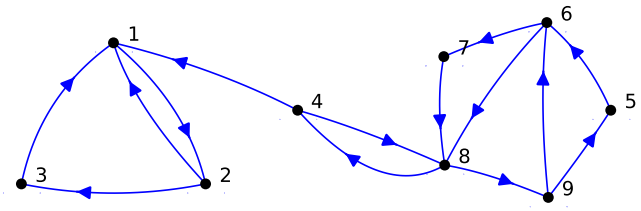
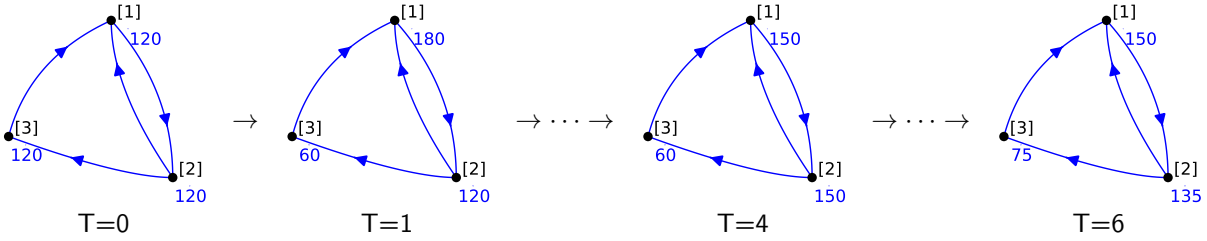
Cadenas de Markov

Un modelo para redes

Supongamos una red de comunicaciones en la que hay diferentes nodos conectados de forma que cada "cable" tiene una dirección en la que fluye la información. El ejemplo que tenemos en mente son las páginas web que componen internet con todos los enlaces que hay entre ellas. La estructura matemática que corresponde a esta situación es un *grafo dirigido*, lo que geométricamente no es más que un conjunto de vértices conectados por aristas con una dirección asignada.



Pensando en el caso de internet, un modelo natural consiste en suponer que en primera aproximación un visitante de una página web escoge al azar uno de los enlaces que se le ofrecen para continuar navegando. Por supuesto hay enlaces más destacados que otros pero eso se podría introducir más adelante en el modelo. Siguiendo con esta idea, para saber qué páginas de la red son más transitadas, se puede hacer una simulación de un *paseo aleatorio* (aunque en este caso bien determinista) asignando un número de visitantes a cada página y suponiendo que en cada unidad de tiempo discretizado se reparten equitativamente entre los diferentes enlaces. Por ejemplo, a continuación se muestra la evolución de un sencillo diagrama al poner 120 personas en cada vértice:



En cada paso, todos los del vértice [1] pasan al [2] y todos los de [3] pasan al [1], mientras que los de [2] se reparten al 50 % entre los otros vértices.

Si se prolonga la simulación durante unos pasos más, parece que en el límite los vértices [1], [2] estarán visitados por 144 personas y el [3] por 72. Por supuesto, la elección de 120 es convencional, se podrían haber tomado valores iniciales de $1/3$, que están más de acuerdo con la idea probabilista. En este caso la *distribución límite* sería $(2/5, 2/5, 1/5)$ que sugiere que un internauta navegando al azar en esta red en miniatura, tiene a la larga una probabilidad de $2/5$ de estar en cada uno de los vértices [1] y [2], y $1/5$ de estar en [3]. En cualquier caso, la conclusión sería que los vértices [1] y [2] son el doble de importantes, tienen el doble de visitantes, que [3]. Todavía más, esta conclusión parece independiente de la *distribución inicial*. Por ejemplo si ponemos a las 360 personas en [3], hay unas oscilaciones iniciales pero en 8 unidades de tiempo ya hay 135 en [1] y [2] y 90 en [3], mientras que esperando unas decenas de unidades de tiempo ya está meridianamente claro que nos acercamos a la distribución límite $(2/5, 2/5, 1/5)$.

Una observación básica es que la distribución $(2/5, 2/5, 1/5)$ que hemos obtenido como límite es *estacionaria*. Esto es, si ponemos $2/5$ de las personas en [1], lo mismo en [2] y el resto en [3], en los instantes siguientes el número de personas en cada vértice no varía.

La pregunta es si este procedimiento para clasificar la importancia de páginas web siempre nos llevará a un resultado. Podemos descomponer la pregunta en varias que concretaremos más adelante con definiciones y resultados matemáticos.

P1. ¿Existe siempre una distribución estacionaria?

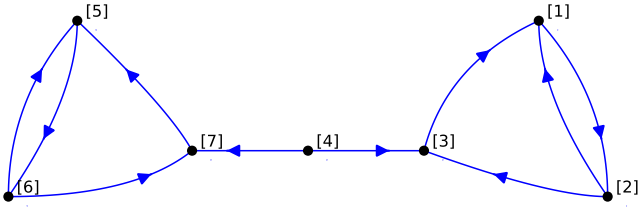
P2. Si existe una distribución estacionaria, ¿es única?

P3. El procedimiento descrito, ¿siempre da lugar a una distribución límite?

P4. La distribución límite, cuando existe, ¿es independiente de la distribución inicial?

