Centro de Investigación en Matemáticas A.C

CIMAT

El $\Lambda-$coalescente e Inferencia sobre el Espectro de Frecuencia de Sitio

T E S I S

QUE PARA OBTENER EL GRADO DE:
MAESTRÍA EN CIENCIAS CON ESPECIALIDAD EN PROBABILIDAD Y ESTADÍSTICA

PRESENTA:
DHYANA DEL SOL SILVA AGUILERA

Dirigida por
Dr. Arno Siri-Jégousse

2016

Agradecimientos

A mis amigos, a mis compañeros, a mis profesores, a mi asesor de tesis, a mis sinodales, a mis padres, a mis hermanos y a mi gatita.

Este párrafo, en color azul, ha sido agregado a la versión original para que la tesis cumpla con el reglamento de CIMAT de agradecer a CONACYT por el apoyo económico brindado para realizar los estudios de maestría. Además aprovecho la oportunidad para aclarar que se me informó puntualmente sobre este reglamento. Sin embargo, deseaba mantener la autenticidad que debería caracterizar la expresión de cualquier sentimiento. Aclarado esto, dejo el siguiente manifiesto: "Agradezco a Conacyt por el apoyo económico brindado para realizar mis estudios de maestría."

Índice general

Introducción

Una de las principales motivaciones para el estudio del $\Lambda-$coalescente se encuentra en sus aplicaciones a importantes modelos en genética de poblaciones. Es por esto que, a modo de preámbulo, a continuación se dará una breve contextualización histórica y genetista, obtenida en su mayoría de los libros [7] y [8].

La **genética** está definida como la rama de la biología que estudia la herencia y la variación de los rasgos y características de los seres vivos. El término "genética" fue usado por primera vez por **William Bateson** en 1905. Bateson redescubrió el trabajo de **Gregor Mendel** (1822-1884) convirtiéndose desde entonces en uno de sus más grandes defensores.

Alrededor de 20 años después del trabajo de Mendel los avances en los microscopios permitieron el descubrimiento e investigación de los cromosomas (en griego *chroma* significa color). Mediante colorantes, los científicos observaron los núcleos de las **células ecuariotas**, que son células con núcleo delimitado por una doble capa lipídica. Cuando estas células se encuentran en estado de reposo (es decir, cuando no están en proceso de dividirse), su núcleo tenía una débil apariencia reticular a la que llamaron **cromatina**. Mientras que al dividirse, mediante la **mitosis** o la **meiosis**, la cromatina experimentaba un cambio brusco transformándose en pequeños "hilos" que se repartían en las células resultantes. Dichos hilos fueron llamados **cromosomas**.

A principios del siglo XX, **Walter Sutton** y **Theodore Boveri** observaron, de manera independiente muchas similitudes entre el comportamiento de los genes y los cromosomas. En base a esto propusieron que los genes se encuentran en los cromosomas. Esta propuesta forma la base de la **Teoría Cromosómica de la Herencia**. Esta teoría fue muy controvertida hasta 1915, cuando **Thomas Hunt Morgan** con-

siguió que fuera universalmente aceptada después de sus estudios realizados en la mosca de la fruta (**Drosophila melanogaster**).

Por el mismo tiempo en que la teoría cromosómica de la herencia fue propuesta, se descubrió una mosca de la fruta de ojos blancos en una botella que solo contenía moscas con ojos rojos. Esta variación fue producida por una mutación en uno de los genes que controla el color de ojos. En el frasco de moscas de la fruta ahora se encontraban al menos dos formas del mismo gen. Una de las formas causaba que los ojos fueran rojos y otra que los ojos fueran blancos. Este es un ejemplo de un alelo. Un **alelo** está definido como una de las posibles formas de un mismo gen. Alelos distintos pueden o no producir características distintas observables en los organismos. El **genotipo** de un organismo corresponde a toda su información genética, incluyendo los alelos. Mientras que el **fenotipo** se refiere a los rasgos observables del organismo, los cuales dependen además del ambiente en que el organismo se desarrolla. Por lo que los alelos siempre representan cambios en el genotipo, pero pueden o no representar cambios en el fenotipo. Una **mutación** se define como cualquier cambio en el genotipo de un organismo.

Estudios sobre la mosca de la fruta de ojos blancos mostraron que este rasgo mutante podía atribuirse a un solo cromosoma. De esta manera se confirmó que los genes se encuentran en los cromosomas. Ahora la pregunta era ¿Qué componente de los cromosomas es el que contiene la información genética? Se sabía los cromosomas estaban compuestos principalmente de dos tipos de polímeros: las proteínas y el **ADN** (**ácido desoxirribonucleico**). La gran variedad y abundancia de proteínas en la célula hacían creer a los científicos que la información genética estaba en las proteínas. En 1944, **Oswald Avery**, **Colin MacLeod** y **Maclyn McCarty** publican sus experimentos probando que la información genética de las bacterias estaba contenida en el ADN. Tomaría algunos años de controversia científica aceptar al ADN como la substancia química que contiene la información genética de todos los seres vivos.

