UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

VARIANTES ALEATORIAS DE LA SUBSECUENCIA COMÚN MÁS GRANDE

JOSÉ SOTO SAN MARTÍN

2006

UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

VARIANTES ALEATORIAS DE LA SUBSECUENCIA COMÚN MÁS GRANDE

JOSÉ SOTO SAN MARTÍN

COMISIÓN EXAMINADORA		CALIFICACIONES: NOTA (n°)	(Letras)	FIRMA
PROFESOR GUÍA SR. MARCOS KIWI	:	.....	.....	.....
PROFESOR CO-GUÍA SR. MARTÍN MATAMALA	:	.....	.....	.....
PROFESOR INTEGRANTE SR. MARTIN LOEBL	:	.....	.....	.....
NOTA FINAL EXAMEN DE TÍTULO	:	.....	.....	.....

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL MATEMÁTICO

SANTIAGO - CHILE
JULIO - 2006

RESUMEN DE LA MEMORIA
PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL MATEMÁTICO
POR: JOSÉ SOTO SAN MARTÍN
FECHA: 13 DE JULIO DE 2010
PROF. GUÍA: SR. MARCOS KIWI

VARIANTES ALEATORIAS DE LA SUBSECUENCIA COMÚN MÁS GRANDE

El problema de la *subsecuencia común más grande* (LCS por sus siglas en inglés: *longest common subsequence*) es un problema combinatorio que surge naturalmente en distintas aplicaciones prácticas, como búsqueda de patrones en moléculas de ADN, alineamiento de proteínas, comparación de archivos, etc.

El objetivo general de este trabajo de título es estudiar algunos aspectos de la *distribución del largo de la subsecuencia común más grande* de varias palabras, cuando sus letras han sido escogidas de manera aleatoria.

Se presenta un marco histórico sobre el problema. Se generalizan resultados de Chvátal y Sankoff [12] y de Alexander [9] concernientes al comportamiento asintótico del largo esperado de la LCS de dos palabras al caso de varias palabras. En particular, se prueba que para toda ley de probabilidad μ sobre un alfabeto fijo, el largo esperado normalizado de la LCS de varias palabras, cuyas letras son escogidas de acuerdo a μ, tiende a una constante y se da una estimación para la velocidad de convergencia.

A continuación se muestran algunas cotas numéricas encontradas para esta constante para el caso en el que μ es la distribución uniforme sobre un alfabeto pequeño y se muestra una aplicación de estas cotas para refutar una conjetura de Steele [8] que data de los ochenta.

Se presenta además un problema más general denominado problema del subhipergrafo monótono más grande y un resultado que relaciona este problema con el problema de la secuencia creciente más larga de varias permutaciones. Este resultado prueba y extiende, al caso de varias palabras, una conjetura de Sankoff y Mainville recientemente demostrada por Kiwi, Loebl y Matoušek [7] referente al comportamiento asintótico del largo esperado de la LCS de dos palabras cuando el tamaño del alfabeto se va a infinito.

Finalmente se estudian algunas variantes de este problema y se generalizan los resultados obtenidos a dichas variantes.

Agradecimientos

Quiero agradecer a mis padres y hermanos por el apoyo incondicional que me han brindado en todas las metas que he trazado en la vida y en las que vendrán y por todo el cariño que me han entregado siempre.

Aprovecho de agradecer a Giannina, una persona muy especial para mí, no sólo por su incuestionable afecto sino también por su alegría y comprensión en esta etapa de cambios importantes, tanto en lo personal como en lo profesional.

Expreso mi agradecimiento a mi profesor guía, Sr. Marcos Kiwi, por su apoyo y confianza permanentes, así como por todo su valiosa ayuda en mi desarrollo académico. Mención aparte merecen su disposición, consejos y tiempo invertidos durante el desarrollo de esta memoria, los cuales sin lugar a dudas posibilitaron una mejor realización de la misma.

Asimismo agradezco a los profesores integrantes de la comisión por su tiempo e interés en la temática propuesta, y a todos los profesores y académicos que me guiaron durante la realización de mis estudios superiores.

En estas líneas, no quisiera dejar de mencionar a mis amigos y compañeros, no sólo por todos los gratos momentos junto a ellos sino también por su valioso apoyo y compañía.

Finalmente quiero agradecer a Conicyt via FONDAP en Matematicas Aplicadas y al proyecto Anillo en Redes ACT08 por el financiamiento otorgado.

Índice general

1. Introducción **1**

1.1. Organización del trabajo 2

2. Preliminares **4**

2.1. Comportamiento asintótico del largo esperado de la LCS 4

2.2. La subsecuencia diagonal más grande. Estimación de la velocidad de convergencia del largo normalizado de una LCS 5

3. Cotas inferiores para las constantes de Chvátal y Sankoff **15**

3.1. Una cota inferior simple 15

3.2. Una mejor cota inferior 17

3.2.1. Subsecuencia común más larga entre dos palabras en un alfabeto binario . . . 17

3.2.2. Subsecuencia común más larga de varias palabras en un alfabeto binario . . . 19

3.2.3. Encontrando una cota inferior 22

3.2.4. Implementación y cotas obtenidas 24

3.2.5. Extensiones 25

3.3. Aplicación: Conjetura de Steele 26

4. Problema del subhipergrafo monótono de tamaño máximo **28**

4.1. Subhipergrafo monótono de tamaño máximo 29

4.2. Aproximación de la mediana. Teorema Principal 31

4.3. Demostración de las cotas inferiores en el Teorema Principal 34

4.4. Demostración de la cotas superiores en el Teorema Principal 40

4.4.1. Notación y definiciones 40

4.4.2. Demostración para el caso $N < t^{\beta}$ 41

4.4.3. Demostración para el caso $N \geq t^{\beta}$ 42

4.4.4. Demostración de la cota superior para la esperanza y mediana 50

4.5. Resultado para el modelo de d palabras aleatorias 52

4.5.1. Problema de la secuencia creciente más larga o problema de Ulam 52

4.5.2. Reducción al problema de Ulam 53

4.5.3. Aplicación del Teorema Principal al modelo de d-palabras aleatorias 54

4.6. Resultado para el modelo binomial 60

5. Variantes simétricas del problema de la LCS **66**

5.1. Modelo simétrico 66

5.1.1. Reducción al problema de la secuencia creciente más larga de una involución 69

5.1.2. Resultado para el modelo simétrico 70

5.2. Modelo antisimétrico 74

Capítulo 1

Introducción

Dadas dos palabras s y t, diremos que s es una *subsecuencia* de t si la primera puede obtenerse a partir de la segunda borrando cero o más letras sin cambiar la posición relativa de las letras restantes. Dada una lista de d palabras no necesariamente distintas, $(s_i)_{i=1}^{d}$, llamaremos *subsecuencia común más grande* de todos ellas a toda subsecuencia común de largo máximo.

El problema de encontrar una *subsecuencia común más grande* (LCS por sus siglas en inglés: *longest common subsequence*) de un conjunto de palabras es un problema combinatorio que surge cada vez que necesitamos buscar similitudes entre distintos textos.

Una motivación importante a este problema nace del área de la biología molecular. Las moléculas de ADN pueden ser representadas como palabras en un alfabeto de 4 caracteres: $\{A, C, G, T\}$ correspondientes a los nucleótidos presentes en la molécula y, vistas como palabras, pueden constar de millones de caracteres. Calcular una LCS de varias secuencias de ADN nos da una buena idea de cuan similares son dichas moléculas. No sólo eso, en muchos casos, por ejemplo en los virus asociados a ciertas enfermedades, éstas moléculas mutan con facilidad y sería importante conocer cual es el trozo de código genético que tienen en común para poder entender como defenderse de nuevas mutaciones.

Otro campo de aplicación importante es la comparación de archivos. Un ejemplo es el programa "diff" de Unix usado para comparar dos versiones diferentes del mismo archivo. Este programa funciona básicamente considerando las líneas de cada archivo como un carácter y luego calcula la LCS de dichos caracteres retornando las líneas que no pertenecen a dicha subsecuencia. Otras aplicaciones en este campo incluyen la detección de plagio de textos o de códigos fuentes de programas y el uso de control de versiones (CVS) para la creación de un software.

El problema de encontrar una LCS entre una palabra de largo m y otra de largo n puede ser resuelto usando programación dinámica en tiempo $O(nm)$. El algoritmo fue originalmente propuesto por Wagner y Fischer [15] en los setenta y mejorado a lo largo de los años. Una lista de mejoras a este algoritmo puede encontrarse en [3].

El caso general de encontrar una LCS de varias palabras de largo $n_1, \ldots, n_d$ puede ser resuelto usando programación dinámica en tiempo $O(n_1 \cdots n_d)$. Sin embargo, si d no es fijo, el problema resulta ser NP-duro [16].

En esta memoria nos concentraremos en estudiar versiones aleatorias del problema de determinar el largo de una LCS. Principalmente se estudiarán algunos aspectos de la *distribución del largo de la subsecuencia común más grande* de varias palabras, cuando sus letras han sido escogidas de manera aleatoria dentro de un alfabeto conocido.

Un primer paso para de este estudio es conocer como se comporta el largo esperado de la LCS de varias palabras al variar el largo de las mismas. Chvátal y Sankoff [1] en los setentas probaron que el largo de una LCS de dos palabras del mismo largo es asintóticamente lineal con respecto al largo de las palabras pero no fueron capaces de determinar la constante de proporcionalidad entre ambos largos. Esta constante depende del número de palabras y del tamaño del alfabeto del cual las letras son elegidas. Aún en el caso más simple de 2 palabras en un alfabeto binario dicha constante, denominado en la literatura como la constante de Chvátal y Sankoff, no ha podido ser determinada. Se han realizado muchos esfuerzos (ver [2, 3, 4, 5, 6]) para encontrar buenas cotas numéricas para dicha constante, teniendo ninguna de ellas una forma cerrada.

En los 80, Steele [8] conjeturó una relación algebraica entre la constante de Chvátal y Sankoff y su símil para el caso de 3 palabras. Dicha relación no parece ser acertada y de hecho en los 90, Dančík [3] hace una demostración de la falsedad de una generalización de esta conjetura, pero nada se ha dicho para el caso original.

A principios de esta década, Kiwi, Loebl y Matoušek [7] demostraron una conjetura de Sankoff y Mainville [12] referente al comportamiento del radio cuando el tamaño del alfabeto tiende a infinito, relacionando con su demostración este problema con el problema de la secuencia creciente más larga de una permutación, también conocido como problema de Ulam en dos dimensiones.

1.1. Organización del trabajo

El presente trabajo se encuentra dividido de la siguiente manera. En el Capítulo 2 presentaremos algo de nomenclatura y un marco sobre algunos resultados conocidos para el problema de la subsecuencia común más grande de dos palabras, los cuales extenderemos al caso de varias palabras. Específicamente, se probará que para toda ley de probabilidad μ sobre un alfabeto fijo, el largo esperado normalizado de la LCS de varias palabras cuyas letras son escogidas de acuerdo a μ tiende a una constante y se dará una estimación para la velocidad de convergencia.

En el Capítulo 3 se estudiará el método usado por Lueker [6] para encontrar cotas inferiores para la constante de Chvátal y Sankoff y se modificará para ser aplicado al caso de varias palabras. Además se mostrará una aplicación de estas cotas para refutar la conjetura de Steele [8].

En el Capitulo 4 se presentará un problema más general denominado *problema del subhipergrafo*

monótono más grande y un resultado que relaciona este problema con el problema de la secuencia creciente más larga de varias permutaciones. Este resultado prueba y extiende, al caso de varias palabras, la conjetura de Sankoff y Mainville. Finalmente, en el Capítulo 5 se estudiarán algunas variantes de este problema y se extenderán los resultados obtenidos a dichas variantes.

El lector podrá encontrar al principio de cada capitulo una reseña más detallada de los principales resultados que se obtienen en cada uno.

Capítulo 2

Preliminares

Decimos que una palabra s es una *subsecuencia* de una palabra t si la primera puede obtenerse a partir de la segunda borrando cero o más letras sin cambiar la posición relativa de las letras restantes. Dada una lista de d palabras no necesariamente distintas, $(s_i)_{i=1}^{d}$, llamamos *subsecuencia común más grande* de todos ellas a toda subsecuencia común de largo máximo. Además, en lo que sigue al número de palabras, d, lo llamaremos *dimensión* del problema.

En este capítulo estaremos interesados en estudiar el comportamiento del largo de una LCS de d palabras cuando sus letras son elegidas de manera aleatoria de un alfabeto fijo. Específicamente mostraremos que el largo esperado de dicha subsecuencia, normalizado por el largo de las palabras originales, converge a una constante y daremos una estimación para la velocidad de convergencia hacia esa constante.

2.1. Comportamiento asintótico del largo esperado de la LCS

Un primer paso para estudiar la *distribución del largo de la subsecuencia común más grande*, es conocer el comportamiento del largo esperado de dicha subsecuencia. En lo que sigue, denotaremos por $\mathcal{L}_{n,k}^{(d)}(\mu)$ al largo de una LCS de d palabras aleatorias de largo n, donde sus letras son elegidas independientes e idénticamente distribuidas (i.i.d.) de un alfabeto Σ de tamaño k, siguiendo una ley de probabilidad μ. En general, μ será la distribución uniforme sobre Σ, y en este caso, para ahorrar notación, llamaremos $\mathcal{L}_{n,k}^{(d)}$ a la variable aleatoria mencionada anteriormente. Además, en el caso bidimensional ($d = 2$), omitiremos el superíndice d.

Chvátal y Sankoff [1] iniciaron el estudio del comportamiento asintótico en n del largo esperado de una LCS de dos palabras cuyas letras son elegidas de manera uniforme sobre un alfabeto fijo. En nuestra nomenclatura, ellos probaron que:

Teorema 2.1. *Para todo $k \geq 2$, existe una constante $\gamma_{k,2}$ tal que*

$$\gamma_{k,2} = \lim_{n\to\infty} \frac{\mathbb{E}\mathcal{L}_{n,k}}{n} = \sup_{n>0} \frac{\mathbb{E}\mathcal{L}_{n,k}}{n}. \tag{2.1.1}$$

Daremos una demostración de la siguiente extensión del teorema anterior al caso de d palabras elegidas bajo cualquier distribución de probabilidad μ:

Teorema 2.2. *Sea $k \geq 2$ y Σ un alfabeto de tamaño k. Sea μ una distribución de probabilidad sobre Σ, entonces para todo $d \geq 2$, existe $\gamma_{k,d}(\mu)$ tal que*

$$\gamma_{k,d}(\mu) = \lim_{n\to\infty} \frac{\mathbb{E}\mathcal{L}^{(d)}_{n,k}(\mu)}{n} = \sup_{n\in\mathbb{N}} \frac{\mathbb{E}\mathcal{L}^{(d)}_{n,k}(\mu)}{n}. \tag{2.1.2}$$

En el caso que μ sea la distribución uniforme, denotaremos al límite anterior simplemente $\gamma_{k,d}$.

La constante $\gamma_{2,2}$ es llamada comúnmente en la literatura *constante de Chvátal y Sankoff*. Por extensión, llamaremos *constantes de Chvátal y Sankoff* a la familia de constante $\gamma_{k,d}(\mu)$.

El valor exacto de estas constantes, incluso para el caso uniforme, es desconocido. Chvátal y Sankoff [1] dieron las primeras cotas para el valor de $\gamma_{k,2}$, para k pequeño. Para el caso de un alfabeto binario, nuevas cotas y técnicas para encontrarlas fueron creadas posteriormente por Deken [2], Dančík y Paterson [3, 4], Baeza-Yates, Navarro, Gavaldá y Scheihing [5] y Lueker [6].

Para demostrar el Teorema 2.2, basta probar que la cantidad $\mathbb{E}\mathcal{L}^{(d)}_{n,k}(\mu)$ es superaditiva en n, es decir basta probar el siguiente lema:

Lema 2.3 (Superaditividad de $\mathbb{E}\mathcal{L}^{(d)}_{n,k}(\mu)$)**.** *Para todo $m, n \in \mathbb{N}$, $m, n \geq 1$,*

$$\mathbb{E}\mathcal{L}^{(d)}_{n,k}(\mu) + \mathbb{E}\mathcal{L}^{(d)}_{m,k}(\mu) \leq \mathbb{E}\mathcal{L}^{(d)}_{n+m,k}(\mu).$$

Demostración. Sea Σ un alfabeto fijo de tamaño k y μ una ley de probabilidad sobre Σ. Consideremos una secuencia $s_1, s_2, \ldots, s_d$ de palabras de largo n y una secuencia $t_1, t_2, \ldots, t_d$ de palabras de largo m, cuyas letras son elegidas de acuerdo a μ. Es fácil ver que

$$L(s_1, s_2, \ldots, s_d) + L(t_1, t_2, \ldots, t_d) \leq L(s_1t_1, s_2t_2, \ldots, s_dt_d),$$

pues la concatenación de cualquier LCS de $(s_1, s_2, \ldots, s_d)$ con una LCS de $(t_1, t_2, \ldots, t_d)$ resulta ser una subsecuencia común de $(s_1t_1, s_2t_2, \ldots, s_dt_d)$. Tomando esperanza, se concluye el resultado. ■

2.2. La subsecuencia diagonal más grande. Estimación de la velocidad de convergencia del largo normalizado de una LCS

En esta sección definiremos una noción relacionada a la subsecuencia común más grande que nos permitirá estimar cuan rápido converge $\mathbb{E}\mathcal{L}^{(d)}_{n,k}(\mu)/n$ a $\gamma_{k,d}$ y posteriormente, encontrar algunas cotas inferiores para $\gamma_{k,d}$.

Sean X e Y, dos palabras sobre un alfabeto Σ de largo al menos n. Denotemos $X(i)$ a la subpalabra de los primeros i caracteres de X, (respectivamente $Y(j)$ será la subpalabra de los primeros j caracteres de Y). Denotemos por $D_n(X,Y)$ al largo de la subsecuencia común más larga de todos los pares de prefijos que podemos obtener de X e Y tal que el largo conjunto de dichos prefijos sea n. Es decir:

$$D_n(X,Y) = \text{máx}\{L(X(i),Y(j)) \ \mid \ i+j=n\}.$$

Siguiendo la nomenclatura usada por Dančík [3], llamaremos a esta cantidad el largo de la subsecuencia diagonal más grande (DCS por sus siglas en inglés, *diagonal common subsequence*) entre $X(n)$ e $Y(n)$. (Cuando X e Y son palabras de largo n, entonces la denotamos simplemente largo de la DCS entre X e Y).

Además, denotaremos $\mathcal{D}_{n,k}(\mu)$ a la variable aleatoria correspondiente al largo de la subsecuencia diagonal más grande entre dos palabras de largo n cuyas letras son elegidas independientemente de acuerdo a una ley de probabilidad μ.

Alexander [9] probó una relación entre el comportamiento asintótico de la LCS de dos palabras aleatorias de largo n y el comportamiento asintótico de la DCS de dos palabras aleatorias de largo n. En nuestra notación, Alexander probó el siguiente resultado:

Proposición 2.4. *Existe una constante α tal que para todo $n \in \mathbb{N}$,*

$$2\mathbb{E}\mathcal{D}_{n,k}(\mu) - \alpha\sqrt{n\ln n} \leq \mathbb{E}\mathcal{L}_{n,k}(\mu) \leq \mathbb{E}\mathcal{D}_{2n,k}(\mu).$$

Además, para todo k positivo, la cantidad

$$\delta_{k,2}(\mu) = \lim_{n\to\infty} \frac{\mathbb{E}\mathcal{D}_{n,k}(\mu)}{n}$$

está bien definida, y de hecho $\delta_{k,2}(\mu) = \gamma_{k,2}(\mu)/2$.

El resultado anterior es importante pues permite, en el momento de estimar el largo esperado de una LCS de dos palabras, omitir la condición de que ambas palabras involucradas tengan largo n. Sólo basta que la suma de ambos largos sea $2n$.

Veremos que podemos extender este resultado al caso de d palabras. Para ello definamos primero el análogo al largo de la subsecuencia diagonal más grande en el caso d-dimensional. Sean $X_1,\ldots,X_d$ palabras sobre un alfabeto Σ de largo al menos n. Definamos $D_n(X_1,X_2,\ldots,X_d)$, análogamente al caso anterior como sigue:

$$D_n(X_1,X_2,\ldots,X_d) = \text{máx}\Big\{L(X_1(i_1),X_2(i_2),\ldots,X_d(i_d)) \ \mid \ \sum_{j=1}^{d} i_j = n\Big\}.$$

Al igual que en el caso bidimensional, denotamos a esta cantidad la DCS de $X_1(n),\ldots,X_d(n)$. Además, llamaremos $\mathcal{D}^{(d)}_{n,k}(\mu)$ a la variable aleatoria correspondiente al largo de la subsecuencia diagonal más grande de d palabras de largo n cuyas letras son elegidas independientemente de acuerdo a una ley de probabilidad μ. Probaremos una proposición similar a la probada por Alexander [9].

Proposición 2.5. *Existe una constante α tal que para todo $n \geq 2$,*

$$d \cdot \mathbb{E}\mathcal{D}_{n,k}^{(d)}(\mu) - d^{3/2}\alpha\sqrt{n \ln n} \leq \mathbb{E}\mathcal{L}_{n,k}^{(d)}(\mu) \leq \mathbb{E}\mathcal{D}_{nd,k}^{(d)}(\mu).$$

Además, para todo k positivo, y $d \geq 2$, la cantidad

$$\delta_{k,d}(\mu) = \lim_{n\to\infty} \frac{\mathbb{E}\mathcal{D}_{n,k}^{(d)}(\mu)}{n}$$

está bien definida, y de hecho $\delta_{k,d}(\mu) = \gamma_{k,d}(\mu)/d$.

Para probar esta proposición, necesitaremos de la siguiente versión de la desigualdad de Azuma [13].

Lema 2.6 (Desigualdad de Azuma)**.** *Sean $Z_1, \ldots, Z_N$ variables aleatorias independientes, con todos los Z_k tomando valores en un conjunto Λ. Sea $X = f(Z_1, \ldots, Z_N)$, con $f : \Lambda^N \to \mathbb{R}$ una función c-Lipschitz, es decir, una función tal que si $z, z' \in \Lambda^N$ difieren sólo en una coordenada, entonces $|f(z) - f(z')| \leq c$. Entonces, la variable X satisface para todo $t \geq 0$,*

$$\mathbb{P}[X \geq \mathbb{E}X + t] \leq e^{-2t^2/Nc^2},$$
$$\mathbb{P}[X \leq \mathbb{E}X - t] \leq e^{-2t^2/Nc^2}.$$

Demostración Proposición 2.5. Sea $(X_i)_{i=1}^d$, una secuencia de d palabras de largo n. Notemos que si cambiamos un carácter de alguna de las palabras X_i, entonces los valores de $L(X_1, \ldots, X_d)$ y de $D_n(X_1, \ldots, X_d)$ cambian en a lo más 1. Es decir, $\mathcal{L}_{n,k}^{(d)}(\mu)$ y $\mathcal{D}_{n,k}^{(d)}(\mu)$ son 1-Lipschitz (vistos como funciones de Σ^{dn} en $\mathbb{R}$). Es decir, se tienen las hipótesis de la desigualdad de Azuma para ambas cantidades.

Luego,

$$\mathbb{P}\left[\mathcal{D}_{n,k}^{(d)}(\mu) \leq \mathbb{E}\mathcal{D}_{n,k}^{(d)}(\mu) - \sqrt{\frac{nd\ln(2)}{2}}\right] \leq \exp\left(-\frac{2(nd\ln(2)/2)}{nd}\right) = \frac{1}{2}.$$

Con esto, si denotamos $\lambda = \mathbb{E}\mathcal{D}_{n,k}^{(d)}(\mu) - \sqrt{nd\ln(2)/2}$,

$$\begin{aligned}
\frac{1}{2} &\leq \mathbb{P}\left[\mathcal{D}_{n,k}^{(d)}(\mu) > \lambda\right] \\
&= \mathbb{P}\left[L(X_1(i_1), \ldots, X_d(i_d)) > \lambda, \text{ para algún } (i_1, \ldots, i_d)\right] \\
&\leq \sum_{\substack{0<i_1,i_2,\ldots,i_d<n,\\ i_1+i_2+\cdots+i_d=n}} \mathbb{P}\left[L(X_1(i_1), \ldots, X_d(i_d)) > \lambda\right].
\end{aligned}$$

La última desigualdad se tiene pues el máximo de los largos de $L(X_1(i_1), \ldots, X_d(i_d))$ se alcanza en los índices donde ningún prefijo es vacío. Llamemos I al número de términos de la suma del lado

derecho. Luego I es el número de formas de separar n en d sumandos positivos, es decir, $I = \binom{n-1}{d-1}$. Con lo anterior, existe algún multi-índice $(j_1, \ldots, j_d)$ que participa en la suma tal que

$$\frac{1}{2I} \leq \mathbb{P}\left[L(X_1(j_1), \ldots, X_d(j_d)) > \lambda\right].$$

Notemos ahora que si $[d]$ denota, como es habitual, al conjunto $\{1, \ldots, d\}$ y si τ es una permutación cualquiera de $[d]$, entonces $L(X_1(j_{\tau(1)}), \ldots, X_d(j_{\tau(d)}))$ tiene la misma distribución que $L(X_1(j_1), \ldots, X_d(j_d))$. En otras palabras la distribución del largo de la LCS no se ve afectada por el orden relativo entre las palabras. Además, esta variable resulta ser invariante bajo traslación. Es decir, si denotamos $X[a, b]$ a la subpalabra de X ubicada entre los símbolos a-ésimo y b-ésimo (ambos incluídos), y si para todo $1 \leq m \leq d$ se tienen índices $1 \leq a_m \leq b_m \leq n$ tal que $b_m - a_m + 1 = j_m$, entonces $L\left(X_1([a_1, b_1]), \ldots, X_d([a_d, b_d])\right)$ también tiene la misma distribución que $L(X_1(j_1), \ldots, X_d(j_d))$.

Fijemos ahora τ a ser el ciclo $(1\,2\,\ldots\,d)$, es decir $\tau : [d] \to [d]$ es la permutación tal que:

$$\tau(m) = \begin{cases} m+1, & \text{si } m \leq d, \\ 1, & \text{si } m = d. \end{cases}$$

Las observaciones anteriores, permiten concluir que los eventos

$$L(X_1(j_1), \ldots, X_d(j_d)) > \lambda, \tag{B_0}$$

$$L\left(X_1[j_1+1, j_1+j_{\tau(1)}], \ldots, X_d[j_d+1, j_d+j_{\tau(d)}]\right) > \lambda, \tag{B_1}$$

$$\vdots$$

$$L\left(X_1\left[\sum_{l=0}^{m-1} j_{\tau^l(1)} + 1, \sum_{l=0}^{m} j_{\tau^l(1)}\right], \ldots, X_d\left[\sum_{l=0}^{m-1} j_{\tau^l(1)} + 1, \sum_{l=0}^{m} j_{\tau^l(1)}\right]\right) > \lambda, \tag{B_m}$$

$$\vdots$$

$$L\left(X_1\left[\sum_{l=0}^{d-2} j_{\tau^l(1)} + 1, \sum_{l=0}^{d-1} j_{\tau^l(1)}\right], \ldots, X_d\left[\sum_{l=0}^{d-2} j_{\tau^l(1)} + 1, \sum_{l=0}^{d-1} j_{\tau^l(1)}\right]\right) > \lambda. \tag{B_{d-1}}$$

tienen la misma probabilidad, son independientes (pues dependen de caracteres distintos) y además, si se cumplen simultaneamente, entonces $L(X_1, \ldots, X_d) > d\lambda$.

Con esto:

$$\left(\frac{1}{2I}\right)^d \leq (\mathbb{P}(B_0))^d = \mathbb{P}(B_0, B_1, \ldots, B_{d-1}) \leq \mathbb{P}(\mathcal{L}_{n,k}^{(d)}(\mu) > d\lambda).$$

Sin embargo, usando la desigualdad de Azuma obtenemos que:

$$\mathbb{P}\left(\mathcal{L}_{n,k}^{(d)}(\mu) \geq \mathbb{E}\mathcal{L}_{n,k}^{(d)}(\mu) + \sqrt{\frac{d^2 n \ln(2I)}{2}}\right) \leq \exp\left(-\frac{2d^2 n \ln(2I)/2}{dn}\right) = \left(\frac{1}{2I}\right)^d.$$

Como $\lambda = \mathbb{E}\mathcal{D}_{n,k}^{(d)}(\mu) - \sqrt{\frac{nd\ln(2)}{2}}$, se tiene, gracias a las dos desigualdades anteriores, que

$$\mathbb{P}\left(\mathcal{L}_{n,k}^{(d)}(\mu) \geq \mathbb{E}\mathcal{L}_{n,k}^{(d)}(\mu) + \sqrt{\frac{d^2 n\ln(2I)}{2}}\right) \leq \mathbb{P}\left(\mathcal{L}_{n,k}^{(d)}(\mu) > d\left(\mathbb{E}\mathcal{D}_{n,k}^{(d)}(\mu) - \sqrt{\frac{nd\ln(2)}{2}}\right)\right),$$

y luego,

$$\mathbb{E}\mathcal{L}_{n,k}^{(d)}(\mu) + \sqrt{\frac{d^2 n\ln(2I)}{2}} > d\cdot\mathbb{E}\mathcal{D}_{n,k}^{(d)}(\mu) - d\sqrt{\frac{nd\ln(2)}{2}}.$$

Usando la desigualdad $I = \binom{n-1}{d-1} \leq \left(\frac{e(n-1)}{d-1}\right)^{d-1}$, y que $\ln(2) < 1$ se concluye que:

$$\begin{aligned} d\cdot\mathbb{E}\mathcal{D}_{n,k}^{(d)}(\mu) - d\sqrt{\frac{nd\ln(2)}{2}} &< \mathbb{E}\mathcal{L}_{n,k}^{(d)}(\mu) + \sqrt{\frac{d^2 n(\ln(2) + (d-1)\ln(e(n-1)/(d-1)))}{2}} \\ &\leq \mathbb{E}\mathcal{L}_{n,k}^{(d)}(\mu) + d\sqrt{\frac{n(1+(d-1)(1+\ln(n)))}{2}} \\ &\leq \mathbb{E}\mathcal{L}_{n,k}^{(d)}(\mu) + d\sqrt{nd\ln(n)}, \end{aligned}$$

y luego usando que para $n \geq 2$, $\ln(2) \leq \ln(n)$,

$$d\cdot\mathbb{E}\mathcal{D}_{n,k}^{(d)}(\mu) - d\sqrt{nd\ln(n)\left(\frac{1}{2}+1\right)} \leq \mathbb{E}\mathcal{L}_{n,k}^{(d)}(\mu).$$

Tomando $\alpha = \sqrt{3/2}$, se concluye la primera desigualdad.

La segunda desigualdad se tiene pues para toda secuencia de d palabras $X_1, \ldots, X_d$ de largo nd,

$$L(X_1(n), \ldots, X_d(n)) \leq D_{nd}(X_1, \ldots, X_d).$$

Tomando esperanza se concluye.

Demostremos la segunda parte de la proposición. Por lo recién probado, para todo n,

$$\frac{\mathbb{E}\mathcal{L}_{n,k}^{(d)}(\mu)}{n} \leq \frac{\mathbb{E}\mathcal{D}_{nd,k}^{(d)}(\mu)}{n} \leq \frac{\mathbb{E}\mathcal{L}_{nd,k}^{(d)}(\mu) + \alpha\sqrt{nd\ln(nd)}}{nd}.$$

Lo anterior dice que el siguiente límite existe e indica su valor:

$$\lim_{n\to\infty}\frac{\mathbb{E}\mathcal{D}_{nd,k}^{(d)}(\mu)}{n} = \gamma_{k,d}(\mu). \tag{2.2.1}$$

Por otro lado, $\mathbb{E}\mathcal{D}_{n,k}^{(d)}(\mu)$ es no decreciente en n, luego,

$$\frac{\lfloor n/d\rfloor}{n/d}\cdot\frac{\mathbb{E}\mathcal{D}_{d\lfloor n/d\rfloor,k}^{(d)}(\mu)}{\lfloor n/d\rfloor} \leq \frac{\mathbb{E}\mathcal{D}_{n,k}^{(d)}(\mu)}{n/d} \leq \frac{\mathbb{E}\mathcal{D}_{d\lceil n/d\rceil,k}^{(d)}(\mu)}{\lceil n/d\rceil}\cdot\frac{\lceil n/d\rceil}{n/d}.$$

Por (2.2.1), tanto el término de la izquierda como el de la derecha convergen a $\gamma_{k,d}(\mu)$, lo cual concluye la demostración. ■

Alexander [9] probó además el siguiente teorema para estimar la velocidad de convergencia de $\mathbb{E}\mathcal{L}_{n,k}(\mu)$:

Teorema 2.7. *Existe una constante A tal que para todo k, todo μ y $n \geq 1$,*

$$\gamma_{k,2}(\mu)n \geq \mathbb{E}\mathcal{L}_{n,k}(\mu) \geq \gamma_{k,2}(\mu) - A(n\log n)^{1/2}.$$

Extenderemos este teorema al caso de d palabras, para ello necesitaremos probar dos lemas:

Lema 2.8. *Sean $n, m \geq 0$. Luego para toda secuencia de d palabras, $(X_i)_{i=1}^d$, de largo al menos $n+m$ y toda secuencia $i_1, \ldots, i_d$ tal que $i_1 + \cdots + i_d = n+m$,*

$$L(X_1(i_1), \ldots, X_d(i_d))) \leq \max_{\substack{0 \leq j_l \leq i_l, l=1,\ldots,d, \\ j_1+j_2+\cdots+j_d=n}} \Big\{ L(X_1(j_1), \ldots, X_d(j_d)) + L(X_1[j_1+1, i_1], \ldots, X_d[j_d+1, i_d] + 1 \Big\}.$$

Demostración. Sea $M = m_1 \ldots m_N$ una subsecuencia común de $(X_1(i_1), \ldots, X_d(i_d))$ de largo máximo. Denotemos $c^{(k)} = (c_1^{(k)}, \ldots, c_N^{(k)})$ al vector de $\mathbb{N}^N$ que indica la posición de cada carácter de M como subsecuencia de la palabra k-ésima y denotemos para cada $1 \leq l \leq N$, $e_l = (c_l^{(1)}, \ldots, c_l^{(d)})$ la posición del l-ésimo carácter de M en cada palabra. Además, por conveniencia definamos $e_0 = (0, \ldots, 0)$ y $e_{N+1} = (i_1 + 1, \ldots, i_d + 1)$. Asignemos además a cada vector $e \in \mathbb{N}^d$, su módulo $|e|$ definido como la suma de las componentes de e. Por ejemplo, si consideramos la LCS $M = \mathbf{bbac}$, marcada con negrita en las siguientes palabras:

$$\begin{array}{lccccccc} X_1: & \mathbf{b} & \mathrm{c} & \mathrm{a} & \mathbf{b} & \mathbf{a} & \mathbf{c} & \mathrm{c} \\ X_2: & \mathrm{c} & \mathbf{b} & \mathbf{b} & \mathbf{a} & \mathrm{a} & \mathrm{a} & \mathbf{c} \\ X_2: & \mathbf{b} & \mathrm{a} & \mathrm{a} & \mathbf{b} & \mathrm{c} & \mathbf{a} & \mathbf{c} \end{array}$$

Entonces $c_1 = (1,4,5,6)$, $c_2 = (2,3,4,7)$ y $c_3 = (1,4,6,7)$, y luego $e_0 = (0,0,0), e_1 = (1,2,1)$, $e_2 = (4,3,4), e_3 = (5,4,6), e_4 = (6,7,7), e_5 = (8,8,8)$.

