Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

¿Las Secuencias de Bruijn son Brownianas?

Tesis de Licenciatura en Ciencias de la Computación

Tomás Spognardi

Director: Nicolás Álvarez

Codirector: Verónica Becher

Buenos Aires, 2026

Índice

1. Definiciones Preliminares

En esta sección, introducimos las definiciones y notaciones para las principales estructuras que vamos a tratar en el trabajo: las **cadenas** y los **grafos**.

1.1. Cadenas

Las cadenas son una de las estructuras más simples y fundamentales en la teoría de la computación.

Definición 1.1 (Cadenas y Operaciones Relacionadas)**.** Dado un **alfabeto** finito $\mathcal{A}$, una **cadena** w sobre $\mathcal{A}$ es la concatenación de una cantidad finita de elementos de $\mathcal{A}$.

Llamamos $\mathcal{A}^*$ al conjunto de todas las cadenas sobre $\mathcal{A}$.

Vamos a utilizar las siguientes convenciones:

- La **longitud** $|w|$ de una cadena w es la cantidad de símbolos que la conforman.
- La **concatenación** de una cadena $u = u_1u_2 \cdots u_n$ y una cadena $v = v_1v_2 \cdots v_m$ es la cadena:

$$uv = u_1u_2 \cdots u_nv_1v_2 \cdots v_m$$

- Una cadena $u \in \mathcal{A}^*$ es **subcadena** de otra $w \in \mathcal{A}^*$ cuando existen $x, y \in \mathcal{A}^*$ tales que $w = xuy$.
- Utilizamos la notación $w[i : j]$, con $1 \leq i \leq j \leq n$ para la subcadena comprendida entre la i-ésima y j-ésima posición (ambas inclusive) de w.
- Un **prefijo** de w es una subcadena de la forma $w[1 : j]$, y un **sufijo** es una subcadena de la forma $w[i : n]$.
- Dada una cadena $w \in \mathcal{A}^*$ y una subcadena $u \in \mathcal{A}^*$, la **cantidad de apariciones** de u en w, denotada $|w|_u$, es la cantidad de veces que u aparece como subcadena en w.

$$|w|_u = |\{i \in \{1, \cdots, |w| - |u| + 1\} \mid w[i : i + |u| - 1] = u\}|$$

En este trabajo usamos el **alfabeto binario** $\mathbb{B} = \{0, 1\}$ y las cadenas sobre este alfabeto que llamamos **cadenas binarias**.

1.2. Grafos

En este trabajo, nos enfocamos en los **digrafos** (o grafos dirigidos), que son grafos donde las aristas tienen una dirección asociada. Los grafos son esenciales tanto en la teoría de la computación como en la teoría de grafos. Nos referiremos a los digrafos simplemente como **grafos**.

Definición 1.2 (Grafo)**.** Un **grafo** G es una tupla (V, E), donde V es el conjunto de **vértices** y $E \subseteq V \times V$ es el conjunto de **aristas**.

Utlizamos la notación $u \to v$ para referirnos a la arista (u, v), y llamamos **vértice de entrada** a u y **vértice de salida** a v. Cuando las aristas se denotan como $e \in E$, usamos $head(e)$ para el vértice de entrada y $tail(e)$ para el de salida.

Definición 1.3 (Grado de Entrada y de Salida)**.** Dado un grafo $G = (V, E)$ y un vértice $v \in V$, el **grado de entrada** de v, denotado $d^+(v)$, está definido como:

$$d^+(v) = |\{u \in V \mid u \to v \in E\}|$$

Por otro lado, el **grado de salida** de v, denotado $d^-(v)$, está definido como:

$$d^-(v) = |\{u \in V \mid v \to u \in E\}|$$

Es decir, $d^+(v)$ es la cantidad de aristas que entran a v y $d^-(v)$ es la cantidad de aristas que salen de v.

Definición 1.4 (Caminos y Ciclos)**.** Un **camino** sobre un grafo $G = (V, E)$ es una secuencia de vértices $C = v_1 v_2 \cdots v_n$ donde cada par de vértices adyacentes u_i, u_{i+1} cumplen $u_i \to u_{i+1} \in E$. Un **ciclo** es un camino donde $v_1 = v_n$.

En particular, vamos a trabajar con 2 tipos de caminos/ciclos:

Definición 1.5 (Camino/Ciclo Hamiltoniano y Euleriano)**.** Un camino es **hamiltoniano** cuando pasa por cada **vértice** exactamente una vez. Por otro lado, un camino es **euleriano** cuando pasa por cada **arista** exactamente una vez.

Un ciclo **hamiltoniano** (respectivamente **euleriano**) es un camino hamiltoniano (respectivamente euleriano) que empieza y termina en el mismo vértice.

Llamamos $\mathcal{H}(G)$ al conjunto de ciclos hamiltonianos de G, y $\mathcal{E}(G)$ al conjunto de ciclos eulerianos de G.

Definición 1.6. Una arborescencia es un grafo $\mathcal{T} = (V, E)$ con un vértice distinguido $r \in V$, denominado **raíz**, donde existe un único camino entre r y cada vértice $v \in V$.

La última operación de grafos que vamos a definir es la de tomar el **grafo de línea** de un grafo.

Definición 1.7 (Grafo de Línea)**.** Dado un grafo $G = (V, E)$, el **grafo de línea** $\mathcal{L}(G) = (V', E')$ es el grafo donde V' y E' están dados por:

$$V' = \{(u, v) \mid u \to v \in E\}$$
$$E' = \{(u, v) \to (v, w) \mid u \to v, v \to w \in E\}$$

Es decir, los vértices de $\mathcal{L}(G)$ son las aristas de G, y la arista $u \to v$ está presente cuando el vértice de salida de la arista asociada a u es el vértice de entrada de la arista asociada a v.

1.3. Cadenas De Bruijn

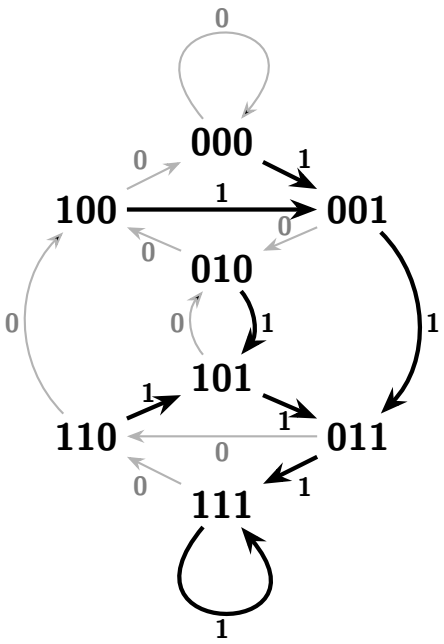

Figura 1: Representación gráfica del grafo $G_{\mathcal{DB}_3}$.

El objeto central de esta tesis son las llamadas **cadenas De Bruijn**, que llevan este nombre por el trabajo del matemático Nicolaas Govert de Bruijn[7], pero fueron formuladas múltiples veces[4]. Para ser justos, debemos mencionar también el trabajo de principios de los años 1950 de Nikolay Korobov [12] [13]. Una presentación de estas cadenas aparece en el libro de Knuth [11] y, recientemente, fue publicado el libro de Etzion [9].

A pesar de que se pueden definir para alfabetos de cualquier tamaño, este trabajo trata únicamente con cadenas De Bruijn sobre **alfabetos binarios**.

Definición 1.8 (Cadena De Bruijn binaria)**.** Una cadena w es una **cadena De Bruijn** binaria de orden n cuando cada bloque binario de longitud n tiene exactamente una aparición en $ww[1:n-1]$. En otras palabras, debe cumplir:

$$\forall b \in \mathbb{B}^n,\ |ww[1:n-1]|_b = 1$$

La concatenación del prefijo de longitud $n-1$ al final de la cadena tiene el efecto de contar bloques que aparecen "partidos" entre el principio y el final. Por ejemplo, la cadena `0110` es De Bruijn de orden 2, donde el bloque `00` aparece entre la cuarta y primera posición.

Observación 1. Una cadena De Bruijn de orden n tiene longitud 2^n, ya que cada posición debe dar inicio a un bloque de longitud n distinto, y hay 2^n bloques posibles.

A priori, no es claro que existan cadenas De Bruijn de orden n para todo n. Sin embargo, veremos una construcción que nos permite generar cadenas para cualquier orden y, más aún, todas las cadenas se pueden generar de esta forma. Para ello, vamos a definir un grafo estrechamente relacionado con dichas cadenas, que llamamos **grafo De Bruijn**.

Definición 1.9 (Grafo De Bruijn)**.** Dado $n \in \mathbb{N}$, el **grafo De Bruijn** de orden n, es un grafo $G_{\mathcal{DB}_n} = (V_n, E_n)$ dado por:

$$V_n = \mathbb{B}^n$$
$$E_n = \{a_1a_2\cdots a_{n-1}a_n \to a_2a_3\cdots a_ns \mid a_1a_2\cdots a_{n-1}a_n \in \mathbb{B}^n, s \in \mathbb{B}\}$$

Este grafo tiene un nodo por cada bloque de longitud n, y el arco $u \to v$ está presente cuando u es el bloque que se obtiene al eliminar el primer bit de v, y agregar un bit al final.

Observación 2. Para cada $n \in \mathbb{N}$, el grafo $G_{\mathcal{DB}_n}$ es **2-regular**: cada nodo tiene exactamente dos arcos salientes, uno que corresponde a agregar un 0 y otro que corresponde a agregar un 1, y dos arcos entrantes, uno que corresponde a eliminar un 0 y otro que corresponde a eliminar un 1.

2. Discrepancia de Cadenas de Bruijn

Este trabajo comenzó con el objetivo de estudiar la distribución de la **discrepancia** de las cadenas De Bruijn. Para enteder esto, primero debemos definir la noción de discrepancia:

Definición 2.1 (Discrepancia)**.** La **discrepancia de 1 símbolo**, o simplemente **discrepancia**, de una cadena w es la máxima diferencia entre la cantidad de 1s y 0s que hay entre todas sus subcadenas, donde w se considera circular. Formalmente, la podemos definir como:

$$D(w) = \max_{u \in C(w)} ||u|_1 - |u|_0|$$

Donde $C(w)$ es el conjunto de subcadenas de w, junto con las cadenas que aparecen partidas entre el final y el principio de w. Concretamente,

$$C(w) = \{u \mid u \leq w\} \cup \{zx \mid w = xyz\}$$

Esta propiedad es de interés teórico, y ha sido abordado en varios estudios recientes. En [5] se estudia la discrepancia de la cadena De Bruijn lexicográficamente mínima. Luego, en [10], los autores extienden este trabajo, estudiando la discrepancia de nuevas construcciones de cadenas De Bruijn, y presentan técnicas para obtener cadenas que alcanzan la cota asintótica mínima de $\Theta(n)$ y la máxima de $\Theta\left(\frac{2^n}{\sqrt{n}}\right)$. Más aún, en [3] se desarolla un nuevo método que permite obtener, para cualquier orden n, una cadena De Bruijn que alcanza la mínima discrepancia posible de n, estableciendo una cota inferior exacta.

En nuestro caso, buscamos estudiar el comportamiento de la discrepancia para cadenas De Bruijn muestreadas de forma uniforme.

Definición 2.2 (Variable Aleatoria D_n)**.** Sea $\mathcal{DB}_n$ el conjunto de cadenas De Bruijn de orden n, definido previamente (1.8). Si X_n es una variable aleatoria que toma como valores elementos de $\mathcal{DB}_n$ con distribución $P(X_n = w) = \frac{1}{|\mathcal{DB}_n|}$, entonces llamamos D_n a la variable dada por:

$$D_n = D(X_n)$$

Es decir, la distribución de D_n es la que resulta de tomar la discrepancia de cadenas De Bruijn muestreadas uniformemente.

Decidimos empezar nuestro estudio de esta distribución de forma empírica. Nuestro abordaje inicial fue directo: muestreamos cadenas De Bruijn, y computamos su discrepancia. Sin embargo, para que el muestreo sobre cadenas de órdenes altos resulte tratable, debimos emplear una técnica de muestreo uniforme eficiente.

2.1. Muestreo Uniforme de Cadenas De Bruijn

Para obtener un método eficiente de muestreo, se puede aprovechar la correspondencia entre cadenas De Bruijn y ciclos eulerianos: cada cadena w de orden n corresponde a un único ciclo euleriano sobre el grafo De Bruijn $G_{\mathcal{DB}_n}$ [4]. Esto implica que el problema de muestrear uniformemente sobre $\mathcal{DB}_n$ se reduce a muestrear ciclos eulerianos sobre $G_{\mathcal{DB}_n}$.

2.1.1. Teorema BEST

La enumeración de ciclos eulerianos en grafos dirigidos es un problema extensivamente estudiado. El resultado central del área es el **Teorema BEST** [16], que presenta una fórmula para contar la cantidad de circuitos eulerianos de un grafo en función de los grados de sus vértices. El nombre es un acrónimo de quiénes lo descubrieron: N. G. de Bruijn, Tatyana van Aardenne-Ehrenfest, Cedric Smith y W. T. Tutte.

Teorema 1 (Teorema BEST)**.** *La cantidad de ciclos eulerianos de un grafo euleriano conexo $G = (V, E)$ está dada por:*

$$|\mathcal{E}(G)| = |\mathcal{T}(G)| \prod_{v \in V} \big(d^-(v) - 1\big)!$$

Donde $\mathcal{T}(G)$ es el conjunto de árboles generadores de G.

El teorema BEST puede ser utilizado para contar la cantidad de cadenas De Bruijn de un orden dado: como cada nodo de $G_{\mathcal{DB}_n}$ tiene grado de salida 2, el factor $(d^-(v) - 1)! = 1! = 1 \forall v$, por lo cual se tiene $e(G) = t(G) = 2^{2^{n/2}}$. Aunque este resultado es interesante de por sí, en nuestro caso la utilidad del teorema yace en su demostración. Para ilustrar las ideas relevantes, presentamos una adaptación de un argumento en [1, p. 445–p. 447].

