UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Matemática

Tesis de Licenciatura

Secuencias Poisson genéricas

con una introducción a las secuencias infinitas de De Bruijn

Gabriel Sac Himelfarb

Directora: Dra. Verónica Becher

Fecha de Presentación: 19 de diciembre de 2022

Resumen

Años atrás, Zeev Rudnick definió las secuencias λ- Poisson genéricas como aquellas secuencias infinitas de símbolos en un alfabeto finito, que satisfacen que el número de ocurrencias de palabras largas en segmentos iniciales sigue la distribución de Poisson de parámetro λ. Aunque se sabe que casi todas las secuencias son Poisson genéricas con respecto a la medida uniforme, ningún ejemplo explícito había sido dado hasta el momento.

En esta tesis presentamos una construcción explícita de una secuencia λ- Poisson genérica sobre cualquier alfabeto y para cualquier λ real positivo, excepto en el caso del alfabeto binario, en que se requiere que $\lambda \leq \ln(2)$.

Dado que λ- Poisson genericidad implica normalidad de Borel, las secuencias construidas son Borel normales. Probamos también que la misma construcción instanciada con parámetros diferentes permite obtener secuencias Borel normales que no son λ-Poisson genéricas.

La construcción utiliza las llamadas secuencias infinitas de De Bruijn. Con el objetivo de resolver el caso de alfabetos binarios completamente, definimos las secuencias infinitas cuasi De Bruijn.

Índice general

Agradecimientos

En primer lugar quisiera agradecer a Verónica por haber dirigido esta tesis y mi Beca Estímulo a las Vocaciones Científicas CIN. Por haber confiado en mí, por haberme sugerido hacerla y haberme mostrado el lindo problema sobre el que trabajé. Por transmitir su pasión, su energía y su alegría en cada una de sus clases y en cada una de las reuniones que tuvimos. ¡Gracias!

Quisiera agradecer a Inés y Martín por aceptar formar parte del jurado y por leer esta tesis. También a Inés por escuchar mi presentación en la UMA, y a Martín por escucharla en las reuniones del grupo KAPOW.

Quisiera agradecerle también a mi familia: a mis papás, a mi abuela, a mi hermana. Esto no habría sido posible sin su apoyo constante a lo largo de todos estos años. Gracias por haberme inculcado el placer por la lectura y por haberme introducido con aquellos primeros libros de Paenza y de la colección Ciencia que Ladra al mundo de la matemática y de la ciencia en general. Por transmitirme el valor de la constancia y la dedicación, y enseñareme que no hay nada que surja por generación espontánea, sino que todo se obtiene a través del trabajo y el esfuerzo.

Quiero agradecerle a todas las hermosas personas con las que tuve el placer de compartir este camino. Por haber hecho del CBC y del inicio de la vida universitaria una experiencia hermosa: a Yani, Miki, Gastón, Ezequiel, Leila, Jose, Bauti, Majo, Gabi, Nadir. A mis amigos y compañeros de cursada en mate: Nico, Chino Z., Chino C., Rama, Nicole, Teo, Lola, Yuri, Leo, Marian, Jan, Mati S., Mati A., Fran, Pablo, Mónica, Gonzalo. A los chicos de KAPOW: Ivo, Charles e Ignacio. A mis amigos y compañeros de compu: Sol, Tomi, Franco, Lucas, Felipe, Diego, Kaiel, Lucía. Também à turma de português: Emi, Dali, Juan,... Y por supuesto, a Carol.

También a Iván, Ilan, Mateo, Franco, Fran, Ann, Cami, Curzel, David.

Quiero agradecerle también a Expedición Ciencia, y especialmente a todas las personas que conocí en los campamentos y que lo volvieron una experiencia inolvidable: Juli, Juampi, Cande, Cris, Wallace, Lau, Melu, Pedro Beck, Gus Vassen, Petu, Jose, Calvi, Pira, y muchos muchos más que lamentablemente no puedo colocar por falta de espacio. A todos, ¡gracias!

También quiero agradecer a la Olimpíada Matemática Argentina y sus organizadores, sin la cual posiblemente hubiera estudiado algo totalmente distinto. Por mostrar

en cada una de sus competencias lo hermosa que puede ser la matemática. Y también quiero agradecer a toda la gente increíble que conocí gracias a la OMA: Agus B., Agus M., Lu, Mati, Carla, Joa, Joaco, Vero, Pablo, Santi, Esca, Turko, Mono, Dido, Fran, Gianni, Mateo. Al equipo de Mateclubes por permitirme formar parte de este proyecto tan lindo.

A todos los docentes del DM y del DC que de alguna forma u otra me marcaron y me guiaron en este camino.

A la gente hermosa que conocí en Filo: Agos, Palo, Damián, José, Juani, Ivana,... A Eduardo Barrio, Paula Teijeiro y Bruno da Ré por hacer de Introducción a las Lógicas no Clásicas una hermosa materia y un bello desafío.

A los chicos de ESSLLI que conocí virtualmente: Anna, Marko, Wei, Alessandro, Zeinab.

A mis compañeros y docentes de Musizap: Sebastían, Belén, Natalia (y hay muchos nombres más!). A Raúl Becerra y Gabriel Romero.

A todos, ¡gracias!

Introducción

Si le presentamos a cualquier persona la siguiente secuencia de ceros y unos

$$0\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ldots,$$

y le preguntamos si cree que es aleatoria, casi seguramente dirá que no. ¿Y qué hay de la siguiente secuencia?

$$0\ 1\ 0\ 0\ 0\ 1\ 0\ 1\ 1\ 0\ 1\ 1\ 1\ 0\ 0\ 1\ 0\ 1\ 0\ 0\ldots$$

Aquí puede que dude, pero probablemente dirá que esta parece más aleatoria que la anterior.

El azar es una de aquellas (tantas) cosas que atraviesa nuestras vidas, y a la que hacemos alusión de forma permanente, pero de la cual no solemos tener más que un conocimiento intuitivo (¡incluso muchas veces errado!). Si pensamos en secuencias aleatorias, viene a nuestra mente una secuencia de tiradas de un dado, o de una moneda por ejemplo. Pero dada una secuencia particular, ¿qué significaría que esa secuencia sea aleatoria? Enfrentado a esa pregunta, Borel dio en el año 1909 ([9]) la definición de *normalidad*, la primera noción de aleatoriedad sobre secuencias.

Una secuencia infinita de ceros y unos se dice normal si en el límite, la proporción de ceros y unos converge a $\frac{1}{2}$, las proporciones de 00, 01, 10, 11 convergen a $\frac{1}{4}$, ..., y más en general, si para todo k, la proporción de veces que aparece una secuencia específica de k dígitos converge a $\frac{1}{2^k}$. Borel demostró que casi todas las secuencias infinitas de ceros y unos son normales (respecto a la medida de Lebesgue, si pensamos a las secuencias como el desarrollo binario de reales en $(0,1)$), y planteó el problema de dar un ejemplo concreto. Hubo que esperar al año 1933 para que Champernowne diese la primera construcción explícita de una secuencia normal (ver [10]). La secuencia de Champernowne en base 2 consiste en concatenar los desarrrollos en base dos de los números 0, 1, 2, ...[1]

Intuitivamente, la definición de Borel captura una propiedad necesaria para que una secuencia sea aleatoria. Sin embargo, no es suficiente: no llamaríamos aleatoria a

[1]Si bien Champernowne formuló originalmente su construcción en base 10, es posible reproducirla en cualquier otra base.

la secuencia de Champernowne, dado que es completamente predecible. Recién en los años 60, gracias a los trabajos de Kolmogorov, Chaitin, Martin-Löf, entre otros, los matemáticos dieron con una definición de aleatoriedad sobre secuencias infinitas que es robusta (se conocen muchas formulaciones equivalentes que no parecen guardar relación entre sí a priori).

Hace varios años, Zeev Rudnick dio una nueva noción de aleatoriedad sobre secuencias infinitas, a la que llamó inicialmente *supernormalidad*. Esta noción resulta ser estrictamente más fuerte que la normalidad de Borel y estrictamente más débil que la aleatoriedad de Chaitin, Martin-Löf y Kolmogorov. Con el paso del tiempo, estas secuencias pasaron a llamarse secuencias λ- Poisson genéricas y Poisson genéricas. Benjamin Weiss difundió esta definición en dos charlas [37] y [36] en los años 2010 y 2020 respectivamente. Junto con Yuval Peres, demostró que casi todas las secuencias son λ-Poisson genéricas para todo λ real positivo. La demostración fue transcripta por Álvarez, Becher y Mereb en [1]. Sin embargo, y tal como ocurrió con la noción de normalidad de Borel, nadie hasta el momento había logrado dar un ejemplo de secuencias λ-Poisson genéricas. La mayor contribución de esta tesis consiste en dar la primera construcción explícita de secuencias de este tipo.

La tesis está estructurada del siguiente modo:

Dado que la construcción utiliza fuertemente las llamadas *secuencias infinitas de De Bruijn*, en el Capítulo 1 damos una introducción a las secuencias y grafos de De Bruijn, junto con la demostración de la existencia de secuencias infinitas de De Bruijn.

En el Capítulo 2 presentamos la noción de λ- Poisson genericidad y Poisson genericidad. Dado que, a diferencia de la normalidad de Borel, no resulta inmediatamente clara su motivación, damos múltiples intuiciones que explican por qué se trata de una definición razonable.

En el Capítulo 3 presentamos la construcción de secuencias λ- Poisson genéricas. De hecho, damos una construcción más general, que instanciada en determinados valores de sus parámetros permite obtener las secuencias buscadas. Los resultados de este capítulo fueron presentados en dos charlas, tituladas *A construction of a λ- Poisson generic sequence*. La primera tuvo lugar virtualmente en el marco de la 19th International Conference on Computability and Complexity in Analysis (CCA2022), entre los días 23 y 26 de mayo de 2022. La segunda, en la sesión de Ecuaciones Diferenciales y Probabilidad de la LXXI Reunión de Comunicaciones Científicas de la Reunión Anual de la UMA 2022. La misma se llevó a cabo en la Universidad Nacional del Comahue entre los días 20 y 23 de septiembre de 2022.

En el Capítulo 4, damos un criterio de normalidad cuya demostración es una pequeña generalización de la demostración de Peres y Weiss de que λ-Poisson genericidad implica normalidad. Esta permite probar que todas las secuencias construidas en el Capítulo 3 son normales. Los resultados de estos dos capítulos fueron reunidos en el artículo [7], que fue recientemente aceptado para su publicación en Mathematics of Computation, de la American Mathematical Society.

El Capítulo 5 es el producto de una investigación en proceso. En muchas ocasiones sucede en matemática algo muy bello: una pregunta conduce naturalmente a otra nueva. Como se explica en el Capítulo 1, las secuencias infinitas de De Bruijn en alfabetos binarios no cumplen las mismas propiedades que en alfabetos de tres o más símbolos. Esto trae aparejado que la construcción que damos en el Capítulo 3 tenga una limitación para los posibles valores de λ si el alfabeto tiene dos símbolos. Motivados por la búsqueda de una solución que funcione para todos los valores de λ en alfabetos binarios, definimos las *secuencias infinitas cuasi De Bruijn*, y en el Capítulo 5 nos embarcamos en la búsqueda de una tal secuencia. Presentamos algunos resultados en esa dirección, e intentamos mostrar cuáles son las principales dificultades que surgen.

Un poco de notación

A continuación presentamos la notación que es empleada a lo largo de este trabajo. Dado Ω un alfabeto de b símbolos, $b \geq 2$,

- Denotamos por $\Omega^{\mathbb{N}}$ al conjunto de secuencias infinitas de símbolos de Ω.
- Llamamos *palabras* a las secuencias finitas de símbolos de Ω, y Ω^k es el conjunto de palabras de longitud k.
- Numeramos las posiciones de las palabras y secuencias infinitas comenzando desde 1, y denotamos $w[i]$ a la subsecuencia de w que comienza en la posición i y que termina en la posición j.
- Si w es una palabra, llamamos $|w|$ a su longitud.
- Dadas dos palabras w y v, el número de ocurrencias de w en v es:

  $$|v|_w = \#\{i : v[i \cdots + |w| - 1] = w\}.$$

  Por ejemplo, $|0001|_{00} = 2$.
- Dada una palabra w de longitud b^k, decimos que δ es un *bloque* en w si es una subsecuencia de w y $|\delta| = b^j \leq b^k$ para algún $j \in \mathbb{N}_0$.
- Decimos que un bloque δ en w tiene *longitud absoluta* $|\delta|$ y *longitud relativa* $|\delta| b^{-k}$ *con respecto a* w.
- Dado un número real $y \in [0, 1)$, denotamos por $\{y\}_k$ al truncamiento a k dígitos del único desarrollo en base b de y que no termina en una cola infinita de $(b-1)$. En el caso de $y = 1$ tomamos la representación $\sum_{i \geq 1} (b-1) b^{-i}$.

Capítulo 1

Secuencias infinitas de De Bruijn

¿Tiene algo de especial la siguiente secuencia de ceros y unos?

0 0 0 1 1 1 0 1 0 0

¿Y la siguiente secuencia de letras A, G, C, T?

A A T T C T G T A C C G C A G G A

Si observamos con atención, en la primera de las dos secuencias, todas las palabras de tres dígitos en el alfabeto $\{0, 1\}$ aparecen exactamente una vez, y en la segunda, todas las palabras de dos letras en el alfabeto {A, G, C, T} aparecen también exactamente una vez.

El uso de palabras de estas características se remonta mucho tiempo atrás: por ejemplo, Fredricksen relata en [17] que percusionistas indios utilizaban hace más de mil años la palabra mnemotécnica *yamátárájahánsalagám*, que contiene entre sus diez sílabas todas las posibles combinaciones de tres golpes largos o cortos (indicados por las sílabas acentuadas y no acentuadas respectivamente).

Los ejemplos anteriores motivan la siguiente definición:

Definición 1. *Dado un alfabeto finito Ω de b símbolos, decimos que $x \in \Omega^{b^k+k-1}$ es una palabra de De Bruijn de orden k si cada palabra de longitud k aparece exactamente una vez en x. Decimos que $x \in \Omega^{b^k}$ es una palabra cícilica de De Bruijn de orden k si cada palabra de longitud k aparece exactamente una vez en la palabra circular determinada por x*[1].

Estas palabras fueron descubiertas y redescubiertas en numerosas ocasiones en los últimos ciento cincuenta años: Flye Sainte-Marie respondió en 1894 una pregunta de A. de Rivière al mostrar la existencia de palabras de De Bruijn en base 2 de cualquier

[1]Por ejemplo, en la palabra circular determinada por 0110, aparecen las palabras de longitud dos 01, 11, 10 y 00.

orden, e incluso probó que existen $2^{2^{n-1}-n}$ de tales palabras. En 1934, Monroe Martin propuso por primera vez la generación de palabras de De Bruijn lexicográficamente mínimas a través de un algoritmo *greedy*. En 1943, y de forma independiente, el problema fue planteado nuevamente por Klaas Posthumus. En 1944, Nicolaas Govert de Bruijn estudió el problema y redescubrió los resultados de Sainte-Marie (ver [12]). Es él, por supuesto, de quien tomaron estas famosas palabras su nombre. Good y Korobov también resolvieron el problema en los años 1946 y 1950, respectivamente (ver [19] y [23]). Para un desarrollo histórico más detallado de este tema, el lector puede consultar [8].

Nuestro primer objetivo es reproducir una demostración de la existencia de palabras de De Bruijn (cíclicas y no cíclicas) de cualquier orden.

1.1. Preliminares

En esta sección introducimos las definiciones y resultados básicos sobre grafos que empleamos en el resto del capítulo.

Definición. *Dado un grafo dirigido $G = (V, E)$,*

- *decimos que G es euleriano si existe un ciclo dirigido que recorre todas sus aristas pasando por cada una exactamente una vez;*
- *decimos que G es hamiltoniano si existe un ciclo dirigido que recorre todos sus vértices pasando por cada uno exactamente una vez;*
- *decimos que G es conexo si el grafo no dirigido subyacente es conexo;*
- *decimos que G es fuertemente conexo si para cada par de vértices existe un camino dirigido que los une;*

Los siguientes dos lemas, que enunciamos sin demostración, proporcionan una caracterización de los grafos dirigidos eulerianos. El primero es un resultado clásico que puede consultarse por ejemplo en [20]. El segundo es parte del *folklore* del área, y una demostración puede leerse en [6].

