = \sum_{i=1}^{k}\frac{(X_i-E_i)^2}{E_i} \tag{1.54}$$

se distribuye según la pdf $f(x;n) = \chi^2(n) = \mathsf{Gamma}(n/2,1/2)$ donde n es el número de grados de libertad, igual a número de datos, k, menos el número de ligaduras. Por ejemplo, si comparamos una predicción teórica con un solo dato entonces $n=1$, pero si comparamos una curva teórica con k bines de datos entonces $n=k-1$, pues en este caso fijamos la normalización total.

En general, para cuantificar el nivel de acuerdo entre los datos y nuestra hipótesis se define el *p-value*,

$$p = \int_{\chi^2}^\infty \mathrm{d}t\ f(t;n) = 1 - F(\chi^2;n) \tag{1.55}$$

(uno menos la cdf de la $\chi^2(n)$) que da la probabilidad de encontrar un valor t de la variable de prueba que sea menos compatible con los datos que el valor observado χ^2. Es decir, mide la probabilidad de que el resultado de las medidas se deba a una fluctuación

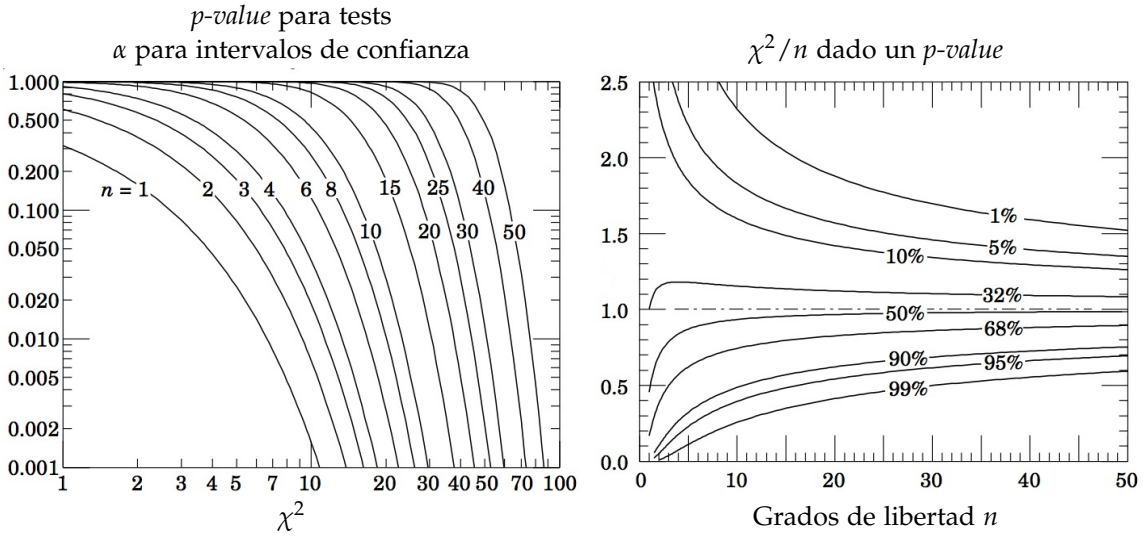


Figura 1.3: Curvas que relacionan χ^2, *p-value* para tests o α para intervalos de confianza, y número de grados de libertad n. Extraído de [5].

estadística (*hipótesis nula*). Véase la figura 1.3. Esto no significa que la probabilidad de nuestra predicción es $1-p$. Como el valor medio de la pdf $\chi^2(n)$ es n, uno espera obtener $\chi^2 \approx n$ en un experimento "razonable".

Por otro lado, podemos usar este test de la χ^2 para estimar el rango de valores de las variables X_i que pueden excluirse con un nivel de confianza α: el resto de los valores constituyen el *intervalo de confianza* con un nivel de confianza $\mathrm{CL} = 1-\alpha$. Por ejemplo, si consideramos una sola variable X que se distribuye según una normal con media μ y varianza σ^2 entonces[a]

$$1-\alpha = \frac{1}{\sqrt{2\pi}\sigma}\int_{\mu-\delta}^{\mu+\delta} \mathrm{d}x\ \mathrm{e}^{-(x-\mu)^2/(2\sigma^2)} = \mathrm{erf}\left(\frac{\delta}{\sqrt{2}\sigma}\right) \tag{1.56}$$

es la probabilidad de que el valor medido x caiga dentro de $\mu \pm \delta$, o bien, puesto que la distribución es simétrica bajo el intercambio de x y μ, es también la probabilidad de que el intervalo $x \pm \delta$ incluya el valor μ. La elección $\delta = \sigma$ da un intervalo llamado *error estándar* o *una desviación estándar* (1σ), que tiene $1-\alpha = 68.27\,\%$. Otros valores representativos pueden encontrarse en la tabla 1.1. En la figura 1.4 se muestra un ejemplo. La relación (1.56) puede reescribirse usando la cdf de la distribución $\chi^2(n)$ como

$$\alpha = 1 - F(\chi^2; n) \tag{1.57}$$

para $\chi^2 = (\delta/\sigma)^2$ y $n = 1$ grado de libertad, pues como $\chi^2(1) = \dfrac{1}{\sqrt{2\pi}} t^{-1/2}\mathrm{e}^{-t/2}$ tenemos que, en efecto

$$\begin{aligned} 1-\alpha = \mathrm{erf}\left(\frac{\delta}{\sqrt{2}\sigma}\right) &= \frac{2}{\sqrt{\pi}}\int_0^{\delta/(\sqrt{2}\sigma)} \mathrm{d}z\ \mathrm{e}^{-z^2} \\ &= \frac{1}{\sqrt{2\pi}}\int_0^{(\delta/\sigma)^2} \mathrm{d}t\ t^{-1/2}\mathrm{e}^{-t/2} = F(\chi^2; n)\big|_{\chi^2=(\delta/\sigma)^2;\ n=1}\ . \end{aligned} \tag{1.58}$$

[a] La función error se define como $\mathrm{erf}(x) = \dfrac{2}{\sqrt{\pi}}\displaystyle\int_0^x \mathrm{d}z\ \mathrm{e}^{-z^2}$.

Tabla 1.1: Área α de las colas fuera de $\pm\delta$ desde el valor medio de una distribución normal. Extraída de [5].

α	δ
0.3173	1σ
4.55×10^{-2}	2σ
2.7×10^{-3}	3σ
6.3×10^{-5}	4σ
5.7×10^{-7}	5σ
2.0×10^{-9}	6σ

α	δ
0.2	1.28σ
0.1	1.64σ
0.05	1.96σ
0.01	2.58σ
0.001	3.29σ
10^{-4}	3.89σ

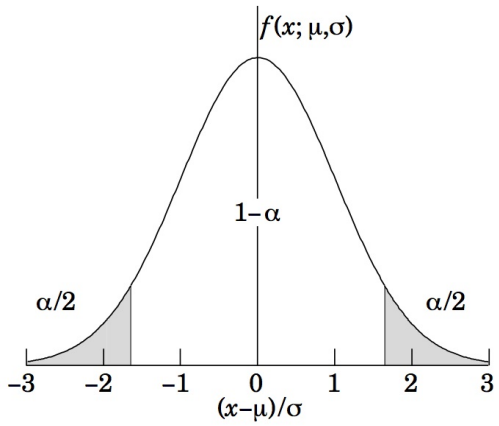


Figura 1.4: Ilustración de un intervalo de confianza del 90 % (zona no sombreada) para la medida de una sola variable gaussiana. Corresponde a $\alpha = 0.1$. Extraída de [5].

Si en lugar de una tenemos varias variables aleatorias gaussianas que ajustar se obtiene un resultado análogo con $n > 1$. Así que de (1.55) y (1.57) vemos que las curvas de la figura (1.3) expresan *p-values* para tests o α para intervalos de confianza. De hecho, se suelen expresar los *p-values* en número de sigmas (tabla 1.1). Por ejemplo, el 4 de julio de 2012 el CERN anunció que dos experimentos del LHC, ATLAS y CMS, habían encontrado una señal compatible con un bosón de Higgs de unos 125 GeV, incompatible con una fluctuación estadística con un nivel de confianza de aproximadamente 5σ, es decir menor que una parte en un millón. Esto *no significa* que se trata de un Higgs al 99.9999 %, pero se habla de "descubrimiento" cuando la probabilidad de que sea una fluctuación es de al menos 5σ, y "evidencia" si son más de 3σ.

1.2.7 Teoremas importantes

Ley de los grandes números

El promedio de N variables aleatorias iid X_i, $\overline{X}_N = S_N/N$ converge a $\mathrm{E}[X]$ cuando N es grande, es decir,

$$\mathrm{P}\left(\lim_{N\to\infty} \overline{X}_N = \mathrm{E}[X]\right) = 1. \tag{1.59}$$

Esta ley justifica la interpretación intuitiva de que la esperanza de una variable aleatoria converge al promedio a largo plazo al hacer un muestreo repetitivo. Así, el valor esperado de una magnitud física converge al promedio de los resultados al repetir muchas medidas de la misma. Por ejemplo, para hallar el valor medio de los resultados al lanzar un dado, lanzamos muchos dados (vale también que muchas personas tiren un dado cada una) y hacemos el promedio de los resultados; o para hallar la vida media de una partícula o un núcleo inestable, especialmente si es muy larga, tomamos una muestra con muchos de ellos y deducimos la vida media a partir de la ley exponencial que sigue el número de desintegraciones que vamos observando (así se ha logrado poner una cota inferior a la vida media de un protón del orden de 10^{29} años, mucho mayor que la edad del universo, aunque los experimentos llevan buscando desintegraciones solamente unos pocos años.) Veamos a qué ritmo se produce esta convergencia.

Desigualdad de Markov

Supongamos una variable aleatoria X que sólo puede tomar valores no negativos y sea f su pdf. Entonces para cualquier $x > 0$,

$$\mathrm{E}[X] = \int_0^x \mathrm{d}t\ tf(t) + \int_x^\infty \mathrm{d}t\ tf(t) \geq \int_x^\infty \mathrm{d}t\ tf(t) \geq \int_x^\infty \mathrm{d}t\ xf(t) = x\mathrm{P}(X \geq x)$$

$$\Rightarrow \quad \mathrm{P}(X \geq x) \leq \frac{\mathrm{E}[X]}{x}. \quad (1.60)$$

Desigualdad de Chebyshev

Si X tiene esperanza μ y varianza σ^2 entonces la desigualdad de Markov aplicada a la variable aleatoria $D^2 = (X-\mu)^2$ (cuya esperanza es también σ^2) nos dice que $\mathrm{P}(D^2 \geq x^2) \leq \sigma^2/x^2$. Por tanto,

$$\mathrm{P}(|X-\mu| \geq x) \leq \frac{\sigma^2}{x^2}. \quad (1.61)$$

Teorema fundamental del Monte Carlo

Consideremos la variable aleatoria G_N, promedio de una función $g(X_i)$ de variables iid,

$$G_N = \frac{1}{N}\sum_{i=1}^N g(X_i), \quad (1.62)$$

cuya esperanza y varianza son respectivamente

$$\mathrm{E}[G_N] = \mathrm{E}[g(X)], \quad \mathrm{var}(G_N) = \frac{\mathrm{var}(g(X))}{N}. \quad (1.63)$$

Al promedio G_N se le llama *estimador* de $E[g(x)]$, pues su esperanza vale

$$\mathrm{E}[G_N] = \mathrm{E}[g(X)] = \int_{-\infty}^\infty \mathrm{d}x\ g(x)f(x) \quad (1.64)$$

donde $X_i \sim f$. Es decir podemos *evaluar la integral* anterior generando un conjunto de N variables aleatorias X_i según $f(x)$ y hallando $g(x)$ para cada una. El estimador (1.62) (la media aritmética de los $g(x)$ generados) nos da el valor de la integral (1.64). Veamos que la varianza del estimador disminuye al crecer N. De hecho, aplicando la desigualdad de Chebyshev (1.61) a la variable aleatoria G_N con $\sigma^2 = \mathrm{var}(G_N)$, $x^2 = \sigma^2/\delta$ y $\delta > 0$ tenemos

$$\mathrm{P}\left(|G_N - E[G_N]| \geq \left[\frac{\mathrm{var}(G_N)}{\delta}\right]^{\frac{1}{2}}\right) \leq \delta, \tag{1.65}$$

o bien, usando (1.63),

$$\mathrm{P}\left(|G_N - E[g(X)]| \geq \left[\frac{\mathrm{var}(g(X))}{N\delta}\right]^{\frac{1}{2}}\right) \leq \delta, \tag{1.66}$$

lo que significa que, generando una muestra suficientemente grande ($N \gg 1/\delta$), la probabilidad de que el estimador se aleje del valor esperado de $g(X)$ es tan pequeña como se desee. Y aún puede decirse más...