Notemos que los $(e_l)_{0 \leq l \leq N+1}$ son vectores en $\mathbb{N}^d$ que forman una cadena creciente con respecto al orden parcial natural de $\mathbb{N}^d$, donde un vector es menor que otro si la desigualdad se tiene componente a componente.

Llamemos ahora T al índice del primer vector de la cadena $(e_l)_{0 \leq l \leq N+1}$ tal que $|e_T| \geq n$, y llamemos $\bar{e} = (\bar{e}_1, \ldots, \bar{e}_d)$ a algún vector de módulo n menor o igual que e_T y mayor o igual que e_{T-1} (componente a componente). Como $|e_{T-1}| < n$ dicho vector existe. En el ejemplo anterior, si n fuera 13, entonces T sería 3 (pues $e_2 = (4,2,4)$ tiene módulo 11 y $e_3 = (5,4,6)$ tiene módulo 15), y luego podemos definir $\bar{e}$ como cualquier vector de módulo 13 entre $(4,3,4)$ y $(5,4,6)$, por ejemplo $\bar{e} = (4,4,5)$.

Dividamos cada una de las palabras en dos subpalabras, las subpalabras formadas por los caracteres que están antes de la posición $\bar{e}$ (incluido) y las subpalabras formadas por los caracteres

que están después de $\bar{e}$ (sin incluir). Con esto, los caracteres asociados a e_{T-1} están completamente contenidos en el primer bloque de subpalabras, y, para todo $l \geq T+1$, los caracteres asociados a e_l quedan completamente contenido en el segundo bloque de palabras. Sin embargo, al ser $\bar{e}$ menor o igual componente a componente que e_T, es posible que al hacer la división, los caracteres asociados a e_T queden separados. En nuestro ejemplo las palabras divididas quedan:

$$\begin{array}{lccccc|ccc} X_1 & \mathbf{b} & \mathrm{c} & \mathrm{a} & \mathbf{b} & & \mathbf{a} & \mathbf{c} & \mathrm{c} \\ X_2 & \mathrm{c} & \mathbf{b} & \mathbf{b} & \mathbf{a} & & \mathrm{a} & \mathrm{a} & \mathbf{c} \\ X_2: & \mathbf{b} & \mathrm{a} & \mathrm{a} & \mathbf{b} & \mathrm{c} & & \mathbf{a} & \mathbf{c} \end{array}$$

y luego uno de los alineamientos (el carácter a asociado a $e_3 = (5, 4, 6)$) queda cortado.

Con esto, el largo de la LCS del primer bloque, $L(X_1(\bar{e}_1), \ldots, X_d(\bar{e}_d))$ resulta ser mayor o igual al numero de aristas de la cadena (exceptuando e_0) que quedan en el primer bloque, es decir $T-1$, y el largo de la LCS del segundo bloque, $L(X_1[\bar{e}_1 + 1, i_1], \ldots, X_d[\bar{e}_d + 1, i_d])$ resulta ser mayor o igual al numero de las aristas de la cadena (exceptuando e_{N+1}) que quedan en el segundo bloque, es decir, $N - T$. Por lo tanto,

$$\begin{aligned} \max_{\substack{0 \leq j_l \leq i_l, l=1,\ldots,d, \\ j_1+j_2+\cdots+j_d=n}} & \Big\{ L(X_1(j_1), \ldots, X_d(j_d)) + L(X_1[j_1+1, i_1], \ldots, X_d[j_d+1, i_d] + 1 \Big\} \\ & \geq L(X_1(\bar{e}_1), \ldots, X_d(\bar{e}_d) + L(X_1[\bar{e}_1 + 1, i_1], \ldots, X_d[\bar{e}_d + 1, i_d] + 1 \\ & \geq T - 1 + N - T + 1 = L(X_1(i_1), \ldots, X_d(i_d)). \end{aligned}$$

■

Lema 2.9. *Para $n \geq 1$ y $\beta > 0$ se definen las funciones generatrices*

$$g_n(\beta) = \ln \left(\sum_{\substack{0 \leq i_1, \ldots, i_d, \\ i_1+i_2+\cdots+i_d=nd}} \mathbb{E} \exp\left(\beta[L(X_1(i_1), \ldots, X_d(i_d)) + 1]\right) \right).$$

Para todo $\beta > 0$, la secuencia $\{g_n(\beta) : n \geq 1\}$ es subaditiva, es decir, $g_{n+m}(\beta) \leq g_n(\beta) + g_m(\beta)$.

Demostración. Por el lema anterior y la invarianza bajo traslación del valor esperado del largo de

una LCS,

$$\begin{aligned}
&\sum_{\substack{0\le i_1,\dots,i_d,\\ i_1+i_2+\cdots+i_d=d(n+m)}} \mathbb{E}\exp\left(\beta[L(X_1(i_1),\dots,X_d(i_d))+1]\right)\\
&\qquad\le \sum_{\substack{0\le i_1,\dots,i_d,\\ i_1+i_2+\cdots+i_d=d(n+m)}} \mathbb{E}\exp\left(\beta\Big[\max_{\substack{0\le j_l\le i_l, l=1,\dots,d,\\ j_1+j_2+\cdots+j_d=dn}} L(X_1(j_1),\dots,X_d(j_d))\right.\\
&\qquad\qquad\left.+L(X_1[j_1+1,i_1],\dots,X_d[j_d+1,i_d])+2\Big]\right)\\
&\qquad\le \sum_{\substack{0\le i_1,\dots,i_d,\\ i_1+i_2+\cdots+i_d=d(n+m)}}\ \sum_{\substack{0\le j_l\le i_l, l=1,\dots,d,\\ j_1+j_2+\cdots+j_d=dn}} \mathbb{E}\exp\left(\beta\Big[L(X_1(j_1),\dots,X_d(j_d))+1\right.\\
&\qquad\qquad\left.+L(X_1[j_1+1,i_1],\dots,X_d[j_d+1,i_d])+1\Big]\right)\\
&\qquad= \sum_{\substack{0\le i_1,\dots,i_d,\\ i_1+i_2+\cdots+i_d=d(n+m)}}\ \sum_{\substack{0\le j_l\le i_l, l=1,\dots,d,\\ j_1+j_2+\cdots+j_d=dn}} \mathbb{E}\exp\left(\beta\Big[L(X_1(j_1),\dots,X_d(j_d))+1\Big]\right)\\
&\qquad\qquad\cdot\mathbb{E}\exp\left(\beta\Big[L(X_1[1,i_1-j_1],\dots,X_d[1,i_d-j_d])+1\Big]\right).
\end{aligned}$$

Haciendo el cambio de variable $r_1=i_1-j_1,\dots,r_d=i_d-j_d$, lo anterior queda igual al producto

$$\begin{aligned}
&\sum_{\substack{0\le j_1,\dots,j_d,\\ j_1+j_2+\cdots+j_d=dn}} \mathbb{E}\exp\left(\beta\Big[L(X_1(j_1),\dots,X_d(j_d))+1\Big]\right)\\
&\cdot\sum_{\substack{0\le r_1,\dots,r_d,\\ r_1+r_2+\cdots+r_d=dm}} \mathbb{E}\exp\left(\beta\Big[L(X_1(r_1),\dots,X_d(r_d))+1\Big]\right).
\end{aligned}$$

Tomando logaritmo se concluye. ■

De los lemas anteriores podemos deducir el siguiente teorema de estimación de la velocidad de convergencia de $\mathbb{E}\mathcal{L}_{n,d}(\mu)$:

Teorema 2.10. *Existe una constante A tal que para todo k, todo μ y $n\ge 2d$,*

$$n\gamma_{k,d}(\mu)\ge\mathbb{E}\mathcal{L}_{n,d}(\mu)\ge n\gamma_{k,d}(\mu)-Ad^2(n\log n)^{1/2}.$$

Demostración. La primera desigualdad se tiene por superaditividad de $\mathbb{E}\mathcal{L}_{n,d}(\mu)$. Para probar la segunda desigualdad primero acotaremos superior e inferiormente $g_n(\beta)$ y nos aprovecharemos de

que esta cantidad es subaditiva. Denotemos S al número de términos que aparece en la suma correspondiente a la definición de $g_n(\cdot)$. Es decir, $S = |\{0 \leq (i_1, \ldots, i_d) : i_1 + \cdots + i_d = nd\}| = \binom{nd+d-1}{d}$. Como todos los términos de la suma son menores o iguales que $\mathbb{E}\exp\left(\beta[\mathcal{D}^{(d)}_{nd,k}(\mu)+1]\right)$,

$$\mathbb{E}\exp\left(\beta[\mathcal{D}^{(d)}_{nd,k}(\mu)+1]\right)) \leq \exp(g_n(\beta)) \leq S\mathbb{E}\exp\left(\beta[\mathcal{D}^{(d)}_{nd,k}(\mu)+1]\right)).$$

Usando la desigualdad de Jensen vemos que

$$g_n(\beta) \geq \ln \mathbb{E}\exp\left(\beta[\mathcal{D}^{(d)}_{nd,k}(\mu)+1]\right)) \geq \beta\mathbb{E}[\mathcal{D}^{(d)}_{nd,k}(\mu)+1]$$

y luego, por subaditividad de $g_n(\cdot)$, se tiene,

$$\frac{g_n(\beta)}{\beta n} \geq \lim_n \frac{g_n(\beta)}{\beta n} \geq \lim_n \frac{\mathbb{E}\mathcal{D}^{(d)}_{nd,k}(\mu)+1}{n} = d\delta_{k,d}(\mu) = \gamma_{k,d}(\mu).$$

Con lo cual, para todo β y todo n (notar que β puede depender de n),

$$\frac{g_n(\beta)}{\beta} \geq n\gamma_{k,d}(\mu).$$

Ahora acotemos superiormente $g_n(\beta)$. Llamemos, para evitar notación innecesaria, $\mathcal{D} = \mathcal{D}^{(d)}_{nd,k}(\mu)$, y sea $\mathbb{E}\mathcal{D} < \lambda < n$ un valor a ser especificado más tarde. Se tiene,

$$\begin{aligned}
\mathbb{E}\exp(\beta\mathcal{D}) &= \mathbb{E}\Big[\exp(\beta\mathcal{D})\mathbb{1}_{\{\mathcal{D}\leq\lambda\}}\Big] + \mathbb{E}\Big[\exp(\beta\mathcal{D})\mathbb{1}_{\{\lambda\leq\mathcal{D}\leq n\}}\Big] \\
&\leq e^{\beta\lambda}\mathbb{P}[\mathcal{D}\leq\lambda] + \Big(-e^{\beta x}\mathbb{P}[\mathcal{D}>x]\Big)\Big|_\lambda^n + \int_\lambda^n \beta e^{\beta x}\mathbb{P}[\mathcal{D}>x]dx \\
&= e^{\beta\lambda} + \int_\lambda^n \beta e^{\beta x}\mathbb{P}[\mathcal{D}>x]dx.
\end{aligned}$$

Usando la desigualdad de Azuma, lo anterior es menor o igual que

$$\begin{aligned}
& e^{\beta\lambda} + \int_\lambda^n \beta e^{\beta x}\exp\left(\frac{-2(x-\mathbb{E}\mathcal{D})^2}{nd^2}\right)dx \\
&= e^{\beta\lambda} + \beta\int_\lambda^n \exp\left(\frac{-2[x-(\mathbb{E}\mathcal{D}+\beta nd^2/4)]^2}{nd^2} + \beta(\mathbb{E}\mathcal{D}+\beta nd^2/8)\right)dx \\
&\leq e^{\beta\lambda} + \beta e^{\beta(\mathbb{E}\mathcal{D}+\beta nd^2/4)}\int_\lambda^n \exp\left(\frac{-2[x-(\mathbb{E}\mathcal{D}+\beta nd^2/4)]^2}{nd^2}\right)dx.
\end{aligned}$$

Tomando $\lambda = \mathbb{E}\mathcal{D} + \beta nd^2/4$ y usando el cambio de variable $z = \sqrt{2/(nd^2)}(x-\lambda)$, lo anterior es igual a

$$e^{\beta\lambda} + \beta e^{\beta\lambda}\int_\lambda^n \exp\left(\frac{-2[x-\lambda]^2}{nd^2}\right)dx \leq e^{\beta\lambda} + \beta e^{\beta\lambda}\sqrt{\frac{nd^2}{2}}\int_0^\infty \exp\left(-z^2\right)dz.$$

Es decir,

$$\mathbb{E}\exp(\beta\mathcal{D}) \leq \exp(\beta\mathbb{E}\mathcal{D} + \beta^2 nd^2/4)\left[1 + \beta\sqrt{\frac{nd^2\pi}{8}}\right].$$

Luego, recordando la primera cota para $g_n(\beta)$ y que $\ln(1+z) < z$ para todo $z > 0$,

$$\begin{aligned} g_n(\beta) &\leq \ln\left(S\mathbb{E}\exp\left(\beta[\mathcal{D}+1]\right)\right) = \ln\left(Se^{\beta}\mathbb{E}\exp(\beta\mathcal{D})\right) \\ &\leq \ln\left(S\exp(\beta + \beta\mathbb{E}\mathcal{D} + \beta^2 nd^2/4)\left[1 + \beta\sqrt{\frac{nd^2\pi}{8}}\right]\right) \\ &\leq \ln(S) + \beta + \beta\mathbb{E}\mathcal{D} + \beta^2 nd^2/4 + \beta\sqrt{\frac{nd^2\pi}{8}}. \end{aligned}$$

Usando la desigualdad $S = \binom{nd+d-1}{d} \leq \left(\frac{e(nd+d-1)}{d}\right)^d \leq (e(n+1))^d < (nd)^{2d}$ si $n \geq 2$, la Proposición 2.5, y la cota inferior para $g_n(\beta)$, tenemos:

$$\begin{aligned} \mathbb{E}\mathcal{L}^{(d)}_{nd,k}(\mu) &\geq -d^{3/2}\alpha\sqrt{nd\ln(nd)} + d\cdot\mathbb{E}\mathcal{D}^{(d)}_{nd,k}(\mu) \\ &\geq -d^2\alpha\sqrt{n\ln(nd)} + d\left(\frac{g_n(\beta)}{\beta} - \frac{\ln(S)}{\beta} - 1 - \beta nd^2/4 - \sqrt{\frac{nd^2\pi}{8}}\right) \\ &\geq -d^2\alpha\sqrt{n\ln(nd)} + d\left(n\gamma_{k,d}(\mu) - \frac{2d\ln(n)}{\beta} - 1 - \beta nd^2/4 - \sqrt{\frac{nd^2\pi}{8}}\right) \\ &= dn\gamma_{k,d}(\mu) - d^2\alpha\sqrt{n\ln(nd)} - d^2\left(\frac{2\ln(n)}{\beta} + 1/d + \beta nd/4 + \sqrt{\frac{n\pi}{8}}\right). \end{aligned}$$

Finalmente como β lo podemos hacer depender de n, definamos $\beta = \sqrt{\ln(nd)/(nd)} \geq 0$, con lo cual

$$\begin{aligned} \mathbb{E}\mathcal{L}^{(d)}_{nd,k}(\mu) &\geq dn\gamma_{k,d}(\mu) - d^2\alpha\sqrt{n\ln(nd)} - d^2\left(2\sqrt{nd\ln(nd)} + 1/d + \frac{1}{4}\sqrt{nd\ln(nd)} + \sqrt{\frac{n\pi}{8}}\right) \\ &\geq dn\gamma_{k,d}(\mu) - d^2\sqrt{nd\ln(nd)}\,(\alpha+3). \end{aligned}$$

Con esto tenemos la cota buscada para todos los múltiplos de d mayores o iguales que $2d$. Para extender esta cota notemos que para todo $n \geq 2d$, $n \geq \left\lfloor\frac{n}{d}\right\rfloor \cdot d \geq \frac{n}{2}$. Con esto, por superaditividad de $\mathbb{E}\mathcal{L}^{(d)}_{nd,k}(\mu)$ y el resultado recién probado,

$$\begin{aligned} \mathbb{E}\mathcal{L}^{(d)}_{n,k}(\mu) = n\frac{\mathbb{E}\mathcal{L}^{(d)}_{n,k}(\mu)}{n} &\geq n\frac{\mathbb{E}\mathcal{L}^{(d)}_{\left\lfloor\frac{n}{d}\right\rfloor d,k}(\mu)}{\left\lfloor\frac{n}{d}\right\rfloor d} \\ &\geq n\gamma_{k,d}(\mu) - \frac{nd^2(\alpha+3)\sqrt{\left\lfloor\frac{n}{d}\right\rfloor d\ln(\left\lfloor\frac{n}{d}\right\rfloor d)}}{\left\lfloor\frac{n}{d}\right\rfloor d} \\ &\geq \gamma_{k,d}(\mu) - 2(\alpha+1)d^2\sqrt{n\ln n}. \end{aligned}$$

■

Capítulo 3

Cotas inferiores para las constantes de Chvátal y Sankoff

En este capítulo mostraremos algunas cotas inferiores para las constantes de Chvátal y Sankoff. Daremos primero una cota inferior simple que no depende del número de palabras, sino explícitamente de la distribución μ con la que se eligen los caracteres de las palabras. Luego nos enfocaremos en acotar inferiormente las constantes correspondientes al caso en el que μ es la distribución uniforme sobre un alfabeto pequeño.

Las mejores cotas conocida para el caso de 2 palabras elegidas uniformemente sobre un alfabeto binario, es decir, para la constante $\gamma_{2,2}$, fueron obtenidas hace algunos años por Lueker [6]. Lueker probó que

$$0{,}788071 \leq \gamma_{2,2} \leq 0{,}826280.$$

En este capítulo extenderemos las Ideas y técnicas usadas por Lueker para probar la cota inferior de $\gamma_{2,2}$, para poder encontrar nuevas cotas en el caso de d palabras.

3.1. Una cota inferior simple

En esta sección mostraremos una cota inferior para las constantes de Chvátal y Sankoff, $\gamma_{k,d}(\mu)$, que no depende de d. Para ello usaremos la siguiente versión de la desigualdad de Chernoff [13, Remark 2.5].

Lema 3.1 (Desigualdad de Chernoff)**.** *Sean $X_1, \ldots, X_n$, variables aleatorias independientes del tipo Bernoulli de parámetro p y sea $X = \sum_{i=1}^{n} X_i$. Entonces para todo $t > 0$,*

$$\mathbb{P}(X \leq np - t) \leq \exp\left(\frac{-2t^2}{n}\right).$$

Para una palabra aleatoria de largo n suficientemente grande, el número de apariciones de un carácter σ fijo en dicha palabra es aproximadamente $n\mu(\sigma)$. Intuitivamente lo anterior quiere decir que si n es suficientemente grande y disponemos de d palabras aleatorias de largo n, la palabra formada por $n\mu(\sigma)$ caracteres 'σ' es, con alta probabilidad, una subsecuencia común de todas las palabras. Veremos a continuación que lo anterior es cierto y que, por lo tanto podemos acotar inferiormente $\gamma_{k,d}(\mu)$ por la probabilidad de σ bajo μ para cualquier carácter σ del alfabeto.

Proposición 3.2. *Para todo $d \geq 2$, $k \geq 2$ y toda distribución de probabilidad μ sobre un alfabeto Σ de tamaño k.*

$$\gamma_{k,d}(\mu) \geq \max_{\sigma \in \Sigma} \mu(\sigma).$$

Demostración. Sea $\sigma \in \Sigma$ un símbolo del alfabeto. Tomemos $S_1, \ldots, S_d$, palabras aleatorias de largo n sobre Σ elegidas independientemente bajo μ. Definamos Y_i como el número de apariciones de σ en la palabra S_i e $Y = \min\{Y_1, \ldots, Y_d\}$. Sea además $0 < \varepsilon < 1$ y $p = \mu(\sigma)$. Por desigualdad de Markov, y de la definición e independencia de los Y_i tenemos que

$$\begin{aligned}
\mathbb{E}(Y) &\geq np(1-\varepsilon)\mathbb{P}(Y \geq np(1-\varepsilon)) \\
&= np(1-\varepsilon)\mathbb{P}(\forall 1 \leq i \leq d, Y_i \geq np(1-\varepsilon)) \\
&= np(1-\varepsilon)\left[\mathbb{P}(Y_1 \geq np(1-\varepsilon))\right]^d \\
&\geq np(1-\varepsilon)\left[1 - \mathbb{P}(Y_1 \leq np(1-\varepsilon))\right]^d.
\end{aligned}$$

Notemos ahora que Y_1 es suma de las indicatrices asociadas a los eventos consistentes en que el j-ésimo carácter de Y_1 sea σ. Dichas indicatrices son variables Bernoulli independientes de parámetro p. Luego, por el lema anterior, $\mathbb{P}(Y_1 \leq np - np\varepsilon) \leq \exp(-2n(p\varepsilon)^2)$. Sigue que:

$$\mathbb{E}(Y) \geq np(1-\varepsilon)\left[1 - \exp\left(-2n(p\varepsilon)^2\right)\right]^d.$$

Tomando n suficientemente grande, digamos $n \geq -\dfrac{1}{2(p\varepsilon)^2} \ln\left(1 - \left[\dfrac{1-2\varepsilon}{1-\varepsilon}\right]^{1/d}\right)$, se obtiene que:

$$\mathbb{E}(Y) \geq np(1-\varepsilon)\left(\frac{1-2\varepsilon}{1-\varepsilon}\right) = np(1-2\varepsilon).$$

Notemos ahora que por definición de Y, cada palabra S_i tiene al menos Y apariciones de σ, es decir, la palabra σ^Y, que corresponde a la palabra de largo Y cuyos caracteres son todos iguales a σ, es una subsecuencia común de $S_1, \ldots, S_d$. Por lo tanto el largo de la subsecuencia común más grande entre ellas debe ser mayor que Y. Es decir, $L(S_1, \ldots, S_d) \geq Y$. Con esto,

$$\frac{\mathbb{E}\mathcal{L}_{n,k}^{(d)}(\mu)}{n} = \frac{\mathbb{E}(L(S_1, S_2, \ldots, S_d))}{n} \geq \frac{\mathbb{E}(Y)}{n} \geq p(1-2\varepsilon) = \mu(\sigma)(1-2\varepsilon).$$

Tomando límite en n y luego haciendo tender ε a 0 se deduce que para todo σ,

$$\gamma_{n,k}^{(d)}(\mu) \geq \mu(\sigma).$$

■

Aplicando la proposición anterior al caso de distribución uniforme se obtiene directamente el siguiente corolario.

Corolario 3.3. *Para todo $d \geq 2$ y $k \geq 2$,*

$$\gamma_{k,d} \geq \frac{1}{k}.$$

3.2. Una mejor cota inferior

En esta sección aplicaremos la técnica usada por Lueker [6] para encontrar cotas inferiores para el caso bidimensional en un alfabeto binario (con distribución uniforme) y la extenderemos al caso d-dimensional, lo que nos permitirá obtener cotas para $\gamma_{2,d}$, para d pequeño.

En la siguiente subsección daremos las definiciones usadas por Lueker para el caso de dos palabras.

3.2.1. Subsecuencia común más larga entre dos palabras en un alfabeto binario

Sean X_1 y X_2 dos palabras aleatorias de largo n sobre un alfabeto binario cuyas letras son elegidas de manera uniforme. Para s y t, dos palabras fijas en dicho alfabeto, definimos

$$W_n(s,t) = \mathbb{E}\left(\max_{i+j=n} L\left(sX_1(i), tX_2(j)\right)\right), \tag{3.2.1}$$

donde $X(i)$ representa la subpalabra de los primeros i caracteres de X, y $L(a,b)$ es, como antes, el largo de una subsecuencia común más larga entre a y b. Informalmente $W_n(s,t)$ representa el largo esperado de una LCS entre dos palabras cuyos prefijos son s y t y cuyos sufijos consisten en total de n caracteres aleatorios.

La Proposición 2.4 mostrada en el capítulo anterior, referente a la subsecuencia diagonal más larga, permite concluir que, sin importar s o t,

$$\gamma_{2,2} = \lim_{n\to\infty} \frac{1}{n} W_{2n}(s,t). \tag{3.2.2}$$

La idea es aproximar $\gamma_{2,2}$ por $W_{2n}(s,t)$. Fijemos, para ello un $l \in \mathbb{N}$, l será el largo de las palabras s y t usadas. Llamemos w_n al vector de 2^{2l} coordenadas cuyas componentes son todos los valores de $W_n(s,t)$ cuando s y t varían sobre todas las palabras de largo l. En adelante, todos los vectores que usaremos tendrán coordenadas indexadas por secuencias de palabras. Para evitar confusiones reservaremos la notación de paréntesis cuadrados para cuando queramos referirnos a una coordenada de un vector y paréntesis redondos para evaluación de funciones.

Por ejemplo w_n tendrá 16 coordenadas si $l = 2$:

$$w_n = \begin{pmatrix} w_n[00,00] \\ w_n[00,01] \\ \vdots \\ w_n[11,10] \\ w_n[11,11] \end{pmatrix} = \begin{pmatrix} W_n(00,00) \\ W_n(00,01) \\ \vdots \\ W_n(11,10) \\ W_n(11,11) \end{pmatrix}.$$

Lueker [6] estableció una cota inferior para cada una de las componentes de w_n en función de w_{n-1} y w_{n-2}. Para poder ver dicha cota, introduzcamos la siguiente notación: Si $s = s^1 s^2 \dots s^l$ es una palabra de largo $l \geq 2$, llamaremos $I(s)$ (por inicio) a la primera letra de s y $C(s)$ (por cola) a la subpalabra que queda al eliminar la primera letra de s, es decir: $I(s) = s^1, C(s) = s^2 \dots s^l$, y luego $s = I(s)C(s)$.

Con esto podemos encontrar fácilmente la siguiente relación entre w_n, w_{n-1} y w_{n-2}.

$\diamond$ Si $I(s) = I(t)$, entonces:

$$w_n[s,t] \geq 1 + \frac{1}{4} \sum_{(c,c') \in \{0,1\}^2} w_{n-2}[C(s)c, C(t)c']. \tag{3.2.3}$$

$\diamond$ Si $I(s) \neq I(t)$, entonces:

$$w_n[s,t] \geq \text{máx} \left\{ \frac{1}{2} \sum_{c \in \{0,1\}} w_{n-1}[C(s)c, t], \frac{1}{2} \sum_{c \in \{0,1\}} w_{n-1}[s, C(t)c] \right\}. \tag{3.2.4}$$

Usando las desigualdades anteriores es fácil definir una función T tal que para todo $n \geq 2$

$$w_n \geq T(w_{n-1}, w_{n-2}), \tag{3.2.5}$$

donde la desigualdad es componente a componente. Más aún, la función T se puede descomponer en dos funciones más simples: $T_=$ y $T_\neq$, tales que, si $\Pi_=(w)$ (respectivamente $\Pi_\neq(w)$) es la proyección sobre las componentes que corresponden a los pares de palabras cuya primera letra coincide (respectivamente no coincide), entonces para todo $n \geq 2$:

$$\Pi_=(w_n) \geq T_=(w_{n-2}), \tag{3.2.6}$$
$$\Pi_\neq(w_n) \geq T_\neq(w_{n-1}). \tag{3.2.7}$$

Antes de proceder a la generalización a varias palabras, veamos un par de ejemplos. Si queremos calcular $w_n[001, 011]$, puesto que los caracteres iniciales coinciden, usaremos la transformación $T_=$:

$$\begin{aligned} w_n[001,011] &\geq T_=(w_{n-2})[001,011] \\ &= 1 + \frac{1}{4}\big(w_{n-2}[010,110] + w_{n-2}[010,111] + w_{n-2}[011,110] + w_{n-2}[011,111]\big). \end{aligned}$$

Si queremos, en cambio, calcular $w_n[001, 111]$, como los caracteres iniciales difieren, usaremos la transformación $T_{\neq}$:

$$\begin{aligned} w_n[001,111] &\geq T_{\neq}(w_{n-1})[001,111] \\ &= \text{máx}\left\{\frac{1}{2}\big(w_{n-1}[010,111] + w_{n-1}[011,111]\big), \frac{1}{2}\big(w_{n-1}[001,110] + w_{n-1}[001,111]\big)\right\}. \end{aligned}$$

3.2.2. Subsecuencia común más larga de varias palabras en un alfabeto binario

Podemos extender las definiciones de la subsección anterior fácilmente a d palabras de la siguiente manera: Sea $X = (X_i)_{i=1}^d$ una colección de d palabras aleatorias de largo n sobre un alfabeto binario y sea $S = (s_i)_{i=1}^d$ una colección de d palabras fijas en dicho alfabeto. Definimos

$$W_n^{(d)}(s_1, \ldots, s_d) = \mathbb{E}\left(\max_{i_1+i_2+\ldots+i_d=n} L\left(s_1X_1(i_1), s_2X_2(i_2), \ldots, s_dX_d(i_d)\right)\right). \tag{3.2.8}$$

Esta cantidad representa el largo esperado de una LCS entre d palabras de prefijos $s_1, \ldots, s_d$ y sufijos consistentes en un total de n caracteres aleatorios repartidos en las d palabras.

La Proposición 2.5 referente a la subsecuencia diagonal más larga de varias palabras demostrada en el capítulo anterior, permite concluir que para toda secuencia de d palabras $S = (s_i)_{i=1}^d$:

$$\gamma_{2,d} = \lim_{n\to\infty} \frac{1}{n} W_{dn}^{(d)}(s_1, \ldots, s_d). \tag{3.2.9}$$

Por claridad, omitiremos en adelante el superíndice d.

Al igual que en el caso de dos palabras descrito en la sección anterior, fijemos un $l \in \mathbb{N}$. Llamemos w_n al vector de 2^{ld} coordenadas, cuyas componentes son todos los valores de $W_n(s_1, \ldots, s_d)$ cuando $s_1, \ldots, s_d$ varían sobre todas las secuencias de d palabras de largo l. Siguiendo la notación de la subsección anterior, definamos $w_n[s_1, \ldots, s_d] = W_n(s_1, \ldots, s_d)$.

En el caso de dos palabras encontramos una función que permitía establecer cotas para un vector w_n en función de los dos vectores w_{n-1} y w_{n-2}. Ahora necesitaremos d vectores.

Para $S = (s_i)_{i=1}^d$ una secuencia de d palabras llamemos $J_0(S)$ (respectivamente $J_1(S)$) a los índices j en $[d] = \{1, \ldots, d\}$ tales que $I(s_j) = 0$ (respectivamente, tales que $I(s_j) = 1$). Usaremos además la notación $\overline{J_0}(S) = [d] \setminus J_0(S) = J_1(S)$ y $\overline{J_1}(S) = [d] \setminus J_1(S) = J_0(S)$

Es fácil ver que si las d palabras comienzan con el mismo carácter, es decir $|J_0(S)| = d$ ó $|J_1(S)| = d$, entonces

$$w_n[s_1, \ldots, s_n] \geq 1 + \frac{1}{2^d} \sum_{\vec{c}\in\{0,1\}^d} w_{n-d}[C(s_1)c_1, \ldots, C(s_d)c_d]. \tag{3.2.10}$$

Informalmente, la desigualdad anterior dice que si todas las palabras de S comienzan con el mismo carácter, entonces el largo de la subsecuencia común más larga entre ellas (permitiendo

sufijos de n caracteres aleatorios) es al menos 1 (el carácter en que coinciden) más el largo de la subsecuencia común más larga entre las palabras obtenidas al eliminar el primer carácter y tomar prestado d caracteres (los eliminados) de los n caracteres aleatorios.

Si no todas las palabras comienzan con el mismo carácter también podemos encontrar una cota, pero para ello necesitaremos algo de notación extra.