Lema 1. *Un grafo dirigido es euleriano si y solo si es fuertemente conexo, y para cada vértice, su grado de salida es igual a su grado de entrada.*

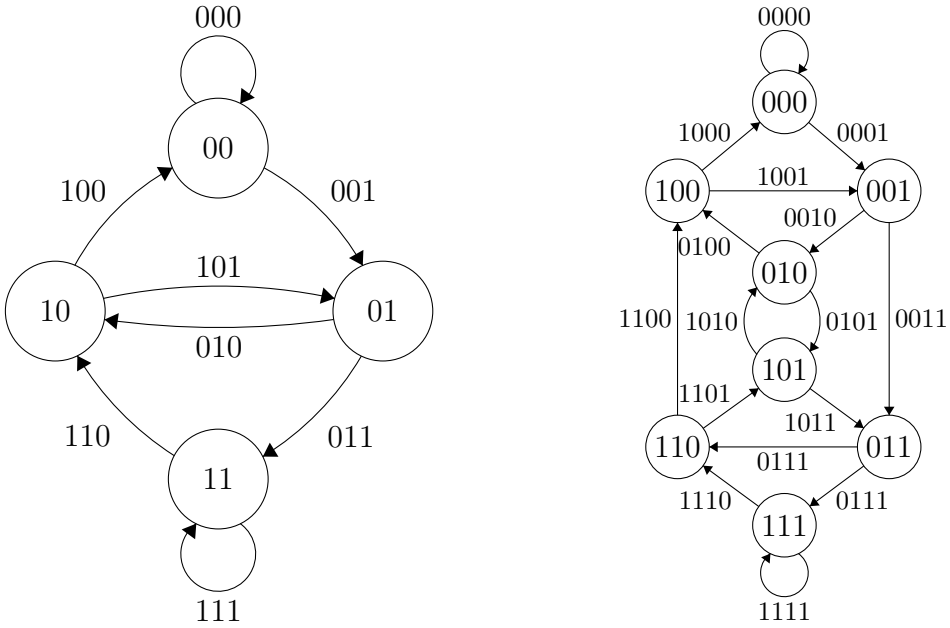
Lema 2. *Un grafo dirigido en que todo vértice tiene el mismo grado de salida que de entrada, es fuertemente conexo si y solo si su grafo subyacente es conexo.*

1.2. Grafos de De Bruijn

Para demostrar la existencia de palabras de De Bruijn de órdenes arbitrarios sobre cualquier alfabeto, resultan especialmente útiles los llamados grafos de De Bruijn, que definimos a continuación.

Definición. *Fijado Ω un alfabeto de b símbolos, el digrafo de De Bruijn de orden n, $G_n = (V_n, E_n)$, se construye del siguiente modo:*

- *V_n, el conjunto de vértices, tiene b^n elementos, cada uno etiquetado con una palabra distinta $x_1 \dots x_n \in \Omega^n$ de n caracteres.*
- *E_n, el conjunto de aristas, consta de b^{n+1} elementos. Por cada par de vértices $v_1 = x_1x_2 \dots x_n$, $v_2 = x_2 \dots x_{n+1}$, hay una arista que los une, y que se etiqueta con $x_1x_2 \dots x_nx_{n+1}$.*



(a) Orden 2 (b) Orden 3

Figura 1.1: Grafos de De Bruijn, alfabeto $\{0, 1\}$

Observación. *Observar que G_n tiene b^n vértices (uno por cada palabra de longitud n en Ω) y b^{n+1} aristas (uno por cada palabra de longitud $n+1$ en Ω). Además, G_n es b-regular, es decir, todos los vértices tienen grado de salida y de entrada igual a b. De cada vértice salen b aristas, uno por cada elemento del alfabeto. Notar además que G_n es fuertemente conexo. Esto puede probarse directamente (sin recurrir necesariamente*

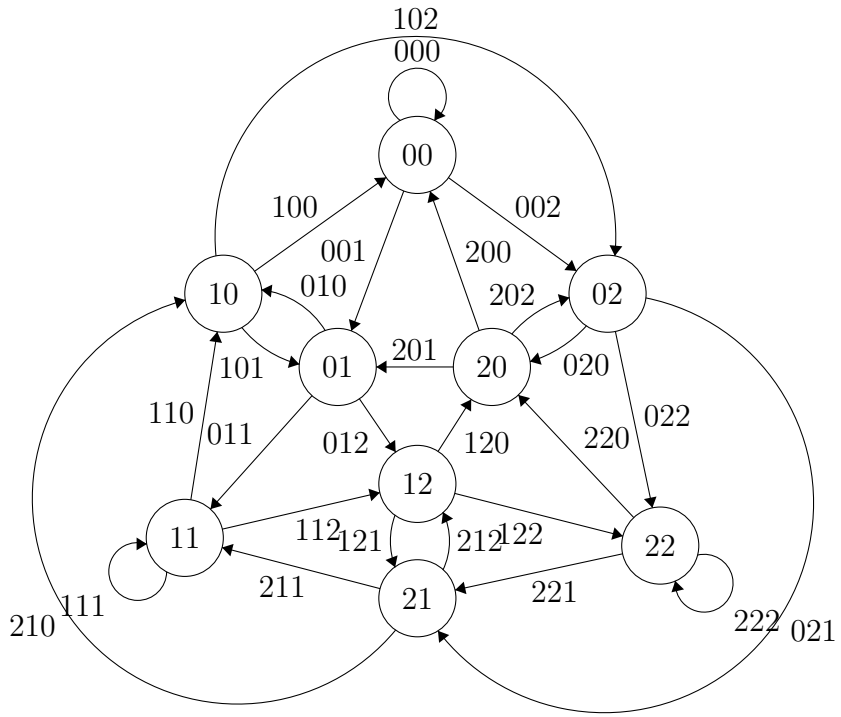


Figura 1.2: Grafo de De Bruijn de orden 2, alfabeto $\{0, 1, 2\}$

al Lema 2) observando que si $x_1 \dots x_n$, $y_1 \dots y_n$ son dos vértices arbitrarios, el camino formado por las aristas $x_1 \dots x_n y_1$, $x_2 \dots x_n y_1 y_2$, $\dots$,$x_n y_1 \dots y_n$ une los dos vértices.

A modo de ejemplo, en la Figura 1.1 pueden observarse los grafos de De Bruijn de orden 2 y 3 para un alfabeto de dos símbolos, y en la Figura 1.2, el grafo de De Bruijn de orden 2 para un alfabeto de tres símbolos.

Los grafos de De Bruijn satisfacen una relación recursiva especialmente útil involucrando sus grafos de línea.

Definición. *Dado un grafo dirigido $G = (V, E)$, su grafo de línea $L(G) = (V', E')$ se construye de la siguiente forma:*

- *Cada vértice en V' representa una arista de G.*
- *Dados $e, f \in V'$, hay una arista de e a f en G' si y solo si las aristas correspondientes e y f en G forman un camino dirigido de longitud 2.*

Observación. *Un ciclo en un grafo G determina un ciclo de la misma longitud en su grafo de línea $L(G)$: si el ciclo en G está formado por las aristas $e_1, e_2, \dots e_k, e_1$, tomamos el ciclo de vértices $e_1, e_2, \dots, e_k, e_1$ en $L(G)$. En la Figura 1.3 puede observarse un ejemplo de un grafo G y su grafo de línea $L(G)$, en que se marca un ciclo distinguido en G y el ciclo que determina en $L(G)$.*

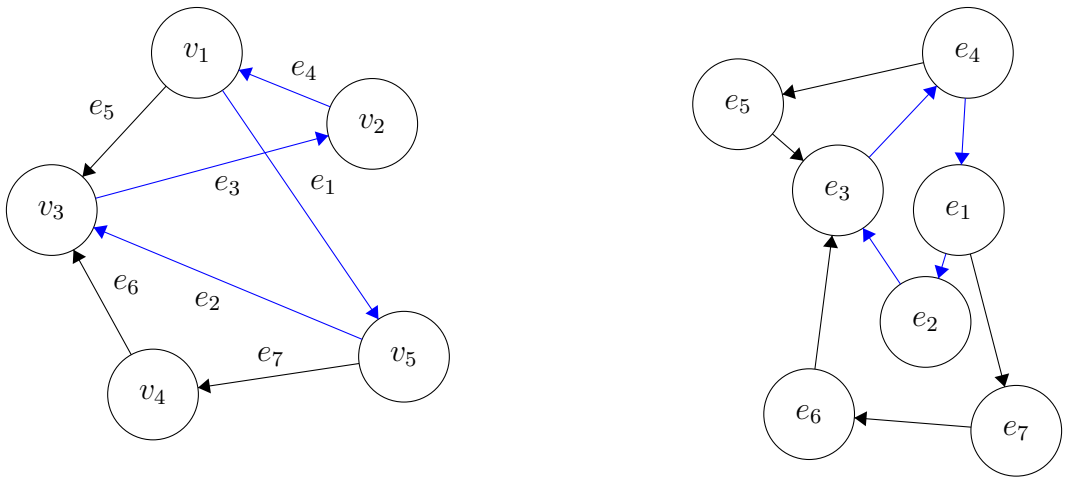


Figura 1.3: Izquierda: G. Derecha: L(G). En azul un ciclo en G y su ciclo correspondiente en $L(G)$.

Los grafos de De Bruijn satisfacen la siguiente propiedad:

Lema 3. *Fijado un alfabeto Ω, los grafos de De Bruijn sobre Ω satisfacen la relación*

$$L(G_n) = G_{n+1}.$$

Demostración. Para verificar que efectivamente $L(G_n) = G_{n+1}$, basta con observar que por definición, cada vértice de G_{n+1} se identifica a través de su etiqueta con una arista de G_n. Además, si dos aristas e_1 y e_2 en G_n forman un camino de longitud dos, existen vértices v_1, v_2 y v_3 de forma que e_1 une v_1 con v_2 y e_2 une v_2 con v_3. Si los vértices están etiquetados con $v_1 = x_1x_2\dots x_n$, $v_2 = x_2\dots x_{n+1}$, $v_3 = x_3\dots x_{n+2}$, entonces $e_1 = x_1\dots x_nx_{n+1}$ y $e_2 = x_2\dots x_{n+2}$. Esto significa que en G_{n+1}, los vértices correspondientes a e_1 y e_2 están unidos por una arista $x_1\dots x_nx_{n+1}x_{n+2}$, como queríamos. Puede verificarse fácilmente que también vale a la inversa: si dos vértices en G_{n+1} están unidos por una arista, entonces sus aristas correspondientes en G_n forman un camino de longitud dos. $\square$

A continuación establecemos una relación entre las palabras de De Bruijn, y ciclos hamiltonianos y eulerianos en los grafos de De Bruijn.

Observación. *Fijemos Ω un alfabeto. Dado un camino cerrado $v_1, \dots, v_k, v_1$ en el grafo G_n, podemos etiquetarlo con $v_1[1]v_2[1]\dots v_k[1]$, es decir, concatenar los primeros símbolos de las etiquetas de cada vértice.*

No resulta difícil demostrar el siguiente lema:

Lema. *Dado un ciclo hamiltoniano en G_n, y fijado cualquiera de los vértices v_1, la etiqueta del ciclo empezando desde v_1 es una palabra circular de De Bruijn de orden n. Si copiamos al final de la etiqueta los primeros $n-1$ símbolos, obtenemos una secuencia de De Bruijn de orden n.*

Por lo tanto, el problema de hallar secuencias de De Bruijn de órdenes arbitrarios puede traducirse a encontrar ciclos hamiltonianos en los grafos G_n. En general, el problema de determinar si un grafo tiene algún ciclo hamiltoniano es computacionalmente difícil: es NP-completo. Sin embargo, para el caso de los grafos de De Bruijn, el Lema 3 brinda una forma simple de demostrar que G_n es hamiltoniano para todo n.

Lema. *Fijemos Ω un alfabeto. Entonces los grafos de De Bruijn G_n son hamiltonianos para todo n.*

Demostración. Puede verificarse fácilmente a mano que G_1 es hamiltoniano. Por otro lado, notar que un ciclo hamiltoniano en G_{n+1} ($n \geq 1$) se corresponde con un ciclo euleriano en G_n, ya que al ser $L(G_n) = G_{n+1}$, podemos identificar un camino en G_{n+1} que recorre todos sus vértices con un camino en G_n que recorre todas sus aristas.

Como G_n es fuertemente conexo, y para cada vértice el grado de entrada es igual al de salida, por el Lema 1 resulta que es euleriano, y por la observación del párrafo anterior, resulta que G_{n+1} es hamiltoniano. □

Finalmente, de los dos lemas anteriores se deduce:

Teorema (De Bruijn [12])**.** *Fijado Ω un alfabeto arbitrario, existen palabras de De Bruijn (circulares y no circulares) de cualquier orden sobre ese alfabeto.*

1.3. Construcción de secuencias infinitas de De Bruijn

Dado Ω un alfabeto, ¿es posible construir una secuencia $x \in \Omega^{\mathbb{N}}$ que satisfaga que todo prefijo de longitud b^k, con $k \in \mathbb{N}$, es una secuencia de De Bruijn?

En esta sección mostramos que en el caso de alfabetos de 3 o más símbolos, la respuesta a la pregunta anterior es afirmativa. Probamos también que en el caso de un alfabeto de 2 símbolos no existe ninguna secuencia que satisfaga la condición previa, e introducimos secuencias que satisfacen una variante de la propiedad. Seguimos esencialmente la presentación de Becher y Heiber en [6], aunque profundizamos en algunos detalles omitidos en ese trabajo.

Comenzamos por los siguientes dos lemas:

Lema 4 ([6, Lemma 3])**.** *Un ciclo hamiltoniano en un grafo de De Bruijn G_n sobre un alfabeto de 3 o más símbolos puede extenderse siempre a un ciclo euleriano en el mismo grafo.*

Demostración. Sea H un ciclo hamiltoniano en G_n. Sea C el grafo que resulta de eliminar las aristas de H de G_n. Veamos que C es euleriano. En tal caso, basta con agregar a H un ciclo euleriano de C para obtener la extensión deseada.

Para ver que C es euleriano, notemos en primer lugar que cada vértice tiene grado de salida y de entrada igual a $b-1$. Por los Lemas 1 y 2 bastaría con probar que el grafo subyacente a C es conexo. Veámoslo.

Dados u y v dos vértices arbitrarios, definimos recursivamente una secuencia de pares de vértices u_i, v_i, para $0 \leq i \leq n$, de forma que:

1. $u = u_0$ y $v = v_0$.
2. Para cada $i < n$, hay una arista de u_i a u_{i+1} en C, y análogamente para v_i.
3. Los últimos i símbolos de u_i y v_i coinciden.

Comenzamos con $u_0 = u$ y $v_0 = v$. Si ya construimos u_j y v_j para $j \leq i$, veamos cómo construir u_{i+1} y v_{i+1}. Notemos que u_i tiene b sucesores en G_n: $u_i[2\ldots n]0$, $u_i[2\ldots n]1$, $\ldots$ y $u_i[2\ldots n](b-1)$. Análogamente para v_i. Como H utiliza exactamente una de las aristas que sale de u_i y una de las aristas que sale de v_i, y como $b \geq 3$, existe un símbolo a_{i+1} en Ω tal que las aristas de u_i a $u_{i+1} = u_i[2\ldots n]a_{i+1}$ y de v_i a $v_{i+1} = v_i[2\ldots n]a_{i+1}$ están ambas en C. Además, si los últimos $i < n$ símbolos de u_i y v_i coincidían, ahora coinciden los últimos $i+1$ símbolos (los i que coincidían antes, más el símbolo a_{i+1}). Por la condición 3, $v_n = u_n$, y por lo tanto u y v están conectados por un camino en el grafo no dirigido subyacente a C. Como esto vale para dos vértices arbitrarios u y v, se obtiene que el grafo subyacente a C es conexo, como queríamos. □

Lema 5 ([6, Lemma 5])**.** *Un ciclo hamiltoniano en un grafo de De Bruijn G_n sobre un alfabeto de 2 símbolos puede extenderse siempre a un ciclo euleriano en el grafo de De Bruijn de orden siguiente G_{n+1}.*

Demostración. Sea H un ciclo hamiltoniano en G_n. Como G_{n+1} es el grafo de línea de G_n, H determina un ciclo simple $\widetilde{H}$ en G_{n+1}. Observar que $\widetilde{H}$ pasa por la mitad de los vértices de G_{n+1}.

Sea C el grafo que resulta de eliminar las aristas de $\widetilde{H}$ de G_{n+1}. Veamos que C es euleriano. De esa forma, la extensión buscada puede obtenerse concatenando a $\widetilde{H}$ un ciclo euleriano de C. Notar que los vértices de C tienen o bien grado de entrada y salida igual a 2 (aquellos vértices no usados por $\widetilde{H}$), o bien grado de entrada y salida igual a 1. Por lo tanto, por los Lemas 1 y 2 bastará ver que C es conexo. Probemos en primer lugar la siguiente afirmación:

Afirmación: Todo vértice en G_{n+1} tiene un sucesor que está en $\widetilde{H}$ y un sucesor que no lo está.

Prueba: Un vértice v en G_{n+1} tiene dos sucesores: $v[2\ldots n+1]0$ y $v[2\ldots n+1]1$. En G_n, $v[2\ldots n+1]$ se corresponde con un vértice, y $v[2\ldots n+1]0$ y $v[2\ldots n+1]1$ son las

dos aristas salientes. Como H es hamiltoniano en G_n, solo utiliza una de las dos aristas. Eso significa que uno de los dos vértices de G_{n+1}, $v[2 \dots n+1]0$ y $v[2 \dots n+1]1$, estará en $\widetilde{H}$, y el otro no.

