Métodos MCMC para estadísticas

La Habana – Abril 2019

Xavier Gendre

Métodos MCMC para estadísticas

La Habana – Abril 2019

Yavier Gendre

Índice general

Prólogo

Este documento corresponde a las notas del curso *Métodos MCMC para estadísticas* que se imparte en la Universidad de La Habana del 22 al 26 de abril de 2019. Esto representa 15 horas de clases y dos sesiones prácticas con Python de 2 horas dedicadas al algoritmo de Metrópolis-Hastings y al recocido simulado.

La organización de este curso involucró a varias personas que deben ser agradecidas aquí. Así, agradecemos a Cécile Hardouin, Stéphane Mischler y Madalina Olteanu para la gestión del programa de cooperación científica entre Cuba y Francia. También agradecemos a Marie-Laure Ausset, Carlos Bouza, Josué Corujo, Sébastien Gadat, Dafne Garcia, Willy Rodríguez, Florian Simatos y Vivian Sistachs.

El curso fue impartido por Xavier Gendre del Institut Supérieur de l'Aéronautique et de l'Espace y del Institut de Mathématiques de Toulouse. Para cualquier solicitud o comentario, contáctese con el autor a xavier.gendre@math.univ-toulouse.fr.

1 — Motivaciones

1.1 Simulación de variables aleatorias

Dada una variable aleatoria X de distribución de probabilidad μ con valores sobre un espacio E, el **problema de la simulación** consiste en la construcción de un algoritmo que permite obtener realizaciones de X, *i.e.* generar una muestra de distribución μ. Este problema ocupa un lugar central en los métodos que estudiaremos en este curso. Por supuesto, esta pregunta será aun más interesante cuando la distribución μ sea complicada (modelos de la física estadística, cálculo bayesiano con una distribución a posteriori, ...). Sin embargo, proponemos presentar algunos enfoques simples en un primer momento para introducir ideas fundamentales para lo que sigue.

El primer paso para tratar de resolver el problema de la simulación es generalmente tener una fuente de números (pseudo) aleatorios. Es decir, necesitamos un algoritmo que produzca grandes secuencias de valores indistinguibles de realizaciones independientes de una variable uniforme en $[0,1]$. Tales algoritmos existen y están disponibles en la mayoría de los lenguajes informáticos, pero sus presentaciones están fuera del alcance de este curso. El principio general para producir realizaciones de cierta distribución μ es, por lo tanto, transformar tales muestras de distribución uniforme.

Cuando la variable aleatoria X de distribución μ toma sus valores en $E = \mathbb{R}$, existen varias metodologías clásicas para abordar el problema de la simulación. Por ejemplo, el **método de inversión** considera la función de distribución F de X,

$$\forall x \in \mathbb{R},\ F(x) = \mu\left((-\infty, x]\right) = \mathbb{P}(X \leqslant x)$$

cuya función inversa F^{-1} es llamada **función cuantil** y dada por

$$\forall u \in [0,1],\ F^{-1}(u) = \text{ínf}\left\{x \in \mathbb{R} \text{ tal que } F(x) \geqslant u\right\}.$$

El papel de la distribución uniforme $\mathscr{U}([0,1])$ para simular una realización de la variable X se ilustra con el siguiente resultado.

Teorema 1.1. *Sea μ una distribución de probabilidad sobre $\mathbb{R}$ cuya función cuantil es F^{-1}. Si $U \sim \mathscr{U}([0,1])$ entonces la distribución de la variable $F^{-1}(U)$ es μ.*

Demostración. Sean $x \in \mathbb{R}$ y $0 < u \leqslant 1$, tenemos la equivalencia

$$u \leqslant F(x) \iff F^{-1}(u) \leqslant x$$

donde F es la función de distribución asociada a μ. En efecto, si $u \leqslant F(x)$ entonces $F^{-1}(u) \leqslant x$ por definición de la función de cuantiles. Recíprocamente, si $F^{-1}(u) \leqslant x$ entonces sabemos que $F(F^{-1}(u)) \leqslant F(x)$ porque la función F es creciente. Ya que $F^{-1}(u)$ es un ínfimo, sabemos que existe una secuencia decreciente $(x_n)_{n\in\mathbb{N}}$ que converge a $F^{-1}(u)$ tal que $u \leqslant F(x_n)$, para cualquier $n \in \mathbb{N}$. Dado que la función F es continua a la derecha, deducimos que $u \leqslant F(F^{-1}(u))$ y la equivalencia es probada. Como $F(x) \in [0,1]$,

$$\mathbb{P}(F^{-1}(U) \leqslant x) = \mathbb{P}(U \leqslant F(x)) = F(x).$$

La función de distribución de la variable $F^{-1}(U)$ es F, por lo tanto su distribución es μ. □

ALGORITMO 1.2 – Método de inversión
Inicialización :
- n : tamaño de la muestra a simular
- F^{-1} : función cuantil asociada con la distribución μ

Para $k = 1$ hasta n :
Generar $U \sim \mathscr{U}([0,1])$
Definir $X_k = F^{-1}(U)$
Devolver los valores $X_1, \ldots, X_n$

EJEMPLO 1.3 (Distribución exponencial). La distribución exponencial con un parámetro $\lambda > 0$, expresada $\mathscr{E}(\lambda)$, está dada por su función de densidad f con respecto a la medida de Lebesgue,

$$\forall x \in \mathbb{R},\ f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x > 0, \\ 0 & \text{si no,} \end{cases}$$

y su función de distribución F,

$$\forall x \in \mathbb{R},\ F(x) = \int_{-\infty}^{x} f(t)\,dt = \begin{cases} 1 - e^{-\lambda x} & \text{si } x > 0, \\ 0 & \text{si no.} \end{cases}$$

La función de cuantiles F^{-1} se deduce fácilmente,

$$\forall 0 < u \leqslant 1,\ F^{-1}(u) = -\frac{1}{\lambda}\ln(1-u).$$

Sea $U \sim \mathscr{U}([0,1])$, la distribución de la variable $1-U$ también es $\mathscr{U}([0,1])$ y deducimos que

$$-\frac{1}{\lambda}\ln(U) \sim \mathscr{E}(\lambda).$$

■

Desde un punto de vista teórico, el método de inversión es interesante porque es simple y permite una **simulación exacta**, es decir la distribución simulada es μ y no una distribución aproximada. En la práctica, si la función de cuantiles no es explícita, se plantean varios problemas. Por ejemplo, un cálculo aproximado de los cuantiles (capacidad computacional, ...) puede impedir la aparición de ciertos valores posibles. Sin embargo, estas dificultades no son el principal problema y el método de inversión encuentra las mismas limitaciones que los otros enfoques clásicos (composición, convolución, ...). En el caso real, estos métodos son inutilizables cuando la distribución μ no se conoce completamente.

1.2 Método de rechazo

Un marco estadístico común en el que no se conoce explícitamente la distribución de probabilidad μ a simular es cuando la constante de normalización de esta distribución es imposible o demasiado costosa de calcular. Como mencionamos anteriormente, el método de inversión no se puede utilizar en este caso y veremos varios ejemplos de dichos marcos estadísticos en lo que sigue. En los años 1950, el matemático y físico húngaro-estadounidense John von Neumann introdujo el **método de rechazo** basado en el vínculo entre el valor de una probabilidad y el área bajo una curva para evitar esta dificultad.

El **método de rechazo** es un primer ejemplo de **método de simulación indirecto** en el sentido de que una **distribución candidata** ν es utilizada para generar un valor que debe pasar una prueba para ser aceptado como una realización de la distribución μ. Idealmente, el problema de la simulación tiene que ser más fácil para ν que para μ. Este tipo de enfoque es muy flexible y puede utilizarse para simular un gran número de distribuciones de probabilidad. Los algoritmos estocásticos estudiados en los siguientes capítulos se basan esencialmente en el mismo principio.

El siguiente teorema da una primera versión simple del método de rechazo para resolver el problema de la simulación para la distribución uniforme en un conjunto medible $B \subset \mathbb{R}^2$ con medida de Lebesgue finita.

Teorema 1.4. *Sean $B \subset A$ dos subconjuntos de $\mathbb{R}^2$ tales que sus medidas de Lebesgue verifican $0 < \lambda(B) \leqslant \lambda(A) < +\infty$. Consideremos una sucesión $(X_n)_{n\geqslant 1}$ de variables aleatorias independientes de misma distribución uniforme en A y pongamos*

$$T = \text{ínf}\{n \geqslant 1 \text{ tal que } X_n \in B\}.$$

Entonces,

- *T sigue la distribución geométrica $\mathcal{G}(p)$ con parámetro $p = \lambda(B)/\lambda(A)$,*

$$\forall k > 0,\ \mathbb{P}(T = k) = (1-p)^{k-1}p,$$

- *las variables T y X_T son independientes,*
- *X_T sigue la distribución uniforme en B.*

Demostración. Por definición de la distribución uniforme en A, sabemos

$$\forall n \geqslant 1,\ \mathbb{P}(X_n \in B) = p.$$

La distribución de la variable T se deduce utilizando la independencia,

$$\begin{aligned}\forall k > 0,\ \mathbb{P}(T = k) &= \mathbb{P}(X_1 \notin B, \ldots, X_{k-1} \notin B, X_k \in B)\\ &= \mathbb{P}(X_1 \notin B) \times \cdots \times \mathbb{P}(X_{k-1} \notin B) \times \mathbb{P}(X_k \in B)\\ &= (1-p)^{k-1}p.\end{aligned}$$

Sea $C \subset B$ un subconjunto medible, los mismos argumentos llevan a

$$\forall k > 0,\ \mathbb{P}(X_T \in C \text{ y } T = k) = (1-p)^{k-1}\frac{\lambda(C)}{\lambda(A)}.$$

La independencia entre las variables T y X_T proviene del producto. Al final, la distribución de la variable X_T se logra haciendo la suma,

$$\mathbb{P}(X_T \in C) = \sum_{k>0} \mathbb{P}(X_T \in C \text{ y } T = k) = \sum_{k>0} (1-p)^{k-1} \frac{\lambda(C)}{\lambda(A)} = \frac{\lambda(C)}{\lambda(B)}.$$

□

ALGORITMO 1.5 – Método de rechazo
Inicialización :
- n : tamaño de la muestra a simular
- $B \subset A$ dos subconjuntos de $\mathbb{R}^2$ de área finito

Para $k = 1$ hasta n :
  Repetir
    Generar una variable U uniformemente distribuida en A
  Hasta que $U \in B$
  Definir $X_k = U$
Devolver los valores $X_1, \ldots, X_n$

Es interesante notar que el método de rechazo no requiere conocer el valor de $\lambda(B)$, *i.e.* la constante de normalización $1/\lambda(B)$ de la distribución uniforme en B a simular no es explícita. La distribución candidata aquí es la distribución uniforme en A que es fácil de simular si, por ejemplo, A es un producto de intervalos compactos. El número promedio de rechazos para obtener una realización de la distribución uniforme en B es $p^{-1} = \lambda(A)/\lambda(B) \geqslant 1$. Dado que cada valor rechazado representa tiempo de cálculo no utilizado, es crucial minimizar el área de $A \setminus B$ para tener un buen desempeño.

EJEMPLO 1.6 (Distribución uniforme en el círculo). El círculo de radio 1 es definido por

$$\mathscr{C} = \left\{(x,y) \in \mathbb{R}^2 \text{ tal que } x^2 + y^2 \leqslant 1\right\}.$$

Para simular un valor uniformemente distribuido en $\mathscr{C} \subset [0,1]^2$ con el método de rechazo, podemos tomar dos variables $U, V \sim \mathscr{U}([0,1])$ independientes y rechazar el vector (U,V) cuando $U^2 + V^2 > 1$. ■

Utilizando el método de rechazo, es relativamente sencillo establecer un algoritmo de simulación para una distribución continua sobre $\mathbb{R}$ basado en el siguiente lema.

Lema 1.7. *Sea μ una distribución continua sobre $\mathbb{R}$ de función de densidad f_μ con respecto a la medida de Lebesgue. Si el vector aleatorio (X,Y) sigue una distribución uniforme en*

$$B_\mu = \left\{(x,y) \in \mathbb{R}^2 \textit{ tal que } 0 < y < f_\mu(x)\right\}$$

entonces la abscisa X sigue la distribución μ.

Demostración. Sea un intervalo $I \subset \mathbb{R}$, definimos el conjunto

$$C = \left\{(x,y) \in B_\mu \text{ tal que } x \in I\right\}.$$

Dado que f_μ es una función de densidad de probabilidad, la medida de Lebesgue de B_μ es igual a $\lambda(B_\mu) = 1$ y la de C es $\lambda(C) = \mu(I)$. Por definición de la distribución uniforme en B_μ, obtenemos

$$\mathbb{P}((X,Y) \in C) = \frac{\lambda(C)}{\lambda(B_\mu)} = \mu(I). \tag{1.1}$$

Por otra parte, sabemos que $Y < f_\mu(X)$ casi seguramente, entonces

$$\mathbb{P}((X,Y) \in C) = \mathbb{P}\left(X \in I \text{ y } Y < f_\mu(X)\right) = \mathbb{P}(X \in I).$$

Por lo tanto $\mathbb{P}(X \in I) = \mu(I)$ y X sigue la distribución μ. □

En la práctica, este lema y el método de rechazo permiten reducir el problema de la simulación de la distribución continua μ al de la distribución uniforme en un conjunto $A \subset \mathbb{R}^2$ que contiene B_μ. El caso más simple es el de una función de densidad f_μ que tiene una cota superior $M > 0$ y un soporte incluido en un intervalo I finito. Entonces, basta con tomar $A = I \times [0,M]$. La elección de la cota M puede ser difícil en la práctica y requiere a veces una evaluación costosa. Por supuesto, siempre es posible tomar una cota más grande que lo necesario, pero esto resulta en un aumento del número de valores candidatos rechazados y, por lo tanto, un tiempo de cálculo mayor. Por el contrario, la cota M no se debe subestimar porque la función de densidad quedariá truncada. También notamos que solo puede usarse una función $\tilde{f}_\mu$ proporcional a la función de densidad f_μ para obtener el mismo resultado, *i.e.* existe una constante $K > 0$ tal que $f_\mu(x) = K\tilde{f}_\mu(x)$ para cualquier $x \in \mathbb{R}$. En este caso, el conjunto B_μ puede ser reemplazado por

$$\tilde{B}_\mu = \left\{(x,y) \in \mathbb{R}^2 \text{ tal que } 0 < y < \tilde{f}_\mu(x)\right\}.$$

El lema sigue siendo cierto y la abscisa X sigue la distribución μ porque la constante K desaparece en el cociente de las medidas de Lebesgue (1.1). El siguiente algoritmo detalla los pasos de este método para su implementación.

ALGORITMO 1.8 – Método de rechazo por una densidad mayorada con soporte finito

Inicialización :
- n : tamaño de la muestra a simular
- $\tilde{f}_\mu$: función proporcional a la densidad de la distribución a simular
- $A = [a,b] \times [0,M]$ tal que $\tilde{B}_\mu \subset A$

Para $k = 1$ hasta n :
  Repetir
    Generar un vector (U,V) uniformemente distribuido en A
  Hasta que $V < \tilde{f}_\mu(U)$
  Definir $X_k = U$

Devolver los valores $X_1, \ldots, X_n$

EJEMPLO 1.9 (Distribución beta 1). Sean $\alpha, \beta > 0$, la distribución beta $\mathscr{B}e(\alpha,\beta)$ en $[0,1]$ está dada por la función de densidad

$$\forall x \in [0,1],\ f_{\alpha,\beta}(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

donde $\Gamma(z) = \displaystyle\int_0^{+\infty} t^{z-1}e^{-t}dt$, $z > 0$. Consideramos la función proporcional

$$\forall x \in [0,1],\ \tilde{f}_{\alpha,\beta}(x) = x^{\alpha-1}(1-x)^{\beta-1}.$$

Por ejemplo, con $\alpha = 2{,}7$ y $\beta = 6{,}3$, un simple estudio de la función $\ln(\tilde{f}_{\alpha,\beta})$ muestra que $M = 0{,}021$ es suficiente para simular a partir de la distribución uniforme en $A = [0,1] \times [0,M]$ por el método de rechazo. ■

Idealmente, para reducir el número de valores rechazados y mejorar el rendimiento del algoritmo, tenemos que considerar un conjunto A más cercano a $\tilde{B}_\mu$ que el producto de intervalos compactos. El siguiente lema propone una especie de recíproco al lema 1.7 que nos permite sortear esta dificultad.

Lema 1.10. *Sea ν una distribución continua sobre $\mathbb{R}$ de función de densidad f_ν con respecto a la medida de Lebesgue. Si $X \sim \nu$ y $U \sim \mathcal{U}([0,1])$ son independientes, entonces el vector aleatorio $(X, f_\nu(X)U)$ sigue la distribución uniforme en el conjunto*

$$A_\nu = \left\{(x,y) \in \mathbb{R}^2 \text{ tal que } 0 < y < f_\nu(x)\right\}.$$

Demostración. Sea un conjunto medible $C \subset A_\nu$ de medida de Lebesgue $\lambda(C)$ del que sacamos las rebanadas,

$$\forall x \in \mathbb{R},\ C_x = \{y \in \mathbb{R}_+ \text{ tal que } (x,y) \in C\}.$$

Si $f_\nu(x) = 0$, entonces C_x es el conjunto vacío y su medida de Lebesgue es cero. Por la independencia entre U y X, obtenemos

$$\begin{aligned}
\mathbb{P}((X, f_\nu(X)U) \in C) &= \int_{\mathbb{R}} \mathbb{P}((X, f_\nu(X)U) \in C \mid X = x)\, f_\nu(x)dx \\
&= \int_{\mathbb{R}} \mathbb{P}(f_\nu(x)U \in C_x)\, f_\nu(x)\mathbf{1}_{f_\nu(x)>0}\, dx \\
&= \int_{\mathbb{R}} \frac{\lambda(C_x)}{f_\nu(x)} f_\nu(x)\mathbf{1}_{f_\nu(x)>0}\, dx \\
&= \lambda(C).
\end{aligned}$$

Entonces $(X, f_\nu(X)U)$ sigue la distribución uniforme en A_ν ya que f_ν es una función de densidad, *i.e.* $\lambda(A_\nu) = 1$. □

En particular, este lema nos permite de usar el método de rechazo para simular una distribución continua μ sobre $\mathbb{R}$ con cualquier distribución candidata ν continua sobre $\mathbb{R}$ tal que $B_\mu \subset A_\nu$. Este método se llama **método de rechazo comparativo** y se formaliza mediante el teorema siguiente.