Gracias a **James Watson**, **Francis Crick** y a su trabajo publicado en 1953, hoy conocemos los detalles de la estructura del ADN. Desde el punto de vista químico,

el ADN es un polímero lineal de **nucleótidos** (cierto tipo de moléculas orgánicas). Un polímero lineal es un compuesto formado por muchas unidades simples conectadas entre sí, como si fuera un largo tren formado por vagones. En el ADN, cada vagón es un nucleótido, y cada nucleótido, a su vez, está formado por un azúcar (la desoxirribosa), una base nitrogenada (que puede ser **A**denina, **T**imina, **C**itosina o **G**uanina) y un grupo fosfato que actúa como enganche de cada vagón con el siguiente. Por lo tanto, lo que distingue a un nucleótido de otro es la base nitrogenada. En este sentido se puede distinguir una **secuencia de ADN** simplemente nombrando la primer letra de cada base nitrogenada de la cadena. Por ejemplo una secuencia de ADN podría ser AGCATTAGCATTAAGCC..... Actualmente existen diversos métodos y técnicas bioquímicas para determinar fragmentos de la secuencia de ADN de un individuo.

La posibilidad de conocer y comparar fragmentos de secuencia de ADN en diferentes individuos nos permite inferir la genealogía de estos individuos y las mutaciones a lo largo de esta, creando así un ambiente fructífero para el desarrollo de modelos probabilistas. A principios de la década de 1930 **Sewall Wright** y **Rodald Fisher** describen independientemente a un modelo probabilista discreto para la genealogía de una población de N individuos. En 1982, **John Kingman** [16] estudia la convergencia de las genealogías del modelo propuesto por Wright y Fisher para N grande, a un modelo continuo llamado **coalescente de Kingman**.

Detrás del coalescente de Kingman se encuentran algunas hipótesis muy fuertes sobre la reproducción de la población a estudiar. Considerar otras hipótesis nos lleva a modelos más generales, como son los Λ-coalescentes y los Ξ-coalecentes.

En esta tesis se presentan modelos para genética de poblaciones basados tanto en el coalescente de Kingman, como en el $\Lambda-$coalescente o el $\Xi-$coalescente. Sin embargo, se pone énfasis en el estudio del $\Lambda-$coalescente. Esto, ya que para algunas poblaciones los modelos basados en este coalescente tienen un mejor ajuste que los modelos basados en el coalescente de Kingman. Además, es más sencillo hacer inferencia sobre los parámetros de modelos que utilizan el $\Lambda-$coalsecente, que hacer inferencia sobre los parámetros de los modelos que utilizan el $\Xi-$coalescente.

La inferencia sobre los parámetros de dichos modelos se hace en base a un impor-

tante estadístico llamado “Espectro de Frecuencia de Sitio”, el cual también explica en esta tesis.

Esta tesis está pensada para ser de utilidad práctica a cualquier lector con conocimientos muy básicos de probabilidad y que tenga interés en alguno de los temas aquí presentados. Dichos temas se muestran de la manera más independiente posible. Por lo anterior, se sugiere al lector no poner énfasis en ningún concepto cuya comprensión se muestre difícil en una primera lectura.

La tesis está estructurada mediante tres capítulos y un apéndice.

El primer capítulo de la tesis habla sobre un teorema probabilista, el Teorema de De Finetti. Este teorema es de importancia fundamental en muchas áreas importantes de probabilidad y estadística, como lo son la estadística bayesiana y la teoría la coalescencia. Aunque el teorema es muy general, en esta tesis se demuestra solo un caso particular, ya que este caso es suficiente para la construcción del $\Lambda-$coalescente.

En el segundo capítulo de la tesis se definen y construyen el coalescente de Kingman, el Λ-coalescente y el $\Xi-$coalescente. La construcción se hace explicando de manera muy intuitiva el espacio sobre el cual está definido y su topología. Este capítulo finaliza con algunos ejemplos de estos coalescentes así como una breve mención de los modelos en los cuales surgen.

En el tercer capítulo se presenta y estudia al Espectro de Frecuencia de Sitio (EFS). El EFS es un estadístico que condensa la información obtenida de las secuencias de ADN de una muestra de una población. Además, en este capítulo se explica cómo utilizar este estadístico para hacer inferencia a los parámetros asociados algunos modelos de genética de poblaciones basados en el $\Lambda-$coalescente.

Por último, en el apéndice se da una breve mención a los conceptos y teoremas básicos y de mayor importancia para la comprensión topológica del espacio de estados sobre el cual está definido un proceso coalescente.

Se recomienda ampliamente al lector hacer una revisión a las Referencias Bibliográficas asociadas a los temas de su interés particular.

Capítulo 1

El Teorema de De Finetti

Una noción importante tanto en estadística bayesiana como en la teoría de coalescencia es el concepto de intercambiabilidad. El teorema más importante referente a la intercambiabilidad es el **Teorema de De Finetti**.

1.1. El enunciado del Teorema

En estadística es muy común encontrarse con problemas en los que el orden de los datos no es tan importante como los datos en sí. Algunas veces incluso el orden de los datos es solo una etiqueta que se les asigna para poder manejarlos. Cuando el orden de nuestros datos no es importante decimos que contamos con la propiedad de intercambiabilidad.

Definición 1. *Un vector aleatorio finito es* ***intercambiable*** *si tiene la misma distribución que cualquier permutación de él. Es decir, cuando*

$$(X_1, \ldots, X_n) \stackrel{d}{=} (X_{\sigma(1)}, \ldots, X_{\sigma(n)}),$$

para cada permutación σ *de* $(n) = \{1, \ldots, n\}$.

Definición 2. *Una sucesión infinita de variables aleatorias* $\{X_i\}_{i=1}^{\infty}$ *es* ***intercambiable*** *cuando* $(X_1, \ldots, X_n)$ *es un vector finito intercambiable, para cada* $n \in \mathbb{N}$.