Ahora, dados dos vértices u y v en G_{n+1}, construimos una secuencia de pares u_i, v_i, con $0 \leq i \leq n+1$, de modo que:

1. Para cada i, o bien u_i, o bien v_i, no está en $\widetilde{H}$.
2. Hay una arista de u a u_0 en C y una arista de v a v_0 en C[2].
3. Para cada $i \leq n$, hay una arista de u_i a u_{i+1} en C. Análogamente para v_i.
4. Los últimos i símbolos de u_i y v_i coinciden.

Sea u_0 el sucesor de u que no está en $\widetilde{H}$ y v_0 el sucesor de v que no está en $\widetilde{H}$. Si ya elegimos $u_0, u_1, \dots u_i$ y $v_0, v_1, \dots, v_i$, veamos cómo elegir u_{i+1} y v_{i+1}. Si $v_i \notin \widetilde{H}$, sea a_{i+1} tal que $u_i[2 \dots n+1]a_{i+1}$ no está en $\widetilde{H}$. Si $u_i \in \widetilde{H}$, por la condición 1., $v_i \notin \widetilde{H}$. Tomemos en ese caso a_{i+1} tal que $v_{i+1} = v[2 \dots n+1]a_{i+1}$ no está en $\widetilde{H}$. Por definición, alguno de los extremos de la arista de u_i a u_{i+1}, no está en $\widetilde{H}$, y por lo tanto, la arista está en C. Y lo mismo ocurre para la arista de v_i a v_{i+1}. Además, si u_i y v_i compartían los últimos i símbolos, por construcción u_{i+1} y v_{i+1} comparten sus últimos $i+1$ dígitos.

Por la condición 4., resulta que $u_{n+1} = v_{n+1}$. Por lo tanto, u y v están conectados en el grafo no dirigido subyacente a C. Como esto vale para cualesquiera dos vértices u y v, concluimos que el grafo subyacente a C es conexo. □

Definición 2. *Dado un ciclo hamiltoniano H en G_n, llamamos* ***grafo remanente*** *(de H) al grafo que resulta de eliminar las aristas de H en G_n.*

Observación 1. ***¿Por qué falla el Lema 4 en base*** 2***?*** *Esencialmente, porque el grafo remanente G de H en G_n nunca es conexo: notar que cualquier ciclo hamiltoniano en G_n utiliza las aristas de 10^{n-1} a 0^n, y de 0^n a $0^{n-1}1$, pero no utiliza el loop en 0^n. Por lo tanto, este loop estará aislado en G, y lo mismo ocurre con el loop en 1^n. Es decir, el grafo G tiene siempre como mínimo tres componentes conexas. Este fenómeno puede visualizarse en la Figura 1.4. Un estudio más detallado de las componentes conexas del grafo remanente puede encontrarse en el Capítulo 5.*

Mediante los Lemas 4 y 5, es posible demostrar el siguiente resultado, que usaremos en el Capítulo 3:

Lema 6 ([6, Theorem 1])**.** *Sea Ω un alfabeto de b símbolos.*

[2] En [6] falta asegurar la pertenencia a C de esas aristas

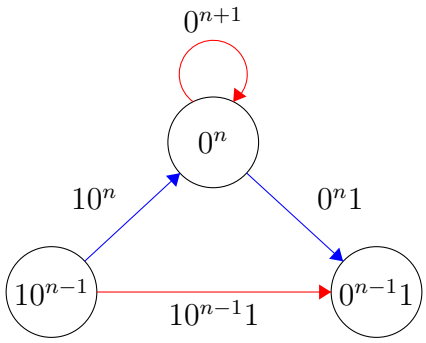


Figura 1.4: Fragmento de G_n en base 2. Hamiltoniano H en azul, grafo remanente G en rojo.

(1) Si $b \geq 3$, es posible construir una secuencia infinita x sobre Ω, que satisface que para todo k natural, $x[1 \dots b^k]$ es una secuencia cíclica de De Bruijn de orden k, y que $x[1 \dots b^k + k - 1]$ es una secuencia de De Bruijn de orden k.

(2) Si $b = 2$, es posible construir una secuencia infinita x sobre Ω que satisface que para todo k natural, $x[1 \dots b^{2k-1}]$ es una secuencia cíclica de De Bruijn de orden $2k - 1$, y $x[1 \dots b^{2k-1} + 2k - 2]$ es una secuencia de De Bruijn de orden $2k - 1$.

El resultado anterior motiva la definición siguiente:

Definición 3. *Decimos que $x \in \Omega^{\mathbb{N}}$ es una secuencia infinita de De Bruijn si satisface la condición (1) del Lema 6 cuando $b \geq 3$, o si satisface la condición (2) si $b = 2$.*

Ahora sí, demostramos el Lema 6.

Demostración. Usamos los Lemas 4 y 5 para construir recursivamente la secuencia buscada en cada caso.

Comenzamos con algún ciclo hamiltoniano H_1 en G_1, con etiqueta x_1, que es una secuencia de De Bruijn (circular y no circular) de orden 1. Decimos que $v_1 = 0$ es el vértice distinguido inicial.

Caso 1: Si $b \geq 3$. En este caso, después del paso k de la construcción, obtenemos un ciclo hamiltoniano H_k en G_k, con un vértice distinguido v_k y una palabra de De Bruijn circular x_k correspondiente a la etiqueta de H_k comenzando desde el vértice distinguido.

Dado H_k, por el Lema 4 se lo puede extender a un ciclo euleriano en G_k. Este ciclo euleriano a su vez se corresponde con un ciclo hamiltoniano en G_{k+1}, que será H_{k+1}. Consideremos en H_k la arista $v_k a_k$ (donde a_k es un caracter del alfabeto $\{0, 1, \dots b-1\}$) que parte de v_k. El vértice distinguido v_{k+1} de H_{k+1} será $v_k a_k$, es decir, el correspondiente a la arista de H_k que parte del vértice distinguido v_k. Finalmente, x_{k+1} será la etiqueta de H_{k+1} comenzando desde v_{k+1}.

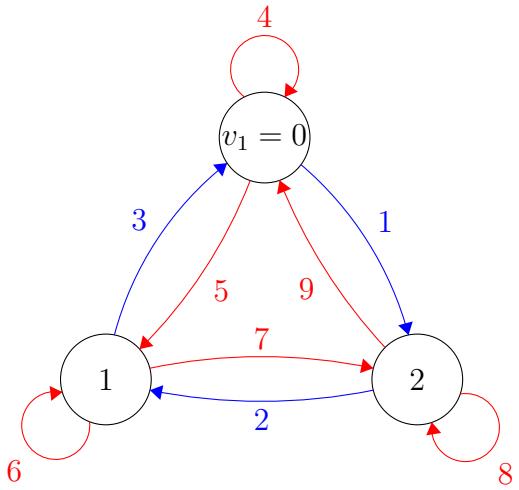
Por construcción, se sigue que x_k es una palabra de De Bruijn circular para todo k. Notemos que $x_{k+1}[1 \dots b^k] = x_k$, es decir, que x_{k+1} es una extensión de x_k. Esto ocurre porque la etiqueta de un camino cerrado en G_k es la misma que la del camino correspondiente en G_{k+1} cuando en este último se comienza desde el vértice correspondiente a la primera arista del camino en G_k. De esta forma, podemos definir una palabra infinita x cuyos prefijos sean exactamente los x_k. Además, $x[1 \dots b^k + k - 1]$ es una palabra de De Bruijn pues para todo k, los $k-1$ dígitos $x[b^k + 1 \dots b^k + k - 1]$ son iguales a los primeros $k-1$ dígitos $x[1 \dots k-1]$: si u_1, u_2,..., u_{b^k}, u_1 es el ciclo hamiltoniano H_k, el ciclo euleriano que es una extensión suya en G_k comienza de la misma forma, y por lo tanto, el ciclo hamiltoniano H_{k+1} en G_{k+1} comienza con $u_1u_2[1]$, $u_2u_3[1]$, ..., $u_{b^k}u_1[1]$. Notemos que como u_k se une a u_1 en G_k, $u_{b^k}[2 \dots k] = u_1[1 \dots k-1]$. Entonces, la etiqueta de H_{k+1} comienza con $u_1[1]u_2[1] \dots u_{b^k}[1] = u_1u_{k+1}[1]u_{k+2}[1] \dots u_{b^k}[1]$ y continúa con $u_{b^k}[2]u_1[1]$, es decir, $u_1[1 \dots k-1]u_1[1]$. Por lo tanto, los primeros $k-1$ dígitos de la continuación coinciden con los primeros.

En la Figura 1.5 puede observarse el inicio de una posible construcción para $b = 3$ y el alfabeto $\{0, 1, 2\}$.

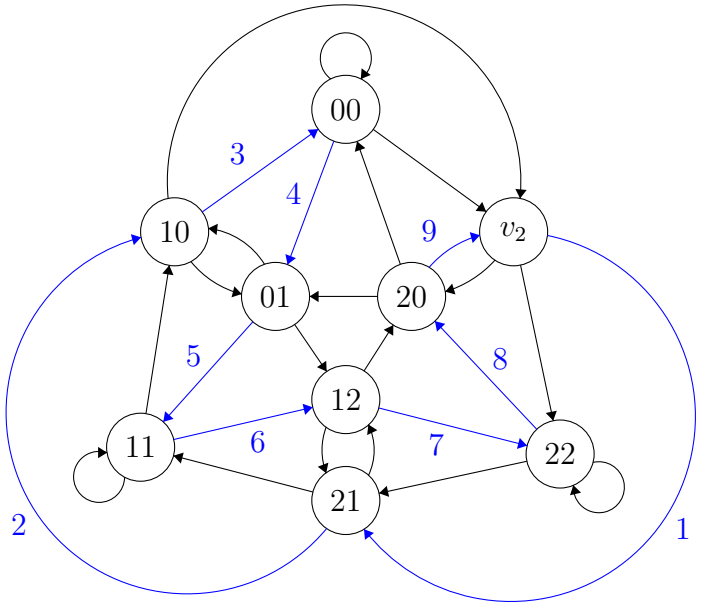
Caso 2: Si $b = 2$. En este caso, después del paso k de la construcción, obtenemos un ciclo hamiltoniano H_{2k-1} en G_{2k-1}, con un vértice distinguido v_{2k-1} y una palabra de De Bruijn circular x_{2k-1} correspondiente a la etiqueta de H_{2k-1} comenzando desde el vértice distinguido.

Dado H_{2k-1}, por el Lema 5 puede extenderse a un ciclo euleriano en G_{2k}. Este ciclo se corresponde a su vez con un ciclo hamiltoniano en G_{2k+1}. Este será el ciclo H_{2k+1}. En H_{2k-1}, del vértice distinguido v_{2k-1} sale una arista, digamos $v_{2k-1}a_{2k-1}$. En H_{2k-1}, esta arista tiene una arista sucesora, digamos $v_{2k-1}[2 \dots 2k-1]a_{2k-1}b_{2k-1}$. El vértice distinguido v_{2k+1} lo definimos como $v_{2k-1}a_{2k-1}b_{2k-1}$. Además, x_{2k+1} será la etiqueta de H_{2k+1} comenzando desde v_{2k+1}.

De forma análoga al caso anterior puede demostrarse que x_{2k-1} es prefijo de x_{2k+1} para todo k, y que la secuencia infinita x que tiene a los x_{2k-1} como prefijos, satisface las condiciones requeridas. $\square$



(a) H_1 en azul. En rojo una extensión a un euleriano. Los números de las aristas indican el orden de recorrido. $x_1 = 021$



(b) H_2 en azul. En este caso $v_2 = 02$. Los números de las aristas indican el orden de recorrido. $x_2 = 021001122$.

Figura 1.5: Construcción de una palabra infinita de De Bruijn, $b = 3$

Capítulo 2

Secuencias Poisson genéricas

Dejamos momentáneamente de lado las palabras y grafos de De Bruijn para introducir las secuencias de Poisson genéricas. Años atrás, Zeev Rudnick las definió como aquellas secuencias infinitas que satisfacen que la ocurrencia de palabras largas en sus segmentos iniciales sigue la distribución de Poisson. La definición puede pensarse como una noción de aleatoriedad para secuencias infinitas de símbolos en un alfabeto finito, al igual que la definición de normalidad de Borel. Sin embargo, es una noción estrictamente más fuerte que la normalidad de Borel.

2.1. Definición

Definimos en primer lugar las siguientes funciones contadoras:

Definición 4. *Dados $x \in \Omega^{\mathbb{N}}$, $\lambda \in \mathbb{R}^+$, $i \in \mathbb{N}_0$ y $k \in \mathbb{N}$, definimos la función contadora $Z^{\lambda}_{i,k}(x)$ como la proporción de palabras de longitud k que ocurren exactamente i veces en $x[1 \dots \lfloor \lambda b^k \rfloor + k - 1]$,*

$$Z^{\lambda}_{i,k}(x) = \frac{\#\{w \in \Omega^k : |x[1 \dots \lfloor \lambda b^k \rfloor + k - 1]|_w = i\}}{b^k}.$$

Ahora sí estamos en condiciones de definir las secuencias Poisson genéricas:

Definición 5. *Sea $\lambda \in \mathbb{R}^+$. Una secuencia $x \in \Omega^{\mathbb{N}}$ se dice λ-Poisson genérica si para cada $i \in \mathbb{N}_0$,*

$$\lim_{k \to \infty} Z^{\lambda}_{i,k}(x) = e^{-\lambda} \frac{\lambda^i}{i!}.$$

Una secuencia se dice Poisson genérica si es λ-Poisson genérica para todo λ real positivo.

2.2. Motivación e intuiciones sobre la definición

Secuencias lacunarias

Zeev Rudnick definió la secuencias Poisson genéricas motivado por un resultado sobre secuencias lacunarias que obtuvo junto con Zaharescu ([33]) y que enunciamos a continuación.

Definición. *Una secuencia lacunaria es una secuencia de enteros* $(a(n))_{n\in\mathbb{N}}$ *que satisface la siguiente condición:*

$$\liminf_{n\to\infty} \frac{a(n+1)}{a(n)} > 1$$

El resultado de Rudnick y Zaharescu es el siguiente:

Teorema ([33, Theorem 1.1])**.** *Sea* $a(n)$ *una secuencia lacunaria. Entonces, para casi todo* α *real (respecto a la medida de Lebesgue), la secuencia* $(\alpha a(n))_{n\in\mathbb{N}}$ *satisface que*[1]*:*

Dado $\lambda > 0$*, la probabilidad de encontrar exactamente* i *elementos de la secuencia* $\{\alpha a(n) \mod 1 : n \leq N\}$ *en un intervalo elegido al azar de longitud* λ/N*, converge a* $e^{-\lambda}\frac{\lambda^i}{i!}$ *cuando* N *tiende a infinito.*

Si tomamos $\alpha \in [0,1)$ y $a(n) = b^n$, la secuencia $(\alpha a(n) \mod 1)$ no es más que efectuar sucesivos corrimientos de la coma hacia la derecha en la representación de α en base b (quitando la parte entera). Cuando $N = \lfloor \lambda b^k \rfloor$, el resultado anterior nos dice que para casi todo α, las probabilidades de que haya exactamente i elementos en un intervalo de longitud $\lambda/\lfloor \lambda b^k \rfloor \approx b^{-k}$ elegido al azar son las determinadas por la distribución de Poisson de parámetro λ.

La definición de Rudnick de secuencias Poisson genéricas se inspira en esta propiedad, pero no considera intervalos arbitrarios de longitud b^{-k}, sino los intervalos específicos

$$\left[0, \frac{1}{b^k}\right), \left[\frac{1}{b^k}, \frac{2}{b^k}\right), \ldots, \left[\frac{j}{b^k}, \frac{j+1}{b^k}\right), \ldots, \left[\frac{b^k-1}{b^k}, 1\right).$$

Para verlo, notar que la pertenencia de un número al intervalo $\left[\frac{j}{b^k}, \frac{j+1}{b^k}\right)$ equivale a que los primeros k dígitos después de la coma del desarrollo en base b del número sean precisamente la escritura en base b de j.

[1] El teorema original abarca más propiedades. Incluimos únicamente aquella que es relevante para la definición de secuencias Poisson genéricas.

Ubicación de bolitas en cajitas

La propiedad de λ- Poisson genericidad puede pensarse en términos de ubicación aleatoria de bolitas en cajas, donde las $N = \lfloor \lambda b^k \rfloor$ palabras iniciales de longitud k de una secuencia aleatoria son las bolitas, y las b^k posibles palabras de longitud k son las cajas.

Imaginemos por un momento que esta alocación de N bolitas en b^k cajas se realiza de forma independiente. En ese caso, la proporción esperada de las b^k cajas que contienen exactamente i bolitas, $0 \leq i \leq N$ (es decir, la proporción de palabras de longitud k que aparecen exactamente i veces) es

$$\binom{N}{i} p^i (1-p)^{N-i},$$

donde $p = b^{-k}$. Como Np converge a λ, una constante fija, cuando $N \to \infty$, tenemos que la distribución de Poisson con parámetro λ surge naturalmente al tomar límite de las distribuciones binomiales:

$$\lim_{\substack{N \to \infty \\ pN \to \lambda}} \binom{N}{i} p^i (1-p)^{N-i} = \frac{e^{-\lambda} \lambda^i}{i!}$$

Más aún, se puede demostrar (ver [16] por ejemplo) que si $\chi_k^{(i)}$ representa la cantidad de cajas que contienen exactamente i bolitas cuando hay b^k cajas y $N = \lfloor \lambda b^k \rfloor$ bolitas, entonces para cada i fijo,

$$\frac{1}{b^k} \chi_k^{(i)} \xrightarrow{P} \frac{e^{-\lambda} \lambda^i}{i!},$$

donde $\xrightarrow{P}$ representa convergencia en probabilidad.