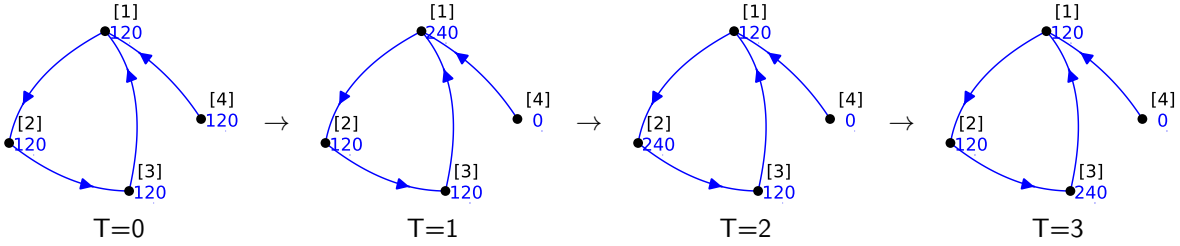
Veamos con una serie de contraejemplos que las tres últimas preguntas no pueden tener una respuesta incondicionalmente afirmativa.

Si consideramos dos copias del ejemplo anterior de tres vértices conectadas del siguiente modo:



Estas copias no se "comunican" y tanto $(1/3, 1/3, 1/3, 0, 0, 0, 0)$ como $(0, 0, 0, 0, 1/3, 1/3, 1/3)$ son distribuciones estacionarias, así como cualquier combinación convexa de ambas. Esto da una respuesta negativa a P2 y, como en cada mitad tenemos convergencia a una distribución límite, también a P4. Cualquiera de las distribuciones antes citadas podrían resultar como límite, de hecho son todos los posible límites, dependiendo de los valores iniciales asignados.

Para P3, consideremos un triángulo de vértices conectados en sentido positivo y un cuarto vértice con una arista hacia este triángulo. Digamos, como antes, que situamos 120 personas en cada vértices y estudiamos la evolución. Como no hay aristas que lleguen al último vértice, tras el tiempo inicial se quedará sin visitantes, mientras que el resto de los vértices muestra un comportamiento oscilante.



Si la simulación funcionase hasta T=1000 parecería que el primer vértice es el doble de importante que los otros pero si funcionase hasta T=1001, llegaríamos a una conclusión similar respecto al segundo vértice. Hay una falta de convergencia que resta valor a esta forma de sacar conclusiones a partir de la simulación y da una respuesta negativa a P3. Uno podría considerar la solución de promediar en el tiempo el número de visitantes de cada vértices obteniendo una distribución límite promediada $(1/3, 1/3, 1/3, 0)$ que parece razonable, pero el ejemplo empleado para P2 y P4 no permite dar por supuesto que este promedio no se vea afectado por la elección de la distribución inicial.

Hay teoremas que dan condiciones para asegurar una respuesta afirmativa a P1-P4. Una sencilla modificación de la idea anterior perturbando ligeramente el grafo para que cumpla estas condiciones, da lugar al *page rank algorithm* que es empleado por el buscador más famoso para ordenar la relevancia de las páginas web.

Cadenas de Markov

Intuitivamente, una *cadena de Markov* es un proceso para el cual lo que prevemos que ocurra mañana depende con cierta probabilidad de lo que ocurre hoy, sin que importe ningún conocimiento añadido sobre la historia anterior. Por ejemplo, si he tirado un dado cada día durante una semana y llevo acumulados 26 puntos, mañana obtendré 32 puntos con probabilidad 1/6 y esto independientemente de cómo haya conseguido los 26 puntos. Evidentemente, hay alguna información implícita acerca de la historia, por ejemplo, no he sacado un 1 todos los días, pero eso es irrelevante para calcular la probabilidad de obtener 32 puntos mañana.

El modelo matemático para este tipo de procesos es utilizar el tiempo discretizado, a través de los naturales (con el cero), como índice de una sucesión de variables aleatorias, lo que lleva a una definición sintética de cadena de Markov, aunque un poco oscura sin la ayuda de un ejemplo.

Las cadenas de Markov fueron introducidas por Andrei Markov a comienzos del siglo XX. También es relevante la contribución de Andrei Kolmogorov quien, dicho sea de paso, en 1933 axiomatizó la teoría de probabilidades de la manera que todavía hoy se explica en los cursos para matemáticos.

Definición. *Una* cadena de Markov *es una sucesión de variables aleatorias* $\{X_n\}_{n=0}^{\infty}$ *que toman valores en un conjunto numerable* S*, el* conjunto de estados*, tales que*

$$\text{Prob}\big(X_{n+1} = v | X_n = u\big) = \text{Prob}\big(X_{n+1} = v | X_n = u, X_{n-1} = u_{n-1}, \dots \ X_0 = u_0\big) \tag{1}$$

para cualesquiera $n \geq 0$ *y* $u, v, u_0, \dots, u_{n-1} \in S$*. Además supondremos que la probabilidad indicada en* (1) *es independiente de* n*.*

En el ejemplo anterior, X_n es la puntuación el día n, el conjunto de estados S son los naturales (o los enteros o racionales, si se prefiere) y la fórmula de la definición lo que dice es que si sabemos que la puntuación el día n es u, podremos calcular la probabilidad de que mañana sea v sin que importe lo que ha ocurrido en días anteriores. Otro ejemplo es un *game* de tenis en el que el primer jugador tiene una probabilidad p de ganar un punto. En este caso hay 20 estados que corresponden a 15 puntuaciones numéricas, *deuce*, *advantage* para el primer o segundo jugador y victoria para el primer o segundo jugador. Por cierto, en [KS76] se analiza esta cadena de Markov y se concluye que si el primer jugador es mejor, $p > 1/2$, su probabilidad de victoria es $p^4(1 - 16q^4)/(p^4 - q^4)$ con $q = 1 - p$, de donde una pequeña superioridad en cada punto se traduce en una muy grande en *matches*.