Teorema 1.11. *Sean μ y ν dos distribuciones continuas sobre $\mathbb{R}$ de función de densidad respectivas f_μ y f_ν con respecto a la medida de Lebesgue tales que existe una constante $M \geqslant 1$ que verifica*

$$\forall x \in \mathbb{R},\ f_\mu(x) \leqslant M f_\nu(x).$$

Consideremos una sucesión $(U_n)_{n\geqslant 1}$ de variables aleatorias independientes de distribución $\mathscr{U}([0,1])$ y una sucesión $(X_n)_{n\geqslant 1}$ de variables aleatorias independientes de distribución ν. Definimos

$$T = \text{ínf}\{n \geqslant 1 \text{ tal que } M f_\nu(X_n) U_n < f_\mu(X_n)\}.$$

Si las sucesiones $(U_n)_{n\geqslant 1}$ y $(X_n)_{n\geqslant 1}$ son independientes, entonces

- *T sigue la distribución geométrica $\mathscr{G}(1/M)$,*
- *las variables T y X_T son independientes,*
- *X_T sigue la distribución μ.*

Demostración. Sea $n \geqslant 1$, el lema 1.10 implica que el vector aleatorio $(X_n, f_\nu(X_n)U_n)$ sigue la distribución uniforme en A_ν. Se deduce que el vector aleatorio $(X_n, M f_\nu(X_n)U_n)$ sigue la distribución uniforme en

$$A = \left\{(x,y) \in \mathbb{R}^2 \text{ tal que } 0 < y < M f_\nu(x)\right\}.$$

Por hipótesis, $B_\mu \subset A$ y las conclusiones se deducen del lema 1.7 y del teorema 1.4. □

El método de rechazo comparativo reduce el problema de la simulación de la distribución μ al de la distribución ν. Si la distribución candidata ν se elige correctamente, este método permite reducir el número de valores rechazados y, por lo tanto, mejorar el rendimiento. Este desempeño será mejor si M está cerca de 1, el caso extremo $M = 1$ corresponde a $\nu = \mu$. Como antes, notamos que la función de densidad f_μ puede ser conocida solo a una constante multiplicativa, *i.e.* usando una función $\tilde{f}_\mu$ tal que existe una constante $K > 0$ y $f_\mu(x) = K\tilde{f}_\mu(x)$ para cualquier $x \in \mathbb{R}$. En la práctica, tener un buen valor para la constante $\tilde{M} = M/K$ que contiene la constante de normalización puede ser costoso, pero no debe ser subestimada, o la distribución a simular quedariá truncada.

ALGORITMO 1.12 – Método de rechazo comparativo

Inicialización :
- n : tamaño de la muestra a simular
- $\tilde{f}_\mu$: función proporcional a la densidad de la distribución a simular
- f_ν : función de densidad candidata tal que existe $\tilde{M} > 0$ que verifica $\tilde{f}_\mu \leqslant \tilde{M} f_\nu$

Para $k = 1$ hasta n :
  Repetir
    Generar $U \sim \mathscr{U}([0,1])$ y $X \sim \nu$ independientemente
  Hasta que $\tilde{M} f_\nu(X) U < \tilde{f}_\mu(X)$
  Definir $X_k = X$

Devolver los valores $X_1, \ldots, X_n$

EJEMPLO 1.13 (Distribución beta 2). Retomemos el problema de la simulación de la distribución beta $\mathscr{B}e(2{,}7,6{,}3)$ cuya función de densidad es proporcional a

$$\forall x \in [0,1],\ \tilde{f}(x) = x^{1,7}(1-x)^{5,3}.$$

Un simple estudio de función es suficiente para establecer que $\tilde{M} = 0{,}0141$ es tal que

$$\forall x \in [0,1],\ \tilde{f}(x) \leqslant \tilde{M} f_{2,6}(x)$$

donde $f_{2,6}$ es la función de densidad de la distribución beta $\mathscr{B}e(2,6)$. Simular distribuciones beta con parámetros enteros es fácil con realizaciones independientes de la distribución exponencial como las del ejemplo 1.3. En efecto, si $E_1,\ldots,E_8$ son variables aleatorias independientes de distribución $\mathscr{E}(1)$, entonces se puede demostrar que

$$\frac{E_1+E_2}{E_1+\cdots+E_8} \sim \mathscr{B}e(2,6).$$

Así, podemos usar el método de rechazo comparativo para simular la distribución $\mathscr{B}e(2{,}7,6{,}3)$ rechazando menos valores que en el ejemplo 1.9. ■

El método de rechazo es sobradamente utilizado en la práctica. Tiene la ventaja de producir una simulación exacta y es fácilmente generalizado a distribuciones continuas sobre $\mathbb{R}^d$. El hecho de poder usar este método para distribuciones cuya función de densidad se conoce solo por una constante multiplicativa es una virtud particularmente útil en varios ambitos. La desventaja de este enfoque radica en la dificultad de calibrar correctamente la constante del procedimiento. Si se subestima la constante, la función de densidad a simular se trunca y la simulación es incorrecta. Si se sobreestima la constante, el número de valores rechazados es grande y el rendimiento malo. Encontraremos el mismo principio de distribución candidata y valores a rechazar en los métodos desarollados en los capítulos siguientes.

1.3 Estimación de integrales por el método de Monte Carlo

Para concluir este primer capítulo, presentamos una aplicación de la simulación de una distribución para la estimación del valor de una integral. En efecto, cuando una variable X de distribución μ sobre un espacio E interviene en un problema de modelización o en estadísticas, a menudo nos interesan cantidades que se expresan como integrales de funciones $h : \Theta \times E \to \mathbb{R}$,

$$\forall \theta \in \Theta,\ H(\theta) = \mathbb{E}[h(\theta,X)] = \int_{\mathscr{X}} h(\theta,x)d\mu(x).$$

No faltan ejemplos de tales problemas (constante de normalización de una función de densidad, momentos de una variable aleatoria, estimador bayesiano para la función de pérdida cuadrática, valores de cuantiles, ...) y es natural querer aproximarse estas cantidades usando realizaciones de X obtenidas por el experimento o por simulación.

A finales de los años 1940, el matemático polaco-americano Stanislaw Ulam trabajó en el mismo equipo que John von Neumann para la concepción del arma nuclear en Los Álamos. Es en el marco de estos trabajos que Ulam formalizó un método ya conocido para estimar el valor de una integral. Dada la naturaleza de sus actividades, Ulam y von Neumann necesitaron un nombre en clave para hablar sobre este método. Nicholas Metropolis, un otro colega de Los Álamos de quien hablaremos más adelante, les sugirió el nombre de **Monte Carlo** en referencia al casino de Mónaco.

El principio del **método de Monte Carlo** es relativamente simple. Dada una distribución de probabilidad μ sobre un espacio E, el objetivo es calcular un valor numérico aproximado de la integral de una función $h : E \to \mathbb{R}$,

$$J = \int_E h(x)\, d\mu(x). \tag{1.2}$$

Para ello, el método consiste en tomar una sucesión $(X_n)_{n\geqslant 1}$ de variables aleatorias independientes de distribución μ y definir

$$\forall n \geqslant 1,\ J_n = \frac{1}{n}\sum_{k=1}^{n} h(X_k).$$

Dicha sucesión puede, por ejemplo, obtenerse mediante un método de rechazo como en la sección anterior.

Este método se basa en la ley de los grandes números que garantiza que J_n converge casi seguramente a J cuando el número n de realizaciones va al infinito,

$$J_n \xrightarrow[n\to+\infty]{c.s.} J.$$

Además, si la función h es de cuadrado integrable,

$$\int_{\mathbb{R}^d} h(x)^2\, d\mu(x) < +\infty,$$

entonces la varianza $\sigma^2 = \mathbb{V}(h(X_1))$ está definida y la desigualdad de Chebyshev implica

$$\forall \varepsilon > 0,\ \mathbb{P}\left(|J_n - J| \geqslant \sqrt{\frac{\sigma^2}{n\varepsilon}}\right) \leqslant \varepsilon.$$

En otras palabras, si conocemos una cota superior para la varianza (o si esta varianza puede estimarse), entonces tenemos un intervalo de confianza no asintótico para medir la calidad de la aproximación de J por J_n. Por el teorema central del límite, también sabemos que la distribución de $\sqrt{n}(J_n - J)$ converge a la distribución normal $\mathcal{N}(0,\sigma^2)$,

$$\sqrt{n}(J_n - J) \xrightarrow[n\to+\infty]{d} \mathcal{N}(0,\sigma^2),$$

y también conocemos el comportamiento asintótico de J_n.

ALGORITMO 1.14 – Método de Monte Carlo
Inicialización :
- n : número de iteraciones del algoritmo
- μ : distribución sobre E
- $h : E \to \mathbb{R}$: función integrable

Para $k = 1$ hasta n :
  Generar $X_k \sim \mu$
  Calcular $h(X_k)$
Devolver el valor $(h(X_1) + \cdots + h(X_n))/n$

Este método se generaliza a las **cadenas de Markov** que son sucesiones de variables aleatorias que no son independientes, ni se distribuyen de manera idéntica. Estos resultados serán el tema del próximo capítulo.

EJEMPLO 1.15 (Valor numérico aproximado de una integral). Sea la función h definida en $[0,1]$ por

$$\forall x \in [0,1],\ h(x) = (\cos(50x) + \sin(20x))^2 .$$

Para tener un valor numérico aproximado de la integral $J = \displaystyle\int_0^1 h(x)\,dx$, consideramos una sucesión $(U_n)_{n\geqslant 1}$ de variables aleatorias independientes y uniformemente distribuidas en $[0,1]$ y consideramos

$$\forall n \geqslant 1,\ J_n = \frac{1}{n}\sum_{k=1}^{n} h(U_k).$$

Una cota superior burda de la intergral del cuadrado de h nos da $\mathbb{V}(h(U_1)) \leqslant 16$ y, con la desigualdad de Chebyshov, deducimos con una probabilidad mayor que 90 %,

$$|J_n - J| \leqslant \frac{12{,}65}{\sqrt{n}}.$$

Este resultado no es muy fino pero nos asegura que el primer dígito después de la coma es preciso con gran probabilidad en cuanto el número de realizaciones supere 27723. Se puede también usar el teorema central del límite para tener un intervalo de confianza sobre el valor J.

■

EJEMPLO 1.16 (Función de distribución). La función de distribución de la distribución normal estándar $\mathcal{N}(0,1)$ está dada por

$$\forall x \in \mathbb{R},\ \varphi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

Esta función necesita ser evaluada a menudo en estadísticas, por ejemplo para obtener los cuantiles útiles para construir intervalos de confianza asintóticos. La función φ no es explícita y es necesario aproximarla en la práctica. Con una sucesión $(Z_n)_{n\in\mathbb{N}}$ de variables normales estándar, el método de Monte Carlo nos lleva a considerar la función de distribución empírica

$$\forall n \in \mathbb{N},\ \forall x \in \mathbb{R},\ \varphi_n(x) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{Z_k \leqslant x}.$$

Puntualmente, la ley de los grandes números nos asegura que $\varphi_n(x)$ converge casi seguramente a $\varphi(x)$ para cualquier $x \in \mathbb{R}$. El teorema de Glivenko-Cantelli da el mismo resultado uniformemente en $\mathbb{R}$,

$$\sup_{x\in\mathbb{R}} |\varphi_n(x) - \varphi(x)| \xrightarrow[n\to+\infty]{c.s.} 0.$$

■

Figura 1.1: John von Neumann (1903-1957) y Stanislaw Ulam (1909-1984).

2 — Un paseo con Markov

2.1 Muestreo de rebanada

En el capítulo anterior discutimos algoritmos para el problema de la simulación para una distribución μ determinada. Para motivar la introducción de objetos matemáticos que jugarán un papel central en lo que sigue, consideramos aquí un caso particular que hace posible generar realizaciones de una distribución cercana a μ sin tener que estimar una constante costosa a evaluar. La principal diferencia con los métodos mencionados en el capítulo anterior reside en el hecho de que estas realizaciones ya no serán independientes.

Empezamos recordando el resultado del lema 1.7. Dada una distribución μ sobre $\mathbb{R}$ de función de densidad f con respecto a la medida de Lebesgue, si sabemos generar un vector aleatorio (X,Y) uniformemente distribuido en

$$B = \left\{(x,y) \in \mathbb{R}^2 \text{ tal que } 0 < y < \tilde{f}(x)\right\}$$

donde $\tilde{f} : \mathbb{R} \to \mathbb{R}_+$ es una función proporcional a f, entonces la variable X sigue la distribución μ. El principio que vamos a desarrollar es construir un **camino aleatorio** en B, *i.e.* una sucesión $((X_n,Y_n))_{n\in\mathbb{N}}$ de vectores aleatorios en B construida por recurrencia. La idea es que si este camino explora B adecuadamente, entonces es posible que la distribución de (X_n,Y_n) esté cerca de la distribución uniforme en B cuando n se hace grande y, por lo tanto, la distribución X_n esté cerca de μ. Si el camino aleatorio es tal que, para cualquier $n \in \mathbb{N}$, el punto (X_{n+1},Y_{n+1}) solo depende de la posición anterior (X_n,Y_n), es entonces un ejemplo simple de una **cadena de Markov**.

Hay varias maneras de construir un camino aleatorio en B, pero una forma relativamente natural es alternar movimientos verticales y horizontales sin dejar de pertenecer al conjunto B. Así, a partir de un punto $(X_0,Y_0) \in B$ arbitrario, se puede generar una realización Y_1 uniforme en las ordenadas,

$$Y_1 \mid X = X_0 \sim \mathscr{U}\left(\{y \in \mathbb{R} \text{ tal que } 0 < y < \tilde{f}(X_0)\}\right)$$

y después generar una realización X_1 uniforme en las abscisas,

$$X_1 \mid Y = Y_1 \sim \mathscr{U}\left(\{x \in \mathbb{R} \text{ tal que } 0 < Y_1 < \tilde{f}(x)\}\right).$$

Una iteración del camino aleatorio consiste entonces en moverse del punto (X_0,Y_0) al punto (X_1,Y_1). Al iterar este procedimiento, construimos un camino aleatorio $((X_n,Y_n))_{n\in\mathbb{N}}$ en B que coincide con el **muestreo de rebanada** (o **slice sampling**) tal como fue introducido en [DWW99] y [Nea03].

ALGORITMO 2.1 – Muestreo de rebanada

Inicialización :

- $\tilde{f}$: función proporcional a la densidad de la distribución a simular
- $(X_0, Y_0) \in \mathbb{R}^2$ tal que $0 < Y_0 < \tilde{f}(X_0)$

En el paso $n \geqslant 1$:

Generar $Y_n \sim \mathscr{U}([0, \tilde{f}(X_{n-1})])$

Generar X_n uniformemente distribuida en

$$A_n = \{x \in \mathbb{R} \text{ tal que } 0 < Y_{n-1} < \tilde{f}(x)\}$$

Devolver los valores $X_0, X_1, \ldots$

Notemos que este algoritmo solo devuelve las abscisas. A diferencia del método de rechazo, no perdemos tiempo de cálculo al rechazar valores, ya que el resultado de cada iteración se usa en el camino aleatorio. Además, para cualquier $n \in \mathbb{N}$, la variable X_{n+1} depende solo del valor anterior X_n y, por lo tanto, el muestreo de rebanada es un método que produce una **cadena de Markov** en el sentido indicado anteriormente. En la práctica, generar una variable uniforme en A_n puede ser complicado y existen variaciones para tratar de sortear esta dificultad. Para mantener esta introducción simple, no discutiremos este problema aquí, pero el lector interesado encontrará algunas pistas en el capítulo 8 de [RC04].

El punto de partida (X_0, Y_0) del muestreo de rebanada puede ser elegido arbitrariamente o generado según cualquier distribución sobre B. Las consecuencias de esta elección no son negativas y es una propiedad de **pérdida de memoria** de la cadena de Markov que formalizaremos más adelante.

El hecho de que $(X_n)_{n\in\mathbb{N}}$ sea una cadena de Markov es la propiedad importante aquí. Esto se formaliza por la **distribución de transición** que, para una iteración $n \in \mathbb{N}$, describe los valores posibles de X_{n+1} condicionalmente a X_n. En el caso del muestreo de rebanada, es más sencillo describir esta distribución a partir de las variables $\tilde{f}(X_n)$ que verifican, par cualquier $t, v \in \mathbb{R}_+$,

$$\begin{aligned}
\mathbb{P}(\tilde{f}(X_{n+1}) \leqslant t \mid \tilde{f}(X_n) = v) &= \int_0^v \mathbb{P}(\tilde{f}(X_{n+1}) \leqslant t \mid \tilde{f}(X_n) = v \text{ y } Y_{n+1} = u) \frac{du}{v} \\
&= \frac{1}{v} \int_0^v \frac{m(u) - m(t)}{m(u)} \mathbf{1}_{u<t} \, du \\
&= \frac{1}{v} \int_0^v \left(1 - \frac{m(t)}{m(u)}\right)_+ du
\end{aligned}$$

donde $x_+ = \max\{x, 0\}$ y, para cualquier $u > 0$, $m(u)$ es la medida de Lebesgue

$$m(u) = \lambda\left(\{x \in \mathbb{R} \text{ tal que } \tilde{f}(x) > u\}\right).$$

Es interesante notar que esta distribución de transición no depende de n y sigue siendo válida en cada iteración. Se habla de cadena de Markov **homogénea** para designar esta propiedad.