Dados A y B dos conjuntos, definimos A^B como el conjunto de todas las funciones de B en A. Para una secuencia de d palabras $S = (s_i)_{i=1}^d$, y $c \in \{0,1\}^{\overline{J_0}(S)}$ una función a valores en $\{0,1\}$ definida en el conjunto de índices i donde la palabra s_i no comienza con 0, definimos $\tau_0(S,c)$ a la *traslación de S que mantiene los 0 iniciales y concatena c*, es decir, a la secuencia de d palabras que resulta al cambiar cada palabra de S que no comience por 0 por la palabra que resulta al eliminar su primer carácter y concatenar, al final, el carácter correspondiente por c. Explícitamente, $\tau_0(S,c) = \big(\tau_0(S,c)_i\big)_{i=1}^d$ con

$$\tau_0(S,c)_i = \begin{cases} s_i, & \text{si } i \in J_0(S), \\ C(s_i)c(i), & \text{si } i \notin J_0(S). \end{cases} \tag{3.2.11}$$

Similarmente, si $c \in \{0,1\}^{\overline{J_1(S)}}$ definimos $\tau_1(S,c)$ a la *traslación de S que mantiene los 1 iniciales y concatena c*. Es decir $\tau_1(S,c) = \big(\tau_1(S,c)_i\big)_{i=1}^d$ con

$$\tau_1(S,c)_i = \begin{cases} s_i, & \text{si } i \in J_1(S), \\ C(s_i)c(i), & \text{si } i \notin J_1(S). \end{cases} \tag{3.2.12}$$

Usando la notación anterior, podemos ver que si S es una secuencia de d palabras tal que no todas comienzan con el mismo carácter (es decir, $0 < |J_0(S)| < d$ y luego $0 < |J_1(S)| < d$). Entonces:

$$w_n[S] \geq \text{máx} \begin{cases} \dfrac{1}{2^{|\overline{J_0}(S)|}} \displaystyle\sum_{c \in \{0,1\}^{\overline{J_0}(S)}} w_{n-|\overline{J_0}(S)|}[\tau_0(S,c)], \\ \dfrac{1}{2^{|\overline{J_1}(S)|}} \displaystyle\sum_{c \in \{0,1\}^{\overline{J_1}(S)}} w_{n-|\overline{J_1}(S)|}[\tau_1(S,c)]. \end{cases} \tag{3.2.13}$$

Intuitivamente, el primer término del máximo anterior representa el largo esperado de la LCS de las palabras que obtendríamos al deshacernos de los caracteres 1 iniciales (buscando con esto que el primer alineamiento sea un 0) y concatenar en aquellas palabras uno de los n caracteres aleatorios. Análogamente, el segundo término representa lo que obtendríamos al buscar que el primer alineamiento sea un 1. Con esto, la desigualdad anterior dice que el largo de la subsecuencia común más grande de las palabras de S (aumentadas con n caracteres aleatorios) es al menos el máximo entre el largo esperado de la LCS de las palabras que obtendríamos al descartar los 1 iniciales y el largo esperado de la LCS de las palabras que obtendríamos al descartar los 0 iniciales.

Para dejar más claro lo anterior, veamos un ejemplo para el caso de $d = 4$ palabras:

$$w_n[001, 011, 101, 001] \geq \text{máx} \begin{cases} \frac{1}{2} & \sum_{c \in \{0,1\}^{\{3\}}} w_{n-1}[001, 011, 01c(3), 001], \\ \frac{1}{2^3} & \sum_{c \in \{0,1\}^{\{1,2,4\}}} w_{n-3}[01c(1), 11c(2), 101, 01c(4)]. \end{cases}$$

En este ejemplo, sólo una palabra, la tercera, no empieza con 0. Luego, el primer término del máximo resulta ser el promedio entre los valores de w_{n-1} en las 2 secuencias posibles de palabras que se obtienen de S al borrar el 1 inicial de la tercera palabra y concatenar un carácter al final. Por otro lado, las otras tres palabras de S empiezan con 0 (es decir, no empiezan con 1). Luego, el segundo término del máximo resulta ser el promedio sobre el valor que toma w_{n-3} en todas las secuencias de palabras que se forman a partir de S al borrar los 0 iniciales y concatenar, en aquellas palabras, un carácter al final.

Notemos ahora que si usamos los lados derechos de las dos desigualdades descritas para w_n en (3.2.10) y (3.2.13) es posible definir una función T tal que para todo $n \geq d$

$$w_n \geq T(w_{n-1}, w_{n-2}, \ldots, w_{n-d}), \tag{3.2.14}$$

donde la desigualdad es componente a componente. Más aún, al igual que en la subsección 3.2.2, podemos descomponer T, esta vez en $d+1$ funciones más simples denotadas T_r, tales que, si $\Pi_r(w)$ es la proyección sobre las componentes que corresponden a las secuencias tales que r de las d palabras comienzan con 0, entonces para todo $n \geq d$

$$\begin{array}{rrcl} & \Pi_0(w_n) & \geq & T_0(w_{n-d}), \\ & \Pi_d(w_n) & \geq & T_d(w_{n-d}), \\ \forall r,\ 0 < r < d & \Pi_r(w_n) & \geq & T_r(w_{n-r}, w_{n-d+r}). \end{array} \tag{3.2.15}$$

Usando los conceptos de traslación que mantiene un carácter que definimos antes, podemos reescribir T de una manera que nos facilitará su programación y análisis. Para $i \in \{0, 1\}$ definimos la transformación lineal $A^i : (\mathbb{R}^{2^{ld}})^d \longmapsto \mathbb{R}^{2^{ld}}$, que toma d vectores de tamaño 2^{ld}, digamos $(v_1, v_2, \ldots v_d)$, y devuelve un vector $A^i(v_1, v_2, \ldots v_d)$, tal que si $S = (s_i)_{i=1}^d$ es una secuencia de d palabras de tamaño l (que representa una de las 2^{ld} secuencias posibles), entonces:

$$A^i(v_1, v_2, \ldots, v_d)[S] = \frac{1}{2^{|\overline{J_i}(S)|}} \sum_{c \in \{0,1\}^{\overline{J_i}(S)}} v_{|\overline{J_i}(S)|}[\tau_i(S, c)]. \tag{3.2.16}$$

Decimos que A^0 es la *parte glotona de T que mantiene los 0* (y elimina los 1), y que A^1 es la *parte glotona de T que mantiene los 1* (y elimina los 0). Además, convengamos como es habitual que la suma sobre un conjunto vacío es 0 (para incluir el caso $|J_i(S)| = 0$). Con esto, si $v_1, v_2, \ldots, v_d$ es una secuencia de d vectores de tamaño 2^{ld}, entonces de las desigualdades (3.2.10) y (3.2.13) se concluye que podemos definir T en función de sus partes glotonas como:

$$T(v_1, v_2, \ldots, v_d) = b + \text{máx}\{A^0(v_1, v_2, \ldots, v_d), A^1(v_1, v_2, \ldots, v_d)\}, \tag{3.2.17}$$

donde b es el vector de $\mathbb{R}^{2^{ld}}$ que vale 1 en las coordenadas correspondientes a las secuencias de d palabras de largo l que comienzan con el mismo carácter (todas con 0 o todas con 1) y 0 en las demás coordenadas.

3.2.3. Encontrando una cota inferior

Al principio de la subsección anterior, vimos que para toda secuencia de palabras $s_1, \ldots, s_d$, se tiene $\gamma_{2,d} = \lim_{n\to\infty} \frac{1}{n} w_{dn}[s_1, \ldots, s_d]$. Un primer intento para encontrar una cota para dicha cantidad podría basarse en la monotonicidad de la transformación T recién definida. Gracias a ella se puede realizar el siguiente procedimiento: x

Lamentablemente en el procedimiento anterior no disponemos de un análisis adecuado para estimar a partir de que valor de n tenemos efectivamente una cota inferior para $\gamma_{2,d}$, (la cantidad $v_{dn}[s_1, \ldots, s_d]/n$ no es siquiera creciente.)

Necesitamos una estrategia distinta. Para esto, usaremos el siguiente resultado, que es una generalización al caso de d palabras de un resultado obtenido por Lueker [6] para el caso bidimensional:

Lema 3.4. *Sea $T : (\mathbb{R}^{2^{ld}})^d \longmapsto \mathbb{R}^{2^{ld}}$ una transformación que cumple con las siguientes propiedades:*

1. ***Monotonía**:*

$$(v_1, v_2, \ldots, v_d) \leq (w_1, w_2, \ldots, w_d) \Longrightarrow T(v_1, v_2, \ldots, v_d) \leq T(w_1, w_2, \ldots, w_d). \tag{3.2.18}$$

2. ***Invarianza bajo traslación**:*
   Si $r \in \mathbb{R}$, $\mathbb{1}$ es el vector de unos en $\mathbb{R}^{2^{ld}}$, y $\vec{\mathbb{1}} = (\mathbb{1}, \mathbb{1}, \ldots, \mathbb{1})$ es el vector de unos en $(\mathbb{R}^{2^{ld}})^d$, entonces para todo vector $(v_1, v_2, \ldots, v_d) \in (\mathbb{R}^{2^{ld}})^d$:

$$T((v_1, v_2, \ldots, v_d) + r\vec{\mathbb{1}}) = T(v_1, v_2, \ldots, v_d) + r\mathbb{1}. \tag{3.2.19}$$

3. ***Factibilidad**: Existe un trío (z, r, ε) (que denotaremos trío factible) con $z \in \mathbb{R}^{2^{ld}}$, r un real y $\varepsilon \geq 0$ tal que:*

$$T(z + (d-1)r\mathbb{1}, \ldots, z + 2r\mathbb{1}, z + r\mathbb{1}, z) \geq z + dr\mathbb{1} - \varepsilon\mathbb{1}. \tag{3.2.20}$$

Entonces, para toda secuencia $(v_n)_{n\in\mathbb{N}}$ de vectores de $\mathbb{R}^{2^{ld}}$ tales que para todo $n \geq d$ se satisface:

$$v_n \geq T(v_{n-1}, v_{n-2}, \ldots, v_{n-d}). \tag{3.2.21}$$

se tiene que existe un vector z_0 tal que para todo $n \geq 0$,

$$v_n \geq z_0 + n(r - \varepsilon)\mathbb{1}. \tag{3.2.22}$$

Demostración. Sea T una transformación monótona, invariante bajo traslación y (z, r, ε) un trío factible para T. Sea v_n una secuencia como en la hipótesis y α cualquier real suficientemente grande tal que para todo j menor o igual que $d - 1$

$$v_j + \alpha\mathbb{1} \geq z + j(r - \varepsilon)\mathbb{1}.$$

Por ejemplo si consideramos el vector $\max_{0\leq j\leq(d-1)}(z+j(r-\varepsilon)\mathbb{1}-v_j)$, podemos elegir α como el valor más grande de dicho vector.

Con esto $z_0 = z-\alpha\mathbb{1}$ cumple (3.2.22) para todo $n \leq (d-1)$. Probemos por inducción que esto se extiende a todo $n \geq d$. Supongamos que (3.2.22) se cumple hasta $n-1$, veamos que se cumple para n. Por hipótesis de inducción:

$$\begin{aligned}(v_{n-1},\ldots,v_{n-d}) &\geq (z_0+(n-1)(r-\varepsilon)\mathbb{1},\ldots,z_0+(n-j)(r-\varepsilon)\mathbb{1},\ldots,z_0+(n-d)(r-\varepsilon)\mathbb{1})\\ &= (z+(d-1)r\mathbb{1},\ldots,z+(d-j)r\mathbb{1}+(j-1)\varepsilon\mathbb{1},\ldots,z+(d-1)\varepsilon\mathbb{1})\\ &\quad +((n-d)(r-\varepsilon)-(d-1)\varepsilon-\alpha)\,\vec{\mathbb{1}}\\ &\geq (z+(d-1)r\mathbb{1},\ldots,z+(d-j)r\mathbb{1},\ldots,z)\\ &\quad +((n-d)(r-\varepsilon)-(d-1)\varepsilon-\alpha)\,\vec{\mathbb{1}}.\end{aligned}$$

Aplicando T a ambos lados de la desigualdad, se obtiene, por monotonía e invarianza bajo traslación:

$$\begin{aligned}v_n \geq T(v_{n-1},v_{n-2},\ldots,v_{n-d}) &\geq T(z+(d-1)r\mathbb{1},\ldots,z+(d-j)r\mathbb{1},\ldots,z)\\ &\quad +((n-d)(r-\varepsilon)-(d-1)\varepsilon-\alpha)\,\mathbb{1}.\end{aligned}$$

Usando que (z,r,ε) es un trío factible de T se tiene que

$$\begin{aligned}v_n &\geq z+(dr-\varepsilon)\mathbb{1}+((n-d)(r-\varepsilon)-(d-1)\varepsilon-\alpha)\mathbb{1}\\ &= z-\alpha\mathbb{1}+n(r-\varepsilon)\mathbb{1}\\ &= z_0+n(r-\varepsilon)\mathbb{1}.\end{aligned}$$

Lo que concluye la demostración del lema. ■

De (3.2.16) y (3.2.17) se concluye fácilmente que la transformación T definida en la subsección anterior es monótona e invariante bajo traslación. Luego, si probamos la existencia de un trío (z,r,ε) factible para T podremos concluir, gracias al lema, que la sucesión de vectores $(w_n)_{n\in\mathbb{N}}$ satisface $w_n \geq z_0+n(r-\varepsilon)\mathbb{1}$ para todo n y luego, gracias a (3.2.9) deducir que:

$$\gamma_{2,d} \geq d(r-\varepsilon). \tag{3.2.23}$$

Sigue que, para encontrar buenas cotas inferiores para $\gamma_{2,d}$ basta encontrar buenos tríos factibles, en particular algún trío que tenga r grande y ε lo más pequeño posible. Una forma de conseguir esto es modificar un poco la estrategia explicada al principio de esta subsección. De hecho, al igual que en dicho procedimiento, basta iterar la transformación T partiendo de d vectores iniciales cualquiera $v_0,v_1,\ldots,v_{d-1}$ calculando $v_n = T(v_{n-1},v_{n-2},\ldots,v_{n-d})$ para $n \geq d$.

La transformación T tiene buenas propiedades y, al menos empíricamente, se ve que los v_n definidos como antes son tales que v_n/n converge a un vector con todas sus coordenadas iguales. Luego, si se toma un n suficientemente grande, los vectores v_n y v_{n-1} diferirán prácticamente en una constante por el vector $\mathbb{1}$. Es decir, existe una constante r tal que $v_n - v_{n-1} \sim r\mathbb{1}$ para todo n

grande. Notemos ahora que $T(v_{n+d-1}, \ldots, v_{n+1}, v_n) = v_{n+d}$ por definición, luego, de la observación anterior y la continuidad de T, se tendrá que

$$T(v_n + (d-1)r\mathbb{1}, v_n + (d-2)r\mathbb{1}, \ldots, v_n + r\mathbb{1}, v_n) \sim v_n + dr\mathbb{1}.$$

Sigue que para encontrar un trío factible podemos tomar un n suficientemente grande, de modo que la diferencia entre v_n y v_{n-1} sea un vector con todas sus coordenadas practicamente iguales, definir z como v_n, tomar r como el máximo valor tal que $v_n - v_{n-1} \geq r\mathbb{1}$, y luego definir ε lo más pequeño posible de modo que el trío (z, r, ε) resulte factible. Con esto, gracias a (3.2.23), se obtiene una (buena) cota para $\gamma_{2,d}$.

Es importante recalcar que no necesitamos probar la propiedad de convergencia de v_n/n para obtener una cota para $\gamma_{2,d}$. Sólo basta exhibir un buen trío factible como testigo en virtud del lema anterior.

3.2.4. Implementación y cotas obtenidas

A continuación describiremos el procedimiento usado para obtener un trío factible (z, r, ε) para T y, en conclusión, una cota inferior para $\gamma_{2,d}$. Este procedimiento tiene como entradas el número de palabras a usar d, el largo de cada palabra, l y una tolerancia δ que regula el momento en el que terminamos el procedimiento:

trioFactible(d, l, δ)
Definir para $i = 0, \ldots, d-1$, el vector v_i como el vector de ceros en $\mathbb{R}^{2^{ld}}$.
Iterar desde $i = d$:
Definir $v_i \leftarrow T(v_{i-1}, v_{i-2}, \ldots, v_{i-d})$.
Definir $\Delta \leftarrow v_i - v_{i-1}$. $M \leftarrow \max_j \Delta[j]$. $m \leftarrow \min_j \Delta[j]$.
Si $M - m < \delta$, salir de la iteración.
Si no, actualizar $i \leftarrow i + 1$.
Definir $z \leftarrow v_i$, $r \leftarrow M$, $W \leftarrow z + dr\mathbb{1} - T(z + (d-1)r\mathbb{1}, \ldots, z + r\mathbb{1}, z)$, $\varepsilon \leftarrow \max_j W[j]$.
Retornar (z, r, ε).

cotaInferior(d, l, δ)
Definir $(z, r, \varepsilon) \leftarrow$**trioFactible**(d, l, δ)
Retornar $d(r - \varepsilon)$.

Usando la definición de T dada por sus partes glotonas (3.2.17) es sencillo escribir una implementación del procedimiento anterior en cualquier lenguaje de programación. Aprovechamos la linealidad de A^0 y A^1 para almacenarlas como un conjunto de matrices ralas, lo cual acelera mucho cada iteración. A medida que aumentamos l, como es de esperarse, la cota va mejorando, sin embargo la memoria usada crece demasiado (recordar que cada vector es de tamaño 2^{ld}). De hecho, un análisis simple de la definición de A^0 y A^1 permite concluir que sólo para almacenar ambas matrices de una manera rala es necesario guardar al menos $2^{ld}2^d = 2^{(d+1)l}$ valores distintos de 0. Por dicha razón sólo es posible realizar este procedimiento para valores pequeños de d y l.

Los resultados obtenidos se muestran en la Tabla 3.1.

Valores para $d = 2$

l	Cota para $\gamma_{2,2}$
1	0.66666666418314
2	0.72727271808253
3	0.74792242549591
4	0.75857669196192
5	0.76544695856955
6	0.77027385445717
7	0.77397507816660
8	0.77686064534113
9	0.77925933824914

Valores para $d = 3$

l	Cota para $\gamma_{2,3}$
1	0.66666665424904
2	0.67391302685782
3	0.68741051941928
4	0.69295022383657
5	0.69773770403703
6	0.70131710654988
7	0.70447344812276

Valores para $d = 4$

l	Cota para $\gamma_{2,4}$
1	0.615384595264409
2	0.643216010100026
3	0.651309094992399
4	0.657241281735180
5	0.661274795991150

Valores para $d = 5$

l	Cota para $\gamma_{2,5}$
1	0.615384368494691
2	0.626505887219029
3	0.632165355051306

Valores para $d = 6$

l	Cota para $\gamma_{2,6}$
1	0.592592356497814
2	0.610924985301239
3	0.617761437978714

Valores para $d = 7$

l	Cota para $\gamma_{2,7}$
1	0.592592481860063
2	0.602493036014458

Tabla 3.1: Cotas inferiores para $\gamma_{2,d}$

Los resultados se obtuvieron con nuestra implementación usando $\delta = 10^{-8}$. Es importante notar que el método converge rápidamente al resultado y de hecho, en ningún caso fueron necesarias más de 150 iteraciones. La parte más lenta del algoritmo corresponde al cálculo de las matrices ralas.

3.2.5. Extensiones

Es relativamente simple extender los resultados anteriores a alfabetos de cualquier tamaño. Veamos como hacer esto. Sea Σ un alfabeto finito cualquiera. Definamos $W_n^d(\cdot)$ y w_n de manera análoga al caso de alfabeto binario, con la salvedad que ahora la colección de palabras aleatorias $(X_i)_{i=1}^d$ es tomada sobre Σ y los vectores w_n tienen $|\Sigma|^{ld}$ coordenadas.

Al igual que antes w_n estará acotada inferiormente por una función T de los últimos d vectores. La idea es escribir T en función de sus *partes glotonas*. La definición será similar para ello extendamos la definición de traslaciones que *mantengan* ciertos caracteres (y reemplacen los demás).

Para una palabra s sobre Σ de largo l definimos, al igual que antes $I(s)$ como el primer carácter de s y $C(s)$ como su cola, es decir, la subpalabra que resulta al eliminar dicho carácter. Además, para todo $\sigma \in \Sigma$ y toda secuencia $S = (s_i)_{i=1}^d$ de d palabras de largo l sobre Σ, definimos $J_\sigma(S)$ como el conjunto de índices j tales que la palabra s_j comienza con σ, es decir $I(s_j) = \sigma$. Definamos además $\overline{J_\sigma}(S) = [d] \setminus J_\sigma(S)$ los índices de las palabras que no comienzan con σ.

Para $\sigma \in \Sigma$ y $c \in \Sigma^{\overline{J_\sigma}(S)}$ una asignación definimos la *traslación de S que mantiene los caracteres*

iniciales σ y cambia los demás por c como $\tau_\sigma(S.c) = \left(\tau_\sigma(S,c)_i\right)_{i=1}^{d}$ con:

$$\tau_\sigma(S,c)_i = \begin{cases} s_i, & \text{si } i \in J_\sigma(S), \\ C(s_i)c(i), & \text{si } i \notin J_\sigma(S). \end{cases}$$

Con esto, para cada $\sigma \in \Sigma$, definimos la transformación lineal $A^\sigma : \left(\mathbb{R}^{|\Sigma|^{ld}}\right)^d \mapsto \mathbb{R}^{|\Sigma|^{ld}}$, que toma d vectores de tamaño $|\Sigma|^{ld}$, digamos $(v_1, v_2, \dots v_d)$, y devuelve $A^\sigma(v_1, v_2, \dots, v_d)$, tal que si $S = (s_1, s_2, \dots, s_d)$ es una secuencia de d palabras de tamaño l (que representa una de las $|\Sigma|^{ld}$ secuencias posibles), entonces:

$$A^\sigma(v_1, v_2, \dots, v_d)[S] = \frac{1}{|\Sigma|^{|\overline{J_\sigma}(S)|}} \sum_{c \in \Sigma^{\overline{J_\sigma}(S)}} v_{|\overline{J_\sigma}(S)|}\,[\tau_\sigma(S,c)].$$

Llamamos, haciendo una analogía al caso de dos palabras, *parte glotona de T que mantiene σ* a la transformación A^σ. Notemos que esta definición es consistente con la definición de A^0 y A^1 realizada para el caso de $\Sigma = \{0,1\}$ en la subsección 3.2.2.

Finalmente definamos T como la transformación que a una secuencia $(v_1, v_2, \dots v_d)$ de d vectores de $\mathbb{R}^{|\Sigma|^{ld}}$ le asocia

$$T(v_1, v_2, \dots v_d) = b + \max_{\sigma \in \Sigma} B^\sigma(v_1, v_2, \dots v_d),$$

con b el vector de $\mathbb{R}^{|\Sigma|^{ld}}$ que vale 1 en las coordenadas correspondientes a secuencias de d palabras de largo l que comienzan con el mismo carácter y 0 en las demás. Con esto, para todo $n \geq d$,

$$w_n \geq T(w_{n-1}, w_{n-2}, \dots, w_{n-d}).$$

Sigue que, aplicando una versión del Lema 3.4, se puede obtener una cota inferior para $\gamma_{|\Sigma|,d}$ del mismo modo que antes.

3.3. Aplicación: Conjetura de Steele

Steele [8] conjeturó en 1986 que la constante $\gamma_{2,3}$ asociada a la LCS de 3 palabras sobre un alfabeto binario se podía obtener a partir de la constante $\gamma_{2,2}$ asociada a la LCS de sólo 2 palabras mediante la siguiente relación: $\gamma_{2,3} = (\gamma_{2,2})^2$. De hecho, él hace una generalización de su conjetura, diciendo que $\gamma_{2,d} = (\gamma_{2,2})^{d-1}$. Dančík [3] lo cita de manera incorrecta, afirmando que la conjetura planteada por Steele es para alfabetos de cualquier tamaño. En otras palabras, Dančík argumenta que la conjetura de Steele es $\gamma_{k,d} = (\gamma_{k,2})^{d-1}$ para todo k y luego prueba que tal conjetura es falsa al observar que [3]:

$$1 \leq \liminf_{k\to\infty} k^{1-1/d}\gamma_{k,d} \leq \limsup_{k\to\infty} k^{1-1/d}\gamma_{k,d} \leq e. \tag{3.3.1}$$

Sin embargo, esto sólo prueba la falsedad de la conjetura extendida, pues los resultados usados son ciertos asintóticamente en k. Hasta ahora, no existen referencias a la veracidad o falsedad de la conjetura original de Steele (que se refiere sólo a alfabetos binarios).

Daremos un argumento simple para probar que la conjetura de Steele general ($\gamma_{2,d} = (\gamma_{2,2})^{d-1}$) es falsa y que de hecho no es cierta para ningún $d \geq 3$.

Lueker [6] probó las siguientes cotas para $\gamma_{2,2}$:

$$L = 0,788071 \leq \gamma_{2,2} \leq 0{,}826280 = U.$$

Usando la cota inferior simple (Corolario 3.3) mostrada al principio del capítulo, tenemos que $1/2 \leq \gamma_{2,d}$ para todo d. Luego, si la conjetura de Steele fuera cierta se tendría que

$$1/2 \leq \gamma_{2,d} = (\gamma_{2,2})^{d-1} \leq U^{d-1},$$

lo cual es falso para todo $d \geq 1 + \ln(1/2)/\ln U \approx 4{,}6$. Con esto hemos probado que la conjetura general es falsa para todo $d \geq 5$. Para ver el caso $d = 3$ (que es precisamente el caso original de la conjetura de Steele) y el caso $d = 4$ basta observar los resultados numéricos obtenidos para las cotas inferiores de $\gamma_{k,d}$ en la Tabla 3.2.

d	Cota inferior para $\gamma_{2,d}$	U^{d-1} (Cota superior asumiendo conjetura de Steele)
3	0.7044734481	0.6827386384
4	0.6612747959	0.5641332822
5	0.6321653550	0.4661320484
6	0.6177614379	0.3182463600
7	0.6024930360	0.2629606023

Tabla 3.2: Tabla con las mejores cotas obtenidas para distintos d. $U = 0{,}826280$ es la mejor cota superior conocida para $\gamma_{2,2}$.

Capítulo 4

Problema del subhipergrafo monótono de tamaño máximo

Sankoff y Mainville [12] conjeturaron en los ochenta que $\lim_{k\to\infty} \gamma_{k,2}\sqrt{k} = 2$. Kiwi, Loebl y Matoušek [7] probaron recientemente la veracidad de la conjetura anterior relacionando el problema de la subsecuencia común más grande con el problema de Ulam o de la secuencia creciente más grande de una permutación.

Dančík [3] señala que una extensión natural a la conjetura de Sankoff y Mainville al caso de d palabras sería $\lim_{k\to\infty} \gamma_{k,d}k^{1-1/d} = 2$. Sin embargo esto no parece acertado pues en el argumento usado para probar la veracidad de la conjetura original, se usa explicitamente que 2 es el valor de la constante c_2 correspondiente al problema de la secuencia creciente más larga, también conocido como problema de Ulam en dos dimensiones.

Por lo anterior, es razonable pensar que la buena extensión para la conjetura de Sankoff y Mainville sería $\lim_{k\to\infty} \gamma_{k,d}k^{1-1/d} = c_d$, donde c_d es la constante para el problema de Ulam en d dimensiones. Esto último será probado en este capítulo como corolario del estudio de un nuevo problema que realizaremos en las siguientes secciones.

Específicamente, para el problema de la subsecuencia común más grande, probaremos el siguiente teorema.

Teorema 4.1. *Para todo d entero positivo y todo $\varepsilon > 0$ existen k_0 y A suficientemente grandes tal que para todo $k \geq k_0$ y todo $n \geq Ak^{1-1/d}$ se tiene:*

$$(1-\varepsilon)\frac{c_d n}{k^{1-1/d}} \leq \mathbb{E}\big[\mathcal{L}_{n,k}^{(d)}\big] \leq (1+\varepsilon)\frac{c_d n}{k^{1-1/d}}.$$

Por otro lado, si $\mathrm{Med}\big[\mathcal{L}_{n,k}^{(d)}\big]$ *es una mediana de* $\mathcal{L}_{n,k}^{(d)}$,

$$(1-\varepsilon)\frac{c_d n}{k^{1-1/d}} \leq \mathrm{Med}\big[\mathcal{L}_{n,k}^{(d)}\big] \leq (1+\varepsilon)\frac{c_d n}{k^{1-1/d}}.$$

Además, existe una constante absoluta $C > 0$ *tal que, para* k *y* n *como antes, se tiene la siguiente cota exponencial para la cola de la distribución,*

$$\mathbb{P}\left[\mathcal{L}_{n,k}^{(d)} \leq (1-\varepsilon)\frac{c_d n}{k^{1-1/d}}\right] \leq \exp\left(-C\varepsilon^2 \cdot \frac{c_d n}{k^{1-1/d}}\right),$$
$$\mathbb{P}\left[\mathcal{L}_{n,k}^{(d)} \geq (1+\varepsilon) \cdot \frac{c_d n}{k^{1-1/d}}\right] \leq 2\exp\left(-\frac{C\varepsilon^2}{1+\varepsilon} \cdot \frac{c_d n}{k^{1-1/d}}\right).$$

Corolario 4.2. *Para todo* d *entero positivo,*

$$\lim_{k\to\infty} \gamma_{k,d} k^{1-1/d} = c_d.$$

Este resultado se obtendrá en este capítulo como corolario de un teorema más general que abarca no sólo estimar el comportamiento de la distribución de $\mathcal{L}_{n,k}^{(d)}$ sino que de varias distribuciones similares.

4.1. Subhipergrafo monótono de tamaño máximo

Replantearemos el problema de la LCS de varias palabras, siguiendo el enfoque usado en [7] para el caso de 2 palabras. En dicho trabajo la noción de subsecuencia común de dos palabras fue analogada a la noción de un cierto tipo de subgrafos (denominados subgrafos planares) de un grafo bipartito cuyos vértices representan los caracteres de las palabras. Para el caso general de d palabras, la extensión natural requiere que trabajemos con hipergrafos en vez de grafos.

Definición. Dados d conjuntos disjuntos, $A_1, A_2, \ldots, A_d$, diremos que un hipergrafo $H = (V, E)$ es d-partito de clases $A_1, A_2, \ldots A_d$ y d-uniforme (o simplemente d-hipergrafo) si se cumplen las siguientes condiciones:

1. $V = \bigdot\bigcup_{j=1}^{d} A_j$.
2. $E \subseteq \{\{v^{(1)}, v^{(2)}, \ldots, v^{(d)}\} \mid v^{(j)} \in A_j, j = 1, \ldots, d\}$.

Sin pérdida de generalidad, identificaremos E con el subconjunto de $A_1 \times A_2 \times \cdots \times A_d$ que contiene las d-tuplas ordenadas asociadas a las aristas de E. Además, usaremos la notación habitual $E(H) = E$ y $V(H) = V$.

En general, de aquí en adelante, llamaremos $n_j = |A_j|$ al tamaño del j-ésimo conjunto y supondremos además que los elementos de A_j están dotados de un orden total $\leq$ (supondremos, de hecho, abusando un poco de notación, que los vértices están numerados $1, 2, \ldots, n_j$).

Denotaremos $K_{n_1,n_2,\ldots,n_d}$ al d-hipergrafo completo sobre las clases anteriores, es decir al hipergrafo tal que $E(K_{n_1,n_2,\ldots,n_d}) = A_1 \times A_2 \times \cdots \times A_d$. En el caso que todas las clases tengan el mismo cardinal, digamos $n_1 = n_2 = \cdots = n_d = n$, lo denotaremos simplemente $K_n^{(d)}$.

En $A_1 \times A_2 \times \cdots \times A_d$ consideramos la relación de orden parcial natural $\leq$ definida por:

$$\left(v^{(1)}, v^{(2)}, \ldots, v^{(d)}\right) \leq \left(w^{(1)}, w^{(2)}, \ldots, w^{(d)}\right) \Longleftrightarrow v^{(j)} \leq w^{(j)}, \text{ para todo } 1 \leq j \leq d.$$

Diremos que un conjunto de aristas de un d-hipergrafo es un *emparejamiento monótono* si todo par de aristas distintas $\{e, f\}$ es comparable vía este orden, es decir si $e \leq f$ ó $f \leq e$. Si un d-hipergrafo H es tal que $E(H)$ es un *emparejamiento monótono*, diremos simplemente que H es un hipergrafo monótono. Además, denotaremos $L(H)$ al número de aristas de un subhipergrafo monótono de H de tamaño máximo.

Llamamos *modelo de d-hipergrafo aleatorio* a toda familia de distribuciones $\mathcal{D} = (\mathcal{D}(K_{n_1,\ldots,n_d}))$ donde cada $\mathcal{D}(K_{n_1,\ldots,n_d})$ es una distribución de probabilidad sobre los subhipergrafos de $K_{n_1,\ldots,n_d}$ que cumple las siguientes propiedades:

1. **[Monotonicidad]** Si H se distribuye de acuerdo a $\mathcal{D}(K_{n_1,\ldots,n_d})$, y H' es un subhipergrafo inducido de H tal que $V(H')$ contiene exactamente n'_j vértices de la clase A_j, entonces H' se distribuye de acuerdo a $\mathcal{D}(K_{n'_1,\ldots,n'_d})$.
2. **[Independencia por bloques]** Si H_1 y H_2 son dos subhipergrafos inducidos de H disjuntos entonces las distribuciones inducidas por H_1 y H_2 son independientes (y luego, $L(H_1)$ y $L(H_2)$ también lo son).

Los modelos de hipergrafo más simples (y en los que enfocaremos nuestra atención) son los siguientes:

1. Denotemos $\Sigma(K_{n_1,\ldots,n_d}, k)$ a la distribución sobre todos los subhipergrafos de $K_{n_1,\ldots,n_d}$ obtenidos al asignar a cada uno de sus vértices un símbolo de $\{1, \ldots, k\}$ de manera uniforme e independiente y quedarse sólo con las aristas cuyos elementos tengan asignado el mismo símbolo. Llamaremos al modelo $\Sigma(\cdot, k)$, *modelo de d palabras aleatorias sobre un alfabeto de tamaño k*.

   El nombre de este modelo proviene del hecho que si H es elegido de acuerdo a $\Sigma(K_{n_1,\ldots,n_d}, k)$ entonces $L(H)$ es precisamente el largo de la LCS de las d palabras que se pueden "leer" en los símbolos de las clases en que está particionado H, y luego $L(\Sigma(K_n^{(d)}, k))$ tiene la misma distribución que $\mathcal{L}_{n,k}^{(d)}$.