La última suposición de la definición, no la exigen todos los autores aunque es muy común en las aplicaciones (en la jerga, se dice que la cadena de Markov es *homogénea* o estacionaria [PK10], aunque esta última notación es confusa). Intuitivamente indica que las "reglas del juego" con las que calculamos probabilidades no cambian con el tiempo. En el ejemplo, da igual que los 26 puntos se obtengan después de 7 tiradas o de 5, siempre darán lugar a 32 el día posterior con probabilidad 1/6. No sería así si por ejemplo sustituyéramos el dado cúbico por uno tetraédrico los jueves.

Si $|S| < \infty$ se dice que la cadena de Markov es *finita*. En caso contrario se dice que es *infinita*. El ejemplo del dado es del segundo tipo y el del tenis, del primero. Como la definición exige que el conjunto de estados sea numerable, renombrando sus elementos podemos suponer sin pérdida de generalidad $S = \{1, 2, \dots, N\}$ en el primer caso y $S = \mathbb{Z}^+$ en el segundo. De esta forma, es natural y habitual escribir i y j en vez de u y v en (1). La probabilidad p_{ij} de pasar del estado i al j (en un paso) se llama *probabilidad de transición* de i a j y es justamente la expresión que aparece en (1):

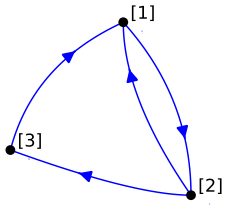
$$p_{ij} = \text{Prob}\big(X_{n+1} = j | X_n = i\big). \tag{2}$$

Por la hipótesis final de la definición de cadena de Markov, no depende de n.

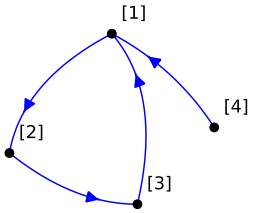
Se dice que las probabilidades de transición conforman la *matriz de transición* P, que es en rigor una matriz (finita) sólo para cadenas de Markov finitas pero conservamos el nombre y la notación también para las infinitas. Cada fila de esta matriz debe sumar 1 por propiedades básicas de la probabilidad:

$$\sum_{j \in S} p_{ij} = \sum_{j \in S} \text{Prob}\big(X_1 = j | X_0 = i\big) = \text{Prob}\big(X_1 \in S | X_0 = i\big) = 1. \tag{3}$$

El modelo para redes discutido en el apartado anterior se puede considerar una cadena de Markov donde S son los vértices del grafo dirigido y p_{ij} es el inverso del número de aristas salientes desde i cuando hay una arista de i a j y cero en otro caso.



$$\to P = \begin{pmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 1 & 0 & 0 \end{pmatrix}$$



$$\to P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

Al vector fila (π_0) cuyas coordenadas son $\text{Prob}(X_0 = i)$ se le llama *distribución inicial*. En parte por tradición y en parte por necesidades teóricas, a diferencia de lo que ocurre en álgebra lineal, los vectores en la teoría de cadenas de Markov son vectores fila. Para hacer hincapié en ello, los denotaremos entre paréntesis. De nuevo, si la cadena de Markov es infinita, no es estrictamente un vector sino una sucesión.

Por la ley de probabilidad total

$$\text{Prob}(X_1 = j) = \sum_{i \in S} \text{Prob}(X_0 = i)\text{Prob}\big(X_1 = j | X_0 = i\big) = \sum_{i \in S} \text{Prob}(X_0 = i) p_{ij} \tag{4}$$

y el segundo miembro puede entenderse como la coordenada j-ésima de $(\pi_0)P$. La iteración de esta idea lleva a la relación fundamental

$$(\pi_n) = (\pi_0)P^n \qquad \text{donde} \qquad (\pi_n) = \big\{\text{Prob}(X_n = i)\big\}_{i \in S}. \tag{5}$$

Es decir, que para calcular la probabilidad de pasar de un estado a otro en n pasos basta calcular una potencia n-ésima de una matriz. En el caso infinito no es difícil ver que P^n tiene sentido, empleando (3) y que $0 \le p_{ij} \le 1$.

El problema matemático al que nos enfrentamos es saber en qué situaciones existe $\lim(\pi_n)$, en ese caso diremos que el resultado es la *distribución límite*. La formula (5) reduce su estudio a límites de potencias de una matriz. La forma canónica es de gran ayuda en esta tarea. En los ejemplos anteriores, escribiendo $z = e^{3\pi i/4}/\sqrt{2}$ y $w = e^{2\pi i/3}$,

$$\begin{pmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 1 & 0 & 0 \end{pmatrix}^n = C^{-1} \begin{pmatrix} 1 & & \\ & z & \\ & & \bar{z} \end{pmatrix}^n C \to \begin{pmatrix} 2/5 & 2/5 & 1/5 \\ 2/5 & 2/5 & 1/5 \\ 2/5 & 2/5 & 1/5 \end{pmatrix} \tag{6}$$

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}^n = \widetilde{C}^{-1} \begin{pmatrix} 1 & & & \\ & w & & \\ & & \bar{w} & \\ & & & 0 \end{pmatrix}^n \widetilde{C} \quad \text{oscila} \quad (P^4 = P) \tag{7}$$

La clave ha sido que en el primer caso $|z| < 1$ y por tanto $z^n \to 0$, mientras que $|w| = 1$ y, de hecho, w^n tiene periodo tres.

En (6) el límite tiene todas las filas iguales y como las coordenadas de (π_0) deben sumar 1, la relación (5) asegura que la distribución límite no depende de la distribución inicial (π_0), lo que sugerían nuestros experimentos. Por otro lado, (7) muestra que en general $\lim(\pi_0)P^n$ no existe, aunque sí lo hace en (infinitos) casos particulares como $(\pi_0) = (1/3, 1/3, 1/6, 1/6)$.