Nos interesa aquí la distribución del vector (X_n, Y_n) cuando n tiende a infinito. Una consecuencia de la homogeneidad es que si esta distribución converge a una distribución π, entonces π debe necesariamente conservarse durante la transición de (X_n, Y_n) a (X_{n+1}, Y_{n+1}). En otras palabras, si la distribución π existe y es tal que $(X_n, Y_n) \sim \pi$, entonces $(X_{n+1}, Y_{n+1}) \sim \pi$. En el vocabulario de las cadenas de Markov, la distribución π se denomina **distribución invariante** y se necesitan resultados teóricos para garantizar su existencia y unicidad. Para el muestreo

de rebanada, notemos que la distribución uniforme en B es invariante, por lo que es un buen candidato para la distribución límite y valida nuestra motivación inicial. De hecho, si (X_n, Y_n) sigue la distribución uniforme en B, entonces sabemos que $X_n \sim \mu$ y la función de densidad del vector (X_n, Y_{n+1}) está dada por

$$(t,y) \in \mathbb{R}^2 \mapsto f(t) \times \frac{\mathbf{1}_{0<y<\tilde{f}(t)}}{\tilde{f}(t)} \propto \mathbf{1}_{0<y<\tilde{f}(t)}$$

porque existe una constante $C > 0$ tal que, para cualquier $t \in \mathbb{R}$, $\tilde{f}(t) = f(t)/C$. También, la función de densidad del vector (X_n, X_{n+1}, Y_{n+1}) puede escribirse

$$(t,x,y) \in \mathbb{R}^3 \mapsto C\mathbf{1}_{0<y<\tilde{f}(t)} \times \frac{\mathbf{1}_{y<\tilde{f}(x)}}{m(y)}.$$

Integrando con respecto a X_n, obtenemos la función de densidad del vector (X_{n+1}, Y_{n+1}),

$$(x,y) \in \mathbb{R}^2 \mapsto C\mathbf{1}_{0<y<\tilde{f}(x)} \int_{\mathbb{R}} \frac{\mathbf{1}_{y<\tilde{f}(t)}}{m(y)} dt \propto \mathbf{1}_{0<y<\tilde{f}(x)}.$$

Así, el vector aleatorio (X_{n+1}, Y_{n+1}) también sigue la distribución uniforme en B que por lo tanto es invariante para el muestreo de rebanada.

Sea $x \in \mathbb{R}$, la notación $\mathbb{P}^x$ corresponde a la distribución del muestreo de rebanada $(X_n)_{n\in\mathbb{N}}$ con $X_0 = x$. Si la función $\tilde{f}$ está acotada y su soporte está incluido en un intervalo finito, es posible probar que la cadena de Markov $(X_n)_{n\in\mathbb{N}}$ es **uniformemente ergodica**, es decir que verifica

$$\forall x \in \mathbb{R},\ \|\mathbb{P}^x(X_n \in \cdot) - \mu\|_{VT} = \sup_A |\mathbb{P}^x(X_n \in A) - \mu(A)| \xrightarrow[n\to+\infty]{} 0 \tag{2.1}$$

donde el supremo se toma sobre todos los subconjuntos medibles $A \subset \mathbb{R}$. La norma $\|\cdot\|_{VT}$ se llama **norma en variación total** y se utilizará para caracterizar la convergencia hacia la distribución invariante. La prueba de este resultado va más allá del alcance de esta introducción pero nos permite resolver aproximadamente el problema de la simulación para la distribución μ. Por ejemplo, se puede considerar K realizaciones independientes $(X_{1,n})_{n\in\mathbb{N}}, \ldots, (X_{K,n})_{n\in\mathbb{N}}$ del muestreo de rebanada con $X_{1,0} = \cdots = X_{K,0} = x \in \mathbb{R}$ y tener una muestra de distribución cercana a μ con la iteración $n \in \mathbb{N}$ de cada una de estas sucesiones,

$$\forall A \subset \mathbb{R} \text{ medible},\ \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}_{X_{k,n}\in A} \simeq \mathbb{P}^x(X_n \in A) \simeq \mu(A)$$

para n y K bastante grandes. Otra propiedad útil es la convergencia de los promedios empíricos que generalizan la ley de los grandes números,

$$\frac{1}{n}\sum_{k=0}^{n} h(X_n) \xrightarrow[n\to+\infty]{c.s.} \int_{\mathbb{R}} h(x) f(x)\, dx \tag{2.2}$$

para cualquier función integrable $h : \mathbb{R} \to \mathbb{R}$.

Algoritmos como el muestreo de rebanada que producen cadenas de Markov cuya distribución invariante es la distribución a simular μ son el objetivo principal de este curso y se llaman **métodos de Monte Carlo con cadenas de Markov** (o **métodos MCMC** por **Monte Carlo Markov Chains**). La referencia al método de Monte Carlo se justifica por propiedades como (2.2).

Definición 2.2. Llamamos **métodos de Monte Carlo con cadenas de Markov** (o **métodos MCMC**) a todos algoritmos estocásticos de simulación de una distribución μ que producen una cadena de Markov con distribución invariante μ.

2.2 Algunas propiedades de las cadenas de Markov

El objetivo de esta sección es presentar importantes objetos probabilísticos, llamados **cadenas de Markov** en referencia al matemático ruso Andrey Markov (1856-1922) que los introdujo alrededor de 1906, así como algunas de sus propiedades que serán útiles para lo que sigue. Esta es solo una introducción a este vasto tema matemático y el lector interesado encontrará las pruebas de los resultados admitidos y más detalles en el libro de referencia [MT09].

2.2.1 Definiciones

En aras de la sencillez, esta presentación se limita al caso de variables aleatorias con valores en un conjunto E finito y dotado de la σ-álgebra de sus subconjuntos. De hecho, cualquier función $f : E \to \mathbb{R}$ es continua y por lo tanto medible. Por definición, una medida de probabilidad μ en $(E, \mathscr{P}(E))$ verifica $\mu(E) = 1$ y

$$\forall A \in \mathscr{P}(E),\ \mu(A) = \sum_{x \in A} \mu(x)$$

donde el valor de la probabilidad del punto $x \in E$ se expresa $\mu(x) = \mu(\{x\})$. Decimos que la medida de probabilidad μ **carga** el punto $x \in E$ cuando $\mu(x) > 0$.

Una función $f : E \to \mathbb{R}$ es **integrable** con respecto a μ si la suma

$$\sum_{x \in E} |f(x)|\mu(x)$$

converge, lo que siempre es cierto para un espacio finito. Definimos la integral de f con respecto a μ como la esperanza de $f(X)$ con $X \sim \mu$,

$$\mathbb{E}[f(X)] = \sum_{x \in E} f(x)\mu(x).$$

La distribución de una sucesión $(X_n)_{n \in \mathbb{N}}$ de variables aleatorias con valores en E se caracteriza por las distribuciones marginales de los vectores aleatorios $(X_0, \ldots, X_m)$ par cualquier $m \in \mathbb{N}$, *i.e.* por las probabilidades $\mathbb{P}(X_0 = x_0, \ldots, X_m = x_m)$ para cualquier $x_0, \ldots, x_m \in E$. Igualmente, es posible definir la distribución de la sucesión $(X_n)_{n \in \mathbb{N}}$ a través de la distribución de X_0 y de las probabilidades condicionadas $\mathbb{P}(X_{n+1} = x_{n+1} \mid X_0 = x_0, \ldots, X_n = x_n)$ para cualquier $n \in \mathbb{N}$ y $x_0, \ldots, x_{n+1} \in E$ porque

$$\mathbb{P}(X_0 = x_0, \ldots, X_m = x_m) = \mathbb{P}(X_0 = x_0) \times \prod_{n=0}^{m-1} \mathbb{P}(X_{n+1} = x_{n+1} \mid X_0 = x_0, \ldots, X_n = x_n).$$

Es común usar un vocabulario espacio-temporal para hablar de los elementos de una sucesión $(X_n)_{n \in \mathbb{N}}$ de variables aleatorias. Así, para cualquier $n \in \mathbb{N}$, la variable X_n representa la posición espacial en el instante n, las variables $X_0, \ldots, X_{n-1}$ representan su pasado y las

variables $X_{n+1}, X_{n+2}, \dots$ su futuro. Con este vocabulario, una **cadena de Markov** se presenta a menudo como una sucesión de variables aleatorias cuya pasado y futuro son independientes condicionalmente al presente, lo que se formaliza mediante la siguiente definición.

Definición 2.3. Una sucesión $(X_n)_{n\in\mathbb{N}}$ de variables aleatorias con valores en un conjunto E finito se llama **cadena de Markov de espacio de estado** E cuando, para cualquier $n \in \mathbb{N}$ y $x_0, \dots, x_{n+1} \in E$,

$$\mathbb{P}(X_{n+1} = x_{n+1} \mid X_0 = x_0, \dots, X_n = x_n) = \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n). \tag{2.3}$$

Además, se dice que la cadena es **homogénea** si la distribución de X_{n+1} condicional a X_n no depende de n,

$$\forall x, y \in E,\ \mathbb{P}(X_{n+1} = y \mid X_n = x) = \mathbb{P}(X_1 = y \mid X_0 = x).$$

En el marco de este curso, una cadena de Markov $(X_n)_{n\in\mathbb{N}}$ será a menudo homogénea. De hecho, podemos introducir una función $P : E \times E \to [0,1]$, denominada **kernel de transición** de la cadena, para codificar la probabilidad de pasar de un estado $x \in E$ a un estado $y \in E$,

$$P(x,y) = \mathbb{P}(X_1 = y \mid X_0 = x).$$

Por construcción, el kernel de transición verifica

$$\forall x \in E,\ \sum_{y\in E} P(x,y) = 1$$

y, si ν es la distribución de X_0, las probabilidades marginales de la cadena de Markov son dadas por

$$\forall m \in \mathbb{N},\ \forall x_0, \dots, x_m \in E,\ \mathbb{P}(X_0 = x_0, \dots, X_m = x_m) = \nu(x_0) \prod_{n=0}^{m-1} P(x_n, x_{n+1}).$$

Así, la distribución de la cadena de Markov es completamente caracterizada por la **distribución inicial** ν y el kernel de transición P.

EJEMPLO 2.4 (Variables independientes). Una sucesión $(X_n)_{n\in\mathbb{N}}$ de variables aleatorias independientes con misma distribución μ en E es un ejemplo elemental de cadena de Markov. En este caso, el kernel de transición ya no depende del estado de inicio,

$$\forall x, y \in E,\ P(x,y) = \mu(y).$$

■

EJEMPLO 2.5 (Camino aleatorio simple en el círculo). Sea un entero $k > 0$, consideramos el espacio $E = \mathbb{Z}/k\mathbb{Z}$, una variable aleatoria X_0 de distribución ν en E y una sucesión $(R_n)_{n\in\mathbb{N}}$ de variables aleatorias independientes de distribución de Rademacher $\mathscr{R}(1/2)$ y independientes de X_0. La sucesión $(X_n)_{n\in\mathbb{N}}$ de variables aleatorias dada por la recursión

$$\begin{aligned} \forall n \in \mathbb{N},\ X_{n+1} &= X_n + R_{n+1} \pmod{k} \\ &= X_0 + R_1 + \cdots + R_{n+1} \pmod{k} \end{aligned}$$

es una cadena de Markov. En efecto, para cualquier $n \in \mathbb{N}$, si tomamos estados $x_0, \dots, x_{n+1} \in E$ tales que $\mathbb{P}(X_0 = x_0, \dots, X_n = x_n) > 0$, entonces la independencia de las variables implica

$$\begin{aligned}\mathbb{P}(X_{n+1} = x_{n+1} \mid X_0 = x_0, \dots, X_n = x_n) &= \frac{\mathbb{P}(X_0 = x_0, \dots, X_n = x_n, X_{n+1} = x_{n+1})}{\mathbb{P}(X_0 = x_0, \dots, X_n = x_n)} \\ &= \frac{\mathbb{P}(R_{n+1} = x_{n+1} - x_n) \times \mathbb{P}(X_0 = x_0, \dots, X_n = x_n)}{\mathbb{P}(X_0 = x_0, \dots, X_n = x_n)} \\ &= \mathbb{P}(R_{n+1} = x_{n+1} - x_n) \\ &= \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n).\end{aligned}$$

Por lo tanto, el kernel de transición es dado por

$$\forall x, y \in \mathbb{Z}/k\mathbb{Z},\ P(x,y) = \begin{cases} 1/2 & \text{si } |x-y| = 1 \pmod{k}, \\ 0 & \text{si no.} \end{cases}$$

Cuando el espacio de estado es finito, el kernel de transición P se maneja como una **matriz de transición** de tamaño $k \times k$,

$$P = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \dots & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & \ddots & \vdots \\ 0 & \frac{1}{2} & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \frac{1}{2} \\ 0 & \dots & 0 & \frac{1}{2} & 0 \end{pmatrix}.$$

■

2.2.2 Distribuciones instantáneas

Es posible definir un kernel de transición independientemente del concepto de cadena de Markov.

Definición 2.6. Sea E un espacio finito, una función $P : E \times E \to [0,1]$ es un **kernel de transición** (o **matriz de transición**) si

$$\forall x \in E,\ \sum_{y \in E} P(x,y) = 1.$$

Es interesante notar que el conjunto de los kernel de transición definidos sobre un espacio E es convexo, *i.e.* si P y Q son dos kernel de transición sobre E, entonces, para cualquier $\alpha \in [0,1]$, la suma $\alpha P + (1-\alpha)Q$ también es un kernel de transición sobre E. Además, el producto PQ de P y Q es definido por

$$\forall x, y \in E,\ PQ(x,y) = \sum_{z \in E} P(x,z)Q(z,y),$$

y es un kernel de transición. El elemento neutro I para este producto es dado por el kernel trivial

$$\forall x, y \in E,\ I(x,y) = \begin{cases} 1 & \text{si } x = y, \\ 0 & \text{si no.} \end{cases} \tag{2.4}$$

En el caso de un espacio E finito en el que los kernel de transición se manipulan como matrices (véase el ejemplo 2.5), el producto PQ corresponde al producto matricial entre P y Q y I es la matriz identidad.

Considere una cadena de Markov $(X_n)_{n\in\mathbb{N}}$ cuyo kernel de transición es P sobre E. Definiendo $P^0 = I$ y $P^1 = P$, la probabilidad de pasar de un estado $x \in E$ a un estado $y \in E$ después de $n \in \mathbb{N}$ iteraciones de la cadena está dada por el kernel de transición P^n. En efecto, si $n \geqslant 2$,

$$\begin{aligned}
\mathbb{P}(X_n = y \mid X_0 = x) &= \sum_{x_1,\dots,x_{n-1}\in E} \mathbb{P}(X_1 = x_1, \dots, X_{n-1} = x_{n-1}, X_n = y \mid X_0 = x) \\
&= \sum_{x_1,\dots,x_{n-1}\in E} P(x,x_1)P(x_1,x_2)\dots P(x_{n-1},y) \\
&= \sum_{x_2,\dots,x_{n-1}\in E} P^2(x,x_2)P(x_2,x_3)\dots P(x_{n-1},y) \\
&= \dots \\
&= P^n(x,y)
\end{aligned}$$

Reencontramos por este cálculo las fórmulas bien conocidas en el marco matricial,

$$P^n = PP^{n-1} = P^{n-1}P.$$

Para expresar la distribución de la cadena de Markov $(X_n)_{n\in\mathbb{N}}$ tal que $X_0 = x \in E$, escribimos $\mathbb{P}^x$. Así, para cualquier $x, y \in E$ y $n \in \mathbb{N}$, $\mathbb{P}^x(X_n = y) = P^n(x,y)$. Más generalmente, si la distribución de X_0 es ν, la distribución de la cadena se expresa $\mathbb{P}^\nu$,

$$\forall y \in E,\ \mathbb{P}^\nu(X_n = y) = \sum_{x\in E} P^n(x,y)\nu(x).$$

Estas notaciones naturalmente se generalizan a las esperanzas, para cualquier estado $x \in E$, distribución ν sobre E y función integrable $f : E \to \mathbb{R}$,

$$\mathbb{E}^x[f(X_n)] = \mathbb{E}[f(X_n) \mid X_0 = x] = \sum_{y\in E} f(y)P^n(x,y)$$

y

$$\mathbb{E}^\nu[f(X_n)] = \sum_{x\in E} \mathbb{E}[f(X_n) \mid X_0 = x]\nu(x) = \sum_{x\in E}\sum_{y\in E} f(y)P^n(x,y)\nu(x).$$

2.2.3 Medidas de probabilidad invariantes

Sea una cadena de Markov $(X_n)_{n\in\mathbb{N}}$ homogénea de espacio de estado E finito y de kernel de transición P. Para cualquier $n \in \mathbb{N}$, si μ_n es la distribución de la variable X_n, sabemos

$$\forall y \in E,\ \mu_{n+1}(y) = \sum_{x\in E} P(x,y)\mu_n(x).$$

Las medidas de probabilidad invariantes por esta acción del kernel de transición desempeñarán un papel clave en el comportamiento asintótico de las cadenas de Markov como hemos visto con el ejemplo del muestreo de rebanada.

Definición 2.7. Una medida de probabilidad μ sobre E es **invariante** para la cadena de Markov $(X_n)_{n\in\mathbb{N}}$ de kernel de transición P si

$$\forall y \in E,\ \mu(y) = \sum_{x\in E} P(x,y)\mu(x).$$

Si la distribución inicial de X_0 es una probabilidad invariante μ, entonces, para cualquier $n \in \mathbb{N}$, la distribución de X_n también es $\mu_n = \mu$. En otras palabras, una medida de probabilidad invariante corresponde a un estado estacionario de la cadena de Markov.