Teorema del límite central

Si la variable aleatoria G_N toma valores g_N y definimos

$$t_N = \frac{g_N - \mathrm{E}[g(X)]}{\sqrt{\mathrm{var}(G_N)}} = \frac{g_N - \mathrm{E}[g(X)]}{\sqrt{\mathrm{var}(g(X))}/\sqrt{N}} \tag{1.67}$$

el teorema del límite central establece que

$$\lim_{N\to\infty} \mathrm{P}(a < t_N < b) = \int_a^b \mathrm{d}t\, \frac{\mathrm{e}^{-\frac{1}{2}t^2}}{\sqrt{2\pi}}. \tag{1.68}$$

Es decir, cuando N es muy grande los valores de G_N siguen una *distribución normal* de media $\mu = \mathrm{E}[g(X)]$ y varianza σ^2/N con $\sigma^2 = \mathrm{var}(g(X))$. En efecto, podemos hacer el cambio de variable:

$$t_N = \frac{g_N - \mu}{\sigma/\sqrt{N}}, \quad \mathrm{d}t_N = \frac{\mathrm{d}g_N}{\sigma/\sqrt{N}} \tag{1.69}$$

y notar que (1.68) nos dice que, en el límite de N grande, G_N se distribuye según

$$f_{G_N}(g_N) \to \frac{1}{\sqrt{2\pi}(\sigma/\sqrt{N})} \exp\left\{-\frac{1}{2}\left(\frac{g_N - \mu}{\sigma/\sqrt{N}}\right)^2\right\}, \quad N \gg 1 \tag{1.70}$$

y esto es así *sea cual sea la distribución* f_X.

Ilustremos el teorema del límite central con un par de ejemplos. Consideremos como estimador el promedio $\overline{X}_N$ de variables aleatorias iid cuya pdf es (a) $\mathsf{U}[0,1]$ ($\mu = 1/2$, $\sigma^2 = 1/12$) y (b) $\mathsf{Exp}(1)$ ($\mu = \sigma^2 = 1$). En ambos casos $f_{\overline{X}_N}$ tiende a una distribución normal $\mathsf{N}(\mu, \sigma^2/N)$ conforme vamos aumentando N, como puede verse en la figura 1.5. Las distribuciones se han hallado usando (1.47) y los resultados que hemos encontrado para la suma de variables uniformemente distribuidas (1.44–1.46) y exponenciales ($S_N \sim$

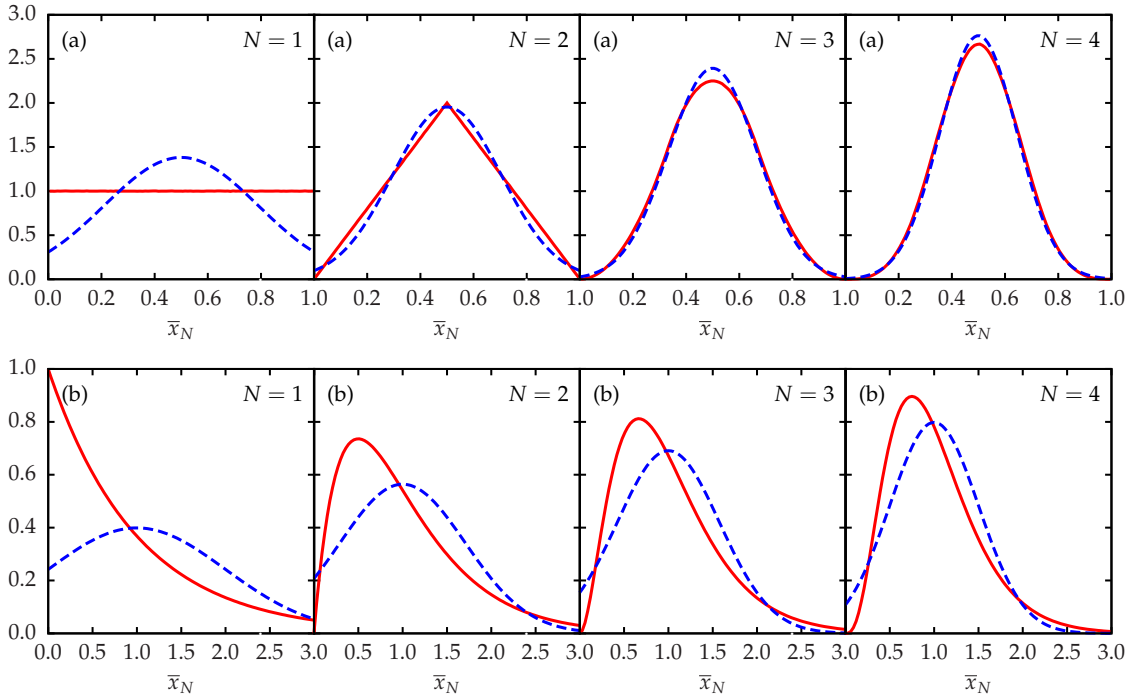


Figura 1.5: Distribución del promedio de N variables aleatorias independientes idénticamente distribuidas (a) uniformemente en $[0, 1]$ y (b) exponencialmente con $\lambda = 1$, para varios valores de N (líneas continuas). Al aumentar N, los promedios tienden a la distribución normal correspondiente (líneas discontinuas).

Gamma(N, λ)). Por tanto, podemos aplicar lo que sabemos sobre intervalos de confianza de la distribución normal (véase §1.2.6) a los resultados de un cálculo Monte Carlo. Así, si N es suficientemente grande, el estimador G_N está dentro de una *desviación estándar* (1σ) de $\mathrm{E}[g(X)]$ (es decir, $\sigma/\sqrt{N}$) un 68.3 % de las veces, dentro de 2σ un 95.4 %, dentro de 3σ un 99.7 %, etc.

Ejercicios

1. *Sobre el experimento de las gotas de lluvia*

   Nótese que podemos dividir la figura 1.1 en cuatro cuadrantes y que la probabilidad de que las gotas caigan dentro del círculo viene dada también por el cociente (nos fijamos sólo en el cuadrante superior derecho):

   $$\mathrm{P} = \frac{\text{área del sector circular}}{\text{área del cuadrado pequeño}} = \int_0^1 \mathrm{d}x \sqrt{1-x^2} = \frac{\pi}{4}.$$

   a) Escribe un programa de ordenador que genere aleatoriamente N pares de coordenadas $x, y \in [0,1]$. Estima el valor de la integral anterior a partir de la fracción de puntos generados (x,y) que cumplan $y < f(x) = \sqrt{1-x^2}$.

   b) Comprueba que el resultado de la integral converge lentamente como $1/\sqrt{N}$.

2. *Sobre el experimento de las agujas de Buffon*

   Sea φ el ángulo que forma una aguja (de longitud L) con la perpendicular a las rectas paralelas (separadas $d > L$).

   a) Muestra que la probabilidad de que una aguja que forma un ángulo φ cruce una recta es

   $$p(\varphi) = \frac{L\cos\varphi}{d}.$$

   b) Como todos los ángulos son equiprobables, demuestra la ecuación 1.2 integrando $p(\varphi)$ para todos los ángulos y dividiendo por el rango.

   c) Escribe un programa de ordenador que simule el lanzamiento de N agujas. Si p_N es la fracción de agujas que cortan las rectas paralelas, comprueba que

   $$\pi_N = \frac{2L}{d}\frac{1}{p_N}$$

   converge a π como $1/\sqrt{N}$. ¿Cuántas veces hay que "tirar la aguja"para obtener las primeras tres, cinco o siete cifras decimales de π?

3. *Coeficiente de correlación*

   Demuestra que el *coeficiente de correlación*

   $$\varrho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

   cumple $-1 \le \varrho(X,Y) \le 1$ y que si X e Y son independientes entonces $\varrho(X,Y) = 0$.

4. *Distribuciones más habituales*

   Dibuja las distribuciones de §1.2.4 y familiarízate con las curvas.

5. *Suma de variables aleatorias discretas uniformes: tirada de dos dados*

   Encuentra cómo se distribuye la puntuación total al tirar dos dados no trucados y compara la distribución con la de la suma de variables uniformes continuas de la ecuación (1.44).

6. *Suma de variables continuas uniformes*

   Comprueba que la distribución de la suma de tres variables iid uniformes en $[0,1]$, $S_3 = X_1 + X_2 + X_3$, viene dada por (1.46).

Tema 2

Muestreo de distribuciones e integración Monte Carlo

2.1 Números pseudoaleatorios

Hemos visto que un Monte Carlo es un proceso estocástico numérico que nos permite resolver un problema. Para ello se requiere sortear variables aleatorias según una distribución de probabilidad. Por ejemplo, para hallar $\int \mathrm{d}x\ f(x)g(x)$ los valores de X deben distribuirse según $f(x)$ y la integral es el valor medio de $g(x)$ sobre el conjunto de X. En este procedimiento es fundamental *muestrear* X siguiendo $f(x)$. Veamos qué quiere decir exactamente muestrear. Consideremos un espacio Ω formado por todos los posibles valores de una o varias variables aleatorias $X = X_1, X_2, \ldots$ y sea $f(x)$ su pdf, que cumple

$$\int_\Omega \mathrm{d}x\ f(x) = 1\ . \tag{2.1}$$

El muestreo es un algoritmo que *produce una secuencia de variables aleatorias* X tales que para cualquier $\Omega' \subset \Omega$,

$$\mathrm{P}\{X \in \Omega'\} = \int_{\Omega'} \mathrm{d}x\ f(x) \leq 1\ . \tag{2.2}$$

En particular, si tenemos una sola variable aleatoria X que se distribuye según una pdf unidimensional definida sobre el intervalo $[0,1]$, lo anterior significa que

$$\mathrm{P}\{X \in (a,b)\} = \int_a^b \mathrm{d}x\ f(x)\ , \quad 0 < a < b < 1\ , \tag{2.3}$$

o, más informalmente, si $\mathrm{d}x = b - a$,

$$\mathrm{P}\{X \in \mathrm{d}x\} = f(x)\mathrm{d}x\ . \tag{2.4}$$

Podemos producir variables aleatorias $X_1, X_2, \ldots$ según una pdf cualquiera si disponemos de variables aleatorias $\xi_1, \xi_2, \ldots$ que se distribuyan *uniformemente* en el intervalo $[0,1]$. En la sección §2.2 veremos cómo hacer esto. Las variables aleatorias uniformemente distribuidas se *imitan* mediante un ordenador, pues siendo generadas por una rutina determinista, un *generador de números pseudoaleatorios* (prng), no son por tanto realmente aleatorias. Sin embargo son *satisfactorias* siempre que cumplan ciertos criterios de calidad. Estudiaremos los números pseudoaleatorios (prn) en esta sección.