Si bien no es cierto que en nuestro caso la alocación de distintas bolitas en cajas se haga de forma independiente (debido al solapamiento entre las palabras), la probabilidad de que dos palabras de longitud k elegidas al azar aparezcan en en dos posiciones i y j específicas, es igual a b^{-2k}, es decir, igual que si se tratara de eventos independientes, incluso cuando $|i - j| < k$.

Para demostrar esta última afirmación, notemos que para que dos palabras $x, y \in \Omega^k$ aparezcan en las posiciones i y j respectivamente, con $i < j$ y $j - i < k$, deben coincidir en las $i - j + k$ posiciones j, $j+1$, $\ldots$, $i + k - 1$, como se muestra en la Figura 2.1. La probabilidad de que dos palabras de longitud k elegidas al azar satisfagan esa condición es $b^{-(i-j+k)}$.

Por otro lado, dadas dos palabras que coinciden en j, $j+1$,$\ldots$, $i + k - 1$, la probabilidad de que aparezcan en las posiciones i y j es la probabilidad de que aparezcan los dígitos correctos en las $j - i + k$ posiciones totales, es decir $b^{-(j-i+k)}$.

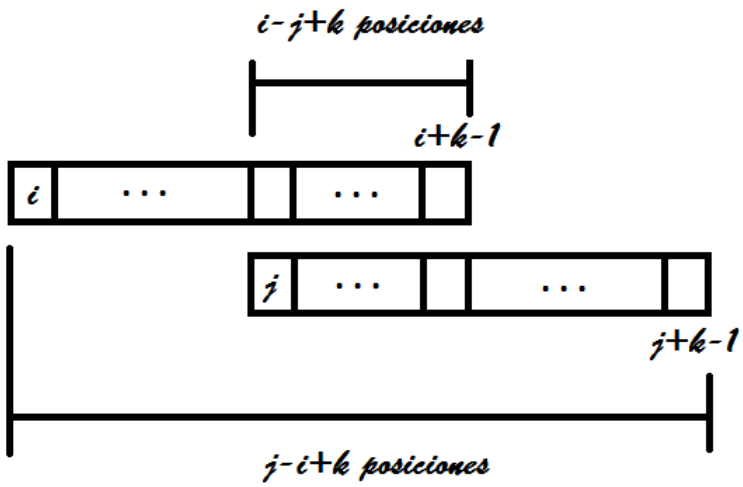


Figura 2.1: Posiciones superpuestas y casi independencia

Concluimos entonces que la probabilidad de que dos palabras elegidas al azar aparezcan en posiciones i y j determinadas es $b^{-(i-j+k)}b^{-(j-i+k)} = b^{-2k}$.

Sobre el tamaño de las palabras y los prefijos

En este apartado intentamos responder la siguiente pregunta:

¿Por qué en la definición de secuencias Poisson genéricas, contamos ocurrencias de palabras de tamaño logarítmico respecto al prefijo considerado?

Esta pregunta en parte puede responderse con las observaciones del apartado anterior. Sin embargo, es posible dar un mayor *insight* recurriendo a un teorema demostrado por Flajolet, Kirschenhofer y Tichy en [15]. Para enunciarlo, introducimos las siguientes definiciones:

Definición. *Dada una palabra w en un alfabeto de b símbolos, con $|w| = k$, se define su ℓ-discrepancia como:*

$$D_\ell(w) = \max_{|u|=\ell} \left| \frac{|w|_u}{k} - \frac{1}{b^\ell} \right|$$

Es decir, la ℓ-discrepancia cuantifica cuánto se aleja una palabra respecto de la equifrecuencia para la ocurrencia de subpalabras de longitud ℓ.

Definición. *Sea $w \in \Omega^{\mathbb{N}}$, y sea $\ell(k)$ una secuencia no decreciente de enteros positivos. Decimos que w es $\ell(k)$-uniformemente distribuida si*

$$\lim_{k\to\infty} b^{\ell(k)} D_{\ell(k)}(w[1\dots k]) = 0$$

A continuación enunciamos el teorema al que nos referimos en los párrafos anteriores:

Teorema ([15, Corollary 1])**.** *Casi todas las secuencias infinitas en un alfabeto de dos símbolos (respecto a la medida de Lebesgue) son $\ell(k)$-uniformemente distribuidas, para $\ell(k) = \lfloor (1-\varepsilon)\log_2 k \rfloor$, $\varepsilon > 0$.*

Es decir, si en la definción de las secuencias 1-Poisson genéricas en base 2 hubiésemos tomado funciones contadoras que analizaran ocurrencias de palabras de longitud $\lfloor (1-\varepsilon)k \rfloor$ en prefijos de longitud 2^k, observaríamos para casi toda secuencia que hay convergencia a la equidistribución. Por lo tanto, en ese sentido, el tamaño logarítmico de las palabras en función de la longitud del prefijo, es el "tamaño justo" que hay que considerar para observar algo distinto a la equidistribución.

2.3. Propiedades

Medida del conjunto de secuencias Poisson genéricas

Consideremos el espacio $\Omega = \{0, 1, \dots, b-1\}$ junto con la medida uniforme μ, y el espacio de secuencias infinitas en el alfabeto Ω, $\Omega^{\mathbb{N}}$, con la medida producto $\mu^{\mathbb{N}}$ (la medida uniforme). Yuval Peres y Benjamin Weiss [36] demostraron que casi todas las secuencias respecto a la medida uniforme son Poisson genéricas. De hecho, probaron un resultado más fuerte, que enunciamos a continuación. A pesar de conocerse este hecho desde hace tiempo, ningún ejemplo explícito había sido hallado previamente a la realización de este trabajo. La transcripción de la demostración de Peres y Weiss fue recientemente publicada en [1].

Peres y Weiss definieron procesos de puntos $M_k^x(\cdot)$ en $\mathbb{R}^+$ para cada $x \in \Omega^{\mathbb{N}}$ y para cada $k \in \mathbb{N}$ del siguiente modo:

$$M_k^x(S)(\omega) = \#\{j \in \mathbb{N} \cap b^k S : x[j \dots j+k-1] = \omega\},$$

donde $S \subseteq \mathbb{R}^+$ es un conjunto boreliano, y $\omega \in \Omega^k$.

Con estas definiciones, Peres y Weiss demostraron el siguiente teorema:

Teorema ([1, Theorem 1])**.** *Para casi todo $x \in \Omega^{\mathbb{N}}$ con respecto a la medida producto $\mu^{\mathbb{N}}$, el proceso de puntos $M_k^x(\cdot)$ converge en distribución a un proceso de Poisson en $\mathbb{R}^+$ cuando k tiende a infinito.*

En Ω^k consideramos la medida uniforme μ^k. Notemos que tomando $S = (0, \lambda]$, se recuperan las funciones contadoras Z:

$$Z_{i,k}^{\lambda}(x) = \mu^k(\{\omega \in \Omega^k : M_k^x((0,\lambda]) = i\})$$

Usando esta formulación de las funciones contadoras en términos de los procesos $M_k^x(\cdot)$, se deduce que casi todas las secuencias son Poisson genéricas respecto a la medida uniforme.

¿Cuán aleatorias son las secuencias de Poisson genéricas?

En el año 1909, Émile Borel presentó la noción de normalidad [9]. Si bien Borel dio su definición en términos de números reales, ésta se traduce naturalmente al lenguaje de las secuencias:

Definición. *Sea Ω un alfabeto de b símbolos, $b \geq 2$. Una secuencia $x \in \Omega^{\mathbb{N}}$ es normal en base b si cada palabra w ocurre en x con la misma frecuencia límite que el resto de las palabras de su misma longitud. Es decir,*

$$\lim_{n\to\infty} \frac{|x[1\ldots n]|_w}{n} = b^{-|w|}.$$

Borel demostró que casi todas las secuencias (respecto a la medida de Lebesgue o la medida uniforme) son normales, y planteó el problema de exhibir un ejemplo. Recién en el año 1933, Champernowne dio la primera instancia explícita de una tal secuencia (ver [10]), que consiste en la concatenación de la escritura en base b de los números $0, 1, 2, \ldots$. Por ejemplo, en base 2, la secuencia de Champernowne es:

$$0110111001011101111000100110 10\ldots$$

Si bien la normalidad es una condición deseable a la hora de clasificar a una secuencia como "aleatoria", no parece ser suficiente. Sin ir más lejos, la secuencia de Champernowne es altamente predecible, y dudaríamos mucho en llamarla aleatoria.

En los años '60 y '70, fueron los trabajos de Martin-Löf, Kolmogorov, Chaitin, y Schnorr, los primeros en sentar bases sólidas que permitieron llegar a la que hoy en día es la definición de aleatoriedad algorítmica. Esta definción es altamente robusta: existen múltiples formulaciones equivalentes que a priori no parecen estar necesariamente relacionadas. Delahaye acuña el término *Tesis de Martin-Löf-Chaitin* en [13], estableciendo una analogía con la Tesis de Turing-Church para el concepto de cómputo efectivo. La Tesis de Martin-Löf-Chaitin puede formularse como:

El concepto intuitivo informal de secuencias de ceros y unos aleatorias es adecuadamente capturado por la definición de Martin Löf

Sin más preámbulos, enunciamos a continuación una de las tantas definiciones de aleatoriedad algorítmica, en la versión de Chaitin:

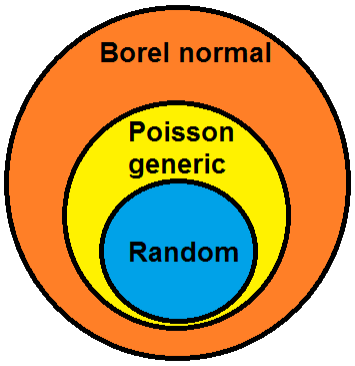


Figura 2.2: Jerarquía de nociones de aleatoriedad

Definición. *Una secuencia es aleatoria si esencialmente la única forma de describir computacionalmente sus segmentos iniciales es explícitamente.* [2]

Es posible demostrar que, como era de esperar, toda secuencia aleatoria es normal, pero no vale la recíproca.

¿Dónde se insertan las secuencias Poisson genéricas en esta jerarquía de aleatoriedad? Por un lado, Peres y Weiss ([37],[36]) demostraron que toda secuencia 1- Poisson genérica es normal (la demostración la exponemos con una pequeña generalización en el Capítulo 4). Además, probaron que la secuencia de Champernowne no es Poisson genérica a pesar de ser normal. Por otro lado, Becher, Álvarez y Mereb demostraron el siguiente teorema.

Teorema ([1, Theorem 3])**.** *Toda secuencia Martin-Löf aleatoria es Poisson genérica.*

A partir de un resultado que enunciamos en la próxima sección, es posible deducir que existen secuencias Poisson genéricas que no son aleatorias. Una representación gráfica de estos resultados puede apreciarse en la Figura 2.2. Cada inclusión es estricta.

Podemos concluir entonces que la noción de Poisson genericidad es estrictamente más fuerte que la de normalidad de Borel, y estrictamente más débil que la de aleatoriedad de Martin-Löf.

Secuencias Poisson genéricas computables

Introducimos la siguiente definición:

Definición. *Una secuencia infinita $x = a_1a_2\ldots$ de símbolos en un alfabeto finito se dice computable cuando la función $k \mapsto a_k$ es computable.*

[2] No entramos en los detalles técnicos que permiten formalizar esta definición.

Un avance en la dirección de exhibir instancias de secuencias Poisson genéricas es el siguiente teorema de Álvarez, Becher y Mereb:

Teorema ([1, Theorem 2])**.** *Existen numerables secuencias Poisson genéricas computables en* $\Omega^{\mathbb{N}}$*.*

La demostración de este teorema es, en esencia, similar a la construcción de Turing (ver [35] y [5]), o a la versión computable de la construcción de Sierpiński [4] de números absolutamente normales[3]. En estas demostraciones, se construye un número absolutamente normal como el punto de interesección de una sucesión computable de intervalos encajados cuya medida converge a cero. El algoritmo es iterativo: en el paso k-ésimo se elige un subintervalo del intervalo actual que no contiene números "malos" de nivel k, es decir, aquellos cuyo segmento inicial de tamaño k en su expansión en ciertas bases, está desbalanceado. La correctitud de estos algoritmos puede demostrarse gracias al uso de cotas para el tamaño de los conjuntos de números "malos". La demostración del teorema de Álvarez, Becher y Mereb se basa en una cota de este estilo pero para la propiedad de Poisson genericidad.

El teorema anterior no proporciona secuencias "explícitas" como la de Champernowne. Este es el problema del que nos ocupamos en el próximo capítulo.

[3]Un número es absolutamente normal si su expansión en cualquier base entera es normal.

Capítulo 3

Una construcción de secuencias λ-Poisson genéricas

En el presente capítulo introducimos la principal contribución original de esta tesis: damos una construcción explícita de una secuencia λ-Poisson genérica, para cualquier $\lambda \in \mathbb{R}^+$ fijo y $b \geq 3$. En el caso de que $b = 2$, la construcción funciona siempre que $\lambda \leq \ln(2)$. Demostramos un teorema de mayor generalidad que da una técnica de construcción de secuencias normales. Para una elección determinada de los parámetros, se obtienen las secuencias de Poisson buscadas.

3.1. Presentación de resultados

El Teorema principal que demostramos en este capítulo es el siguiente:

Teorema 1. *Sea λ un real positivo y Ω un alfabeto de b símbolos. Sea $(p_i)_{i\in\mathbb{N}_0}$ una secuencia de números reales no negativos tales que $\sum_{i\geq 0} p_i = 1$ y $\sum_{i\geq 0} ip_i = \lambda$. Si $b = 2$, adicionalmente requerimos que $p_0 \geq 1/2$. Entonces es posible dar una construcción de una secuencia infinita x sobre el alfabeto Ω que satisface para cada $i \in \mathbb{N}_0$,*

$$\lim_{k\to\infty} Z^{\lambda}_{i,k}(x) = p_i.$$

Tomando $p_i = e^{-\lambda}\lambda^i/i!$, x es λ-Poisson genérica.

Corolario 1. *Sea Ω un alfabeto de b símbolos. Si $b \geq 3$, fijemos λ un real positivo cualquiera. Si $b = 2$, tomemos $\lambda \leq \ln(2)$. Entonces es posible dar una construcción de una secuencia λ-Poisson genérica $x \in \Omega^{\mathbb{N}}$.*

Observación. ***Sobre la computabilidad de las secuencias generadas*** *Dado que el conjunto de secuencias computables es numerable, tiene medida de Lebesgue 0, por*

lo que la existencia de secuencias de Poisson genéricas computables no se desprende necesariamente de que el conjunto de secuencias Poisson genéricas tenga medida 1. *El Teorema 1 permite obtener una instancia computable cuando* $(p_i)_{i\in\mathbb{N}}$ *es una secuencia computable de números reales, es decir, cuando la función* $(i,n) \mapsto$ *el* n*-ésimo dígito en el desarrollo en base* b *de* p_i*, es computable.*

En particular, el Teorema 1 brinda una instancia computable de una secuencia 1*-Poisson genérica, también llamada simply- Poisson generic ([37],[36]).*

3.2. La construcción

Sean λ un real positivo y Ω un alfabeto de b símbolos. Sea $(p_i)_{i\in\mathbb{N}_0}$ una secuencia de reales no negativos tales que $\sum_{i\geq 0} p_i = 1$ and $\sum_{i\geq 0} ip_i = \lambda$.

La construcción requerirá los siguientes elementos:

- En primer lugar, fijamos una secuencia infinita de De Bruijn A sobre el alfabeto Ω (recordar la Definición 3). Definimos A_k como $A[1\ldots b^k]$.
- Definimos también la función $g : \mathbb{N} \to \mathbb{N}$ como $g(k) = \lceil \frac{k}{2} \rceil$.
- Definimos inductivamente las secuencias de números reales $(p_i^k)_{i\geq 0, k\geq 1}$ del siguiente modo:

  Para cada $i \geq 1$,

$$p_i^1 = \{p_i\}_{g(1)},$$
$$p_0^1 = 1 - \sum_{i\geq 1} p_i^1.$$

  Y para cada $k \geq 1$ y $i \geq 1$,

$$p_i^{k+1} = \frac{1}{b} p_i^k + \left\{ \frac{b-1}{b} p_i \right\}_{g(k+1)}$$
$$p_0^{k+1} = 1 - \sum_{i\geq 1} p_i^{k+1}.$$

La construcción se realizará en pasos. Denotamos por x_k a la secuencia obtenida luego del Paso k de la construcción.

Comenzamos con x_0 la palabra vacía.