EJEMPLO 2.8 (Variables independientes). Si $(X_n)_{n\in\mathbb{N}}$ es una sucesión de variables aleatorias independientes de misma distribución μ sobre E, entonces $(X_n)_{n\in\mathbb{N}}$ es una cadena de Markov y μ es la única medida de probabilidad invariante. ■

EJEMPLO 2.9 (Cadena con dos estados). Un kernel de transición P sobre el espacio binario $E = \{0,1\}$ se escribe

$$P(0,1) = 1 - P(0,0) = p_0 \quad \text{y} \quad P(1,0) = 1 - P(1,1) = p_1$$

donde $p_0, p_1 \in [0,1]$. Desde el punto de vista matricial, tenemos

$$P = \begin{pmatrix} 1-p_0 & p_0 \\ p_1 & 1-p_1 \end{pmatrix}.$$

Una medida de probabilidad μ sobre E es invariante para este kernel de transición si, y sólo si, $p_0\mu(0) = p_1\mu(1)$. Cuando $p_0 + p_1 > 0$, la única medida de probabilidad invariante está dada por

$$\mu(0) = \frac{p_1}{p_0+p_1} \quad \text{y} \quad \mu(1) = \frac{p_0}{p_0+p_1}.$$

Si $p_0 = p_1 = 0$, la cadena es constante y todas las medidas de probabilidad son invariantes. ■

Teorema 2.10. *Una cadena de Markov de espacio de estado E tiene al menos una medida de probabilidad invariante.*

Demostración. Sea $|E|$ la cardinalidad finita de E, consideramos una medida de probabilidad ν sobre E. Se puede definir una sucesión $(\nu_n)_{n\geqslant 1}$ de medidas de probabilidad por

$$\forall n \geqslant 1,\ \forall y \in E,\ \nu_n(y) = \frac{1}{n}\sum_{k=1}^{n}\sum_{x\in E} P^k(x,y)\nu(x).$$

Las valores de los vectores $(\nu_n(y))_{y\in E}$ pertenecen al espacio compacto $[0,1]^{|E|}$. Por lo tanto existe una subsucesión $(\nu_{n_k})_{k\geqslant 1}$ que converge hacia una medida μ sobre E,

$$\forall y \in E,\ \lim_{k\to+\infty} \nu_{n_k}(y) = \mu(y).$$

Dado que E es finito, la medida μ es una probabilidad y tenemos, para cualquier $n \geqslant 1$ y $z \in E$,

$$\begin{aligned}
\sum_{y\in E} P(y,z)\nu_n(y) &= \frac{1}{n}\sum_{k=1}^{n}\sum_{x\in E} P^{k+1}(x,z)\nu(x) \\
&= \nu_n(z) + \frac{1}{n}\sum_{k=1}^{n}\sum_{x\in E}\left(P^{k+1}(x,z) - P^k(x,z)\right)\nu(x) \\
&= \nu_n(z) + \frac{1}{n}\sum_{x\in E}\left(P^{n+1}(x,z) - P(x,z)\right).
\end{aligned}$$

Como $|P^{n+1}(x,z) - P(x,z)| \leqslant 1$, deducimos que μ es invariante porque

$$\forall z \in E,\ \sum_{y\in E} P(y,z)\mu(y) - \mu(z) = \lim_{k\to+\infty} \sum_{y\in E} P(y,z)\nu_{n_k}(y) - \nu_{n_k}(z) = 0.$$

□

El resultado anterior se limita a los espacios finitos y no se puede generalizar. También se debe tener en cuenta que este teorema no dice nada acerca de la unicidad de la medida de probabilidad invariante. Se necesita más estructura en la cadena de Markov para establecer tales resultados.

2.2.4 Irreducibilidad

En los ejemplos de la subsección anterior, hemos visto que la existencia de probabilidades invariantes para una cadena de Markov depende de su kernel de transición y que algunos kernel de transición tienen una única probabilidad invariante. Para estudiar este vínculo entre un kernel de transición y sus medidas de probabilidad invariantes, se debe describir cómo el kernel relaciona los varios estados que la cadena de Markov puede tomar. Esta estructura relacional tiene consecuencias en el soporte de una probabilidad invariante.

Definición 2.11. Sea P un kernel de transición sobre un espacio de estado E finito. Para cualquier $x, y \in E$, decimos que

- $x \to y$: x **comunica** con y si existe $n \in \mathbb{N}$ tal que $P^n(x,y) > 0$,
- $x \leftrightarrow y$: x y y **comunican** si $x \to y$ y $y \to x$.

La relación de comunicación $\leftrightarrow$ es una relación de equivalencia que reagrupa los elementos de E en clases de equivalencia disociadas llamadas **clases irreducibles**.

Definición 2.12. Sea P un kernel de transición sobre un espacio de estado E finito. Se dice que el kernel de transición P es **irreducible** si el espacio entero E es la única clase irreducible para la relación $\leftrightarrow$, *i.e.* si para cualquier estados x y y, existe $n \in \mathbb{N}$ tal que $P^n(x,y) > 0$. Por extensión, se dice que una cadena de Markov homogénea de kernel de transición irreducible es una **cadena de Markov irreducible**.

No todas las cadenas de Markov son irreducibles y es posible considerar clases irreducibles de naturalezas distintas en las que la cadena tiene un comportamiento particular. Por ejemplo, cuando una clase irreducible se reduce a un solo punto $\{x_0\}$, la cadena de Markov no puede escaparse y se dice que el estado x_0 es **absorbente**. Sin embargo, para las aplicaciones que nos interesan en este curso, limitaremos nuestro estudio a las cadenas irreducibles en lo que sigue.

EJEMPLO 2.13 (Variables independientes). Si $(X_n)_{n\in\mathbb{N}}$ es una sucesión de variables aleatorias independientes de misma distribución μ sobre un espacio E finito, sabemos que $(X_n)_{n\in\mathbb{N}}$ es una cadena de Markov. Además, si suponemos que μ carga todos los estados, entonces el kernel de transición P es irreducible porque, para cualquier $x, y \in E$, $P(x,y) = \mu(y) > 0$. En este ejemplo trivial, donde el kernel de transición y la distribución inicial se confunden, se

puede caracterizar la única probabilidad invariante μ. En efecto, para cualquier estado $x \in E$, el **tiempo de regreso** T_x^+ en x dado por

$$T_x^+ = \text{ínf}\{n \geqslant 1 \text{ tal que} X_n = x\},$$

tiene una distribución geométrica con parámetro $\mu(x) > 0$,

$$\forall n \geqslant 1,\ \mathbb{P}^x(T_x^+ = n) = (1-\mu(x))^{n-1}\mu(x).$$

En particular, tenemos $\mathbb{E}^x[T_x^+] = 1/\mu(x) < +\infty$ y así

$$\forall x \in E,\ \mu(x) = \frac{1}{\mathbb{E}^x[T_x^+]}.$$

■

La caracterización de la medida de probabilidad invariante en el ejemplo anterior se generaliza a cualquier cadena de Markov de espacio de estado finito bajo la hipótesis de irreducibilidad. Este teorema permite establecer un vínculo entre la probabilidad invariante y la frecuencia con la que la cadena visita los diferentes estados.

Teorema 2.14. *Cualquier cadena de Markov irreducible en un espacio de estado finito E verifica*

$$\forall x,y \in E,\ \mathbb{E}^x[T_y^+] < +\infty$$

y la única medida de probabilidad invariante está dada por

$$\forall x \in E,\ \mu(x) = \frac{1}{\mathbb{E}^x[T_x^+]}.$$

Demostración. Admitido. □

2.2.5 Reversibilidad

Por definición, encontrar una medida de probabilidad invariante para un kernel de transición P sobre un espacio E finito es como resolver el sistema de ecuaciones lineales

$$\forall y \in E,\ \mu(y) = \sum_{x \in E} P(x,y)\mu(x)$$

cuyo el número de incógnitas es proporcional a la cardinalidad de E. El teorema 2.14 caracteriza μ pero no da una solución analítica. En la práctica, el espacio de estado suele ser muy grande y es difícil determinar μ explícitamente. El concepto de **reversibilidad** representa una condición suficiente y fácilmente verificable para garantizar la existencia de una medida de probabilidad invariante.

Definición 2.15. Se dice que un kernel de transición P sobre un espacio E finito es **reversible** con respecto a una medida de probabilidad μ si verifica

$$\forall x,y \in E,\ \mu(x)P(x,y) = \mu(y)P(y,x).$$

Por extensión, se dice que una cadena de Markov homogénea de kernel de transición reversible con respecto a μ es una **cadena de Markov reversible**.

La reversibilidad caracteriza un equilibrio en el sentido de que una trayectoria de la cadena de Markov tiene la misma probabilidad que la trayectoria inversa. En otras palabras, si $(X_n)_{n\in\mathbb{N}}$ es una cadena de Markov reversible con respecto a una probabilidad μ sobre un espacio E finito, entonces, para cualquier $n \in \mathbb{N}$ y $x_0,\dots,x_n \in E$,

$$\mathbb{P}^\mu(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}^\mu(X_0 = x_n, X_1 = x_{n-1}, \dots, X_n = x_0).$$

En efecto, por definición de una cadena de Markov,

$$\begin{aligned}\mathbb{P}^\mu(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) &= \mu(x_0)P(x_0,x_1)P(x_1,x_2)\dots P(x_{n-1},x_n)\\ &= P(x_1,x_0)\mu(x_1)P(x_1,x_2)\dots P(x_{n-1},x_n)\\ &= \dots\\ &= P(x_1,x_0)P(x_2,x_1)\dots P(x_n,x_{n-1})\mu(x_n)\\ &= \mathbb{P}^\mu(X_0 = x_n, X_1 = x_{n-1}, \dots, X_n = x_0).\end{aligned}$$

EJEMPLO 2.16 (Camino aleatorio simple en el círculo). Sean un entero $k > 0$ y $p \in [0,1]$, consideramos el kernel de transición siguiente sobre el espacio $E = \mathbb{Z}/k\mathbb{Z}$,

$$\forall x,y \in \mathbb{Z}/k\mathbb{Z},\ P(x,y) = \begin{cases} p & \text{si } y = x+1 \pmod k,\\ 1-p & \text{si } y = x-1 \pmod k,\\ 0 & \text{si no.}\end{cases}$$

Cualquiera que sea p, este kernel de transición admite la medida de probabilidad uniforme $\mu(x) = 1/k$ como única probabilidad invariante,

$$\forall y \in \mathbb{Z}/k\mathbb{Z},\ \sum_{x\in\mathbb{Z}/k\mathbb{Z}} P(x,y)\mu(x) = \frac{1}{k}(p+1-p) = \frac{1}{k} = \mu(y).$$

Se nota que gracias al teorema 2.14, podemos deducir que $\mathbb{E}^x[T_x^+] = k$. La cadena es reversible con respecto a μ si

$$\forall x \in \mathbb{Z}/k\mathbb{Z},\ \mu(x)P(x,x+1) = \mu(x+1)P(x+1,x),$$

es decir, si $p = 1-p$, lo que solo es cierto en el caso del camino aleatorio simple con $p = 1/2$ (véase el ejemplo 2.5). En efecto, si $p \neq 1/2$, la cadena tenderá a girar en la misma dirección y una trayectoria en la dirección inversa tendrá una probabilidad menor. Si $p = 1/2$, el camino aleatorio es simétrico y las trayectorias en una u otra dirección son equiprobables. Este ejemplo muestra que existen probabilidades invariantes para las cuales el kernel de transición no es reversible. ■

Teorema 2.17. *Si un kernel de transición P sobre un espacio E finito es reversible con respecto a una medida de probabilidad μ, entonces μ es invariante.*

Demostración. Sea $y \in E$, como $P(y,\cdot)$ es una medida de probabilidad, la reversibilidad implica

$$\sum_{x\in E} P(x,y)\mu(x) = \sum_{x\in E} P(y,x)\mu(y) = \mu(y).$$

□

2.2.6 Convergencia

Como dicho en el ejemplo del muestreo de rebanada, esta introducción a las cadenas de Markov es motivada por propiedades de convergencia. En particular, la referencia al método de Monte Carlo en el nombre de los métodos MCMC es relacionada con la siguiente generalización de la ley de los grandes números para las cadenas de Markov.

Teorema 2.18 (Teorema ergódico)**.** *Sea $(X_n)_{n\in\mathbb{N}}$ una cadena de Markov irreducible sobre un espacio de estado E finitio con distribución inicial ν y kernel de transición P. Se nota μ su única probabilidad invariante y consideramos dos funciones $f : E \to \mathbb{R}$ y $g : E \times E \to \mathbb{R}$ tales que*

$$\sum_{x\in E} |f(x)|\mu(x) < +\infty \quad y \quad \sum_{x,y\in E} |g(x,y)|P(x,y)\mu(x) < +\infty.$$

Entonces, los promedios a lo largo de las trayectorias convergen casi seguramente,

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \xrightarrow[n\to+\infty]{c.s.} \sum_{x\in E} f(x)\mu(x)$$

y

$$\frac{1}{n}\sum_{k=1}^{n} g(X_{k-1},X_k) \xrightarrow[n\to+\infty]{c.s.} \sum_{x,y\in E} g(x,y)P(x,y)\mu(x).$$

Demostración. Admitido. □

Dado un estado $x \in E$, una consecuencia inmediata del teorema ergódico para la función indicatriz $f(y) = \mathbf{1}_{y=x}$ permite establecer un vínculo entre la probabilidad invariante μ y la frecuencia de las visitas del estado x,

$$\frac{1}{n}\sum_{k=0}^{n-1} \mathbf{1}_{X_k=x} \xrightarrow[n\to+\infty]{c.s.} \mu(x) = \frac{1}{\mathbb{E}^x[T_x^+]}. \tag{2.5}$$

Estos resultados son interesantes para obtener valores aproximados de integrales con respecto a μ que son difíciles de calcular explícitamente. Sin embargo, esto no es suficiente para resolver el problema de la simulación de la distribución μ para el cual necesitamos resultados de convergencia sobre la distribución de la variable X_n. Para ilustrar eso, tomemos el ejemplo 2.9 de la cadena con dos estados. Si $p_0, p_1 \in]0,1[$, hemos visto que la única probabilidad invariante está dada por

$$\mu(0) = \frac{p_1}{p_0+p_1} \quad \text{y} \quad \mu(1) = \frac{p_0}{p_0+p_1}.$$

La matriz de transición P se diagonaliza fácilmente

$$P = \begin{pmatrix} 1-p_0 & p_0 \\ p_1 & 1-p_1 \end{pmatrix} = \begin{pmatrix} 1 & -\mu(1) \\ 1 & \mu(0) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1-p_0-p_1 \end{pmatrix} \begin{pmatrix} \mu(0) & \mu(1) \\ -1 & 1 \end{pmatrix}.$$

Por lo tanto, las potencias de P se calculan y podemos deducir la convergencia siguiente,

$$P^n = \begin{pmatrix} 1 & -\mu(1) \\ 1 & \mu(0) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & (1-p_0-p_1)^n \end{pmatrix} \begin{pmatrix} \mu(0) & \mu(1) \\ -1 & 1 \end{pmatrix} \xrightarrow[n\to+\infty]{} \begin{pmatrix} \mu(0) & \mu(1) \\ \mu(0) & \mu(1) \end{pmatrix}.$$

En otras palabras, cuando n va al infinito, las probabilidades de transición convergen exponencialmente rápidamente hacia la medida de probabilidad invariante,

$$\forall x, y \in \{0,1\},\ \lim_{n\to+\infty} \mathbb{P}^x(X_n = y) = \mu(y).$$

Como en el teorema ergódico, el límite no depende de la distribución inicial. En tiempo largo, la cadena olvida la posición de la que proviene y, en el vocabulario de las cadenas de Markov, se habla de **pérdida de memoria**.

La convergencia de las probabilidades de transición hacia la probabilidad invariante μ es una propiedad general de las cadenas de Markov sobre un espacio finito. Sin embargo, para establecer resultados generales y cuantificar la velocidad de esta convergencia, es necesario protegerse contra ciertos casos patológicos y verificar algunas condiciones. De hecho, considerando la esperanza de (2.5), sabemos que

$$\forall x, y \in E,\ \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{P}^x(X_k = y) \xrightarrow[n\to+\infty]{} \mu(y)$$

pero eso no es suficiente para establecer que

$$\forall x, y \in E,\ \lim_{n\to+\infty} \mathbb{P}^x(X_n = y) = \mu(y). \tag{2.6}$$

Para convencerse de esto, podemos tomar el ejemplo 2.5 del camino aleatorio simple en el círculo con un número $k = 2\ell$ par de estados. Si este camino aleatorio empeza en 0, la distribución de X_{2n} solo carga los estados pares y la de X_{2n+1} solo carga los estados impares. Para obtener resultados de convergencia como (2.6), es necesario evitar los kernel de transición que tienen tales propiedades periódicas.

Definición 2.19. Se dice que una cadena de Markov $(X_n)_{n\in\mathbb{N}}$ irreducible de kernel de transición P sobre un espacio de estado E finito es **aperiódica** si, para cualquier $x, y \in E$, existe $n(x,y) \in \mathbb{N}$ tal que

$$\forall n \geqslant n(x,y),\ \mathbb{P}^x(X_n = y) = P^n(x,y) > 0.$$

Esta definición evita los problemas anteriores. Además, si una cadena de Markov irreducible de kernel de transición P no es aperiódica, es fácil cambiarla en una cadena de Markov aperiódica considerando una variante cuyo kernel de transición es dado por

$$Q = \frac{I+P}{2}$$

donde I es el kernel de transición trivial (2.4). En todo instante, esta variante se queda en el lugar con una probabilidad $1/2$ o se mueve a un nuevo estado según P con una probabilidad $1/2$. Es obvio que una probabilidad invariante para P también es invariante para Q. En la práctica, la irreducibilidad de una cadena de Markov ayuda para probar su aperiodicidad con la siguiente proposición.

Proposición 2.20. *Sea una cadena de Markov $(X_n)_{n\in\mathbb{N}}$ irreducible de kernel de transición P sobre un espacio de estado E finito. Si un estado $x \in E$ es aperiódico, es decir que existe un entero n_x tal que*

$$\forall n \geqslant n_x,\ \mathbb{P}^x(X_n = x) = P^n(x,x) > 0,$$

entonces la cadena es aperiódica.

Demostración. Sean $y, z \in E$, la irreducibilidad implica que existen dos enteros r y s tales que $P^r(y,x) > 0$ y $P^s(x,z) > 0$. Deducimos que, para cualquier $n \geqslant n_x$,

$$P^{r+n+s}(y,z) \geqslant P^r(y,x)P^n(x,x)P^s(x,z) > 0.$$

Por lo tanto, la cadena es aperiódica. □

Una consecuencia importante de la aperiodicidad está dada por el siguiente resultado que asegura la convergencia hacia la probabilidad invariante.

Teorema 2.21. *Si $(X_n)_{n\in\mathbb{N}}$ es una cadena de Markov irreducible y aperiódica con única probabilidad invariante μ sobre un espacio de estado E finito, entonces, para cualquier distribución inicial ν, la distribución de X_n converge hacia μ cuando n va al infinito,*

$$\forall x \in E,\ \lim_{n\to+\infty} \mathbb{P}^\nu(X_n = x) = \mu(x).$$

Demostración. Admitido. □

Así, para cualquier estado $x \in E$, se puede aproximar el valor de la probabilidad invariante $\mu(x)$ por $\mathbb{P}^\nu(X_n = x)$ cuando n es grande cualquiera que sea la distribución inicial ν. Esta probabilidad $\mathbb{P}^\nu(X_n = x)$ puede ser acercada por la ley de los grandes números considerando varias realizaciones independientes $(X_n^{(1)})_{n\in\mathbb{N}}, \ldots, (X_n^{(K)})_{n\in\mathbb{N}}$ de la cadena de Markov,

$$\frac{1}{K}\sum_{k=1}^{K} \mathbf{1}_{X_n^{(k)}=x} \xrightarrow[K\to+\infty]{c.s.} \mathbb{P}^\nu(X_n = x).$$

De hecho, hay dos enfoques para estimar la medida de probabilidad invariante μ :

- dado un estado x, tomar el medio empírico (2.5) del número de visitas de x a lo largo de una trayectoria por el teorema ergódico.
- dado un instante n, construir el histograma de los valores tomados en el momento n con varias realizaciones independientes de la cadena.