2. Sea $G(K_{n_1,\ldots,n_d}, p)$ la distribución sobre todos los subhipergrafos de $K_{n_1,\ldots,n_d}$ donde la probabilidad de que una arista e esté en el subhipergrafo es p, y estos eventos son independientes. Llamaremos al modelo $G(\cdot, p)$ *modelo binomial, de parámetro p, de d-hipergrafos.*

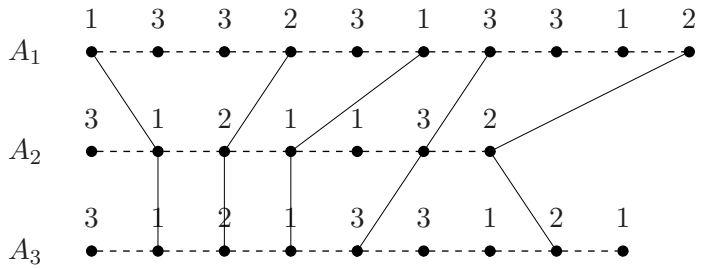

Figura 4.1: Representación esquemática de un subhipergrafo monótono de tamaño máximo de H, donde H es elegido de acuerdo a $\Sigma(K_{10,7,9}, 3)$. Las aristas están representadas por líneas poligonales que pasan por sus vértices. Sobre cada vértice se indica el símbolo elegido en la realización de H y, en este caso, $L(H) = 5$. Intuitivamente la monotonicidad es equivalente a que las aristas no se crucen ni compartan vértices

Finalmente, dado un modelo, llamaremos *parámetro interno* del modelo a cualquier valor que sea constante para todas las distribuciones que participan del modelo. Por ejemplo, k y p (o cualquier función de ellos) son parámetros internos de los modelos $\Sigma(\cdot, k)$ y $G(\cdot, p)$ respectivamente.

En este capítulo probaremos un teorema que permitirá estimar el valor esperado de $L(H)$ así como el valor de sus medianas, cuando H es elegido de acuerdo a algún modelo de hipergrafo aleatorio con ciertas condiciones especiales (en particular, los modelos anteriormente descritos). Usando que $L(\Sigma(K_n^{(d)}, k))$ se distribuye como $\mathcal{L}_{n,k}^{(d)}$, obtendremos como corolario el Teorema 4.1 enunciado al comienzo de este capítulo.

4.2. Aproximación de la mediana. Teorema Principal

En los modelos que estudiaremos, la siguiente desigualdad de Talagrand será de vital importancia:

Lema 4.3 (Desigualdad de Talagrand)**.** *Supongamos que $Z_1, \ldots, Z_N$ son variables aleatorias independientes que toman valores en un cierto conjunto Λ. Sea $X = f(Z_1, \ldots, Z_N)$, con $f : \Lambda^N \to \mathbb{R}$ una función tal que las siguientes condiciones se cumplen para un cierto l_0 y una cierta función Φ.*

1. *(f es Lipschitz) Si $z, z' \in \Lambda^N$ difieren solo en una coordenada, entonces $|f(z) - f(z')| \leq l_0$.*
2. *(Existencia de testigos) Si $z \in \Lambda^N$ y $r \in \mathbb{R}$ tal que $f(z) \geq r$, entonces existe un conjunto de índices $J \subseteq \{1, \ldots, N\}$, $|J| \leq \Phi(r)/l^2$ y un testigo $(w_j : j \in J)$ tal que para todo $y \in \Lambda^N$ con $y_j = z_j$ cuando $j \in J$, se tiene $f(y) \geq r$.*

Sea $\mathrm{Med}(X)$ *una mediana de* X*. Entonces, para todo* $t \geq 0$*, se cumplen las siguientes desigualdades:*

$$\mathbb{P}[X \leq \mathrm{Med}(X) - t] \leq 2e^{-t^2/4\Phi(\mathrm{Med}(X))},$$
$$\mathbb{P}[X \geq \mathrm{Med}(X) + t] \leq 2e^{-t^2/4\Phi(\mathrm{Med}(X)+t)}.$$

Notemos que para el modelo de d palabras aleatorias, el valor de $L(H)$ depende exclusivamente de los símbolos asignados a los vértices (que son independientes), y además si cambiamos uno de los símbolos, entonces el valor de $L(H)$ cambia a lo más en 1. Además, si $L(H) \geq r$, entonces existen dr vértices testigos (los extremos de las r aristas) que garantizan que para cualquier asociación de caracteres que preserve la asociación correspondiente a los vértices testigos, el hipergrafo resultante posee un subhipergrafo monótono de al menos r aristas. Luego por Talagrand, si Med es una mediana de $L(H)$, entonces para todo $s \geq 0$,

$$\mathbb{P}[L(H) \leq (1-s)\,\mathrm{Med}] \leq 2\exp\left(-\frac{s^2}{4d}\,\mathrm{Med}\right),$$
$$\mathbb{P}[L(H) \geq (1+s)\,\mathrm{Med}] \leq 2\exp\left(-\frac{s^2}{4d(1+s)}\,\mathrm{Med}\right).$$

Similarmente, para el modelo binomial, el valor de $L(H)$ depende de la existencia o no de las aristas (que son independientes), con lo cual al igual que antes el valor de $L(H)$ es 1-Lipschitz. Además, si $L(H) \geq r$, entonces existen r aristas testigos que garantizan que $L(H) \geq r$, para cualquier hipergrafo H que contenga dichas aristas. luego, si Med es una mediana de $L(H)$, entonces para todo $s \geq 0$,

$$\mathbb{P}[L(H) \leq (1-s)\,\mathrm{Med}] \leq 2\exp\left(-\frac{s^2}{4}\,\mathrm{Med}\right),$$
$$\mathbb{P}[L(H) \geq (1+s)\,\mathrm{Med}] \leq 2\exp\left(-\frac{s^2}{4(1+s)}\,\mathrm{Med}\right).$$

Este tipo de cotas aparecerá en todos los modelos que estudiaremos, en virtud de esto, la siguiente definición nos será útil.

Definición. Diremos que un modelo $\mathcal{D}$ tiene una *constante de concentración* h si para toda mediana Med de $L(H)$ se cumple que para todo $s \geq 0$,

$$\mathbb{P}[L(H) \leq (1-s)\,\mathrm{Med}] \leq 2\exp\left(-hs^2\,\mathrm{Med}\right),$$
$$\mathbb{P}[L(H) \geq (1+s)\,\mathrm{Med}] \leq 2\exp\left(-h\frac{s^2}{(1+s)}\,\mathrm{Med}\right).$$

Si podemos estimar una mediana de $L(H)$ para los modelos a estudiar, y logramos probar que la media y la mediana están cerca, entonces gracias a la desigualdad anterior, tendremos una buena cota de concentración con respecto a la media. Lamentablemente estimar una mediana no es fácil

en este contexto, pero sí podremos hacerlo cuando los tamaños n_j de las clases, satisfagan ciertas condiciones. En particular, veremos que cuando los tamaños de las clases se encuentran en cierto rango, entonces podremos encontrar una aproximación de la mediana que resultará ser proporcional al promedio geométrico de los tamaños de las clases del hipergrafo.

Definición ((c, λ, θ)-mediana)**.** Diremos que un modelo de d-hipergrafo $\mathcal{D}$ con parámetro interno t admite una (c, λ, θ)-*aproximación de la mediana* (o simplemente, una (c, λ, θ)-mediana) si para todo $\delta > 0$ existen constantes $a(\delta)$, $b(\delta)$ y $t'(\delta)$ suficientemente grandes tales que para todo $t \geq t'$ y todo conjunto de valores $n_1, \ldots, n_d$ de promedio geométrico N y suma S, que cumplan

$$N \geq t^{\lambda} a, \qquad \text{(Cota inferior en el tamaño)}$$
$$Sb \leq t^{\theta}. \qquad \text{(Cota superior en el tamaño)}$$

se tiene que si $\text{Med}[L(\mathcal{D}(K_{n_1,\ldots,n_d}))]$ es una mediana de $L(\mathcal{D}(K_{n_1,\ldots,n_d}))$,

$$(1-\delta)\frac{cN}{t^{\lambda}} \leq \text{Med}[L(\mathcal{D}(K_{n_1,\ldots,n_d}, t))] \leq (1+\delta)\frac{cN}{t^{\lambda}}.$$

En otras palabras, si $\mathcal{D}$ es un modelo de parámetro interno t que admite una (c, λ, θ)-mediana, y además los valores $n_1, \ldots, n_d$ de promedio geométrico N y suma S son tales que $N = \Omega(t^{\lambda})$ y $S = O(t^{\theta})$, entonces, para t grande, toda mediana de $L(\mathcal{D}(K_{n_1,\ldots,n_d}))$ estará cerca de $cNt^{-\lambda}$. Cabe recalcar que la definición anterior depende explícitamente del parámetro interno t elegido para $\mathcal{D}$.

La definición anterior podrá parecer en estos momentos algo artificial. Sin embargo, más adelante veremos que no es difícil obtener aproximaciones para la mediana de este tipo en nuestros modelos. Esto se logrará relacionando el problema de determinar el tamaño de un subhipergrafo monótono de tamaño máximo con otro problema bastante estudiado en la literatura, el problema de la secuencia creciente más larga. Todo esto se verá en las secciones 4.5 y 4.6.

Volvamos a la definición anterior. La importancia de esta noción es que, cuando el modelo a estudiar admite además una constante de concentración, nos permitirá encontrar una cota de concentración en torno a la aproximación de la mediana.

Gracias a las propiedades de independencia por bloques y monotonicidad de estos modelos demostraremos además que podemos extender dicha cota de concentración a un caso mucho más general: esencialmente reemplazaremos la condición $Sb \leq t^{\theta}$ dada por la definición de (c, λ, θ)-mediana por una relación de "balanceo" de los tamaños n_j de las clases del hipergrafo, pidiendo que su suma S no sea demasiado grande con respecto a su promedio geométrico N. Además, en dicho caso, probaremos que la mediana y la media no están muy lejos y luego la cota anterior se convierte además en una estimación de la media. El siguiente teorema es el teorema principal de este capítulo.

Teorema 4.4 (Teorema Principal)**.** *Sea $\mathcal{D}$ un modelo de d-hipergrafo con parámetro interno t, con constante de concentración h y que admite una (c, λ, θ)-mediana. Sea también $g : \mathbb{R} \to \mathbb{R}$ una función tal que $g(t) = O(t^\eta)$ para un cierto $0 \leq \eta < \min\{\lambda/(d-1), \theta - \lambda\}$.*

Para todo $\varepsilon > 0$, existen t_0 y A suficientemente grandes tales que si $t \geq t_0$ y los valores $n_1, \ldots, n_d$ de promedio geométrico N y suma S cumplen:

$$N \geq t^\lambda A, \qquad \text{(Condición de tamaño)}$$
$$S \leq g(t)N, \qquad \text{(Condición de balanceo)}$$

se tiene, definiendo $M = cN/t^\lambda$,

$$(1-\varepsilon)M \leq \mathbb{E}[L(\mathcal{D}(K_{n_1,\ldots,n_d}))] \leq (1+\varepsilon)M. \tag{4.2.1}$$

Por otro lado, si $\mathrm{Med}[L(\mathcal{D}(K_{n_1,\ldots,n_d}))]$ *es una mediana de $L(\mathcal{D}(K_{n_1,\ldots,n_d}))$,*

$$(1-\varepsilon)M \leq \mathrm{Med}[L(\mathcal{D}(K_{n_1,\ldots,n_d}))] \leq (1+\varepsilon)M. \tag{4.2.2}$$

Además, existe una constante absoluta K, tal que para t y $n_1, \ldots, n_d$ como antes, se tiene la siguiente cota exponencial para la cola de la distribución:

$$\mathbb{P}\left[L(\mathcal{D}(K_{n_1,\ldots,n_d})) \leq (1-\varepsilon)M\right] \leq \exp\left(-Kh\varepsilon^2 M\right), \tag{4.2.3}$$
$$\mathbb{P}\left[L(\mathcal{D}(K_{n_1,\ldots,n_d})) \geq (1+\varepsilon)M\right] \leq \exp\left(-Kh\frac{\varepsilon^2}{1+\varepsilon}M\right). \tag{4.2.4}$$

4.3. Demostración de las cotas inferiores en el Teorema Principal

En esta sección probaremos las cotas inferiores dadas por el teorema, es decir, probaremos la cota inferior en (4.2.1), (4.2.2) y la desigualdad (4.2.3).

Notemos primero que si $\varepsilon \geq 1$ todas estas cotas son triviales, luego asumiremos en esta parte que $0 < \varepsilon < 1$. Sea entonces $\mathcal{D}$ un modelo de hipergrafo aleatorio como en la hipótesis del Teorema Principal y sean $\eta < \min\{\lambda/(d-1), \theta - \lambda\}$ y g tal que $g(t) = O(t^\eta)$. En particular sabemos que existe una constante $g_0 > 1$ tal que $g(t) \leq g_0 t^\eta$ cuando t es grande.

Sea δ suficientemente pequeño para que

$$(1-\delta)^2(1-2\delta) \geq (1-\varepsilon/2) \qquad \text{(Definición de } \delta\text{)}$$

y sean $a = a(\delta)$, $b = b(\delta)$ y $t' = t'(\delta)$ dados por la definición de (c, λ, θ)-mediana.

Consideremos además A una constante suficientemente grande dependiente sólo de δ, ε y los datos del problema y elijamos $t_0 > t'(\delta)$ suficientemente grande de modo que para todo $t \geq t_0$,

$$g(t) \leq g_0 t^\eta \quad \text{y} \quad g_0 bA t^\eta \leq t^{\theta-\lambda}. \tag{4.3.1}$$

Sean ahora $t > t_0$ y $n_1, n_2, \ldots, n_d$ de promedio geométrico N y suma S que satisfagan las condiciones de tamaño y balanceo para estas constantes, es decir, tales que:

$$N \geq At^{\lambda} \quad \text{y} \quad S \leq g(t)N \leq g_0 N t^{\eta}$$

y sea H un hipergrafo elegido de acuerdo a $\mathcal{D}(K_{n_1,\ldots,n_d})$.

Si los n_j satisfacen las cotas de tamaño en la definición de (c, λ, θ)-mediana, tendríamos gracias a que el modelo posee una constante de concentración, una cota de concentración en torno a $cNt^{-\lambda}$ que nos serviría para proseguir. Sin embargo, puede que S sea demasiado grande para satisfacer la cota superior en el tamaño. Por dicha razón, descompondremos H en subhipergrafos que tengan tamaño adecuado.

La idea es dividir H en subhipergrafos del mismo tamaño, que denotaremos bloques, y de tal forma que, en cierto modo, sean proporcionales a H. La división no será exacta puesto que el número de vértices de cada clase no necesariamente será divisible por el número de bloques. Sin embargo podremos suplir este inconveniente haciendo que el bloque final de la división más pequeño que los demás.

Sea $q = \lceil N/(At^{\lambda}) \rceil$. Dividiremos H en q bloques del mismo tamaño (y probablemente un bloque adicional más pequeño). Para cada $1 \leq j \leq d$ definamos $n'_j = \lfloor n_j/q \rfloor$ a ser el largo de la clase j de un bloque y llamemos N' al promedio geométrico de dichos largos y S' a su suma.

Definamos con esto para todo $1 \leq i \leq q$ el hipergrafo H_i inducido por los vértices:

$$\begin{aligned}
&\text{Clase } A_1\text{:} \quad (i-1)n'_1{+}1, \ldots, in'_1.\\
&\text{Clase } A_2\text{:} \quad (i-1)n'_2{+}1, \ldots, in'_2.\\
&\qquad\qquad\vdots\\
&\text{Clase } A_d\text{:} \quad (i-1)n'_d{+}1, \ldots, in'_d.
\end{aligned}$$

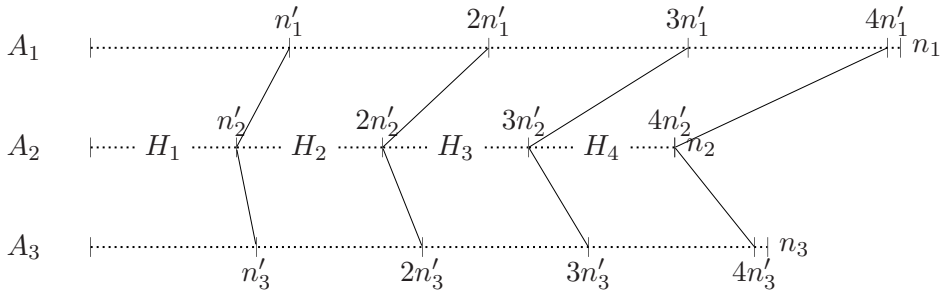

Figura 4.2: División de un hipergrafo H en $q = 4$ bloques del mismo tamaño. Debido a que las clases de H no son múltiplos de 4 en algunas de ellas sobran vértices. Las clases de los 4 bloques H_1, H_2, H_3, H_4 tienen largos prácticamente proporcionales a los de las clases correspondientes de H.

Con esto, para todo i, H_i se distribuye de acuerdo a $\mathcal{D}(K_{n'_1,\ldots,n'_d})$. Además los $L(H_i)$ al ser disjuntos son independientes entre sí. Notemos además que la unión de todos estos bloques no

necesariamente cubre todo H. De hecho en la clase j hay $n_j - q\lfloor n_j/q \rfloor \geq 0$ vértices sin usar. No nos preocuparemos de dichos vértices pues todo lo que necesitamos de ahora en adelante es que $L(H) \geq \sum_{i=1}^{q} L(H_i)$.

Por otro lado, por definición

$$\frac{N}{At^\lambda} \leq q \leq \frac{N}{At^\lambda} + 1 \leq \frac{2N}{At^\lambda}. \qquad \text{(Estimación de } q\text{)}$$

Necesitaremos del siguiente lema para estimar N',

Lema 4.5. *Para toda colección de reales positivos* $(x_j)_{1\leq j\leq d}$*,*

$$\prod_{j=1}^{d}(x_j - 1) \geq \prod_{j=1}^{d} x_j - \left[\sum_{j=1}^{d} x_j\right]^{d-1}.$$

Demostración. Lo probaremos por inducción en d. Para $d = 1$ se tiene igualdad, y en general para $d \geq 2$ si asumimos que el lema es cierto para todo valor menor que d, se tendrá

$$\begin{aligned}
\prod_{j=1}^{d} x_j - \prod_{j=1}^{d}(x_j - 1) &= (x_d - 1)\left[\prod_{j=1}^{d-1} x_j - \prod_{j=1}^{d-1}(x_j - 1)\right] + \prod_{j=1}^{d-1} x_j \\
&\leq (x_d - 1)\left[\sum_{j=1}^{d-1} x_j\right]^{d-2} + \left[\sum_{j=1}^{d-1} x_j\right]^{d-1} \\
&= \left[\sum_{j=1}^{d-1} x_j\right]^{d-2}\left(x_d - 1 + \sum_{j=1}^{d-1} x_j\right) \leq \left[\sum_{j=1}^{d} x_j\right]^{d-1}.
\end{aligned}$$

■

Gracias al lema anterior, a la estimación de q y a la condición de balanceo,

$$\begin{aligned}
\frac{N}{q} = \prod_{j=1}^{d}\left(\frac{n_j}{q}\right)^{1/d} \geq N' \geq \left[\prod_{j=1}^{d}\left(\frac{n_j}{q} - 1\right)\right]^{1/d} &\geq \left[\prod_{j=1}^{d}\frac{n_j}{q} - \left(\sum_{j=1}^{d}\frac{n_j}{q}\right)^{d-1}\right]^{1/d} \\
&= \frac{1}{q}\left[N^d - qS^{d-1}\right]^{1/d} \geq \frac{1}{q}\left[N^d - \frac{2N}{At^\lambda} g_0^{d-1} N^{d-1} t^{\eta(d-1)}\right]^{1/d} \\
&= \frac{N}{q}\left[1 - \frac{2}{A} g_0^{d-1} t^{\eta(d-1)-\lambda}\right]^{1/d}.
\end{aligned}$$

Ahora, usando que $\eta(d-1) < \lambda$ e imponiendo A suficientemente grande, se tiene:

$$\frac{N}{q} \geq N' \geq \frac{N}{q}\left[1 - \frac{2}{A} g_0^{d-1}\right]^{1/d} \geq \frac{N}{q}(1 - \delta). \qquad \text{(Estimación de } N'\text{)}$$

Usando la estimación anterior y la estimación para q, probaremos que $n'_1, \ldots, n'_d$ cumplen las cotas de tamaño dadas por la definición de (c, λ, θ)-mediana. En efecto recordando (4.3.1) e imponiendo $A \geq 2a/(1-\delta)$,

$$N' \geq \frac{N}{q}(1-\delta) \geq \frac{1}{2}At^{\lambda}(1-\delta) \geq at^{\lambda},$$
$$S'b \leq \frac{Sb}{q} \leq \frac{SbAt^{\lambda}}{N} \leq g_0 bAt^{\lambda+\eta} \leq t^{\theta}.$$

Con esto, si H' se distribuye de acuerdo a $\mathcal{D}(K_{n'_1,\ldots,n'_d})$ y llamamos Med' a una mediana de $L(H')$, se tendrá en virtud de la definición de (c, λ, θ)-mediana, que $cN't^{-\lambda}(1-\delta) \leq \mathrm{Med}' \leq cN't^{-\lambda}(1+\delta)$.

Luego, por desigualdad de Markov y recordando que h es constante de concentración de $\mathcal{D}$ y que $N' \geq At^{\lambda}(1-\delta)/2$,

$$\begin{aligned}
\mathbb{E}[L(H')] &\geq (1-2\delta)\frac{cN'}{t^{\lambda}}\mathbb{P}\left[L(H') \geq (1-2\delta)\frac{cN'}{t^{\lambda}}\right] \\
&\geq (1-2\delta)\frac{cN'}{t^{\lambda}}\left(1 - \mathbb{P}\left[L(H') \leq \left(\frac{1-2\delta}{1-\delta}\right)\mathrm{Med}'\right]\right) \\
&= (1-2\delta)\frac{cN'}{t^{\lambda}}\left(1 - \mathbb{P}\left[L(H') \leq \left(1-\frac{\delta}{1-\delta}\right)\mathrm{Med}'\right]\right) \\
&\geq (1-2\delta)\frac{cN'}{t^{\lambda}}\left(1 - 2\exp\left(-h\frac{\delta^2}{(1-\delta)^2}\mathrm{Med}'\right)\right) \\
&\geq (1-2\delta)\frac{cN'}{t^{\lambda}}\left(1 - 2\exp\left(-h\frac{\delta^2}{(1-\delta)}\frac{cN'}{t^{\lambda}}\right)\right) \\
&\geq (1-2\delta)\frac{cN'}{t^{\lambda}}\left(1 - 2\exp\left(-\frac{h\delta^2 c}{2}A\right)\right).
\end{aligned}$$

Como A lo podemos tomar suficientemente grande, lo anterior se puede hacer mayor o igual que $(1-2\delta)(1-\delta)cN't^{-\lambda}$. Luego, usando que $L(H) \geq \sum_{i=1}^{q} L(H_i)$, la estimación de N' y la definición de δ,

$$\begin{aligned}
\mathbb{E}[L(H)] &\geq q\mathbb{E}[L(H')] \geq (1-2\delta)(1-\delta)\frac{qcN'}{t^{\lambda}} \geq (1-2\delta)(1-\delta)^2\frac{cN}{t^{\lambda}} \\
&\geq (1-\varepsilon/2)M \geq (1-\varepsilon)M,
\end{aligned}$$

lo que concluye la demostración de la cota inferior para la esperanza dada por (4.2.1).

Para probar la cota para la cola inferior (4.2.3) notemos que:

$$\mathbb{P}\left[L(H) \leq (1-\varepsilon)\frac{cN}{t^{\lambda}}\right] \leq \sum_{\substack{(s_1,\ldots,s_q)\in\mathbb{N}^q \\ s_1+\cdots+s_q \leq (1-\varepsilon)cNt^{-\lambda}}} \mathbb{P}\left[L(H_i) = s_i, i = 1 \ldots q\right].$$

Llamemos $\mathcal{T}$ al conjunto de índices de la suma anterior y para todo $T = (s_1, \ldots, s_q) \in \mathcal{T}$, denotemos $P_T = \mathbb{P}\left[L(H_i) = s_i, i = 1 \ldots q\right]$. Probaremos que P_T es exponencialmente pequeña con respecto a $cNt^{-\lambda}$. En efecto, como los $L(H_i)$ son independientes y están idénticamente distribuidos,

$$P_T = \prod_{i=1}^{q} \mathbb{P}[L(H_i) = s_i] \leq \prod_{i=1}^{q} \mathbb{P}[L(H') \leq s_i].$$

Recordemos que H' tiene tamaño adecuado para usar la aproximación de la mediana. Luego, por definición de (c, λ, θ)-mediana, si llamamos como antes Med' a una mediana de $L(H')$ se tendrá que para todos aquellos i tales que $s_i \leq (1-\delta)cN't^{-\lambda} \leq \mathrm{Med}' \leq (1+\delta)cN't^{-\lambda} \leq 2cN't^{-\lambda}$,

$$\begin{aligned}
\mathbb{P}[L(H') \leq s_i] &= \mathbb{P}\left[L(H') \leq \left(1 - \frac{\mathrm{Med}' - s_i}{\mathrm{Med}'}\right)\mathrm{Med}'\right] \\
&\leq 2\exp\left(-h\frac{(\mathrm{Med}' - s_i)^2}{\mathrm{Med}'}\right) \\
&\leq 2\exp\left(-\frac{ht^{\lambda}}{2cN'}((1-\delta)cN't^{-\lambda} - s_i)^2\right).
\end{aligned}$$

Luego, para todo $1 \leq i \leq q$,

$$\mathbb{P}[L(H') \leq s_i] \leq 2\exp\left(-\frac{ht^{\lambda}}{2cN'}\,\mathrm{máx}(0, (1-\delta)cN't^{-\lambda} - s_i)^2\right)$$

y entonces

$$\begin{aligned}
-\ln P_T &\geq -\sum_{i=1}^{q} \ln \mathbb{P}[L(H') \leq s_i] \\
&\geq -q\ln(2) + \frac{ht^{\lambda}}{2cN'}\sum_{i=1}^{q} \mathrm{máx}(0, (1-\delta)cN't^{-\lambda} - s_i)^2.
\end{aligned}$$

Pero por Cauchy-Schwartz, recordando que $N'q \geq N(1-\delta)$ y la definición de δ,

$$\begin{aligned}
\sqrt{q\sum_{i=1}^{q} \mathrm{máx}(0, (1-\delta)cN't^{-\lambda} - s_i)^2} &\geq \sum_{i=1}^{q} \mathrm{máx}(0, (1-\delta)cN't^{-\lambda} - s_i) \\
&\geq (1-\delta)cN'qt^{-\lambda} - \sum_{i=1}^{q} s_i \\
&\geq (1-\delta)^2 cNt^{-\lambda} - (1-\varepsilon)cNt^{-\lambda} \\
&\geq cNt^{-\lambda}\left((1-\delta)^2(1-2\delta) - (1-\varepsilon)\right) \\
&\geq \frac{cN\varepsilon}{2t^{\lambda}}.
\end{aligned}$$

Notando que lo anterior es positivo y recordando que $N \geq N'q$, se tiene,

$$\begin{aligned}
-\ln P_T &\geq -q\ln(2) + \frac{ht^\lambda}{2cN'}\sum_{i=1}^{q} \text{máx}(0,(1-\delta)cN't^{-\lambda} - s_i)^2 \\
&\geq -q\ln(2) + \frac{ht^\lambda}{2cN'q}\cdot\frac{c^2N^2\varepsilon^2}{4t^{2\lambda}} \\
&\geq -q\ln(2) + \frac{hcN\varepsilon^2}{8t^\lambda}.
\end{aligned}$$

Con esto, usando la estimación $\binom{a}{b} \leq (ea/b)^b$,

$$\begin{aligned}
\mathbb{P}\left[L(H) \leq (1-\varepsilon)\frac{cN}{t^\lambda}\right] &\leq \left|\left\{(s_1,\ldots,s_q)\in\mathbb{N}^d \;\mid\; s_1+\cdots+s_q \leq (1-\varepsilon)\frac{cN}{t^\lambda}\right\}\right| \cdot \max_{T\in\mathcal{T}} P_T \\
&\leq \binom{\lfloor(1-\varepsilon)cNt^{-\lambda}\rfloor}{q} \cdot \max_{T\in\mathcal{T}} P_T \\
&\leq \exp\left(q\ln(e(1-\varepsilon)cNt^{-\lambda}/q) + q\ln 2 - \frac{hcN\varepsilon^2}{8t^\lambda}\right) \\
&\leq \exp\left(q\ln\left(\frac{2e(1-\varepsilon)cNt^{-\lambda}}{q}\right) - \frac{hcN\varepsilon^2}{8t^\lambda}\right).
\end{aligned}$$

Ahora, recordando que $N \leq qAt^\lambda \leq 2N$ e imponiendo A suficientemente grande de modo que $32\ln(2e(1-\varepsilon)cA) \leq Ahc\varepsilon^2$, se tiene que:

$$\begin{aligned}
\mathbb{P}\left[L(H) \leq (1-\varepsilon)\frac{cN}{t^\lambda}\right] &\leq \exp\left(\frac{2N}{At^\lambda}\ln\left(2e(1-\varepsilon)cA\right) - \frac{hcN\varepsilon^2}{8t^\lambda}\right) \\
&\leq \exp\left(\frac{hNc\varepsilon^2}{16t^\lambda} - \frac{hcN\varepsilon^2}{8t^\lambda}\right) \\
&= \exp\left(-\frac{h\varepsilon^2cN}{16t^\lambda}\right).
\end{aligned}$$

Lo que concluye la demostración de la cota para la cola inferior (4.2.3). Para demostrar la cota inferior para la mediana en (4.2.2) basta recordar que $N \geq At^\lambda$, con lo cual el lado derecho de la última desigualdad es menor o igual a

$$\exp\left(-\frac{hc\varepsilon^2}{16}A\right).$$

Imponiendo A suficientemente grande, lo anterior se puede hacer menor o igual a $1/2$. Es decir, hemos probado que $\mathbb{P}\left[L(H) \leq (1-\varepsilon)cNt^{-\lambda}\right] \leq 1/2$, lo que implica que toda mediana de $L(H)$ es mayor o igual a $(1-\varepsilon)cNt^{-\lambda}$.

Comentarios.

1. En la demostración que acabamos de concluir encontramos que la la cota para la cola inferior (4.2.3) se satisface para cualquier constante $K \leq 1/16$. En realidad, rehaciendo la demostración con cotas más precisas, e imponiendo δ más pequeño y A más grande, podemos hacer que K tome cualquier valor estrictamente menor que 1.

2. La demostración de las cotas inferiores del Teorema Principal también puede ser aplicada a otras familias de distribuciones que no sean necesariamente modelos de d-hipergrafo como tal. En particular, la misma demostración es válida para familias del tipo $\mathcal{D} = \mathcal{D}(K_n^{(d)})$, es decir, para familias de distribuciones restringidas a los hipergrafos tales que todas sus clases tienen el mismo largo. La única condición que necesitamos es que estas familias mantengan la propiedad de monotonicidad e independencia por bloques, entendiendo que los únicos subhipergrafos validos son los que tienen sus clases del mismo largo.

   La razón por la cual la demostración sigue funcionando en dicho caso, es que si H es un hipergrafo con todas sus clases del mismo largo, entonces al dividirlo en bloques "proporcionales" al original, los hipergrafos resultantes siguen teniendo sus clases del mismo largo. Este argumento nos servirá para poder aplicar esta demostración a la variante simétrica que se estudiará en el Capítulo 5.

4.4. Demostración de la cotas superiores en el Teorema Principal

En esta sección probaremos las cotas superiores dadas por el Teorema Principal, es decir, probaremos la cota superior de (4.2.1), la cota superior de (4.2.2) y la desigualdad (4.2.4).

Demostraremos en detalle (4.2.4). La demostración de esta cota es bastante extensa, por lo cual la hemos dividido en tres partes. En la primera parte definimos algunas variables que nos serán de utilidad y algunas desigualdades que usaremos frecuentemente. Estas definiciones son relativamente técnicas por lo cual se recomienda usar esta sección más que nada como referencia. A continuación demostraremos las cotas para el caso cuando el promedio geométrico de las clases del hipergrafo, N no es muy grande, en particular cuando $N = o(t^{\theta-\eta})$. Finalmente trataremos el caso cuando N es grande.

4.4.1. Notación y definiciones

Consideremos $\mathcal{D}$ un modelo de d-hipergrafo aleatorio con parámetro interno t, constante de concentración h y que admite una (c, λ, θ)-mediana. Sean además $\varepsilon > 0$, $\eta < \min\{\lambda/(d-1), \theta - \lambda\}$ y $g : \mathbb{N} \to \mathbb{N}$ como en la hipótesis del teorema, en particular, sabemos que existe una constante $g_0 > 1$ tal que $g(t) \leq g_0 t^{\eta}$ cuando t es suficientemente grande.

Definamos

$$\delta = \text{mín}\left\{1, \frac{\varepsilon^2}{(1+\varepsilon)}, \frac{\varepsilon}{6}\right\}, a = a(\delta), b = b(\delta), t' = t'(\delta), A = \text{máx}\left\{\frac{a}{\delta}, \frac{8\ln 2}{h\delta c}\right\}, \tag{4.4.1}$$

con $a(\delta)$, $b(\delta)$ y $t'(\delta)$ los valores entregados por la definición de (c, λ, θ)-mediana.

Elijamos además dos constantes arbitrarias α y β tales que:

$$\lambda < \alpha < \beta < \theta - \eta. \tag{4.4.2}$$

La idea detrás de la definición de estas constantes es aprovechar el hecho que si $\rho_1 < \rho_2$ entonces $t^{\rho_1} = o(t^{\rho_2})$.