Paso 1. En este primer paso, consideramos la expansión en base b de p_i^1, para $i \geq 1$,

$$p_i^1 = 0, c_i$$

Para cada c_i, $i \geq 1$, seleccionamos c_i bloques de longitud relativa b^{-1} con respecto a A_1, esto es, bloques de longitud absoluta 1 (si $c_i = 0$ no elegimos ningún bloque). Los bloques escogidos no deben superponerse. Esto es posible gracias a que $\sum_{i\geq 1} p_i^1 \leq 1$. La salida de la construcción después del Paso 1, x_1, será la concatenación de los bloques elegidos, en cualquier orden, donde para cada $i \geq 1$, cada uno de los c_i bloques seleccionados se repite exactamente i veces.

Paso k+1. Consideramos la expansión en base b de los números $\left\{\frac{b-1}{b}p_i\right\}_{g(k+1)}$ para $i \geq 1$:

$$\left\{\frac{b-1}{b}p_i\right\}_{g(k+1)} = 0, a_{i,1}a_{i,2}..a_{i,g(k+1)}$$

donde $a_{i,j} \in \{0, 1, 2, \ldots, b-1\}$.

Ahora seleccionamos bloques del siguiente modo: para cada $a_{i,j}$, $i \geq 1$, $j \leq g(k+1)$, elegimos $a_{i,j}$ bloques de longitud relativa b^{-j} con respecto a A_{k+1}. Si $a_{i,j} = 0$ no seleccionamos ningún bloque. Notar que solo finitos bloques son seleccionados. Además, los bloques tienen que ser tales que no se superpongan. Esto es posible gracias a que

$$\sum_{i\geq 1}\sum_{1\leq j\leq g(k+1)} a_{i,j}\frac{1}{b^j} = \sum_{i\geq 1}\left\{\frac{b-1}{b}p_i\right\}_{g(k+1)} \leq \frac{b-1}{b} = \frac{|A_{k+1}[b^k+1\ldots b^{k+1}]|}{b^{k+1}}$$

En el caso de que $b = 3$, los bloques podemos elegirlos en cualquier sitio en A_{k+1}, por ejemplo, en $A_{k+1}[b^k + 1 \ldots b^{k+1}]$. En el caso de $b = 2$, los bloques los elegimos de forma de no dejar huecos entre los bloques del paso k y los del paso $k + 1$.

La construcción concatena ahora los bloques elegidos al final de x_k, para obtener x_{k+1}. Para cada $i \geq 1$, $j \leq g(k + 1)$, cada uno de los $a_{i,j}$ bloques es repetido exactamente i veces. Decimos que cada uno de los bloques elegidos en A son *segmentos constitutivos* en la salida x_{k+1}. También decimos que la concatenación de las i copias de un segmento constitutivo correspondiente a $a_{i,j}$ es una *racha* en la salida.

3.3. Un ejemplo

Para ilustrar el modo en que funciona la construcción, damos un ejemplo de tres pasos de la ejecución. Para preservar la claridad de la explicación, tomamos $g(k) = k$ en esta sección. Sea $p_0 = 0$, $p_1 = 1/2$, $p_2 = 5/18$, $p_3 = 2/9$, y $p_i = 0$ para $i \geq 4$. En este caso $\lambda = 31/18$. Fijemos ahora $b = 3$, $\Omega = \{0, 1, 2\}$ y

$$A = 0121100220102000111202122 21\ldots$$

$p_1 = 0{,}1111\ldots$ y $\frac{2}{3}p_1 = 0{,}1000\ldots$

$p_2 = 0{,}0211\ldots$ y $\frac{2}{3}p_2 = 0{,}0120\ldots$

$p_3 = 0{,}0200\ldots$ y $\frac{2}{3}p_3 = 0{,}0110\ldots$

Step 1:

$$A = \boxed{0}12\Big|110022\Big|010200011120212221\Big|\ldots$$

$$x_1 = \boxed{0}$$

Step 2:

$$A = 012\Big|\boxed{110}\boxed{0}\boxed{2}2\Big|010200011120212221\Big|\ldots$$

$$x_2 = 0\boxed{110}\boxed{0}\boxed{0}\boxed{2}\boxed{2}\boxed{2}$$

En este caso $\boxed{0}$, $\boxed{110}$, $\boxed{0}$ y $\boxed{2}$ son los segmentos constitutivos de x_2.

Step 3:

$$A = 012\Big|110022\Big|\boxed{010200011}\boxed{120}\boxed{021}\boxed{2}\boxed{2}\boxed{1}\Big|\ldots$$

$$x_3 = 011000222\boxed{010200011}\boxed{120}\boxed{120}\boxed{212}\boxed{212}\boxed{212}\boxed{2}\boxed{2}\boxed{2}\boxed{2}\boxed{1}\boxed{1}\boxed{1}$$

En este caso, $\boxed{120}\boxed{120}$ es la racha correspondiente al segmento constitutivo $\boxed{120}$, y $\boxed{212}\boxed{212}\boxed{212}$ es la racha correspondiente al segmento constitutivo $\boxed{212}$.

3.4. Correctitud

Para demostrar que la construcción presentada arroja una secuencia que satisface la condición del Teorema 1, usamos cinco lemas que enunciamos y probamos a continuación.

Observación 2. *Para cada $y \in [0, 1)$ y cada $k \geq 1$, $y - \frac{1}{b^k} < \{y\}_k \leq y$. En el caso que $y = 1$, $\{y\}_k = y - \frac{1}{b^k}$.*

Lema 7. *Sea $b = 2$ y $k \geq 2$. Entonces, en el paso k de la construcción, siempre es posible tomar todos los bloques necesarios de $A[1 \dots 2^{k-1}]$.*

Demostración. En primer lugar, notemos que para $k \geq 1$, la longitud relativa con respecto a A_{k+1} de los bloques que deben ser seleccionados en el paso $k+1$ es

$$\sum_{i \geq 1} \sum_{1 \leq j \leq g(k+1)} a_{i,j} \frac{1}{2^j} = \sum_{i \geq 1} \left\{ \frac{1}{2} p_i \right\}_{g(k+1)} \leq \frac{1}{2} \sum_{i \geq 1} p_i = \frac{1}{2}(1-p_0) \leq \frac{1}{4} = \frac{|A[2^{k-1}+1 \dots 2^k]|}{2^{k+1}},$$

donde $a_{i,j}$ tiene el mismo significado que en la construcción, y en la última desigualdad usamos la hipótesis $p_0 \geq \frac{1}{2}$.

Esto significa que $A[2^{k-1}+1 \dots 2^k]$ tiene suficiente espacio para todos los bloques que deben elegirse en el paso $k+1$. Entonces, solo es necesario verificar que $A[2^{k-2}+1 \dots 2^{k-1}]$ está libre en el paso k para todo $k \geq 2$. Podemos probarlo inductivamente. En el primer paso, la proporción usada de A_1 es

$$\sum_{i \geq 1} p_i^1 \leq \sum_{i \geq 1} p_i = 1 - p_0 \leq \frac{1}{2},$$

donde la última desigualdad vale gracias a que $p_0 \geq 1/2$. Entonces, al menos la mitad de $A_1 = A[1 \dots 2]$ queda sin usar después del paso 1, con lo que $A[2^{2-2}+1 \dots 2^{2-1}] = A[2 \dots 2]$ está disponible en el paso 2. Esto prueba el caso base.

Ahora supongamos que en el paso k, $A[2^{k-2}+1 \dots 2^{k-1}]$ está libre. Gracias a la primera observación, podemos elegir todos los bloques necesarios allí. Esto deja $A[2^{k-1}+1 \dots 2^k]$ libre para el paso $k+1$. □

Lema 8. *Para cada $i \geq 1$, p_i^k es la suma de las longitudes relativas respecto a A_k de todos los segmentos constitutivos en la salida x_k que son repetidos exactamente i veces.*

Demostración. Se demuestra fácilmente por inducción en k, usando la definición de p_i^k y el modo en que opera la construcción.

Si $k = 1$, es verdadero por el Paso 1 de la construcción.

Asumiendo que el enunciado es verdadero para k, veamos que también lo es para $k+1$. Notar que los bloques que ocurren en x_k tienen una longitud relativa respeco a A_{k+1} que es $1/b$ de su longitud relativa en A_k. Los bloques extra añadidos contribuyen con $\left\{ \frac{b-1}{b} p_i \right\}_{g(k+1)}$ a la suma. Entonces, la suma de las longitudes relativas con respecto a A_{k+1} es

$$\frac{1}{b} p_i^k + \left\{ \frac{b-1}{b} p_i \right\}_{g(k+1)} = p_i^{k+1}.$$

□

Lema 9. *Para cada $i \in \mathbb{N}_0$, $\lim_{k\to\infty} p_i^k = p_i$. De hecho, para cada $i \geq 1$, $k \geq 1$, valen las siguientes desigualdades:*

$$p_i - \frac{k}{b^{g(k)}} \leq p_i^k \leq p_i. \tag{3.1}$$

Demostración. Para $i \geq 1$ probamos (3.1) por inducción en k. Si $k = 1$ se sigue inmediatamente de la definición de p_i^1 y de la Observación 2. Para el paso inductivo, notar que

$$p_i^{k+1} = \frac{1}{b}p_i^k + \left\{\frac{b-1}{b}p_i\right\}_{g(k+1)} \leq \frac{1}{b}p_i + \frac{b-1}{b}p_i \leq p_i.$$

$$\begin{aligned}
p_i - p_i^{k+1} = p_i - \left(\frac{1}{b}p_i^k + \left\{\frac{b-1}{b}p_i\right\}_{g(k+1)}\right) &\leq p_i - \frac{1}{b}\left(p_i - \frac{k}{b^{g(k)}}\right) - \left\{\frac{b-1}{b}p_i\right\}_{g(k+1)} \\
&\leq \frac{b-1}{b}p_i - \left\{\frac{b-1}{b}p_i\right\}_{g(k+1)} + \frac{k}{b^{1+g(k)}} \\
&\leq \frac{1}{b^{g(k+1)}} + \frac{k}{b^{1+g(k)}} \\
&\leq \frac{k+1}{b^{g(k+1)}},
\end{aligned}$$

donde en la última desigualdad usamos que $g(k+1) \leq g(k) + 1$.

En el caso de $i = 0$,

$$|p_0^k - p_0| = \left|1 - \sum_{i\geq 1} p_i^k - \left(1 - \sum_{i\geq 1} p_i\right)\right| = \sum_{i\geq 1}(p_i - p_i^k).$$

Dado $\varepsilon > 0$, existe $N > 0$ tal que $\sum_{i\geq N+1} p_i < \frac{\varepsilon}{2}$. Entonces,

$$|p_0^k - p_0| \leq \sum_{i=1}^{N}(p_i - p_i^k) + \sum_{i\geq N+1} p_i < \frac{kN}{b^{g(k)}} + \frac{\varepsilon}{2}.$$

Si k es suficientemente grande, entonces $\frac{kN}{b^{g(k)}} < \frac{\varepsilon}{2}$ y $|p_0^k - p_0| < \varepsilon$, como queríamos. □

Lema 10. *Sea x_k la salida de la construcción al finalizar el Paso k. Entonces,*

$$\lim_{k\to\infty} \frac{|\lfloor \lambda b^k \rfloor + k - 1 - |x_k||}{b^k} = 0.$$

Demostración. Por el Lemma 8, $|x_k| = b^k \sum_{i\geq 1} ip_i^k$. Entonces,

$$\frac{|\lfloor \lambda b^k \rfloor + k - 1 - |x_k||}{b^k} \leq \frac{k-1}{b^k} + \frac{|\lfloor \lambda b^k \rfloor - \lambda b^k|}{b^k} + \left|\lambda - \sum_{i\geq 1} ip_i^k\right| \leq \frac{k}{b^k} + \left|\lambda - \sum_{i\geq 1} ip_i^k\right|.$$

Basta con probar que el último término converge a cero. Recordemos que $(p_i)_{i\in\mathbb{N}_0}$ satisface $\lambda = \sum_{i\geq 1} ip_i$. Entonces,

$$\left|\lambda - \sum_{i\geq 1} ip_i^k\right| = \sum_{i\geq 1} i(p_i - p_i^k).$$

Dado $\varepsilon > 0$, tomemos N suficientemente grande de forma que $\sum_{i\geq N+1} ip_i < \frac{\varepsilon}{2}$. Gracias a la Ecuación 3.1 del Lemma 9,

$$\sum_{i\geq 1} i(p_i - p_i^k) \leq \sum_{i=1}^{N} i\frac{k}{b^{g(k)}} + \sum_{i\geq N+1} ip_i < \frac{k}{b^{g(k)}}\frac{N(N+1)}{2} + \frac{\varepsilon}{2}.$$

Claramente, cuando k es suficientemente grande, $\frac{k}{b^{g(k)}}\frac{N(N+1)}{2} < \frac{\varepsilon}{2}$. □

Recordemos que cada uno de los bloques de A usados en el paso k, es un segmento constitutivo de la salida x_k, y que la concatenación de i copias de un segmento constitutivo correspondiente a $a_{i,j}$ es una racha.

Sea B_k el número de rachas en la salida x_k.

Lema 11. *Las cantidades B_k satisfacen*

$$\lim_{k\to\infty} \frac{kB_k}{b^k} = 0.$$

Demostración. Notemos que para cada $\ell \geq 2$,

$$\begin{aligned}
\frac{1}{b^{g(\ell)}}(B_\ell - B_{\ell-1}) &= \frac{1}{b^{g(\ell)}} \sum_{i\geq 1}\sum_{j=1}^{g(\ell)} a_{i,j} \\
&\leq \sum_{i\geq 1}\sum_{j=1}^{g(\ell)} \frac{1}{b^j} a_{i,j} \\
&\leq \sum_{i\geq 1} \frac{b-1}{b} p_i \\
&\leq 1.
\end{aligned}$$

Recordemos que $g(\ell) = \left\lceil \frac{\ell}{2} \right\rceil$. Notemos que

$$B_k = B_1 + \sum_{\ell=2}^{k} B_\ell - B_{\ell-1}.$$

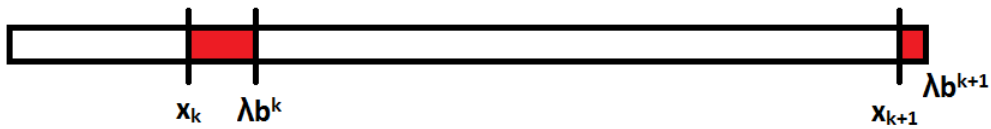


Figura 3.1: Primera fuente de error

Entonces obtenemos

$$\begin{aligned}\frac{kB_k}{b^k} \leq & \frac{k\sum_{\ell=2}^{k} b^{g(\ell)} + kB_1}{b^k} \\ \leq & \frac{k\sum_{\ell=0}^{k} b^{1+\ell/2} + kB_1}{b^k} \\ \leq & \frac{b(b^{1/2}-1)^{-1}k(b^{(k+1)/2}-1) + kB_1}{b^k},\end{aligned}$$

que converge a 0 cuando k tiende a infinito. □

Observación. *Existen muchas otras alternativas para la elección de la función g. Por ejemplo, cualquier función que satisfaga las siguientes condiciones es una opción viable:*

- $g(k) \leq k/m$ *para cada* $k \in \mathbb{N}$, *donde* m *es una constante mayor que* 1.
- $g(k+1) \leq g(k) + 1$ *para cada* $k \in \mathbb{N}$.
- $\frac{k}{b^{g(k)}} \xrightarrow{k\to\infty} 0$

La primera y la segunda condición aseguran que los Lemas 11 y 9 sigan valiendo, respectivamente. La tercera condición garantiza que $\lim_{k\to\infty} p_i^k = p_i$. *Por ejemplo,* $g(k) = \lceil\sqrt{k}\rceil$ *es una posible elección, pero* $g(k) = \lceil\log_b(k)\rceil$ *no lo es, porque falla la tercera condición.*

Ahora estamos en condiciones de dar la demostración del Teorema 1.

Teorema 1. En primer lugar consideramos el caso $i \geq 1$, y estimamos el valor de $Z_{i,k}^{\lambda}(x)$. Como estamos interesados únicamente en el valor de $Z_{i,k}^{\lambda}(x)$ cuando k tiende a infinito, por el Lema 10 es suficiente contar la ocurrencia de palabras en x_k en lugar de $x[1\ldots\lfloor\lambda b^k\rfloor + k - 1]$. En la Figura 3.1 puede verse una representación gráfica de esta

Figura 3.2: Segunda fuente de error

aproximación. Se marcan en rojo las palabras que deberían ser contadas en el paso k pero que son ignoradas.

Veamos que los segmentos constitutivos elegidos hasta el paso k no tienen palabras de longitud k en común. Por la Definición 3, si $b \geq 3$, cada palabra de longitud k ocurre exactamente una vez en A_k, de donde se sigue la afirmación. Si $b = 2$ y k es impar, $A[1 \ldots 2^k]$ es una palabra de De Bruijn de orden k, por lo que no repite palabras de longitud k. Si $b = 2$ y k es par, $A[1 \ldots 2^{k-1}]$ es una palabra de De Bruijn de orden $k-1$, con lo que no repite palabras de longitud $k-1$ (y en consecuencia, tampoco repite palabras de longitud k). Como hasta el paso k inclusive, la construcción seleccionó bloques únicamente de $A[1 \ldots 2^{k-1}]$ (gracias al Lema 7), dos segmentos constitutivos distintos no comparten ninguna palabra de longitud k.