2.1.1 Tests de calidad de números de pseudoaleatorios

Un buen prng debe satisfacer las siguientes condiciones:

- *Equidistribución*. Los prn deben repartirse por igual, como correspondería a una verdadera distribución uniforme.
- *Largo periodo*. Todos los prng tienen un periodo a partir del cual la secuencia de números se vuelve a repetir. Para evitar correlaciones no deseadas es importante que el periodo sea largo para no llegar a agotar la secuencia en un cálculo concreto.
- *Repetibilidad*. A veces se necesita repetir un cálculo con exactamente los mismos prn (para hacer una comprobación, por ejemplo). Así que conviene que el generador permita almacenar su estado.
- *Largas subsecuencias disjuntas*. Si la simulación es muy extensa resulta conveniente subdividirla en otras más pequeñas, para lo que es importante que sean estadísticamente independientes y así se puedan recombinar sin introducir correlaciones.
- *Portabilidad*. La rutina debe generar exactamente la misma secuencia de prn no solamente por distintos lenguajes de programación sino también en distintas máquinas.
- *Eficiencia*. La generación de cada prn debe consumir muy poco tiempo.

Para comprobar la bondad de un prng se pueden realizar una serie de tests empíricos. Mostramos a continuación los dos más habituales.

- *Test de la frecuencia*. Sirve para comprobar la equidistribución de N valores generados. Se divide el intevalo $[0,1]$ en k subintervalos. Uno esperaría encontrar N/k valores en cada uno. Sea N_j el número de valores generados en cada subintervalo $j \in \{1,\ldots,k\}$. Entonces la siguiente χ^2 permite hallar la probabilidad de que la distribución generada sea compatible con una verdadera distribución aleatoria uniforme,

$$\chi^2 = \frac{k}{N}\sum_{j=1}^{k}\left(N_j - \frac{N}{k}\right)^2 , \tag{2.5}$$

  que se comportará asintóticamente como una $\chi^2(k-1)$.

- *Test de la serie*. Sirve para comprobar la independencia entre sucesivos prn en una secuencia. Es una generalización del test anterior. Dividimos el intevalo $[0,1]$ en r subintervalos y miramos ternas de $s \geq 2$ puntos $\boldsymbol{x}_n = (x_n, x_{n+1}, \ldots, x_{n+s-1})$ consecutivamente generados. Cada una de las N ternas $\boldsymbol{x}_n$ generadas caerá dentro de uno de los r^s bines en los que se divide este espacio s-dimensional. Si llamamos $N_{j_1,j_2\ldots,j_s}$, con $j_i \in \{1,\ldots,r\}$, al número de valores que caen en el bin $(j_1, j_2, \ldots, j_s)$, la siguiente χ^2 nos permite hallar la probabilidad de que la distribución generada sea uniforme,

$$\chi^2 = \frac{r^s}{N}\sum_{\{j_1,j_2,\ldots,j_s\}}^{r}\left(N_{j_1,j_2,\ldots,j_s} - \frac{N}{r^s}\right)^2 , \tag{2.6}$$

  que se comportará asintóticamente como una $\chi^2(r^s-1)$.

2.1.2 Distintos tipos de generadores pseudoaleatorios

Todos los algoritmos para generar números pseudoaleatorios producen enteros positivos hasta un máximo m. Para obtener números reales entre 0 y 1 basta dividir por m. La mayoría se basan en métodos recurrentes lineales en los que[a]

$$x_{i+1} = a_0 x_i + a_1 x_{i-1} + \cdots + a_r x_{i-r} + b \mod m \,, \tag{2.7}$$

donde inicialmente hay que especificar $r+1$ valores de partida $x_0, x_1, \ldots, x_r$. El periodo y las propiedades estadísticas de las secuencias de prn así generadas dependen de los parámetros $a_0, \ldots, a_r, b$ y m.

- *Generadores congruentes lineales*

  Consisten en generar una secuencia de números mediante la fórmula recursiva

  $$x_{i+1} = (a x_i + c) \mod m \,, \tag{2.8}$$

  donde el valor inicial x_0 se llama *semilla* y a, c y m son parámetros enteros positivos llamados *multiplicador*, *incremento* y *módulo*, respectivamente. Cada x_i puede tomar valores en $\{0, 1, \ldots, m-1\}$ y la secuencia $x_0, x_1, \ldots$ se repetirá al cabo de un número de pasos (*periodo*) que será como máximo m. Para conseguir este máximo los parámetros deben satisfacer ciertos criterios.

- *Generadores congruentes multiplicativos*

  Es el caso especial de generadores congruentes lineales con $c = 0$,

  $$x_{i+1} = (a x_i) \mod m \,. \tag{2.9}$$

  Para conseguir una secuencia de periodo suficientemente largo conviene elegir para m un número primo muy grande.[b] Por ejemplo, para ordenadores de 32-bits, IBM implementaba en los sesenta y los setenta $m = 2^{31} - 1$ y $a = 7^5$. Sin embargo, este tipo de prng no es muy satisfactorio pues se sabe que s-tuplas de puntos generados consecutivamente $(x_n, x_{n+1}, \ldots, x_{n+s-1})$ tienden a agruparse en hiperplanos de este espacio s-dimensional.

- *Generadores de Fibonacci retardados (Lagged Fibonacci Congruential Generators)*

  Son generalizaciones de la secuencia de Fibonacci $x_{i+2} = x_{i+1} + x_i$ de la forma

  $$x_i = (x_{i-p} + x_{i-q}) \mod m \,, \tag{2.10}$$

  que para $m = 2^\beta$ tiene un periodo máximo $(2^q - 1)2^{\beta-1}$ con $q > p$. Un caso particular muy utilizado es el *generador de Mitchell-Moore*,[c]

  $$x_i = (x_{i-24} + x_{i-55}) \mod 2^{32} \,. \tag{2.11}$$

[a] La operación módulo-m consiste en dividir por m y tomar el resto.

[b] Curiosidad: un número primo M_n que puede expresarse como $2^n - 1$ se llama número primo de Mersenne. Es condición necesaria pero no suficiente que n sea primo. El $M_{31} = 2147483647$ tiene 10 dígitos, es el octavo en la serie y fue descubierto por Euler en 1772. El siguiente es M_{61}, que tiene 19 dígitos.

[c] Los primeros 55 números pueden hallarse con cualquier otro método recurrente.

- *Desplazamiento de registros con retroalimentación lineal (Feedback Shift Register)*

  La relación de recurrencia genérica (2.7) puede usarse para generar dígitos, e.g. *bits* de números binarios, $b_i \in \{0,1\}$,

  $$b_i = a_1 b_{i-1} + \cdots + a_r b_{i-r} \quad \text{mód } 2\,. \tag{2.12}$$

  El *Q-bit* (secuencia de Q dígitos binarios),

  $$y_i = b_i b_{i-1} \cdots b_{i-Q+1} \tag{2.13}$$

  es un número entero *en base 2* entre 0 y 2^Q. Hay que especificar los valores de los r primeros bits de la recurrencia (2.12). En cuanto a la fórmula, la mayoría de los generadores usan para los a_j una secuencia de 0s y 1s correspondiente a

  $$b_i = b_{i-p} \oplus b_{i-(p-q)} \tag{2.14}$$

  donde $\oplus$ denota la disyunción exclusiva XOR entre bits (*suma con desplazamiento*),

  $$0 \oplus 0 = 1 \oplus 1 = 0\,, \quad 0 \oplus 1 = 1 \oplus 0 = 1\,. \tag{2.15}$$

  Una elección conveniente es $p = 250$, $q = 103$ que conduce a una serie de periodo $2^{250} - 1$.

2.2 Algoritmos generales para muestrear distribuciones

2.2.1 Variables aleatorias continuas

Necesitamos muestrear una variable aleatoria $Y \sim f_Y(y)$ a partir de una variable aleatoria uniforme $X = \xi \sim f_\xi(\xi)$,

$$f_\xi(\xi) = \begin{cases} 1\,, & \text{si } 0 \le \xi \le 1 \\ 0\,, & \text{en otro caso} \end{cases} \tag{2.16}$$

que imitamos mediante un prng. Veamos distintas formas de hacerlo.

Transformación

Aplicando (1.41) tenemos que si Y es una función monótona de $y(X)$ con $X = \xi$ entonces

$$f_Y(y) = f_\xi(\xi) \left|\frac{\mathrm{d}y}{\mathrm{d}\xi}\right|^{-1} = \left|\frac{\mathrm{d}y}{\mathrm{d}\xi}\right|^{-1}\,. \tag{2.17}$$

Veamos algunos ejemplos:

- $Y = a + (b-a)\xi$ con $a < b$.

  $$f_Y(y) = \frac{1}{(b-a)}\,, \quad a < y < b\,. \tag{2.18}$$

  Vemos que Y se distribuye uniformemente en $[a,b]$.

- $Y = \xi^r$. Dependiendo de $r \neq 0$ generamos una familia de distribuciones,

$$f_Y(y) = \left|\frac{1}{r}\right| y^{1/r-1} \quad \text{con} \begin{cases} 0 < y < 1\,, & \text{si } r > 0 \\ 1 < y < \infty\,, & \text{si } r < 0\,. \end{cases} \tag{2.19}$$

  Si $r > 1$ la pdf diverge en $y = 0$ porque las potencias de y son negativas, pero a pesar de ser singular es integrable (lo es porque el exponente sea mayor que -1). Así, a partir de una variable uniforme podemos muestrear una *ley de potencias* y^α, con $0 < y < 1$ (si $\alpha > -1$) o $y > 1$ (si $\alpha < -1$).

- $Y = -\ln \xi$.

$$f_Y(y) = \mathrm{e}^{-y}\,, \quad 0 < y < \infty\,. \tag{2.20}$$

  El logaritmo natural de la variable uniforme se distribuye exponencialmente.

Inversión

Para distribuciones de una variable podemos usar la siguiente técnica de inversión. Hemos visto en (1.36) que si $Y = y(X)$ es una función *no decreciente* de X entonces su cdf $F_Y(y)$ cumple

$$F_Y(y) = F_X(x)\,. \tag{2.21}$$

Como la cdf de $X = \xi$ es

$$F_\xi(\xi) = \begin{cases} 0\,, & \text{si } \xi < 0 \\ \xi\,, & \text{si } 0 \leq \xi \leq 1 \\ 1\,, & \text{si } \xi \geq 1\,, \end{cases} \tag{2.22}$$

la función $Y(\xi)$ que estamos buscando se halla despejándola de la ecuación (invirtiendo)

$$F_Y(y) = \xi\,. \tag{2.23}$$

Por ejemplo:

- $f_Y(y) = 2y$, con $0 < y < 1$. Si queremos muestrear una variable Y que se distribuye según esta pdf necesitamos despejar

$$F_Y(y) = \int_0^y \mathrm{d}u\; 2u = y^2 = \xi \quad \Rightarrow \quad Y = \sqrt{\xi}\,. \tag{2.24}$$

Veremos más ejemplos enseguida.

Composición

A veces la técnica de transformación o de inversión conduce a complicadas ecuaciones que hay que resolver numéricamente. Entonces puede ser útil generar dos o más variables aleatorias y combinarlas de forma apropiada para muestrear la que necesitamos. Un ejemplo obvio es la suma de variables aleatorias. Veamos otro ejemplo:

- $f_Y(y) = 6y(1-y)$, con $0 < y < 1$. Si queremos muestrar una variable Y que se distribuye según esta pdf, basta escribir Y como el valor que está en medio de tres valores aleatorios ξ_1, ξ_2, ξ_3 distribuidos uniformemente en $[0,1]$,

$$Y = \text{mid}(\xi_1, \xi_2, \xi_3) \; . \tag{2.25}$$

  En efecto. Supongamos primero que $\xi_1 < \xi_2 < \xi_3$, es decir $Y = \xi_2$. La probabilidad de que Y esté en $\mathrm{d}y = \mathrm{d}\xi_2$ con $\xi_1 < \xi_2 < \xi_3$ es

$$\begin{aligned} f_Y(y)\mathrm{d}y = \mathrm{P}(\xi_1 \le \xi_2)\mathrm{P}(\xi_2 \le \xi_3)\mathrm{d}\xi_2 &= \xi_2(1-\xi_2)\mathrm{d}\xi_2 \\ &= y(1-y)\mathrm{d}y \; , \quad \text{si } \xi_1 < \xi_2 < \xi_3 \; , \end{aligned} \tag{2.26}$$

  pues $\mathrm{P}(\xi_2 \le \xi_3) = 1 - \mathrm{P}(\xi_3 \le \xi_2) = 1 - \xi_2$. Como hay 6 posibilidades de ordenación de ξ_1, ξ_2, ξ_3 igualmente probables tenemos finalmente que

$$f_Y(y) = 6y(1-y) \; . \tag{2.27}$$

Veremos otros ejemplos enseguida.