Claramente cualquier sucesión finita o infinita de variables aleatorias independientes e idénticamente distribuidas (i.i.d) es intercambiable. ¿Será que solo las variables aleatorias i.i.d. son intercambiables?. Si existen otras variables aleatorias intercambiables, ¿Podremos construirlas? ¿Podremos caracterizar a todas? La respuesta a estas preguntas la tenemos en el Teorema de De Finetti.

Notación 1. *Si E es un espacio topológico, entonces $\mathcal{B}(E)$ denotará a la σ-álgebra de Borel de dicho espacio.*

Teorema 1 (Teorema de De Finetti)**.** *Consideremos a una sucesión de variables aleatorias reales $\{X_i\}_{i=1}^{\infty}$, definidas en el espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$. Sea además $\mathcal{M}(\Omega)$ el espacio de todas las medidas de probabilidad en $(\Omega, \mathcal{F})$. Entonces la sucesión es intercambiable si y solo si existe una medida de probabilidad α, definida sobre $\mathcal{M}(\Omega)$, tal que para cualesquiera $B_1, B_2 \ldots, B_n \in \mathcal{B}(\mathbb{R})$ se tiene que*

$$\mathbb{P}[X_1 \in B_1, \ldots, X_n \in B_n] = \int_{\mathcal{M}(\Omega)} \prod_{i=1}^{n} P[X_i \in B_i] \alpha(dP).$$

*A la medida α se le llama **medida de De Finetti** para la sucesión intercambiable $\{X_i\}_{i=1}^{\infty}$.*

Intuitivamente el Teorema de De Finetti nos dice que condicionado a una medida aleatoria, las variables aleatorias intercambiables son i.i.d. La demostración del Teorema de De Finetti se puede encontrar en [1] y una versión más general (para variables aleatorias que no necesariamente toman valores reales) se puede encontrar en [14]. Aunque la prueba sea bastante elemental, nosotros solo la construiremos para el caso de sucesiones de variables aleatorias binarias (variables aleatorias Bernoulli). Para esto nos basaremos en [12], ya que la prueba proporcionada en este libro será más ilustrativa para la construcción del Λ-coalescente. El resto de lemas, teoremas y corolarios del capítulo estarán enfocados en realizar la prueba de dicho teorema.

1.2. Las sucesiones de incrementos

El objetivo de este capítulo es construir la medida de De Finetti para el caso binario. La construcción de dicha medida se dará mediante sucesiones completamente monótonas. En esta sección se definirán las sucesiones completamente monónotonas mediante sucesiones de incrementos y se darán algunos resultados que se requieren para dicha construcción.

Definición 3. *Dada una sucesión de números reales* $\{a_i\}_{i=0}^{\infty}$ *y* $r \geq 0$, *definimos inductivamente a su* r***-ésima sucesión de incrementos*** *mediante:*

$$\Delta^0 a_i := a_i, \quad \Delta^1 a_i := a_{i+1} - a_i$$

y

$$\Delta^{r+1} a_i := \Delta^r a_{i+1} - \Delta^r a_i.$$

De manera natural, las derivadas múltiples de una función se relacionan con las sucesiones de incrementos.

Proposición 1. *Sea* f *una función de clase* C^{k+1}. *Para cada* $x \in \mathbb{R}$, *definimos la sucesión* $\left\{x_k^{(n)}\right\}_{k=0}^{\infty}$ *mediante* $x_k^{(n)} = f(x + k/n)$. *Entonces*

$$f^{(k)}(x) = \lim_{n\to\infty} n^k \Delta^k x_0^{(n)},$$

donde $f^{(k)}(x)$ *denota la* $k-$*ésima derivada de* f *en* x.

La prueba de la proposición anterior se puede obtener utilizando los primeros términos de la aproximación de Taylor a la k-ésima derivada $f^{(k)}$ alrededor del punto $(x + 1/n)$.

La siguiente proposición nos dice como obtener r-ésima sucesión de incrementos sin necesidad de inducción.

Proposición 2. *Para toda sucesión* $\{a_i\}_{i=0}^{\infty}$ *y* $r \geq 0$ *se tiene que*

$$\Delta^r a_i = \sum_{j=0}^{r} \binom{r}{j} (-1)^{r-j} a_{i+j}. \tag{1.1}$$

Demostración. Probaremos esto para toda i y por inducción sobre r. Los casos $r = 0$ y $r = 1$ se cumplen por definición. Supongamos ahora que para $r \geq 1$ se cumple (1.1) para toda i. Entonces tenemos

$$\begin{aligned}
\Delta^{r+1} a_i =& \Delta^r a_{i+1} - \Delta^r a_i \\
=& \sum_{j=0}^{r} \binom{r}{j} (-1)^{r-j} a_{i+j+1} - \sum_{j=0}^{r} \binom{r}{j} (-1)^{r-j} a_{i+j} \\
=& \sum_{j=1}^{r+1} \binom{r}{j-1} (-1)^{r-j+1} a_{i+j} - \sum_{j=0}^{r} \binom{r}{j} (-1)^{r-j} a_{i+j} \\
=& a_{i+r+1} + (-1)^{r+1} a_i + \sum_{j=1}^{r} \frac{(r+1)!}{(r+1-j)!j!} \frac{j}{r+1} (-1)^{r+1-j} a_{i+j} \\
& - \sum_{j=1}^{r} \frac{(r+1)!}{(r+1-j)!j!} \frac{r+1-j}{r+1} (-1)^{r-j} a_{i+j} \\
=& a_{i+r+1} + (-1)^{r+1} a_i + \sum_{j=1}^{r} \binom{r+1}{j} (-1)^{r+1-j} a_{i+j} \left[\frac{j+r+1-j}{r+1} \right] \\
=& \sum_{j=0}^{r+1} \binom{r+1}{j} (-1)^{r_1-j} a_{i+j},
\end{aligned}$$

por lo que la proposición se cumple para toda r y para toda i. □

El siguiente Teorema da una importante relación entre las sucesiones de incrementos de dos series cualesquiera.