Durante la demostración encontraremos dos constantes K_1 y K_2, dependientes sólo de la dimensión d del problema. Con ellas, podremos elegir t_0 mayor que t' de modo que para todo $t \geq t_0$, se tenga $g(t) \leq g_0 t^\eta$,

$$9At^\lambda < t^\alpha < t^\beta < \frac{1}{bdg_0} t^{\theta-\eta} \tag{4.4.3}$$

y

$$\frac{2\alpha K_1}{\delta c K_2 h} t^\lambda < \frac{t^\alpha}{\ln t}. \tag{4.4.4}$$

Las constantes t_0 y el A recién definidos serán los que entregará esta parte del teorema.

Sean ahora $t \geq t_0$ y un conjunto de enteros positivos $n_1, n_2, \ldots, n_d$, de promedio geométrico N y suma S que satisfagan las condiciones de tamaño ($N \geq At^\lambda$) y de balanceo ($S \leq g(t)N$). Sea además $M = cNt^{-\lambda}$ y H elegido de acuerdo a $\mathcal{D}(K_{n_1,\ldots,n_d})$. Dividiremos la demostración de la cota para la cola en dos casos de acuerdo a si N es mayor o menor que t^β.

4.4.2. Demostración para el caso $N < t^\beta$

Si $N < t^\beta$ podremos probar que H tiene tamaño adecuado para aplicar la definición de (c, λ, θ)-mediana con lo cual, recordando que $\mathcal{D}$ posee una constante de concentración h, encontraremos una cota exponencial para $\mathbb{P}\left(L(H) \geq (1+\varepsilon)M\right)$.

Por las condiciones de tamaño y balanceo, la definición de A en (4.4.1) y la desigualdad (4.4.3),

1. $N \geq At^\lambda \geq at^\lambda/\delta \geq at^\lambda$.
2. $Sb \leq g(t)bN \leq g_0 bt^{\eta+\beta} \leq t^\theta/d \leq t^\theta$.

Lo anterior prueba que $n_1, \ldots, n_d$ satisfacen las cotas de tamaño dadas por la definición de (c, λ, θ)-mediana. Luego, si H es elegido de acuerdo a $\mathcal{D}(K_{n_1,\ldots,n_d})$, entonces toda mediana de $L(H)$ está

entre $M(1-\delta)$ y $M(1+\delta)$. Sigue que para Med una mediana de $L(H)$,

$$\begin{aligned}\mathbb{P}\left(L(H) \geq (1+\varepsilon)M\right) &\leq \mathbb{P}\left(L(H) \geq \left(1 + \frac{(1+\varepsilon)M - \text{Med}}{\text{Med}}\right)\text{Med}\right) \\ &\leq 2\exp\left(-h\frac{((1+\varepsilon)M - \text{Med})^2/\,\text{Med}^2}{[(1+\varepsilon)M/\,\text{Med}]}\,\text{Med}\right) \\ &= 2\exp\left(-h\frac{((1+\varepsilon)M - \text{Med})^2}{(1+\varepsilon)M}\right) \\ &\leq \exp\left(\ln 2 - h\frac{(\varepsilon-\delta)^2}{(1+\varepsilon)}M\right).\end{aligned}$$

Recordando la definición de A y que $\delta \leq \varepsilon/2 < \varepsilon/6$, lo anterior es menor a:

$$\begin{aligned}\exp\left(\frac{Ah\delta c}{8} - \frac{h\varepsilon^2}{4(1+\varepsilon)}M\right) &\leq \exp\left(\frac{\varepsilon^2}{(1+\varepsilon)}\frac{hnc}{8t^\lambda} - \frac{h\varepsilon^2}{4(1+\varepsilon)}M\right) \\ &= \exp\left(-\frac{h\varepsilon^2}{8(1+\varepsilon)}M\right),\end{aligned}$$

y luego la cota buscada para la concentración se tiene tomando K cualquier constante menor o igual a $1/8$. Hemos probado así que la cota se tiene cuando $N < t^\beta$.

4.4.3. Demostración para el caso $N \geq t^\beta$

En esta subsección, probaremos la cota para el caso $N \geq t^\beta$. En esta ocasión no podremos probar que H tiene un tamaño adecuado para aplicar la definición de (c, λ, θ)-mediana y así obtener la cota exponencial buscada. Sin embargo, de manera similar a como lo hicimos para probar las cotas inferiores del Teorema Principal, podremos dividir H en bloques que si tengan esta condición.

Partición en bloques

Definamos $l = t^\alpha$, $L = g_0 t^{\eta+\alpha}$ y

$$m_{\max} = \lceil (1+\varepsilon)M \rceil\,. \tag{4.4.5}$$

En lo que sigue acotaremos superiormente la cantidad $\mathbb{P}\left(L(H) \geq m_{\max}\right)$, es decir, la probabilidad de que H posea un subhipergrafo monótono con al menos $m_{\max}$ aristas.

Notemos que ningún subhipergrafo monótono de H puede tener un número de aristas mayor que el tamaño de alguna de las clases de H. Esto se debe a que cada arista usa un vértice de cada clase y en un subhipergrafo monótono las aristas no comparten vértices. En consecuencia $L(H)$ debe ser menor que N, el promedio geométrico de los largos de las clases y por lo tanto, si $m_{\max}$

es mayor que N la probabilidad anterior resulta ser 0. Por esto en lo que sigue asumiremos además que $m_{\max} \leq N$.

Sea J un subhipergrafo monótono sobre el mismo conjunto de vértices que $K_{n_1,\ldots,n_d}$ con $m_{\max}$ aristas. Definiremos una partición de $E(J)$ en bloques de aristas consecutivas, de modo que cada bloque no ocupe más de L nodos consecutivos de cada clase de partición y cada bloque no tenga demasiadas aristas. Específicamente, si llamamos

$$s_{\max} = \left\lfloor \frac{l}{N} m_{\max} \right\rfloor \tag{4.4.6}$$

pediremos que cada bloque de la partición no tenga más de $s_{\max}$ aristas.

Dadas dos aristas e y $\widetilde{e}$ de J, con $e \leq \widetilde{e}$, denotamos $[e, \widetilde{e}]$ al subconjunto de aristas ubicadas entre e y $\widetilde{e}$, usando el orden natural (coordenada a coordenada) definido anteriormente. Es decir

$$[e, \widetilde{e}] = \{f \in E(J) \ \mid \ e \leq f \leq \widetilde{e}\}.$$

Denotaremos partición en bloques de $E(J)$ al conjunto $\mathcal{P}(J) = \{[e_i, \widetilde{e}_i] \ \mid \ 1 \leq i \leq q\}$, con $e_i = \left(v_1^{(i)}, v_2^{(i)}, \ldots v_d^{(i)}\right)$, $\widetilde{e}_i = \left(\widetilde{v}_1^{(i)}, \widetilde{v}_2^{(i)}, \ldots \widetilde{v}_d^{(i)}\right)$, construido de la siguiente forma:

1. e_1 es la primera arista de $E(J)$.
2. Una vez que e_i está definida, $\widetilde{e}_i$ es la mayor arista de $E(J)$ que cumple:
   a) $[e_i, \widetilde{e}_i]$ tiene a lo más $s_{\max}$ elementos.
   b) $\widetilde{v}_j^{(i)} - v_j^{(i)} \leq L$ para todo $1 \leq j \leq d$. (Esto es, con un poco de abuso de notación, recordando que cada clase de partición de K tiene sus nodos numerados $1, 2, \ldots, n_j$).
3. Una vez definido $\widetilde{e}_i$, si todavía quedan aristas mayores en $E(J)$, tomamos e_{i+1} como la arista inmediatamente mayor a $\widetilde{e}_i$.

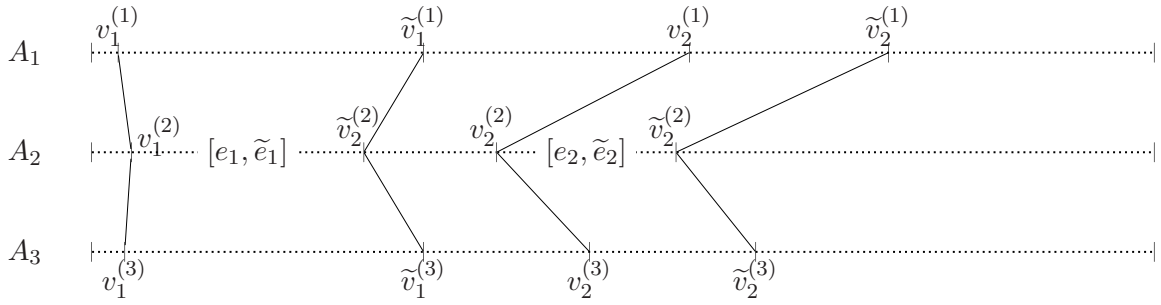

Figura 4.3: Partición en bloques de un hipergrafo H. Cada bloque $[e_i, \widetilde{e}_i]$ posee a lo más $s_{\max}$ aristas y cada clase de un bloque no tiene más de L vértices.

La partición en bloques definida de esta forma, es una partición de $E(J)$ en q bloques consecutivos. Probaremos las siguientes cotas para q:

$$\frac{N}{l} \leq q \leq \frac{3N}{l}. \qquad \text{(Estimación de } q\text{)}$$

En efecto:

- ⋄ Cada bloque tiene a lo más $s_{\max}$ aristas, luego $q \geq m_{\max}/s_{\max} \geq N/l$.
- ⋄ Denotemos *bloques cortos* al bloque final $[e_q, \widetilde{e}_q]$ y a todos aquellos bloques que tengan exactamente $s_{\max}$ aristas. Sea $B_0 \subseteq [q]$ el conjunto de índices de los bloques cortos. Luego $m_{\max} \geq (|B_0|-1)s_{\max} \geq (|B_0|-1)(m_{\max}l/N-1)$. Como $m_{\max}l/N$ es suficientemente grande, sigue que $|B_0| \leq 2N/l$.
- ⋄ Denotemos *bloques regulares* a aquellos bloques que no son cortos y sea $B_1 = [q] \setminus B_0$. Además, llamemos cubrimiento del bloque a todos los nodos contenidos entre la primera arista del bloque (inclusive) y la primera arista del bloque siguiente (sin incluir).

  Por definición de la partición, si el bloque i-ésimo es regular, entonces se tendrá que para alguna de las clases j, $v_j^{(i+1)} - v_j^{(i)} \geq L$. Sigue que $\sum_{j=1}^{d}(v_j^{(i+1)} - v_j^{(i)}) \geq L$. Es decir, el número de nodos que *cubre* el bloque es al menos L. Como todos los nodos son cubiertos por exactamente un bloque, se concluye que $|B_1| \leq S/L \leq Ng_0t^{\eta}/L = N/l$.
- ⋄ Las observaciones anteriores dicen que $q = |B_0| + |B_1| \leq 3N/l$.

Tipos de particiones

Sea s_i el número de aristas de $E(J)$ en el bloque i. Definimos el *tipo* de la partición $T = T(M)$ a la $3q$-tupla

$$T = (e_1, \widetilde{e}_1, s_1, \ldots, e_q, \widetilde{e}_q, s_q).$$

Sea además $\mathcal{T}$ el conjunto de todos los posibles tipos de particiones de un subhipergrafo monótono de K con $m_{\max}$ aristas.

Lema 4.6. *Para alguna constante K_1 que sólo depende de d,*

$$|\mathcal{T}| \leq \exp\left(K_1 \frac{N}{l} \ln l\right).$$

Demostración. Notemos que las aristas e_i se determinan completamente al indicar sus vértices. Luego el número de formas de elegir $e_1, \ldots, e_q$ es a lo más el número de elegir q elementos en cada uno de las clases de partición de los nodos, es decir $\prod_{i=1}^{d}\binom{n_i}{q}$. Similarmente, el número de formas de elegir $\widetilde{e}_1, \ldots, \widetilde{e}_q$ está acotado por la misma cantidad.

Por otro lado, el número de elecciones para los s_i es menor que el número de particiones de $m_{\max}$ en q sumandos positivos. Como asumimos al principio de la subsección que $m_{\max} \leq N$, se tiene que la cantidad anterior es menor que $\binom{N}{q}$. Usando la estimación $\binom{a}{b} \leq (ea/b)^b$ se obtiene que, para un q fijo, el número de tipos está acotado por:

$$\binom{N}{q}\left(\prod_{i=1}^{d}\binom{n_i}{q}\right)^2 \leq (eN/q)^q\left(\prod_{i=1}^{d}(en_i/q)^q\right)^2 = (eN/q)^q\left(\prod_{i=1}^{d}en_i/q\right)^{2q} = (eN/q)^{q+2qd}.$$

Usando las estimaciones obtenidas para q se obtiene que:

$$|\mathcal{T}| \leq \sum_{q=\lceil N/l\rceil}^{\lfloor 3N/l\rfloor}(eN/q)^{q(1+2d)} \leq 3\frac{N}{l}\left(\frac{eN}{(N/l)}\right)^{(1+2d)3N/l} = 3\frac{N}{l}(el)^{(1+2d)3N/l},$$

y luego:

$$\ln|\mathcal{T}| \leq \ln\left(\frac{3N}{l}\right) + (1+2d)3\frac{N}{l}\left(1+\ln(l)\right) \leq (2+2d)3\frac{N}{l}\left(1+\ln(l)\right) \leq (1+d)12\frac{N}{l}\ln l. \qquad \blacksquare$$

Probabilidad de un subhipergrafo monótono con un tipo de partición de bloques

A continuación veremos que para cada tipo fijo T la probabilidad de que un hipergrafo elegido según $\mathcal{D}(K_{n_1,\ldots,n_d})$ contenga un subhipergrafo monótono del tipo T con $m_{\max}$ aristas es exponencialmente pequeña con respecto a M.

Lema 4.7. *Para cada tipo $T \in \mathcal{T}$, la probabilidad P_T de que un hipergrafo elegido según $\mathcal{D}(K_{n_1,\ldots,n_d})$ contenga un subhipergrafo monótono J con m_{max} aristas, tal que $T(J) = T$ cumple,*

$$P_T \leq \exp\left(-K_2 h\frac{\varepsilon^2}{(1+\varepsilon)}M\right),$$

para alguna constante absoluta, $K_2 > 0$.

Demostración. Sea $T = (e_1, \widetilde{e}_1, s_1, \ldots, e_q, \widetilde{e}_q, s_q)$ un tipo de partición y, como antes, para todo i, llamemos $e_i = \left(v_1^{(i)}, v_2^{(i)}, \ldots v_d^{(i)}\right)$ y $\widetilde{e}_i = \left(\widetilde{v}_1^{(i)}, \widetilde{v}_2^{(i)}, \ldots \widetilde{v}_d^{(i)}\right)$. Sea H elegido de acuerdo a $\mathcal{D}(K_{n_1,\ldots,n_d})$, y sea H_i el subhipergrafo inducido por los vértices entre e_i y $\widetilde{e}_i$, es decir por los vértices

$$\begin{gathered} v_1^{(i)}, v_1^{(i)}+1, \ldots, \widetilde{v}_1^{(i)}, \\ v_2^{(i)}, v_2^{(i)}+1, \ldots, \widetilde{v}_2^{(i)}, \\ \vdots \\ v_d^{(i)}, v_d^{(i)}+1, \ldots, \widetilde{v}_d^{(i)}. \end{gathered}$$

Notemos que H_i se distribuye de acuerdo a $\mathcal{D}\Big(K_{n_1^{(i)},n_2^{(i)},\ldots,n_d^{(i)}}\Big)$, donde $n_j^{(i)} = \widetilde{v}_j^{(i)} - v_j^{(i)} + 1$ son los largos en cada clase de partición. Además. si existe un subhipergrafo J de H tal que $T(J) = T$, entonces se debe tener que $L(H_i) \geq s_i$ para todo $i = 1, \ldots, q$.

Como los eventos $L(H_i) \geq s_i$ son independientes para distintos i (recordar que $\mathcal{D}$ es un modelo independiente por bloques), se tiene que:

$$P_T \leq \prod_{i=1}^{q} \mathbb{P}\left[L\Big(\mathcal{D}\Big(K_{n_1^{(i)},n_2^{(i)},\ldots,n_d^{(i)}}\Big)\Big) \geq s_i\right].$$

Denotemos por $N_i = \left(\prod_{j=1}^{d} n_j^{(i)}\right)^{1/d}$ al promedio geométrico de los largos de las clases de H_i y por $S_i = \sum_{j=1}^{d} n_j^{(i)}$ a su suma. Si para cada i, los tamaños de las clases de H_i cumplen las cotas de tamaño en la definición de (c, λ, θ)-mediana, entonces todas las probabilidades anteriores serán pequeñas.

Por construcción de los bloques: $n_1^{(i)}, n_2^{(i)}, \ldots, n_d^{(i)} \leq L$. Luego

$$S_i b \leq dbL = g_0 dbt^{\eta+\alpha} < t^{\theta}.$$

Sin embargo, la cota inferior en el tamaño, $N_i \geq at^{\lambda}$, puede fallar por lo cual artificialmente *aumentaremos* el tamaño de los bloques donde esta condición falle. Específicamente, definamos para todo i:

$$\begin{aligned}
\bar{n}_1^{(i)} &= \text{máx}(\delta n_1 At^{\lambda}/N, n_1^{(i)}),\\
\bar{n}_2^{(i)} &= \text{máx}(\delta n_2 At^{\lambda}/N, n_2^{(i)}),\\
&\vdots\\
\bar{n}_d^{(i)} &= \text{máx}(\delta n_d At^{\lambda}/N, n_d^{(i)}).
\end{aligned}$$

Como es costumbre, llamemos $\overline{N}_i = \left(\prod_{j=1}^{d} \bar{n}_j^{(i)}\right)^{1/d}$ al promedio geométrico de los nuevos largos de las clases y $\overline{S}_i = \sum_{j=1}^{d} \bar{n}_j^{(i)}$ a su suma.

Notemos que si aumentamos los tamaños de las clases de partición de un hipergrafo, por monotonicidad, la probabilidad de que posea un emparejamiento monótono de s_i aristas aumenta. Luego

$$P_T \leq \prod_{i=1}^{q} \mathbb{P}\left[L\Big(\mathcal{D}\Big(K_{n_i^{(1)},n_i^{(2)},\ldots,n_i^{(d)}}\Big)\Big) \geq s_i\right] \leq \prod_{i=1}^{q} \mathbb{P}\left[L\Big(\mathcal{D}\Big(K_{\bar{n}_i^{(1)},\bar{n}_i^{(2)},\ldots,\bar{n}_i^{(d)}}\Big)\Big) \geq s_i\right].$$

Veamos que los nuevos largos sí cumplen las cotas de tamaño buscadas. En efecto, recordando la desigualdad (4.4.3), podemos ver que para todo j,

$$\delta n_j At^{\lambda}/N \leq \delta n_j t^{\alpha}/N \leq \delta St^{\alpha}/N \leq \delta g(t) t^{\alpha} \leq \delta g_0 t^{\alpha+\eta} = \delta L \leq L.$$

Luego, como los $n_j^{(i)}$ también son menores que L se concluye que los $\bar{n}_j^{(i)}$ son menores que L y con esto, al igual que antes de agrandar los bloques, $\overline{S}_i b \leq t^\theta$.

Por otro lado, recordando que $A \geq a/\delta$,

$$\overline{N}_i = \left(\prod_{j=1}^{d} \bar{n}_j^{(i)}\right)^{1/d} \geq \frac{\delta A t^\lambda}{N}\left(\prod_{j=1}^{d} n_j\right)^{1/d} = \delta A t^\lambda \geq a t^\lambda.$$

Sigue que los nuevo tamaños cumplen las cotas dadas por la definición de (c, λ, θ)-mediana.

Sea ahora $\overline{\text{Med}}_i$ una mediana de $L(\mathcal{D}(K_{\bar{n}_i^1, \bar{n}_i^2, \ldots, \bar{n}_i^d}), t))$. Por definición de (c, λ, θ)-mediana,

$$(1-\delta)c\overline{N}_i t^{-\lambda} \leq \overline{\text{Med}}_i \leq (1+\delta)c\overline{N}_i t^{-\lambda}.$$

Luego, para los i tales que $s_i \geq (1+\delta)c\bar{N}_i t^{-\alpha} \geq \overline{\text{Med}}_i$, usando que h es constante de concentración para el modelo, se tendrá

$$\begin{aligned}
\mathbb{P}\left[L\Big(\mathcal{D}\Big(K_{\bar{n}_i^{(1)}, \bar{n}_i^{(2)}, \ldots, \bar{n}_i^{(d)}}\Big)\Big) \geq s_i\right] &\leq 2\exp\left(-h\frac{\big(s_i \,/\, \overline{\text{Med}}_i\big) - 1\big)^2}{s_i \,/\, \overline{\text{Med}}_i}\overline{\text{Med}}_i\right) \\
&= 2\exp\left(-h\frac{(s_i - \overline{\text{Med}}_i)^2}{s_i}\right) \\
&\leq 2\exp\left(-h\frac{(s_i - (1+\delta)c\overline{N}_i t^{-\lambda})^2}{s_i}\right).
\end{aligned}$$

Luego para todo i, recordando que s_i, el número de aristas del bloque i es a lo más $s_{\max}$,

$$\begin{aligned}
\mathbb{P}\left[L\Big(\mathcal{D}\Big(K_{\bar{n}_i^{(1)}, \bar{n}_i^{(2)}, \ldots, \bar{n}_i^{(d)}}\Big)\Big) \geq s_i\right] &\leq 2\exp\left(-h\frac{\text{máx}(0, s_i - (1+\delta)c\overline{N}_i t^{-\lambda})^2}{s_i}\right) \\
&\leq 2\exp\left(-h\frac{\text{máx}(0, s_i - (1+\delta)c\overline{N}_i t^{-\lambda})^2}{s_{\max}}\right).
\end{aligned}$$

Con esto:

$$\begin{aligned}
-\ln P_T &\geq -\ln\left(\prod_{i=1}^{q} \mathbb{P}\left[L\Big(\mathcal{D}\Big(K_{\bar{n}_i^{(1)}, \bar{n}_i^{(2)}, \ldots, \bar{n}_i^{(d)}}\Big)\Big) \geq s_i\right]\right) \\
&= -\sum_{i=1}^{q} \ln\left(\mathbb{P}\left[L\Big(\mathcal{D}\Big(K_{\bar{n}_i^{(1)}, \bar{n}_i^{(2)}, \ldots, \bar{n}_i^{(d)}}\Big)\Big) \geq s_i\right]\right) \\
&\geq -q\ln(2) + \frac{h}{s_{\max}}\sum_{i=1}^{q} \text{máx}(0, s_i - (1+\delta)c\overline{N}_i t^{-\lambda})^2.
\end{aligned}$$

Acotemos inferiormente la suma del lado derecho de la última expresión obtenida. Para ello usaremos la siguiente desigualdad de Hölder generalizada:

Lema 4.8 (Desigualdad de Hölder generalizada)**.** *Para toda colección de números positivos* $x_{i,j}$, $1 \leq i \leq q$, $1 \leq j \leq d$,

$$\left(\sum_{i=1}^{q}\prod_{j=1}^{d} x_{i,j}\right)^{d} \leq \prod_{j=1}^{d}\sum_{i=1}^{q} x_{i,j}^{d}.$$

Usando la desigualdad anterior, para la colección $x_{i,j} = \left(\bar{n}_j^{(i)}\right)^{1/d}$, se tiene que

$$\begin{aligned}\sum_{i=1}^{q}\overline{N}_i = \sum_{i=1}^{q}\prod_{j=1}^{d} x_{i,j} &\leq \left(\prod_{j=1}^{d}\sum_{i=1}^{q}\bar{n}_j^{(i)}\right)^{1/d} \\ &\leq \left(\prod_{j=1}^{d}\sum_{i=1}^{q}\left(n_j^{(i)} + \delta n_j A t^{\lambda}/N\right)\right)^{1/d} \\ &\leq \prod_{j=1}^{d}\left(n_j + \delta q n_j A t^{\lambda}/N\right)^{1/d} = \left(\prod_{j=1}^{d} n_j^{1/d}\right)\left(1 + \delta q A t^{\lambda}/N\right).\end{aligned}$$

Recordando las cotas obtenidas para q y la desigualdad (4.4.3), se concluye que:

$$\sum_{i=1}^{q}\overline{N}_i \leq N(1 + 3\delta A t^{\lambda-\alpha}) \leq N(1 + \delta/3) \leq N(1+\delta).$$

Usando la desigualdad anterior, la desigualdad de Cauchy-Schwartz y recordando que la suma de los s_i es exactamente $m_{\max} = \lceil (1+\varepsilon)M \rceil$, se tiene:

$$\begin{aligned}\sqrt{q\sum_{i=1}^{q}\max\left(0, s_i - (1+\delta)c\overline{N}_i t^{-\lambda}\right)^2} &\geq \sum_{i=1}^{q}\max\left(0, s_i - (1+\delta)c\overline{N}_i t^{-\lambda}\right) \\ &\geq m_{\max} - (1+\delta)ct^{-\lambda}\sum_{i=1}^{q}\overline{N}_i \\ &\geq M(1+\varepsilon) - M(1+\delta)^2.\end{aligned}$$

Veamos que el lado derecho es positivo, para eso recordemos que por definición $\delta \leq \varepsilon/6$. Con esto,

$$(1+\varepsilon) - (1+\delta)^2 = \varepsilon - 2\delta - \delta^2 \geq \varepsilon - 3\delta \geq \varepsilon/2.$$

Es decir:

$$\sqrt{q\sum_{i=1}^{q}\max\left(0, s_i - (1+\delta)c\overline{N}_i t^{-\lambda}\right)^2} \geq \frac{\varepsilon M}{2}.$$

Lo anterior implica que:

$$\begin{aligned} -\ln P_T &\geq -q\ln(2) + \frac{h}{s_{\max}} \sum_{i=1}^{q} \text{máx}\left(0, s_i - (1+\delta)c\overline{N}_i t^{-\lambda}\right)^2 \\ &\geq -q\ln(2) + \frac{h}{s_{\max} q} \cdot \frac{\varepsilon^2 M^2}{4}. \end{aligned}$$

Además, por definición,

$$s_{\max} \leq (1+\varepsilon)\frac{lM}{N}.$$

Luego:

$$-\ln P_T \geq -q\ln(2) + \frac{hN\varepsilon^2 M}{4q(1+\varepsilon)l}.$$

Finalmente, recordando que $q \leq 3N/l$, y que $l \geq 9At^{\lambda}$, se concluye que:

$$\begin{aligned} -\ln P_T &\geq -\frac{N\ln(2)}{3At^{\lambda}} + \frac{h\varepsilon^2 M}{12(1+\varepsilon)} \\ &= \frac{cN}{t^{\lambda}}\left(\frac{h\varepsilon^2}{12(1+\varepsilon)} - \frac{\ln(2)}{3Ac}\right). \end{aligned}$$

Dado que $A \geq 8\ln(2)/(hc\delta)$ y $\delta \leq \varepsilon^2/(1+\varepsilon)$,

$$\begin{aligned} -\ln P_T &\geq \frac{cN}{t^{\lambda}}\left(\frac{h\varepsilon^2}{12(1+\varepsilon)} - \frac{h\delta}{24}\right) \\ &\geq \frac{\varepsilon^2}{(1+\varepsilon)}\frac{hM}{24}, \end{aligned}$$

con lo cual la constante K_2 buscada en el Lema 4.7 resulta ser positiva y mayor o igual que $1/24$. ■

Demostración de la cota para la cola superior para el caso $N \geq t^{\beta}$

Demostración de (4.2.4). Tenemos que

$$\mathbb{P}\left[L(\mathcal{D}(K_{n_1,n_2,\ldots,n_d})) \geq m_{\max}\right] \leq \sum_{T\in\mathcal{T}} P_T \leq |\mathcal{T}| \max_{t\in\mathcal{T}} P_T.$$

De los Lemas 4.6 y 4.7, la desigualdad (4.4.4) y la definición de δ,

$$\begin{aligned}\mathbb{P}\left[L(\mathcal{D}(K_{n_1,n_2,\ldots,n_d}),t)) \geq m_{\max}\right] &\leq \exp\left(K_1\frac{N}{l}\ln l - K_2 h\frac{\varepsilon^2}{(1+\varepsilon)}M\right)\\ &= \exp\left(K_1\alpha N\frac{\ln t}{t^\alpha} - K_2 h\frac{\varepsilon^2}{(1+\varepsilon)}M\right)\\ &\leq \exp\left(\frac{\delta K_2 h}{2}\frac{cN}{t^\lambda} - K_2 h\frac{\varepsilon^2}{(1+\varepsilon)}M\right)\\ &\leq \exp\left(-\frac{K_2 h\varepsilon^2}{2(1+\varepsilon)}M\right).\end{aligned}$$

Recordando que K_2 era una constante absoluta (1/24), se tiene que cualquier constante K menor o igual a 1/24 sirve para la cota de concentración dada por (4.2.4). ■

4.4.4. Demostración de la cota superior para la esperanza y mediana en el Teorema Principal.

Ahora probaremos las cotas superiores de (4.2.1) y (4.2.2). Para la demostración de la cota superior de la esperanza, observemos primero que en la demostración anterior, para d y η fijos, A y t_0 dependen solamente de δ y que para todo $\varepsilon \geq 2$, $\delta = \delta(\varepsilon) = 1$ (Ver (4.4.1)).

Sean ahora $\varepsilon_0 > 0$, d, η y g como antes y K la constante entregada por la desigualdad (4.2.4) que acabamos de demostrar. Sea además, $\delta_0 = \delta(\varepsilon_0/2)$ como en (4.4.1) y A_0 una constante suficientemente grande como para que

$$\exp\left(-\frac{KhcA_0\varepsilon_0^2}{4(1+\varepsilon_0/2)}\right) \leq \frac{\varepsilon_0}{12} \qquad \text{y} \qquad \frac{8}{c^2A_0^2Kh} \leq \varepsilon_0. \qquad \text{(Elección de } A_0\text{)}$$

Definamos $\widetilde{A} = \max\{A(\delta_0), A(1), A_0\}$ y $\widetilde{t_0} = \max\{t_0(\delta_0), t_0(2)\}$.

Sean ahora $t \geq \widetilde{t_0}$ y valores $n_1, \ldots, n_d$ de promedio geométrico N y suma S que satisfagan las condiciones de tamaño y balanceo para estas constantes $\widetilde{t_0}$ y $\widetilde{A}$ recién definidas, es decir, tales que:

$$N \geq \widetilde{A}t^\lambda \quad \text{y} \quad Sb \leq g(t)N.$$

La elección de $\widetilde{t_0}$ y $\widetilde{A}$ garantiza que la desigualdad (4.2.4) es válida tanto para $\varepsilon = \varepsilon_0/2$ como para todo $\varepsilon \geq 2$.

Sea H elegido de acuerdo a $\mathcal{D}(K_{n_1,\ldots,n_d})$ y $M = cNt^{-\lambda}$. Notemos primero que:

$$\begin{aligned}\mathbb{E}[L(H)] = \mathbb{E}[L(H)\mathbb{1}_{[0,(1+\varepsilon_0/2)M]}(L(H))] \,+\\ \mathbb{E}[L(H)\mathbb{1}_{[(1+\varepsilon_0/2)M,3M]}(L(H))] + \mathbb{E}[L(H)\mathbb{1}_{[3M,\infty]}(L(H))].\end{aligned}$$

Acotemos cada término por separado. El primero queda:

$$\mathbb{E}[L(H)\mathbb{1}_{[0,(1+\varepsilon_0/2)M]}(L(H))] \leq (1+\varepsilon_0/2)M.$$

El segundo queda gracias a la desigualdad (4.2.4),

$$\begin{aligned}\mathbb{E}[L(H)\mathbb{1}_{[(1+\varepsilon_0/2)M,3M]}(L(H))] &\leq 3M\mathbb{P}[L(H) > (1+\varepsilon_0/2)M] \\ &\leq 3M\exp\left(-\frac{Kh\varepsilon_0^2/4}{(1+\varepsilon_0/2)}\cdot\frac{cN}{t^\lambda}\right) \\ &\leq 3M\exp\left(-\frac{Kh\varepsilon_0^2/4}{(1+\varepsilon_0/2)}cA_0\right) \leq \frac{M\varepsilon_0}{4}.\end{aligned}$$

Notando que para $\varepsilon > 2$, $\varepsilon^2/(1+\varepsilon) \geq \varepsilon/2$, se tiene que

$$\begin{aligned}\mathbb{E}[L(H)\mathbb{1}_{[3M,\infty]}(L(H))] &\leq \int_2^\infty \mathbb{P}[L(H) > (1+\varepsilon)M]\,d\varepsilon \\ &\leq \int_2^\infty \exp\left(-\frac{Kh\varepsilon^2}{(1+\varepsilon)}M\right)\,d\varepsilon \\ &\leq \int_2^\infty \exp\left(-\frac{Kh\varepsilon}{2}M\right)\,d\varepsilon \\ &= \frac{2}{MKh}\exp(-MKh) \leq \frac{2M}{M^2Kh} \leq \frac{2M}{c^2A_0^2Kh} \leq \frac{M\varepsilon_0}{4}.\end{aligned}$$

En resumen,

$$\mathbb{E}[L(H)] \leq (1+\varepsilon_0/2)M + M\varepsilon_0/4 + M\varepsilon_0/4 = (1+\varepsilon_0)M.$$

Lo que completa la demostración de la cota superior para la esperanza.