En principio, utilizando este tipo de argumentos de álgebra lineal podríamos resolver los problemas relativos a cadenas de Markov finitas pero en la práctica calcular la forma canónica de una matriz grande es demasiado costoso. Es necesario desarrollar alguna teoría para abordar P1-P4. La teoría es todavía más necesaria en el caso infinito, puesto que el análisis funcional nos muestra lo difícil que puede ser tener un teorema espectral para operadores en espacios de dimensión infinita.

Los problemas que surgen en P1-P4 están relacionados con la falta de interconexión entre diferentes estados. Los experimentos sugieren que cuando "casi todos" los estados están conectados, es habitual que haya una distribución límite independiente de la distribución inicial. Así el contraejemplo a P2 y P4 tenía dos mitades entre las cuales no era posible ninguna comunicación. En el contexto de las cadenas de Markov hay dos versiones de esta conexión. La más débil se llama *irreducible* y la más fuerte *regular*. Estos nombres no son muy afortunados pero están demasiado asentados como para ser modificados.

Definición. *Se dice que una cadena de Markov es* irreducible *si es posible ir de cualquier estado a cualquier otro en un número finito de pasos. Equivalentemente, es* irreducible *si para cada i y j existe $k \in \mathbb{Z}^+$ tal que el elemento de ij de P^k es no nulo. Se dice que es* regular *si existe un k tal que es posible ir de cualquier estado a cualquier otro en exactamente k pasos. Equivalentemente, es* regular *si todos los elementos de P^k son no nulos para algún $k \in \mathbb{Z}^+$.*

Para abordar P1, definimos una *distribución estacionaria* como un posible valor (π) de (π_0) tal que $(\pi) = (\pi)P$. Según (5) esto asegura que $(\pi_n) = (\pi_0)$.

Teorema. *Una cadena de Markov finita siempre tiene al menos una distribución estacionaria.*

Quizá lo más sorprendente de este resultado es que su prueba no es inmediata. Aunque hay una elemental y breve (pero nada obvia), aquí veremos una todavía más breve basada en topología.

Demostración. Digamos que $|S| = N$. Consideremos el subconjunto compacto de $\mathbb{R}^N$ (el *simplex*) $K = \{(x) \in \mathbb{R}^N \ : \ x_i \geq 0, \ \sum x_i = 1\}$, el cual es homeomorfo a la bola cerrada $(N-1)$-dimensional. La función lineal $f(x) = (x)P$ aplica K en K y el teorema de Brouwer asegura que tiene un punto fijo. □

El teorema anterior no es cierto en general para cadenas de Markov infinitas porque siempre se pueden "alejar" suficientemente las probabilidades para lleguen a desaparecer. Por ejemplo, si tomamos $p_{ij} = 1$ si $j = i + 1$ y cero en otro caso, la ecuación $(x) = (x)P$ implica $x_i = x_{i+1}$ para $n \geq 1$ que no tiene solución con $\sum x_i = 1$.

El resultado que nos va a ser de más utilidad para nuestro propósito es el siguiente, que también puede interpretarse como una consecuencia de un resultado (nada sencillo) de álgebra lineal debido a Oskar Perron y Georg Frobenius [LM12, 15.2] [FG04] [Lax07].

Teorema. *Para una cadena de Markov finita regular, existe* $\lim_{n\to\infty}(\pi_n)$ *donde* $(\pi_n) = (\pi_0)P^n$ *y el resultado es la única distribución estacionaria. En particular, el límite no depende de la distribución inicial* (π_0).

De nuevo, daremos una demostración con aires topológicos a través de una variación del teorema de la aplicación contractiva para aplicaciones lineales.

Lema. *Sea $F : \mathbb{R}^m \longrightarrow \mathbb{R}^m$ una aplicación lineal y $\vec{0} \in \Omega$ un conjunto compacto tal que $F(\Omega) \subset \mathrm{Int}(\Omega)$, entonces para cualquier $\vec{x}_0 \in \Omega$, la sucesión definida por $\vec{x}_n = F(\vec{x}_{n-1})$ tiende a $\vec{0}$.*

Demostración. Basta probar que existe $0 < \delta < 1$ tal que $F(\Omega) \subset (1-\delta)\Omega = \{(1-\delta)\vec{x} \ : \ \vec{x} \in \Omega\}$ porque en este caso $F^n(\Omega) = \big(F \circ \overset{n \text{ veces}}{\cdots} \circ F\big) \subset (1-\delta)^n\Omega$ y $(1-\delta)^n\vec{x} \to \vec{0}$ para $\vec{x} \in \Omega$.

Veamos que $\delta = \mathrm{dist}\big(\partial\Omega, F(\Omega)\big)\|\vec{a}\|^{-1}$ es un valor válido donde $\vec{a}$ es el punto de Ω, necesariamente en $\partial\Omega$, más lejano al origen. Como $\partial\Omega$ y $F(\Omega)$ son compactos disjuntos, su distancia está bien definida y es positiva. Si $F(\Omega) \not\subset (1-\delta)\Omega$, entonces $\exists\vec{x} \notin \Omega$ tal que $(1-\delta)\vec{x} \in F(\Omega)$. Por la convexidad, existe $\lambda \in (1-\delta, 1)$ tal que $\lambda\vec{x} \in \partial\Omega$, por tanto por la elección de δ, $\delta \leq \|\lambda\vec{x} - (1-\delta)\vec{x}\|\|\vec{x}\|^{-1} = \lambda - (1-\delta)$, lo que contradice $\lambda < 1$. □