Demostración. Supongamos que $G = (V, E)$ es un grafo euleriano, y sea $s \to s' \in E$ una arista inicial arbitraria. Vamos a exhibir una biyección ϕ entre las arborescencias $\mathcal{T}$ con raíz en s y conjuntos de ciclos eulerianos $\mathcal{C}(\mathcal{T})$, cada uno de los cuales es de tamaño $\prod_{v \in V} (d^-(v) - 1)!$.

I. Primero, observemos que se puede construir una arborescencia con raíz en s a partir de cualquier ciclo euleriano $C \in \mathcal{E}(G)$. Para lograrlo, tomemos a la arista $s \to s' = e_1$ como el "inicio" del ciclo, y consideremos el ordenamiento $C = e_1, \cdots e_{|E|}$. Para formar una arborescencia, se pueden seleccionar las aristas $E' = \{e_i \in E \mid head(e_i) \neq s \wedge \forall i < j \leq |E|, head(e_i) \neq head(e_j)\}$, es decir, las últimas en salir de cada vértice (excepto por s). El grafo resultante $T(C) = (V, E')$ tiene $|V| - 1$ aristas, y es una arborescencia por construcción:

   - Para cada arista $u \to v \in E'$, el vértice de entrada u es único en E', porque hay una única última arista saliente de u en el ordenamiento C. Esto implica que $T(C)$ es acíclico.
   - Para cada vértice $u \in V \setminus \{s\}$, hay alguna arista $u \to v$ en E'. Combinado con el punto anterior, esto hace que todos los caminos en $T(C)$ terminen en s.

II. Por otro lado, tomemos una arborescencia T con raíz en s. Para cada v, consideramos el conjunto de aristas salientes $\{e \in E \mid head(e) = v\}$, y las ordenamos respetando 2 condiciones:

   a) Si e está en T, está última en el orden. Hay exactamente 1 arista saliente de cada nodo en el árbol, excepto por la raíz s, que no tiene ninguna.

   b) Si $v = s$, la primera arista en el orden debe ser $s \to s'$.

   Para cada $v \in V$, hay $(d^-(v) - 1)!$ órdenes que cumplen estas reglas. Si se fija un orden particular para cada nodo, se puede construir una caminata C por medio del siguiente proceso: se empieza en s y, al estar en un nodo v, se elije la primer arista dentro del orden para v que todavía no fue elegida. La caminata resultante debe cumplir:

   - C **termina en s y, como también empieza en s, es un ciclo**: Como G es euleriano, se

tiene $d^-(v) = d^+(v)\ \forall v \in V$, y esto implica que siempre que se entra a un vértice, queda alguna arista de salida; la única excepción es s, que es el vértice inicial.

- C **es un ciclo euleriano**: Sabemos que C es un ciclo, y que por construcción no repite ninguna arista. Para comprobar que es euleriano, sólo falta verificar que usa todas las aristas de E. Supongamos que no lo hace, es decir, que se tiene alguna arista $u_0 \to u_1 \notin C$. Luego, como la arista $u_0 \to u_1$ dentro de T es la última en el orden, esta tampoco debe estar en C. Como $d^-(u_1) = d^+(u_1)$, alguna arista saliente de u_1 tampoco está en C, y por consecuencia otra arista del árbol $u_1 \to u_2$ está ausente. Este argumento se puede repetir hasta llegar a la raíz s. Sin embargo, para haber construido C, se tienen que haber usado todas las aristas de entrada y salida de s. Por ende, la suposición inicial es absurda, y todas las aristas de E aparecen en C.

Llamemos $\mathcal{C}(T)$ a los ciclos eulerianos construidos a partir de T. Es claro que $|\mathcal{C}(T)| = \prod_{v \in V}(d^-(v) - 1)!$. Por otra parte, notemos que cada ciclo euleriano $C \in \mathcal{E}(G)$ está en uno, y solo uno, de los conjuntos $\mathcal{C}(T)$: de hecho, la construcción de un T' a partir del C en (I.) resulta en uno que cumple $C \in \mathcal{C}(T')$. Esto implica que los conjuntos $\mathcal{C}(T)$ particionan a $\mathcal{E}(G)$ en partes iguales, por lo cual:

$$|\mathcal{E}(G)| = |\mathcal{T}(G)| \prod_{v \in V} (d^-(v) - 1)!$$

La demostración sugiere un método para muestrear los ciclos eulerianos de un grafo $G = (V, E)$ uniformemente:

Algorithm 1 Muestreo Uniforme de Ciclos Eulerianos

Muestrear una arista inicial $s \to s' \in V$.
Muestrear una arborescencia T con raíz en s.
for all $v \in V$ **do** ▷ *Elegimos una secuencia $S(v)$ de vecinos salientes para cada v.*
    if $v = s$ **then**
        $S(s) \leftarrow s' \oplus \textsc{Shuffle}(N^+(v) \setminus s')$ ▷ *$s \to s'$ debe ser la primer arista que toma s.*
    else
        Sea $v \to p \in T$ la arista que conecta a v con su padre en T
        $S(v) \leftarrow \textsc{Shuffle}(N^+(v) \setminus p) \oplus p$ ▷ *$v \to p$ debe ser la última arista que toma v.*
Construir C por medio de una caminata por el grafo. Se debe empezar por s y, al estar en un vértice v, tomar el siguiente elemento de $S(v)$ que aún no se haya tomado.
return C

Es decir, el método consiste en muestrear primero una arborescencia T, y luego muestrear uno de los caminos de $\mathcal{C}(T)$. Como todos los conjuntos $\mathcal{C}(T)$ tienen el mismo tamaño y particionan a $\mathcal{E}(G)$, esto produce un muestreo uniforme de sus elementos.

Elegir órdenes en los cuales tomar las aristas es una tarea trivial, y por lo tanto el único componente restante para tener un algoritmo concreto sería un muestro uniforme sobre las arborescencias de G enraizadas en s.

2.1.2. Muestreo de Arborescencias

El problema de muestrear arborescencias de forma uniforme en un grafo ha sido ampliamente estudiado, y existen múltiples métodos en la literatura. En nuestro trabajo, utilizamos el *Algoritmo de Wilson* [6]: es una modificación del algoritmo de Aldous-Broder, con un tiempo de ejecución promedio de $\mathcal{O}(\tau)$, siendo

τ el tiempo de cobertura esperado del grafo de entrada. Para muestrear una arborescencia $T \in \mathcal{T}_s(G)$, el procedimiento consiste en:

Algorithm 2 Algoritmo de Wilson

```
Fijar T = (s, ∅), y elegir un orden v_1, v_2, ..., v_{n-1} para recorrer el resto de los vértices.
while |T| ≠ |G| do
    Tomar el primer v_i que aún no está en T.
    W ← [v_i]
    repeat
        Elegir un vecino aleatorio u de cabeza(W)
        Añadir u a W
    until cabeza(W) ∈ T
    T ← T ∪ LoopErase(W).
return T
```

La operación LoopErase(W) elimina los ciclos dentro de W: para cada vértice, se eliminan todas las apariciones excepto la última.

2.2. Distribución de la Discrepancia de las Cadenas De Bruijn

Una vez establecido un método eficiente para el muestreo uniforme de cadenas De Bruijn, realizamos una serie de experimentos con el fin de estudiar el comportamiento de $D(B_n)$.

2.2.1. Comparación de la Discrepancia para Distintos Órdenes

Orden	Media	Desv. est.	Mínimo	Máximo
3	3,00	0,00	3,00	3,00
4	4,50	0,50	4,00	5,00
5	6,55	0,83	5,00	8,00
6	9,44	1,45	6,00	13,00
7	13,57	2,36	7,00	22,00
8	19,39	3,67	9,00	36,00
9	27,65	5,56	12,00	56,00
10	39,32	8,15	16,00	82,00
11	55,89	11,89	23,00	124,00
12	79,33	17,11	33,00	172,00
13	112,56	24,36	46,00	247,00
14	159,45	34,55	68,00	366,00

Cuadro 1: Estadísticos básicos de la discrepancia de $\mathcal{DB}_n$ (para $n \leq 5$) o muestras de 1,000,000 elementos (para $n \geq 6$).

Para nuestro primer experimento, comparamos los valores de discrepancia obtenidos para muestreos de $\mathcal{DB}_n$ con $n \leq 14$. Para órdenes $n \leq 5$, computamos los conjuntos de forma extensiva, pero para valores mayores utilizamos una muestra de 1,000,000 elementos.

Comenzamos con un análisis descriptivo de los datos obtenidos, incluyendo una comparación entre las distribuciones de los distintos órdenes. En la tabla 1 podemos ver una comparación de sus estadísticos, incluyendo los primeros momentos y el mínimo/máximo valor encontrado.

En el cuadro 1 se puede observar un crecimiento progresivo de tanto la media como la desviación estándar a medida que aumenta el orden de las cadenas. En la figura 2 presentamos un diagrama *boxplot* que compara las distintas distribuciones, mientras que en la figura 3 se puede ver una serie de histogramas que demuestran la forma de cada una.

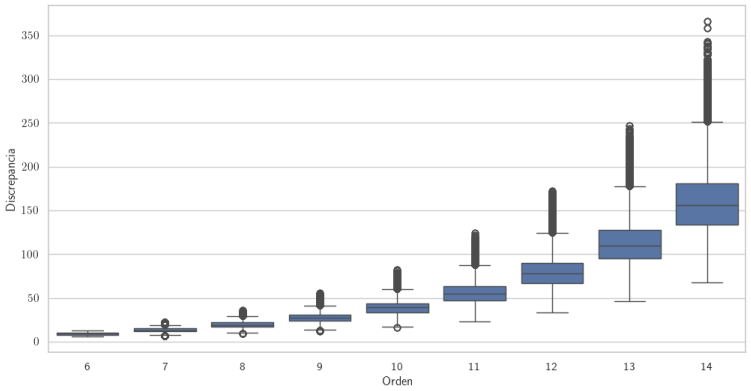

Figura 2: Boxplots de las muestras de discrepancia obtenidas para los conjuntos $\mathcal{DB}_n$ con n entre 6 y 14.

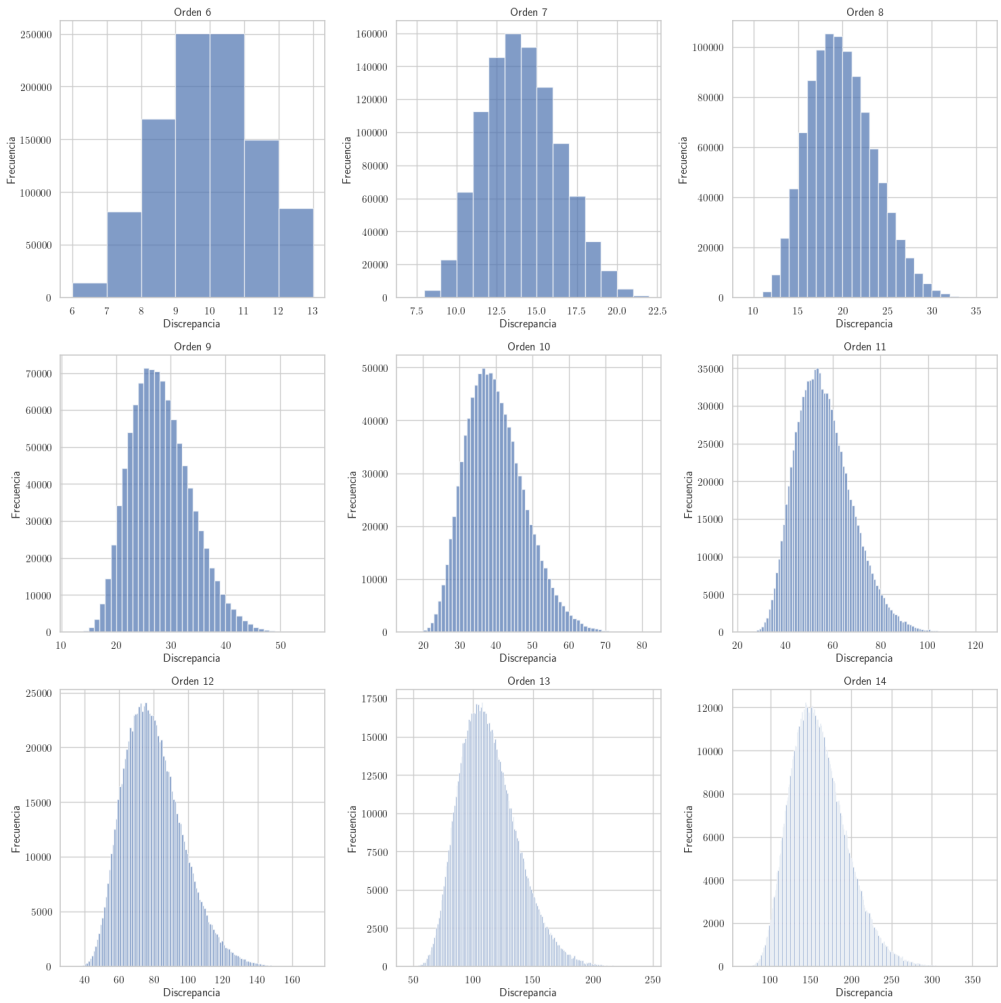

Figura 3: Histogramas de las muestras de discrepancia obtenidas para los conjuntos $\mathcal{DB}_n$ con n entre 6 y 14.

En el gráfico de histogramas de la figura 3 se puede ver que, a partir de $n = 10$, las distribuciones empiezan a tener una forma similar. Para validar esto de forma visual, graficamos las versiones **estandarizadas** de cada una. El proceso de estandarización consiste en aplicar la transformación $x \mapsto (x - \hat{\mu}_n)/\hat{\sigma}_n$ a cada valor, donde $\hat{\mu}_n$ y $\hat{\sigma}_n$ son la media y desviación estándar muestrales para cada orden n.