Para ver la cota superior para la Mediana, sea $\varepsilon > 0$ y sean A y t_0 las constantes asociadas a este ε en la demostración de la cota para la cola superior (4.2.4). Si definamos

$$A' = \text{máx}\left\{A, \frac{(1+\varepsilon)\ln(2)}{Khc\varepsilon^2}\right\},$$

entonces para todo $t \geq t_0$ y todo $n_1, \ldots, n_d$, de promedio geométrico N y suma S que satisfagan las condiciones de tamaño y balanceo para esta nueva constante A' recién definida, es decir, tales que:

$$N \geq A't^\lambda \quad \text{y} \quad Sb \leq g(t)N,$$

se tendrá:

$$\mathbb{P}[L(H) \geq (1+\varepsilon)M] \leq \exp\left(-Kh\frac{\varepsilon^2}{1+\varepsilon}\cdot\frac{cN}{t^\lambda}\right) \leq \exp\left(-Kh\frac{\varepsilon^2}{1+\varepsilon}cA'\right) \leq 1/2,$$

lo que implica que toda mediana de $L(H)$ es menor o igual que $(1+\varepsilon/2)M$. Esto concluye la demostración de la cota superior para la mediana, lo cual finaliza la demostración del Teorema Principal.

Comentarios. En la demostración anterior se probó que la cota para la cola superior (4.2.4) se satisface para cualquier constante $K \leq 1/48$. Sin embargo, al igual que en la demostración para la cola inferior, si se rehace la demostración con cotas más precisas, se puede hacer que K tome cualquier valor estrictamente menor que 1.

4.5. Resultado para el modelo de d palabras aleatorias

En esta sección probaremos que el modelo de d-palabras aleatorias admite una (c, λ, θ)-mediana, lo cual, gracias al Teorema Principal nos permitirá obtener estimaciones para la media y la mediana del tamaño del subhipergrafo monótono más grande para este modelo.

Probaremos en particular que la constante c en la definición de (c, λ, θ)-mediana para este modelo corresponde exactamente a la constante de Ulam para el problema de la secuencia creciente más larga en d-dimensiones, denotada generalmente como c_d.

4.5.1. Problema de la secuencia creciente más larga o problema de Ulam

Dada una permutación $\pi \in S_n$, llamamos secuencia creciente de largo L a toda secuencia $1 \leq i_1 < i_2 < \cdots < i_L \leq n$ tal que $\pi(i_1) < \pi(i_2) < \cdots < \pi(i_L)$.

Ulam [18] en 1961, fue al parecer el primero en proponer la siguiente pregunta: ¿Cuan rápido crece el largo esperado de la secuencia creciente más larga de una permutación en S_n con respecto a n? Por dicha razón, el problema de determinar el comportamiento de dicho largo también es conocido en la literatura como el problema de Ulam.

Denotemos LIS(n) a la variable aleatoria correspondiente al largo de la secuencia creciente más larga de una permutación escogida de manera uniforme en S_n. Usamos esta notación por las siglas en inglés del problema *Longest Increasing Sequence*. Ulam [18] dio un análisis preliminar para determinar el comportamiento de esta variable. Usando simulaciones de Monte Carlo, Ulam notó que el valor esperado de LIS(n), a medida que n crece se aproxima a $2\sqrt{n}$. Hammersley [19] en 1972, dio una demostración rigurosa de la convergencia en probabilidad de $\text{LIS}(n)/\sqrt{n}$ a una constante, conjeturando que dicho valor es 2. Cinco años más tarde, trabajos de Logan y Shepp [20] y de Vershik y Kerov [21] permitieron demostrar la veracidad de dicha conjetura.

Reescribamos el problema anterior de una manera que nos será más útil para nuestros propósitos. Sea $\pi \in S_n$ una permutación y sea $X = X(\pi)$ el conjunto $X(\pi) = \{(i, \pi(i)) \mid 1 \leq i \leq n\}$. Con esto, toda secuencia creciente de π puede verse como una cadena (subconjunto totalmente ordenado) de X usando como orden el orden natural en $\mathbb{N}^2$ (es decir: $(a, b) \leq (c, d) \iff a \leq c \wedge b \leq d$).

Usando la interpretación anterior, podemos extender este concepto a más dimensiones (el caso original, corresponde a 2 dimensiones). Para d un entero positivo, y $n \in \mathbb{N}$, consideremos $d-1$ permutaciones, $\pi_1, \ldots, \pi_{d-1} \in S_n$. Sea además X el conjunto $\{(i, \pi_1(i), \ldots, \pi_{d-1}(i)) \mid 1 \leq i \leq n\}$.

Llamaremos secuencia creciente de largo L de las permutaciones $\pi_1, \ldots, \pi_{d-1}$ a toda cadena de largo L en $(X, \leq)$, donde $(a_1, \ldots, a_d) \leq (b_1, \ldots, b_d) \iff a_i \leq b_i$ para todo i.

Análogamente al caso 2-dimensional, llamemos $\mathrm{LIS}_d(n)$ a la variable aleatoria correspondiente al largo de la secuencia creciente más larga de $d-1$ permutaciones $\pi_1, \ldots, \pi_{d-1}$ elegidas de manera uniforme e independientemente en S_n. El problema de Ulam d-dimensional corresponde entonces a estudiar la distribución de la variable $\mathrm{LIS}_d(n)$.

El problema anterior puede ser visto también como un problema geométrico de la siguiente manera. Sean d y n enteros positivos fijos. Consideremos $\vec{x}(1), \vec{x}(2), \ldots, \vec{x}(n)$ a ser n puntos elegidos de manera independiente y uniforme en el cubo unitario d-dimensional $[0,1]^d$. En este espacio definimos el orden parcial natural $\leq$ como: $\vec{y} \leq \vec{z} \iff y_i \leq z_i$ para todo $1 \leq i \leq d$. Denotamos $H_d(n)$ al largo de la cadena más larga que podemos formar en este orden parcial. Se puede probar que $H_d(n)$ y $LIS_d(n)$ siguen la misma distribución.

Bollobas y Winkler [11] probaron en 1992 que para todo d existe una constante c_d tal que $H_d(n)/\sqrt[d]{n}$ (y luego $\mathrm{LIS}_d(n)/\sqrt[d]{n}$) tiende a c_d en esperanza y en probabilidad cuando $n \to \infty$. Los valores para dichas constantes, denotadas constantes de Ulam, no se conocen salvo por las dos primeras: $c_1 = 1$ y $c_2 = 2$, pero se sabe [11] que para todo d, $c_d < e$ y que $\lim_{d\to\infty} c_d = e$.

4.5.2. Reducción al problema de Ulam

Volvamos a nuestro problema. Sea H un hipergrafo elegido de acuerdo al modelo de d-palabras aleatorias $\Sigma(K_{n_1,n_2,\ldots,n_d}, k)$, y sea H' el subhipergrafo de H obtenido al remover todas las aristas incidentes en nodos de grado 2 o mayor. Denotemos E y E' a $E(H)$ y $E(H')$ respectivamente. Para estimar una mediana de $L(H)$ será necesario más adelante estimar $L(H')$. Notemos, sin embargo, que $L(H')$ es precisamente el largo de la cadena más larga que podemos formar en E' usando el orden natural de aristas. En otras palabras, $L(H')$ es el largo de la secuencia creciente más larga de $d-1$ permutaciones en $\{1, 2, \ldots, |E'|\}$.

La observación anterior nos permitirá utilizar los resultados existentes para el problema de Ulam en el análisis de nuestro problema. En particular, el siguiente teorema de Bollobas y Brightwell [10] de concentración de la secuencia creciente más larga d-dimensional nos será de utilidad.

Teorema 4.9. *Para cada entero $d \geq 2$, existe una constante D_d tal que, para m suficientemente grande,*

$$\mathbb{P}\left(|\,\mathrm{LIS}_d(m) - \mathbb{E}\,\mathrm{LIS}_d(m)| > \frac{\lambda D_d m^{1/2d} \log m}{\log\log m} \right) \leq 80\lambda^2 e^{-\lambda^2}$$

para todo λ con $2 < \lambda < m^{1/2d}/\log\log m$.

No usaremos directamente el teorema, sino que demostraremos y usaremos el siguiente corolario:

Corolario 4.10. *Para todo entero $d \geq 2$ y todo $t > 0$ y $\alpha > 0$, existe un $m_0(t, \alpha, d)$ suficientemente*

grande tal que si $m \geq m_0$, entonces

$$\mathbb{P}\left(|\operatorname{LIS}_d(m) - c_d m^{1/d}| > t c_d m^{1/d}\right) \leq \alpha.$$

con c_d la constante para el problema de Ulam d-dimensional.

Demostración. Sean d, t y α como en la hipótesis y sea D_d la constante dada por el Teorema 4.9. De la definición de la constante de Ulam [11], sabemos que $\lim_{n\to\infty} \mathbb{E}\operatorname{LIS}_d(m)/\sqrt[d]{m} = c_d$. Luego, podemos elegir un $m_0 = m_0(t, \alpha, d)$ suficientemente grande tal que para todo $m \geq m_0$ las siguientes condiciones se cumplan:

1. $|\mathbb{E}\operatorname{LIS}_d(m) - c_d m^{1/d}| < t c_d m^{1/d}/2$.
2. $\lambda = \lambda(m) \stackrel{\text{def}}{=} \dfrac{t c_d}{2 D_d} \cdot \dfrac{m^{1/2d} \log\log m}{\log m} \leq \dfrac{m^{1/2d}}{\log\log m}$.
3. $80\lambda^2 e^{-\lambda^2} \leq \alpha$
4. El Teorema 4.9 es válido para m.

Estas condiciones se puede obtener pues $(\log\log m)^2 = o(\log m)$ y $\lim_{m\to\infty} \lambda(m) = \infty$.

Sigue que, para todo $m > m_0$.

$$\begin{aligned}
&\mathbb{P}\left(|\operatorname{LIS}_d(m) - c_d m^{1/d}| > t c_d m^{1/d}\right) \\
&\quad\leq \mathbb{P}\left(|\operatorname{LIS}_d(m) - \mathbb{E}\operatorname{LIS}_d(m)| + |\mathbb{E}\operatorname{LIS}_d(m) - c_d m^{1/d}| > t c_d m^{1/d}\right) \\
&\quad\leq \mathbb{P}\left(|\operatorname{LIS}_d(m) - \mathbb{E}\operatorname{LIS}_d(m)| > \frac{t c_d m^{1/d}}{2}\right) \\
&\quad= \mathbb{P}\left(|\operatorname{LIS}_d(m) - \mathbb{E}\operatorname{LIS}_d(m)| > \frac{\lambda D_d m^{1/2d} \log m}{\log\log m}\right) \\
&\quad\leq 80\lambda^2 e^{-\lambda^2} \leq \alpha.
\end{aligned}$$

■

4.5.3. Aplicación del Teorema Principal al modelo de d-palabras aleatorias

Como ya se dijo al principio de la sección, demostraremos que el modelo de d-palabras aleatorias admite una (c, λ, θ)-mediana. Para ello necesitaremos un par de lemas que nos permitirán estimar $L(H)$.

Lema 4.11 (Desigualdad de Chebyshev para indicatrices)**.** *Sean $X_1, \ldots, X_m$ variables aleatorias que sólo toman valores 0 y 1 y sea $X = \sum_{i=1}^m X_m$ su suma. Sea además $\Delta = \sum_{(i,j):i\neq j} \mathbb{E}[X_i X_j]$. Para todo $t \geq 0$,*

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq \frac{1}{t^2}\left(\mathbb{E}[X](1 - \mathbb{E}[X]) + \Delta\right)$$

Demostración. Por desigualdad de Chebyshev, $\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq \mathrm{Var}[X]/t^2$. Además,

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \sum_{i=1}^{m}\sum_{j=1}^{m} \mathbb{E}[X_i X_j] - \mathbb{E}[X]^2 = \Delta + \sum_{i=1}^{m} \mathbb{E}[X_i^2] - \mathbb{E}[X]^2.$$

Como los X_i son tales que $X_i^2 = X_i$, lo anterior es igual a $\Delta + \mathbb{E}[X] - \mathbb{E}[X]^2$, con lo que se concluye la demostración. ■

Lema 4.12. *Si $n_1, \ldots, n_d$ son los largos de las clases de H, y llamamos $S = \sum_{i=1}^{d} n_i$ a su suma y $N = \left(\prod_{i=1}^{d} n_i\right)^{1/d}$ a su promedio geométrico, entonces:*

$$\mathbb{E}[|E|] = \frac{N^d}{k^{d-1}}. \tag{4.5.1}$$

$$\mathbb{E}[|E'|] = \frac{N^d}{k^{d-1}}\left(\frac{k-1}{k}\right)^{S-d} \geq \frac{N^d}{k^{d-1}}\left(1 - \frac{S}{k}\right). \tag{4.5.2}$$

$$\mathbb{E}[|E \setminus E'|] \leq \frac{N^d S}{k^d}. \tag{4.5.3}$$

Además, para todo $\eta > 0$,

$$\mathbb{P}[|E' - \mathbb{E}[|E'|]| \geq \eta \mathbb{E}[|E'|]] \leq \frac{1}{\eta^2 \mathbb{E}[|E'|]} + \frac{1}{\eta^2}\left(\left(\frac{k-1}{k-2}\right)^{2d-1} - 1\right). \tag{4.5.4}$$

Demostración. Sea $K = K_{n_1,\ldots,n_d}$, y para cada arista $e \in E(K)$, llamemos X_e e Y_e a las indicatrices de los eventos $e \in E$ y $e \in E'$ respectivamente. Con esto,

$$|E| = \sum_{e \in E(K)} X_e \qquad \text{y} \qquad |E'| = \sum_{e \in E(K)} Y_e.$$

Además,

$$\mathbb{E}[X_e] = k\left(\frac{1}{k}\right)^d = \frac{1}{k^{d-1}}.$$
$$\mathbb{E}[Y_e] = k\left(\frac{1}{k}\right)^d\left(1 - \frac{1}{k}\right)^{S-d} = \frac{1}{k^{d-1}}\left(\frac{k-1}{k}\right)^{S-d}.$$

Notando que $|E(K)| = \prod_{i=1}^{d} n_i = N^d$ se concluye (4.5.1) y la primera igualdad en (4.5.2).

Por otro lado, usando una desigualdad de Bernoulli,

$$\mathbb{E}[|E'|] = \frac{N^d}{k^{d-1}}\left(1 - \frac{1}{k}\right)^{S-d} \geq \frac{N^d}{k^{d-1}}\left(1 - \frac{S-d}{k}\right) \geq \frac{N^d}{k^{d-1}}\left(1 - \frac{S}{k}\right)$$

y luego

$$\mathbb{E}[|E \setminus E'|] = \mathbb{E}[|E|] - \mathbb{E}[|E'|] \leq \frac{N^d S}{k^d}.$$

Esto concluye (4.5.2) y (4.5.3).

Para la última parte del lema, sean e y f dos aristas distintas de $E(K)$. Tenemos dos casos:

Caso 1. Si $e \cap f \neq \varnothing$, entonces e y f no pueden aparecer juntas en E', por lo tanto $\mathbb{E}(Y_e Y_f) = 0$.

Caso 2. Si $e \cap f = \varnothing$ entonces

$$\begin{aligned}\mathbb{E}(Y_e Y_f) = \mathbb{P}(e \in E', f \in E') &= \sum_{i=1}^{k} \sum_{j \neq i} \left(\frac{1}{k}\right)^{2d} \left(1 - \frac{2}{k}\right)^{S-2d} \\ &= \frac{k(k-1)(k-2)^{S-2d}}{k^S} = \frac{(k-1)^{2S-2d}}{k^{2S-2}} \cdot \frac{k^{S-1}(k-2)^{S-2d}}{(k-1)^{2S-2d-1}} \\ &= \mathbb{E}(Y_e)\mathbb{E}(Y_f) \cdot \left(\frac{k(k-2)}{(k-1)^2}\right)^{S-1} \cdot \left(\frac{k-1}{k-2}\right)^{2d-1}.\end{aligned}$$

Notando que $k(k-2) \leq (k-1)^2$ para todo k, lo anterior es menor o igual que:

$$\mathbb{E}(Y_e)\mathbb{E}(Y_f) \cdot \left(\frac{k-1}{k-2}\right)^{2d-1}.$$

Del análisis anterior se deduce que :

$$\Delta \stackrel{\text{def}}{=} \sum_{(e,f):e \neq f} \mathbb{E}(Y_e Y_f) \leq |(e,f) : e \cap f = \varnothing| \cdot \left(\frac{\mathbb{E}[|E'|]}{N^d}\right)^2 \left(\frac{k-1}{k-2}\right)^{2d-1} \leq \mathbb{E}[|E'|]^2 \left(\frac{k-1}{k-2}\right)^{2d-1}.$$

Aplicando la desigualdad de Chebyshev para indicatrices (Lema 4.11) a las variables Y_e y llamando $Y = |E'|$, se obtiene:

$$\begin{aligned}\mathbb{P}(|Y - \mathbb{E}(Y)| \geq \eta \mathbb{E}(Y)) &\leq \frac{1}{(\eta \mathbb{E}(Y))^2} \left(\mathbb{E}(Y) + \Delta - \mathbb{E}(Y)^2\right) \\ &\leq \frac{1}{\eta^2 \mathbb{E}(Y)} + \frac{1}{\eta^2} \left(\left(\frac{k-1}{k-2}\right)^{2d-1} - 1\right).\end{aligned}$$

■

Gracias a los lemas recién probados podemos mostrar la siguiente estimación para la mediana de $L(H)$ cuando H es elegido de acuerdo a $\Sigma(K_{n_1,n_2,\ldots,n_d}, k)$.

Proposición 4.13. *Sea $\delta > 0$, $d \geq 2$, y un conjunto $n_1, n_2, \ldots, n_d$ de enteros positivos de suma $S = \sum_{i=1}^{d} n_i$ y promedio geométrico $N = \left(\prod_{i=1}^{d} n_i\right)^{1/d}$. Definamos $M = c_d N / k^{1-1/d}$ con c_d la constante de Ulam d-dimensional. Existen constantes $C = C(\delta)$ y $K = K(\delta)$ suficientemente grandes tal que:*

1. *Si* $k \geq K$, $N \geq Ck^{1-1/d}$, $12S^d \leq \delta c_d k^{d-1+1/d}$ *y* $S \leq k/2$, *entonces toda mediana de la variable* $L(\Sigma(K_{n_1,\ldots,n_d}, k))$ *es menor que* $(1+\delta)M$.
2. *Si* $k \geq K$, $n \geq Ck^{1-1/d}$ *y* $S \leq \delta k/2$, *entonces toda mediana de* $L(\Sigma(K_{n_1,\ldots,n_d}, k))$ *es mayor que* $(1-\delta)M$.

Demostración. Sea H un hipergrafo elegido de acuerdo a $\Sigma(K_{n_1,\ldots,n_d}, k)$ y asumamos las hipótesis de la proposición. Probar la proposición es equivalente a mostrar que

$$\mathbb{P}[L(H) \geq (1+\delta)M] \leq 1/2. \tag{4.5.5}$$

y que

$$\mathbb{P}[L(H) \leq (1-\delta)M] \leq 1/2. \tag{4.5.6}$$

Probemos (4.5.5), para ello notemos primero que $L(H) \leq L(H') + |E \setminus E'|$. Con esto:

$$\begin{aligned}
\mathbb{P}\left[L(H) \geq (1+\delta)M\right] &\leq \mathbb{P}\left[L(H') + |E \setminus E'| \geq (1+\delta)M\right] \\
&\leq \mathbb{P}\left[L(H') + |E \setminus E'| \geq (1+\delta)M, |E \setminus E'| \geq M\delta/2\right] \\
&\quad + \mathbb{P}\left[L(H') + |E \setminus E'| \geq (1+\delta)M, |E \setminus E'| < M\delta/2\right] \\
&\leq \mathbb{P}\left[|E \setminus E'| \geq M\delta/2\right] + \mathbb{P}\left[L(H') \geq (1+\delta/2)M\right] \\
&\leq \mathbb{P}\left[|E \setminus E'| \geq M\delta/2\right] + \mathbb{P}\left[|E'| \geq (1+\delta/2)M^d/c_d^d\right] \\
&\quad + \mathbb{P}\left[L(H') \geq (1+\delta/2)M, |E'| < (1+\delta/2)M^d/c_d^d\right].
\end{aligned}$$

Acotaremos los tres términos de la derecha uno a uno. El primer término lo acotaremos usando la desigualdad de Markov, la desigualdad $N < S$ y el Lema 4.12 como sigue:

$$\mathbb{P}\left[|E \setminus E'| \geq \frac{M\delta}{2}\right] \leq \frac{2}{M\delta}\mathbb{E}(|E \setminus E'|) \leq \frac{2k^{1-1/d}}{\delta c_d n} \cdot \frac{N^d S}{k^d} = \frac{2SN^{d-1}}{\delta c_d k^{d-1+1/d}} \leq \frac{2S^d}{\delta c_d k^{d-1+1/d}} \leq \frac{1}{6}.$$

El segundo término lo acotamos usando el Lema 4.12, obteniendo:

$$\begin{aligned}
\mathbb{P}\left[|E'| \geq (1+\delta/2)M^d/c_d^d\right] &\leq \mathbb{P}\left[|E'| \geq (1+\delta/2)\mathbb{E}[|E'|]\right] \\
&\leq \frac{4}{\delta^2 \mathbb{E}[|E'|]} + \frac{4}{\delta^2}\left(\left(\frac{k-1}{k-2}\right)^{2d-1} - 1\right).
\end{aligned}$$

Recordando que $S \leq k/2$ e imponiendo $K = K(\delta)$ suficientemente grande lo anterior se puede hacer menor que:

$$\frac{4k^{d-1}}{\delta^2 N^d(1-S/k)} + \frac{4}{\delta^2}(48\delta^2) \leq \frac{8k^{d-1}}{\delta^2 N^d} + \frac{1}{12} \leq \frac{8}{\delta^2 C^d} + \frac{1}{12}.$$

Tomando $C^d \geq 96/\delta^2$, lo anterior es menor o igual que $1/6$.

Para el tercer término, consideremos $m = \lfloor (1+\delta/2)M^d/c_d^d \rfloor$. Usando la desigualdad de Bernoulli $(1+x)^a \leq 1+ax$ para $x \geq -1$ y $0 < a < 1$, se tiene:

$$\begin{aligned}
\mathbb{P}\left[L(H') \geq (1+\delta/2)M, |E'| < (1+\delta/2)M^d/c_d^d\right] &\leq \mathbb{P}\left[\mathrm{LIS}_d(m) \geq (1+\delta/2)M\right] \\
&\leq \mathbb{P}\left[\mathrm{LIS}_d(m) \geq \frac{(1+\delta/2)c_d m^{1/d}}{(1+\delta/2)^{1/d}}\right] \\
&\leq \mathbb{P}\left[\mathrm{LIS}_d(m) \geq \frac{1+\delta/2}{1+\delta/2d} c_d m^{1/d}\right] \\
&\leq \mathbb{P}\left[\mathrm{LIS}_d(m) \geq \left(1+\frac{(d-1)\delta}{2d+\delta}\right) c_d m^{1/d}\right].
\end{aligned}$$

Necesitamos que N sea suficientemente grande para poder aplicar el Corolario 4.10. Específicamente si llamamos $t = (d-1)\delta/(2d+\delta)$ e imponemos $C^d \geq m_0 + 1$, con $m_0 = m_0(t, 1/6, d)$ como en el corolario 4.10, se tendrá que:

$$m = \lfloor (1+\delta/2)M^d/c_d^d \rfloor = \lfloor (1+\delta/2)N^d/k^{d-1} \rfloor \geq \lfloor C^d \rfloor \geq m_0$$

y luego, usando la conclusión de dicho corolario,

$$\mathbb{P}\left[L(H') \geq (1+\delta/2)M, |E'| < (1+\delta/2)M^d/c_d^d\right] \leq 1/6.$$

En resumen,

$$\mathbb{P}[L(H) \geq (1+\delta)M] \leq \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2},$$

que es lo que queríamos probar.

Probemos ahora (4.5.6). Para ello notemos primero que si $\delta \geq 1$, el resultado se tiene pues toda mediana de $L(H)$ es positiva. Luego asumamos en esta parte que $\delta < 1$. Como $L(H') \leq L(H)$, se tiene que

$$\begin{aligned}
\mathbb{P}\left[L(H) \leq (1-\delta)M\right] &\leq \mathbb{P}\left[L(H') \leq (1-\delta)M, |E'| \leq (1-\delta)M^d/c_d^d\right] \\
&\quad + \mathbb{P}\left[L(H') \leq (1-\delta)M, |E'| > (1-\delta)M^d/c_d^d\right] \\
&\leq \mathbb{P}\left[|E'| \leq (1-\delta)M^d/c_d^d\right] \\
&\quad + \mathbb{P}\left[L(H') \leq (1-\delta)M, |E'| > (1-\delta)M^d/c_d^d\right].
\end{aligned}$$

Al igual que antes, acotamos los términos por separado. El primer término lo acotamos de manera similar al segundo término para el caso anterior, obteniendo:

$$\begin{aligned}
\mathbb{P}\left[|E'| \leq (1-\delta)M^d/c_d^d\right] = \mathbb{P}\left[|E'| \leq (1-\delta)N^d/k^{d-1}\right] &\leq \mathbb{P}\left[|E'| \leq \frac{1-\delta}{1-S/k}\mathbb{E}[|E'|]\right] \\
&\leq \mathbb{P}\left[|E'| \leq \frac{1-\delta}{1-\delta/2}\mathbb{E}[|E'|]\right] \leq \mathbb{P}\left[|E'| \leq \left(1-\frac{\delta}{2}\right)\mathbb{E}[|E'|]\right] \\
&\leq \frac{4}{\delta^2\mathbb{E}[|E'|]} + \frac{4}{\delta^2}\left(\left(\frac{k-1}{k-2}\right)^{2d-1} - 1\right).
\end{aligned}$$

Recordando que $S \leq \delta k/2 < k/2$ y que podemos tomar $K = K(\delta)$ suficientemente grande lo anterior se puede hacer menor que:

$$\frac{4k^{d-1}}{\delta^2 N^d(1-S/k)} + \frac{4}{\delta^2}(32\delta^2) \leq \frac{8k^{d-1}}{\delta^2 N^d} + \frac{1}{8}.$$

Tomando $C^d \geq 64/\delta^2$, lo anterior es menor que $1/4$.

Para el segundo término, tomemos $m = \lceil (1-\delta)M^d/c_d^d \rceil$. Usando una desigualdad de Bernoulli,

$$\begin{aligned}\mathbb{P}\left[L(H') \leq (1-\delta)M, |E'| > (1-\delta)M^d/c_d^d\right] &\leq \mathbb{P}\left[\mathrm{LIS}_d(m) \leq (1-\delta)M\right] \\ &= \mathbb{P}\left[\mathrm{LIS}_d(m) \leq (1-\delta)^{1-1/d} c_d m^{1/d}\right] \\ &\leq \mathbb{P}\left[\mathrm{LIS}_d(m) \leq (1-(1-1/d)\delta) c_d m^{1/d}\right].\end{aligned}$$

Ahora necesitamos que m sea suficientemente grande para poder aplicar el Corolario 4.10. Específicamente si llamamos $t = (1-1/d)\delta$ e imponemos $C \geq \left(m_0/(1-\delta)\right)^{1/d}$ con $m_0 = m_0(t, 1/4, d)$ como en el corolario 4.10, se tendrá que:

$$m \geq (1-\delta)M^d/c_d^d = (1-\delta)N^d/k^{d-1} \geq (1-\delta)C^d \geq m_0.$$

Y luego, usando la conclusión de dicho corolario,

$$\mathbb{P}\left[L(H') \leq (1-\delta)M, |E'| > (1-\delta)M^d/c_d^d\right] \leq 1/4.$$

Resumiendo:

$$\mathbb{P}[L(H) \leq (1-\delta)M] \leq \frac{1}{4} + \frac{1}{4} = \frac{1}{2},$$

lo que concluye la demostración. ■

Gracias a la proposición anterior tenemos el siguiente corolario:

Corolario 4.14. *El modelo Σ de d-palabras aleatorias de parámetro interno k admite una (c, λ, θ)-mediana, con*

$$(c, \lambda, \theta) = (c_d, 1-1/d, 1-1/d+1/d^2),$$

donde c_d la constante para el problema de Ulam d-dimensional.

Demostración. Sean $n_1, \ldots, n_d$ con promedio geométrico N y suma S a especificar más adelante, y sea H elegido de acuerdo a $\Sigma(K_{n_1,\ldots,n_d}, k)$. Sean además $M = cN/k^\lambda = c_d N/k^{1-1/d}$, $\delta > 0$, $C(\delta)$ y $K(\delta)$ como en la proposición anterior.

Definamos $a(\delta) = C(\delta)$, $b(\delta) = (12/(\delta c_d))^{1/d}$ y $k'(\delta) > K(\delta)$ suficientemente grande de modo que para todo $k > k'(\delta)$, $k^{1-1/d+1/d^2} < \min\{\delta/2, 1/2\} b(\delta) k$. Con esto, si $k > k'(\delta)$, y se satisfacen las cotas de tamaño para estas constantes, es decir, $N \geq a(\delta)k^{1-1/d}$ y $Sb(\delta) \leq k^{1-1/d+1/d^2}$, se tendrán todas las hipótesis de la proposición anterior. Con esto toda mediana de $L(H)$ estará entre $(1-\delta)M$ y $(1+\delta)M$. ■

Gracias a este último corolario podemos aplicar el Teorema Principal al modelo de d palabras aleatorias, recordando que en este caso $h = 1/(4d)$ es constante de concentración para el modelo.

Teorema 4.15 (Teorema de estimación para el modelo de d-palabras aleatorias)**.** *Sean $\varepsilon > 0$ y $g : \mathbb{N} \to \mathbb{R}$ una función tal que $g(k) = O(k^\eta)$ para un cierto $0 \leq \eta < 1/d^2$. Existen k_0 y A suficientemente grandes tales que si $k \geq k_0$ y los valores $n_1, \ldots, n_d$ de promedio geométrico N y suma S cumplen*

$$N \geq k^{1-1/d}A, \qquad \text{(Condición de tamaño)}$$
$$S \leq g(k)N, \qquad \text{(Condición de balanceo)}$$

se tiene, definiendo $M = c_d N/k^{1-1/d}$, con c_d la constante de Ulam d-dimensional,

$$(1-\varepsilon)M \leq \mathbb{E}[L(\Sigma(K_{n_1,\ldots,n_d}, k))] \leq (1+\varepsilon)M. \tag{4.5.7}$$

Por otro lado, si $\mathrm{Med}[L(\Sigma(K_{n_1,\ldots,n_d}, k))]$ *es una mediana de $L(\Sigma(K_{n_1,\ldots,n_d}, k))$,*

$$(1-\varepsilon)M \leq \mathrm{Med}[L(\Sigma(K_{n_1,\ldots,n_d}, k))] \leq (1+\varepsilon)M. \tag{4.5.8}$$

Además, existe una constante absoluta $C > 0$ tal que para k y $n_1, \ldots, n_d$ como antes:

$$\mathbb{P}\left[L(\Sigma(K_{n_1,\ldots,n_d}, k)) \leq (1-\varepsilon)M\right] \leq \exp\left(-\frac{C}{d}\varepsilon^2 M\right), \tag{4.5.9}$$

$$\mathbb{P}\left[L(\Sigma(K_{n_1,\ldots,n_d}, k)) \geq (1+\varepsilon)M\right] \leq \exp\left(-\frac{C}{d}\frac{\varepsilon^2}{1+\varepsilon}M\right). \tag{4.5.10}$$

4.6. Resultado para el modelo binomial

En esta sección veremos que, al igual que el modelo anterior, el modelo binomial admite una (c, λ, θ)-mediana y luego podemos aplicar el Teorema Principal en este modelo.

Consideremos en lo que sigue H elegido de acuerdo al modelo binomial $G(K_{n_1,\ldots,n_d}, p)$, y H' el subhipergrafo de H obtenido al remover todas las aristas incidentes en nodos de grado 2 o mayor. Denotemos E y E' a $E(H)$ y $E(H')$. Usando el mismo método de la sección anterior, encontraremos una estimación de la mediana de $L(G(K_{n_1,\ldots,n_d}, p)$. Para ello, necesitaremos los siguientes lemas:

Lema 4.16. *Para todo conjunto de números positivos $n_1, \ldots, n_d$, si definimos $S = \sum_{j=1}^d n_j$, $N = \left(\prod_{j=1}^d n_j\right)^{1/d}$ y $\widetilde{N} = \left(\prod_{j=1}^d (n_j - 1)\right)^d$ entonces $N^d - \widetilde{N}^d \leq S^{d-1}$.*

Demostración. Aplicación directa del Lema 4.5 ■

Lema 4.17. *Si* $S = \sum_{j=1}^{d} n_j$, $N = \left(\prod_{j=1}^{d} n_j\right)^{1/d}$ *y* $\widetilde{N} = \left(\prod_{j=1}^{d}(n_j - 1)\right)^{d}$ *entonces:*

$$\mathbb{E}[|E|] = N^d p, \tag{4.6.1}$$

$$\mathbb{E}[|E'|] = N^d p(1-p)^{N^d - \widetilde{N}^d} \geq N^d p(1 - S^{d-1}p), \tag{4.6.2}$$

$$\mathbb{E}[|E \setminus E'|] \leq N^d S^{d-1} p^2. \tag{4.6.3}$$

Además, para todo $\eta > 0$,

$$\mathbb{P}[|E'| - \mathbb{E}[|E'|]| \geq \eta \mathbb{E}[|E'|]] \leq \frac{1}{\eta^2 \mathbb{E}[|E'|]}. \tag{4.6.4}$$

Demostración. Sea $K = K_{n_1,\ldots,n_k}$, y para cada arista $e \in E(K)$, llamemos X_e e Y_e a las indicatrices de los eventos $e \in E$ y $e \in E'$ respectivamente. Con esto,

$$|E| = \sum_{e \in E(K)} X_e \qquad \text{y} \qquad |E'| = \sum_{e \in E(K)} Y_e.$$

Además, para todo e, se tiene que $\mathbb{E}[X_e] = p$ y $\mathbb{E}[Y_e] = p(1-p)^{N^d - \widetilde{N}^d}$. Notando que $|E(K)| = N^d$ se concluyen (4.6.1) y la primera igualdad en (4.6.2).