Teorema 2 (Fórmula General de Reciprocidad)**.** *Para cualesquiera sucesiones* $\{a_i\}_{i=0}^{\infty}$ *y* $\{c_i\}_{i=0}^{\infty}$ *se tiene que*

$$\sum_{r=0}^{v} c_r \binom{v}{r} \Delta^r a_i = \sum_{j=0}^{v} a_{i+j} \binom{v}{j} (-1)^{v-j} \Delta^{v-j} c_j.$$

Demostración. Tenemos:

$$\begin{aligned}
\sum_{r=0}^{v} c_r \binom{v}{r} \Delta^r a_i &= \sum_{r=0}^{v} c_r \binom{v}{r} \sum_{j=0}^{r} \binom{r}{j} (-1)^{r-j} a_{i+j}, \text{ por Proposición 2,} \\
&= \sum_{j=0}^{v} \sum_{r=0}^{v-j} c_{r+j} \binom{v}{r+j} \binom{r+j}{j} (-1)^r a_{i+j} \\
&= \sum_{j=0}^{v} \sum_{r=0}^{v-j} c_{r+j} \binom{v}{j} \binom{v-j}{r} (-1)^{j-v} (-1)^{v-j-r} a_{i+j} \\
&= \sum_{j=0}^{v} a_{i+j} \binom{v}{j} (-1)^{v-j} \Delta^{v-j} c_j, \text{ de nuevo por Proposición 2,}
\end{aligned}$$

por lo que la proposición es cierta. □

La Proposición 2 nos dice como obtener todas las r-ésimas sucesiones de incrementos de $\{a_i\}$ directamente. El siguiente es un corolario del Teorema 2, que nos dice como obtener $\{a_i\}$ directamente de los primeros términos de sus r-ésimas sucesiones de incrementos.

Corolario 1 (Fórmula de Inversión)**.** *Para cualquier sucesión $\{b_i\}_{i=0}^{\infty}$ y para cualquier $v \in \mathbb{N}$ se tiene que*

$$b_k = \sum_{j=0}^{v} \binom{v}{j} (-1)^{v-j} \Delta^{v-j} b_{j+k}.$$

Demostración. Aplicar la Fórmula General de Reciprocidad (Teorema 2) a las sucesiones $a_i = 1$ y $c_i = b_{i+k}$. □

La siguiente observación es un caso particular de la Fórmula de Inversión. Más adelante se utilizará dicha observación para la construcción de una serie de medidas de probabilidad que convergen débilmente a la medida de De Finetti buscada en este capítulo.

Observación 1. *En particular se tiene que*

$$b_0 = \sum_{j=0}^{v} \binom{v}{j} (-1)^{v-j} \Delta^{v-j} b_j.$$

1.3. Los Momentos y las Sucesiones Completamente Monótonas

El Teorema 3, que veremos en esta sección, da la relación entre las sucesiones completamente monótonas, que también se definirán en esta sección, y los momentos de una distribución. Estudiaremos este resultado con el enfoque probabilista de Feller [12], pero utilizando conceptos más actuales. Para probar dicho teorema haremos uso de algunos lemas.

Lema 1. *Si* $\{c_k\}_{k=0}^{\infty}$ *es la sucesión de momentos de una variable aleatoria* X, *entonces*

$$(-1)^r \Delta^r c_k = \mathbb{E}\left[X^k (1-X)^r\right].$$

Demostración. Tenemos:

$$\begin{aligned}
(-1)^r \Delta^r c_k &= (-1)^r \sum_{j=0}^{r} \binom{r}{j} (-1)^{r-j} \mathbb{E}\left[X^{k+j}\right], \text{ por Proposición 2,} \\
&= \sum_{j=0}^{r} \binom{r}{j} (-1)^{j} \mathbb{E}\left[X^{k+j}\right] \\
&= \mathbb{E}\left[X^k \sum_{j=0}^{r} \binom{r}{j} (-X)^j\right] \\
&= \mathbb{E}\left[X^k (1-X)^r\right],
\end{aligned}$$

por lo que dicho lema se cumple. □

Definición 4. *Una sucesión* $\{a_k\}_{k=0}^{\infty}$ *se llama* ***completamente monótona*** *si, para toda* i *y* r, *cumple que*

$$(-1)^r \Delta^r a_i \geq 0.$$

Observación 2. *Por el Lema 1, si* X *es una variable aleatoria tal que* $0 \leq X \leq 1$ *casi seguramente, entonces su sucesión de momentos es completamente monótona.*

Como nos interesa relacionar variables aleatorias con sucesiones completamente monótonas y viceversa, durante el resto del capítulo consideraremos solo variables

aleatorias acotadas con probabilidad 1 en $[0,1]$.

Lema 2. *Sea $\{c_i\}_{i=0}^{\infty}$ una sucesión completamente monótona tal que $c_0 = 1$. Para cada $n \in \mathbb{N}$ definimos los valores*

$$p_j^{(n)} = \binom{n}{j}(-1)^{n-j}\Delta^{n-j}c_j \quad 0 \leq j \leq n.$$

Entonces $\left\{p_j^{(n)}\right\}_{j=0}^{n}$ son los pesos de una medida de probabilidad discreta.