Del párrafo anterior, deducimos que si un segmento constitutivo w de longitud relativa b^{-j} con respecto a A_k, y longitud absoluta b^{k-j}, es repetido exactamente i veces en x_k, entonces podemos asumir que contribuye con b^{k-j} palabras al numerador de la función contadora $Z^{\lambda}_{i,k}(x)$, esto es, contribuye a $Z^{\lambda}_{i,k}(x)$ con su longitud relativa con respecto a A_k. Para verlo, notemos que las palabras de longitud k que podrían hacer que el conteo real difiera de esta aproximación pertenecen a alguno de los dos grupos siguientes:

1) **En rachas correspondientes a un segmento constitutivo de longitud mayor o igual que** k: Debemos considerar las palabras que ocurren entre los segmentos constitutivos dentro de la racha, que son a lo sumo k distintas; y las palabras entre el final de la racha y el inicio de la siguiente, que también son a lo sumo k.

   Una representación gráfica de este error puede observarse en la Figura 3.2. En naranja se representa un bloque constitutivo que es repetido tres veces, formando una racha. Las palabras que introducen el error a cuantificar son las señaladas por los óvalos rojos. Notar que las palabras entre el segundo y tercer segmento naranja son las mismas que las que ocurren entre el primero y el segundo.

2) **En rachas correspondientes a un segmento constitutivo de longitud** $s < k$: En este caso, las primeras s palabras de longitud k se repiten a lo largo de toda la racha. Hay menos de k palabras extra que ocurren entre el final de la

racha y el comienzo de la siguiente. Hay entonces a lo sumo $s + k < 2k$ palabras distintas que contar por cada racha en este caso.

Concluimos entonces que este segundo error está acotado por $2kB_k$. Gracias al Lema 11, sabemos que converge a cero.

Usando el hecho de que el error de aproximar $Z^{\lambda}_{i,k}(x)$ por p_i^k converge a cero para $i \geq 1$, podemos calcular:

$$\lim_{k\to\infty} Z^{\lambda}_{i,k}(x) = \lim_{k\to\infty} p_i^k = p_i,$$

para $i \geq 1$, donde la última igualdad vale por el Lema 9.

Para concluir, veamos el caso $i = 0$. Necesitamos estimar

$$Z^{\lambda}_{0,k}(x) = \frac{\#\{w \in \Omega^k : |x[1 \ldots \lfloor \lambda b^k \rfloor + k - 1]|_w = 0\}}{b^k}.$$

El numerador es igual al número total de palabras de longitud k menos el número de palabras de longitud k que ocurren al menos una vez. Estimamos entonces el cociente como la longitud relativa de la porción de A_k que queda sin usar luego del paso k, es decir:

$$1 - \sum_{i\geq 1} p_i^k = p_0^k.$$

Podemos verificar que el error de esta estimación converge a cero empleando los mismos argumentos que antes. Deducimos entonces que:

$$\lim_{k\to\infty} Z^{\lambda}_{0,k}(x) = \lim_{k\to\infty} p_0^k = p_0,$$

donde la última igualdad vale por el Lema 9. Esto concluye la demostración. □

3.5. Una posible mejora

En lugar de escoger múltiples bloques para cada p_i en el paso k-ésimo de la construcción ($k \geq 2$), podemos tomar un único bloque[1] de longitud relativa $\sum_{j=1}^{g(k)} a_{i,j} b^{-j}$ (donde los $a_{i,j}$ se definen como antes).

Realizando la construcción de este modo, podemos tomar $g(k) = k$. Para verlo, sea $f(k)$ la cantidad de índices $i \geq 1$ tales que $\frac{b-1}{b} p_i \geq b^{-k}$. Es decir, $f(k)$ se corresponde con la cantidad de probabilidades p_i con $i \geq 1$ tales que $\frac{b-1}{b} p_i$ tiene algún dígito no nulo en el truncamiento a k dígitos de su desarrollo en base b.

Entonces:

$$B_k - B_{k-1} = f(k),$$

y por lo tanto

[1] Usamos ahora el término bloque sin restringirnos a subpalabras de tamaño potencia de b.

$$B_k = B_1 + \sum_{j=2}^{k} B_j - B_{j-1} = B_1 + \sum_{j=2}^{k} f(j).$$

Notemos por otro lado que

$$\lambda = \sum_{i \geq 1} i p_i \geq \sum_{j=1}^{f(k)} i \frac{b}{b-1} b^{-k},$$

de donde

$$f(k)^2 \leq f(k)(f(k)+1) \leq 2\frac{b-1}{b}\lambda b^k.$$

De esta última desigualdad deducimos que:

$$f(k) \leq C b^{k/2},$$

con C una constante que depende de b y λ.

Entonces

$$\frac{kB_k}{b^k} \leq \frac{kB_1}{b^k} + \frac{k}{b^k}\sum_{j=2}^{k} C b^{j/2} \leq \frac{kB_1}{b^k} + \frac{kC(b^{(k+1)/2}-1)(b^{1/2}-1)^{-1}}{b^k},$$

y esta última expresión converge a 0 cuando k tiende a infinito. Esto demuestra el Lema 11 para esta versión de la construcción. A partir de aquí, la demostración del Teorema 1 es idéntica.

3.6. Limitaciones de la construcción

La construcción expuesta en las secciones previas resuelve el problema de exhibir una secuencia λ-Poisson genérica para cualquier λ fijo. Sin embargo, no nos permite generar una secuencia de Poisson genérica a secas. Esto se debe en parte a que utilizamos una secuencia infinita de De Bruijn para la construcción. Supongamos que construimos x para $\lambda = 1$. Entonces las frecuencias para $\lambda = 1/b$, $i \geq 0$, satisfacen,

$$\lim_{k\to\infty} Z^{1/b}_{i,k+1}(x) - \frac{1}{b} Z^{1}_{i,k}(x) = 0.$$

Pero esta relación no se satisface en el caso de las probabilidades puntuales de la distribución de Poisson:

$$e^{-1/b}\frac{1}{b^i i!} \neq e^{-1}\frac{1}{bi!}.$$

Queda abierta entonces la pregunta de si es posible adaptar la construcción para producir una secuencia de Poisson genérica, cambiando por ejemplo la secuencia A que es utilizada como fuente para los bloques.

Capítulo 4

Un criterio de normalidad

Como ya mencionamos en el Capítulo 2, Peres y Weiss [37, 36] demostraron que las secuencias 1-Poisson genéricas son normales. En este capítulo presentamos el segundo aporte de esta tesis: un criterio de normalidad cuya demostración es una pequeña generalización de la de Peres y Weiss. Seguimos esencialmente la versión de la demostración dada por Puterman en [29].

4.1. Presentación de resultados

Enunciamos a continuación el teorema principal de este capítulo.

Teorema 2. *Sea Ω un alfabeto de b símbolos, $b \geq 2$, y sea $x \in \Omega^{\mathbb{N}}$. Sea λ un real positivo fijo, y para cada $i \in \mathbb{N}_0$, sea $p_i = \liminf_{k\to\infty} Z^{\lambda}_{i,k}(x)$. Si los números p_i satisfacen $\sum_{i\geq 0} ip_i = \lambda$, entonces x es normal en base b.*

Observación. *Se puede verificar fácilmente, empleando el Lema de Fatou por ejemplo, que si los números p_i se definen como en el enunciado del Teorema 2, entonces siempre ocurre que $\sum\limits_{i\geq 0} ip_i \leq \lambda$.*

Empleando el criterio de normalidad presentado, es posible demostrar que todas las secuencias obtenidas por medio de la construcción del Teorema 1 son normales. Por lo tanto, obtenemos el siguiente corolario:

Corolario 2. *Toda secuencia λ-Poisson genérica es normal, pero las dos nociones no coinciden. La construcción del Teorema 1 arroja infinitas secuencias normales que no son λ-Poisson genéricas (basta tomar una distribución de esperanza λ distinta a la de Poisson).*

4.2. Demostración del Teorema 2

Preliminares

Para la demostración empleamos el siguiente resultado clásico de Pyatetskii-Shapiro [30] que puede leerse en [27, Theorem A] o [2, Theorem 1.1].

Lema 12 (Pyatetskii-Shapiro)**.** *Sea Ω un alfabeto de b símbolos, $b \geq 2$. Sea $x \in \Omega^{\mathbb{N}}$. Si existe una constante positiva C tal que para cada $\ell \in \mathbb{N}$ y para cada palabra finita $w \in \Omega^{\ell}$,*

$$\limsup_{N\to\infty} \frac{|x[1\ldots N]|_w}{N} \leq Cb^{-\ell},$$

entonces x es normal en base b.

Sea w una palabra fija. Definimos el conjunto $Bad(k, w, \varepsilon)$ como el conjunto de palabras de longitud k en que la frecuencia de w difiera de la frecuencia esperada por la normalidad en más de ε.

$$Bad(k, w, \varepsilon) = \left\{v \in \Omega^k : \left||v|_w - kb^{-|w|}\right| \geq \varepsilon k\right\}$$

El cardinal del conjunto $Bad(k, w, \varepsilon)$ tiene decaimiento exponencial en k. Esto fue demostrado en los primeros trabajos sobre números normales, como [11, 30], y desde entonces distintos autores han calculado cotas superiores similares. Para la demostración del Teorema 2 empleamos la siguiente versión (que puede hallarse en [3]):

Lema 13. *Consideremos un alfabeto de b símbolos. Sean k y ℓ enteros positivos, y sea ε tal que $6/\lfloor k/\ell\rfloor \leq \varepsilon \leq 1/b^{\ell}$. Entonces, para cada palabra w de longitud ℓ,*

$$|Bad(k, w, \varepsilon)| < 4\ell b^{k+\ell} e^{-b^{\ell}\varepsilon^2 k/(6\ell)}.$$

Demostración

Debemos probar que x es normal en base b, dado que para cada $i \in \mathbb{N}_0$,

$$\liminf_{k\to\infty} Z^{\lambda}_{i,k}(x) = p_i,$$

$$\sum_{i\geq 0} ip_i = \lambda.$$

Fijemos un real positivo ε. Por hipótesis sabemos que $\sum_{i\geq 0} ip_i = \lambda$. Sea i_0 tal que $\sum_{i>i_0} ip_i < \frac{\lambda\varepsilon}{2}$. Se sigue que

$$\sum_{i=0}^{i_0} ip_i > \lambda\left(1 - \frac{\varepsilon}{2}\right).$$

Sea k_0 tal que para cada $k > k_0$ y $0 \leq i \leq i_0$,

$$Z^{\lambda}_{i,k}(x) > p_i - \frac{\lambda\varepsilon}{2i_0^2}.$$

Consideremos el prefijo $x[1 \ldots \lfloor \lambda b^k \rfloor]$. Decimos que una posición en este prefijo es *problemática* si la palabra de longitud k que comienza en esa posición ocurre más de i_0 veces en el prefijo $x[1 \ldots \lfloor \lambda b^k \rfloor + k - 1]$ de x. Para cada $k > k_0$, podemos acotar el número de posiciones problemáticas entre 1 y $\lfloor \lambda b^k \rfloor$ del siguiente modo:

$$\begin{aligned}
\lfloor \lambda b^k \rfloor - \sum_{i=0}^{i_0} i \; Z^{\lambda}_{i,k}(x) b^k &< \lambda b^k - b^k \sum_{i=0}^{i_0} \left(i p_i - \frac{\lambda\varepsilon}{2i_0} \right) \\
&< \lambda b^k + \frac{b^k \lambda\varepsilon}{2} - b^k \sum_{i=0}^{i_0} i p_i \\
&< \lambda b^k + \frac{b^k \lambda\varepsilon}{2} - b^k \lambda \left(1 - \frac{\varepsilon}{2} \right) \\
&< \lambda\varepsilon b^k.
\end{aligned}$$

Ahora cubrimos las posiciones de 1 a $\lfloor \lambda b^k \rfloor + k - 1$ con palabras de longitud k que no se superpongan, de modo que ninguna palabra comienza en una posición problemática, y toda posición no problemática está cubierta por exactamente una palabra. Nos referimos a estas palabras como *palabras cubridoras*. Notemos que una palabra cubridora podría contener posiciones problemáticas, en tanto no sean la primera.

Dada una palabra cualquiera w, las ocurrencias de w en $x[1 \ldots \lfloor \lambda b^k \rfloor + k - 1]$ pueden clasificarse en las siguientes cuatro categorías:

- ocurrencias de w comenzando en una posición problemática. El número total de tales ocurrencias está acotado por el número de posiciones poblemáticas, que es a lo sumo $\lambda\varepsilon b^k$ (cuando $k > k_0$).
- ocurrencias de w que no se encuentran completamente contenidas en una palabra cubridora. Como hay a lo sumo $k^{-1}\lambda b^k$ palabras cubridoras, el número de estas ocurrencias está acotado por $|w| k^{-1} \lambda b^k$.
- ocurrencias de w contenidas en una palabra cubridora que está en el conjunto $Bad(k, w, \varepsilon)$. Cada palabra cubridora puede ocurrir a lo sumo i_0 veces, y puede contener a lo sumo k ocurrencias de w. Entonces hay a lo sumo $i_0 k |Bad(k, w, \varepsilon)|$ ocurrencias de w en esta categoría. Notemos que para k suficientemente grande, $\varepsilon \geq 6/\lfloor k/|w| \rfloor$, por lo que podemos usar la cota del Lema 13.
- ocurrencias contenidas en una palabra cubridora que no está en $Bad(k, w, \varepsilon)$. Cada una de tales palabras contiene a lo sumo $kb^{-|w|} + \varepsilon k$ ocurrencias de w. Como hay

a lo sumo $k^{-1}\lambda b^k$ palabras cubridoras, el número total de ocurrencias en este caso es como máximo $\lambda b^k(b^{-|w|}+\varepsilon)$.

Combinando las cotas para cada categoría obtenemos una cota para la cantidad de ocurrencias de w en $x[1\dots\lfloor\lambda b^n\rfloor+k-1]$,

$$\frac{|x[1\dots\lfloor\lambda b^k\rfloor+k-1]|_w}{\lambda b^k}\leq\varepsilon+\frac{1}{k}|w|+\frac{4i_0|w|b^{|w|}}{\lambda}ke^{-\varepsilon^2kb^{|w|}/(6|w|)}+b^{-|w|}+\varepsilon.$$

Tomando límite superior,

$$\limsup_{k\to\infty}\frac{|x[1\dots\lfloor\lambda b^k\rfloor+k-1]|_w}{\lambda b^k}\leq 2\varepsilon+b^{-|w|}.$$

Como esto vale para cada $\varepsilon\leq 1/b^{|w|}$, se sigue que

$$\limsup_{k\to\infty}\frac{|x[1\dots\lfloor\lambda b^k\rfloor+k-1]|_w}{\lambda b^k}\leq b^{-|w|}.$$

Para mostrar que x es normal, aplicamos el Lema 12. Fijemos N y sea k tal que $\lambda b^{k-1}\leq N<\lambda b^k$. Usando entonces las cotas obtenidas antes:

$$\limsup_{N\to\infty}\frac{|x[1\dots N]|_w}{N}\leq\limsup_{n\to\infty}\frac{|x[1\dots\lfloor\lambda b^k\rfloor+k-1]|_w}{\lambda b^{k-1}}\leq b^{1-|w|}.$$

Concluimos así que x es normal en base b.

Capítulo 5

Secuencias infinitas cuasi De Bruijn

En este capítulo introducimos la noción de *secuencias infinitas cuasi De Bruijn*. Si fuésemos capaces de obtener una tal secuencia, podríamos usarla para construir secuencias λ- Poisson genéricas, sin restricciones sobre el valor de λ, incluso en base 2.

5.1. Secuencias infinitas cuasi De Bruijn

Hasta donde tenemos conocimiento, en la literatura no se ha introducido previamente la siguiente definición:

Definición 6. *Dado Ω un alfabeto de b símbolos, decimos que una secuencia $x \in \Omega^{\mathbb{N}}$ es infinita cuasi De Bruijn si satisface que*

$$\lim_{k\to\infty} Z^1_{1,k}(x) = 1,$$

es decir, si la proporción de palabras que aparecen exactamente una vez en los prefijos de longitud $b^k + k - 1$ converge a 1.

En primer lugar observemos que toda secuencia infinita de De Bruijn en un alfabeto de 3 o más símbolos es una secuencia infinita cuasi De Bruijn. En el caso $b = 2$ la definición toma mayor sentido: una secuencia infinita de De Bruijn en base 2 (de acuerdo a la Defincion 3) no tiene por qué ser una secuencia infinita cuasi De Bruijn, y tampoco vale a la inversa.

¿Por qué introducimos esta definición? Puede verificarse fácilmente el siguiente resultado:

Observación 3. *Es posible reproducir en base 2 la construcción del Teorema 1 para bases $b \geq 3$, empleando como fuente de bloques una palabra A infinita cuasi De Bruijn. De esta forma, no es necesario requerir que $p_0 \geq 1/2$ para base 2 y puede obtenerse una secuencia λ-Poisson genérica en base 2 para cualquier $\lambda > 0$.*

Para probarlo, basta con agregar en la demostración del Teorema 1 una tercera fuente de error, correspondiente a aquellas palabras que no aparecen exactamente una vez en $A[1 \dots 2^k]$. Por definición, ese error convergiría a cero.