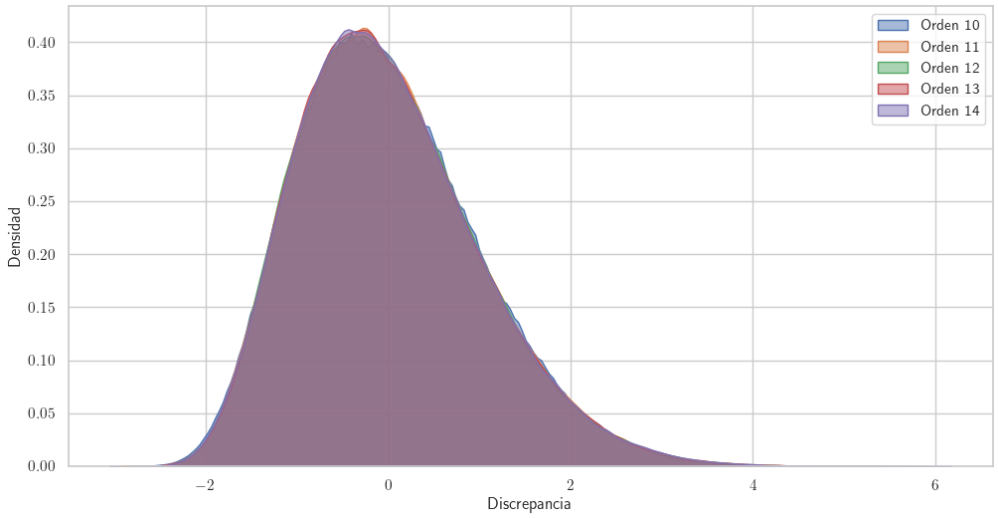

Figura 4: KDE (estimación de densidad por kernel) de las muestras de discrepancia normalizadas para cadenas De Bruijn de órdenes entre 10 y 14.

Esta nueva comparación valida nuestra observación anterior: las distribuciones en cuestión tienen la misma forma, difiriendo solamente en su escala y ubicación. Esto podría ser un indicio de que, a medida que crece el orden n, $D(B_n)$ tiende a una distribución límite (salvo algún factor de ubicación/escala). Para indagar en esto, decidimos comparar las discrepancias obtenidas con un caso más estudiado: el de cadenas binarias generales.

2.2.2. Comparación con Cadenas Generales

El conjunto más básico para el cual se puede analizar la discrepancia es $\mathbb{B}^\ell$, el conjunto de cadenas binarias de un tamaño ℓ dado. Como notamos anteriormente , $\mathcal{DB}_n \subseteq \mathbb{B}^{2^n}$, por lo cual comparamos cada orden $\mathcal{DB}_n$ con el superconjunto correspondiente $\mathbb{B}^{2^n}$. Los resultados se pueden observar en la siguiente figura:

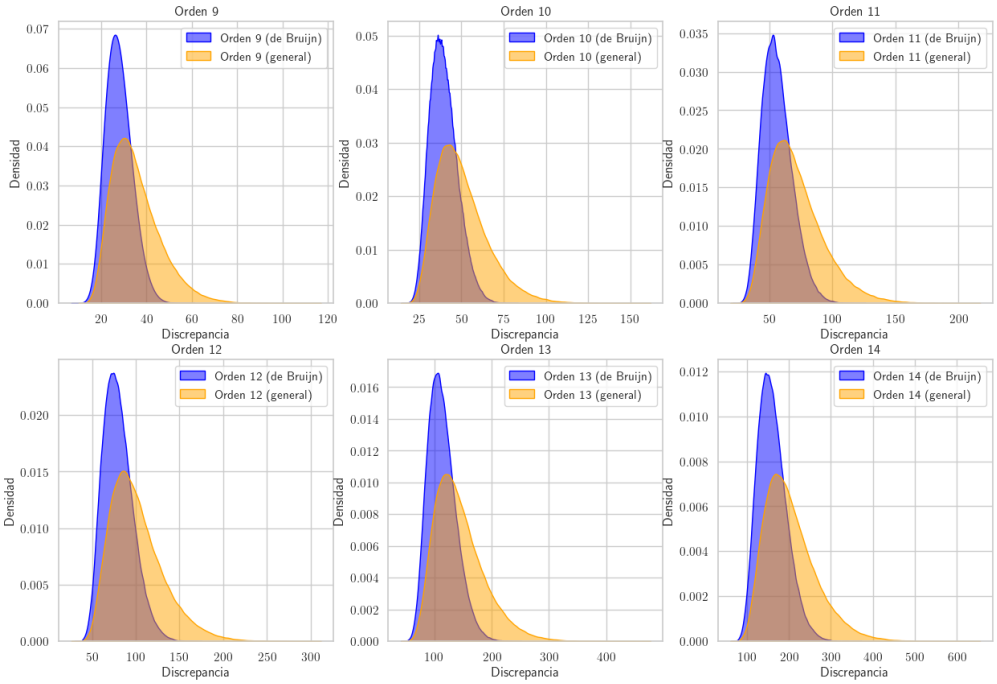

Figura 5: Comparaciones entre las KDEs de la discrepancia muestreada de los conjuntos $\mathcal{DB}_n$ y $\mathbb{B}^{2^n}$.

Se puede ver que las distribuciones para cada orden son claramente distintas: las cadenas de $\mathbb{B}^{2^n}$ tienden a tener una discrepancia mayor. A pesar de estas diferencias, las cadenas generales también parecen converger a una distribución límite. Al igual que en el caso anterior, podemos comprobar esto visualmente al estandarizar las distribuciones y graficarlas en el mismo eje.

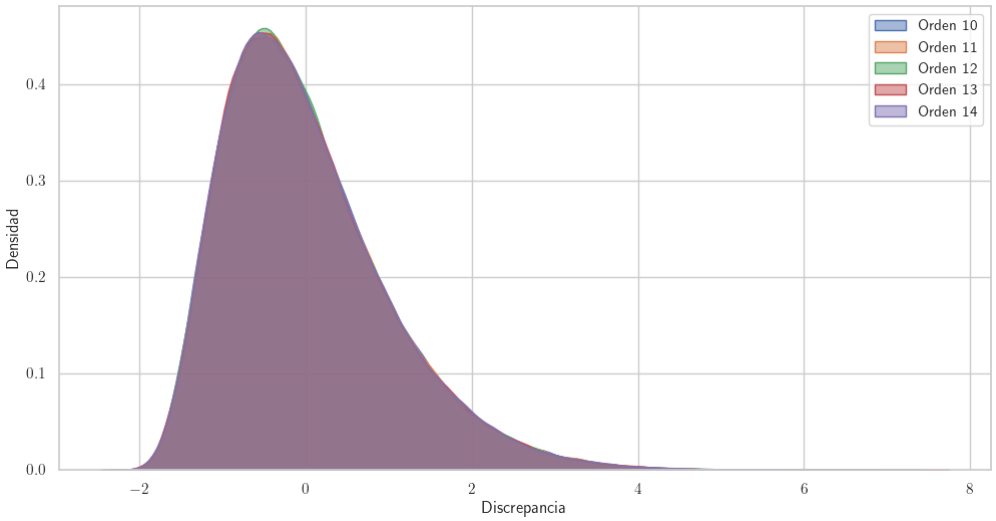

Figura 6: KDE de las muestras de discrepancia normalizadas para cadenas de los conjuntos $\mathbb{B}^{2^n}$ con n entre 10 y 14.

Efectivamente, la discrepancia de cadenas generales parece exhibir el mismo comportamiento de convergencia. Sería razonable suponer que la distribución límite es la misma que para las cadenas De Bruijn, y cada orden

difiere únicamente en su media/escala. Para comprobar esta hipótesis, podemos comparar las distribuciones estandarizadas de la discrepancia para un orden alto (en este caso, $n = 14$).

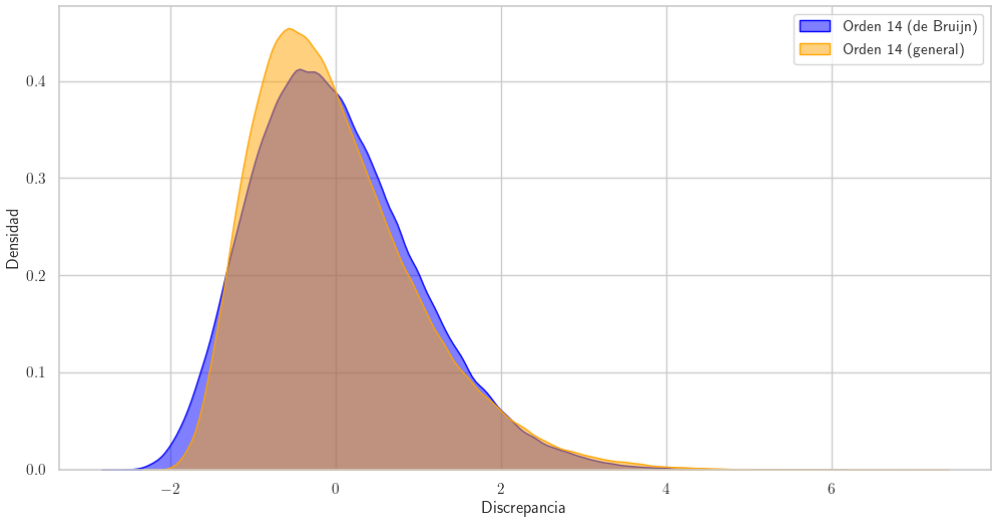

Figura 7: Comparación entre KDEs de las muestras de discrepancia para los conjuntos $\mathcal{DB}_14$ y $\mathbb{B}^{2^14}$

Como se puede ver en la figura 7, las distribuciones de ambos conjuntos de cadenas difieren aún después de ser estandarizadas.

2.2.3. Comparación con Cadenas Balanceadas

Habiendo confirmado que la discrepancia de cadenas De Bruijn sigue una distribución fundamentalmente distinta a la de las cadenas generales, decidimos comparar esta propiedad en un conjunto distinto: el de las **cadenas balanceadas**.

> **Definición 2.3.** Una cadena $w \in \mathbb{B}^*$ es **balanceada** cuando $|w|_0 = |w|_1$, es decir, tiene la misma cantidad de 0s que de 1s.

Llamaremos Bal_ℓ a la variable que toma como valores cadenas balanceadas de longitud ℓ de forma equiprobable.

Es fácil ver que, de forma análoga al caso de las cadenas generales, todas las cadenas De Bruijn son balanceadas.

Por ende, comparamos la discrepancia de las variables $\mathcal{DB}_n$ y Bal_{2^n} para distintos valores de n. Cabe destacar que el muestreo uniforme de cadenas balanceadas es trivial: simplemente se debe elegir un orden para una secuencia con 2^{n-1} 1s y 2^{n-1} 0s, lo cual se puede lograr utilizando un algoritmo de shuffling como el de Fisher-Yates [8]. Los resultados de la comparación se pueden ver a continuación:

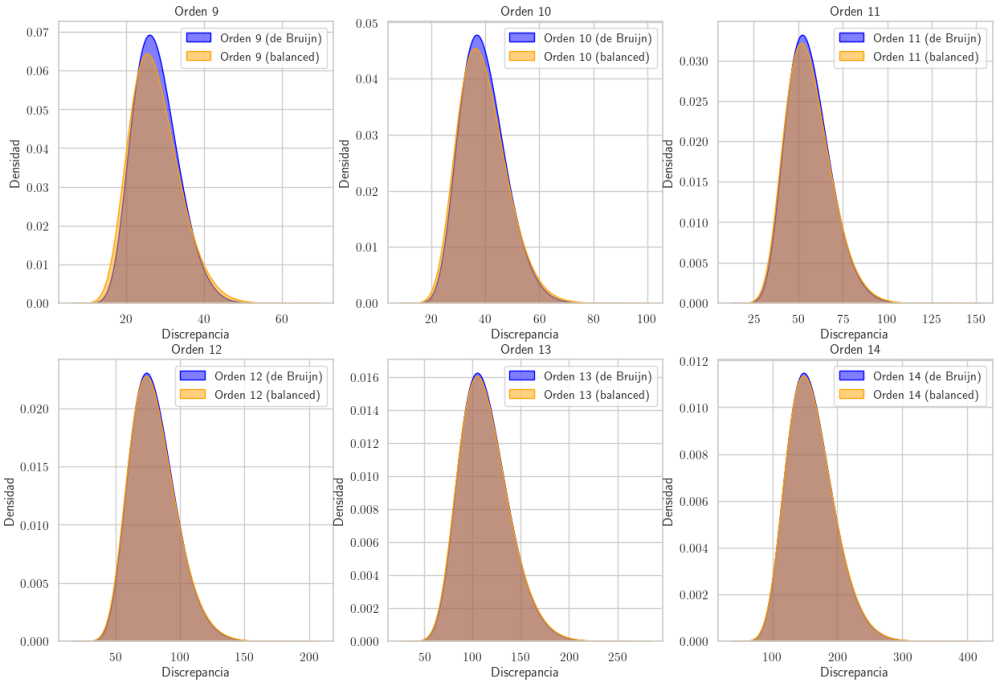

Figura 8: Comparaciones entre las KDEs de la discrepancia muestreada de los conjuntos $\mathcal{DB}_n$ y Bal_{2^n}.

En este caso, además de exhibir un comportamiento de convergencia, la discrepancia de las cadenas balanceadas parece seguir una distribución cada vez más similar a la de las De Bruijn a medida que crece el orden.

Dicha similitud puede ser cuantificada, por ejemplo, utilizando la **distancia de Wasserstein** [18], una función de distancia definida entre distribuciones que puede ser computada fácilmente. Al calcular dicha distancia entre las distribuciones de la discrepancia para cada par de conjuntos, queda claro que ésta decrece a medida que aumenta el orden n.