Demostración. Ya que $\{c_j\}_{j=0}^{\infty}$ es completamente monótona entonces $p_j^{(n)} \geq 0$. Solo falta ver que dichos valores suman uno. Por la Observación 1, de la fórmula de inversión, tenemos que:

$$\sum_{i=0}^{n} p_j^{(n)} = \sum_{i=0}^{n} \binom{n}{j}(-1)^{n-j}\Delta^{n-j}c_j = c_0.$$

□

Durante el resto de la sección denotaremos mediante μ_n a las medidas cuyos pesos $p_j^{(n)}$ están asociados a los puntos j/n. Además denotaremos mediante F_n a las funciones de distribución asociadas a dichas medidas y mediante X_n a las variables aleatorias con función de distribución respectiva F_n. Notamos además que μ_n, F_n y X_n están definidas en función de una sucesión completamente monótona $\{c_j\}_{j=0}^{\infty}$. Además, podemos definir al r-ésimo momento de X_n mediante $\mathbb{E}[X_n^r] := \int x^r \mu_n(dx)$.

Lema 3. *El r-ésimo momento de X_n converge a c_r cuando n tiende a infinito.*

$$\lim_{n\to\infty} \mathbb{E}[X_n^r] = c_r.$$

Demostración. Tenemos

$$
\begin{aligned}
\lim_{n\to\infty} \mathbb{E}[X_n^r] &= \lim_{n\to\infty} \sum_{j=0}^{\infty} \left(\frac{j}{n}\right)^r p_j^{(n)} \mathbb{1}_{\{j\leq n\}} \\
&= \lim_{n\to\infty} \sum_{j=0}^{\infty} \left(\frac{j}{n}\right)^r \binom{n}{j} (-1)^{n-j} \Delta^{n-j} c_j \mathbb{1}_{\{j\leq n\}} \\
&= \lim_{n\to\infty} \sum_{j=0}^{\infty} c_j \binom{n}{j} \Delta^j a_0 \mathbb{1}_{\{j\leq n\}}, \text{ por Teorema 2 con } a_j = \left(\frac{j}{n}\right)^r, \text{ para } 0 \geq j \geq n, \\
&= \lim_{n\to\infty} \sum_{j=0}^{\infty} \left[\frac{c_j}{j!}\right] \left[\frac{n!}{(n-j)!} \frac{1}{n^j}\right] \left[n^j \Delta^j a_0\right] \mathbb{1}_{\{j\leq n\}} \\
&= \sum_{j=1}^{\infty} \left[\frac{c_j}{j!}\right] \lim_{n\to\infty} \left[\frac{n!}{(n-j)!} \frac{1}{n^j}\right] \lim_{n\to\infty} \left[n^j \Delta^j a_0\right] \mathbb{1}_{\{j\leq n\}} \\
&= \sum_{j=1}^{\infty} \left[\frac{c_j}{j!}\right] \lim_{n\to\infty} \left[n^j \Delta^j a_0\right] \\
&= \sum_{j=1}^{\infty} \left[\frac{c_j}{j!}\right] f^{(j)}(0), \text{ esto de acuerdo a la Proposición 1 con } f(x) = x^r, \\
&= \left[\frac{c_r}{r!} r!\right] \\
&= c_r,
\end{aligned}
$$

por lo que tenemos la convergencia. □

Teniendo los lemas anteriores, podemos concluir la sección con el siguiente teorema, el cual nos dice que a cada sucesión completamente monótona le corresponde una medida finita. Para dicho teorema denotaremos mediante X_μ a la variable aleatoria con ley μ.

Teorema 3. *Si μ es una medida de probabilidad concentrada en $[0,1]$, entonces los momentos $\{c_r\}_{r=0}^{\infty}$ de X_μ forman una sucesión completamente monótona con $c_0 = 1$. Recíprocamente, para cada sucesión completamente monótona $\{c_r\}_{r=0}^{\infty}$, con $c_0 = 1$, existe una única medida μ de probabilidad en $[0,1]$ tal que $\{c_r\}_{r=0}^{\infty}$ es la sucesión de momentos de X_μ.*

Demostración. Si $\{c_r\}_{r=0}^{\infty}$ es la sucesión de momentos de X_μ entonces por el Lema 1

la sucesión es completamente monótona.

Supongamos ahora que $\{c_r\}_{r=0}^{\infty}$ es completamente monótona. Entonces por el Lema 2, tenemos definidas a las variables aleatorias X_n y a sus respectivas funciones de distribución F_n. En esta demostración veremos que las variables aleatorias X_n convergen en distribución. Esta prueba está basada en el ejemplo d) del Teorema 2 del capítulo XIII del libro [12] (página 251).

Ya que las medidas μ_n están concentradas en $[0,1]$, entonces para ver la convergencia débil de dichas medidas solo necesitamos probar que $\int g d\mu_n$ converge a un límite finito para cada g continua en $[0,1]$. Ya que toda función continua en $[0,1]$ puede ser aproximada uniformemente en $[0,1]$ mediante una sucesión de polinomios, entonces solo nos falta ver que $\int x^k d\mu_n$ converge a un límite finito para toda k. Por el Lema 3, tenemos que F_n converge débilmente a una distribución digamos F. Ya que cada F_n tiene soporte en $[0,1]$ y F_n converge a F, entonces F también está concentrada en $[0,1]$.

Definimos entonces a la medida μ mediante la medida asociada a la función de distribución F. Nos falta ver que efectivamente la sucesión de momentos de X_μ es $\{c_r\}_{r=0}^{\infty}$. Sin embargo, está propiedad también se deduce de la convergencia en distribución, del Lema 3 y del hecho de que todas las medidas están concentradas en [0,1].