Surge entonces naturalmente la necesidad de resolver el siguiente problema:

Problema. ¿Es posible construir una secuencia infinita cuasi De Bruijn en un alfabeto de dos símbolos?

Si bien no fuimos capaces de resolverlo, exponemos a continuación una serie de resultados parciales que pueden tener interés por sí mismos y servir para trabajos futuros. En lo que sigue Ω será un alfabeto de dos símbolos.

Como primera observación, veamos que resulta posible dar dos construcciones bastante simples de secuencias $x^1, x^2 \in \Omega^{\mathbb{N}}$, que garantizan respectivamente:

- $\liminf_{k\to\infty} Z^1_{1,k}(x^1) \geq 1/2$, y
- $\liminf_{k\to\infty} Z^1_{1,k}(x^2) \geq 2/3$.

Podemos tomar x_1 como una secuencia infinita de De Bruijn en base 2 (Definición 3). En ese caso, se tiene que

$$Z^1_{1,2k-1}(x^1) = 1.$$

Además, como $x^1[1 \dots 2^{2k-1} + 2k - 2]$ no repite palabras de longitud $2k - 1$, tampoco repite palabras de longitud $2k$, y por lo tanto, en $x^1[1 \dots 2^{2k} + 2k - 1]$ aparece al menos la mitad de las palabras de longitud $2k$:

$$Z^1_{1,2k}(x^1) \geq 1/2.$$

Se deduce entonces que

$$\liminf_{k\to\infty} Z^1_{1,k}(x^1) \geq 1/2,$$

como queríamos.

Para construir x^2, partimos de una secuencia infinita de De Bruijn en base 2, B. A continuación, para cada k, dividimos la subsecuencia $B[2^{2k-1}+1 \dots 2^{2k+1}]$ en 3 palabras consecutivas de longitud $2^{2k-1} = \frac{2^{2k+1}-2^{2k-1}}{3}$. Como la secuencia $B[1 \dots 2^{2k-1}]$ es de De Bruijn de orden $2k - 1$, no repite palabras de orden $2k$, y faltan 2^{2k-1} palabras de longitud $2k$ para completar todas. Como sabemos que $B[1 \dots 2^{2k+1}]$ es de De Bruijn de orden $2k + 1$, allí aparecen todas las palabras de longitud $2k + 1$, y en particular, todas las de longitud $2k$. Por lo tanto, en $B[2^{2k-1} + 1 \dots 2^{2k+1}]$ aparecen las 2^{2k-1} palabras de longitud $2k$ que no están en $B[1 \dots 2^{2k-1}]$. Pero entonces alguna de las tres secciones de la división que hicimos tiene que tener al menos $2^{2k-1}/3$ de esas palabras. Tomamos esa sección y la intercambiamos con la que aparecía en primer lugar, como se indica

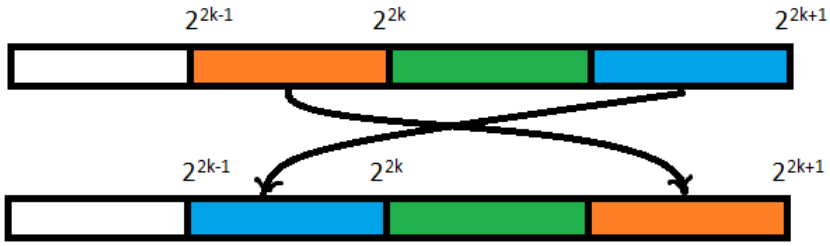


Figura 5.1: Construcción de x^2

en la Figura 5.1. De esta forma, en el prefijo $B[1 \dots 2^{2k}]$ garantizamos que aparecen al menos $\frac{2^{2k-1}+2^{2k-1}/3}{2^{2k}} = 2/3$ de todas las palabras de longitud $2k$.

Repetimos este procedimiento para cada k. Para realizar los conteos de palabras e intercambios de bloques asumimos que las modificaciones anteriores no fueron realizadas.

Puede verificarse fácilmente que la palabra x^2 así construída satisface que

$$\liminf_{k\to\infty} Z^1_{1,k}(x^2) \geq 2/3.$$

5.2. Ciclos en grafos de De Bruijn

Una estrategia posible para obtener secuencias infinitas cuasi De Bruijn consiste en realizar un proceso iterativo de extensiones análogo al descripto en la demostración del Lema6 en el Capítulo 1. El principal inconveniente es que los grafos remanentes no son conexos. Frente a esta situación, surge la necesidad de estudiar en mayor profundidad la cantidad de componentes conexas del grafo remanente de una secuencia de De Bruijn, y técnicas de unión de ciclos en grafos de De Bruijn que permitan eventualmente unir estas componentes.

Ciclos en grafos de De Bruijn

Los ciclos PCR y CCR

Presentamos a continuación dos descomposiciones distintas de los vértices del grafo de De Bruijn G_n en ciclos simples disjuntos, que resultarán de relevancia en la próxima sección.

La primera descomposición se conoce como PCR (pure cycling register), y se construye del siguiente modo: el sucesor de un vértice $v = a_1 \dots a_n$ es $a_2 \dots a_n a_1$.

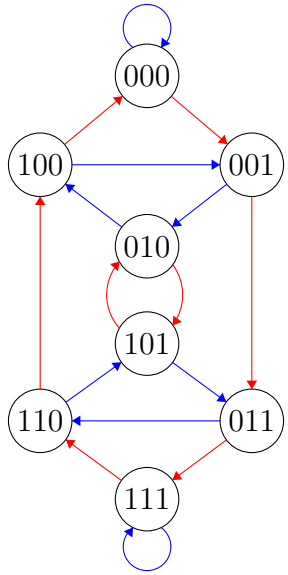


Figura 5.2: Ciclos en G_3. En azul, ciclos del PCR. En rojo, ciclos del CCR.

Puede demostrarse (utilizando por ejemplo el Lema de Burnside, ver [18]) que la cantidad de ciclos $\mathcal{Z}(n)$ del PCR es:

$$\mathcal{Z}(n) = \frac{1}{n} \sum_{d|n} \varphi(d) 2^{\frac{n}{d}},$$

donde $\varphi(k)$ es la función indicatriz de Euler que cuenta la cantidad de números coprimos con k menores que k. Notar que $\mathcal{Z}(n) \geq 2^n/n$.

La segunda descomposición se conoce como CCR (complemented cycling register), y se construye con la siguiente regla: el sucesor de un vértice $v = a_1 \dots a_n$ es $a_2 \dots a_n \overline{a_1}$.

Puede demostrarse que la cantidad de ciclos $\mathcal{Z}^*(n)$ del CCR es:

$$\mathcal{Z}^*(n) = \frac{1}{2}\mathcal{Z}(n) - \frac{1}{2n} \sum_{2d|n} \varphi(2d) 2^{\frac{n}{2d}}.$$

Notar que $\mathcal{Z}^*(n) = \frac{1}{2}Z(n)$ si n es impar.

En la Figura 5.2 se observa la descomposición en ciclos de PCR y CCR del grafo G_3.

Unión de ciclos

Una forma de construir ciclos en grafos de De Bruijn consiste en unir ciclos más pequeños de vértices disjuntos. Se trata de un método conocido, utilizado por ejemplo por Golomb en su libro clásico del área [18] y por Lempel en [24].

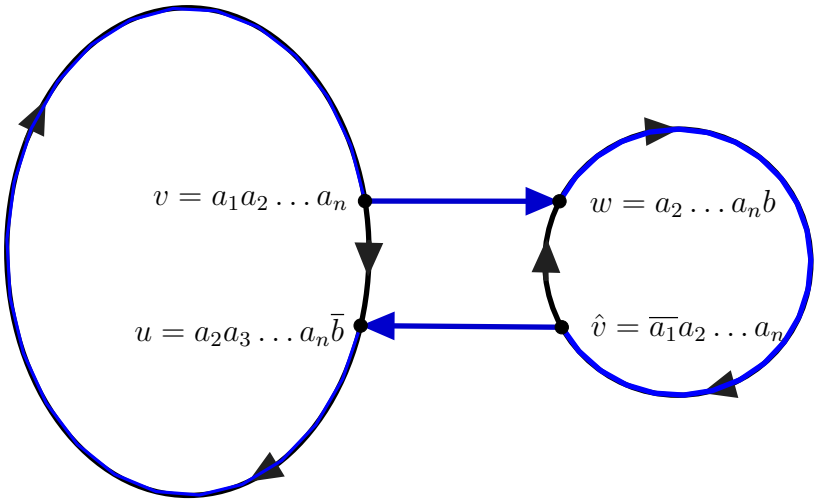


Figura 5.3: Unión de ciclos en G_n. En azul, el ciclo unión.

Imitando a Lempel, decimos que el vértice *conjugado* de un vértice $v = a_1a_2 \dots a_n$ es $\hat{v} = \overline{a_1}a_2 \dots a_n$. Decimos que dos ciclos C_1 y C_2 en G_n, sin aristas en común, son *adyacentes*, si existe un vértice v en C_1 cuyo conjugado $\hat{v}$ está en C_2. En la Figura 5.3 puede observarse un ejemplo de dos ciclos conjugados. Es inmediato verificar que si dos ciclos son conjugados, siempre habrá cuatro vértices distinguidos u, v, $\hat{v}$ y w como los de la Figura 5.3, con u y v en C_1, $\hat{v}$ y w en C_2, tales que v se une a u y w, y $\hat{v}$ se une a w y a u. El ciclo unión de C_1 y C_2 se obtiene tomando a w como sucesor de v en el ciclo, y a u como sucesor de $\hat{v}$ en el ciclo.

Observemos que dada una descomposición de los vértices de G_n en ciclos simples disjuntos, es posible unirlos para formar un ciclo hamiltoniano utilizando repetidas veces la técnica anterior. Lo hacemos en pasos: mientras que queden al menos dos ciclos, como G_n es fuertemente conexo, hay un camino dirigido que permite llegar de un ciclo C_1 a otro C_2. Esto nos dice que, si bien es posible que el camino tenga longitud mayor que 1, hay al menos una arista que sale de un vértice de C_1 y llega a un vértice de un ciclo distinto a C_1, digamos C_3. Se ve entonces que C_1 y C_3 tienen dos vértices conjugados y es posible aplicar el procedimiento anterior. De esta forma, en cada paso la cantidad de ciclos disminuye en uno, hasta que finalizamos con un único ciclo hamiltoniano.

Componentes conexas del grafo remanente

En la Observación 1 del Capítulo 1 probamos que al eliminar las aristas de cualquier ciclo hamiltoniano en G_n, el grafo resultante, al que denominamos *grafo remanente*, es siempre disconexo. En esta sección intentamos arrojar un poco de luz sobre las

características del grafo remanente, en particular la cantidad de componentes conexas que posee.

Una cota para la cantidad de componentes

Golomb conjeturó en [18] que la mayor cantidad de ciclos de aristas disjuntos en que puede descomponerse el grafo de De Bruijn G_n es exactamente $\mathcal{Z}(n)$. Lempel planteó una generalización de esta conjetura en [25], que fue finalmente probada por Mykkeltveit en [28]:

Teorema. *El mínimo número de vértices que puede removerse*[1] *del grafo G_n de forma que no contenga ningún ciclo es $\mathcal{Z}(n)$.*

Lempel llamó V-conjunto a un conjunto de vértices de G_n que, al ser eliminados, dejan al grafo sin ciclos. El teorema anterior afirma entonces que es posible hallar un V-conjunto de $\mathcal{Z}(n)$ elementos, y además, que es el tamaño de V-conjunto más pequeño posible. Se deducen los siguientes dos corolarios, el primero de los cuales es la conjetura de Golomb:

Corolario. *La máxima cantidad de ciclos de vértices disjuntos en que pueden descomponerse los vértices de G_n es exactamente $\mathcal{Z}(n)$.*

Demostración. Basta con observar que un V-conjunto debe tener al menos un vértice en cada ciclo de la descomposición. Por lo tanto, la cantidad de ciclos es a lo sumo el tamaño más chico de V-conjunto, es decir, $\mathcal{Z}(n)$. □

Corolario. *La cantidad de componentes conexas del grafo remanente de un ciclo hamiltoniano en G_n es a lo sumo $\mathcal{Z}(n)$.*

Demostración. Basta con notar que las componentes conexas del grafo remanente forman una descomposición de los vértices de G_n en ciclos disjuntos. □

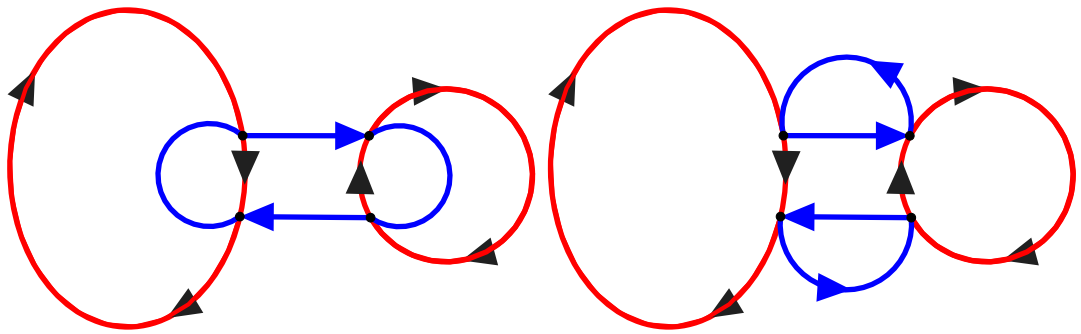
El peor caso

Si bien vimos que $\mathcal{Z}(n)$ es una cota superior para la cantidad de componentes conexas del grafo remanente, ¿se alcanza este peor caso? Es decir: ¿es posible hallar un ciclo hamiltoniano en G_n tal que el grafo remanente correspondiente tenga $\Theta(\mathcal{Z}(n))$ componentes conexas[2]? A continuación probamos que la respuesta a esta pregunta es afirmativa.

Consideremos la descomposición en ciclos CCR de los vértices del grafo G_n. Aplicando la técnica de unión de ciclos presentada en la sección previa, se obtiene un ciclo

[1] Al eliminar un vértice en un grafo, eliminamos también las aristas que llegan y salen de él.

[2] La notación asintótica $f \in \Theta(g(n))$ indica que existen constantes c_1 y c_2 tales que $f(n) \leq c_1 g(n)$ y $g(n) \leq c_2 f(n)$ para n suficientemente grande.



(a) Caso 1 (b) Caso 2

Figura 5.4: Unión de ciclos CCR. En rojo, los ciclos a ser unidos. En azul, parte del grafo remanente.

hamiltoniano. ¿Cuál es la cantidad de componentes conexas del grafo remanente de este hamiltoniano?

En cada paso de unión de ciclos pueden ocurrir dos cosas: o bien las dos aristas que conectan los ciclos pertenecen a un mismo ciclo de las aristas remanentes (Caso 1), o bien pertenecen a ciclos distintos de las aristas remanentes (Caso 2). En la Figura 5.4 están representados gráficamente los dos casos. En el primero, la cantidad de ciclos del remanente aumentará en uno (el ciclo azul se divide en dos cuando se unen los ciclos rojos). En el segundo caso, la cantidad de ciclos del remanente disminuirá en uno (los ciclos azules se unen en uno solo).

Al comienzo del proceso de unión, las aristas remanentes que no pertenecen a los ciclos CCR son exactamente las de los ciclos PCR. Por lo tanto, al inicio hay $\mathcal{Z}(n)$ ciclos en el remanente. En cada paso, la cantidad de ciclos del remanente aumenta o disminuye en uno. Como se realizan $\mathcal{Z}^*(n)-1$ pasos de unión, se deduce que al finalizar, el remanente del hamiltoniano obtenido tiene al menos

$$\mathcal{Z}(n) - (\mathcal{Z}^*(n) - 1) = \frac{1}{2}\mathcal{Z}(n) + 1 + \frac{1}{2n}\sum_{2d|n} \varphi(2d)2^{\frac{n}{2d}} \geq \frac{1}{2}\mathcal{Z}(n)$$

componentes conexas. Concluimos entonces que el peor caso se alcanza (en el sentido asintótico que dijimos antes).

El mejor caso

Por la Observación 1, sabemos que la menor cantidad de componentes conexas que puede tener el grafo remanente de un ciclo hamiltoniano es 3. ¿Puede alcanzarse este

óptimo? Es decir, ¿siempre existe algún ciclo hamiltoniano en G_n cuyo remanente tenga tres componentes conexas?

La respuesta a esta pregunta es que sí. En su trabajo no publicado [26], McConnell llama a una secuencia que satisface esa condición *doble hélice* y demuestra que cualquier ciclo hamiltoniano generado por un LFSR de período máximo[3] es una doble hélice. Un resultado análogo fue demostrado mucho antes por Rowley y Bose en [32], aunque con una formulación ligeramente distinta.