Generación de algunas distribuciones habituales

- *Exponencial*. La pdf de la exponencial es

$$f_Y(y) = \lambda \mathrm{e}^{-\lambda y} \; , \quad 0 < y < \infty \; . \tag{2.28}$$

  Por tanto, aplicando (2.23),

$$F_y(y) = \int_0^y \mathrm{d}u \; \lambda \mathrm{e}^{-\lambda u} = 1 - \mathrm{e}^{-\lambda y} = \xi \quad \Rightarrow \quad Y = -\frac{1}{\lambda}\ln(1-\xi) \; . \tag{2.29}$$

  Esta expresión es computacionalmente equivalente a

$$Y = -\frac{1}{\lambda}\ln\xi \tag{2.30}$$

  que hemos usado en (2.20), pues si ξ es uniforme en $[0,1]$ entonces $1-\xi$ también lo es. Esta $Y(\xi)$, a diferencia de la primera, es una función decreciente en ξ, como hemos presupuesto en (2.21).

- *Normal*. Como la cdf de una normal es

$$F_y(y) = \frac{1}{2}\left(1 + \mathrm{erf}(y/\sqrt{2})\right) = \xi \; , \tag{2.31}$$

  la aplicación directa de la técnica de inversión para hallar $Y(\xi)$ no es práctica.

  Podemos sin embargo usar el *método de Box-Muller* basado en la técnica de composición consistente en generar dos variables uniformes ξ_1 y ξ_2 en $[0,1]$. Si tenemos dos variables aleatorias independientes Y_1 e Y_2 que se distribuyen según $\mathsf{N}(0,1)$,

$$f(y_1, y_2) = f(y_1)f(y_2) = \frac{1}{2\pi}\exp\left[-\frac{y_1^2 + y_2^2}{2}\right] \tag{2.32}$$

podemos cambiar Y_1, Y_2 por R, Φ,

$$Y_1 = R\cos\Phi\ , \quad Y_2 = R\sin\Phi \tag{2.33}$$

y entonces

$$f(y_1)f(y_2)\mathrm{d}y_1\mathrm{d}y_2 = \left(r\exp\left[-\frac{r^2}{2}\right]\mathrm{d}r\right)\left(\frac{1}{2\pi}\mathrm{d}\phi\right)\ . \tag{2.34}$$

Las variables R y Φ son independientes. La variable R se distribuye según

$$f_R(r) = r\exp\left[-\frac{r^2}{2}\right]\ , \quad 0 < r < \infty \tag{2.35}$$

y la podemos generar aplicando (2.23),

$$F_R(r) = \int_0^r \mathrm{d}u\ u\exp\left[-\frac{u^2}{2}\right] = 1-\exp\left[-\frac{r^2}{2}\right] = \xi_1 \quad \Rightarrow \quad R = [-2\ln(1-\xi_1)]^{\frac{1}{2}} \tag{2.36}$$

que es computacionalmente equivalente a

$$R = [-2\ln\xi_1]^{\frac{1}{2}}\ . \tag{2.37}$$

La variable Φ se distribuye uniformemente en $[0, 2\pi)$ así que la generamos con

$$\Phi = 2\pi\xi_2\ . \tag{2.38}$$

Por tanto, podemos muestrear una variable $Y \sim \mathsf{N}(0,1)$ usando dos variables uniformes ξ_1, $\xi_2 \sim \mathsf{U}[0,1]$ a partir de

$$Y = [-2\ln\xi_1]^{\frac{1}{2}}\cos(2\pi\xi_2) \tag{2.39}$$

o bien

$$Y = [-2\ln\xi_1]^{\frac{1}{2}}\sin(2\pi\xi_2)\ . \tag{2.40}$$

Si queremos muestrear $Z \sim \mathsf{N}(\mu,\sigma^2)$ basta con tomar Y de (2.39) o (2.40) y usar

$$Z = \sigma Y + \mu\ , \tag{2.41}$$

como se demuestra trivialmente aplicando (1.41).

Alternativamene, puede generarse de forma aproximada una variable aleatoria gaussiana $Y \sim \mathsf{N}(0,1)$ invocando el teorema del límite central: muestreando la suma de un número N grande de variables aleatorias uniformes ξ_i,

$$Y = \sqrt{\frac{12}{N}}\left(\sum_{i=1}^{N}\xi_i - \frac{N}{2}\right)\ . \tag{2.42}$$

En efecto, vimos en §1.2.7 que el promedio

$$\overline{X}_N = \frac{1}{N}\sum_{i=1}^{N}\xi_i \sim \mathsf{N}\left(\mu = \frac{1}{2}, \sigma^2 = \frac{1}{12N}\right)\ , \quad N \gg 1\ . \tag{2.43}$$

Por tanto, usando (2.41),

$$\overline{X}_N = \sigma Y + \mu \sim \mathsf{N}(\mu, \sigma^2) \quad \Rightarrow \quad Y = \frac{\overline{X}_N - \mu}{\sigma} \sim \mathsf{N}(0,1)\ , \tag{2.44}$$

que es la variable Y de (2.42). En la práctica $N = 12$ suele ser suficiente y además cancela el factor $\sqrt{12/N}$ (véase en la figura 1.5a cómo el promedio de N variables uniformes converge muy rápidamente a una gaussiana al aumentar N).

- *Gamma*. Sabemos que $Y \sim \mathsf{Gamma}(\alpha, \lambda)$ con $\alpha = N$ entero (*distribución de Erlang*) es equivalente a la suma de N variables aleatorias exponenciales iid $Y_i \sim \mathsf{Exp}(\lambda)$ (véase §1.2.5). Por tanto, podemos generar N variables aleatorias ξ_i uniformes en $[0,1]$ y aplicar la técnica de composición usando (2.30),

$$Y = -\frac{1}{\lambda}\sum_{i=1}^{N} \ln \xi_i = -\frac{1}{\lambda} \ln \prod_{i=1}^{N} \xi_i\ . \tag{2.45}$$

  Si α no es entero existe un método basado en generar dos variables $Y_1 \sim \mathsf{N}(0,1)$ e $Y_2 \sim \mathsf{U}[0,1]$ (véase [2] p. 60) que involucra el método de aceptación-rechazo, descrito a continuación.

Método de aceptación-rechazo

Se trata de un método de composición en el que *se selecciona y se somete a prueba una variable aleatoria*. Si pasa la prueba se acepta (es decir, se usa). De lo contrario, se rechaza y se repite el ciclo hasta que sea aceptada. Tiene la ventaja de que *no se necesita conocer la normalización* de la distribución y el incoveniente de que puede ser ineficiente, si se rechazan muchos intentos antes de que uno sea aceptado.

Supongamos que la pdf f que queremos muestrear es nula fuera de un intervalo $[a,b]$ (véase figura 2.1, arriba) y sea

$$c = \sup\{f(x)\ ;\ x \in [a,b]\}\ . \tag{2.46}$$

Entonces, para generar una variable $Z \sim f$ basta seguir los siguientes pasos:

1. Generar $X \sim \mathsf{U}[a,b]$.
2. Generar $Y \sim \mathsf{U}[0,c]$.
3. Si $Y \leq f(X)$ tomar $Z = X$. De lo contrario volver al paso 1.

Nótese que los puntos (X,Y) *generados* se distribuyen uniformemente sobre el rectángulo $[a,b] \times [0,c]$ y por tanto los puntos (X,Y) *aceptados* se distribuirán uniformemente bajo la curva f. Como la probabilidad de que un punto caiga bajo la curva es igual al cociente entre número de puntos aceptados y generados, está claro que la distribución de valores aceptados de X tiene a f como pdf. (Recuérdese el experimento de las gotas de lluvia de §1.1.)

Podemos generalizar el procedimiento. Sea g una función tal que $\phi(x) = Cg(x)$ *mayoriza* $f(x)$ para alguna constante C, es decir, $\phi(x) \geq f(x)$, $\forall x$ (véase figura 2.1, abajo). Nótese que necesariamente $C \geq 1$. Decimos que $g(x)$ es la pdf *propuesta* y *supondremos que es fácil generar variables aleatorias según* g. Entonces, podemos mejorar la eficiencia del procedimiento anterior del siguiente modo:

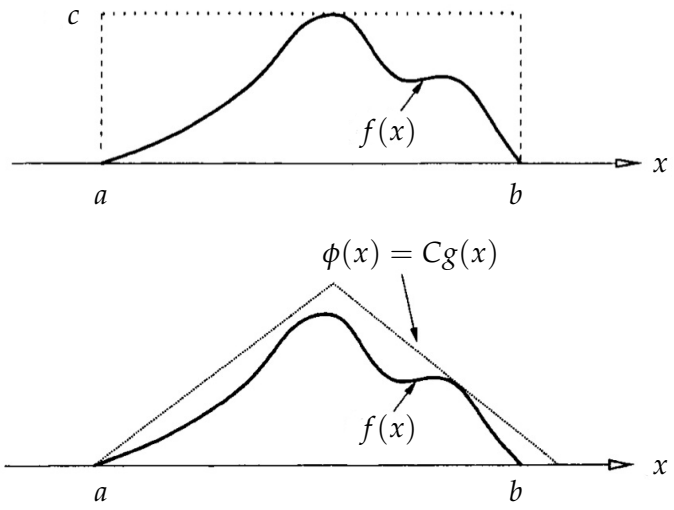


Figura 2.1: Método de aceptación-rechazo. Extraído de [2].

1. Generar X según $g(x)$.
2. Generar $Y \sim \mathrm{U}[0, Cg(X)]$, es decir, $Y = \xi Cg(X)$ con $\xi \sim \mathrm{U}[0,1]$.
3. Si $Y \leq f(X)$ tomar $Z = X$. De lo contrario, volver al paso 1.

O equivalentemente:

1. Generar X según $g(x)$.
2. Generar $Y \sim \mathrm{U}[0,1]$, independiente de X.
3. Si $Y \leq \dfrac{f(X)}{Cg(X)}$ tomar $Z = X$. De lo contrario, volver al paso 1.

Si llamamos $\mathcal{A}$ y $\mathcal{B}$ a las áreas bajo las curvas $Cg(x)$ y $f(x)$, respectivamente, entonces la eficiencia del algoritmo es

$$\mathrm{P}((X,Y) \text{ sea aceptado}) = \frac{\mathcal{B}}{\mathcal{A}} = \frac{1}{C}\,, \tag{2.47}$$

tomando f y g normalizadas a uno. Para conseguir una eficiencia grande (C cerca de 1) hay que conseguir que $g(x)$ sea lo más parecida posible a $f(x)$. (Idealmente $g(x) = f(x)$, lo que exigiría saber muestrar esa pdf.)

Veamos un ejemplo que ya nos resulta familiar. Supongamos que queremos generar una variable aleatoria Z que se distribuye según la pdf dada por el semicírculo de radio unidad,

$$f(x) = \frac{2}{\pi}\sqrt{1-x^2}\,, \quad -1 \leq x \leq 1\,. \tag{2.48}$$

Tomemos como pdf propuesta una distribución uniforme en $[-1,1]$,

$$g(x) = \frac{1}{2}\,, \quad -1 \leq x \leq 1\,, \tag{2.49}$$

y tomemos $C = 4/\pi$, que es el valor más pequeño de C que cumple $Cg(x) \geq f(x)$. En este caso el algoritmo tiene una eficiencia de $1/C = \pi/4 = 0.785$. Es decir, el 78.5 % de los puntos generados uniformemente bajo $Cg(x)$ están también bajo $f(x)$ y son aceptados. Con esta propuesta el algoritmo queda:

1. Generar $X = 1 - 2\xi_1$.
2. Generar $Y = \xi_2$.
3. Si $Y \leq \dfrac{f(X)}{Cg(X)} = \sqrt{1 - X^2}$ tomar $Z = X$. De lo contrario, volver al paso 1.