Para concluir la demostración de este teorema notamos que la unicidad se debe de nuevo a que los momentos caracterizan a una distribución con soporte acotado. $\square$

La relación entre las sucesiones completamente monótonas y la sucesión de momentos asociada a una distribución de probabilidad es poco intuitiva. La intuición pudiera encontrarse en teorema mucho más importante que asocia a las funciones completamente monótonas (versión continua de las sucesiones completamente monótonas) y las transformadas de Laplace. Para leer más sobre esta relación y sus aplicaciones a procesos estocásticos se puede consultan consultar las notas de Zenghu [17].

1.4. Teorema de De Finetti para Sucesiones Binarias

Ahora sí estamos en condiciones de probar el Teorema de De Finetti para el caso binario.

Teorema 4. *Si $\{X_i\}_{i=1}^{\infty}$ es una sucesión de variables aleatorias Bernoulli intercambiables, existe una única medida de probabilidad μ en $[0,1]$ tal que, para cada $0 \leq k \leq n < \infty$*

$$\mathbb{P}[X_1 = 1, \ldots, X_k = 1, X_{k+1} = 0, \ldots, X_n = 0] = \int_0^1 p^k (1-p)^{n-k} d\mu(p). \tag{1.2}$$

Demostración. Denotemos al lado izquierdo de la ecuación (1.2) mediante $p_{k,n}$. Se tiene la relación

$$p_{n-1,n-1} = p_{n-1,n} + p_{n,n}, \tag{1.3}$$

para $n \geq 2$. Sea $\{c_k\}$ la sucesión dada mediante $c_0 = 0$ y, para $n \geq 1$ $c_n = p_{n,n}$. Entonces la ecuación (1.3) se puede escribir mediante

$$p_{n-1,n} = -(p_{n,n} - p_{n-1,n-1}) = -\Delta c_{n-1}. \tag{1.4}$$

Probaremos ahora, mediante inducción sobre $m = n - k$, que en general se tiene

$$p_{k,n} = (-1)^{n-k} \Delta^{n-k} c_k.$$

Cuando $m = n - k = 1$ la hipótesis de inducción se cumple gracias a la ecuación (1.4). Supongamos ahora que la hipótesis se cumple para $m = n - k$. Entonces, por intercambiabilidad tenemos

$$\begin{aligned} p_{k,n+1} &= p_{k,n} - p_{k+1,n+1} \\ &= (-1)^{n-k} \Delta^{n-k} c_k - (-1)^{n-k} \Delta^{n-k} c_{k+1} \\ &= (-1)^{n+1-k} \Delta^{n+1-k} c_k. \end{aligned} \tag{1.5}$$

Por lo tanto tenemos que $p_{k,n} = (-1)^{n-k}\Delta^{n-k}c_k$, para todas $0 \leq k \leq n$. Además tenemos que todas las constantes $p_{k,n}$ son probabilidades de eventos, por lo que son no negativas. Entonces $\{c_k\}$ es una sucesión completamente monótona. Luego, por el Teorema 3, $\{c_k\}$ es la sucesión de momentos de X_μ. Entonces por el Lema 1 se cumple el teorema. $\square$

Intuitivamente este teorema nos dice que si tenemos una sucesión de variables aleatorias Bernoulli intercambiables, entonces condicionando sobre una $p \in [0,1]$ aleatoria, estas variables aleatorias son independientes e idénticamente distribuidas Bernoulli de parámetro p. Este es un caso particular del Teorema de De Finetti (Teorema 1) que intuitivamente nos dice que las variables aleatorias intercambiables son iid condicionando con la medida de De Finetti.

Capítulo 2

Λ-Coalescentes

En este capítulo se construyen los Λ-coalescentes, como lo introdujo Pitman en [20].

2.1. Las particiones y los coalescentes

Los coalescentes son procesos estocásticos con espacio de estados en particiones. Es por esto que es importante entender a las particiones y a su topología.

2.1.1. Particiones de $\mathbb{N}$

En este capítulo estudiaremos procesos estocásticos de partículas que se van fusionando con el tiempo. La forma de caracterizar dichas partículas será mediante particiones de los números naturales.

Notación 2. *Denotaremos al conjunto de los primeros n naturales mediante $(n) := \{1, 2, \ldots, n\}$.*

Definición 5. *Una **partición** π de $A \subset \mathbb{N}$ es una relación de equivalencia entre los elementos de A. A las clases de equivalencia de dicha relación los llamaremos **bloques** de π. Dichos bloques, $\pi^1, \pi^2 \ldots$, se considerarán enlistados en el orden creciente que se obtiene en base al mínimo elemento de cada bloque. Es decir, el mínimo*

elemento de A se encuentra en π^1, el mínimo elemento de $A \setminus \pi^1$ se encuentra en π^2 y así sucesivamente.

Ejemplo 1. *Consideremos a la partición π de (n), con $n = 6$, dada mediante:*

$$\pi = \Big\{\pi^1 := \{1,4\}, \pi^2 := \{2,3,5\}, \pi^3 := \{6\}\Big\}.$$

En este ejemplo los mínimos elementos de cada bloque son 1,2 y 6 respectivamente. Por lo cual la partición es presentada en dicho orden.