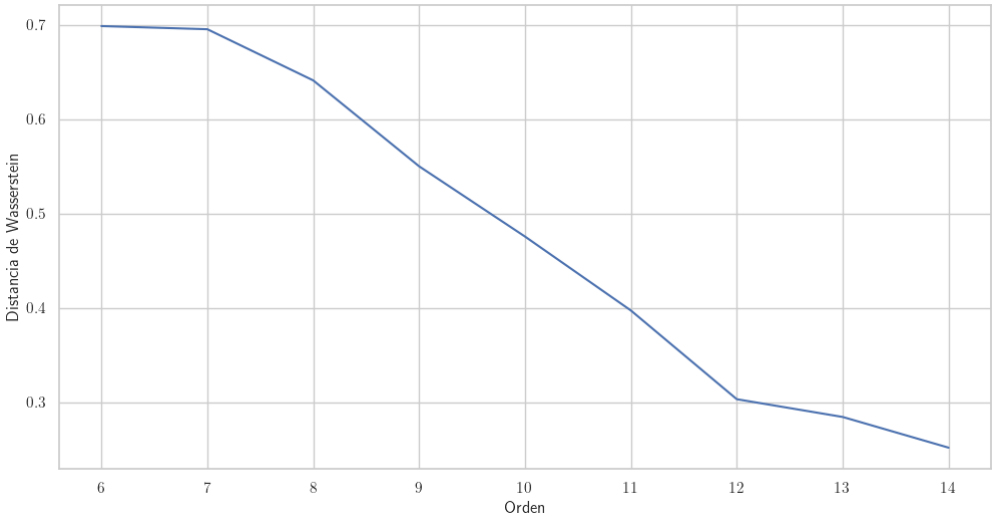

Figura 9: Evolución de la distancia de Wasserstein estimada entre las distribuciones de discrepancia de $\mathcal{DB}_n$ y Bal_{2^n}.

3. Conjetura: Las Cadenas De Bruijn son Brownianas

Las similitudes en el comportamiento de la discrepancia para $\mathcal{DB}_n$ y Bal_{2^n} nos resultaron sorprendentes, por lo cual decidimos investigar si dichos conjuntos coinciden en otros aspectos estadísticos.

Para estudiar otras propiedades sobre cadenas, es conveniente introducir una nueva perspectiva: la de los **procesos estocásticos**.

> **Definición 3.1** (Proceso Estocástico)**.** Un **proceso estocástico**, o función aleatoria, es una familia de variables aleatorias $X(t)$, donde t está en un conjunto de índices $\mathcal{I}$ y los valores que toma cada $X(t)$ pertenecen a un espacio medible S. Visto de otra manera, X es una variable aleatoria cuyas realizaciones son funciones $\mathcal{I} \to S$.

Para analizar conjuntos de cadenas bajo el paradigma de procesos estocásticos, podemos hacer uso de una correspondencia natural entre las cadenas binarias de longitud ℓ y las funciones $\{0, \ldots, \ell\} \to \mathbb{Z}$, asociando a la cadena $w \in \mathbb{B}^\ell$ con la función:

$$f_w(k) = \sum_{i=1}^{k} (\mathbb{1}_{w_i=1} - \mathbb{1}_{w_i=0})$$

Esta función lleva la cuenta de la diferencia entre la cantidad de 1s y 0s que hay en cada prefijo de la cadena.

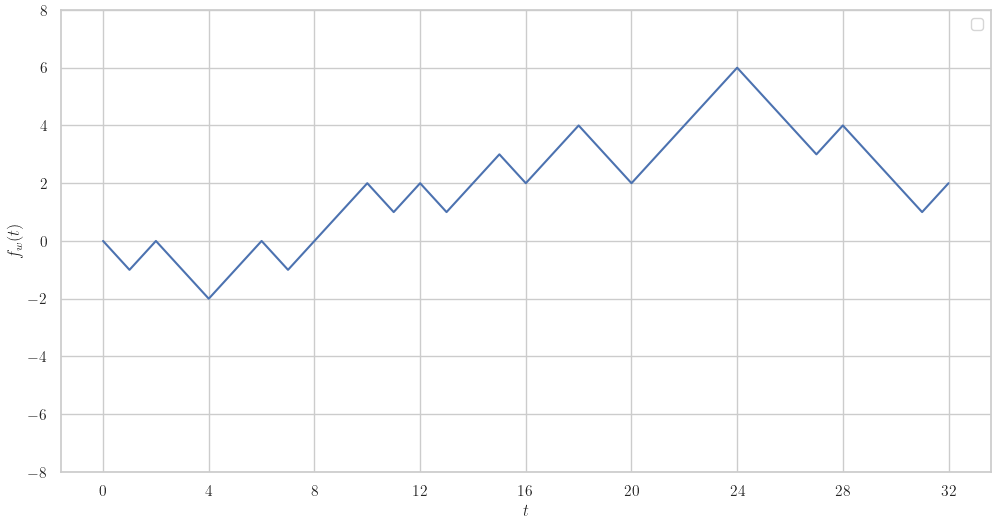

Figura 10: Representación de la función $f_w(k)$ para $w =$ 0100110111010110110011111000100 01

> **Observación 3.** f_w nos permite dar una nueva interpretación a la discrepancia: el valor de $D(w)$ coincide con la diferencia entre el máximo y el mínimo valor de $f_w(k)$ a lo largo de su dominio.
>
> $$D(w) = \left(\max_{i=0,\ldots,\ell} f_w(k)\right) - \left(\min_{i=0,\ldots,\ell} f_w(k)\right)$$

Esta correspondencia entre cadenas binarias y funciones basta para trabajar dentro del marco de procesos estocásticos: cualquier variable aleatoria X sobre $\mathbb{B}^\ell$ puede verse como un proceso sobre funciones $\{0, \ldots, \ell\} \to \mathbb{Z}$. No obstante, como nos interesa comparar el comportamiento asintótico de distintas familias de cadenas, adoptaremos una normalización que ponga a la misma escala colecciones de distintas longitudes.

Definición 3.2 (Proceso W_S)**.** Sea $S \subseteq \mathbb{B}^\ell$ un conjunto de cadenas de longitud ℓ. Tomando la medida de probabilidad uniforme sobre S, llamamos $W_S : S \times [0,1] \to \mathbb{R}$ al proceso estocástico dado por:

$$W_S(w,t) = \frac{1}{\sqrt{\ell}} \sum_{i=1}^{\lfloor t\ell \rfloor} (\mathbb{1}_{w_i=1} - \mathbb{1}_{w_i=0})$$

A diferencia de las funciones f_w, el proceso W_S es de **tiempo continuo**: el dominio de sus realizaciones es $[0,1]$. En la figura 11 se puede ver que, para valores bajos de ℓ, éstas tienen saltos abruptos en los puntos de la forma k/ℓ, donde se agrega un nuevo sumando a la fórmula de 3.2. Sin embargo, a medida que la longitud de las cadenas crece, las funciones correspondientes empiezan a parecer continuas (pero no diferenciables).

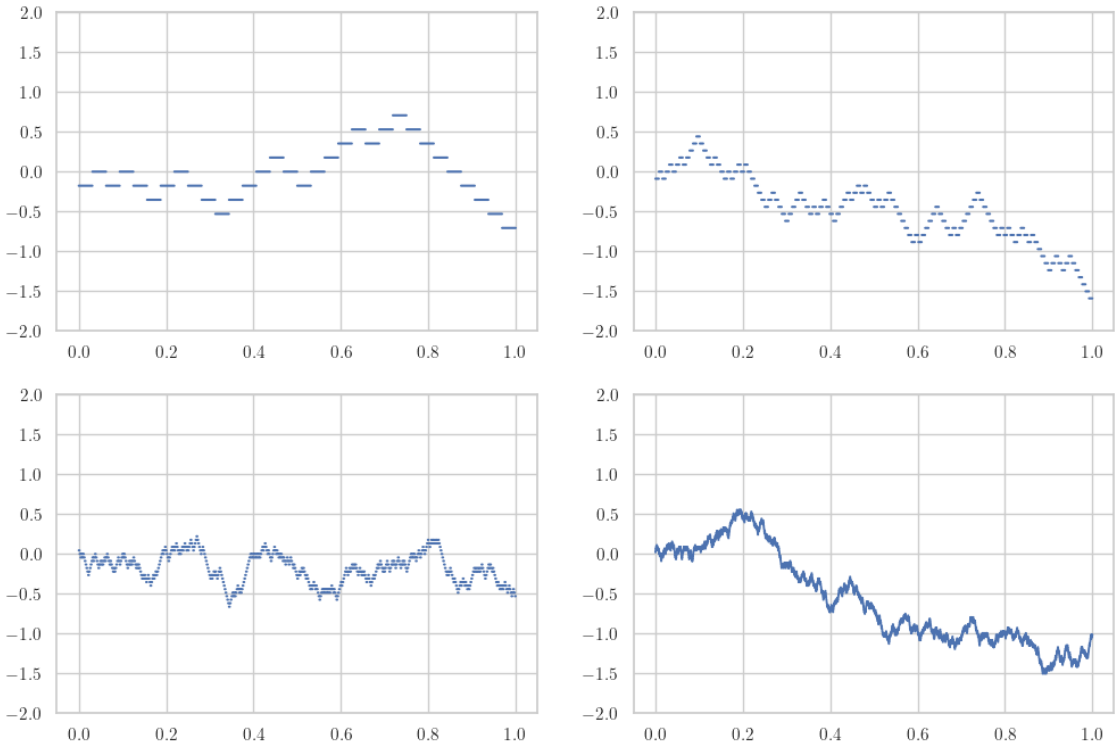

Figura 11: Gráfico de ejemplos de resultados de W_S con cadenas de longitudes 32, 128, 512 y 2048.

3.1. Comparación de Estadísticos

Habiendo establecido el marco teórico que nos permite interpretar las familias de cadenas como procesos estocásticos, procederemos a examinar si las similitudes observadas en la discrepancia entre $\mathcal{DB}_n$ y Bal_{2^n} se extienden a otras propiedades estadísticas. Para ello, realizaremos un análisis empírico sistemático de los procesos $W_{\mathcal{DB}_n}$ y $W_{Bal_{2^n}}$, evaluando diversos estadísticos que caracterizan el comportamiento de ambos procesos.

3.1.1. Media y Varianza por Tiempo

Las características más básicas que se pueden definir sobre un proceso estocástico son la media y la varianza para cada tiempo t:

Definición 3.3 (Media y Varianza en tiempo t)**.** Dado un proceso estocástico $W = \{W(t) : t \in \mathcal{I}\}$ y un tiempo $t \in \mathcal{I}$:

- La **media en el tiempo** t se define como $\mathbb{E}[W(t)]$.
- La **varianza en el tiempo** t se define como $\mathrm{Var}[W(t)]$.

Dado que podemos muestrear uniformemente de tanto $\mathcal{DB}_n$ como Bal_{2^n} sin mucha dificultad, estimar estos valores para $W_{\mathcal{DB}_n}$ y $W_{Bal_{2^n}}$ resulta fácil: basta con tomar, para cada tiempo $t \in \{1/2^{14}, ..., 2^{14}/2^{14}\}$, los estimadores típicos de la media y varianza muestrales. Por lo tanto, si tomamos una muestra M de cadenas de $\mathcal{DB}_n$, computamos los siguientes estadísticos para $W_{\mathcal{DB}_n}$:

$$\hat{\mu}_t = \frac{1}{|M|} \sum_{w \in M} W_{\mathcal{DB}_n}(w, t)$$
$$\hat{\sigma}_t^2 = \frac{1}{|M| - 1} \sum_{w \in M} \left(W_{\mathcal{DB}_n}(w, t) - \hat{\mu}_t\right)^2$$

Tomamos muestras de los conjuntos $\mathcal{DB}_n$ y Bal_{2^n} de distintos órdenes n, observando el mismo comportamiento que en casos anteriores: a medida que el orden crece, las muestras tomadas se asemejan cada vez más. Para este experimento, y todos los que le siguen en esta sección, reportamos solamente los resultados del orden más alto analizado, donde los valores son prácticamente indistinguibles. En este caso, es una muestra de 1,000,000 de cadenas de $\mathcal{DB}_{14}$ y $Bal_{2^{14}}$, obteniendo los resultados que se pueden visualizar en la figura 12.

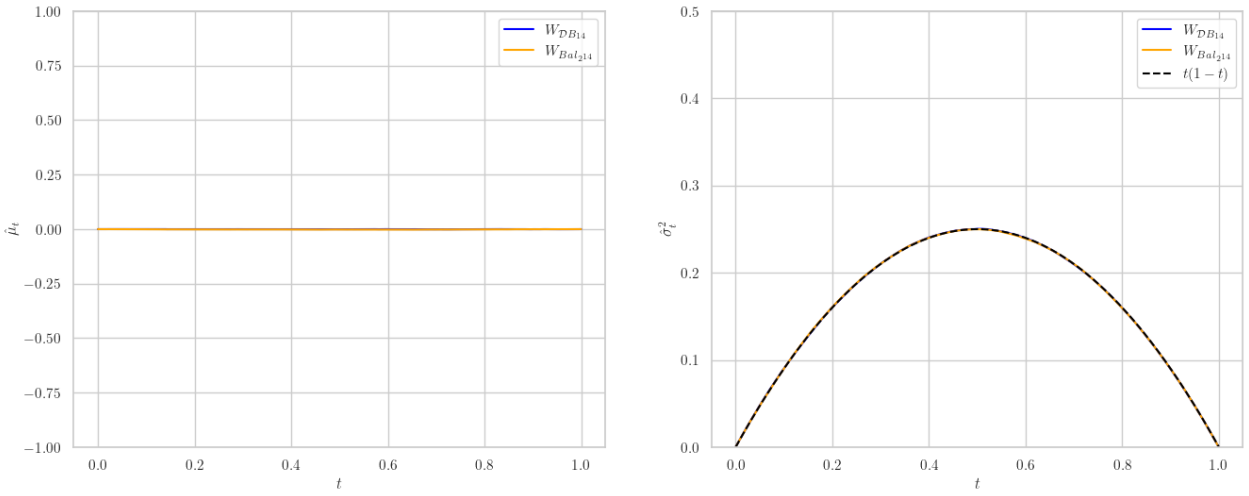

Figura 12: Izquierda: comparación de $\hat{\mu}_t$ para muestras de $\mathcal{DB}_{14}$ y $Bal_{2^{14}}$. Derecha: comparación de $\hat{\sigma}_t^2$ para muestras de $\mathcal{DB}_{14}$ y $Bal_{2^{14}}$, junto con la función $f(t) = t(1-t)$ de referencia.