Demostración (del teorema). Sean K y f como en la prueba del primer teorema y sea (π) una distribución estacionaria. El conjunto $\Omega = \{(x) - (\pi) \ : \ (x) \in K\}$ es un conjunto compacto y convexo como subconjunto del hiperplano $x_1 + \cdots + x_N = 0$, que puede identificarse con $\mathbb{R}^{N-1}$. Se cumple $f(\Omega) \subset \Omega$ porque $f(K) \subset K$ y $f(\pi) = (\pi)$. Además $\mathrm{Int}(\Omega) = \{(x) - (\pi) \ : \ (x) \in K, \ x_i > 0\}$ y al ser cadena de Markov regular, existe k tal

que $f^k(\Omega) \subset \mathrm{Int}(\Omega)$. Al aplicar el lema con $F = f^k$ se tiene que $(\pi_{nk}) - (\pi)$ tiende a cero cuando $n \to \infty$ y de ahí, $\lim_{n\to\infty}(\pi_n) = (\pi)$, ya que f es continua y $f(\pi) = (\pi)$. □

La condición de regularidad no es fácil de comprobar para cadenas de Markov que tengan muchos estados. Intuitivamente el único problema que puede impedir la convergencia en el caso finito es que haya una oscilación periódica y para impedir esta situación, librándonos de cualquier hipótesis adicional, basta promediar (como en la sumación de Cesàro empleada a veces en análisis).

Teorema. *Para una cadena de Markov finita, el límite*

$$\lim_{n\to\infty} \frac{(\pi_0) + (\pi_1) + \cdots + (\pi_n)}{n+1}, \qquad \textit{con } (\pi_n) = (\pi_0)P^n, \tag{8}$$

siempre existe y es una distribución estacionaria.

Demostración. Consideremos $S_n = (n+1)^{-1}\sum_{k=0}^{n} P^k$. Las filas de P^k están compuestas por elementos positivos de suma 1 y, consecuentemente, también las de S_n. Por el teorema de Bolzano-Weierstrass, existe una subsucesión convergente S_{n_j}, digamos $\lim S_{n_j} = L_1$. Si no existiera el límite de S_n, habría otra subsucesión con $\lim S_{m_j} = L_2 \neq L_1$. Es fácil ver que $L_1P = L_1$ y $PL_2 = L_2$ ya que S_nP y PS_n son iguales a S_n salvo el primer y el último término. De aquí $L_1 = L_1S_{m_j}$ y $L_2 = S_{n_j}L_2$. Tomando límites $j \to \infty$, se sigue $L_1 = L_2$.

El límite del que habla el enunciado es $\lim(\pi_0)S_n = (\pi_0)L_1$ y $L_1P = L_1$ asegura que es una distribución estacionaria. □

Hay tantos posibles límites (8) como distribuciones estacionarias, lo que nos lleva a estudiar su unicidad. Para cadenas de Markov finitas se prueba [Doo53, p.181] que la unicidad equivale a la irreducibilidad. En las cadenas de Markov infinitas, surge el problema al que antes hemos apuntado de que las probabilidades se pueden alejar, incluso en el caso irreducible. Hay una forma elegante de tratar el problema introduciendo el *tiempo medio de retorno* a un estado i dado por

$$m_i = \mathbb{E}[T_i | X_0 = i] \qquad \text{donde} \quad T_i = \text{ínf}\left\{n > 0 \,:\, X_n = i\right\}. \tag{9}$$

En cadenas de Markov irreducibles finitas, como es fácil de sospechar, m_i está bien definido, esto es, $m_i < \infty$, pero no es así en las infinitas.

Llamando $N_n(i)$ al número de veces que hemos vuelto a i en n unidades de tiempo, matemáticamente la variable aleatoria $\left|\left\{0 < j \leq n \,:\, X_j = X_0 = i\right\}\right|$, parece natural esperar $N_n(i)m_i \sim n$. Esto se puede concretar en [HPS72, §2.3]

$$\lim_{n\to\infty} \frac{N_n(i)}{n} = \frac{1}{m_i} \qquad \text{casi seguro,} \tag{10}$$

siempre que volver a i tenga probabilidad 1. La prueba es una aplicación de la *ley fuerte de los grandes números*.

Justamente, que se vuelva con cierta frecuencia a un estado es la condición que asegura la unicidad en las cadenas de Markov irreducibles.

Teorema. *Una cadena de Markov irreducible tiene una distribución estacionaria si y sólo si m_i, definido en* (9)*, es finito para algún estado. Además, en este caso, la distribución estacionaria es única y su coordenada j-ésima es $1/m_j$ para cada $j \in S$.*

La demostración puede consultarse en [Dur99, p.86]. La idea fundamental es que, salvo una normalización, es posible construir la distribución estacionaria a partir de un estado i con $m_i < \infty$ tomando como coordenada j la suma de las probabilidades de que en n pasos se vaya de i a j y posteriormente se vuelva por primera vez a i.

El *page rank algorithm*

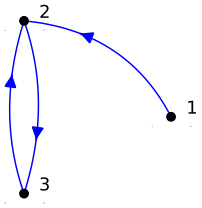
Aunque es un algoritmo patentado (U.S. Patent 6,285,999) e incluso ha dado lugar a una marca registrada, es una sencilla variación de métodos matemáticos bien conocidos (y muchas veces aplicados) que se remontan a los trabajos de Markov de hace más de 100 años.