Notación 3. *Al conjunto de todas las posibles particiones de (n) lo denotaremos mediante $\mathcal{P}^n$ y al conjunto de todas las posibles particiones de $\mathbb{N}$ lo denotaremos mediante $\mathcal{P}^\infty$. Hablaremos de $\mathcal{P}^n$, con $n \leq \infty$, para referiros a particiones que pueden ser de $n \in \mathbb{N}$ o de todo $\mathbb{N}$.*

Ejemplo 2. $\mathcal{P}^3 = \{\pi_1, \pi_2, \pi_3, \pi_4, \pi_5\}$, *donde* $\pi_1 = \Big\{\{1\},\{2\},\{3\}\Big\}$, $\pi_2 = \Big\{\{1\},\{2,3\}\Big\}$, $\pi_3 = \Big\{\{1,2\},\{3\}\Big\}$, $\pi_4 = \Big\{\{1,3\},\{2\}\Big\}$ *y* $\pi_5 = \Big\{\{1,2,3\}\Big\}$.

Notación 4. *Como cada partición $\pi \in \mathcal{P}^n$, $n \leq \infty$, nos representa a una relación de equivalencia, denotaremos a dicha relación mediante*

$$i \overset{\pi}{\sim} j,$$

o simplemente $i \sim j$, en caso de no haber ambigüedad. Es decir, $i \sim j$ significa que tanto i como j pertenecen al mismo bloque $B \in \pi$.

Definición 6. *Dada una partición $\pi_1 \in \mathcal{P}^k$ con $k \in \mathbb{N} \cup \{\infty\}$ conformada por los bloques $\pi_1 = \{A_1, A_2, \ldots, A_m\}$, con $m \leq \infty$, decimos que $\pi_1 \in \mathcal{P}^k$ es un* ***refinamiento*** *de una partición $\pi_2 \in \mathcal{P}^k$ si todo bloque $B \in \pi_2$ puede ser expresado de la forma*

$$B = \bigcup_{j \in J} A_j,$$

para algún subconjunto $J \subseteq \{1, \ldots m\}$.

Ejemplo 3. *El conjunto de los singuletes* $\{\{1\},\{2\},\dots,\{k\}\}$ *es refinamiento de cualquier partición de* $\mathcal{P}^k$*, mientras que cualquier partición de* $\mathcal{P}^k$ *es un refinamiento del conjunto total* $(k)=\{1,2,\dots k\}$.

2.1.2. La topología de las particiones

Para poder trabajar con procesos estocásticos definidos en $\mathcal{P}^\infty$, necesitaremos asignarle una topología a las particiones. En esta subsección construiremos y entenderemos la topología dada por Pitman ([20]), a la que llamaremos **topología usual**. Para poder dar la construcción se hará uso de los resultados y definiciones dados en el Apéndice A. Además será necesario fijar la definición y la notación de las proyecciones entre particiones.

Definición 7. *Dados* $1\leq m\leq n\leq\infty$*, se le llama* ***proyección*** *de* $\mathcal{P}^n$ *a* $\mathcal{P}^m$ *a la función* $\phi_m^n:\mathcal{P}^n\to\mathcal{P}^m$ *tal que, para cualesquiera* $i,j\in(m)$ *y* $\pi\in\mathcal{P}^n$*, se tiene que*

$$i\overset{\phi_m^n(\pi)}{\sim}j\iff i\overset{\pi}{\sim}j.$$

Ejemplo 4. $\phi_3^5\left(\left\{\{1,4\},\{2,3\},\{5\}\right\}\right)=\left\{\{1\},\{2,3\}\right\}$.

Comencemos por dotar a cada espacio $\mathcal{P}^k$ con la topología discreta (ver Apéndice A) y por lo tanto a $\prod_{k=1}^\infty\mathcal{P}^k$ con la topología producto que resulta de considerar dichas topologías discretas. Consideramos ahora a la función $\phi:\mathcal{P}^\infty\to\prod_{k=1}^\infty\mathcal{P}^k$ dada por

$$\phi(\pi)=\left(\phi_1^\infty(\pi),\phi_2^\infty(\pi)\dots\right).$$

La imagen de esta función $\phi(\mathcal{P}^\infty)$ es subconjunto de $\prod_{k=1}^\infty\mathcal{P}^k$, por lo que está dotada de manera natural con la topología de subconjunto de $\prod_{k=1}^\infty\mathcal{P}^k$. Siguiendo esta construcción natural la topología usual de $\mathcal{P}^\infty$ es simplemente la topología heredada por la función ϕ. Esto quiere decir que un abierto en $\mathcal{P}^\infty$ es la imagen inversa de un abierto en $\phi(\mathcal{P}^\infty)$. Observamos que ϕ es inyectiva, por lo que cada abierto de $\phi(\mathcal{P}^\infty)$ corresponde a un único abierto de $\mathcal{P}^\infty$.

Observación 3. *En $\mathcal{P}^\infty$ las proyecciones ϕ_i^∞ coinciden con las proyecciones ϕ_i de la Definición 34 el Apéndice A y además $\phi_j^i = \phi_j \circ \phi_i^{-1}$.*

Observación 4. *El espacio $\prod_{k=1}^\infty \mathcal{P}^k$ es compacto.*

Demostración. Cada espacio $\mathcal{P}^k$ es finito y tiene la topología discreta, por lo tanto cada espacio $\mathcal{P}^k$ es compacto. Por el Teorema de Tychonoff (del Apéndice A), tenemos que $\prod_{k=1}^\infty \mathcal{P}^k$ es un espacio compacto. □

Ya que $\prod_{k=1}^\infty \mathcal{P}^k$ es un espacio compacto, entonces solo tenemos que probar que $\phi(\mathcal{P}^\infty)$ es cerrado para tener que $\mathcal{P}^\infty$ es compacto con la topología usual.