Claramente, las procesos inducidos por ambos conjuntos coinciden en tanto media como varianza para cada tiempo t.

- En el caso de la media, se tiene $\mathbb{E}[W_{\mathcal{DB}_{14}}(t)] = \mathbb{E}[W_{Bal_{2^{14}}}(t)] = 0$ para todo t. Esto es fácil de demostrar: como ambos conjuntos están **cerrados por complemento**, en cada posición i debe haber la misma cantidad de cadenas con $w_i = 0$ que con $w_i = 1$.

- En el caso de la varianza, los resultados del experimento sugieren que $\text{Var}[W_{\mathcal{DB}_{14}}(t)] = \text{Var}[W_{Bal_{2^{14}}}(t)] = t(1-t)$. Mientras que, como veremos más adelante, el caso de $W_{Bal_{2^{14}}}$ es un teorema, para el caso de $W_{\mathcal{DB}_{14}}$ es una conjetura.

3.1.2. Distribuciones Marginales

Como ya establecimos, los procesos estocásticos pueden verse como familias indexadas de variables aleatorias.

> **Definición 3.4** (Distribución Marginal de Proceso)**.** Dado un proceso $W : \Omega \times \mathcal{I} \to \mathcal{S}$ y un tiempo $t \in \mathcal{I}$, denominamos **distribución marginal en tiempo** t a la variable aleatoria $X(\omega) = W(\omega, t)$

El muestreo de distribuciones marginales en tiempo t para el proceso $W_{\mathcal{DB}_n}$ consiste en tomar una muestra M del conjunto $\mathcal{DB}_n$ y computar los valores de $W_{\mathcal{DB}_n}(w, t)$ para cada $w \in M$. El caso de $W_{Bal_{2^n}}$ es análogo.

De forma análoga al experimento anterior, tomamos una muestra de tamaño 1,000,000 para las distribuciones marginales de $W_{\mathcal{DB}_{14}}$ y $W_{Bal_{2^{14}}}$ en los tiempos $t \in \{0{,}25, 0{,}50, 0{,}75\}$.

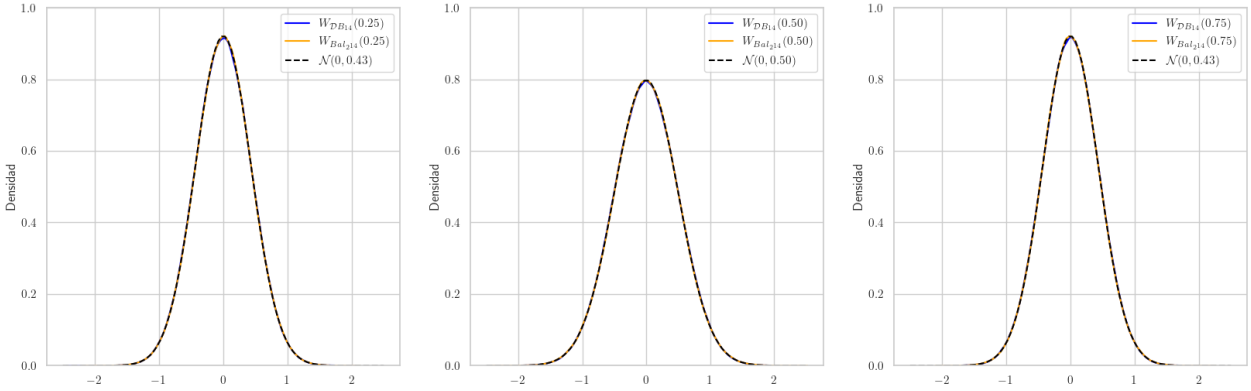

Figura 13: Comparaciones de las estimaciones KDE a partir de muestras de $W_{\mathcal{DB}_n}$ y $W_{Bal_{2^n}}$ en los tiempos $t \in \{0{,}25, 0{,}50, 0{,}75\}$, junto con las funciones de densidad para variables normales $\mathcal{N}(0, t(1-t))$.

En la figura 13 se puede ver que, para cada tiempo t, tanto $W_{\mathcal{DB}_{14}}(t)$ como $W_{Bal_{2^{14}}}(t)$ se ajustan perfectamente a una distribución normal $\mathcal{N}(0, t(1-t))$.

3.1.3. Valor Máximo

Otra medida de interés que se puede estudiar es el valor máximo de a lo largo del proceso.

> **Definición 3.5.** Sea $W : \Omega \times \mathcal{I} \to \mathcal{S}$ un proceso estocástico. El **valor máximo** $\mathcal{M}$ del proceso W es la variable aleatoria dada por $\mathcal{M}(\omega) = \text{máx}_{t \in I} W(\omega, t)$

Para los procesos $W_{\mathcal{DB}_n}$ y $W_{Bal_{2^n}}$, el valor máximo se corresponde con la longitud del prefijo con la mayor diferencia entre cantidad de 1s y 0s.

A partir de una muestra M de cadenas de $\mathcal{DB}_n$, obtenemos una muestra del valor máximo de $W_{\mathcal{DB}_n}$ al calcular $\text{máx}_{t \in [0,1]} W_{\mathcal{DB}_n}(w, t)$ para cada cadena $w \in M$. Al igual que en los experimentos anteriores, trabajamos con orden $n = 14$ y muestras de tamaño 1,000,000. Los resultados se pueden ver a continuación:

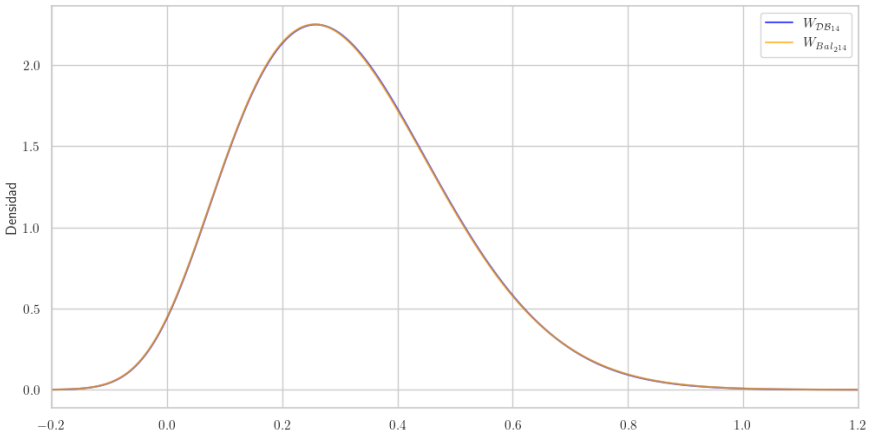

Figura 14: Comparación entre las distribuciones de máximo valor para $W_{\mathcal{DB}_{14}}$ y $W_{Bal_{2^{14}}}$.

Nuevamente, el gráfico de la figura 14 demuestra que los procesos coinciden en esta propiedad.

Cabe destacar que se puede considerar también el valor mínimo de los procesos, pero en este punto no tiene sentido comparar esas distribuciones porque sabemos que también deben coincidir. Esto se debe a que, como ambos conjuntos de cadenas están cerrados por complemento, para cada cadena w que alcanza una diferencia máxima M de 1s y 0s, la cadena w^c alcanza una diferencia mínima $-M$. Esto implica que la función de densidad del valor mínimo de ambos procesos debe ser exactamente la negación de la función de densidad del valor máximo.

3.1.4. Valor Promedio

Otra propiedad relevante a considerar es el valor promedio del proceso a lo largo del tiempo.

> **Definición 3.6** (Valor Promedio de Proceso)**.** Sea $W : \Omega \times \mathcal{I} \to \mathbb{R}$ un proceso estocástico con $\mathcal{I} \subseteq \mathbb{R}$ medible en $\mathbb{R}$. El **valor promedio** $\mathcal{A}$ del proceso W es la variable aleatoria definida como $A(\omega) = \frac{1}{|\mathcal{I}|} \int_{t \in \mathcal{I}} W(\omega, t)\, dt$

En nuestro contexto, donde trabajamos con cadenas w de longitud 2^n, computamos en cambio la suma discreta de los valores del proceso:

$$\hat{A}(w) = \sum_{i=0}^{2^n} W_{\mathcal{DB}_n}(w, i/2^n)$$

Esta técnica es un método básico para aproximar integrales, conocido como **sumas de Riemann**. Sin embargo, como las funciones que consideramos son constantes entre los puntos que tomamos para la suma, en este caso la aproximación **es exacta**.

En este experimento tomamos las mismas muestras que antes: 1,000,000 de cadenas de orden 14 para cada uno de los conjuntos.

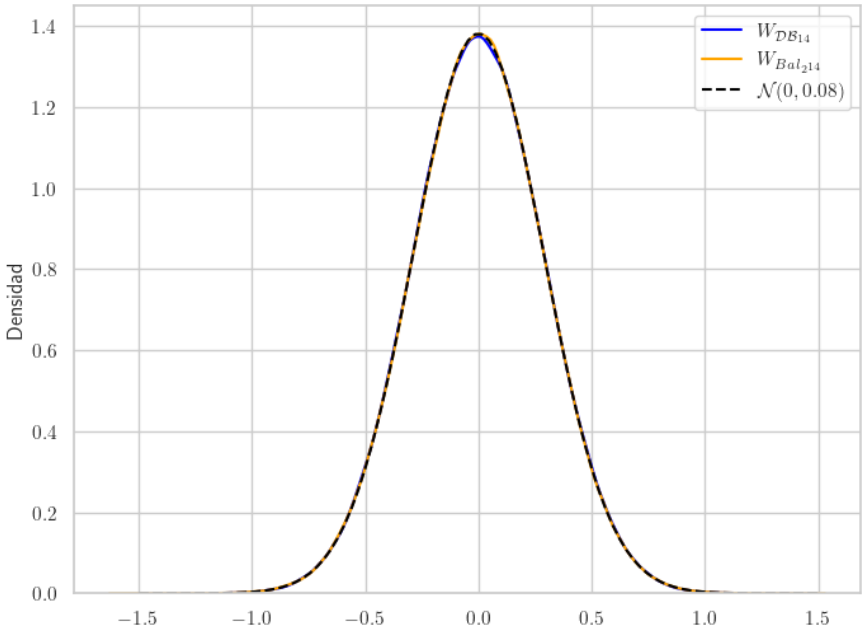

Figura 15: Comparación entre las distribuciones de valor promedio para $W_{\mathcal{DB}_{14}}$ y $W_{Bal_{2^{14}}}$, junto con la densidad de la distribución normal $\mathcal{N}(0, 1/12)$.

Como en las propiedades anteriores, el experimento confirma que el valor promedio de ambos procesos sigue la misma distribución. Además, se puede ver que ésta se ajusta perfectamente a la distribución gaussiana con media 0 y varianza 1/12.

3.2. Convergencia al Puente Browniano

Para explicar las múltiples coincidencias que observamos en los experimentos realizados, decidimos formular una hipótesis más fuerte que "coinciden en estos aspectos particulares". En cambio, proponemos que los procesos $W_{\mathcal{DB}_n}$ y $W_{Bal_{2^n}}$ **convergen a la misma distribución** a medida que crece el orden n, en un sentido preciso que detallamos a continuación.

3.2.1. Teorema de Donsker y Puente Browniano

El resultado central en el análisis asintótico de los procesos estocásticos es el **Teorema de Donsker**, también conocido como el **Teorema Central del Límite Funcional**.

> **Teorema 2** (Teorema de Donsker)**.** *Sean $X_1, X_2, \ldots$ una secuencia de variables aleatorias i.i.d. con media 0 y varianza 1, y sean $S_n = \sum_{i=1}^{n} X_i$ las variables que representan sus sumas parciales. Tomando el siguiente proceso sobre $[0, 1]$:*
>
> $$W^{(n)}(t) = \frac{S_{\lfloor nt \rfloor}}{\sqrt{n}}$$
>
> *La función aleatoria $W^{(n)}$ **converge en distribución** al proceso de Wiener W a medida que $n \to \infty$.*

El **proceso de Wiener**, o movimiento browniano estándar, es un proceso estocástico caracterizado por las siguientes propiedades:

1. $P(W(0) = 0) = 1$.
2. Tiene **incrementos independientes**: para cada tiempo $t > 0$, los incrementos $W(t + u) - W(t)$ con $u \geq 0$ son independientes de los valores previos $W(s)$ con $s < t$.
3. Tiene **incrementos gaussianos**: $W(t + u) - W(t) \sim \mathcal{N}(0, u)$.
4. Para cada t, $W(t)$ es casi seguramente continua en t.