A la luz de los resultados anteriores, la idea es bien simple, modificar ligeramente la matriz de transición cambiando los ceros por números pequeños para asegurar que la cadena de Markov sea regular. Concretamente se reemplaza P por

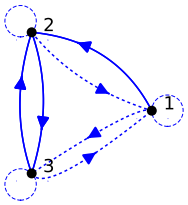
$$P_\epsilon = (1-\epsilon)P + \epsilon E \qquad \text{donde } E = (e_{ij})_{i,j=1}^{N} \text{ con } e_{ij} = 1/N. \tag{11}$$

Es fácil ver que P_ϵ sigue siendo una matriz de transición de una cadena de Markov para $0 < \epsilon < 1$, es decir, sus filas siguen sumando 1 y sus elementos están entre 0 y 1. De cara al modelo inicial, el sumando ϵE significa que permitimos la posibilidad de que un internauta salte al azar a una página sin seguir un enlace (en [LM12] se dice que esta es la "matriz de teletransporte"). Si ϵ es pequeño, damos menos peso a esta posibilidad. En definitiva, creamos artificialmente enlaces "débiles".

Como ejemplo, consideremos la red del primer diagrama. Claramente, tras el primer paso, el paseo aleatorio dará oscilaciones entre los vértices 2 y 3. Esto está relacionado con que P^n no tiene límite. La matriz P_ϵ está asociada de alguna forma al segundo diagrama:



$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \longrightarrow$$



$$P_\epsilon = \begin{pmatrix} \epsilon/3 & 1-2\epsilon/3 & \epsilon/3 \\ \epsilon/3 & \epsilon/3 & 1-2\epsilon/3 \\ \epsilon/3 & 1-2\epsilon/3 & \epsilon/3 \end{pmatrix}$$

Para cualquier $0 < \epsilon < 1$ la matriz P_ϵ corresponde a una cadena de Markov regular y por tanto existe el límite de (π_n) y coincide con la única distribución estacionaria. En este caso se obtiene

$$\lim_{n\to\infty}(\pi_n) = \left(\frac{\epsilon}{3}, \frac{1-2\epsilon/3}{2-\epsilon}, \frac{1-\epsilon+\epsilon^2/3}{2-\epsilon}\right). \tag{12}$$

Si ϵ es pequeño, esto se parece a $(0, 1/2, 1/2)$, lo cual está conforme con la idea natural de que el primer vértice es irrelevante, ya que $\text{Prob}(X_n = 1) = 0$ para $n > 0$, y que los otros dos son intercambiables.

Veamos ahora superficialmente algunas consideraciones prácticas. Los buscadores más completos afirman que tienen indexadas más de 10^{12} páginas. Multiplicar una matriz cuadrada de estas dimensiones por un vector es algo demasiado costoso (sin hablar del almacenamiento). Pensemos de manera ilusoriamente optimista que un ordenador "normal" actual pudiera hacer $2 \cdot 10^9$ operaciones por segundo (casi una por cada ciclo de reloj), entonces como la matriz tiene al menos 10^{24} elementos no nulos, cada iteración con P_ϵ llevaría al menos unos 16 millones de años. Lo que salva este obstáculo es que el primer sumando de (11) es una matriz muy dispersa. Si cada página tiene en media k enlaces, entonces $(\pi_n)P$ requeriría algo comparable a $k \cdot 10^{12}$ operaciones, lo cual es factible. Por otro parte, $(\pi_n)E$ no requiere ninguna operación, todas sus coordenadas son $1/N$.

Según los informes, en la práctica se hacen entre 50 y 100 iteraciones, lo cual da lugar a un cálculo asequible, en el que además se puede emplear computación paralela. Posiblemente el cálculo en sí no consuma tanto tiempo como la recopilación de la información y el acceso a ella. La actualización de los datos se realiza en periodos del orden de un mes y el cálculo, siempre según los informes, lleva varios días [Aus06] y se realiza con $\epsilon = 0.15$. Éste número (seguramente fruto de prueba y error) es una solución de compromiso entre la precisión deseada y el número de iteraciones posibles. Cuanto menor sea ϵ es de esperar una menor velocidad de convergencia [LM12, §6.1]. Por otro lado un ϵ que no sea pequeño aleja el resultado del modelo original. Una posibilidad que parece

no ponerse en práctica es utilizar el tercero de los teoremas anteriores para acelerar la convergencia y tomar ϵ menor.

Es importante notar que el posicionamiento de un sitio web en un buscador tiene consecuencias económicas y que ninguno de los motores de búsqueda con uso significativo en la actualidad está mantenido por entidades sin ánimo de lucro [LM12, p.44]. Por ello, el algoritmo no es el único condicionamiento para la ordenación. Por ejemplo, algunos sitios web con fines comerciales llevan a cabo *link farming*, que, en términos matemáticos, consiste en introducir grafos completos interconectando artificialmente páginas que no guardan relación para incrementar la relevancia de todas ellas. Otras veces, iniciativas populares han establecido enlaces falsos que conectan personas con términos despectivos [LM12, p.51–55], [FG04]. El hecho de que estas acciones hayan tenido corto recorrido una vez que han sido advertidas, es una prueba de que el algoritmo matemático descrito no es una explicación completa de la ordenación de resultados en un buscador.