Nótese que el prefactor $2/\pi$ de $f(x)$ en (2.48), que sirve de normalización, es irrelevante. Si hubiésemos muestreado $\kappa f(x)$, con κ una constante arbitraria, la función $\phi(x)$ que la mayoriza habría sido $\kappa C g(x)$ y el paso 3 quedaría inalterado. La normalización solamente es importante para conocer la eficiencia del algoritmo.

2.2.2 Variables aleatorias discretas

A partir de una o varias variables continuas $X_i = \xi_i \sim \mathsf{U}[0,1]$ generadas e.g. con un prng, pretendemos muestrear una variable aleatoria discreta Y, que toma un conjunto numerable de valores $Y = k$ con probabilidades $f(k)$.

- *Uniforme*. La pdf discreta uniforme $\mathsf{DU}\{1 \ldots n\}$ es

$$f(k) = \frac{1}{n}\,, \quad k = 1, 2, \ldots, n\,. \tag{2.50}$$

  Para muestrear $Y \sim \mathsf{DU}\{1 \ldots n\}$ basta tomar una variable $\xi \sim \mathsf{U}[0,1]$,

$$Y = \lfloor n\xi \rfloor + 1 \tag{2.51}$$

  donde $\lfloor u \rfloor$ es el resultado de truncar al mayor número entero menor que u, es decir tomar la *parte entera por debajo*.

- *Bernoulli*. La pdf $\mathsf{Ber}(p)$ es

$$f(k) = p^k (1-p)^{1-k}\,, \quad k = 0, 1\,. \tag{2.52}$$

  Expresa la probabilidad de obtener uno de los dos posibles resultados ($k = 0, 1$) de *un* experimento aleatorio, si la probabilidad de obtener $k = 1$ (*acierto*) es p. La probabilidad de obtener el otro resultado, $k = 0$ (*error*), es obviamente $1 - p$.

  Para muestrear $Y \sim \mathsf{Ber}(p)$ basta tomar una variable $\xi \sim \mathsf{U}[0,1]$,

$$Y = \begin{cases} k = 0\,, & \text{si } \xi \leq p \\ k = 1\,, & \text{si } \xi > p \end{cases} \qquad \text{o bien} \quad Y = \lfloor \xi + 1 - p \rfloor\,. \tag{2.53}$$

- *Binomial*. La pdf $\mathsf{Bin}(n, p)$ es

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n\,. \tag{2.54}$$

Expresa la probabilidad de acertar k veces en n experimentos aleatorios independientes con dos resultados posibles (*acierto/error*) en cada uno de los cuales se acierta con probabilidad p. Por tanto, podemos escribir $Y \sim \mathsf{Bin}(n,p)$ como suma de n variables $Y_i \sim \mathsf{Ber}(p)$.

Así que, para muestrear $Y \sim \mathsf{Bin}(n,p)$ basta tomar n variables $\xi_i \sim \mathsf{U}[0,1]$,

$$Y = \sum_{i=1}^{n} Y_i \,, \quad Y_i = \lfloor \xi_i + 1 - p \rfloor \,, \quad i = 1,2,\ldots,n \,. \tag{2.55}$$

- *Geométrica*. La pdf $\mathsf{G}(p)$ es

$$f(k) = p(1-p)^{k-1} \,, \quad k = 1,2,\ldots \tag{2.56}$$

Expresa a probabilidad de acertar a la de k intentos en una secuencia de experimentos independientes con dos resultados posibles en cada uno de los cuales se acierta con probabilidad p. Para muestrear una variable discreta $Y \sim \mathsf{G}(p)$ conviene notar la estrecha relación entre la distribución exponencial y la geométrica, que ya habíamos anunciado en §1.2.4, escribiendo

$$(1-p)^x = \mathrm{e}^{-\lambda x} \,. \tag{2.57}$$

Así que si muestreamos una variable $Z \sim \mathsf{Exp}(\lambda)$ con $\lambda = -\ln(1-p)$ tenemos que $Y = \lfloor Z \rfloor + 1$ se distribuye geométricamente. Sabemos que para generar Z a partir de ξ necesitamos $Z = -\frac{1}{\lambda}\ln\xi$. Por tanto, para muestrear $Y \sim \mathsf{G}(p)$ tenemos que tomar

$$Y = \left\lfloor \frac{\ln \xi}{\ln(1-p)} \right\rfloor + 1 \,. \tag{2.58}$$

- *Poisson*. La pdf $\mathsf{Poi}(\lambda)$ es

$$f(k) = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!} \,, \quad k = 0,1,\ldots \tag{2.59}$$

Expresa la probabilidad de que ocurra un suceso aleatorio k veces en un intervalo de tiempo fijo, digamos el intervalo $[0,t]$, si sabemos que este suceso se repite en promedio un número de veces λ independientemente del tiempo transcurrido desde el suceso anterior. Por ejemplo, nos dice cómo se distribuye el número de emails que me llegan al día, N_t, si en promedio recibo λ en $t = 1$ día.

Conviene notar lo siguiente. Si $T_1 < T_2 < \ldots$ son los tiempos de llegada, entonces

$$N_t = \mathrm{máx}\{k \,:\, T_k \le t\} \,, \tag{2.60}$$

y la probabilidad de recibir n mensajes en 1 día está relacionada con la probabilidad

$$\begin{aligned} \mathrm{P}(T_n \le t = 1) = \mathrm{P}(n \le N_t) = \mathrm{P}(N_t \ge n) = 1 - \mathrm{P}(N_t < n) = 1 - \sum_{k=0}^{n-1} f(k) \\ = 1 - \sum_{k=0}^{n-1} \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!} \end{aligned} \tag{2.61}$$

que es precisamente la cdf de la $\mathsf{Gamma}(n,\lambda)$ y, por tanto,

$$T_n \sim \mathsf{Gamma}(n,\lambda)\ . \tag{2.62}$$

En efecto, la cdf de la $\mathsf{Gamma}(n,\lambda)$ es (integrando sucesivamente por partes)

$$\mathrm{P}(T_n \leq t=1) = \int_0^1 \mathrm{d}x\ \frac{\lambda^n \mathrm{e}^{-\lambda x} x^{n-1}}{\Gamma(n)} = 1 - \sum_{k=0}^{n-1} \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}\ . \tag{2.63}$$

Como la $\mathsf{Gamma}(n,\lambda)$ es la distribución de la suma de n variables iid según $\mathsf{Exp}(\lambda)$, concluimos que una variable de Poisson $Y=k$ puede interpretarse como el máximo número de variables exponenciales $Y_i \sim \mathsf{Exp}(\lambda)$ cuya suma no supere 1,

$$Y = \text{máx}\left\{k\ :\ \sum_{i=1}^{k} Y_i \leq 1\right\}\ . \tag{2.64}$$

Por tanto, para muestrear $Y \sim \mathsf{Poi}(\lambda)$ generamos un número suficiente de ξ_i y tomamos

$$Y = \text{máx}\left\{k\ :\ -\frac{1}{\lambda}\sum_{i=1}^{k} \ln \xi_i \leq 1\right\} = \text{máx}\left\{k\ :\ \prod_{i=1}^{k} \xi_i \geq \mathrm{e}^{-\lambda}\right\}\ . \tag{2.65}$$

Vamos probando $k=1,2,\ldots$ hasta hallar el máximo que verifica la desigualdad. Si para $i=1$ no se verifica, tomamos $k=0$. Si $e^{-\lambda}$ es muy pequeño se necesitarán muchas ξ_i para conseguir $\prod_{i=1}^{k}\xi_i \geq \mathrm{e}^{-\lambda}$ con lo que el algoritmo se hace ineficiente. Existen métodos alternativos para esos casos.

2.3 Camino aleatorio y cadena de Markov

Un proceso estocástico $\{X_t, t\}$, donde X_t es un variable aleatoria y t el tiempo, es un *proceso de Markov* si para cualquier intevalo $s>0$ y tiempo t,

$$(X_{t+s}|X_u,\ u\leq t) \sim (X_{t+s}|X_t)\ . \tag{2.66}$$

Es decir, la probabilidad de que ocurra un suceso en un instante futuro $t+s$, descrito por la variable X_{t+s}, está condicionada por el valor de esa variable en el instante actual t, pero es *independiente de su valor en cualquier instante anterior* $u<t$ (*propiedad de Markov*). La dependencia de la probabilidad con el suceso anterior distingue a los procesos de Markov de las series de sucesos independientes, como tirar un dado o lanzar una moneda al aire.

Una *cadena de Markov* es un proceso de Markov *discreto* en el que el "tiempo" toma un conjunto numerable de valores $t=n\in\mathbb{N}$. En este caso, la propiedad de Markov se expresa como

$$\mathrm{P}(X_{n+1}=x_{n+1}|X_0=x_0,\ldots,X_n=x_n) = \mathrm{P}(X_{n+1}=x_{n+1}|X_n=x_n)\ . \tag{2.67}$$

El *camino aleatorio* (*random walk*) es una cadena de Markov en la que las transiciones de probabilidad son temporalmente homogéneas, es decir, no dependen del instante en que ocurren (*transiciones estacionarias*),

$$p_{ij} = \mathrm{P}(X_{n+1}=j|X_n=i)\ ,\quad \text{independiente de } n\ . \tag{2.68}$$

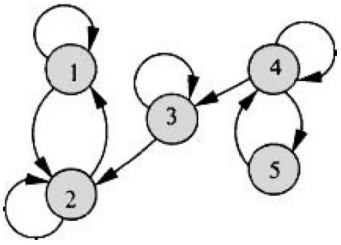


Figura 2.2: Diagrama de transición que representa una cadena de Markov. Extraído de [2].

Los caminos aleatorios se usan frecuentemente como *aproximaciones discretas a procesos físicos continuos*, como el movimiento browniano o procesos de difusión. Otros ejemplos son el "camino de un borracho", la evolución del precio de una acción en bolsa o la situación financiera de un jugador que apuesta hasta arruinarse.

La *matriz de probabilidades de transición en un paso*, $\boldsymbol{P} = (p_{ij})$, especifica completamente una cadena de Markov. Otra forma conveniente de describir una cadena de Markov son los *diagramas de transición*, como el de la figura 2.2, en los que los estados se representan mediante nodos y la transición de probabilidad entre del estado i al j se representa mediante una flecha entre ambos con peso p_{ij}. Como siempre que el sistema abandona el estado i debe ir a parar a otro estado j, entre N posibles, se cumple que

$$\sum_{j=1}^{N} p_{ij} = 1 \,, \quad i = 1, \ldots, N \,. \tag{2.69}$$

Nótese que, en general, p_{ij} no tiene por qué coincidir con p_{ji}. Si el estado j puede alcanzarse desde el estado i en un número finito de pasos, se dice que el estado j es *accesible* desde el estado i. Si además el estado i es accesible desde el estado j, se dice ambos estados se *comunican*. Si dos estados no se comunican, entonces $p_{ij} = 0$ o $p_{ji} = 0$. Los estados que se comunican entre sí forman una clase de equivalencia. Si todos los estados se comunican entre sí, es decir, la cadena está formada por una sola clase de equivalencia tenemos una *cadena de Markov irreducible*. Se dice que i es un *estado recurrente* si siempre se puede regresar a él en un número finito de pasos. De lo contrario, se llama *estado transitorio*. Los estados recurrentes forman una sola clase de equivalencia. Un *estado absorbente* es aquél del que no se puede salir y por tanto es el final de la cadena. Se llama periodo $d(i)$ de un estado i al máximo común divisor del número de pasos que hay que dar por diferentes caminos para regresar hasta él. Si $d(i) = 1$ entonces i es un *estado aperiódico*. Si todos sus estados son aperiódicos tenemos una *cadena de Markov aperiódica*.