Gracias a la aperiodicidad, las probabilidades de transición convergen hacia la medida de probabilidad invariante. La manipulación de una cadena de Markov de probabilidad invariante μ será más fácil ya que esta convergencia será rápida. Una velocidad de convergencia corresponde a una cota superior por la diferencia entre la distribución de la iteración n de la cadena y la medida des probabilidad invariante. Hay varias maneras de cuantificar esta diferencia y en este curso nos interesa a la **distancia en variación total** introducida en (2.1).

Definición 2.22. Dadas dos medidas de probabilidad μ y ν sobre un espacio E medible, la **distancia en variación total** entre μ y ν es definida por

$$\|\mu - \nu\|_{VT} = \sup_{A \subset E} |\mu(A) - \nu(A)|$$

donde se toma el supremo sobre todos los subconjuntos medibles. Si el espacio E es finito, esta definición es equivalente a

$$\|\mu - \nu\|_{VT} = \frac{1}{2} \sum_{x \in E} |\mu(x) - \nu(x)|.$$

Desde el punto de vista de las cadenas de Markov, la distancia en variación total también es interesante porque verifica un principio de contracción.

Proposición 2.23. *Para cualquier kernel de transición P y cualquier medida de probabilidad μ sobre un espacio E finito, podemos definir la medida de probabilidad siguiente,*

$$\forall y \in E,\ \mu P(y) = \sum_{x \in E} P(x,y)\mu(x).$$

Sean dos medidas de probabilidad μ y ν sobre E, la distancia en variación total verifica

$$\|\mu P - \nu P\|_{VT} \leqslant \|\mu - \nu\|_{VT}.$$

Demostración. Este resultado es una consecuencia de la desigualdad triangular y de que $P(x,\cdot)$ sea una medida de probabilidad sobre E porque

$$\begin{aligned}
\|\mu P - \nu P\|_{VT} &= \frac{1}{2} \sum_{y \in E} |\mu P(y) - \nu P(y)| \\
&= \frac{1}{2} \sum_{y \in E} \left| \sum_{x \in E} P(x,y)(\mu(x) - \nu(x)) \right| \\
&\leqslant \frac{1}{2} \sum_{x \in E} \sum_{y \in E} P(x,y)|\mu(x) - \nu(x))| = \|\mu - \nu\|_{VT}.
\end{aligned}$$

□

En particular, si μ es una probabilidad invariante para un kernel de transición P sobre un espacio E finito, entonces, para cualquier medida de probabilidad ν sobre E, esta contracción implica

$$\forall n \in \mathbb{N},\ \|\nu P^n - \mu\|_{VT} \leqslant \|\nu - \mu\|_{VT}.$$

Si $(X_n)_{n \in \mathbb{N}}$ es una cadena de Markov de kernel de transición P y de distribución inicial ν sobre E, sabemos que

$$\forall n \in \mathbb{N},\ \forall y \in E,\ \mathbb{P}^{\nu}(X_n = y) = \nu P^n(y).$$

Por lo tanto, la distribución de la cadena de Markov no puede alejarse de una probabilidad invariante en términos de distancia en variación total. Al cuantificar esta diferencia entre la distribución de X_n y la probabilidad invariante, podemos establecer una velocidad de convergencia como veremos más adelante.

Figura 2.1: Andrey Markov (1856-1922)

3 — Algoritmo de Metrópolis-Hastings

En el marco de las investigaciones de las armas nucleares en Los Álamos a fines de los años 1940, Nicholas Metrópolis y Stanislaw Ulam trabajaron con la **distribución de Boltzmann**. Es una medida de probabilidad utilizada en física estadística para describir el estado de un sistema con respecto a su energía y su temperatura. El conjunto E de los estados posibles para el sistema es finito y, para cualquier estado $x \in E$, esta distribución da una probabilidad proporcional a $\exp(-\varepsilon_x/(k_B T))$ donde ε_x es la energía del estado x, k_B es la **constante de Boltzmann** y T es la temperatura. La constante de normalización de esta distribución está por tanto dada por

$$Z_T = \sum_{x \in E} \exp\left(-\frac{\varepsilon_x}{k_B T}\right).$$

Aunque el conjunto E es finito, su cardenal suele ser muy grande y esta suma puede ser costosa (o imposible) de calcular. Esta dificultad llevó a Metrópolis y Ulam a desarrollar un algoritmo específico para resolver el problema de la simulación para la distribución de Boltzmann. Metrópolis *et al.* describirán correctamente este algoritmo en 1953 en [MRR+53]. Este método será generalizado en 1970 por el estadístico Wilfred K. Hastings para simular realizaciones de una medida de probabilidad μ arbitraria en [Has70]. El algoritmo se llama **Metrópolis-Hastings** y es un método MCMC relativamente universal ya que no impone ninguna hipótesis sobre la distribución μ.

3.1 Construcción del algoritmo

Consideramos une medida de probabilidad μ sobre un espacio E finito para la que deseamos resolver el problema de la simulación. Al restringirnos al soporte de μ, podemos suponer que todos los estados son cargados, *i.e.* $\mu(x) > 0$ para cualquier $x \in E$. El principio general del algoritmo de Metrópolis-Hastings es construir una cadena de Markov $(X_n)_{n \in \mathbb{N}}$ irreducible y aperiódica con valores en E tal que μ sea su única medida de probabilidad invariante. Para esta cadena de Markov, el teorema 2.21 asegura que la distribución de los estados visitados converge hacia μ,

$$X_n \xrightarrow[n \to +\infty]{d} \mu. \tag{3.1}$$

Esta construcción se realiza mediante la definición de un kernel de transición P reversible con respecto a μ. El teorema 2.17 asegura que μ es invariante para este kernel y, dado que el espacio E es finito, la unicidad de la probabilidad invariante proviene de la irreducibilidad.

La idea de Metrópolis es introducir un kernel de transición auxiliar Q que sea irreducible e idealmente fácil de simular tal que

$$\forall x,y \in E,\ Q(x,y) > 0 \Rightarrow Q(y,x) > 0. \tag{3.2}$$

Sean $x,y \in E$, la probabilidad de pasar del estado x al estado y para la cadena de Markov auxiliar de distribución inicial μ y de kernel de transición Q está dada por $\mu(x)Q(x,y)$ y, recíprocamente, la transición de y a x admite una probabilidad $\mu(y)Q(y,x)$. El cociente entre estas probabilidades cuantifica qué transición es más probable a partir de una realización de la distribución μ. Así, una iteración del algoritmo consiste en favorecer aún más la transición del estado x al estado y que su probabilidad es alta y nos llevan a definir las **probabilidades de aceptación** usando estos cocientes,

$$\forall x,y \in E,\ \alpha(x,y) = \text{mín}\left\{1, \frac{\mu(y)Q(y,x)}{\mu(x)Q(x,y)}\right\}.$$

El cociente que define estas probabilidades de aceptación es muy importante porque permite considerar solo una función $\tilde{\mu} : E \to \mathbb{R}_+^*$ proporcional a μ para usar el algoritmo. En efecto, si existe $C > 0$ tal que $\tilde{\mu}(x) = C\mu(x)$ para cualquier $x \in E$, entonces

$$\forall x,y \in E,\ \frac{\mu(y)Q(y,x)}{\mu(x)Q(x,y)} = \frac{\tilde{\mu}(y)Q(y,x)}{\tilde{\mu}(x)Q(x,y)}.$$

Además, en el caso particular de un kernel de transición Q simétrico, *i.e.* para cualquier $x,y \in E$, $Q(x,y) = Q(y,x)$, las probabilidades de aceptación se simplifican y solo se queda el cociente $\mu(y)/\mu(x)$ para cuantificar de manera más intuitiva cuánto un estado candidato $y \in E$ es más probable que el estado corriente $x \in E$ para la distribución μ. Esta es la forma inicial del algoritmo según lo propuesto por Metrópolis.

ALGORITMO 3.1 – Metrópolis-Hastings

Inicialización :
- $\tilde{\mu}$: función proporcional a la distribución μ sobre E finito
- Q : kernel de transición irreducible sobre E tal que (3.2)
- $X_0 \in E$: estado inicial

En el paso $n \geqslant 1$:
- Generar $Y \sim (X_{n-1}, \cdot)$
- Calcular $\alpha = \alpha(X_{n-1}, Y)$
- Generar $U \sim \mathscr{U}([0,1])$
- Regla de rechazo :
  - Si $U \leqslant \alpha$, el candidato es aceptado : $X_n = Y$
  - Si $U > \alpha$, el candidato es rechazado : $X_n = X_{n-1}$

Devolver los valores $X_0, X_1, \ldots$

Para cualquier $n \in \mathbb{N}$, el estado X_{n+1} solo depende de X_n y este algoritmo produce una cadena de Markov de espacio de estado E finito. Esta cadena puede verse como un camino aleatorio en E que permanece en su lugar cuando el candidato Y es rechazado y que tiende a moverse hacia estados de alta probabilidad de lo contrario. En otras palabras, se alienta este camino aleatorio para que visite los áreas más probables de E mientras continúa explorando el resto de los estados con una probabilidad menor.

3.2 Kernel de Metrópolis-Hastings

El kernel de transición P de la cadena de Markov producido por el algoritmo de Metrópolis-Hastings, denominado **kernel de Metrópolis-Hastings**, se deduce de la definición,

$$\forall x,y \in E,\ P(x,y) = \begin{cases} Q(x,y)\alpha(x,y) = \min\left\{Q(x,y), \dfrac{\mu(y)Q(y,x)}{\mu(x)}\right\} & \text{si } x \neq y, \\ 1 - \displaystyle\sum_{z\neq x} Q(x,z)\alpha(x,z) & \text{si no.} \end{cases}$$

Proposición 3.2. *El kernel de Metrópolis-Hastings es reversible con respecto a la medida de probabilidad μ.*

Demostración. Sean dos estados $x,y \in E$ distintos, el resultado proviene de la definición,

$$\mu(x)P(x,y) = \min\{\mu(x)Q(x,y), \mu(y)Q(y,x)\} = \mu(y)P(y,x).$$

□

Por reversibilidad, la distribución μ es invariante para el kernel de Metrópolis-Hastings. Por lo tanto, la convergencia de las probabilidades de transición hacia la medida μ se deduce de la proposición siguiente.

Proposición 3.3. *Si existen dos estados $x_0, y_0 \in E$ distintos tales que*

$$Q(x_0,y_0) > 0 \quad y \quad \alpha(x_0,y_0) < 1$$

entonces el kernel de Metrópolis-Hastings es irreducible y aperiódico.

Demostración. Sean dos estados $x,y \in E$ distintos, la irreducibilidad de Q implica la existencia de un entero $m \in \mathbb{N}$ y de estados intermedios distintos $z_0,\ldots,z_m \in E$ tales que $z_0 = x$ y $z_m = y$ con $Q(z_{k-1},z_k) > 0$ para cualquier $k \in \{1,\ldots,m\}$. La hipótesis (3.2) da $Q(z_k,z_{k-1}) > 0$ y tenemos $\mu(z_k) > 0$ porque μ carga todos los estados. Por lo tanto, se obtiene que $\alpha(z_{k-1},z_k) > 0$ y

$$P^m(x,y) \geqslant \prod_{k=1}^{m} P(z_{k-1},z_k) = \prod_{k=1}^{m} Q(z_{k-1},z_k)\alpha(z_{k-1},z_k) > 0.$$

Así, el kernel de Metrópolis-Hastings es irreducible. Además, por hipótesis, sabemos que la probabilidad de permanecer en x_0 es tal que

$$\begin{aligned} P(x_0,x_0) &= 1 - \sum_{z\neq x_0} Q(x_0,z)\alpha(x_0,z) \\ &= 1 - Q(x_0,y_0)\alpha(x_0,y_0) - \sum_{z\neq x_0,y_0} Q(x_0,z)\alpha(x_0,z) \\ &> 1 - Q(x_0,y_0) - \sum_{z\neq x_0,y_0} Q(x_0,z) = 1 - \sum_{z\neq x_0} Q(x_0,z) = Q(x_0,x_0) \geqslant 0. \end{aligned}$$

La desigualdad estricta permite deducir $P(x_0,x_0) > 0$ y por tanto la aperiodicidad del estado x_0. Por irreducibilidad, se deduce que el kernel de Metrópolis-Hastings es aperiódico gracias a la proposición 2.20. □

3.3 Velocidad de convergencia

La convergencia de las probabilidades de transición hacia la medida de probabilidad invariante μ es particularmente interesante en la práctica si su velocidad es rápida. En efecto, para poder considerar las iteraciones del algoritmo de Metrópolis-Hastings como realizaciones aproximadas de μ gracias a (3.1), es necesario esperar hasta que la cadena haya evolucionado lo suficiente como para olvidar sus condiciones iniciales y haberse acercado a μ. Se habla de **tiempo de mezcla** para designar este fenómeno que será tanto más corto que la convergencia se produce rápidamente. Las primeras iteraciones de la cadena deben ignorarse y el tiempo de cálculo para producirlas se pierde necesariamente. El sacrificio de estos valores antes que la cadena sea mezclada suele denominarse un período de **burn-in**.

Dado que el cardenal $K = |E|$ del espacio de estado E es finito, podemos identificar E a $\{1,\dots,K\}$ y el kernel de Metrópolis-Hastings se maneja como una matriz de tamaño $K \times K$. Para establecer la velocidad de convergencia del algoritmo de Metrópolis-Hastings, vamos a estudiar el espectro de esta matriz, *i.e.* el conjunto de sus valores propios. La motivación para esto radica en el hecho de que, si se escribe $\mu = (\mu_1,\dots,\mu_K)' \in \mathbb{R}_+^*$ el vector de la probabilidades $\mu_k = \mu(k) > 0$ para $k \in E$, la propiedad de invarianza resulta en

$$P\mu = \mu,$$

es decir que μ es un vector proprio de P con valor propio 1. La reversibilidad de P con respecto a μ corresponde a

$$\forall k,\ell \in E,\ \mu_k P_{k\ell} = \mu_\ell P_{\ell k}$$

y, si D es la matriz diagonal dada por $D_{kk} = \sqrt{\mu_k}$ para cualquier $k \in E$, se obtiene que la matriz DPD^{-1} es simétrica porque

$$\forall k,\ell \in E,\ \left(DPD^{-1}\right)_{k\ell} = \frac{\sqrt{\mu_k}}{\sqrt{\mu_\ell}} P_{k\ell} = \frac{\sqrt{\mu_k}}{\sqrt{\mu_\ell}} \times \frac{\mu_\ell}{\mu_k} P_{\ell k} = \frac{\sqrt{\mu_\ell}}{\sqrt{\mu_k}} P_{\ell k} = \left(DPD^{-1}\right)_{\ell k}.$$

Lema 3.4. *Las matrices $I + DPD^{-1}$ y $I - DPD^{-1}$ son positivas.*

Demostración. Por definición de un kernel de transición, sabemos que

$$\forall k \in E,\ \sum_{\ell \in E} P_{k\ell} = 1.$$

Sea $x \in \mathbb{R}^K$, esto y la reversibilidad de P con respecto a μ implican

$$\begin{aligned}
x'(I + DPD^{-1})x &= \frac{1}{2} \sum_{k,\ell \in E} P_{k\ell} x_k^2 + P_{\ell k} x_\ell^2 + 2 \frac{\sqrt{\mu_k}}{\sqrt{\mu_\ell}} P_{k\ell} x_k x_\ell \\
&= \frac{1}{2} \sum_{k,\ell \in E} \mu_k P_{k\ell} \frac{x_k^2}{\mu_k} + \mu_\ell P_{\ell k} \frac{x_\ell^2}{\mu_\ell} + 2\mu_k P_{k\ell} \frac{x_k}{\sqrt{\mu_k}} \frac{x_\ell}{\sqrt{\mu_\ell}} \\
&= \frac{1}{2} \sum_{k,\ell \in E} \mu_k P_{k\ell} \frac{x_k^2}{\mu_k} + \mu_k P_{k\ell} \frac{x_\ell^2}{\mu_\ell} + 2\mu_k P_{k\ell} \frac{x_k}{\sqrt{\mu_k}} \frac{x_\ell}{\sqrt{\mu_\ell}} \\
&= \frac{1}{2} \sum_{k,\ell \in E} \mu_k P_{k\ell} \left(\frac{x_k}{\sqrt{\mu_k}} + \frac{x_\ell}{\sqrt{\mu_\ell}} \right)^2 \geqslant 0.
\end{aligned}$$

Un cálculo similar lleva a la positividad de $I - DPD^{-1}$ ya que

$$x'(I - DPD^{-1})x = \frac{1}{2}\sum_{k,\ell \in E} \mu_k P_{k\ell} \left(\frac{x_k}{\sqrt{\mu_k}} - \frac{x_\ell}{\sqrt{\mu_\ell}} \right)^2 \geqslant 0.$$

□

Ahora podemos establecer el teorema que hace posible vincular las propiedades de un kenel de transición reversible sobre un espacio de estado finito y la matriz asociada.

Teorema 3.5. *Sea P una matriz dada por un kenel de transición reversible con respecto a una medida de probabilidad μ sobre un espacio de estado E finito tal que $\mu(k) > 0$ para cualquier $k \in E$. Se verifica todas las propriedades siguientes :*

(I) El espectro de P es incluido en el intervalo $[-1, 1]$.

(II) El kenel de transición es irreducible si y solo si 1 es un valor proprio simple de P.

(III) Si existe un estado $k \in E$ tal que $P_{kk} > 0$ y que el kernel es irreducible, entonces -1 no es un valor propio de P.