5.3. Principales dificultades

Teniendo en cuenta los resultados de la sección anterior y la idea de imitar la construcción de las secuencias infinitas de De Bruijn, podemos intentar alguno de los caminos de acción que detallamos a continuación. En lo que sigue, decimos que el error absoluto de la construcción en el paso k es la cantidad de palabras de longitud k que no aparecen exactamente una vez en el prefijo de longitud 2^k construido. El error relativo en el paso k es el error absoluto dividido 2^k.

Una primera opción es concatenar las palabras asociadas a cada componente conexa del grafo remanente. Sin embargo, en el peor de los casos, en el paso k hay que pagar un error de k palabras por cada concatenación (si las componentes tienen tamaño menor que k, en realidad la cota para el error es el tamaño de la componente), y como $\mathcal{Z}(k) \geq 2^k/k$, no podríamos garantizar que el error converja a cero.

En lugar de concatenar las palabras asociadas a cada componente conexa, podemos intentar utilizar el método de unión de ciclos descripto en la sección anterior. Por cada unión de dos componentes, utilizaríamos dos aristas que ya habían sido usadas previamente, y dejaríamos dos sin usar. Entonces el error absoluto sería del orden de $C\mathcal{Z}(k)$ con C constante, y no resulta difícil demostrar que

$$\lim_{k\to\infty} \frac{\mathcal{Z}(k)}{2^k} = 0.$$

Si bien el error parece tener un orden de crecimiento adecuado, aparecen otras dificultades. Después del primer paso de unión de componentes, al considerar el ciclo construido en el grafo de orden siguiente, no resulta obvio cómo repetir el proceso: una vez que unimos las componentes conexas del grafo remanente en G_n, algunas aristas serán recorridas dos veces, y por lo tanto, el ciclo visto en G_{n+1} repetiría vértices. Por lo tanto, es posible que haya vértices en G_{n+1} que no pertenezcan al grafo remanente en el próximo paso de la construcción. Esto trae aparejado que pueda haber componentes conexas del grafo remanente que no pueden conectarse a ninguna otra como en la Figura

[3] Linear Feedback Shift Register. Los LFSR constituyen uno de los métodos clásicos de generación de secuencias de De Bruijn. No entramos en más detalles por exceder los límites de este trabajo, pero referimos al lector al libro clásico de Golomb [18].

5.3, y por lo tanto, que no pueda llevarse a cabo el proceso de unión de ciclos como lo vimos.

Incluso aunque hubiese un modo de realizar la unión de ciclos, aparece una cuestión adicional a tener en cuenta: no podemos garantizar que no se repitan caminos de longitud dos luego de realizada la unión de componentes. Esto implica que al pasar al grafo de orden siguiente, pueden repetirse aristas, y por lo tanto, al error de cada paso contribuyen los errores anteriores (una arista repetida en G_n se corresponde con una palabra que aparece más de una vez en la secuencia que está siendo construida). Es posible demostrar que

$$\lim_{k\to\infty} \frac{1}{2^k} \sum_{j=1}^{k} \mathcal{Z}(j) = 0.$$

Esto dice que si fuera posible hacer una construcción en que el error absoluto en cada paso fuera aproximadamente $\mathcal{Z}(k)$, entonces el error relativo total convergería a cero. Sin embargo, si el error fuese un poco mayor en cada paso, podríamos perder control sobre él.

Nos encontramos entonces con dos grandes dificultades: por un lado, el control del error de una posible construcción; por el otro, cómo proceder ante la falta de garantías sobre el prefijo ya construído a partir del segundo paso.

Otro posible camino de acción es utilizar siempre secuencias generadas por un LFSR de período máximo, que sabemos que tienen grafos remanentes con pocas componentes conexas. Sin embargo, no parece evidente cómo sería posible conseguirlo. No tenemos garantías de que la extensión de una secuencia generada por un LFSR corresponda a otra generada por un LFSR. De hecho, la cantidad de estas secuencias (de orden k) es $\varphi(2^k - 1)/k$, a comparación de las $2^{2^k - k}$ secuencias posibles de De Bruijn de orden k.

Para concluir, comentamos al respecto de los ciclos en grafos de De Bruijn. Se sabe muy poco acerca de las distribuciones de las longitudes de los ciclos generados por NLFSRs[4] en los grafos de De Bruijn[5]. Golomb realizó un estudio estadístico de las longitudes y cantidades de ciclos en [18], y propuso un modelo simplificado para estudiarlas. En su modelo, que se adecúa muy bien a datos reales, la cantidad esperada de ciclos generados por un NLFSR es

$$\ln(2^k) + \gamma + o(1) < k + 1 + o(1),$$

donde γ es la constante de Euler $0.5771\ldots$ Es decir, la cantidad esperada de ciclos parece ser muchísimo más pequeña que la dada por el peor caso. Algo similar podría ocurrir con la cantidad de ciclos del grafo remanente. Tener algún tipo de control más fino que el dado por la cota superior del peor caso $\mathcal{Z}(k)$ podría permitir dar construcciones para las que sea posible demostrar que su error converge a cero.

[4]Non Linear Feedback Shift Registers

[5]El estudio específico de ciclos de grafos remanentes de un ciclo hamiltoniano no parece haber sido siquiera abordado en la literatura

5.4. La secuencia de Ehrenfeucht-Mycielski

Para concluir, introducimos la secuencia de Ehrenfeucht-Mycielski. A partir de resultados experimentales, resulta una candidata prometedora a ser una secuencia infinita cuasi De Bruijn. Sin embargo, demostrarlo no parece ser una tarea nada fácil.

Para definir la secuencia, pensemos en lo siguiente: ¿Qué posibles estrategias pueden diseñarse para predecir el próximo dígito de una secuencia binaria a partir de las anteriores? Una forma posible sería la siguiente: si abrigamos la esperanza de que la secuencia tenga algo de regularidad, podemos buscar el sufijo más largo del fragmento conocido de la secuencia, que ya haya aparecido previamente. A continuación, buscamos la anteúltima aparición de ese sufijo, y tomamos como predicción el bit que viene a continuación. Notemos que si la secuencia es eventualmente periódica, este método eventualmente adivina correctamente todos los bits.

Ehrenfeucht y Mycielski definieron en [14] una secuencia binaria, a la que llamamos EM, que es la "más impredecible" respecto a la estrategia descripta en el párrafo anterior.

Definición. *EM se construye de forma recursiva. Comenzamos con el prefijo* 010*. Para definir el dígito n-ésimo, se procede de la siguiente forma: se busca el sufijo de mayor longitud $EM[n-1-i \dots n-1]$ que ya apareció antes del final. Para la anteúltima aparición de ese sufijo $EM[n-1-i-j \dots n-1-j]$, se toma el dígito siguiente $EM[n-j]$. El dígito n-ésimo $EM[n]$ será el complemento, es decir $1-EM[n-j]$.*

¿Es EM una secuencia con buenas propiedades de aleatoriedad? Es posible probar sin demasiado esfuerzo que todas las cadenas finitas aparecen en la secuencia EM infinitas veces. Sin embargo, hasta el día de hoy no se ha podido demostrar siquiera la conjetura, formulada por los mismos Ehrenfeucht y Mycielski, que establece que las frecuencias de ceros y unos en el límite convergen a $1/2$. Se han hecho pocos avances en dirección a probar este resultado, uno de los más significativos el de Kieffer y Szpankowski en [22], donde demostraron que la frecuencia límite de unos, si existe, está entre $1/4$ y $3/4$.

Existe fuerte evidencia empírica que apoya no solo la hipótesis de balance de ceros y unos en EM, sino que también satisface condiciones mucho más fuertes, como la normalidad. Aún más, en [34], Sutner reporta que casi todas las palabras de longitud n aparecen en el prefijo de longitud 2^n en EM. Afirma que los prefijos de longitud exponencial se comportan casi como secuencias de De Bruijn. Pareciera entonces que EM es una gran candidata a ser una secuencia infinita cuasi De Bruijn. En la Figura 5.5 está representada la cantidad de palabras distintas de longitud k en cada prefijo de EM, para $k \in \{9, 10, 11\}$ (extraído de [34]).

Herman y Soltys demostraron en [21] una cota para la longitud del prefijo más pequeño de EM en el que aparecen todas las secuencias de longitud k. Esta cota depende de la función de Ackermann, y como los mismos autores indican, no es muy

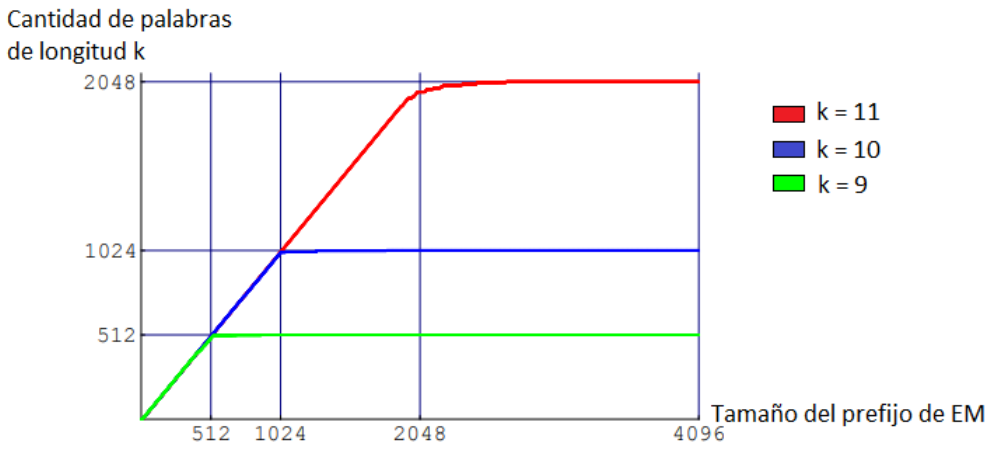


Figura 5.5: Cantidad de palabras de longitud k que aparecen en cada prefijo de EM.

buena dado que resultados experimentales sugieren que el tamaño del menor prefijo es exponencial en k. Sin embargo, es la mejor cota que se conoce hoy en día[6].

En [34], Sutner establece una relación entre los grafos de De Bruijn y la construcción de la secuencia EM, y plantea algunas preguntas abiertas sobre EM en términos de los grafos de De Bruijn. La respuesta a estas preguntas podría terminar de resolver por ejemplo la conjetura de balance. No entramos en más detalles al respecto por exceder los límites de este trabajo.

[6]La función de Ackermann es una función con velocidad de crecimiento extremadamente alta, conocida entre otras cosas por tratarse de una función computable que no es primitiva recursiva. Ver por ejemplo [31].

Bibliografía

[1] Nicolás Alvarez, Verónica Becher, and Martín Mereb. Poisson generic sequences. *International Mathematics Research Notices*, rnac234, 2022.

[2] David H. Bailey and Michal Misiurewicz. A strong hot spot theorem. *Proceedings of the American Mathematical Society*, 134(9):2495–2501, 2006.

[3] V. Becher and O. Carton. Normal numbers and computer science. In V. Berthé and editors M. Rigo, editors, *Sequences, Groups, and Number Theory*, Trends in Mathematics Series. Birkhäuser/Springer, 2018.

[4] Verónica Becher and Santiago Figueira. An example of a computable absolutely normal number. *Theoretical Computer Science*, 270(1-2):947–958, 2002.

[5] Verónica Becher, Santiago Figueira, and Rafael Picchi. Turing's unpublished algorithm for normal numbers. *Theoretical Computer Science*, 377(1-3):126–138, 2007.

[6] Verónica Becher and Pablo Ariel Heiber. On extending de Bruijn sequences. *Information Processing Letters*, 111:930–932, 2011.

[7] Verónica Becher and Gabriel Sac Himelfarb. A construction of a λ - Poisson generic sequence. *Mathematics of Computation in press, to appear in 2023*. Preprint https://arxiv.org/abs/2205.03981.

[8] Jean Berstel and Dominique Perrin. The origins of combinatorics on words. *European Journal of Combinatorics*, 28(3):996–1022, 2007.

[9] E. Borel. Les probabilités denombrables et leurs applications arithmétiques. *Rendiconti del Circolo Matematico di Palermo*, 27:247–271, 1909.

[10] David Champernowne. The construction of decimals normal in the scale of ten. *Journal of London Mathematical Society*, s1-8(4):254–260, 1933.

[11] Arthur H. Copeland and Paul Erdös. Note on normal numbers. *Bulletin American Mathematical Society*, 52:857–860, 1946.

[12] Nicolaas Gover de Bruijn. A combinatorial problem. *Indagationes Mathematicae*, 8:461–467, 1946.

[13] Jean-Paul Delahaye. The Martin-Löf thesis: The identification by recursion theory of the mathematical notion of random sequence. 02 2011.

[14] Andrzej Ehrenfeucht and Jan Mycielski. A pseudorandom sequence – how random is it? *American Mathematical Monthly*, 99(4):373–375, 1992.

[15] P. Flajolet, P. Kirschenhofer, and R. F. Tichy. Deviations from uniformity in random strings. *Probability Theory and Related Fields*, 80(1):139–150, 1988.

[16] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.

[17] Harold Fredricksen. A survey of full length nonlinear shift register cycle algorithms. *SIAM Review*, 24:195–221, 1982.

[18] Solomon W. Golomb. Shift register sequences. With portions co-authored by Lloyd R. Welch, Richard M. Goldstein and Alfred W. Hales. Holden-Day Series in Information Systems. San Francisco: Holden-Day, Inc. xiv, 224 p. (1967)., 1967.

[19] Irving J. Good. Normal recurring decimals. *Journal of London Mathematical Society*, 21:167–169, 1946.

[20] Jonathan L. Gross, Jay Yellen, and Mark Anderson. *Graph theory and its applications*. Textb. Math. Boca Raton, FL: CRC Press, 3rd edition edition, 2019.

[21] Grzegorz Herman and Michael Soltys. On the Ehrenfeucht-Mycielski sequence. *Journal of Discrete Algorithms*, 7(4):500–508, 2009.

[22] John C. Kieffer and W. Szpankowski. On the Ehrenfeucht-Mycielski balance conjecture. In *2007 Conference on analysis of algorithms, AofA 07. Papers from the 13th Conference held in Juan-les-Pins, France, June 17–22, 2007.*, pages 19–28. Nancy: The Association. Discrete Mathematics & Theoretical Computer Science (DMTCS), 2007.

[23] Nikolay. M. Korobov. Concerning some questions of uniform distribution. *Izv. Akad. Nauk SSSR Ser. Mat*, 14(3):215–238, 1950.

[24] Abraham Lempel. On a homomorphism of the de Bruijn graph and its applications to the design of feedback shift registers. *IEEE Transactions on Computers*, 19:1204–1209, 1970.

[25] Abraham Lempel. On extremal factors of the de Bruijn graph. Combinat. Struct. Appl., Proc. Calgary internat. Conf. combinat. Struct. Appl., Calgary 1969, 239-240 (1970)., 1970.

[26] Terry R. McConnell. DeBruijn strings, double helices, and the Ehrenfeucht-Mycielski mechanism. arXiv 1303.6820, 2021.

[27] Nikolay G. Moshchevitin and Ilya D. Shkredov. On the Pyatetskii-Shapiro criterion of normality. *Mathematical Notes*, 73(3):539–550, 2003.

[28] Johannes Mykkeltveit. A proof of Golomb's conjecture for the de Bruijn graph. *Journal of Combinatorial Theory. Series B*, 13:40–45, 1972.

[29] Lucas Puterman. Very normal numbers, 2019. Tesis de Licenciatura.

[30] Iliá Pyatetskii-Shapiro. On the laws of distribution of the fractional parts of the exponential function. *Nauk SSR. Ser. Mat.*, 15:47–52, 1951.

[31] Raphael M. Robinson. Recursion and double recursion. *Bulletin of the American Mathematical Society*, 54:987–993, 1948.

[32] R. Rowley and B. Bose. Edge-disjoint hamiltonian cycles in de bruijn networks. In *The Sixth Distributed Memory Computing Conference, 1991. Proceedings*, pages 707–709, 1991.

[33] Zeev Rudnick and Alexandru Zaharescu. The distribution of spacings between fractional parts of lacunary sequences. *Forum Mathematicum*, 14(5):691–712, 2002.

[34] Klaus Sutner. The Ehrenfeucht-Mycielski sequence. In *Implementation and application of automata. 8th international conference, CIAA 2003, Santa Barbara, CA, USA, July 16–18, 2003. Proceedings*, pages 282–293. Berlin: Springer, 2003.

[35] Alan Turing. A note on normal numbers. In J. L. Britton, editor, *Collected Works of Alan M. Turing, Pure Mathematics*, pages 117–119. North Holland, 1992. Notes of editor, 263–265.

[36] Benjamin Weiss. Poisson generic points. Jean-Morlet Chair 2020 - Conference: Diophantine Problems, Determinism and Randomness, Centre International de Rencontres Mathématiques, November 23 to 29, 2020. Audio- visual resource: doi:10.24350/CIRM.V.19690103.

[37] Benjamin Weiss. Random-like behavior in deterministic systems, 16 June 2010. Institute for Advanced Study Princeton University USA. https://www.youtube.com/watch?v=8AB7591De68ab$_{c}hannel = InstituteforAdvancedStudy$.