Por otro lado, usando una desigualdad de Bernoulli y el lema anterior,

$$\mathbb{E}[|E'|] = N^d p(1-p)^{N^d - \widetilde{N}^d} \geq N^d p(1 - (N^d - \widetilde{N}^d)p) \geq N^d p(1 - S^{d-1}p).$$

Luego,

$$\mathbb{E}[|E \setminus E'|] = \mathbb{E}[|E|] - \mathbb{E}[|E'|] \leq N^d S^{d-1} p^2,$$

lo que concluye (4.6.2) y (4.6.3).

Para la última parte del lema, sean e y f dos aristas distintas de $E(K)$. Tenemos dos casos:

Caso 1. Si $e \cap f \neq \varnothing$, entonces e y f no pueden aparecer ambas en E', por lo tanto $\mathbb{E}(Y_e Y_f) = 0$.

Caso 2. Si $e \cap f = \varnothing$ entonces Y_e e Y_f son independientes y luego,

$$\mathbb{E}(Y_e Y_f) = \mathbb{E}[Y_f]\mathbb{E}[Y_e] = \mathbb{E}[Y_e]^2 = \left(\frac{\mathbb{E}[|E'|]}{N^d}\right)^2.$$

Del análisis anterior,

$$\Delta \stackrel{\text{def}}{=} \sum_{(e,f):e \neq f} \mathbb{E}(Y_e Y_f) = |(e,f) : e \cap f = \varnothing| \cdot \left(\frac{\mathbb{E}[|E'|]}{N^d}\right)^2 \leq \mathbb{E}[|E'|]^2.$$

Aplicando la desigualdad de Chebyshev para indicatrices (Lema 4.11) a las variables Y_e y llamando $Y = |E'|$, se obtiene:

$$\mathbb{P}(|Y - \mathbb{E}(Y)| \geq \eta \mathbb{E}(Y)) \leq \frac{1}{(\eta \mathbb{E}(Y))^2} \left(\mathbb{E}(Y)(1 - \mathbb{E}(Y)) + \Delta\right) \leq \frac{1}{\eta^2 \mathbb{E}(Y)}. \qquad \blacksquare$$

Al igual que en el caso del modelo de d-palabras aleatorias, $L(H')$ corresponde al largo de la secuencia creciente más larga de $d-1$ permutaciones en $\{1,2,\ldots,|E'|\}$ y luego podemos usar el Corolario 4.10 para estimar dicha cantidad. Gracias a esto, podemos demostrar la siguiente cota para la mediana.

Proposición 4.18. *Sean $\delta > 0$, $d \geq 2$ y un conjunto $n_1, n_2, \ldots, n_d$ de enteros positivos de suma $S = \sum_{i=1}^{d} n_i$, y promedio geométrico $N = \left(\prod_{i=1}^{d} n_i\right)^{1/d}$. Definamos $M = c_d N p^{1/d}$ con c_d la constante de Ulam d-dimensional. Existe una constante $C = C(\delta)$ suficientemente grande tal que:*

1. *Si $Np^{1/d} \geq C$, $12S^{2d-2}p^{2-1/d} \leq \delta c_d$ y $S^{d-1}p \leq 1/2$, entonces toda mediana de $L(H)$ es menor que $(1+\delta)M$.*
2. *Si $Np^{1/d} \geq C$ y $S^{d-1}p \leq \delta/2$, entonces toda mediana de $L(H)$ es mayor que $(1-\delta)M$.*

Demostración. La demostración de esta proposición es muy similar a la de la Proposición 4.13. Al igual que entonces, probar la proposición es equivalente a mostrar que

$$\mathbb{P}[L(H) \geq (1+\delta)M)] \leq 1/2 \tag{4.6.5}$$

y que

$$\mathbb{P}[L(H) \leq (1-\delta)M)] \leq 1/2. \tag{4.6.6}$$

Para probar (4.6.5), notemos primero que $L(H) \leq L(H') + |E \setminus E'|$. Con esto:

$$\begin{aligned}
\mathbb{P}\left[L(H) \geq (1+\delta)M\right] &\leq \mathbb{P}\left[L(H') + |E \setminus E'| \geq (1+\delta)M\right] \\
&= \mathbb{P}\left[L(H') + |E \setminus E'| \geq (1+\delta)M, |E \setminus E'| \geq M\delta/2\right] \\
&\quad + \mathbb{P}\left[L(H') + |E \setminus E'| \geq (1+\delta)M, |E \setminus E'| < M\delta/2\right] \\
&\leq \mathbb{P}\left[|E \setminus E'| \geq M\delta/2\right] + \mathbb{P}\left[L(H') \geq (1+\delta/2)M\right] \\
&\leq \mathbb{P}\left[|E \setminus E'| \geq M\delta/2\right] + \mathbb{P}\left[|E'| \geq (1+\delta/2)M^d/c_d^d\right] \\
&\quad + \mathbb{P}\left[L(H') \geq (1+\delta/2)M, |E'| < (1+\delta/2)M^d/c_d^d\right].
\end{aligned}$$

Acotaremos los tres términos de la derecha uno a uno. El primer término lo acotamos usando la desigualdad de Markov, la desigualdad $N < S$ y el Lema 4.12 como sigue:

$$\begin{aligned}
\mathbb{P}\left[|E \setminus E'| \geq \frac{M\delta}{2}\right] &\leq \frac{2}{M\delta}\mathbb{E}(|E \setminus E'|) \leq \frac{2p^2 S^{d-1} N^d}{\delta c_d N p^{1/d}} \\
&= \frac{2p^{2-1/d} S^{d-1} N^{d-1}}{\delta c_d} \leq \frac{2p^{2-1/d} S^{2d-2}}{\delta c_d} \leq \frac{1}{6}.
\end{aligned}$$

El segundo término lo acotamos usando el Lema 4.12 y la desigualdad de Bernoulli $(1+x)^a \geq 1+ax$ para $x \geq -1$ y $a > 1$, obteniendo:

$$\begin{aligned}\mathbb{P}\left[|E'| \geq (1+\delta/2)M^d/c_d^d\right] &\leq \mathbb{P}\left[|E'| \geq (1+\delta/2)\mathbb{E}[|E'|]\right] \\ &\leq \frac{4}{\delta^2\mathbb{E}(|E'|)} \leq \frac{4}{\delta^2 N^d p(1-S^{d-1}p)} \leq \frac{8}{\delta^2 N^d p}.\end{aligned}$$

Tomando $C^d \geq 48/\delta^2$, lo anterior es menor o igual que $1/6$.

Para el tercer término, consideremos $m = \lfloor (1+\delta/2)M^d/c_d^d \rfloor$, de modo que usando la desigualdad de Bernoulli $(1+x)^a \leq 1+ax$ para $x \geq -1$ y $0 < a < 1$, se tiene:

$$\begin{aligned}\mathbb{P}\left[L(H') \geq (1+\delta/2)M, |E'| < (1+\delta/2)M^d/c_d^d\right] &\leq \mathbb{P}\left[\mathrm{LIS}_d(m) \geq (1+\delta/2)M\right] \\ &\leq \mathbb{P}\left[\mathrm{LIS}_d(m) \geq \frac{(1+\delta/2)c_d m^{1/d}}{(1+\delta/2)^{1/d}}\right] \\ &\leq \mathbb{P}\left[\mathrm{LIS}_d(m) \geq \frac{1+\delta/2}{1+\delta/(2d)} c_d m^{1/d}\right] \\ &= \mathbb{P}\left[\mathrm{LIS}_d(m) \geq \left(1+\frac{(d-1)\delta}{2d+\delta}\right) c_d m^{1/d}\right].\end{aligned}$$

Necesitamos que m sea suficientemente grande para poder aplicar el Corolario 4.10. Específicamente si llamamos $t = (d-1)\delta/(2d+\delta)$ e imponemos $C^d \geq m_0 + 1$ con $m_0 = m_0(t, 1/6, d)$ como en el Corolario 4.10, se tendrá que:

$$m = \lfloor (1+\delta/2)M^d/c_d^d \rfloor = \lfloor (1+\delta/2)N^d p \rfloor \geq \lfloor C^d \rfloor \geq m_0.$$

Y luego, usando la conclusión de dicho corolario,

$$\mathbb{P}\left[L(H') \geq (1+\delta/2)M, |E'| < (1+\delta/2)M^d/c_d^d\right] \leq 1/6.$$

Resumiendo,

$$\mathbb{P}[L(H) \geq (1+\delta)M] \leq \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2},$$

que es lo que queríamos probar.

Probemos ahora (4.6.6). Para ello notemos primero que si $\delta \geq 1$, el resultado se tiene pues toda mediana de $L(H)$ es positiva. Luego asumamos en esta parte que $\delta < 1$.

Como $L(H') \leq L(H)$,

$$\begin{aligned}\mathbb{P}\left[L(H) \leq (1-\delta)M\right] &= \mathbb{P}\left[L(H') \leq (1-\delta)M, |E'| \leq (1-\delta)M^d/c_d^d\right] \\ &\quad + \mathbb{P}\left[L(H') \leq (1-\delta)M, |E'| > (1-\delta)M^d/c_d^d\right] \\ &\leq \mathbb{P}\left[|E'| \leq (1-\delta)M^d/c_d^d\right] \\ &\quad + \mathbb{P}\left[L(H') \leq (1-\delta)M, |E'| > (1-\delta)M^d/c_d^d\right].\end{aligned}$$

Al igual que antes, acotamos los términos por separado. El primer término lo acotamos de manera similar al segundo término para el caso anterior, obteniendo:

$$\begin{aligned}\mathbb{P}\left[|E'| \leq (1-\delta)M^d/c_d^d\right] &= \mathbb{P}\left[|E'| \leq (1-\delta)N^d p\right] \leq \mathbb{P}\left[|E'| \leq \frac{(1-\delta)}{(1-S^{d-1}p)}\mathbb{E}[|E'|]\right] \\ &\leq \mathbb{P}\left[|E'| \leq \frac{1-\delta}{1-\delta/2}\mathbb{E}[|E'|]\right] \leq \mathbb{P}\left[|E'| \leq (1-\delta/2)\mathbb{E}[|E'|]\right] \\ &\leq \frac{4}{\delta^2\mathbb{E}[|E'|]} \leq \frac{8}{\delta^2 N^d p}.\end{aligned}$$

Tomando $C^d \geq 32/\delta^2$, lo anterior es menor que $1/4$.

Para el segundo término, tomemos $m = \lceil (1-\delta)M^d/c_d^d \rceil$, luego usando la desigualdad de Bernoulli,

$$\begin{aligned}\mathbb{P}\left[L(H') \leq (1-\delta)M, |E'| > (1-\delta)M^d/c_d^d\right] &\leq \mathbb{P}\left[\mathrm{LIS}_d(m) \leq M(1-\delta)\right] \\ &= \mathbb{P}\left[\mathrm{LIS}_d(m) \leq (1-\delta)^{1-1/d} c_d m^{1/d}\right] \\ &\leq \mathbb{P}\left[\mathrm{LIS}_d(m) \leq (1-(1-1/d)\delta) c_d m^{1/d}\right].\end{aligned}$$

Ahora necesitamos que m sea suficientemente grande para poder aplicar el Corolario 4.10. Específicamente si llamamos $t = (1-1/d)\delta$ e imponemos $C > (m_0/(1-\delta))^{1/d}$ con $m_0 = m_0(t, 1/4, d)$ como en el corolario 4.10, se tendrá que:

$$m \geq (1-\delta)M^d/c_d^d = (1-\delta)N^d p \geq (1-\delta)C^d \geq m_0.$$

Y luego, usando la conclusión de dicho corolario,

$$\mathbb{P}\left[L(H') \leq (1-\delta)M, |E'| > (1-\delta)M^d/c_d^d\right] \leq 1/4.$$

Resumiendo:

$$\mathbb{P}[L(H) \leq (1-\delta)M] \leq \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

Esto concluye la demostración. ■

Gracias a la proposición anterior tenemos el siguiente corolario,

Corolario 4.19. *Si definimos $t = 1/p$, entonces el modelo $G(K_{n_1,\ldots,n_d}, p)$, de parametro interno t admite una (c, λ, θ)-mediana con*

$$(c, \lambda, \theta) = \left(c_d, \frac{1}{d}, \frac{2d-1}{2d(d-1)}\right).$$

Demostración. Sean $n_1, \ldots, n_d$ con promedio geométrico N y suma S a especificar más adelante, y sea H elegido de acuerdo a $G(K_{n_1,\ldots,n_d}, p)$. Sean además $M = cN/t^\lambda = c_d N p^{1/d}$, $\delta > 0$ y $C(\delta)$ como en la proposición anterior.

Definamos ahora $a(\delta) = C(\delta)$, $b(\delta) = (12/(\delta c_d))^{1/(2d-2)}$ y $t'(\delta)$ suficientemente grande de modo que para todo $t > t'(\delta)$, $t^{1-1/(2d)} < \min\{\delta/2, 1/2\} t b(\delta)^{d-1}$.

Con esto, si $t > t'(\delta)$, $n \geq a(\delta) t^{1/d}$ y $Sb(\delta) \leq t^{(2d-1)/(2d(d-1))}$, se tendrán todas las hipótesis de la proposición anterior y luego, toda mediana de $L(H)$ estará entre $(1-\delta)M$ y $(1+\delta)M$. ■

Análogamente al modelo anterior y recordando que $h = 1/4$ es constante de concentración para el modelo binomial, tenemos el siguiente resultado:

Teorema 4.20 (Teorema de estimación para el modelo binomial)**.** *Sea $\varepsilon > 0$ y $g : \mathbb{R} \to \mathbb{R}$ una función tal que $g(t) = O(t^\eta)$ para un cierto $0 \leq \eta < 1/(2d(d-1))$. Existen p_0 suficientemente pequeño y A suficientemente grande tales que si $p \leq p_0$ y los valores $n_1, \ldots, n_d$ de promedio geométrico N y suma S cumplen*

$$Np^{1/d} \geq A. \qquad \text{(Condición de tamaño)}$$
$$S \leq g(1/p)N. \qquad \text{(Condición de balanceo)}$$

se tiene, definiendo $M = c_d N p^{1/d}$ con c_d la constante de Ulam d-dimensional,

$$(1-\varepsilon)M \leq \mathbb{E}[L(G(K_{n_1,\ldots,n_d}, p))] \leq (1+\varepsilon)M. \tag{4.6.7}$$

Por otro lado, si $\mathrm{Med}[L(G(K_{n_1,\ldots,n_d}, p))]$ *es una mediana de $L(G(K_{n_1,\ldots,n_d}, p))$,*

$$(1-\varepsilon)M \leq \mathrm{Med}[L(G(K_{n_1,\ldots,n_d}, p))] \leq (1+\varepsilon)M. \tag{4.6.8}$$

Además, existe una constante $C > 0$ absoluta tal que para p y $n_1, \ldots, n_d$ como antes:

$$\mathbb{P}\left[L(G(K_{n_1,\ldots,n_d}, p) \leq (1-\varepsilon)M\right] \leq \exp\left(-C\varepsilon^2 M\right), \tag{4.6.9}$$

$$\mathbb{P}\left[L(G(K_{n_1,\ldots,n_d}, p) \geq (1+\varepsilon)M\right] \leq \exp\left(-C\frac{\varepsilon^2}{1+\varepsilon}M\right). \tag{4.6.10}$$

Capítulo 5

Variantes simétricas del problema de la LCS

En este capítulo describiremos dos variantes del problema del subhipergrafo monótono más grande estudiado en el capítulo anterior, un modelo que denominaremos modelo simétrico y otro que llamaremos modelo antisimétrico. En ambas variantes consideraremos que d, el número de clases de partición del hipergrafo a considerar, es 2, es decir, nos restringiremos al caso de grafos bipartitos.

En ambos modelos el espacio muestral de grafos ya no será, como lo era en el caso general, el conjunto de todos los subgrafos de un grafo bipartito completo sino que será el conjunto de todos los subgrafos que presentan algún tipo de simetría en sus conjuntos de arcos.

Veremos que restringirnos a estos casos simétricos no cambia el comportamiento asintótico del largo de su subgrafo monótono más grande. De hecho, probaremos que el teorema de estimación que presenta el modelo binomial de 2-hipergrafo se mantiene sin cambios para estos nuevos modelos.

5.1. Modelo simétrico

Consideremos la misma notación usada en el problema del subhipergrafo monótono más grande, pero restringiéndolo al caso de dos palabras, esta vez de igual tamaño.

Sean A y B dos conjuntos disjuntos de tamaño n. Asumiremos, al igual que en el capítulo anterior, que los elementos de ambos conjuntos están numerados de 1 a n y que G es un grafo bipartito con clases de partición A y B, donde identificamos $E(G)$ con el subconjunto de $A \times B$ que contiene los pares ordenados asociados a las aristas de G. Además, usaremos la notación habitual $K_{n,n}$ para denotar al grafo bipartito completo de clases de partición A y B.

Denotemos $S(K_{n,n}, p)$ a la distribución sobre todos los subgrafos de $K_{n,n}$ donde para todo $x < y$,

el evento $\{(x,y),(y,x)\} \subseteq E(G)$, con G una realización, tiene probabilidad p, y todos estos eventos son independientes.

De la definición anterior, cada vez que el arco (x,y) está en una realización, su simétrico (y,x) también lo está. Por esta razón llamaremos a esta distribución, distribución simétrica bipartita de parámetro p.

Nos será de utilidad definir otra distribución relacionada con la anterior. Para ello, sea ahora $K^*_{n,n}$ el grafo bipartito con clases de partición A y B, donde sólo los arcos (x,y) con $x < y$ están presentes y denotemos $O(K^*_{n,n},p)$ a la distribución sobre todos los subgrafos de $K^*_{n,n}$ donde la probabilidad de que un arco (x,y), con $x < y$, esté en el subgrafo es p y todos los eventos son independientes. Llamamos a esta distribución, distribución orientada bipartita de parámetro p.

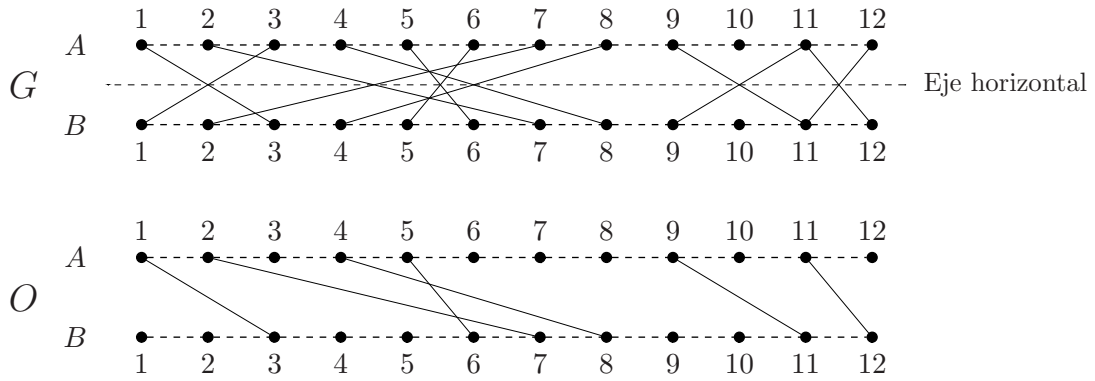

Figura 5.1: Arriba se muestra una representación de un grafo G elegido de acuerdo a la distribución simétrica bipartita $S(K_{12,12},p)$. Intuitivamente un grafo es simétrico si su representación gráfica resulta ser simétrica con respecto a un eje horizontal imaginario ubicado entre ambas clases de partición. Al quedarnos solo con las aristas del tipo (x,y) con $x < y$ obtenemos un grafo que se distribuye de acuerdo al modelo orientado bipartito $O(K^*_{12,12},p)$. Dicho grafo se encuentra representado en la zona inferior de la imagen.

Al igual que en el capítulo anterior, para G un subgrafo de $K_{n,n}$, denotaremos $L(G)$ al largo de su subgrafo monótono más grande. Tenemos el siguiente lema que relaciona las distribuciones simétrica y orientada:

Lema 5.1. *$L(O(K^*_{n,n},p))$ tiene la misma distribución que $L(S(K_{n,n},p))$.*

Demostración. Notemos que para todo grafo O elegido de acuerdo a $O(K^*_{n,n},p)$ podemos asociar de manera única un grafo G con el mismo conjunto de vértices y con conjunto de aristas $E(G) = \{(x,y) : (x,y) \in E(O) \text{ ó } (y,x) \in E(O)\}$, dicho grafo G es un subgrafo simétrico de $K_{n,n}$ y su probabilidad bajo $S(K_{n,n},p)$ es exactamente la misma que la probabilidad de O bajo $O(K^*_{n,n},p)$.

Por otro lado si M es un subgrafo monótono de G, entonces existe otro subgrafo monótono de O (y luego de G), digamos N, del mismo número de aristas tal que todas las aristas son del tipo $x < y$. En efecto basta definir $E(N)$ como $\{(\text{mín}(x,y),\text{máx}(x,y)) \mid (x,y) \in E(M)\}$.

Notemos que como G es simétrico, si $(x, y) \in E(M)$ entonces $((\text{mín}(x, y), \text{máx}(x, y)) \in E(G)$. Luego, N es efectivamente un subgrafo de G. Por otro lado, como M es monótono, si (x, y) está en $E(M)$ entonces (y, x) no puede estarlo, luego N tiene el mismo número de aristas que M. Finalmente es fácil notar que si (x, y) y (z, w) son dos aristas de M (y luego, no se cruzan ni comparten vértices), entonces $(\text{mín}(x, y), \text{máx}(x, y))$ y $(\text{mín}(z, w), \text{máx}(z, w))$ tampoco se cruzan ni comparten vértices, con esto N es monótono. ■

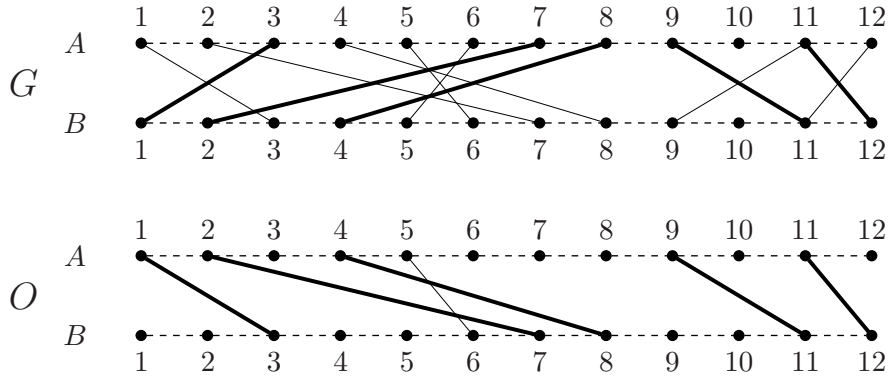

Figura 5.2: Ejemplo ilustrativo del Lema 5.1. El grafo simétrico G indicado arriba es el asociado al grafo monótono O indicado abajo. Las aristas de G indicadas en negrita representan el conjunto de aristas de M, un subgrafo monótono. Las aristas de O indicadas en negrita corresponden a N, el subgrafo monótono derecho asociado a M en el Lema 5.1.

Una observación importante que se obtiene de la demostración anterior es que podemos restringir el estudio de los subgrafos monótonos de grafos simétricos, al caso de subgrafos simétricos orientados, es decir, grafos cuyas aristas son todas del tipo (x, y) con $x < y$.

Queremos obtener un teorema de estimación para el largo del subgrafo monótono de largo máximo de un grafo simétrico. Para ello usaremos el mismo argumento del capítulo anterior.

Comenzamos observando que el modelo simétrico posee una constante de concentración para el largo del subgrafo monótono más grande. Para ello, basta ver que si tenemos un grafo simétrico G y lo modificamos agregando o quitando un par de aristas $\{(u, v), (v, u)\}$ entonces el valor de $L(G)$ cambia en a lo más 1. Además si sabemos que $L(G) \geq r$ entonces podemos mostrar r pares de aristas testigos (las aristas que forman el subgrafo planar junto con sus simétricas) que garantizan que cualquier configuración de pares de aristas que contengan dicho grupo de aristas testigos, darán a lugar un subgrafo planar de tamaño al menos r. Con esto, gracias a la desigualdad de Talagrand, se prueba que el modelo simétrico tiene una constante de concentración $h = 1/4$, es decir, si Med es

una mediana de $L(S(K_{n,n}, p))$,

$$\mathbb{P}\left[L(S(K_{n,n},p) > \mathrm{Med}(1+s)\right] \leq 2\exp\left(-\frac{s^2}{4(1+s)}\right),$$
$$\mathbb{P}\left[L(S(K_{n,n},p) < \mathrm{Med}(1-s)\right] \leq 2\exp\left(-\frac{s^2}{4\,\mathrm{Med}}\right).$$

Luego, al igual que antes, para conseguir una cota de concentración y esperanza, encontraremos una mediana ajustada para este modelo.

5.1.1. Reducción al problema de la secuencia creciente más larga de una involución

Sea G un grafo elegido de acuerdo a $S(K_{n,n}; p)$ y sea G' el grafo obtenido a partir de G eliminando los arcos incidentes en vértices de grado 2 o más. Denotemos $E(G)$ y $E(G')$ por E y E' respectivamente. Notemos que el grafo G' resultante sigue siendo simétrico, en el sentido que si $(a, b) \in E'$, entonces $(b, a) \in E'$. Llamemos $2m$ al número de vértices de la clase de partición A que tiene grado de incidencia 1 en el grafo G' (esta cantidad debe ser par, pues el grafo es simétrico). Análogamente B posee $2m$ vértices con grado de incidencia 1 en G'.

Si nos quedamos sólo con dichos vértices (pues el resto tiene grado 0 en G'), podemos ver los arcos de G' como una involución (permutación autoinversa) de $[2m]$ sin puntos fijos. De hecho, como la distribución de G es invariante bajo permutaciones de los nodos, la distribución de G' también lo es, y luego la involución resultante es arbitraria y uniformemente elegida dentro de todas las involuciones de $[2m]$ sin puntos fijos.

Definamos $\mathcal{I}_{2m}$ como la distribución uniforme sobre todas las involuciones de $[2m]$ sin puntos fijos y denotemos $L(\mathcal{I}_{2m})$ al largo de su secuencia creciente más larga. Notemos que $L(\mathcal{I}_{2m})$ se distribuye como $L(G')$ en este caso. Es sabido [17], que el largo esperado de una involución aleatoria de $[2m]$ sin puntos fijos se comporta asintóticamente como $2\sqrt{2m}$ y de hecho, disponemos de un teorema de concentración para $L(\mathcal{I}_{2m})$ de Kiwi [14, Teorema 5] (Esta es una versión mas débil de su resultado pero es todo cuanto necesitaremos):

Teorema 5.2. *Para m suficientemente grande y todo $0 \leq s \leq 2\sqrt{2m}$,*

$$\mathbb{P}\left[|L(\mathcal{I}_{2m}) - \mathbb{E}(L(\mathcal{I}_{2m}))| \geq s + 32(2m)^{1/4}\right] \leq 4e^{-s^2/16e^{3/2}\sqrt{2m}}.$$

No usaremos directamente el teorema sino que usaremos el siguiente corolario:

Corolario 5.3. *Para todo entero $0 \leq t \leq 1$ y todo $\alpha > 0$ existe un $m_0 = m_0(t, \alpha)$ suficientemente grande tal que, para todo $m \geq m_0$,*

$$\mathbb{P}\left[|L(\mathcal{I}_{2m}) - 2\sqrt{2m}| \geq 2t\sqrt{2m}\right] \leq \alpha.$$

Demostración. Sean $0 \leq t \leq 1$ y $\alpha > 0$. Elijamos $m_0 = m_0(t, \alpha)$ suficientemente grande tal que, para todo $m > m_0$ las siguientes condiciones se cumplan:

1. $|\mathbb{E}(L(\mathcal{I}_{2m})) - 2\sqrt{2m}| + 32(2m)^{1/4} \leq t\sqrt{2m}$.
2. $4e^{-t^2\sqrt{2m}/16e^{3/2}} < \alpha$.
3. El Teorema anterior se cumple para m.

Con esto se tiene que:

$$\begin{aligned}\mathbb{P}\left[|L(\mathcal{I}_{2m}) - 2\sqrt{2m}| \geq 2t\sqrt{2m}\right] &\leq \mathbb{P}\left[|L(\mathcal{I}_{2m}) - \mathbb{E}(L(\mathcal{I}_{2m}))| \geq 2t\sqrt{2m} - |\mathbb{E}(L(\mathcal{I}_{2m})) - 2\sqrt{2m}|\right] \\ &\leq \mathbb{P}\left[|L(\mathcal{I}_{2m}) - \mathbb{E}(L(\mathcal{I}_{2m}))| \geq t\sqrt{2m} + 32(2m)^{1/4}\right] \\ &\leq 4e^{-t^2\sqrt{2m}/16e^{3/2}} \leq \alpha\end{aligned}$$

■

5.1.2. Resultado para el modelo simétrico

Probaremos el siguiente lema que nos permitirá, al igual que como lo hicimos en el capitulo anterior para los modelo de d-palabras y binomial, encontrar una aproximación de la mediana.

Lema 5.4. *Se satisface que*

$$\mathbb{E}[|E|] = pn(n-1), \tag{5.1.1}$$

$$\mathbb{E}[|E'|] = pn(n-1)(1-p)^{2n-4}, \tag{5.1.2}$$

$$\mathbb{E}[|E \setminus E'|] \leq 2p^2 n(n-1)(n-2). \tag{5.1.3}$$

Además, para todo $\eta > 0$,

$$\mathbb{P}[|E - \mathbb{E}[|E|]| \geq \eta\mathbb{E}[|E|]] \leq \frac{2}{\eta^2\mathbb{E}[|E|]}. \tag{5.1.4}$$

Demostración. Para cada $i \neq j$ sea X_{ij} la indicatriz del evento $(i,j) \in E$ y sea Y_{ij} la indicatriz del evento $(i,j) \in E'$. Con esto, para todo $i \neq j$,

$$\begin{aligned}E[X_{ij}] &= p, \\ E[Y_{ij}] &= E[X_{ij}] \prod_{k \notin \{i,j\}} (1 - E[X_{kj}])(1 - E[X_{ik}]) = p(1-p)^{2n-4}.\end{aligned}$$

Notando que $|E| = \sum_{(i,j):i\neq j} X_{ij}$ y $|E'| = \sum_{(i,j):i\neq j} Y_{ij}$ y que en ambas sumas el número de sumandos es $n(n-1)$, se concluyen las desigualdades (5.1.1) y (5.1.2). Por otro lado, usando una desigualdad de Bernoulli,

$$\mathbb{E}[|E \setminus E'|] = \mathbb{E}[|E|] - \mathbb{E}[|E'|] = pn(n-1)[1 - (1-p)^{2n-4}] \leq pn(n-1)(2n-4)p,$$

lo que prueba la desigualdad (5.1.3).

Para la última parte del lema, notemos que $|E|$ también puede escribirse como $2\sum_{i<j} X_{ij}$. Como las variables X_{ij} son variables independientes, se tiene

$$\Delta \stackrel{\text{def}}{=} \sum_{\substack{(i,j),(k,l):i<j,k<l \\ (i,j)\neq(k,l)}} \mathbb{E}(X_{ij}X_{kl}) = \frac{n^2(n-1)^2}{4}p^2 = \frac{\mathbb{E}[|E|]^2}{4}.$$

Aplicando la desigualdad de Chebyshev para indicatrices (Lema 4.11) a las variables $X_{ij}, i<j$, y llamando $X = |E|/2$ a su suma, se obtiene:

$$\begin{aligned}\mathbb{P}(||E| - \mathbb{E}(|E|)| \geq \eta\mathbb{E}(|E|)) &= \mathbb{P}(|X - \mathbb{E}(X)| \geq \eta\mathbb{E}(X)) \\ &\leq \frac{1}{(\eta\mathbb{E}(X))^2}\left(\mathbb{E}(X)(1-\mathbb{E}(X)) + \Delta\right) = \frac{1}{\eta^2\mathbb{E}(X)} \leq \frac{2}{\eta^2\mathbb{E}(|E|)}. \quad \blacksquare\end{aligned}$$

Usemos el lema y el corolario probado en la subsección anterior para estimar una mediana de $L(G)$ cuando G es elegido de acuerdo a $S(K_{n,n}, p)$.

Proposición 5.5. *Sea $\delta > 0$. Existe una constante $C_1 = C_1(\delta)$ suficientemente grande, y dos constantes absolutas C_2 y C_3 suficientemente pequeñas, tal que se cumple lo siguiente:*

1. *Si n y p son tal que*
$$\frac{C_1}{p} \leq n^2 \leq \frac{C_2\delta}{p^{3/2}},$$
entonces toda mediana de $L(S(K_{n,n},p))$ es menor que $2n\sqrt{p}(1+\delta)$.

2. *Si n y p son tal que*
$$\frac{C_1}{p} \leq n^2 \leq \frac{C_3\delta^2}{p^2},$$
entonces toda mediana de $L(S(K_{n,n},p))$ es mayor que $2n\sqrt{p}(1-\delta)$.

Demostración. La demostración es muy similar a la demostración de la Proposición 4.13 referente al modelo de d-palabras aleatorias y a la Proposición 4.18 referente al modelo binomial de d-hipergrafo, así que la realizaremos sin mucho detalle. Sólo incluimos esta demostración por completitud pero el lector puede omitirla sin problemas.