Demostración. Se nota en primer lugar que P y DPD^{-1} tienen los mismos valores propios porque si existen $\lambda \in \mathbb{R}$ y $v \in \mathbb{R}^K \setminus \{0\}$ tales que $Pv = \lambda v$, entonces

$$DPD^{-1}(Dv) = DPv = \lambda(Dv). \tag{3.3}$$

Si $\lambda \in \mathbb{R}$ es un valor proprio de DPD^{-1}, entonces $1 + \lambda$ es un valor proprio de $I + DPD^{-1}$ y $1 - \lambda$ es un valor proprio de $I - DPD^{-1}$. Gracias al lema anterior, se deduce que $1 + \lambda \geqslant 0$ y $1 - \lambda \geqslant 0$, *i.e.* $\lambda \in [-1, 1]$.

Si un vector $v \in \mathbb{R}^K \setminus \{0\}$ es tal que $DPD^{-1}v = v$, entonces $(I - DPD^{-1})v = 0$ y se obtiene como en la demostración del lema anterior que

$$v'(I - DPD^{-1})v = \frac{1}{2}\sum_{k,\ell \in E} \mu_k P_{k\ell} \left(\frac{v_k}{\sqrt{\mu_k}} - \frac{v_\ell}{\sqrt{\mu_\ell}} \right)^2 = 0.$$

Por lo tanto, el espacio propio asociado al valor propio 1 es dado por los vectores v tales que

$$\forall k, \ell \in E,\ P_{k\ell} > 0 \Rightarrow \frac{v_k}{\sqrt{\mu_k}} = \frac{v_\ell}{\sqrt{\mu_\ell}}.$$

Si existe un camino de probabilidad estrictamente positiva entre dos estados $k, \ell \in E$ distintos, entonces todas los componentes asociados a los estados visitados por este camino verifican esta igualdad. En otras palabras, la dimensión del espacio propio asociado con el valor propio 1 (que no es vacío porque μ le pertenece) está dada por el número de clases irreducibles del kernel de transición.

Supongamos que -1 es un valor proprio de P, entonces existe un vector $v \in \mathbb{R}^K \setminus \{0\}$ tal que $DPD^{-1}v = -v$. Se obtiene como arriba que

$$\forall k, \ell \in E,\ P_{k\ell} > 0 \Rightarrow \frac{v_k}{\sqrt{\mu_k}} = -\frac{v_\ell}{\sqrt{\mu_\ell}}.$$

En particular, si existe un estado $k \in E$ tal que $P_{kk} > 0$, entonces $v_k = 0$ y todos los estados $\ell \in E$ tales que $P_{k\ell} > 0$ verifican $v_\ell = 0$. Por irreducibilidad, esto implica que el vector v es cero, lo que es una contradicción. Por lo tanto, para cualquier $k \in E$, el elemento diagonal P_{kk} es cero y la última propriedad es verificada. □

Para el kernel de Metrópolis-Hastings, la existencia de dos estados $k_0, \ell_0 \in E$ distintos tales que $Q(k_0, \ell_0) > 0$ y $\alpha(k_0, \ell_0) < 1$ implica la irreducibilidad del kernel y su aperiodicidad por la proposición 3.3. En la demostración de este resultado, hemos mostrado que $P_{k_0 k_0} > 0$ y por tanto que -1 no es un valor proprio por el teorema anterior. Ahora suponemos que somos en este marco, por lo que sabemos que :

- el valor proprio 1 es simple y el espacio proprio asociado es el espacio $\mathbb{R}\sqrt{\mu}$ generado por la única probabilidad invariante μ,
- todos los otros valores propios pertenecen a $]-1, 1[$.

Así, se puede ordenar los K valores proprios de la matriz P en orden decreciente con la multiplicidad,

$$\lambda_1 = 1 > \lambda_2 \geqslant \cdots \geqslant \lambda_K > -1.$$

Las matrices DP^nD^{-1} donde $n \geqslant 1$ son simétricas y conmutativas, por tanto son diagonalizables en la misma base ortonormal de los vectores propios. Se notan $\phi_1, \ldots, \phi_K \in \mathbb{R}^K$ estos vectores propios asociados a los valores propios $\lambda_1, \ldots, \lambda_K$ respectivamente. Para el caso particular del primer vector propio, obviamente consideramos

$$\phi_1 = (\sqrt{\mu_1}, \ldots, \sqrt{\mu_K})'.$$

Para cualquier $u, v \in \mathbb{R}^K$, se puede desarrolar el producto siguiente gracias a la base ortonormal de los ϕ_k,

$$u'DP^nD^{-1}v = \sum_{k=1}^{K} (u'\phi_k)(v'\phi_k)\lambda_k^n.$$

Sean $k, \ell \in E$, se toma $u = \mathbf{1}_k/\sqrt{\mu_k}$ y $v = \mathbf{1}_\ell\sqrt{\mu_\ell}$ donde $\mathbf{1}_k$ es el vector de $\mathbb{R}^K$ con el componente k igual a 1 y los otros a cero. Usando la igualdad anterior, obtenemos

$$(P^n)_{k\ell} = \mu_\ell + \frac{\sqrt{\mu_\ell}}{\sqrt{\mu_k}} \sum_{j=2}^{K} \phi_{j,k}\phi_{j,\ell}\lambda_j^n \tag{3.4}$$

y, ya que $|\lambda_j| < 1$ para cualquier $j \geqslant 2$, esto establece la convergencia de las probabilidades de transición hacia la distribución μ a velocidad exponencial cuando el número de iteraciones n tiende al infinito. Esta fórmula no es fácil de usar en la práctica y el siguiente teorema da una forma más simple.

Teorema 3.6. *Sea P una matriz dada por un kernel de transición irreducible, aperiódico y reversible con respecto a una medida de probabilidad μ sobre un espacio de estado E finito tal que $\mu(k) > 0$ para cualquier $k \in E$. Si α es el mayor valor propio de P estrictamente menos de 1, entonces, para cualquier entero $n \geqslant 1$,*

$$\forall k \in E,\ \sum_{\ell \in E} \mu_\ell \left(\frac{(P^n)_{k\ell}}{\mu_\ell} - 1 \right)^2 \leqslant \frac{\alpha^{2n}}{\mu_k}.$$

Demostración. Sean $k, \ell \in E$, al reorganizar los términos de (3.4), tenemos

$$\sqrt{\mu_\ell}\left(\frac{(P^n)_{k\ell}}{\mu_\ell} - 1\right) = \frac{1}{\sqrt{\mu_k}} \sum_{j=2}^{K} \phi_{j,k}\phi_{j,\ell}\lambda_j^n.$$

Levantemos esta igualdad al cuadrado y sumemos sobre todos los estados, la ortonormalidad de los vectores $\phi_1, \ldots, \phi_K$ da

$$\begin{aligned}\sum_{\ell\in E} \mu_\ell \left(\frac{(P^n)_{k\ell}}{\mu_\ell} - 1\right)^2 &= \frac{1}{\mu_k} \sum_{\ell\in E} \left(\sum_{j=2}^{K} \phi_{j,k}\phi_{j,\ell}\lambda_j^n\right)^2 \\ &= \frac{1}{\mu_k} \sum_{j=2}^{K}\sum_{j'=2}^{K} \phi_{j,k}\phi_{j',k}\lambda_j^n\lambda_{j'}^n \sum_{\ell\in E} \phi_{j,\ell}\phi_{j',\ell} \\ &= \frac{1}{\mu_k} \sum_{j=2}^{K} \phi_{j,k}^2\lambda_j^{2n}.\end{aligned}$$

Por hipótesis, $\lambda_j^{2n} \leqslant \alpha^{2n}$ para cualquier $j \in \{2, \ldots, K\}$ y obtenemos

$$\sum_{\ell\in E} \mu_\ell \left(\frac{(P^n)_{k\ell}}{\mu_\ell} - 1\right)^2 \leqslant \frac{\alpha^{2n}}{\mu_k} \sum_{j=2}^{K} \phi_{j,k}^2 \leqslant \frac{\alpha^{2n}}{\mu_k}$$

donde la última desigualdad se deduce de la ortogonalidad de la matriz para pasar de la base canónica a la base ortonormal de los ϕ_j porque, por transposición, las líneas de esta matriz también son de norma unitaria. □

Este teorema da la velocidad de convergencia exponencial de las probabilidades de transición $P^n(k,\cdot)$ hacia la distribución μ para la **distancia** χ^2 definida para cualquier probabilidades ν y μ sobre E, con μ estrictamente positiva, por

$$\chi^2(\nu,\mu) = \sum_{\ell\in E} \mu_\ell \left(\frac{\nu_\ell}{\mu_\ell} - 1\right)^2 = \sum_{\ell\in E} \frac{(\nu_\ell - \mu_\ell)^2}{\mu_\ell}.$$

Esta cantidad no es una distancia matemática porque no es simétrica. Sin embargo, permite establecer una cota superior en la distancia en variación total por la desigualdad de Cauchy-Schwarz,

$$\|\nu - \mu\|_{VT} = \frac{1}{2}\sum_{\ell\in E} |\nu_\ell - \mu_\ell| = \frac{1}{2}\sum_{\ell\in E} \frac{|\nu_\ell - \mu_\ell|}{\sqrt{\mu_\ell}}\sqrt{\mu_\ell} \leqslant \frac{1}{2}\sqrt{\chi^2(\nu,\mu)}.$$

Así, se deduce una velocidad de convergencia exponencial hacia la distribución μ para la distancia en variación total del algoritmo de Metrópolis-Hastings,

$$\forall k \in E,\ \|P^n(k,\cdot) - \mu\|_{VT} \leqslant \frac{\alpha^n}{2\sqrt{\mu(k)}}. \tag{3.5}$$

El valor $1-\alpha$ se llama el **agujero espectral** y la velocidad de convergencia establecida es tanto mejor como esta cantidad es grande. La elección del kernel auxiliar Q es crucial en la práctica porque influye los valores propios de la matriz P y, por lo tanto, el valor del agujero espectral.

3.4 Aplicaciones

3.4.1 Modelo de Ising

El **modelo de Ising** es es un modelo simple de física estadística para describir las interacciones locales entre partículas llamadas **espines** que solo pueden tomar dos estados. En cada nodo k de un grafo Λ, hay un espín $s_k \in \{+1,-1\}$ y una configuración de todos los espines se expresa $S = \{s_k\}_{k\in\Lambda}$. Sea S tal configuración, podemos definir una energía utilizando las interacciones locales entre cada espín y sus vecinos,

$$V(S) = \sum_{k\sim\ell} s_k s_\ell$$

donde la suma se refiere a todos los pares de nodos $(k,\ell) \in \Lambda^2$ alejados de distancia 1 en el grafo. Para tener en cuenta las fluctuaciones térmicas, se introduce un parámetro de temperatura $T > 0$ y definimos la medida de probabilidad μ sobre el espacio E de las configuraciones por

$$\forall S \in E,\ \mu(S) = \frac{1}{Z_T}\exp(-V(S)/T).$$

Como la distribución de Boltzmann introducida al principio de este capítulo, esta medida de probabilidad es un caso especial de las **medidas de Gibbs** que nos interesarán en el capítulo siguiente. En la práctica, para tener la probabilidad $\mu(S)$ de una configuración $S \in E$, la constante de normalización Z_T debe ser calculada y esto equivale a calcular los valores de la función de energía V para todas las configuraciones cuyo número a menudo es demasiado grande. El algoritmo de Metrópolis-Hastings brinda un método de simulación de la distribución μ sin tener que estimar esta constante Z_T.

Para un grafo regular bidimensional $\Lambda = \{1,\dots,L\}\times\{1,\dots,L\}$ con $L = 128$, el espacio de las configuraciones $E = \{+1,-1\}^\Lambda$ tiene una cardinalidad $2^{128\times128} \simeq 10^{4932}$. Por lo tanto, es imposible enumerar todas las configuraciones para calcular la constante Z_T y utilizaremos el algoritmo de Metrópolis-Hastings para producir realizaciones aproximadas de la distribución μ. Para ello, introducimos la transformación elemental del nodo $k \in \Lambda$ de una configuración $S \in E$ que solo cambia el signo del espín s_k,

$$\forall \ell \in \Lambda,\ S^{(k)}_\ell = \begin{cases} -s_k & \text{si } \ell = k, \\ s_\ell & \text{si } \ell \neq k. \end{cases}$$

El kernel de transición auxiliar Q que utilizamos es el sorteo de un nodo k uniforme en Λ y esta transformación elemental,

$$\forall S \in E,\ \forall k \in \Lambda,\ Q(S,S^{(k)}) = \frac{1}{|\Lambda|}.$$

Este kernel verifica (3.2) y es simétrico. Desaparece por tanto en las probabilidades de aceptación y obtenemos

$$\begin{aligned} \forall S \in E,\ \forall k \in \Lambda,\ \alpha(S,S^{(k)}) &= \text{mín}\left\{1, \exp\left(-\frac{V(S^{(k)}) - V(S)}{T}\right)\right\} \\ &= \text{mín}\left\{1, \exp\left(-\frac{2s_k}{T}\sum_{\ell\sim k} s_\ell\right)\right\}. \end{aligned}$$

La figura 3.1 es un ejemplo de realización de μ producida por este método con $L = 128$ y $T = 1{,}0$.

ALGORITMO 3.7 – Modelo de Ising en un grafo regular bidimensional

Inicialización :

- $L > 0$: tamaño del grafo
- $T > 0$: temperatura del sistema
- $S_0 \in E$: configuración inicial

En el paso $n \geqslant 1$:

Sortear k uniformemente en Λ

Calcular $\alpha = \alpha(S, S^{(k)})$

Generar $U \sim \mathscr{U}([0,1])$

Regla de rechazo :

- Si $U \leqslant \alpha$, la configuración es aceptada : $S_n = S_{n-1}^{(k)}$
- Si $U > \alpha$, la configuración es rechazada : $S_n = S_{n-1}$

Devolver las configuraciones $S_0, S_1, \ldots$

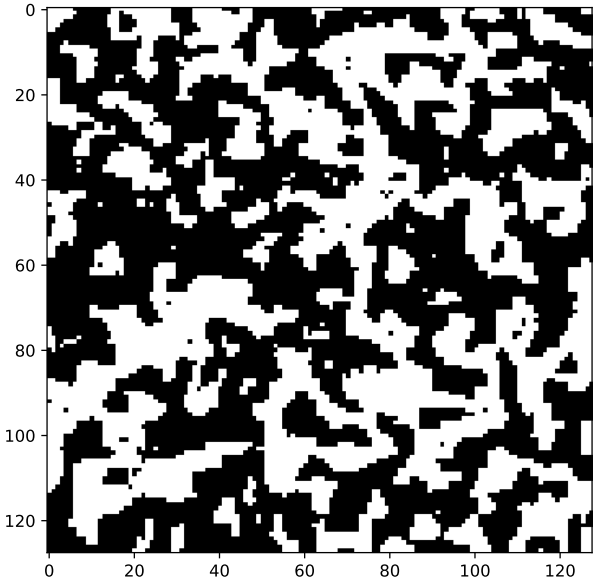

Figura 3.1: Una realización del modelo de Ising producida por el algoritmo de Metrópolis-Hastings en un grafo regular bidimensional de tamaño $L = 128$ para una temperatura $T = 1{,}0$.

3.4.2 Modelo probit

El **modelo probit** es un modelo de regresión estadística en el que la variable explicada es binaria, *i.e.* solo puede tomar dos valores. Para ilustrar esto, consideramos datos de un estudio médico de partos por cesárea. Las variables explicativas también son binarias en este ejemplo :

- x^1 es 1 si la cesárea fue planeada, 0 si no.
- x^2 es 1 si hay factores de riesgo para el parto, 0 si no.
- x^3 es 1 si se han administrado antibióticos, 0 si no.

El vector $x = (x^1, x^2, x^3)' \in \{0,1\}^3$ puede tomar 8 configuraciones distintas. La variable explicada y es 1 cuando se ha observado una infección y la tabla siguiente resume el número total n_x de observaciones y el número Y_x de veces que y es 1 para cada configuración $x \in \{0,1\}^3$.

x^1	x^2	x^3	Y_x	n_x
0	0	0	8	40
0	0	1	0	2
0	1	0	28	58
0	1	1	1	18
1	0	0	0	9
1	0	1	0	0
1	1	0	23	26
1	1	1	11	98

Para medir el riesgo de infección de acuerdo con una configuración, introducimos las probabilidades

$$\forall x \in \{0,1\}^3,\ p_x = \mathbb{P}(y = 1 \mid x),$$

y la distribución de Y_x es por tanto una distribución binomial $\mathscr{B}(n_x, p_x)$. Para representar las probabilidades, el modelo probit utiliza la función de distribución φ de la distribución normal estándar $\mathscr{N}(0,1)$ y coeficientes $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)' \in \mathbb{R}^3$ para definir

$$\forall x \in \{0,1\}^3,\ p_x = \varphi(X_x'\beta),$$

donde $X_x = (1, x^1, x^2, x^3)'$. La verosimilitud de este modelo está dada por

$$\ell(\beta \mid y) \propto \prod_{x \in \{0,1\}^3} \varphi(X_x'\beta)^{Y_x} (1 - \varphi(X_x'\beta))^{n_x - Y_x}.$$

Consideramos un enfoque bayesiano para estimar los coeficientes β para los cuales tomamos un prior gaussiano $\mathscr{N}(0, I/\lambda)$ donde I es la matriz identidad y $\lambda > 0$. Una función proporcional a la densidad $f(\beta \mid y)$ de la distribución a posteriori se deduce fácilmente,

$$f(\beta \mid y) \propto \exp\left(-\frac{\lambda \|\beta\|^2}{2}\right) \times \prod_{x \in \{0,1\}^3} \varphi(X_x'\beta)^{Y_x} (1 - \varphi(X_x'\beta))^{n_x - Y_x}.$$

El estimador bayesiano $\hat{\beta}$ para la función de pérdida cuadrática es dado por la esperanza de esta distribución,

$$\hat{\beta} = \int_{\mathbb{R}^4} \beta f(\beta \mid y)\, d\beta.$$

No hay fórmula explícita de $\hat{\beta}$ y vamos a acercar esta integral gracias al teorema ergódico 2.18 y al algoritmo de Metrópolis-Hastings.