Volviendo a nuestra definición de procesos inducidos por conjuntos de cadenas, el Teorema de Donsker tiene el siguiente corolario:

> **Corolario 1** ($W_{\mathbb{B}^\ell}$ converge W)**.** *El proceso $W_{\mathbb{B}^\ell}$ converge en distribución a W a medida que $\ell \to \infty$.*

Por otro lado, podemos considerar al llamado **puente browniano** B, el proceso resultante de condicionar a W a que termine en 0:

$$P(B(t) = k) = P(W(t) = k \mid W(1) = 0)$$

Existe otro corolario, análogo al anterior, para el proceso inducido por las cadenas balanceadas:

> **Corolario 2** (W_{Bal_ℓ} converge a B)**.** *El proceso W_{Bal_ℓ} converge en distribución a B a medida que $\ell \to \infty$.*

Puesto que será relevante para el análisis de la conjetura que planteamos a continuación, vale la pena mencionar algunas propiedades relevantes del proceso B. Al igual que W, se puede caracterizar completamente por medio de una serie de propiedades:

1. $P(B(0) = 0) = 1$ y $P(B(1) = 0) = 1$.
2. B es un **proceso gaussiano**, es decir, cualquier colección finita de la forma $\{B(t_1), B(t_2), \ldots, B(t_k)\}$ tiene distribución normal multivariada.
3. $\mathbb{E}(B(t)) = 0$ para $t \in [0, 1]$
4. $\text{Cov}(B(t), B(s)) = \text{mín}\{s, t\} - st$ para $s, t \in [0, 1]$

Una propiedad importante de los procesos gaussianos es que quedan completamente determinados por su momento de segundo orden. Si se asume un proceso gaussiano de media cero, la función de covarianza define completamente el comportamiento del proceso.

3.2.2. Conjetura Principal

Las coincidencias observadas en la sección anterior, junto con el teorema de convergencia conocido para W_{Bal_ℓ}, nos llevan a plantear la siguiente conjetura:

> **Conjetura 1** (Las De Bruijn son Brownianas)**.** *El proceso $W_{\mathcal{DB}_n}$ converge en distribución al puente browniano B a medida que $n \to \infty$.*

Esta conjetura resulta sorprendente. A pesar de la estructura local restrictiva impuesta por la propiedad de de Bruijn (cada bloque de longitud n debe aparecer exactamente una vez), el proceso inducido exhibe comportamiento browniano caracterizado únicamente por la restricción global de tener igual cantidad de 0s y 1s.

3.3. Evidencia para la Conjetura Principal

A continuación, presentaremos experimentos que proveen evidencia para la validez de la conjetura 1. Éstos se basan en el hecho de que B es un **proceso gaussiano** y, por ende, una condición necesaria para que $W_{\mathcal{DB}_n}$ converja en distribución a B es que el vector aleatorio $[W_{\mathcal{DB}_n}(t_1), \ldots, W_{\mathcal{DB}_n}(t_m)]$ converja a la multinormal correspondiente para cualquier selección de tiempos $t_1, \cdots, t_m$. Nos dispusimos a verificar esto por medio de 2 nuevos experimentos: en el primero, comprobamos que las **covarianzas** entre los distintos puntos fueran las correctas, mientras que en el segundo verificamos que las **combinaciones lineales** de los valores $W_{\mathcal{DB}_n}(t_i)$ siguieran las distribuciones normales esperadas.

3.3.1. Covarianzas

En este experimento, muestreamos 1,000,000 realizaciones del proceso $W_{\mathcal{DB}_{14}}$, y computamos la matriz de covarianzas para las posiciones $i \in \{1/8, \ldots, 7/8\}$. Luego, bajo la conjetura 1, el valor de las covarianzas en los puntos $i < j$ debería aproximarse a $\mathrm{Cov}(B(i), B(j)) = i(1-j)$. En la siguiente figura podemos visualizar la matriz de covarianza computada, comparada a la matriz correspondiente de B.

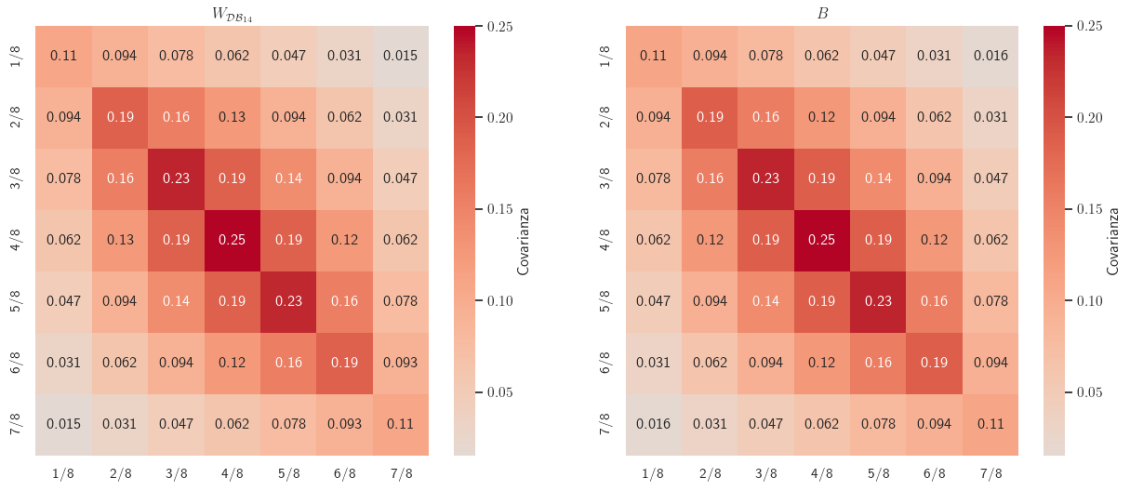

Figura 16: Visualizaciones de matrices de covarianza para los procesos $W_{\mathcal{DB}_{14}}$ y B en las posiciones $t \in \{1/8, \ldots, 7/8\}$.

Podemos ver que los valores de cada entrada de las matrices son muy similares: la máxima diferencia entre cada par es del orden de 10^{-4}. Si realizamos esta misma comparación para pares de procesos de órdenes menores, se puede ver la caída en dicha diferencia máxima, lo cual sugiere un comportamiento de convergencia.

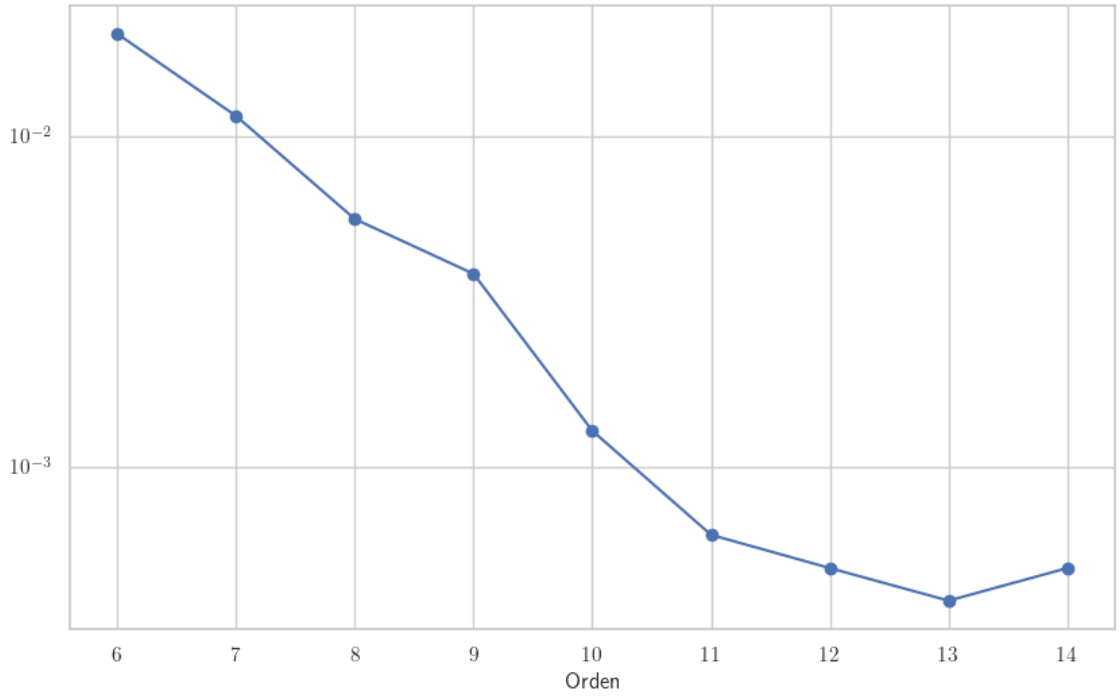

Figura 17: Evolución de la máxima diferencia entre los valores de covarianza de B y la covarianza muestral observada en $W_{\mathcal{DB}_n}$ con $n \in \{6, \ldots, 14\}$.

3.3.2. Combinaciones Lineales

Este experimento fue realizado sobre las mismas muestras que el anterior: 4,000,000 realizaciones del proceso $W_{\mathcal{DB}_{14}}$. Sobre este conjunto, computamos 100 combinaciones lineales aleatorias sobre las posiciones $i \in \{1/8, \ldots, 7/8\}$, cada una con pesos muestreados de $\mathcal{N}(0, 1)$.

Como anticipamos, nuestro objetivo para este experimento fue verificar que, para cada combinación lineal con pesos $\mathbf{a} = [a_1 \cdots a_7]$, el valor de $\mathbf{a}^T\mathbf{X}$ sigue la distribución normal esperada, donde $\mathbf{X} = [X_1 \cdots X_7]$ con $X_i = W_{\mathcal{DB}_{14}}(i)$. Para lograr esto, primero debimos determinar cuál es esa distribución, o más específicamente, qué media y varianza debería tener.

- Por un lado, como $\mathbb{E}[X_i] = 0$ para cada i, la media de $\mathbf{a}^T\mathbf{X} = \sum_{i=1}^{7} a_i X_i$ debe ser $\sum a_i \mathbb{E}[X_i] = 0$.
- La varianza $\mathrm{Var}(\mathbf{a}^T\mathbf{X})$ está definida como:

  $$\mathrm{Var}(\mathbf{a}^T\mathbf{X}) = \mathbb{E}[(\mathbf{a}^T\mathbf{X} - E[\mathbf{a}^T\mathbf{X}])^2]$$

  Como el valor esperado es 0, esto se reduce a:

  $$\mathbb{E}[(\mathbf{a}^T\mathbf{X} - E[\mathbf{a}^T\mathbf{X}])^2] = \mathbb{E}[(\mathbf{a}^T\mathbf{X})^2]$$

  Ahora, ya que $\mathbf{a}^T\mathbf{X}$ es un escalar, es igual a su tranposición $\mathbf{X}^T\mathbf{a}$:

  $$\mathbb{E}[(\mathbf{a}^T\mathbf{X})^2] = \mathbb{E}[(\mathbf{a}^T\mathbf{X})(\mathbf{a}^T\mathbf{X})] = \mathbb{E}[(\mathbf{a}^T\mathbf{X})(\mathbf{X}^T\mathbf{a})] = \mathbb{E}[\mathbf{a}^T\mathbf{X}\mathbf{X}^T\mathbf{a}]$$

  Como el vector de coeficientes $\mathbf{a}$ es constante, se tiene:

  $$\mathrm{Var}(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\mathbb{E}[\mathbf{X}\mathbf{X}^T]\mathbf{a}$$

Por otro lado, la matriz de covarianza de $\mathbf{X}$, que llamaremos $\mathbf{\Sigma}$, está dada por $\mathbb{E}[(\mathbf{X}-\mathbb{E}[\mathbf{X}])(\mathbf{X}-\mathbb{E}[\mathbf{X}])^T]$. Nuevamente, como $\mathbb{E}[\mathbf{X}] = 0$, esto se simplifica a $\mathbb{E}[\mathbf{X}\mathbf{X}^T]$. Por ende, concluimos que la varianza de la combinacion lineal debe ser:

$$\mathrm{Var}(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\mathbf{\Sigma}\mathbf{a}$$

Si nuestra conjetura fuera cierta, entonces las covarianzas de las posiciones deberían coincidir con las del puente browniano. Por ende, la varianza que esperaríamos para la combinación lineal con pesos $\mathbf{a}$ es $\mathbf{a}^T\mathbf{\Sigma}_B\mathbf{a}$, siendo $\mathbf{\Sigma}_B$ la matriz de covarianza correspondiente en el proceso B.

Habiendo determinado los parámetros esperados para cada combinación, empezamos a comparar las distribuciones obtenidas con las esperadas.

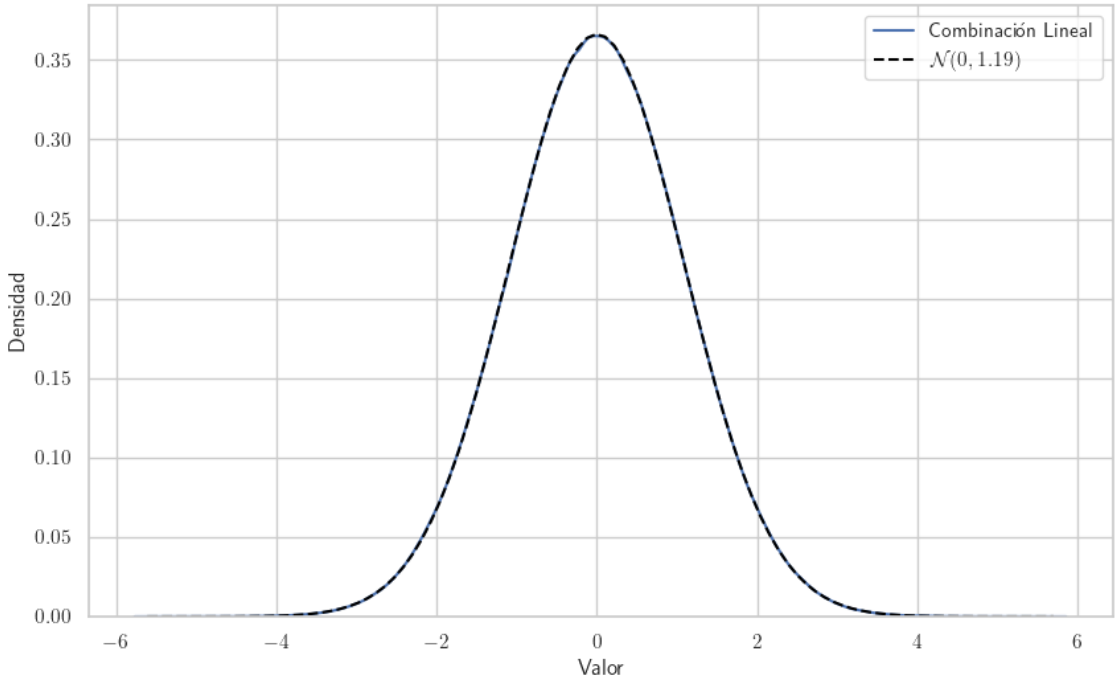

Figura 18: KDE de la distribución de una combinación lineal particular de $W_{\mathcal{DB}_{14}}(1/8), \ldots, W_{\mathcal{DB}_{14}}(7/8)$, comparada con la normal correspondiente.