Movimiento Browniano y paseos aleatorios infinitos

En 1827 el botánico Robert Brown observó que pequeñas partículas de polen suspendidas en una disolución se trasladan siguiendo caminos caóticos, lo que hoy en día llamamos *movimiento browniano* (y que tiene un significado más concreto en el ámbito matemático). Incialmente se interpretó como un signo de vida primaria, pero más tarde el desarrollo de la teoría atómica probó que representaba los empujones en direcciones aleatorias que dan las moléculas a las partículas de polen. Albert Einstein contribuyó decisivamente en el desarrollo del modelo matemático y en su (brevísima) tesis lo empleó (junto con modelos de fluidos) en la aproximación teórica de la *constante de Avogadro* (es poco conocido, que debido a un error, años más tarde señalado por un estudiante, su memoria termina dando el valor $N_A = 2.1 \cdot 10^{23}$ y afirmando que "está de acuerdo con el orden de magnitud de lo obtenido con otros métodos", sin embargo este número es bien diferente del valor real, $6.022 \cdot 10^{23}$, y de lo que sugerían los experimentos de su época).

Para simplificar, nos restringimos al caso unidimensional, esto es, como si las moléculas de un gas estuvieran metidas en un tubo largo y delgado y sólo pudieran ir a la derecha o a la izquierda con la misma probabilidad. Discretizamos además tiempo y espacio para representar la situación como una cadena de Markov. Digamos que los valores del tiempo son 0, h, $2h$, $3h$, etc. y que una partícula se mueve saltando entre los puntos de $\epsilon\mathbb{Z}$. Consideramos $S = \mathbb{Z}$ y X_n será la variable aleatoria que toma el valor j cuando la partícula está en la posición ϵj en el tiempo hn.

Un análisis combinatorio bien conocido del *paseo aleatorio* unidimensional [PK10] muestra que, con la notación de (10), $N_n(i)/\sqrt{n} \to \sqrt{2/\pi}$ casi seguro, por tanto $m_i = \infty$ y no hay una distribución estacionaria (también se puede comprobar directamente). Esto también sugiere que en tiempo $hn = 1$ se aleja del origen del orden de $\epsilon\sqrt{n}$. Si se quiere que en $hn = 1$ la distancia se mantenga acotada (el alejamiento medio por unidad de tiempo sea finito), deberíamos tomar $h^{-1/2}\epsilon$ comparable a una constante. Supongamos convencionalmente $h = \epsilon^2/2$. Nuestro objetivo es estudiar qué ocurre cuando $\epsilon \to 0$, es decir, cuando desaparece la discretización, y entender la evolución del resultado con el tiempo.

Si aleatoriamente una partícula se traslada a la izquierda o a la derecha, se tiene $p_{ij} = 1/2$ si $j = i \pm 1$ y $p_{ij} = 0$ en otro caso. Esto es:

$$\text{Prob}(X_{n+1} = j) = \frac{1}{2}\text{Prob}(X_n = j-1) + \frac{1}{2}\text{Prob}(X_n = j+1) \tag{13}$$

que podemos reescribir como

$$\frac{\text{Prob}(X_{n+1} = j) - \text{Prob}(X_n = j)}{h} = \frac{\text{Prob}(X_n = j-1) + \text{Prob}(X_n = j+1) - 2\,\text{Prob}(X_n = j)}{\epsilon^2} \tag{14}$$

porque $1/2 = h/\epsilon^2$. Esperamos que cuando $\epsilon \to 0$, $\text{Prob}(X_n = j)$ se pueda representar como una función "buena" que dependa del espacio y el tiempo, $u(x,t)$ con $x = j\epsilon$, $t = hn$. De esta forma, la ecuación anterior conduce a

$$\frac{u(x,t+h) - u(x,t)}{h} = \frac{u(x-\epsilon,t) + u(x+\epsilon,t) - 2u(x,t)}{\epsilon^2}. \tag{15}$$

Utilizando la regla de l'Hôpital o el sentido común, se llega a la *ecuación del calor* en $\mathbb{R}$

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \qquad x \in \mathbb{R},\ t > 0. \tag{16}$$

Dada una condición inicial $u(x,0) = f(x)$ su solución general es

$$u(x,t) = (4\pi t)^{-1/2} \int_{-\infty}^{\infty} e^{-(x-y)^2/4t} f(y)\, dy. \tag{17}$$

Cuando $t \to +\infty$, $u(x,t) \to 0$ puntualmente, reafirmándonos en que no hay una distribución estacionaria. Las probabilidades desaparecen escapándose al infinito. Éste es un ejemplo de un *proceso de difusión*.

En cierto modo en la fórmula anterior para resolver (16), lo único que hace es "sumar" (integrar) todas las campanas de Gauss correspondientes a aplicar el teorema central del límite a los paseos aleatorios de cada partícula. La manera de estocástica de mirar a la ecuación del calor está lejos de ser una mera curiosidad y es una muestra más de la estrecha interrelación entre la intuición física y los modelos matemáticos [Kac66].

Generación de números pseudoaleatorios

La inmensa mayoría de los lenguajes de programación principales y paquetes matemáticos tienen algún comando para generar números aleatorios con una distribución uniforme, típicamente $U(0,1)$. Incluso si no programamos, muchas veces el *software* que usamos habitualmente elige opciones al azar (por ejemplo, el reproductor de música cuando escoge una pista de la *playlist*) y hasta nuestra calculadora de bolsillo puede tener una tecla `Ran#`.

Los números generados no son aleatorios sino bien deterministas pero simulan serlo y por ello se dice que son pseudoaleatorios (por cierto, no es fácil definir matemáticamente qué significa *aleatorio*). Seguramente muchos informáticos se sorprenderán al saber que los algoritmos más comunes son extremadamente simples y prácticamente no han cambiado desde los primeros ordenadores (aunque parecen empezar a estar en declive frente a los llamados *Mersenne twisters*). Por ejemplo, el ZX-Spectrum de principios de los 80 (que ahora consideraríamos poco más que una calculadora programable conectable a un televisor) simplemente multiplicaba por 75 módulo $65537 = 2^{16}+1$. Es decir, efectuaba $x_{n+1} = 75x_n \pmod{65537}$ y la n-ésima vez que se pedía un número aleatorio mostraba $(x_n - 1)/65537$. El x_0 dependía del tiempo de inicio. En C++11, la versión de C++ aprobada en 2011, el generador `minstd_rand` simplemente cambia 75 por 48271 y 65537 por $2^{31} - 1$. Es justo añadir que C++11 también tiene implementados otros algoritmos más complejos.