La diferencia fundamental con respecto a lo que vimos anteriormente es que la distribución de densidad $f(\beta \mid y)$ no toma sus valores en un espacio finito pero en $\mathbb{R}^4$. La teoría de las cadenas de Markov se extiende naturalmente a los espacios de estado continuo, pero va más allá del alcance de este curso. El principio general es manejar densidades en lugar de las probabilidades puntuales y la definición 2.6 de un kernel de transición Q se generaliza como una función medible con valores en $\mathbb{R}_+$ tal que

$$\forall \beta \in \mathbb{R}^4,\ \int_{\mathbb{R}^4} Q(\beta, \beta')\, d\beta' = 1.$$

Para nuestro ejemplo, hay que definir un algoritmo de Metrópolis-Hastings que admite la distribución de densidad $f(\beta \mid y)$ como su única medida de probabilidad invariante. La irreducibilidad del kernel de transición auxiliar Q se generaliza al caso de un espacio continuo pero no daremos una definición propia. En la práctica, un kernel gaussiano es suficiente para explorar el espacio de estado de manera adecuada y tomaremos Q tal que la distribución de transición de un estado $\beta \in \mathbb{R}^4$ a un estado $\beta' \in \mathbb{R}^4$ está dada por

$$\beta' \mid \beta \sim \mathcal{N}_4(\beta, \sigma^2 I)$$

donde $\sigma^2 > 0$. En particular, este kernel es simétrico, *i.e.* $Q(\beta, \beta') = Q(\beta', \beta)$ para cualquier $\beta, \beta' \in \mathbb{R}^4$, y las probabilidades de aceptación se simplifican como antes. Así obtenemos el algoritmo siguiente.

ALGORITMO 3.8 – Metrópolis-Hastings con kernel gaussiano

Inicialización :

- parámetros $\lambda > 0$ y $\sigma^2 > 0$
- $\beta^{(0)} \in \mathbb{R}^4$: estado inicial

En el paso $n \geqslant 1$:

Generar $\beta \sim \mathcal{N}_4(\beta^{(n-1)}, \sigma^2 I)$

Calcular la probabilidad de aceptación dada por

$$\alpha = \text{mín}\left\{1, \frac{e^{-\lambda\|\beta\|^2/2}}{e^{-\lambda\|\beta^{(n-1)}\|^2/2}} \prod_{x\in\{0,1\}^3} \frac{\varphi(X_x'\beta)^{Y_x}(1-\varphi(X_x'\beta))^{n_x-Y_x}}{\varphi(X_x'\beta^{(n-1)})^{Y_x}(1-\varphi(X_x'\beta^{(n-1)}))^{n_x-Y_x}}\right\}$$

Generar $U \sim \mathcal{U}([0,1])$

Regla de rechazo :

- Si $U \leqslant \alpha$, el estado es aceptado : $\beta^{(n)} = \beta$
- Si $U > \alpha$, el estado es rechazado : $\beta^{(n)} = \beta^{(n-1)}$

Devolver los vectores $\beta^{(0)}, \beta^{(1)}, \ldots$

La sucesión $(\beta^{(n)})_{n\in\mathbb{N}}$ producida por este algoritmo es una cadena de Markov y es posible demostrar que verifica un teorema ergódico que permite acercar $\hat{\beta}$ por el promedio a lo largo de una trayectoria,

$$\forall n \geqslant 1,\ \hat{\beta}^{(n)} = \frac{1}{n}\sum_{k=0}^{n-1} \beta^{(k)} \xrightarrow[n\to+\infty]{c.s.} \hat{\beta}.$$

La figura 3.2 ilustra esta convergencia de $\hat{\beta}^{(n)}$ hacia el estimador bayesiano $\hat{\beta}$ de los coeficientes del modelo probit para $\lambda = 10$ y $\sigma^2 = 0{,}08$.

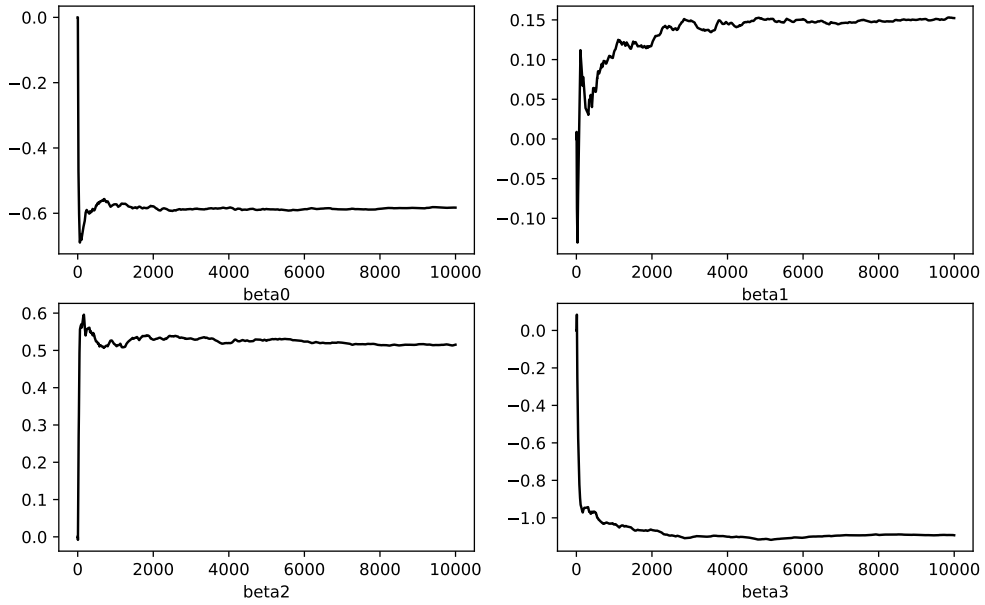

Figura 3.2: Estimación de los coeficientes del modelo probit por el promedio empírico de una trayectoria del algoritmo Metrópolis-Hastings.

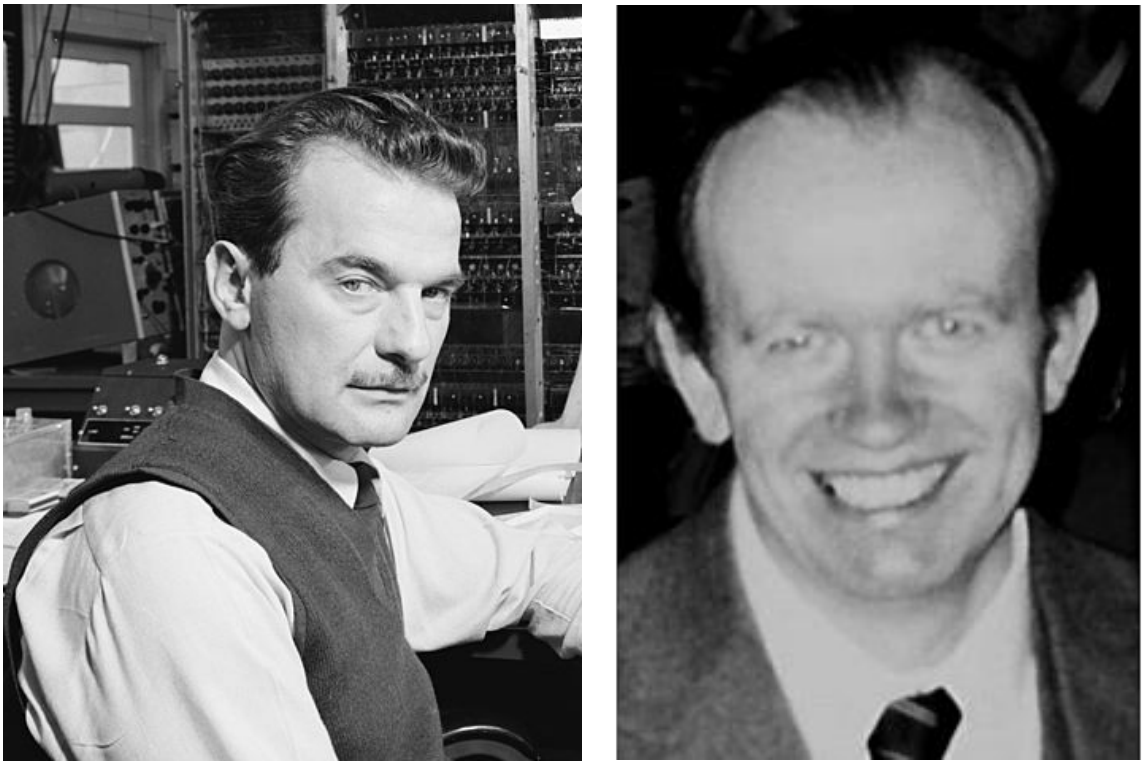

Figura 3.3: Nicholas Metrópolis (1915-1999) y Wilfred Keith Hastings (1930-2016).

4 — Algoritmo de recocido simulado

El **recocido simulado** es un método estocástico para tratar problemas de optimización. Este enfoque es particularmente útil en casos difíciles donde una búsqueda sistemática es imposible dado el tamaño del espacio a explorar o cuando la función a minimizar, por ejemplo, tiene un gran número de mínimos locales que queremos evitar para encontrar un mínimo global. El **recocido simulado** fue introducido independientemente en 1983 por Kirkpatrick *et al.* en [KGJV83] y en 1985 por Černý en [Če85]. Discutiremos su uso para problemas de **optimización combinatoria** que consisten en la minimización de una función f con valores reales definida sobre un espacio E finito pero de tamaño potencialmente muy grande.

4.1 Medida de Gibbs

Las **medidas de Gibbs** son medidas de probabilidad que aparecen en varios problemas en la teoría de la probabilidad o en la física estadística. El nombre es una referencia al físico-químico estadounidense Josiah Willard Gibbs. Una motivación importante para considerar tal clase de medidas de probabilidad es el principio de la física estadística que considera que un sistema aislado alcanza un estado de equilibrio cuando su **entropía** es máxima. La definición siguiente establece el vínculo entre este punto de vista físico y el enfoque probabilístico.

Definición 4.1. Consideramos un espacio E finito y una medida de probabilidad π sobre E. La **entropía** de la distribución π es la cantidad

$$H(\pi) = -\sum_{k\in E} \pi_k \ln(\pi_k)$$

donde $\pi_k = \pi(k)$ para cualquier $k \in E$.

La entropía $H(\pi)$ caracteriza el nivel de imprevisibilidad de la distribución π. En efecto, es mínima y nula para una probabilidad que carga solo un estado. Por otro lado, sin información adicional, la distribución que maximiza la entropía es la distribución uniforme sobre E. Para proporcionar información, es posible buscar la distribución que maximice la entropía tal que la integral con respecto a π de una función $f : E \to \mathbb{R}$ sea fija,

$$\sum_{k\in E} f(k)\pi_k.$$

Por los multiplicadores de Lagrange, se obtiene fácilmente que la distribución π que tiene la mayor entropía bajo esta restricción es de la forma

$$\forall k \in E,\ \pi_k = \alpha \exp(\beta f(k))$$

donde $\alpha, \beta \in \mathbb{R}$ son dos constantes. En física, el opuesto de β se entiende como el inverso de una temperatura $T > 0$ y se tiene la definición de la medida de Gibbs de **función de energía** f a la temperatura T,

$$\forall k \in E,\ \pi_k^T = \frac{1}{Z_T} \exp(-f(k)/T)$$

donde la constante de normalización se llama **función de partición** y está dada por

$$Z_T = \sum_{\ell \in E} \exp(-f(\ell)/T).$$

Una propiedad de las medidas de Gibbs que nos interesa es que a bajas temperaturas se concentran en estados que minimizan la función f.

Proposición 4.2. *Sea una función real f sobre un espacio E finito, el conjunto de los minimizadores de f es definido por*

$$E_{\text{mín}} = \{k \in E \text{ tal que } \forall \ell \in E,\ f(k) \leqslant f(\ell)\}.$$

El cardenal de $E_{\text{mín}}$ se expresa $|E_{\text{mín}}|$ y π^T es la medida de Gibbs de función de energía f a la temperatura $T > 0$. Entonces,

$$\forall k \in E,\ \lim_{T \to 0^+} \pi_k^T = \begin{cases} \dfrac{1}{|E_{\text{mín}}|} & \text{si } k \in E_{\text{mín}}, \\ 0 & \text{si no.} \end{cases}$$

Además, cuando T tiende a infinito, la medida de Gibbs π^T converge hacia la distribución uniforme en E.

Demostración. Por definición de $E_{\text{mín}}$, la función de partición se escribe

$$Z_T = |E_{\text{mín}}| e^{-m/T} + \sum_{\ell \in E \setminus E_{\text{mín}}} e^{-f(\ell)/T} = e^{-m/T} \left(|E_{\text{mín}}| + \sum_{\ell \in E \setminus E_{\text{mín}}} e^{-(f(\ell)-m)/T} \right)$$

donde $m = \min_{k \in E} f(k)$. Si $\ell \in E \setminus E_{\text{mín}}$, entonces $f(\ell) > m$ y sabemos que

$$e^{-(f(\ell)-m)/T} \xrightarrow[T \to 0^+]{} 0.$$

Sea $k \in E$, deducimos el valor límite de la probabilidad π_k^T a baja temperatura,

- si $k \in E_{\text{mín}}$,

$$\pi_k^T = \frac{e^{-m/T}}{Z_T} = \frac{1}{|E_{\text{mín}}| + \sum_{\ell \in E \setminus E_{\text{mín}}} e^{-(f(\ell)-m)/T}} \xrightarrow[T \to 0^+]{} \frac{1}{|E_{\text{mín}}|}.$$

- si $k \notin E_{\text{mín}}$,

$$\pi_k^T = \frac{e^{-m/T} e^{-(f(k)-m)/T}}{Z_T} = \frac{e^{-(f(k)-m)/T}}{|E_{\text{mín}}| + \sum_{\ell \in E \setminus E_{\text{mín}}} e^{-(f(\ell)-m)/T}} \xrightarrow[T \to 0^+]{} 0.$$

A alta temperatura, la convergencia hacia la distribución uniforme en E se deduce de la definición ya que, para cualquier $k \in E$, $\exp(-f(k)/T)$ tiende a 1 cuando $T \to +\infty$. □

La conclusión de la proposición anterior es que a baja temperatura la medida de Gibbs de función de energía f está cerca de la distribución uniforme en $E_{\text{mín}}$. Por lo tanto, una idea natural para minimizar f es simular la distribución π^T para una temperatura suficientemente baja y así obtener realizaciones cercanas a un minimizador de f. Gracias al algoritmo de Metrópolis-Hastings, sabemos que no es necesario conocer el valor de Z_T para simular π^T y eso es la idea detrás del recocido simulado.

Para garantizar la irreducibilidad, suponemos que E tiene una estructura de grafo conexo regular, *i.e.* todos los estados tienen el mismo número de vecinos, lo que suele ser cierto en la práctica. Podemos así considerar el kernel de transición Q del camino aleatorio simétrico sobre E y las probabilidades de aceptación $\alpha(k,\ell) = \text{mín}\{1, \pi_\ell^T / \pi_k^T\}$ para cualquier $k, \ell \in E$. Estas probabilidades simplemente se reescriben a partir de la definición de la distribución π^T,

$$\forall k, \ell \in E,\ \alpha(k,\ell) = \begin{cases} 1 & \text{si } f(\ell) \leqslant f(k), \\ e^{-(f(\ell)-f(k))/T} & \text{si } f(\ell) > f(k). \end{cases}$$

ALGORITMO 4.3 – Recocido simulado simple

Inicialización :
- f : función real a minimizar sobre un espacio E finito
- $T > 0$: temperatura del sistema
- $X_0 \in E$: estado inicial

En el paso $n \geqslant 1$:
- Sortear X' uniformemente en los vecinos de X_{n-1}
- Si $f(X') \leqslant f(X_{n-1})$:
  - Definir $X_n = X'$
- Si no :
  - Calcular $\alpha = e^{-(f(X')-f(X_{n-1}))/T}$
  - Generar $U \sim \mathscr{U}([0,1])$
  - Regla de rechazo :
    - Si $U \leqslant \alpha$, el estado es aceptado : $X_n = X'$
    - Si $U > \alpha$, el estado es rechazado : $X_n = X_{n-1}$

Devolver los estados $X_0, X_1, \ldots$

Este algoritmo puede entenderse como un camino aleatorio que seguramente se va a los valores bajos de la función f y puede también tomar pasos en la dirección equivocada con una pequeña probabilidad. Sin esta capacidad de pasar a estados donde la función f toma un valor mayor, el algoritmo puede quedar bloqueado en mínimos locales, *i.e.* estados $k \in E$ tales que $f(k) < f(\ell)$ para cualquier estado ℓ vecino de k. El problema con esta versión del recocido simulado es que, a temperatura demasiado baja, la probabilidad de elevarse es muy baja y el algoritmo puede permanecer bloqueado durante mucho tiempo alrededor de los mínimos locales.

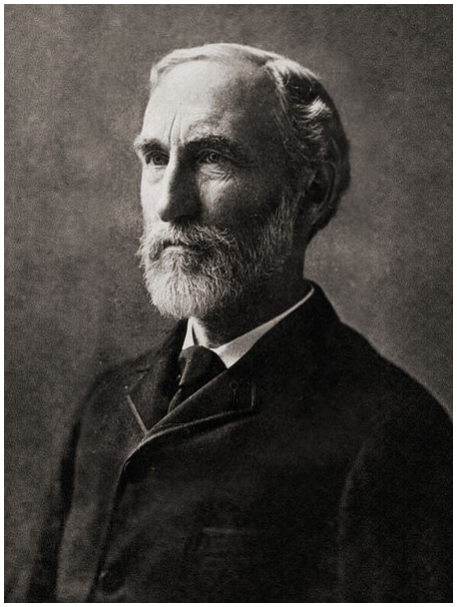

Figura 4.1: Josiah Willard Gibbs (1839-1903)

4.2 Esquema de temperatura

El nombre del **recocido simulado** (**simulated annealing** en inglés) se refiere al proceso de recocido en metalurgia que corresponde a un ciclo de calentamiento de un metal que consiste en un paso de aumento gradual de la temperatura seguido de un descenso controlado para mejorar las cualidades del metal. La idea física es que un descenso demasiado brutal puede bloquear el metal en un estado desfavorable. Es la misma idea que encontramos en el algoritmo de recocido simulado para evitar que la sucesión generada quede bloqueada en los mínimos locales. Así, consideramos un **esquema de temperatura** (**cooling schedule** en inglés) dado por una sucesión de temperaturas decrecientes $(T_n)_{n\geqslant 1}$. El principio es el mismo que en la versión simple 4.3 con la diferencia de que la temperatura cambia según la iteración, lo que implica la falta de homogeneidad de la cadena de Markov producida.