El diagrama de transición de la figura 2.2 representa una cadena de Markov no irreducible (no todos los estados están comunicados) con tres clases de equivalencia: $\{1,2\}$, $\{3\}$ y $\{4,5\}$. Los estados 1 y 2 son recurrentes. Los estados 3, 4 y 5 son transitorios. Ninguno de los estados es absorbente, pero el estado 2 lo sería si no estuviera el estado 1. Todos los estados de esta cadena son aperiódicos, pero el estado 5 tendría periodo $d(5) = 2$ si $p_{44} = 0$.

2.4 Algoritmo de Metropolis

El *algoritmo de Metropolis* [M(RT)^2] es un sofisticado método de muestreo.[d] Está basado en las cadenas de Markov y está relacionado con las técnicas de rechazo, pues hay que proponer un valor de prueba y no se necesita conocer la normalización de la función de distribución a muestrear. El método está motivado por su analogía con el comportamiento de los sistemas cerca del equilibrio en mecánica estadística.

La evolución del sistema (proceso) viene descrita por la probabilidad de que éste vaya de un estado X a un estado Y. La condición de que el sistema se aproxime al equilibrio significa que en promedio el sistema tiene la misma probabilidad de estar en X y pasar a Y que de estar en Y y pasar a X, lo que se expresa por la ecuación de *balance detallado*,

$$f(X)\mathrm{P}(Y|X) = f(Y)\mathrm{P}(X|Y) \ , \tag{2.70}$$

pues $f(X)\mathrm{P}(Y|X)$ es la probabilidad $f(X)$ de encontrar el sistema en la vecindad de X por la probabilidad condicionada $\mathrm{P}(Y|X)$ de que el sistema evolucione de X a Y. En un proceso físico, $P(Y|X)$ es conocido y se trata de hallar $f(X)$. El algoritmo M(RT)^2 pretende lo contrario: dada una $f(X)$ se trata de hallar la probabilidad de transición que lleva el sistema al equilibrio.

La idea es la siguiente. Se *proponen* transiciones de X a Y siguiendo una *distribución de prueba cualquiera* $T(Y|X)$. Se compara $f(Y)$ con $f(X)$ y se acepta Y con probabilidad $A(Y|X)$. Por tanto,

$$P(Y|X) = A(Y|X)T(Y|X) \ . \tag{2.71}$$

Se construye una cadena de Markov formada por los estados $X_0, X_1, X_2, \ldots X_N$ a los que se va llegando en cada transición, a partir de un X_0 inicial. Cada X_n es una variable aleatoria con pdf $\phi_n(X)$ que cumplirá (lo comprobaremos luego)

$$\lim_{n\to\infty} \phi_n(X) = f(X) \ . \tag{2.72}$$

En cada paso del camino aleatorio hay una probabilidad de transición $T(Y|X)$ que está normalizada,

$$\int \mathrm{d}Y \ T(Y|X) = 1 \ . \tag{2.73}$$

Suponiendo que siempre es posible ir de Y a X si es posible ir de X a Y y viceversa,[e] definimos

$$q(Y|X) = \frac{T(X|Y)f(Y)}{T(Y|X)f(X)} \geq 0 \ . \tag{2.74}$$

A partir de esta $q(Y|X)$ podemos tomar como probabilidad de aceptación (no es la única como veremos luego),

$$A(Y|X) = \mathrm{mín}\{q(Y|X), 1\} \ . \tag{2.75}$$

[d] Introducido por N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller y E. Teller en 1953 [6].

[e] El algoritmo M(RT)^2 es una cadena de Markov irreducible, ergódica y reversible de estados recurrentes y aperiódicos. La *irreducibilidad* garantiza que todos los estados estén comunicados. La *ergodicidad* garantiza que cada estado sea accesible desde cualquier otro en un número finito de pasos. La *reversibilidad* es equivalente a la condición de balance detallado.

Para comprobarlo, veamos que en efecto se verifica la ecuación de balance detallado (2.70, 2.71),

$$f(X)A(Y|X)T(Y|X) = f(Y)A(X|Y)T(X|Y) \ , \tag{2.76}$$

usando que $q(X|Y)q(Y|X) = 1$ (a partir de 2.74). Si suponemos que $q(X|Y) \geq 1$ entonces $A(Y|X) = q(Y|X)$, $A(X|Y) = 1$ y sustituyendo la definición de $q(Y|X)$ tenemos

$$f(X)A(Y|X)T(Y|X) = f(X)\frac{T(X|Y)f(Y)}{T(Y|X)f(X)}T(Y|X) = f(Y)T(X|Y) = f(Y)A(X|Y)T(X|Y) \ . \tag{2.77}$$

Suponiendo $q(X|Y) < 1$ habríamos llegado a la misma conclusión.

Comprobemos ahora (2.72). Para ello, nótese que la probabilidad $\phi_{n+1}(X)$ de obtener un valor X en el paso $n+1$ viene dada por la probabilidad de que habiéndose obtenido cualquier Y en el paso n se acepte el cambio de Y a X, más la probilidad de que habiéndose obtenido X en el paso n se rechace el cambio de X a Y, es decir,

$$\phi_{n+1}(X) = \int \mathrm{d}Y \ \phi_n(Y)A(X|Y)T(X|Y) + \phi_n(X)\int \mathrm{d}Y \ [1 - A(Y|X)]T(Y|X) \ . \tag{2.78}$$

La condición de balance detallado implica que cuando n se hace muy grande el sistema tiende al equilibrio, lo que significa que $\phi_n(X)$ debe ser un punto fijo de la iteración y debe corresponder a la distribución de equilibrio $f(X)$. En efecto, así es porque sustituyendo $\phi_n(X) = f(X)$ en la ecuación anterior y teniendo en cuenta (2.76) obtenemos

$$\begin{aligned}\phi_{n+1}(X) &= \int \mathrm{d}Y \ f(Y)A(X|Y)T(X|Y) + \int \mathrm{d}Y \ f(X)[1 - A(Y|X)]T(Y|X) \\ &= \int \mathrm{d}Y \ f(Y)A(X|Y)T(X|Y) + \int \mathrm{d}Y \ f(X)T(Y|X) - \int \mathrm{d}Y \ f(Y)A(X|Y)T(X|Y) \\ &= \int \mathrm{d}Y \ f(X)T(Y|X) = f(X) \ ,\end{aligned} \tag{2.79}$$

donde se ha usado (2.73).

Como último comentario, mostremos que la probabilidad de aceptación no tiene que ser necesariamente la de (2.75). Otra posibilidad que cumple la ecuación de balance detallado, como puede comprobarse fácilmente, que puede mejorar la convergencia en ciertos casos, es

$$A'(Y|X) = \frac{q(Y|X)}{1 + q(Y|X)} = \frac{1}{1 + q(X|Y)} \ . \tag{2.80}$$

Veamos ya la aplicación de la idea en forma de algoritmo. Supongamos una variable X que toma valores $X_0, X_1, \ldots$ con probabilidad

$$\pi_i = \frac{b_i}{C} \ , \tag{2.81}$$

donde $C = \sum_{i=1}^{N} b_i$ puede ser una constante difícil de calcular. Para muestrear X construimos una cadena de Markov $\{X_n, \ n = 0, 1, \ldots\}$ cuya evolución viene dada por g_{ij} que son los valores de una una pdf *condicionada* $T(Y = x_j | X = x_i)$ dada como *función propuesta*. El algoritmo de Metropolis consiste en comenzar por un estado inicial X_0 e ir reemplazando iterativamente cada estado X_n por el siguiente X_{n+1}:

- Dado $X_n = x_i$, generar un variable aleatoria Y con probabilidad g_{ij}. Sea $y = x_j$.
- Tomar

$$X_{n+1} = \begin{cases} x_j & \text{con probabilidad } \alpha_{ij} = \text{mín}\left\{\dfrac{\pi_j\, g_{ji}}{\pi_i\, g_{ij}}, 1\right\} = \text{mín}\left\{\dfrac{b_j\, g_{ji}}{b_i\, g_{ij}}, 1\right\} \\ x_i & \text{con probabilidad } 1 - \alpha_{ij}\ . \end{cases} \tag{2.82}$$

Los α_{ij} determinan la *probabilidad de aceptación*. Los valores de X se acabarán distribuyendo según $f(X)$ en el límite de muchas iteraciones, una vez que el sistema alcance el equilibrio, lo que no dependerá del estado inicial de partida ni de las g_{ij} propuestas. La normalización C resulta irrelevante. Veamos cómo implementarlo en general:

1. Dado $X_n = x$, generar Y según una función propuesta $g(x, y)$. Sea $Y = y$.
2. Generar $\xi \sim \mathsf{U}[0, 1]$ y tomar

$$X_{n+1} = \begin{cases} y\ , & \text{si } \xi \leq \alpha(x, y) \\ x\ , & \text{en otro caso} \end{cases} \qquad \text{donde } \alpha(x, y) = \text{mín}\left\{\frac{f(y)\, g(y, x)}{f(x)\, g(x, y)}, 1\right\}\ . \tag{2.83}$$

3. Iterar.

El algoritmo más simple se basa en suponer como función propuesta $g(x, y)$ una pdf $g(y)$ independiente de x. Entonces (*independence sampler*):

$$g(x, y) = g(y) \quad \Rightarrow \quad \alpha(x, y) = \text{mín}\left\{\frac{f(y)g(x)}{f(x)g(y)}, 1\right\}\ . \tag{2.84}$$

En la mayoría de las aplicaciones se simplifica aún más suponiendo que la probabilidad de transición es constante en el dominio de X (*random walk sampler*):

$$g(x, y) = C \quad \Rightarrow \quad \alpha(x, y) = \text{mín}\left\{\frac{f(y)}{f(x)}, 1\right\}\ . \tag{2.85}$$

La *mayor ventaja* del algoritmo M(RT)^2 es la facilidad con la que se pueden muestrear distribuciones de probabilidad, incluso si dependen de varias variables $f(\boldsymbol{X})$, particularmente si usamos la opción más simple. Basta reemplazar x e y por vectores $\boldsymbol{x}$ e $\boldsymbol{y}$ de varias componentes.

El *inconveniente* está en que la variable X a muestrear se distribuye como $f(X)$ sólo asintóticamente , es decir, hay que descartar los L primeros pasos del camino aleatorio y no está claro cuál es el número L óptimo. Así por ejemplo, un estimador de la variable aleatoria $E[h(X)]$ debe calcularse como

$$E[h(X)] = \int \mathrm{d}x\ h(x) f(x) = \sum_{n=L}^{L+N-1} \frac{h(X_n)}{N}\ . \tag{2.86}$$

2.5 Tests de los algoritmos de muestreo

Es importante comprobar si un algoritmo logra muestrear X según $f(X)$ con suficiente exactitud. Para ello se pueden distribuir en bines las variables aleatorias generadas y aplicar, por ejemplo, un test de la χ^2. Otra posibilidad es evaluar una integral

$$\int \mathrm{d}x\; g(x)f(x) = E[g(X)]\;, \quad X \sim f(X)\;, \tag{2.87}$$

cuyo valor sea conocido y decidir si es correcto dentro del margen de error que queramos asumir. Resulta particularmente útil tomar los *momentos* de la distribución, $g_n(x) = x^n$.