En el primer caso, $x_n = 75^n x_0 \pmod{65537}$. Se puede probar que 75 genera $\mathbb{Z}^*_{2^{16}+1}$ y por tanto tardaremos 2^{16} en ver el mismo número. La aplicación $x \mapsto 75^x$ descoloca "mucho" los elementos de $\mathbb{Z}_{2^{16}}$ en $\mathbb{Z}^*_{2^{16}+1}$ y eso produce la ilusión de un resultado aleatorio. Por supuesto, examinando con un poco de cuidado los presuntos números aleatorios obtenidos, es fácil percatarse de que sólo se está multiplicando por 75. En las aplicaciones en criptografía es importante no poder prever el siguiente número. Con este propósito, el *algoritmo Blum-Blum-Shub* (por el nombre de sus autores) simplemente utiliza $x_{n+1} = x_n^2 \pmod{m}$ donde m es un producto de dos primos muy grandes.

Una vez que se sabe simular una distribución uniforme, se puede simular una normal con el *método de Box-Muller* que no es más que el sencillo teorema o ejercicio que afirma que si $U_1 \sim U(0,1)$ y $U_2 \sim U(0,1)$ son independientes, entonces $\sqrt{-2\log U_1}\cos(2\pi U_2) \sim N(0,1)$. Además también se tiene que $\sqrt{-2\log U_1}\sin(2\pi U_2) \sim N(0,1)$ y es independiente de la anterior (el *método de Marsaglia* es una pequeña modificación un poco más eficiente desde el punto de vista computacional).

Para generar muestras de otras distribuciones a partir de uniformes, hay varios métodos [RK08, Ch.2]. Uno de ellos es el *algoritmo de Metropolis-Hastings* en el que la idea matemática subyacente es buscar una cadena de Markov cuya distribución estacionaria aproxime a la deseada. Aunque este algoritmo es particularmente

ventajoso y popular cuando el número de dimensiones de la distribución a generar es grande, veremos aquí sólo una versión simplificada del caso unidimensional (pero fácilmente generalizable).

Forcemos un poco la definición de cadena de Markov admitiendo $\mathbb{R}$ como conjunto de estados y digamos que queremos generar una muestra de una distribución con función de densidad conocida $f : \mathbb{R} \longrightarrow \mathbb{R}$. A partir del n-ésimo valor de la muestra $x_n = x$ tomamos un posible siguiente valor y de una distribución $N(x_n, \sigma)$ y suponemos una "matriz de transición" con $p_{xy} = \text{mín}\left(f(y)/f(x), 1\right)$ y $p_{xx} = 1 - p_{xy}$, es decir,

$$x_{n+1} = \begin{cases} y & \text{con probabilidad } p_{xy} \\ x & \text{con probabilidad } 1 - p_{xy} \end{cases} \tag{18}$$

Los primeros valores de x_n habitualmente se desprecian. En la jerga se dice que son el *burn-in* (rodaje), para que la distribución inicial que impone el valor de x_0 no tenga influencia. La elección de σ afecta seriamente al rendimiento. En buenas condiciones σ^2 debiera ser comparable a la varianza de f. Si σ^2 es mucho mayor, p_{xy} será habitualmente pequeño y repetirémos muchos términos. Por otro lado, si σ^2 es muy pequeño p_{xy} será habitualmente próximo a 1 y x_{n+1} cercano a x_n, de modo que la muestra no estará bien "mezclada".

Referencias

[Aus06] D Austin. How google finds your needle in the web's haystack. `http://www.ams.org/featurecolumn/archive/pagerank.html`, 2006.

[Doo53] J. L. Doob. *Stochastic processes*. John Wiley & Sons, Inc., New York; Chapman & Hall, Limited, London, 1953.

[Dur99] R. Durrett. *Essentials of stochastic processes*. Springer Texts in Statistics. Springer-Verlag, New York, 1999.

[FG04] P. Fernández Gallardo. El secreto de Google y el álgebra lineal. *Bol. Soc. Esp. Mat. Apl. SẽMA*, (30):115–141, 2004.

[HPS72] P. G. Hoel, S. C. Port, and C. J. Stone. *Introduction to stochastic processes*. Houghton Mifflin Co., Boston, Mass., 1972. The Houghton Mifflin Series in Statistics.

[Kac66] M. Kac. Can one hear the shape of a drum? *Amer. Math. Monthly*, 73(4, part II):1–23, 1966.

[KS76] J. G. Kemeny and J. L. Snell. *Finite Markov chains*. Springer-Verlag, New York-Heidelberg, 1976. Reprinting of the 1960 original, Undergraduate Texts in Mathematics.

[Lax07] Peter D. Lax. *Linear algebra and its applications*. Pure and Applied Mathematics (Hoboken). Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2007.

[LM12] A. N. Langville and C. D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2012.

[PK10] M. A. Pinsky and S. Karlin. *An Introduction to Stochastic Modeling*. Academic Press, fourth edition, 2010.

[RK08] R. Y. Rubinstein and D. P. Kroese. *Simulation and the Monte Carlo method*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2008.