ALGORITMO 4.4 – Recocido simulado con esquema de temperatura

Inicialización :
- f : función real a minimizar sobre un espacio E finito
- $(T_n)_{n\geqslant 1}$: esquema de temperatura tal que $T_n > 0$ para cualquier $n \geqslant 1$
- $X_0 \in E$: estado inicial

En el paso $n \geqslant 1$:
  Sortear X' uniformemente en los vecinos de X_{n-1}
  Si $f(X') \leqslant f(X_{n-1})$:
    Definir $X_n = X'$
  Si no :
    Calcular $\alpha = e^{-(f(X')-f(X_{n-1}))/T_n}$
    Generar $U \sim \mathscr{U}([0,1])$
    Regla de rechazo :
      - Si $U \leqslant \alpha$, el estado es aceptado : $X_n = X'$
      - Si $U > \alpha$, el estado es rechazado : $X_n = X_{n-1}$

Devolver los estados $X_0, X_1, \ldots$

La dificultad aquí es elegir un buen esquema de temperatura. Debemos asegurar que la distribución converja hacia la distribución uniforme en el conjunto $E_{\text{mín}}$ de los minimizadores de la función f y que el tiempo para alcanzar esta distribución esté controlado. En lo que sigue, diremos que el recocido simulado **converge** si la cadena de Markov $(X_n)_{n\in\mathbb{N}}$ producida por el algoritmo 4.4 verifica

$$\lim_{n\to+\infty} \mathbb{P}(X_n \in E_{\text{mín}}) = 1.$$

El principal obstáculo para la convergencia del recocido simulado es la profundidad de los mínimos locales que se debe evitar.

Definición 4.5. Sea una función real f sobre un espacio E finito y $E_{\text{mín}}$ el conjunto de los minimizadores de f. Para cualquier estado $k \in E \setminus E_{\text{mín}}$, decimos que k **comunica** con $E_{\text{mín}}$ a una altura $h > 0$ si existen estados vecinos $k_0, \dots, k_\ell \in E$ entre k y $E_{\text{mín}}$,

$$k_0 = k, \quad k_\ell \in E_{\text{mín}} \quad \text{y} \quad k_{j-1} \sim k_j,\ \forall j \in \{1,\dots,\ell\},$$

tales que

$$\forall j \in \{0,\dots,\ell\},\ f(k_j) \leqslant f(k) + h.$$

La **altura de comunicación** h^* de f es la altura más pequeña a la que cualquier estado de $E \setminus E_{\text{mín}}$ comunica con $E_{\text{mín}}$.

Un resultado teórico importante para la convergencia del recocido simulado fue demostrado por Bruce Hajek en 1988 en [Haj88].

Teorema 4.6. *Sea f una función real sobre un espacio E finito con altura de comunicación $h^* > 0$. El algoritmo de recocido simulado converge para el esquema de temperatura $(T_n)_{n\geqslant 1}$ si, y solo si,*

$$\lim_{n\to+\infty} T_n = 0 \quad y \quad \sum_{n\geqslant 1} \exp\left(-\frac{h^*}{T_n}\right) = +\infty.$$

Demostración. Admitido. □

El esquema de temperatura que parece más natural a la vista del teorema de Hajek es un decrecimiento logarítmico,

$$\forall n \geqslant 1,\ T_n = \frac{h}{\ln(n)} \tag{4.1}$$

donde $h \geqslant h^*$. Es posible demostrar que la convergencia es más rápida cuando h está cerca de h^*. Sin embargo, el valor h^* no se conoce en la práctica y el resultado del teorema no indica el tiempo para alcanzar un minimizador con una precisión determinada. En casos concretos, el parámetro h debe ajustarse experimentalmente.

Para sortear la falta de homogeneidad de la cadena de Markov, una idea simple que se usa comúnmente en la práctica es mantener la temperatura constante durante etapas que aumentan de manera exponencial. Esta variante se llama **recocido simulado por etapas** y corresponde al esquema de temperatura definido por

$$\forall N \geqslant 1,\ \forall n \in \left\{e^{a(N-1)}, \dots, e^{aN} - 1\right\},\ T_n = \frac{b}{N}.$$

donde $a, b > 0$ son parámetros a ajustar. Para tal esquema de temperatura, la cadena de Markov es homogénea sobre intervalos de tiempo más y más largos. Mostraremos que estas etapas de tamaño exponencial son suficientemente largos para que la cadena alcance su equilibrio que corresponde a la medida de Gibbs π^T de función de energía f a la temperatura T de la etapa. Dado que las temperaturas T_n son decrecientes hacia 0, la medida π^{T_n} converge hacia la distribución uniforme en E_{min} y esto asegurará la convergencia del recocido simulado por etapas.

4.3 Convergencia del recocido simulado por etapas

Sea un espacio E con cardinalidad K finita, empezamos estableciendo dos desigualdades para la distancia en variación total entre dos medidas de probabilidad ν y μ sobre E que se utilizarán más adelante. Suponiendo que $\mu_k > 0$ para cualquier $k \in E$, el primer resultado es el enlace con la divergencia de Kullback-Leibler entre ν y μ definida por

$$\mathscr{K}(\nu,\mu) = \sum_{k\in E} \nu_k \ln(\nu_k/\mu_k).$$

Gracias al hecho de que

$$\forall t \geqslant 0,\ 3(t-1)^2 \leqslant 2(2+t)(t\ln(t)+1-t),$$

podemos deducir por la desigualdad de Cauchy-Schwarz,

$$\begin{aligned}
3\|\nu-\mu\|_{VT}^2 &= \frac{1}{4}\left(\sum_{k\in E}\sqrt{3}|\nu_k-\mu_k|\right)^2 = \frac{1}{4}\left(\sum_{k\in E}\mu_k\sqrt{3}\left|\frac{\nu_k}{\mu_k}-1\right|\right)^2 \\
&\leqslant \frac{1}{4}\left(\sum_{k\in E}\mu_k\sqrt{2\left(2+\frac{\nu_k}{\mu_k}\right)\left(\frac{\nu_k}{\mu_k}\ln\left(\frac{\nu_k}{\mu_k}\right)+1-\frac{\nu_k}{\mu_k}\right)}\right)^2 \\
&\leqslant \frac{1}{4}\left(\sum_{k\in E}2\left(2\mu_k+\nu_k\right)\right)\left(\sum_{k\in E}\nu_k\ln\left(\frac{\nu_k}{\mu_k}\right)+\mu_k-\nu_k\right) \\
&= \frac{3}{2}\mathscr{K}(\nu,\mu).
\end{aligned}$$

Así, hemos demostrado que

$$\|\nu-\mu\|_{VT}^2 \leqslant \frac{1}{2}\mathscr{K}(\nu,\mu). \tag{4.2}$$

La otra desigualdad útil se refiere al kernel de Metrópolis-Hastings P, entendido como una matriz, que admite μ como única medida de probabilidad invariante. Para $n \in \mathbb{N}$, introducimos la distribución de la iteración n de una cadena de Markov de distribución inicial ν y de matriz de transición P,

$$\forall k \in E,\ \nu P^n(k) = \sum_{\ell\in E}\nu_\ell (P^n)_{\ell,k}.$$

Consideramos los K valores proprios de la matriz P en orden decreciente con la multiplicidad,

$$\lambda_1 = 1 > \lambda_2 \geqslant \cdots \geqslant \lambda_K > -1.$$

Un estudio similar a lo que hicimos para obtener (3.5) nos lleva a

$$\|\nu P^n-\mu\|_{VT}^2 \leqslant \frac{1}{4}\max\left\{\lambda_2^{2n},\lambda_K^{2n}\right\}\chi^2(\nu,\mu). \tag{4.3}$$

Para establecer una velocidad de convergencia, consideramos una recurrencia sobre las etapas. Por eso, tomamos $a, b > 0$ y el esquema de temperatura (4.1). Para cualquier $N \in \mathbb{N}$, $R_N = \lceil e^{aN} \rceil$ es el entero más pequeño mayor o igual a e^{aN} y, si $N \geqslant 1$, $r_N = R_N - R_{N-1}$ es la duración de la etapa donde la temperatura es constante igual a b/N. El objetivo es minimizar una función $f : E \to \mathbb{R}$ de altura de comunicación $h^* > 0$ para la cual introducimos algunas notaciones,

$$m = \min_{k \in E} f(k), \quad F_- = \min_{\substack{k \in E \text{ t.q.} \\ f(k) \neq m}} \{f(k) - m\} \quad \text{y} \quad F_+ = \max_{k \in E} \{f(k) - m\}$$

Sea un entero $N \geqslant 1$, la medida de Gibbs π_N de función de energía f a la temperatura b/N está dada por

$$\forall k \in E,\ \pi_{N,k} = \frac{1}{Z_{f,N}} \exp(-N(f(k) - m)/b) \quad \text{donde} \quad Z_{f,N} = \sum_{\ell \in E} \exp(-N(f(\ell) - m)/b).$$

En lo que sigue, suponemos que existe $M > 0$ tal que la función de partición es limitada,

$$\sup_{N \geqslant 1} Z_{f,N} \leqslant M.$$

En un espacio E de cardinalidad K finita, siempre es posible tomar $M = K$, aunque cualquier valor inferior ofrece mejores constantes en el resto de la demostración.

Consideramos la cadena de Markov $(X_n)_{n \in \mathbb{N}}$ producida por el recocido simulado por etapas con el esquema (4.1) y la función f. Para cualquier $N \geqslant 1$, la distribución de las variables $X_{R_{N-1}}, \ldots, X_{R_N - 1}$ es la de una cadena de Markov homogénea dada por un algoritmo de Metrópolis-Hastings de kernel P_N reversible con respecto a la medida de probabilidad π_N y que tiene valores propios $\lambda_{N,1}, \ldots, \lambda_{N,K}$. Por la proposición 2.23, sabemos que

$$\forall j \in \{R_{N-1}, \ldots, R_N - 1\},\ \|\mu_j - \pi_N\|_{VT} = \|\mu_{R_{N-1}} P_N^{j - R_{N-1}} - \pi_N\|_{VT} \leqslant \|\mu_{R_{N-1}} - \pi_N\|_{VT}$$

donde, para cualquier $n \in \mathbb{N}$, μ_n es la distribución de X_n. Así, introducimos la sucesión $(u_N)_{N \geqslant 1}$ de los cuadrados de las distancias en variación total entre la distribución inicial y la probabilidad invariante para una etapa dada,

$$\forall N \geqslant 1,\ u_N = \|\mu_{R_{N-1}} - \pi_N\|_{VT}^2.$$

Tomemos una etapa $N \geqslant 1$ tal que $r_N > 0$, nuestro objetivo es establecer una recurrencia con la siguiente etapa,

$$\begin{aligned} u_{N+1} = \|\mu_{R_N} - \pi_{N+1}\|_{VT}^2 &= \|\mu_{R_{N-1}} P_N^{r_N} - \pi_{N+1}\|_{VT}^2 \\ &\leqslant 2\|\mu_{R_{N-1}} P_N^{r_N} - \pi_N\|_{VT}^2 + 2\|\pi_N - \pi_{N+1}\|_{VT}^2 \\ &\leqslant \frac{1}{2} \max\left\{\lambda_{N,2}^{2r_N}, \lambda_{N,K}^{2r_N}\right\} \chi^2(\mu_{R_{N-1}}, \pi_N) + 2\|\pi_N - \pi_{N+1}\|_{VT}^2 \end{aligned}$$

donde la segunda desigualdad se deduce de (4.3). La distancia χ^2 verifica

$$\chi^2(\mu_{R_{N-1}}, \pi_N) = \sum_{k \in E} \frac{(\mu_{R_{N-1},k} - \pi_{N,k})^2}{\pi_{N,k}} \leqslant M e^{NF_+/b} \left(\sum_{k \in E} |\mu_{R_{N-1},k} - \pi_{N,k}| \right)^2 = 4M e^{NF_+/b} u_N.$$

El control de los valores propios de P_N requiere un análisis espectral del kernel de Metrópolis-Hastings que va fuera del alcance de este curso. Admitiremos el resultado siguiente que establece un enlace entre estos valores propios y la altura de la comunicación.

Teorema 4.7. *Sea $T > 0$, la matriz P_T está dada por el kernel de transición de la cadena de Markov homogénea producida por el algoritmo de Metrópolis-Hastings para la medida de Gibbs π^T sobre E de cardinalidad K finita, a la temperatura $T > 0$ y de función de energía f de altura de comunicación $h^* > 0$. Entonces, existe una constante $C > 0$ tal que*

$$\forall j \in \{2,\dots,K\},\ \lambda_{T,j}^2 \leqslant 1 - C\exp\left(-\frac{h^*}{T}\right)$$

donde $\lambda_{T,2},\dots,\lambda_{T,K}$ son los valores propios de P_T estrictamente inferior a 1.

Demostración. Admitido. □

Gracias al teorema anterior, se deduce la desigualdad siguiente entre u_N y u_{N+1},

$$u_{N+1} \leqslant 2Me^{NF_+/b}\left(1 - Ce^{-Nh^*/b}\right)^{2r_N} u_N + 2\|\pi_N - \pi_{N+1}\|_{VT}^2. \tag{4.4}$$

Este resultado permite dar el orden de magnitud de la velocidad de convergencia. En efecto, si consideramos que r_N es bastante grande para aniquilar el primer término, solo se queda el segundo. Por (4.2) y un cálculo simple basado en las hipótesis, existe una constante $C' > 0$ tal que

$$2\|\pi_N - \pi_{N+1}\|_{VT}^2 \leqslant \mathscr{K}(\pi_N, \pi_{N+1}) \leqslant \frac{C'e^{-NF_-/b}}{b}.$$

Por construcción, sabemos que $N \geqslant \ln(n)/a$ y deducimos por tanto una velocidad polinomial dada por

$$\|\mu_n - \pi^{ab/\ln(n)}\|_{VT} \lesssim \frac{C'n^{-F_-/(2ab)}}{b}.$$

Esta heurística se formaliza a través del estudio de la recurrencia dada por (4.4),

$$u_{N+1} \leqslant \beta_N u_N + \delta_N$$

donde, para cualquier $N \geqslant 1$,

$$\beta_N = 2Me^{NF_+/b}\left(1 - Ce^{-Nh^*/b}\right)^{2r_N} \quad \text{y} \quad \delta_N = \frac{C'e^{-NF_-/b}}{b}.$$

Por iteración, obtenemos

$$u_{N+1} \leqslant B_N\left(u_1 + \sum_{j=1}^{N}\frac{\delta_j}{B_j}\right)$$

donde, para cualquier $j \geqslant 1$, $B_j = \prod_{\ell=1}^{j}\beta_\ell$.

No desarrollaremos los cálculos (es un ejercicio riguroso pero factible para el lector), pero un primer paso es controlar B_N. Por definición, existe $\rho_a > 0$ tal que $r_N \geqslant \rho_a e^{aN}$. Si b es bastante pequeña para verificar $h^* > b\ln(C)$, entonces se obtiene

$$\begin{aligned} B_N &= \prod_{j=1}^{N} \exp\left(2r_j \ln(1 - Ce^{-jh^*/b}) + \frac{jF_+}{b} + \ln(2M)\right) \\ &\leqslant \prod_{j=1}^{N} \exp\left(-2C\rho_a e^{j(a-h^*/b)} + \frac{jF_+}{b} + \ln(2M)\right) \\ &= \exp\left(-2C\rho_a \sum_{j=1}^{N} e^{j(a-h^*/b)} + \frac{N(N+1)F_+}{2b} + N\ln(2M)\right) \end{aligned}$$

donde usamos la desigualdad $\ln(1-x) \leqslant -x$ para cualquier $0 < x < 1$. Así, una condición necesaria para garantizar la convergencia de B_N hacia cero es

$$a - h^*/b > 0.$$

Esta condición es equivalente a $ab > h^*$ y, si el producto ab se usa como la cantidad h en el esquema logarítmico (4.1), recobramos la condición $h > h^*$ dada por el teorema de Hajek. Bajo ciertas hipótesis, el estudio de la recurrencia (4.4) nos lleva a la siguiente velocidad de convergencia para el recocido simulado por etapas

$$\|\mu_n - \pi^{ab/\ln(n)}\|_{VT} \leqslant C_1 \ln\ln(n) n^{-C_2 F_-/(ab)}$$

donde $C_1, C_2 > 0$ son constantes.

Bibliografía

[DWW99] Paul Damien, Jon Wakefield, and Stephen Walker. Gibbs sampling for bayeisan non-conjugate and hierarchical models using auxiliary variables. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(2):331–344, 1999.

[Haj88] Bruce Hajek. Cooling schedules for optimal annealing. *Mathematics of Operations Research*, 13(2):311–329, 1988.

[Has70] Wilfred K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[KGJV83] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[MRR+53] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[MT09] Sean Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge Mathematical Library. Cambridge University Press, 2009.

[Nea03] Radford M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 2003.

[RC04] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics, 2004.

[Če85] V. Černý. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45:41–51, 1985.

Créditos fotograficos

• Figura 1.1 (izquierda) : John von Neumann, from period while at Los Alamos National Laboratory, taken from a Los Alamos publication (*Los Alamos: Beginning of an era*, 1943–1945, Los Alamos Scientific Laboratory, 1986).
Dominio público (Terms and Conditions of Use).

• Figura 1.1 (derecha) : Stanislaw Ulam (alrededor de 1945).
Dominio público (Terms and Conditions of Use).

• Figura 2.1 : Photo of mathematician Andrey Markov.
Dominio público.

• Figura 3.3 (izquierda) : Portrait of American computer scientists Nicholas Metropolis (1915–1999) (seated) and James Henry Richardson (1918–1996) at Los Alamos National Laboratory, Los Alamos, New Mexico, November 1953. (Photo by Loomis Dean/The LIFE Picture Collection/Getty Images).
Rights managed, Getty Images content licence agreement.

• Figura 3.3 (derecha) : Wilfred Keith Hastings.
Published in Victoria Times Colonist from May 21 to May 22, 2016.

• Figura 4.1 : Frontispiece of *The Scientific Papers of J. Willard Gibbs*, in two volumes, eds. H. A. Bumstead and R. G. Van Name, (London and New York: Longmans, Green, and Co., 1906).
Dominio público.