Como se puede observar en el ejemplo de la figura 18, las distribuciones de las observaciones son indistinguibles de las que esperaríamos si nuestra conjetura fuera cierta. Para tomar una medida más robusta que la comparación visual, podemos comparar la distribución de cada combinación lineal con la distribución normal esperada por medio del test de Kolmogorov-Smirnov [15]. Éste consiste en tomar el siguiente estadístico:

$$D_M = \sup_x |\hat{F}_M(x) - F(x)|$$

Donde $\hat{F}_M(x)$ es la función de distribución empírica para la muestra M, mientras que $F(x)$ es la función de distribución acumulada para la distribución esperada. Por el teorema de Glivenko-Cantelli [17], D_M converge a 0 con probabilidad 1 a medida que $|M| \to \infty$ cuando las distribuciones coinciden.

Este experimento de comparación entre distribuciones de combinaciones lineales fue repetido para los procesos $W_{\mathcal{DB}_n}$ con órdenes entre 6 y 14. En la figura 19 se puede observar la tendencia descendiente del máximo valor

observado para el estadístico D_M.

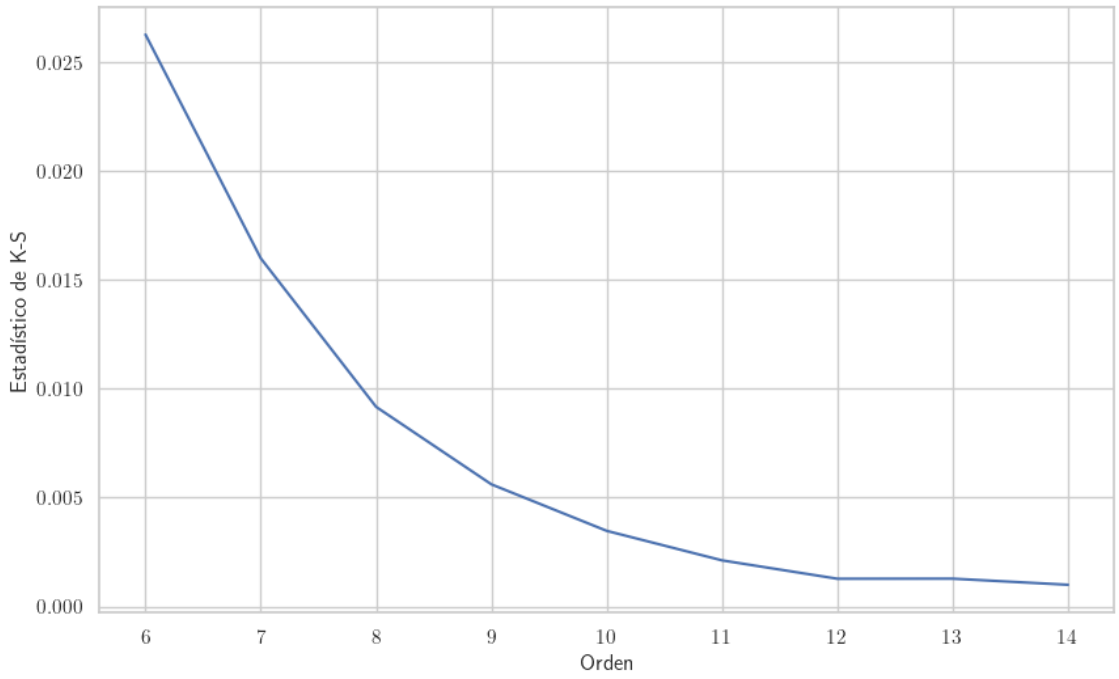

Figura 19: Valor máximo del estadístico del test K-S entre las combinaciones lineales muestreadas para el proceso $W_{\mathcal{DB}_n}$ con n entre 6 y 14.

3.4. No todo es Browniano

Abundan los conjuntos de cadenas donde el muestreo uniforme no induce un proceso Browniano. En particular, experimentamos con los collares aritméticos [2],

Los collares aritméticos de orden n, $\mathcal{C}_n$, son cadenas que se construyen concatenando una tras otra las 2^n cadenas de longitud n que surgen de los términos sucesivos de una progresión aritmética de diferencia impar.

El j-ésimo bloque es la representación en base 2 del número $dj \mod 2^n$, para un d impar. Hay exactamente 2^{n-1} collares aritmétcos. En [2, Teorema 1] se demuestra que los collares aritméticos de orden n cumplen que cada cadena de longitud n ocurre n veces, pero estas n ocurrencias comienzan en posiciones con distinto residuo módulo n, para cualquier convensión de la posición incial.

Computamos las varianzas de $W_{\mathcal{C}_n}(t)$ para $t \in [0,1]$ para $n \in \{5,\ldots,9\}$, y observamos que difiere signficativamente de $t(1-t)$ como es el caso en el puente Browniano.

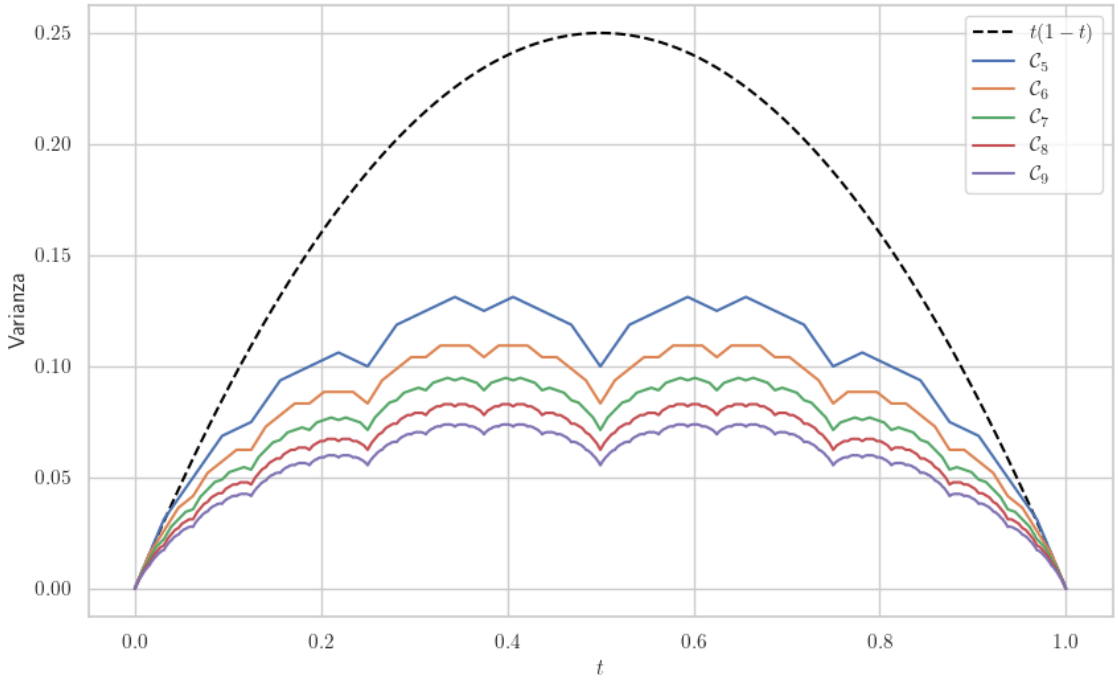

Figura 20: Varianzas de $W_{\mathcal{C}_n}(t)$ para $t \in [0,1]$ y $n \in \{5,\ldots,9\}$, comparadas con la función $t(1-t)$.

4. Otras conjeturas

En nuestra búsqueda de una prueba de la conjetura 1, exploramos distintas variantes y generalizaciones de la misma con la esperanza de que resultaran más fáciles de abordar. En esta sección expondremos algunas de ellas, exhibiendo evidencia que apoya la validez de cada una.

Para lograr interpretar a las siguientes conjeturas como generalizaciones, debemos recordar la estructura de las cadenas De Bruijn. En particular, vamos a centrarnos en su caracterización como ciclos eulerianos del grafo De Bruijn.

4.1. Iteración de Grafo de Línea

Existe una generalización simple de la conjetura original que, a pesar de ser falsa, vale la pena considerar: como el proceso $W_{\mathcal{DB}_n}$ se obtiene a partir de las cadenas de $\mathcal{DB}_n$, que son los ciclos eulerianos de una familia de grafos 2-regulares, uno podría creer que el comportamiento en el límite es el mismo para cualquier otra familia "similar" $\mathcal{F}_n$. Una noción de similitud razonable podría incluir:

- Que cada $G \in \mathcal{F}_n$ sea **euleriano**, y por ende **conexo**.
- Que las aristas de cada $G \in \mathcal{F}_n$ tengan etiquetas $\{0, 1\}$, y que cada nodo tenga una arista saliente etiquetada con 1 y una etiquetada con 0.

Sin embargo, no es difícil construir contraejemplos que refutan esto. Un ejemplo sería la familia de grafos $\mathcal{C} = \{C_n\}_{n\in\mathbb{N}}$ donde cada grafo está conformado por los vértices $v_1, \ldots, v_n$ y cada vértice tiene 2 aristas (una con cada etiqueta) que lo conectan con el siguiente.

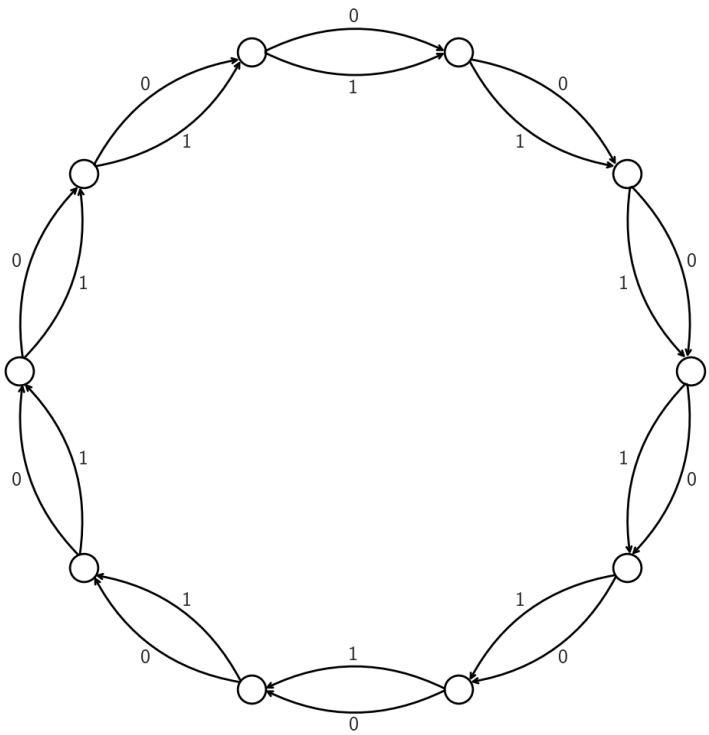

Figura 21: Representación gráfica del grafo C_{10}.

Cualquier ciclo euleriano en este grafo consiste de dos "vueltas" por el grafo, donde la segunda vuelta debe pasar por exactamente las aristas que no se tomaron en la primera. Por ende, las cadenas que se producen en estos ciclos son todas de la forma $w = a\bar{a}$, y esto implica que los procesos no pueden converger a B. Es fácil verlo al estudiar la varianza en cada posición de las caminatas aleatorias inducidas por dichas cadenas:

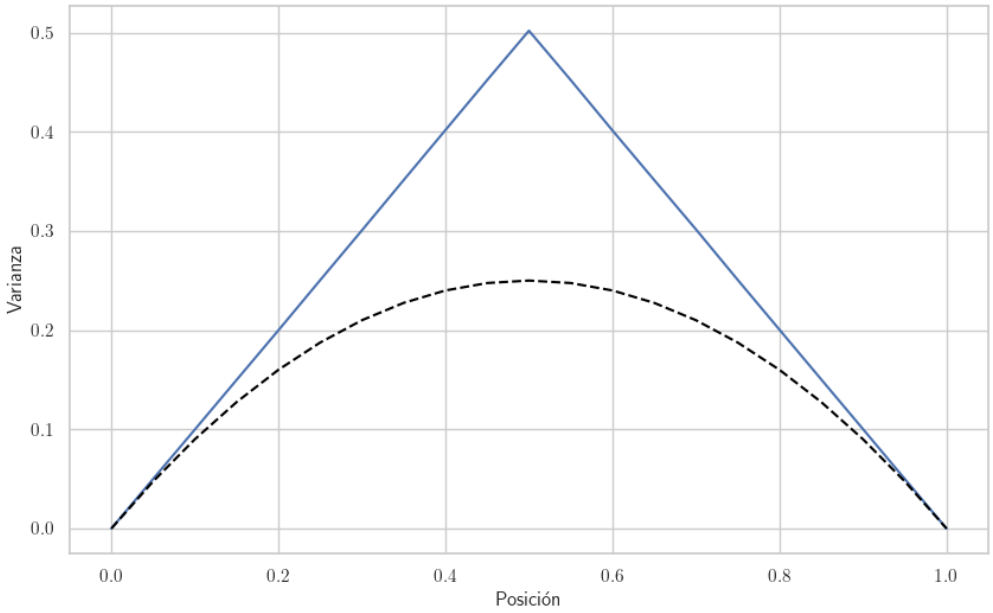

Figura 22: Varianza en cada posición del proceso inducido por tomar cadenas eulerianas de C_{10} (Análogo al gráfico de la varianza en la figura 12). La línea punteada indica la varianza correspondiente para el puente browniano.