En primer lugar denotemos $M = 2\sqrt{pn(n-1)}$. Un breve cálculo permite concluir que:

$$\begin{aligned}\mathbb{P}[L(G) \geq (1+\delta)2n\sqrt{p}] &\leq \mathbb{P}[L(G) \geq (1+\delta)M] \\ &\leq \mathbb{P}\left[|E \setminus E'| \geq \delta M\right] + \mathbb{P}\left[|E'| \geq (1+\delta/2)M^2/4\right] \\ &\quad + \mathbb{P}\left[L(G') \geq (1+\delta/2)M, |E| < (1+\delta/2)M^2/4\right].\end{aligned}$$

El Lema 5.4, la desigualdad de Markov y la desigualdad $n-1 \geq n/2$ permiten concluir que

$$\mathbb{P}\left[|E \setminus E'| \geq \delta M\right] \leq \frac{n^2 p^{3/2}}{\delta} \qquad \text{y} \qquad \mathbb{P}\left[|E| \geq (1+\delta/2)M^2/4\right] \leq \frac{16}{\delta^2 pn^2}.$$

Imponiendo $C_2 \leq 1/6$ y $C_1 \geq 96\delta^2$, se tiene que ambas cantidades son menores que $1/6$.

Por otro lado, definiendo $m = \lfloor (1+\delta/2)M^2/8 \rfloor$ y usando una desigualdad de Bernoulli se tiene que:

$$\begin{aligned}
\mathbb{P}\left[L(G') \geq (1+\delta/2)M, |E'| < (1+\delta/2)M^2/4\right] &\leq \mathbb{P}\left[L(\mathcal{I}_{2m}) \geq (1+\delta/2)M\right] \\
&\leq \mathbb{P}\left[L(\mathcal{I}_{2m}) \geq 2\sqrt{2m}\frac{1+\delta/2}{\sqrt{1+\delta/2}}\right] \\
&\leq \mathbb{P}\left[L(\mathcal{I}_{2m}) \geq 2\sqrt{2m}\left(1+\frac{\delta}{4+\delta}\right)\right].
\end{aligned}$$

Necesitamos que m sea suficientemente grande para poder aplicar el Corolario 5.3. Específicamente, notando que $\delta/(4+\delta) < 1$ e imponiendo $C_1 \geq 4m_0 + 1$, con $m_0 = m_0(\delta/(4+\delta), 1/6)$ como en el Corolario 5.3 se tiene que:

$$m = \lfloor (1+\delta/2)M^2/8 \rfloor \geq \lfloor pn(n-1)/2 \rfloor \geq \lfloor pn^2/4 \rfloor \geq \lfloor 4C_1 \rfloor \geq m_0,$$

y luego, usando la conclusión de dicho corolario,

$$\mathbb{P}\left[L(G') \geq (1+\delta/2)M, |E'| < (1+\delta/2)M^2/4\right] \leq 1/6.$$

Resumiendo,

$$\mathbb{P}[L(G) \geq (1+\delta)2n\sqrt{p}] \leq 1/6 + 1/6 + 1/6 = 1/2.$$

con lo cual toda mediana de $L(G)$ resulta ser menor o igual que $(1+\delta)2n\sqrt{p}$.

Ahora demostraremos que toda mediana de $L(G)$ es mayor o igual que $(1-\delta)2n\sqrt{p}$. Notemos primero que si $\delta \geq 1$, el resultado se tiene pues $L(G) \geq 0$. Luego, supongamos en esta parte que $\delta < 1$. Además, asumamos primero que $C_1 \geq 2/\delta$, con esto:

$$\sqrt{\frac{n}{n-1}} \leq \left(1+\frac{2}{n}\right)^{1/2} \leq 1 + \frac{1}{n} \leq 1 + \frac{\delta\sqrt{p}}{2} \leq 1 + \frac{\delta}{2(1-\delta)} = \frac{1-\delta/2}{1-\delta}.$$

Luego:

$$(1-\delta)2n\sqrt{p} \leq (1-\delta/2)2\sqrt{pn(n-1)} = (1-\delta/2)M.$$

La desigualdad anterior y un poco de álgebra permiten concluir que:

$$\begin{aligned}
\mathbb{P}\left[L(G') \leq (1-\delta)2n\sqrt{p}\right] &\leq \mathbb{P}\left[L(G) \leq (1-\delta/2)M\right] \\
&\leq \mathbb{P}\left[|E \setminus E'| > \delta M^2/16\right] + \mathbb{P}\left[|E| \leq (1-\delta/4)M^2/4\right] + \\
&\quad \mathbb{P}\left[L(G') \leq (1-\delta/2)M, |E'| > (1-\delta/2)M^2/4\right].
\end{aligned}$$

Nuevamente acotemos los términos del lado derecho por separado. Gracias al Lema 5.4, se tiene:

$$\mathbb{P}\left[|E \setminus E'| > \delta M^2/16\right] \leq \frac{8pn}{\delta} \qquad \text{y} \qquad \mathbb{P}\left[|E| \leq (1-\delta/4)M^2/4\right] \leq \frac{64}{\delta^2 pn^2}.$$

Imponiendo $C_3 \leq (1/48)^2$ y $C_1 \geq 384/\delta^2$ se concluye que ambas cantidades son menores que $1/6$.

Para el tercer término, definamos $m = \lceil (1-\delta/2)M^2/8 \rceil$, luego, usando la desigualdad de Bernoulli, se obtiene que:

$$\begin{aligned}\mathbb{P}\left[L(G') \leq (1-\delta/2)M, |E'| > (1-\delta/2)M^2/4\right] &\leq \mathbb{P}\left[L(\mathcal{I}_{2m}) \leq (1-\delta/2)M\right] \\ &\leq \mathbb{P}\left[L(\mathcal{I}_{2m}) \leq 2\sqrt{2m}\sqrt{1-\delta/2}\right] \\ &\leq \mathbb{P}\left[L(\mathcal{I}_{2m}) \leq 2\sqrt{2m}\,(1-\delta/4)\right].\end{aligned}$$

Necesitamos que m sea suficientemente grande para poder aplicar el Corolario 5.3. Específicamente notando que $\delta/4 < 1$ y e imponiendo $C_1 \geq 4m_0/(1-\delta/2)$, que no se indefine pues $\delta < 1$, con $m_0 = m_0(\delta/4, 1/6)$ dado por el Corolario 5.3 se tiene que

$$m \geq (1-\delta/2)M^2/8 \geq (1-\delta/2)pn(n-1)/2 \geq (1-\delta/2)pn^2/4 \geq m_0,$$

y luego, usando la conclusión de dicho corolario,

$$\mathbb{P}\left[L(G') \leq (1-\delta)M, |E'| > (1-\delta)M^2/4\right] \leq 1/6.$$

Resumiendo,

$$\mathbb{P}[L(G) \leq (1-\delta)M] \leq 1/6 + 1/6 + 1/6 = 1/2.$$

Lo que concluye la demostración. ■

Tenemos inmediatamente el siguiente corolario,

Corolario 5.6. *Si definimos $t = 1/p$, el modelo $S(K_{n,n}, p)$ de parámetro interno t admite una $(2, 1/2, 3/4)$-mediana.*

Lamentablemente el modelo simétrico no es en estricto rigor un modelo de (2-hiper)grafo, (pues solo admite grafos con clases del mismo número de vértices) y luego no podemos aplicar directamente el Teorema Principal para obtener una cota de estimación para este modelo. Sin embargo, veremos que la misma demostración basta para probar una parte de lo que queremos.

Teorema 5.7 (Teorema de estimación para el modelo simétrico.)**.** *Para todo $\varepsilon > 0$, existen constantes p_0 suficientemente pequeño y A suficientemente grande tales que, para todo $p \leq p_0$ y todo $n \geq A/\sqrt{p}$,*

$$(1-\varepsilon)2n\sqrt{p} \leq \mathbb{E}[L(S(K_{n,n}, p))] \leq (1+\varepsilon)2n\sqrt{p}. \tag{5.1.5}$$

Por otro lado, si $\mathrm{Med}[L(S(K_{n,n},p))]$ *es una mediana de* $L(S(K_{n,n},p))$

$$(1-\varepsilon)2n\sqrt{p} \leq \mathrm{Med}[L(S(K_{n,n},p))] \leq (1+\varepsilon)2n\sqrt{p}. \tag{5.1.6}$$

Además, existe una constante absoluta $C>0$*, tal que para* p *y* n *como antes:*

$$\mathbb{P}\left[L(G)) \leq (1-\varepsilon)2n\sqrt{p}\right] \leq \exp\left(-C\varepsilon^2 n\sqrt{p}\right), \tag{5.1.7}$$

$$\mathbb{P}\left[L(G)) \geq (1+\varepsilon)2n\sqrt{p}\right] \leq \exp\left(-C\frac{\varepsilon^2}{1+\varepsilon}n\sqrt{p}\right). \tag{5.1.8}$$

Demostración. Para probar la cota inferior de (5.1.5), la cota inferior de (5.1.6) y la desigualdad (5.1.7), usamos que el modelo $S(K_{n,n},p)$ posee una constante de concentración $h=1/4$ y una $(2,1/2,3/4)$-mediana para el parámetro $t=1/p$. Con esto basta repetir la demostración de la cota inferior para el Teorema Principal. El comentario al final de dicha demostración permite aplicar la misma técnica a esta familia de distribuciones.

Para demostrar el resto, sea H un grafo aleatorio bipartito de clases A y B donde cada arista está en $E(H)$ con probabilidad p de manera independiente, y sea O el subgrafo de H obtenido al eliminar las aristas (x,y) tales que $x \geq y$. Como O es subgrafo de H claramente $L(O) \leq L(H)$. Además, por definición, H sigue una distribución binomial $G(K_{n,n},p)$, mientras que O sigue una distribución orientada $O(K^*_{n,n},p)$. Recordando el Lema 5.1, $L(O(K^*_{n,n},p))$ tiene la misma distribución que $L(S(K_{n,n},p))$. Con esto, si n y p cumplen las hipotesis del Teorema de estimación para el modelo binomial,

$$\begin{aligned}\mathbb{E}[L(S(K_{n,n},p))] = \mathbb{E}[L(O)] \leq \mathbb{E}[L(H)] \leq (1+\varepsilon)2n\sqrt{p},\\ \mathrm{Med}[L(S(K_{n,n},p))] = \mathrm{Med}[L(O)] \leq \mathrm{Med}[L(H)] \leq (1+\varepsilon)2n\sqrt{p},\end{aligned}$$

y

$$\begin{aligned}\mathbb{P}\left[L(S(K_{n,n},p)) \geq (1+\varepsilon)2n\sqrt{p}\right] &= \mathbb{P}\left[L(O) \geq (1+\varepsilon)2n\sqrt{p}\right]\\ &\leq \mathbb{P}\left[L(H) \geq (1+\varepsilon)2n\sqrt{p}\right]\\ &\leq \exp\left(-C\frac{\varepsilon^2}{1+\varepsilon}2n\sqrt{p}\right),\end{aligned}$$

Esto que concluye la demostración. ■

5.2. Modelo antisimétrico

El segundo modelo a estudiar posee otro tipo distinto de simetría. Para definirlo consideremos esta vez A y B dos conjuntos disjuntos de tamaño $2n$. Como es habitual denotemos por $K_{2n,2n}$ al grafo bipartito completo de clases de partición A y B. Esta vez, a diferencia del caso simétrico, supondremos que los elementos de ambos conjuntos están numerados de $-n$ a n (sin usar el 0).

Denotemos $A(K_{2n,2n}, p)$ a la distribución sobre todos los subgrafos de $K_{2n,2n}$ donde para todo $x \leq y$, el evento $\{(x, y), (-x, -y)\} \subseteq E(G)$, con G una realización, tiene probabilidad p, y todos estos eventos son independientes.

De la definición anterior, cada vez que el arco (x, y) está en una realización, su antisimétrico $(-x, -y)$ también lo está.

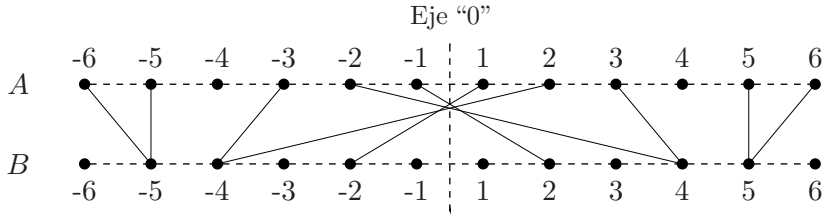

Figura 5.3: Representación de un grafo G elegido de acuerdo a la distribución antisimétrica bipartita $A(K_{12,12}, p)$. Intuitivamente un grafo es antisimétrico si su representación gráfica resulta ser simétrica con respecto a un eje vertical imaginario (el eje "0") ubicado entre los vértices -1 y 1 de cada clase.

Como es habitual, para G elegido de acuerdo a $A(K_{2n,2n}, p)$ denotamos $L(G)$ al largo de su subgrafo monótono más grande. Probaremos la siguiente cota:

Teorema 5.8 (Teorema de estimación para el modelo antisimétrico)**.** *Para todo $\varepsilon > 0$, existen constantes p_0 suficientemente pequeño y A suficientemente grande tales que, para todo $p \leq p_0$ y todo $n \geq A/\sqrt{p}$, se tiene:*

$$(1 - \varepsilon)4n\sqrt{p} \leq \mathbb{E}[L(A(K_{2n,2n}, p))] \leq (1 + \varepsilon)4n\sqrt{p}. \tag{5.2.1}$$

Por otro lado, si $\mathrm{Med}[L(A(K_{2n,2n}, p))]$ *es una mediana de* $L(A(K_{2n,2n}, p))$

$$(1 - \varepsilon)4n\sqrt{p} \leq \mathrm{Med}[L(A(K_{2n,2n}, p))] \leq (1 + \varepsilon)4n\sqrt{p}. \tag{5.2.2}$$

Además, existe una constante absoluta $C > 0$, tal que para p y n como antes:

$$\mathbb{P}\left[L(A(K_{2n,2n}, p)) \leq (1 - \varepsilon)4n\sqrt{p}\right] \leq \exp\left(-C\varepsilon^2 n\sqrt{p}\right), \tag{5.2.3}$$

$$\mathbb{P}\left[L(A(K_{2n,2n}, p)) \geq (1 + \varepsilon)4n\sqrt{p}\right] \leq \exp\left(-C\frac{\varepsilon^2}{1+\varepsilon} n\sqrt{p}\right). \tag{5.2.4}$$

Demostración. Demostraremos esta cota reduciendo este problema a los problemas ya estudiados.

Para probar la cota inferior de (5.2.1), la cota inferior de (5.2.2) y la desigualdad (5.2.3), consideremos G elegido de acuerdo a $A(K_{2n,2n}, p)$ y G' el grafo obtenido al eliminar los arcos que "cruzan el cero" (es decir, los arcos (x, y) tales que x e y tienen distinto signo). Llamemos además G_- al subgrafo de G' inducido por los vértices etiquetados con números negativos y G_+ el subgrafo

de G' inducido por los vértices etiquetados con números positivos. Es fácil ver, por definición de $A(K_{2n,2n}, p)$, que

$$E(G_-) = \{(-x, -y) \mid (x, y) \in E(G_+)\}$$

y que G_+ se distribuye de acuerdo al modelo binomial $G(K_{n.n}, p)$. Con esto, para todo $\varepsilon > 0$, existen p_0, A y C (dados por el Teorema 4.20 de estimación para el modelo binomial) tales que para todo $p \leq p_0$ y $n \geq A/\sqrt{p}$, se tiene:

$$\mathbb{E}[L(A(K_{2n,2n}, p))] \geq \mathbb{E}(L(G_-)) + \mathbb{E}(L(G_+)) = 2\mathbb{E}(L(G(K_{n.n}, p))) \geq (1-\varepsilon)4n\sqrt{p}.$$

Es decir, se prueba (5.2.1). Por otro lado,

$$\mathbb{P}\left[L(A(K_{2n,2n}, p)) \leq (1-\varepsilon)4n\sqrt{p}\right] \leq \mathbb{P}\left[L(G(K_{n,n}, p)) \leq (1-\varepsilon)2n\sqrt{p}\right].$$

Usando el Teorema 4.20, tenemos que el lado derecho de la desigualdad anterior es menor que $1/2$ (pues $\text{Med}[L(G(K_{n,n}, p)] \geq (1-\varepsilon)2n\sqrt{p}$) y también es menor que $\exp\left(-C\varepsilon^2 2n\sqrt{p}\right)$. Esto último concluye la demostración de (5.2.2) y de (5.2.3).

Las cotas restantes requerirán algo más de trabajo, de todas formas al igual que en el caso anterior esta demostración será una reducción al problema asociado al modelo binomial.

Sea $\varepsilon > 0$ y G un grafo elegido de acuerdo a $A(K_{2n,2n}, p)$. Definamos además $A = A(\varepsilon)$ y $p_0 = p_0(\varepsilon)$ dados por el Teorema de estimación para el modelo binomial y $m_{\max} = \lceil (1+\varepsilon)4n\sqrt{p} \rceil$.

Notemos que si J es un subgrafo monótono de G y $e = (x, y)$ es una arista de J que cruza el cero, entonces por monotonicidad de J, todas las aristas de J que cruzan el cero lo hacen en la misma dirección, es decir, si $f = (z, w)$ es otra arista que cruza el cero, entonces $\text{sign}(z) = \text{sign}(x)$ y $\text{sign}(w) = \text{sign}(y)$. Además, si definimos $-J$ como el grafo que contiene las aristas antisimétricas de las aristas de J, entonces $-J$ también es un subgrafo monótono de G.

La observación anterior dice que si J es un subgrafo monótono de largo máximo de G entonces podemos asumir que todas las aristas (x, y) de J que cruzan el 0 son tales que $x < 0$ e $y > 0$. Diremos, en este caso, que J es un subgrafo monótono derecho.

Clasificaremos todos los subgrafos monótonos derechos de G en *tipos*, siguiendo un poco la idea de la demostración del Teorema Principal. Sin embargo esta vez los tipos serán definidos de manera diferente.

Dividiremos la parte positiva de la clase de partición B de G en segmentos de la siguiente manera: Sea $\delta = \text{mín}\{\varepsilon, 1\}$ y para $1 \leq i \leq q = \lceil 1/\delta \rceil$, llamemos $B_i = \{v \in B : n - i\lceil n\delta \rceil < v \leq n - (i-1)\lceil n\delta \rceil\}$.

Para J un grafo monótono derecho con exactamente $m_{\max}$ aristas, definimos $\tau(J)$ como el índice del segmento en el que cae la última arista (usando el orden natural) que cruza el cero (Si J no posee aristas que cruzan el cero entonces definimos $\tau(J) = q$). Además, llamemos J_1 al subgrafo inducido por todos las vértices que se encuentran "a la izquierda" de dicha arista, incluyéndola, (si dicha arista no existe entonces J_1 será el subgrafo inducido por los vértices de índice negativo) y J_2 al subgrafo inducido por los vértices restantes.

Con esto, definimos el *tipo* de J, denotado $T(J)$, como el par (i, s) donde $i = \tau(J)$ y $s = |E(J_1)|$, y llamemos $\mathcal{T}$ al conjunto $\{(i, s) : 1 \leq i \leq q, 1 \leq s \leq m_{\max}\}$ de todos los tipos posibles.

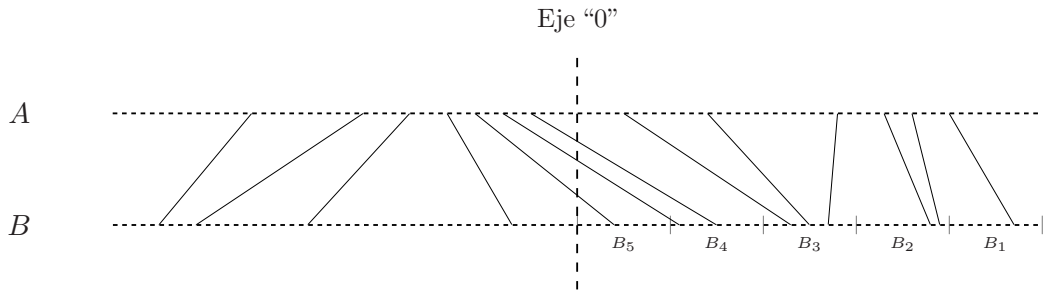

Figura 5.4: Se representa J, un subgrafo monótono derecho de G. La parte positiva de B fue dividida en $q = 5$ segmentos. La última arista de J que cruza el 0 cae en el segmento B_4. Notando que a la izquierda de dicha arista hay exactamente 7 aristas (incluyéndola), se tiene que el tipo de J es $T(J) = (4, 7)$.

Denotemos además P_T a la probabilidad de que G posea un subhipergrafo monótono derecho del tipo T. Con esto,

$$\mathbb{P}[L(G) \geq m_{\max}] \leq \sum_{T \in \mathcal{T}} P_T.$$

Acotemos esta cantidad. Para ello, necesitamos un poco de notación, llamemos $G[a_1, a_2][b_1, b_2]$ al subgrafo de G inducido por los vértices de la clase A con etiquetas en $[a_1, a_2]$ y por los vértices de la clase B con etiquetas en $[b_1, b_2]$, con esto, si $T = (i, s)$, entonces (notando que J_1 y J_2 no son independientes),

$$\begin{aligned} P_T &\leq \min\{\mathbb{P}[L(J_1) \geq s], \mathbb{P}[L(J_2) \geq m_{\max} - s]\} \\ &\leq \min\Big\{\mathbb{P}\left[L\left(G\big[-n, -1\big]\big[-n, n - (i-1)\lceil n\delta\rceil\big]\right) \geq s\right], \\ &\qquad \mathbb{P}\left[L\left(G\big[1, n\big]\big[n - i\lceil n\delta\rceil + 1, n\big]\right) \geq m_{\max} - s\right]\Big\}. \end{aligned}$$

Cabe observar que los dos subgrafos, $G\big[-n, -1\big]\big[-n, n - (i-1)\lceil n\delta\rceil\big]$ y $G\big[1, n\big]\big[n - i\lceil n\delta\rceil + 1, n\big]$, tampoco son independientes. De hecho se sobreponen en algunos vértices de la clase B. Además, en cada uno de los subgrafos, todos los vértices de la clase A tienen el mismo signo. Con esto, al interior de cada uno de ellos, los eventos asociados a sus posibles aristas son independientes y luego ambos subgrafos se distribuyen de acuerdo al modelo binomial asociado al largo de sus clases. Por lo tanto:

$$\begin{aligned} P_T &\leq \min\{\mathbb{P}[L(G(K_{n,2n-(i-1)\lceil n\delta\rceil}, p) \geq s], \mathbb{P}[L(G(K_{n,i\lceil n\delta\rceil}, p) \geq m_{\max} - s]\} \\ &\leq \sqrt{\mathbb{P}[L(G(K_{n,2n-(i-1)\lceil n\delta\rceil}, p) \geq s] \cdot \mathbb{P}[L(G(K_{n,i\lceil n\delta\rceil}, p) \geq m_{\max} - s]}. \end{aligned}$$

A continuación probaremos que ambos grafos en el lado derecho satisfacen las condiciones de tamaño y balanceo requeridos para aplicar el Teorema de estimación para la distribución binomial. En efecto,

para la condición de tamaño basta notar que el promedio geométrico de los tamaño de las clases en cada grafo, digamos $\widetilde{N}_1$ y $\widetilde{N}_2$, es mayor o igual que $n\delta$. Luego, imponiendo $n\sqrt{p} \geq A/\delta$ (Recordando que podemos hacer esta cantidad tan grande como queramos por hipótesis), se tiene que para $j = 1, 2$,

$$\widetilde{N}_j\sqrt{p} \geq n\delta\sqrt{p} \geq A.$$

Por otro lado, para la condición de balanceo basta notar que la suma de los tamaños de las clases de cada grafo, digamos $\widetilde{S}_1$ y $\widetilde{S}_2$ es a lo más el número de vértices de G, es decir, $4n$, con lo cual para $j = 1, 2$

$$\widetilde{S}_j \leq 4n = \frac{4}{\delta}n\delta \leq \frac{4}{\delta}\widetilde{N}_j.$$

Como la cantidad $4/\delta$ es una constante independiente de p, se tiene la condición de balanceo. Luego, aplicando el Teorema de estimación para el modelo binomial se tiene,

$$\mathbb{P}[L(G(K_{n,2n-(i-1)\lceil n\delta\rceil}, p) \geq s] \leq \exp\left(-C\frac{\text{máx}\left(0, s - 2\sqrt{n(2n-(i-1)\lceil n\delta\rceil)p}\right)^2}{s}\right),$$

$$\mathbb{P}[L(G(K_{n,i\lceil n\delta\rceil}, p) \geq m_{\max} - s] \leq \exp\left(-C\frac{\text{máx}\left(0, m_{\max} - s - 2\sqrt{in\lceil n\delta\rceil p}\right)^2}{m_{\max} - s}\right).$$

Como los denominadores al interior de las exponenciales son menores o iguales que $m_{\max}$, y usando la desigualdad de Cauchy-Schwartz,

$$\begin{aligned}
-\ln P_T &\geq \frac{C}{m_{\max}}\left(\text{máx}\left(0, s - 2\sqrt{n(2n-(i-1)\lceil n\delta\rceil)p}\right)^2 + \text{máx}\left(0, m_{\max} - s - 2\sqrt{in\lceil n\delta\rceil p}\right)^2\right) \\
&\geq \frac{C}{2m_{\max}}\,\text{máx}\left(0, s - 2\sqrt{n(2n-(i-1)\lceil n\delta\rceil)p} + m_{\max} - s - 2\sqrt{in\lceil n\delta\rceil p}\right)^2 \\
&= \frac{C}{2m_{\max}}\,\text{máx}\left(0, m_{\max} - 2\sqrt{p}\left(\sqrt{n(2n-(i-1)\lceil n\delta\rceil)} + \sqrt{in\lceil n\delta\rceil}\right)\right)^2 \\
&\geq \frac{C}{2m_{\max}}\,\text{máx}\left(0, m_{\max} - 2\sqrt{p}\sqrt{2n(2n+\lceil n\delta\rceil)}\right)^2.
\end{aligned}$$

Como n es grande, lo anterior es mayor o igual que, digamos

$$\begin{aligned}
&\frac{C}{2m_{\max}}\,\text{máx}\left(0, m_{\max} - 2\sqrt{p}\sqrt{2n(2n+(3/2)n\varepsilon}\right)^2 \\
&\geq \frac{C}{2m_{\max}}\,\text{máx}\left(0, 4n\sqrt{p}\left((1+\varepsilon) - \sqrt{1+(3/2)\varepsilon}\right)\right)^2,
\end{aligned}$$

y usando la desigualdad de Bernoulli, lo anterior es mayor o igual que

$$\frac{C}{2m_{\max}}\,\text{máx}\,(0, n\sqrt{p}\varepsilon)^2 \geq \frac{C}{(1+\varepsilon)8n\sqrt{p}}n^2p\varepsilon^2 = \frac{C\varepsilon^2}{32(1+\varepsilon)}4n\sqrt{p}.$$

Por otro lado, como $n\sqrt{p}$ se puede escoger tan grande como queramos,

$$|\mathcal{T}| = qm_{\max} \leq 4n\sqrt{p}(1+\varepsilon)^2\lceil 1/\delta\rceil \leq \exp\left(\frac{C\varepsilon^2}{64(1+\varepsilon)}4n\sqrt{p}\right).$$

Sigue que

$$\begin{aligned}\mathbb{P}[L(G) \geq m_{\max}] &\leq \sum_{T\in\mathcal{T}} P_T \\ &\leq \exp\left(4n\sqrt{p}\frac{C\varepsilon^2}{64(1+\varepsilon)}\right)\exp\left(-4n\sqrt{p}\frac{C\varepsilon^2}{32(1+\varepsilon)}\right) \\ &= \exp\left(-4n\sqrt{p}\frac{C\varepsilon^2}{64(1+\varepsilon)}\right).\end{aligned}$$

Lo que concluye la demostración de la cota para la cola superior, es decir, la desigualdad (5.2.4). Además, para la cola de la mediana (5.2.2), basta notar que tomando $n\sqrt{p}$ suficientemente grande en la desigualdad anterior, el lado derecho se puede hacer menor o igual a $1/2$.

Ahora sólo falta la cota para la esperanza (5.2.1). Para un ε_0 cualquiera, tomemos ε tal que $(1+\varepsilon)\sqrt{1+3\varepsilon/4}$ sea menor que $(1+\varepsilon_0)$. Luego, usando la misma notación de la parte anterior, es decir, $\delta = \text{mín}\{\varepsilon, 1\}$ y $q = \lceil 1/\delta\rceil$, podemos notar que:

$$\begin{aligned}\mathbb{E}[L(G)] &\leq \max_{1\leq i\leq q}\left\{\mathbb{E}\Big[L\big(G\big[-n,-1\big]\big[-n,n-(i-1)\lceil n\delta\rceil\big]\big)\Big] + \mathbb{E}\Big[L\big(G\big[1,n\big]\big[n-i\lceil n\delta\rceil+1,n\big]\big)\Big]\right\} \\ &= \max_{1\leq i\leq q}\left\{\mathbb{E}\big[L(G(K_{n,2n-(i-1)\lceil n\delta\rceil},p))\big] + \mathbb{E}\big[L(G(K_{n,i\lceil n\delta\rceil},p))\big]\right\}.\end{aligned}$$

Esto se tiene pues, si consideramos J un subgrafo monótono derecho de tamaño máximo de G, la desigualdad anterior se alcanza para aquel índice correspondiente al segmento de B (en la división realizada anteriormente) en el que cae la última arista de J que cruza el 0.

Como todos los grafos que participan en el máximo satisfacen las hipótesis de tamaño y balanceo para el Teorema de Estimación para el modelo binomial,

$$\begin{aligned}\mathbb{E}[L(G)] &\leq \max_{1\leq i\leq q}\left\{(1+\varepsilon)2\sqrt{pn\cdot(2n-(i-1)\lceil n\delta\rceil)} + (1+\varepsilon)2\sqrt{pn\cdot(i\lceil n\delta\rceil)}\right\} \\ &\leq (1+\varepsilon)2\sqrt{p}\max_{1\leq i\leq q}\left\{\sqrt{2n\cdot(2n-(i-1)\lceil n\delta\rceil + i\lceil n\delta\rceil)}\right\} \\ &= (1+\varepsilon)2\sqrt{p}\sqrt{2n\cdot(2n+\lceil n\delta\rceil)} \\ &\leq (1+\varepsilon)2\sqrt{p}\sqrt{2n\cdot(2n+(3/2)n\varepsilon)} \\ &= (1+\varepsilon)4n\sqrt{p}\sqrt{1+3\varepsilon/4} \leq (1+\varepsilon_0)4n\sqrt{p}.\end{aligned}$$

Esto concluye la demostración. ■

Bibliografía

[1] V. Chvátal y D. Sankoff. Longest common subsequences of two random sequences. *Journal of Applied Probability*, 12:306-315, 1975.

[2] J. Deken. Some limits results for longest common subsequences. *Discrete Mathematics*, 26:17-31, 1979.

[3] V. Dančík. Expected length of longest common subsequences. *PhD thesis, CS Dept, Univ. of Warwick, Warwick, UK.* 1994

[4] M. Patterson y V. Dančík Longest common subsequences. *In I. Privara, B. Rovan, and P. Ruzicka, editors, MFCS'*94*, LNCS* , 841:127-142. *Springer-Verlag,* 1994.

[5] R. Baeza-Yates, G. Navarro, R. Gavaldá y R. Scheihing. Bounding the expected length of the longest common subsequences and forests. *Theory of Computing Systems*, 32(4):435-452, 1999.

[6] G. S. Lueker. Improved bounds on the average length of longest common subsequences. *Fourteenth Annual ACM/SIAM Symposium on Discrete Algorithms*, 130-131, 2003.

[7] M. Kiwi, M. Loebl, y J. Matoušek. Expected length of the longest common subsequence for large alphabets. *LATIN 2004: Theoretical Informatics*, 302-311, 2004.

[8] J. M. Steele. An Efron-Stein inequality for nonsymmetric statistics. *The Annals of Statistics*, 14(2):753-758, 1986.

[9] K. S. Alexander. The rate of convergence of the mean length of the longest common subsequence. *The Annals of Applied Probability*, 4(4):1074-1082, 1994.

[10] B. Bollobas y G. Brightwell. The height of a random partial order: Concentration of Measure. *The Annals of Applied Probability*, 2(4):1009-1018, 1992.

[11] B. Bollobas y P. Winkler. The longest chain among random points in euclidean space. *Proc. on the American Mathematical Society*, 103(2):347-353, 1988.

[12] D. Sankoff y J. Kruskal, editors. Common subsequences and monotone subsequences *Addison-Wesley, Reading, Mass.*, chapter 17: 363-365, 1983.

[13] S. Janson, T. Luzak y A. Rucinski. Random Graphs. *Wiley.* 2000.

[14] M. Kiwi A Concentration Bound for the Longest Increasing Subsequence of a Randomly Chosen Involution. *Por aparecer en Discrete Applied Mathematics, 2006.*

[15] R. Wagner y M. Fischer. The string-to-string correction problem. *Journal of the Association for Computing Machinery*, 21(1):168-173, 1974.

[16] D. Maier. The complexity of some problems on subsequences and supersequences. *Journal of the Association for Computing Machinery*, 25(2):322-336, 1978.

[17] J. Baik y E. M. Rains. Symmetrized random permutations. *In Random Matrix Models and Their Applications, volume 40 of Mathematical Sciences Research Institute Publications, pages 1-19, Cambridge Univ. Press, Cambridge*, 2001.

[18] S. M. Ulam. Monte Carlo calculations in problems of mathematical physics. *Modern Mathematics for the Engineer: Second Series. E. F. Beckenbach, Editor. McGraw-Hill, New York*, 1961.

[19] J. M. Hammersley. A few seedlings of research. *In Proc. Sixth Berk. Symp. Math. Stat. and Prob., Vol. 1*, 345 - 394, 1972.

[20] B. F. Logan y L. A. Shepp. A variational problem for random Young tableaux. *Advances in Mathematics*, 26:206-222, 1977.

[21] A. M. Verskik y S. V. Kerov. Asymptotics of the Plancherel measure of the symmetric group and the limiting form of Young tables. *Soviet Math. Dokl.* 18:527-531. *Translation of Dokl. Acad. Nauk. SSSR* 233:1024-1027, 1977