2.6 Técnicas de integración Monte Carlo

2.6.1 Integración Monte Carlo

Ya sabemos que podemos evaluar la integral

$$G = \int_\Omega \mathrm{d}x\; g(x)f(x)\;, \quad \text{con } f(x) \geq 0\;, \quad \int_\Omega \mathrm{d}x\; f(x) = 1\;, \tag{2.88}$$

sorteando N variables aleatorias $X_1, \ldots, X_N$ según $f(X)$ y haciendo la media aritmética

$$G_N = \frac{1}{N}\sum_i g(X_i)\;. \tag{2.89}$$

La cantidad G_N es un estimador de G y el teorema fundamental del Monte Carlo garantiza que, si la integral (2.88) existe,

$$\mathrm{E}[G_N] = G\;. \tag{2.90}$$

Por otro lado, la ley de los grandes números (1.59) nos dice que

$$\mathrm{P}\left(\lim_{N\to\infty} G_N - G\right) = 1 \tag{2.91}$$

y el teorema del límite central (1.68) nos especifica el margen de error, si la varianza existe, con un cierto nivel de confianza correspondiente a $\delta = n\sigma$ (tabla 1.1 y figura 1.4), dado por

$$\lim_{N\to\infty} \mathrm{P}\left(-n\frac{\sigma_1}{\sqrt{N}} < G_N - G < n\frac{\sigma_1}{\sqrt{N}}\right) = \int_{-n}^{n} \mathrm{d}t\; \frac{\mathrm{e}^{-\frac{1}{2}t^2}}{\sqrt{2\pi}} \tag{2.92}$$

donde $\sigma_1^2 = \mathrm{var}(g(X))$, es decir,

$$\sigma_1^2 = \int \mathrm{d}x\; g^2(x)f(x) - G^2\;. \tag{2.93}$$

Tomando una desviación estándar tenemos que el error en la estimación es

$$|G_N - G| = \epsilon \approx \frac{\sigma_1}{\sqrt{N}}\;, \tag{2.94}$$

que escala como $1/\sqrt{N}$. En la práctica no conocemos el valor exacto de la varianza σ_1^2 pero podemos usar su estimador Monte Carlo,

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}[g(X_i) - G_N]^2 \,. \tag{2.95}$$

Nótese que si la integral (2.88) es sobre d variables $x = (u_1, u_2, \ldots, u_d)$ el error del método Monte Carlo sigue escalando como $1/\sqrt{N}$ *independientemente del número de dimensiones* d.

En cambio, si hacemos la integral por un método tradicional basado en tomar incrementos finitos, por ejemplo la regla del trapecio, suponiendo como volumen de integración un hipercubo $[0,1]^d$,

$$\int \mathrm{d}^d u\, f(u_1, u_2, \ldots, u_d) = \frac{1}{n^d}\sum_{j_1=0}^{n}\cdots\sum_{j_d=0}^{n} f\left(\frac{j_1}{n}, \ldots, \frac{j_d}{n}\right) + \mathcal{O}\left(\frac{1}{n^2}\right) \,, \tag{2.96}$$

necesitamos evaluar la función $N = (n+1)^d \approx n^d$ veces, así que el error va como $n^{-2} = N^{-2/d}$. Puede demostrarse que con otros métodos similares la situación no mejora mucho. Por ejemplo, con la regla de Simpson el error escala como $N^{-4/d}$.

Como el tiempo de computación es proporcional al número de veces N que invocamos la función, está claro que si la integración es sobre un número de dimensiones grande ($d > 4$ si comparamos con la regla del trapecio, $d > 8$ si comparamos con la de Simpson) el método Monte Carlo resultará más eficiente.

A pesar de que la convergencia mejora con N más rápidamente que con los métodos tradicionales, un factor $1/\sqrt{N}$ puede no ser suficientemente satisfactorio. Sin embargo, existen distintas formas de mejorar la convergencia disminuyendo el otro factor relevante en (2.94), es decir, logrando una *reducción de la varianza* σ_1^2:

- Muestreo por importancia.
- Uso de valores esperados.
- Métodos de correlación.

Los estudiaremos a continuación.

2.6.2 Muestreo por importancia

Si deseamos evaluar la integral (2.88), la función $f(x)$ no es necesariamente la mejor pdf que podemos usar para muestrear X, aunque aparezca en el integrando. Podemos introducir en su lugar otra pdf $\tilde{f}(x)$ del siguiente modo,

$$G = \int \mathrm{d}x\, \tilde{g}(x)\tilde{f}(x) \,, \quad \text{con } \tilde{g}(x) = \frac{g(x)f(x)}{\tilde{f}(x)} \,, \quad \tilde{f}(x) \geq 0 \,, \quad \int \mathrm{d}x\, \tilde{f}(x) = 1 \tag{2.97}$$

siendo $\tilde{g}(x) < \infty$ (acotada superiormente) excepto tal vez en un conjunto numerable de puntos. La varianza σ_1^2 cuando usamos $\tilde{f}(x)$ es entonces

$$\mathrm{var}(\tilde{g}(X)) = \int \mathrm{d}x\, \tilde{g}^2(x)\tilde{f}(x) - G^2 \,. \tag{2.98}$$

La función $\tilde{f}(x)$ que minimiza esta varianza para un G fijo es

$$\tilde{f}(x) = \frac{g(x)f(x)}{G} \tag{2.99}$$

pero no es práctica pues involucra el valor de G que pretendemos calcular. De hecho, con esta $\tilde{f}(x)$ ideal, el estimador de la integral sería

$$\tilde{G}_N = \frac{1}{N}\sum_{i=1}^{N}\frac{g(X_i)f(X_i)}{\tilde{f}(X_i)} = \frac{1}{N}\sum_{i=1}^{N} G = G\,, \tag{2.100}$$

que tiene varianza cero. Por tanto, en la práctica, lo que se trata es de introducir una $\tilde{f}(x)$ *lo más parecida posible a* $g(x)f(x)$. Veamos un par de ejemplos.

– Consideremos la integral

$$G = \int_0^1 \mathrm{d}x\ \cos\left(\frac{\pi x}{2}\right)\ . \tag{2.101}$$

Lo directo sería muestrear $X_i \sim \mathsf{U}[0,1]$, es decir $f(x) = 1$, y promediar $g(X_i)$ con

$$g(x) = \cos\left(\frac{\pi x}{2}\right)\ . \tag{2.102}$$

La varianza de esta esta distribución se puede hallar analíticamente de (2.93),

$$\sigma_1^2 = \mathrm{var}(g(X)) = 0.0947.... \tag{2.103}$$

Sin embargo, podemos reducir esta varianza si muestreamos los X_i según

$$\tilde{f}(x) = \frac{3}{2}(1-x^2) \tag{2.104}$$

y promediamos $\tilde{g}(X_i)$ con

$$\tilde{g}(x) = \frac{g(x)f(x)}{\tilde{f}(x)} = \frac{2}{3}\frac{\cos\left(\frac{\pi x}{2}\right)}{1-x^2}\ . \tag{2.105}$$

En este caso la varianza es unas cien veces menor! ya que se obtiene

$$\mathrm{var}(\tilde{g}(X)) = 0.000990.... \tag{2.106}$$

Esta elección de $\tilde{f}(x)$ se debe a su parecido con $g(x)f(x)$ para x pequeños, pues

$$\cos\left(\frac{\pi x}{2}\right) = 1 - \frac{\pi^2 x^2}{8} + \frac{\pi^4 x^4}{2^4 4!} - \cdots \approx 1 - x^2 \quad \Rightarrow \quad \tilde{f}(x) = \frac{3}{2}(1-x^2)\ . \tag{2.107}$$

En general, conviene usar una $\tilde{f}(x)$ que se parezca al integrando $g(x)f(x)$ cerca del máximo, para lo que un desarrollo en serie de Taylor suele ayudar, pues eso garantiza que $\tilde{g}(x) < 1$.

– Cuando el integrando $g(x)f(x)$ es *singular*, *la varianza puede no existir*. En ese caso, puede siempre elegirse $\tilde{f}(x)$ de modo que el cociente $\tilde{g}(x) = g(x)f(x)/\tilde{f}(x)$ esté acotado. Por ejemplo,

$$G = \int_0^1 \mathrm{d}x \, \frac{1}{\sqrt{x}} \, . \tag{2.108}$$

Si directamente muestreamos $X \sim \mathsf{U}[0,1]$, es decir $f(x) = 1$, y usamos $g(x) = 1/\sqrt{x}$ la varianza diverge, pues

$$\mathrm{E}[g(X)^2] = \int_0^1 \frac{\mathrm{d}x}{x} \to \infty \quad \Rightarrow \quad \mathrm{var}(g(X)) = \mathrm{E}[g(X)^2] - G^2 \to \infty \, . \tag{2.109}$$

Como alternativa, podemos probar

$$\tilde{f}(x) = (1-u)x^{-u} \, , \quad \text{con } \frac{1}{2} \leq u < 1 \, . \tag{2.110}$$

El estimador resultante ya no es singular,

$$\tilde{g}(x) = \frac{x^{u-\frac{1}{2}}}{1-u} \quad \Rightarrow \quad \mathrm{var}(\tilde{g}(X)) = \frac{1}{u(1-u)} - 2 \, . \tag{2.111}$$

Obviamente la mejor elección será $u = \frac{1}{2}$, que tiene varianza igual a 2.

Vemos que el método de muestreo por importancia es particularmente útil para reducir la varianza en el caso de integrandos singulares.

2.6.3 Uso de valores esperados

Supongamos que queremos evaluar la integral

$$G = \int \mathrm{d}x \mathrm{d}y \; g(x,y) f(x,y) \, . \tag{2.112}$$

Podemos reescribir esta integral como

$$G = \int \mathrm{d}x \; h(x) m(x) \tag{2.113}$$

donde $m(x)$ es la distribución marginal de x,

$$m(x) = \int \mathrm{d}y \; f(x,y) \tag{2.114}$$

y $h(x)$ es el *valor esperado condicionado* de g dado x,

$$h(x) = \frac{1}{m(x)} \int \mathrm{d}y \; g(x,y) f(x,y) = E[g|x] \, . \tag{2.115}$$

Si las integrales $h(x)$ y $m(x)$ son *conocidas*, es conveniente hallar la integral G muestreando según $m(x)$ y usando $h(x)$ como estimador, pues de esta manera se reduce la varianza, ya que

$$\mathrm{var}(g(x,y)) \geq \mathrm{var}(h(x)) \, . \tag{2.116}$$

En efecto,

$$\begin{aligned}\mathrm{var}(g) - \mathrm{var}(h) &= \mathrm{E}[g^2] - \mathrm{E}[h^2] \\ &= \mathrm{E}[g^2|x] - \mathrm{E}[(\mathrm{E}[g|x])^2] \\ &= \mathrm{E}[\mathrm{E}[g^2|x] - (\mathrm{E}[g|x])^2] = \mathrm{E}[\mathrm{var}(g|x)] \geq 0. \end{aligned} \tag{2.117}$$

Como ejemplo consideremos la integral (que se resuelve trivialmente de forma analítica)

$$G = \int_0^1 \mathrm{d}x \int_0^1 \mathrm{d}y \; g(x,y) \; , \quad g(x,y) = \left\{ \begin{array}{ll} 1 \, , & \text{si } x^2 + y^2 \leq 1 \\ 0 \, , & \text{si } x^2 + y^2 > 1 \, . \end{array} \right. \tag{2.118}$$

Si muestreamos uniformemente x e y en $[0,1]$ y promediamos los $g(x,y)$ la varianza de este estimador es

$$\mathrm{var}(g) = \frac{\pi}{4} - \left(\frac{\pi}{4}\right)^2 = 0.168 \, . \tag{2.119}$$

Pero si tomamos hallamos la distribución marginal,

$$m(x) = \int_0^1 \mathrm{d}y = 1 \tag{2.120}$$

y el valor esperado condicionado,

$$h(x) = \int_0^{\sqrt{1-x^2}} \mathrm{d}y = \sqrt{1-x^2} \, , \tag{2.121}$$

la integral se convierte en

$$G = \int_0^1 \mathrm{d}x \; \sqrt{1-x^2} \, . \tag{2.122}$$

Muestreando uniformemente x en $[0,1]$ y promediando los $h(x)$ la varianza de este nuevo estimador es

$$\mathrm{var}(h) = \int_0^1 \mathrm{d}x \; (1-x^2) - \left(\frac{\pi}{4}\right)^2 = 0.050. \tag{2.123}$$

que supone una reducción de la varianza en algo más de un tercio.