No obstante, la 2-regularidad y las etiquetas no son lo único que distingue a la familia de grafos De Bruijn, si no que son producto de una construcción particular: el grafo de orden $n+1$ es el grafo de línea del de orden n. Luego, podemos tomar otra variante de nuestra conjetura:

Conjetura 2 (Familias de Grafos producidas por Iteración de Grafo Línea son Brownianas)**.** *Sea* $\mathcal{F} = \{G_n\}_{n\in\mathbb{N}}$ *una familia de grafos donde:*

- G_1 *es euleriano, y cada uno de sus vértices tiene 2 aristas salientes, una con cada etiqueta en* $\{0, 1\}$*.*
- $G_{n+1} = \mathcal{L}(G_n)$

donde S_n *el conjunto de cadenas que son las etiquetas de los ciclos eulerianos de* G_n*.*

Entonces, el proceso W_{S_n} *converge en distribución al puente browniano* B *cuando* $n \to \infty$*.*

Tomemos al grafo de doble ciclo C_{10} como ejemplo. En el siguiente gráfico se puede ver el comportamiento de la varianza en cada posición de las caminatas aleatorias correspondientes a $\mathcal{L}^{(n)}(C_{10})$ para $n = 0, \ldots, 5$:

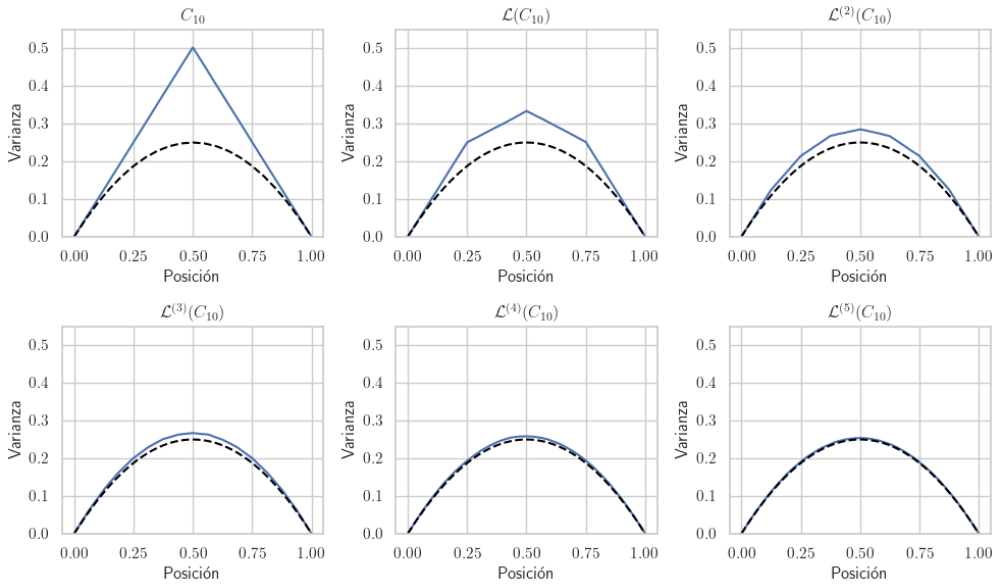

Figura 23: Evolución de la varianza por posición en los procesos inducidos por los ciclos eulerianos de $\mathcal{L}^n(C_{10})$ para $n = 0, \ldots, 5$.

Para comprobar la validez de esta conjetura, muestreamos grafos 2-regulares uniformemente e iteramos $\mathcal{L}$ sobre cada uno 5 veces. Luego, aplicamos el test Kolmogorov-Smirnov sobre las combinaciones lineales de valores en ciertas posiciones que explicamos en la sección anterior.

El muestreo uniforme de grafos 2-regulares se realizó por medio del modelo de configuración [14], que produce grafos con una secuencia de grados dada de forma equiprobable). Los grafos generados fueron posteriormente filtrados para descartar disconexos, y se asignaron las etiquetas 0 y 1 a las aristas aleatoriamente (respetando la restricción establecida en la conjetura). En total, se muestrearon 32 grafos base de 16 nodos, y de cada uno se tomaron 100,000 cadenas para cada iteración de $\mathcal{L}$.

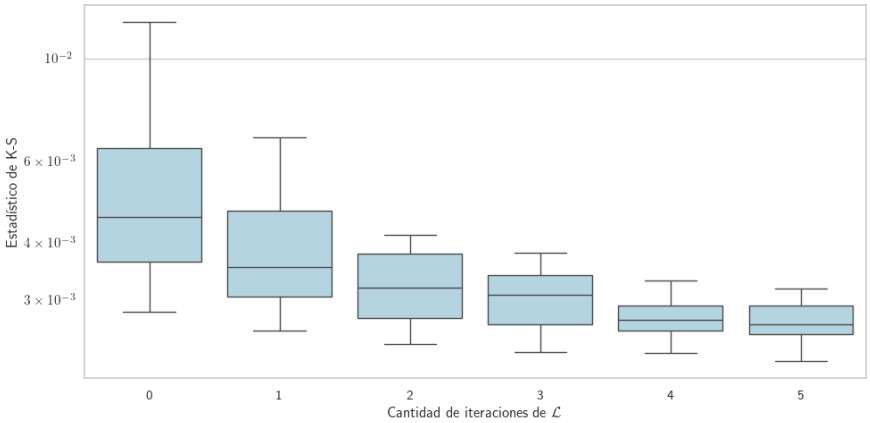

Figura 24: Evolución del estadístico K-S promedio a lo largo de distintos grafos muestreados, a medida que se itera la operación de grafo de línea $\mathcal{L}$.

Los resultados visualizados en 24 apoyan nuestra conjetura: el estadístico tiende a bajar a medida que se toman iteraciones del grafo de línea. También podemos analizar la trayectoria del estadístico para grafos

particulares:

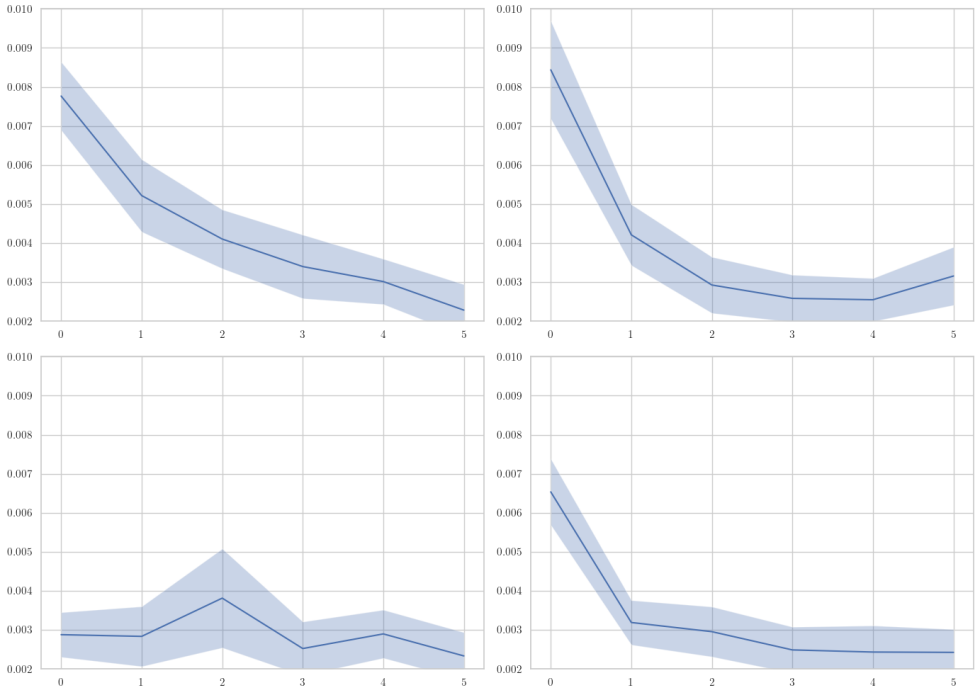

Figura 25: Trayectorias del estadístico del test de Kolmogorov-Smirnov al iterar $\mathcal{L}$ sobre 4 grafos iniciales seleccionados.

En 25 se puede observar que la tendencia a la baja observada en el gráfico anterior es algo inconsistente: en uno de los ejemplos, la aplicación de $\mathcal{L}$ causa una subida inicial del estadístico K-S, aunque el valor promedio finalmente alcanaza un punto menor al inicial después de 5 iteraciones.

4.2. Grafos 2-regulares

Como mencionamos en la sección anterior, sabemos que existen grafos 2-regulares arbitrariamente grandes cuyos ciclos eulerianos no tienen comportamiento browniano. No obstante, esto no nos permite descartar que **la mayoría** de estos grafos exhiban este comportamiento.

Para comprobar esta hipótesis, realizamos un experimento similar al pasado: muestreamos grafos 2-regulares de distintos tamaños, y estudiamos los procesos inducidos por sus ciclos eulerianos. A diferencia del caso anterior, para cada tamaño siempre muestreamos grafos nuevos, independientes a los del tamaño anterior.

Concretamente, muestreamos 32 grafos para tamaños 2^i con $i \in \{6, \dots, 10\}$. Sobre éstos realizamos exactamente el mismo procedimiento que para la otra conjetura: tomamos un conjunto S de 100,000 de cadenas a partir de los ciclos eulerianos de cada uno, y computamos el test K-S comparando la distribución de 100 combinaciones lineales distintas de $W_S(1/8), W_S(2/8), \dots, W_S(7/8)$ y la normal correspondiente.

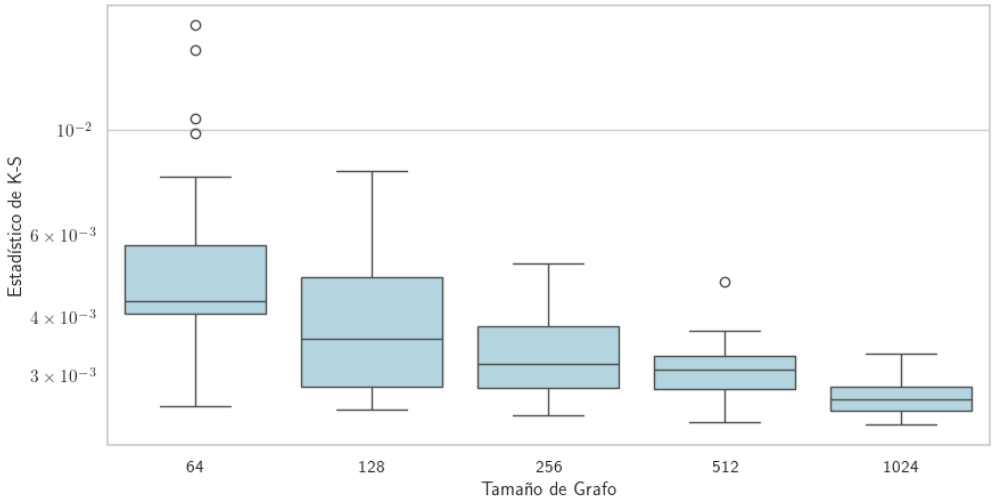

Figura 26: Evolución del estadístico K-S promedio a lo largo de conjuntos de grafos muestreados en distintos órdenes.

A partir de los resultados en 26, observamos el mismo patrón que en experimentos anteriores:

> **Conjetura 3** (Casi todos los grafos 2-regulares dan puente Browniano)**.** *A medida que* $n \to \infty$*, casi todos los grafos* 2*-regulares de* n *vértices inducen procesos (muestreo aleatorio de ciclos eulerianos) que se aproximan cada vez más al puente Browniano.*

Referencias

[1] Martin Aigner. *A Course in Enumeration*. Springer Berlin Heidelberg, 2007.

[2] Nicolás Álvarez, Verónica Becher, Pablo Ferrari, and Sergio Yuhjtman. Perfect necklaces. *Advances of Applied Mathematics*, 80:48–61, 2016.

[3] Nicolás Álvarez, Verónica Becher, Martín Mereb, Ivo Pajor, and Carlos Miguel Soto. De Bruijn sequences with minimum discrepancy. *Mathematics of Computation*, page in press, 2026. arXiv:2407.17367.

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

[5] Joshua Cooper and Christine Heitsch. The discrepancy of the lex-least de Bruijn sequence. *Discrete Mathematics*, 310(6–7):1152–1159, April 2010.

[6] Lucia Costantini. Algorithms for sampling spanning trees uniformly at random. Master's thesis, Universitat Politècnica de Catalunya, 2020.

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

[8] Manuel Eberl. Fisher–Yates shuffle. *Archive of Formal Proofs*, September 2016.

[9] Tuvi Etzion. *Sequences and the de Bruijn graph—properties, constructions, and applications*. Academic Press, London, 2024.

[10] Daniel Gabric and Joe Sawada. Investigating the discrepancy property of de Bruijn sequences. *Discrete Mathematics*, 345(4):112780, 2022.

[11] D. Knuth. *The Art of Computer Programming. Volumen 4, Combinatorial algorithms*. Addison-Wesley, 1973.

[12] N. M. Korobov. Normal periodic systems and their applications to the estimation of sums of fractional parts. *Izv. Akad. Nauk SSSR Ser. Mat.*, 15(1):17–46, 1951.

[13] N. M. Korobov. On normal periodic systems. *Izv. Akad. Nauk SSSR Ser. Mat.*, 16(3):211–216, 1952.

[14] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64:026118, Jul 2001.

[15] N. Smirnov. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin Mathématique de l'Université de Moscou*, 2(2):3–14, 1939.

[16] W.T. Tutte. *Graph Theory*. Encyclopedia of mathematics and its applications. Addison-Wesley Publishing Company, Advanced Book Program, 1984.

[17] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer Science & Business Media, 1996.

[18] C. Villani. *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society, 2003.