2.6.4 Métodos de correlación

Estos métodos consiguen reducir la varianza usando puntos de la muestra que están correlacionados en vez de muestrear puntos independientemente.

Variantes de control

Consisten en escribir

$$G = \int \mathrm{d}x \; g(x) f(x) = \int \mathrm{d}x \; [g(x) - h(x)] f(x) + \int \mathrm{d}x \; h(x) f(x) \tag{2.124}$$

donde $\int \mathrm{d}x\ h(x)f(x)$ es *conocida analíticamente*. Entonces el estimador de G es

$$G \approx \int \mathrm{d}x\ h(x)f(x) + \frac{1}{N}\sum_{i=1}^{N}[g(x_i) - h(x_i)]\ , \tag{2.125}$$

donde se muestrean $g(x)$ y $h(x)$ en los mismos puntos x_i. Es decir, h y g son dos variables aleatorias *correlacionadas*. Evidentemente esta técnica es ventajosa cuando

$$\mathrm{var}(g-h)_f \ll \mathrm{var}(g)_f\ , \tag{2.126}$$

lo que ocurre si $h(x)$ es *similar* a $g(x)$. Esto nos recuerda al muestreo por importancia, pero es completamente diferente.

Por ejemplo, consideremos la integral

$$G = \int_0^1 \mathrm{d}x\ \mathrm{e}^x\ . \tag{2.127}$$

cuya varianza muestreando x uniformemente en $[0,1]$ y usando e^x como estimador es

$$\mathrm{var}(\mathrm{e}^x) = \frac{1}{2}(\mathrm{e}^2-1) - (\mathrm{e}-1)^2 = 0.242\ . \tag{2.128}$$

Una posibilidad es $h(x) = 1+x$, los dos primeros términos de la serie de Taylor de e^x,

$$G = \int_0^1 \mathrm{d}x\ [\mathrm{e}^x - (1+x)] + \int_0^1 \mathrm{d}x\ (1+x) = \int_0^1 \mathrm{d}x\ [\mathrm{e}^x - (1+x)] + \frac{3}{2}\ , \tag{2.129}$$

cuya varianza es

$$\mathrm{var}(\mathrm{e}^x - (1+x)) = 0.0437\ . \tag{2.130}$$

Puede ensayarse $h(x) = 1+\beta x$ y encontrar que $\beta = 1.69$ es óptimo para reducir la varianza, a un valor 0.0039. Como curiosidad, si se usa $\tilde{f}(x) = 1+\beta x$ como función de importancia, el mejor parámetro es $\beta = 1.81$ y conduce a casi la misma varianza, 0.0040.

Variantes antitéticas

Consisten en explotar la reducción de la varianza que se consigue cuando las variables aleatorias están *negativamente correlacionadas*. Supongamos

$$G = \int_0^1 \mathrm{d}x\ g(x)\ . \tag{2.131}$$

Entonces G puede escribirse exactamente como

$$G = \int_0^1 \mathrm{d}x\ \frac{1}{2}[g(x) + g(1-x)]\ , \tag{2.132}$$

que puede evaluarse muestreando x uniformemente en $[0,1]$ y usando el estimador

$$G_N = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}[g(x_i) + g(1-x_i)] \tag{2.133}$$

que tiene varianza cero si $g(x)$ es lineal. Para funciones no muy alejadas de la linealidad este método reduce la varianza.

Por ejemplo, consideremos la integral de antes,

$$G = \int_0^1 \mathrm{d}x\ \mathrm{e}^x\ , \tag{2.134}$$

cuya varianza usando un Monte Carlo directo es 0.242. Si usamos (2.133) es sólo 0.0039. Otro ejemplo más general,

$$G = \int_0^\infty \mathrm{d}x\ \mathrm{e}^{-x} g(x)\ . \tag{2.135}$$

Esta vez usaremos e^{-x} como función de muestreo, tomando el estimador

$$G = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}[g(x_i) + g(x_i')]\ , \quad x = -\ln\xi\ , \quad x' = -\ln(1-\xi)\ . \tag{2.136}$$

Es posible reducir notablemente la varianza si se combinan dos métodos. Por ejemplo, combinando muestreo por importancia y variantes antitéticas,

$$G = \frac{1}{2}\int_0^1 \mathrm{d}x\ \left[\frac{g(x)+g(1-x)}{\tilde{f}(x)}\right]\tilde{f}(x) \tag{2.137}$$

en el caso de $g(x) = \mathrm{e}^x$ con

$$\tilde{f}(x) = \frac{24}{25}\left[1 + \frac{1}{2}\left(x - \frac{1}{2}\right)^2\right]\ , \tag{2.138}$$

que proviene de los tres primeros términos de la serie de Taylor en torno a $x = \frac{1}{2}$ de $\frac{1}{2}[\mathrm{e}^x + \mathrm{e}^{1-x}]$, tiene una varianza de 0.0000012.

Muestreo estratificado

Consiste en dividir la región de integración en subespacios y hacer una integración Monte Carlo completa en cada uno (la integral total es la suma). Para ello se muestrean variables aleatorias independientes en cada subespacio. En general, troceamos la región de integración M en k subespacios M_j, $j = 1, \ldots, k$ y sorteamos N_j puntos en cada una. Si en total sorteamos N puntos, sabemos que el error estimado, en función de la varianza total σ^2 y de las varianzas σ_j^2 en cada subespacio, es

$$\frac{\sigma^2}{N} = \sum_j \frac{\sigma_j^2}{N_j}\ , \quad N = \sum_{j=1}^{k} N_j\ . \tag{2.139}$$

Si los subespacios y el número de puntos sorteados en cada uno se eligen convenientemente, la varianza σ^2 puede reducirse, aunque también aumentarse si no se hace la elección apropiada. Para ver esto, dividamos el dominio de integración en dos regiones con varianzas σ_a^2 y σ_b^2 en las que sorteamos N_a y N_b puntos respectivamente. Entonces

$$\frac{\sigma^2}{N} = \frac{\sigma_a^2}{N_a} + \frac{\sigma_b^2}{N_b}\ , \quad N = N_a + N_b\ . \tag{2.140}$$

Dado N, el mínimo de σ^2 se obtiene derivando respecto a N_a con $N_b = N - N_a$ y vale

$$\frac{N_a}{N} = \frac{\sigma_a}{\sigma_a + \sigma_b} \quad \Rightarrow \quad \sigma^2 = (\sigma_a + \sigma_b)^2 \,. \tag{2.141}$$

Es decir la varianza es mínima si el número de puntos en cada subespacio j es proporcional a σ_j.

2.6.5 Métodos adaptativos

Todos los métodos de reducción de la varianza que hemos visto hasta ahora requieren algún conocimiento previo de la función a integrar. En los métodos adaptativos, se diseña un *algoritmo que va aprendiendo* sobre el integrando en sucesivas iteraciones. Nos centraremos en VEGAS,[f] que se usa extensivamente en física de altas energías.

VEGAS combina muestreo por importancia y muestreo estratificado de forma iterativa, concentrando de forma automática más puntos en las regiones que contribuyen más a la integral. El algoritmo comienza parcelando el dominio de integración en un retículo (*grid*) rectangular y haciendo una integral en cada celda. Los resultados se usan para *reajustar* el retículo en la siguiente iteración, según su contribución a la integral. A partir de estos resultados se intenta aproximar función de importancia $\tilde{f}(x)$, que idealmente sería

$$\tilde{f}(x) = \frac{|g(x)|}{\int \mathrm{d}x |g(x)|} \tag{2.142}$$

pero en la práctica es una función escalón

$$p(x) = p(u_1, u_2, \ldots, u_d) = p_1(u_1) p_2(u_2) \cdots p_d(u_d) \tag{2.143}$$

si la integral es en d dimensiones, que tiene el mismo valor para todos aquellos puntos que caigan dentro de una misma celda y, por construcción, es separable en un factor por cada una de las variables de integración elegidas. Tras una *fase exploratoria* inicial se fija el retículo optimizado al que se ha llegado, se ignoran las estimaciones preliminares de la integral y se procede a calcular la integral con mayor resolución (*fase de evaluación*). En cada iteración $j = 1, 2, \ldots, m$ de la fase de evaluación se estima una integral G_{Nj} (2.89) con varianza S_j^2 (2.95) sorteando uniformemente N_j puntos x_n (cada uno tiene d coordenadas), $n = 1, \ldots, N_j$,

$$G_{Nj} = \frac{1}{N_j} \sum_{n=1}^{N_j} \frac{g(x_n)}{p(x_n)} \,, \quad S_j^2 = \frac{1}{N_j} \sum_{n=1}^{N_j} \left(\frac{g(x_n)}{p(x_n)} \right)^2 - G_{Nj}^2 \,. \tag{2.144}$$

Los resultados de cada iteración de la fase de evaluación se combinan (no se tira ninguno), pesándolos por el número de llamadas de cada iteración N_j y las varianzas S_j^2,

$$G = \frac{\displaystyle\sum_{j=1}^{m} \frac{N_j}{S_j^2} G_{Nj}}{\displaystyle\sum_{j=1}^{m} \frac{N_j}{S_j^2}} \tag{2.145}$$

[f] Introducido por G.P. Lepage en 1978 [7].

de modo que en el promedio contribuyen más los estimados con mayor número de puntos y los que menos varianza tienen. Si los valores de S_j^2 no son fiables (por ejemplo si $g(x)$ no es de cuadrado integrable) conviene pesar simplemente por N_j. Además VEGAS va dando en cada iteración la χ^2 por grado de libertad (χ^2/dof),

$$\chi^2/\mathrm{dof} = \frac{1}{m-1}\sum_{j=1}^{m}\frac{(G_{Nj}-G)^2}{S_j^2}\,, \tag{2.146}$$

que sirven para verificar que las estimaciones en cada iteración son consistentes. Se espera que los χ^2/dof no sean mucho mayor que uno (véase figura 1.3).

La rutina se encuentra por ejemplo en [8]. También es parte de la librería CUBA [9].

Ejercicios

Tema 3

Algunas aplicaciones físicas de los Métodos Monte Carlo

3.1 Generadores de sucesos en física de partículas

3.2 Contraste de hipótesis

Ejercicios

Bibliografía

[1] M. H. Kalos, P. A. Whitlock,
Monte Carlo Methods,
Wiley, 2nd edition, 2008.

[2] R. Y. Rubinstein, D. P. Kroese,
Simulations and the Monte Carlo Method,
Wiley, 2nd edition, 2008.

[3] S. Weinzierl,
Introduction to Monte Carlo Methods,
arXiv: hep-ph/0006269, 2000.

[4] A. M. Johansen, L. Evers,
Simulation and the Monte Carlo Methods — Lecture Notes,
Ed. Nick Whiteley, University of Bristol, 2011.

[5] J. Beringer *et al.* (Particle Data Group),
Statistics en *Review of Particle Physics*,
Phys. Rev. **D86**, 010001 (2012).

[6] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller,
Equation of state calculations by fast computing machines,
Journal of Chemical Physics **21**, 1087 (1953).

[7] G. P. Lepage,
A New Algorithm for Adaptive Multidimensional Integration,
Journal of Computational Physics **27**, 192 (1978).

[8] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery,
Numerical Recipes in Fortran,
Cambridge University Press, 2nd edition, 1992,
http://apps.nrbook.com/fortran/index.html

[9] T. Hahn,
CUBA: A Library for multidimensional numerical integration,
Comput. Phys. Commun. **168**, 78 (2005) [hep-ph/0404043],
http://www.feynarts.de/cuba/