Teorema 5. *El espacio $\mathcal{P}^\infty$ es compacto con la topología usual.*

Demostración. Veamos que el complemento $\phi(\mathcal{P}^\infty)^C$ es abierto.

$$\begin{aligned}
\phi(\mathcal{P}^\infty)^C &= \bigcup_{j=1}^\infty \bigcup_{\pi_j \in \mathcal{P}^j} \bigcup_{i=j+1}^\infty \left\{ \pi \in \prod_{k=1}^\infty \mathcal{P}^k \text{ tal que } \phi_j(\pi) = \pi_j \text{ y } \phi_j(\phi_i^{-1}(\phi_i(\pi))) \neq \pi_j \right\} \\
&= \bigcup_{j=1}^\infty \bigcup_{\pi_j \in \mathcal{P}^j} \bigcup_{i=j+1}^\infty \left[\phi_j^{-1}(\pi_j) \bigcap \left(\phi_i^{-1}(\phi_i(\phi_j^{-1}(\pi_j)))\right)^C \right] \\
&= \bigcup_{j=1}^\infty \bigcup_{\pi_j \in \mathcal{P}^j} \bigcup_{i=j+1}^\infty \left[\phi_j^{-1}(\pi_j) \bigcap \phi_i^{-1}(\left[\phi_i(\phi_j^{-1}(\pi_j))\right]^C) \right].
\end{aligned}$$

Ya que la topología producto está generada por las imágenes inversas de las proyecciones (ver la Definición 35 del Apéndice A), tenemos que $\phi(\mathcal{P}^\infty)^C$ es abierto. Por lo que $\mathcal{P}^\infty$ es compacto con la topología usual. □

Analicemos con más detalle a la subbbase $\mathcal{S}$ que genera a la topología usual de $\mathcal{P}^\infty$. Ya que $\prod_{k=1}^\infty \mathcal{P}^k$ está dotado de la topología producto, entonces es generado por uniones de intersecciones de conjuntos de la forma $\phi_k^{-1}(A)$, donde A es cualquier subconjunto de $\mathcal{P}^k$. Ya que $\mathcal{P}^k$ es finito, entonces existen $\pi_k^1, \ldots, \pi_k^j \in \mathcal{P}^k$ tales que

$A = \bigcup_{i=1}^{j}\{\pi_k^i\}$. Por lo que $\phi_k^{-1}(A) = \phi_k^{-1}(\bigcup_{i=1}^{j}\{\pi_k^i\}) = \bigcup_{i=1}^{j}\phi_k^{-1}(\{\pi_k^i\})$. Esto quiere decir que el conjunto

$$\widetilde{\mathcal{S}} = \{\phi_k^{-1}(\pi_k) : k \geq 1 \text{ y } \pi_k \in \mathcal{P}^k\}$$

es una subbase del producto $\prod_{k=1}^{\infty}\mathcal{P}^k$. Por lo tanto el conjunto

$$\widehat{\mathcal{S}} = \{\phi_k^{-1}(\pi_k) \cap \phi(\mathcal{P}^\infty) : k \geq 1 \text{ y } \pi_k \in \mathcal{P}^k\}$$

es una subbase del subconjunto $\phi(\mathcal{P}^\infty) \subset \prod_{k=1}^{\infty}\mathcal{P}^k$.

Hasta ahora $\widehat{\mathcal{S}}$ parece solo definida mediante notación engorrosa. Para entender un poco más a $\widehat{\mathcal{S}}$, nos fijaremos en el conjunto $\phi_k^{-1}(\pi_k) \cap \phi(\mathcal{P}^\infty)$. Tenemos

$$\phi_k^{-1}(\pi_k) = \left\{(\mu_1, \mu_2, \dots) \in \prod_{l=1}^{\infty}\mathcal{P}^l : \mu_k = \pi_k\right\}.$$

Por lo que

$$\phi_k^{-1}(\pi_k) \cap \phi(\mathcal{P}^\infty) = \{(\mu_1, \mu_2, \dots) \in \phi(\mathcal{P}^\infty) : \mu_k = \pi_k\}.$$

Además

$$\phi^{-1}(\phi_k^{-1}(\pi_k) \cap \phi(\mathcal{P}^\infty)) = \{\pi \in \mathcal{P}^\infty : \phi_k^\infty(\pi) = \pi_k\},$$

por lo que la subbase de $\mathcal{P}^\infty$ con la topología usual es

$$\mathcal{S} = \bigcup_{k=1}^{\infty}\bigcup_{\pi_k \in \mathcal{P}^k}\{\{\pi \in \mathcal{P}^\infty : \phi_k(\pi) = \pi_k\}\}. \tag{2.1}$$

En la Figura 2-1 tenemos dibujado a un árbol infinito que en cada nivel tiene a los espacios de todas las posibles particiones de (k), es decir $\mathcal{P}^k$ o también $\phi_k(\mathcal{P}^\infty)$. En este árbol cada nuevo nivel $n+1$ se obtiene agregando el entero $n+1$ a todas las posibles particiones de n elementos. El espacio $\mathcal{P}^\infty$ puede ser pensado como las hojas de este árbol. Entonces un elemento de la subbase $\mathcal{S}$ puede ser pensado como todos los descendientes en $\mathcal{P}^\infty$ de una partición π_k en $\mathcal{P}^k$. Más adelante veremos que $\mathcal{S}$ es de hecho una base.

Siguiendo la intención de comprender a la topología usual, definiremos una métrica

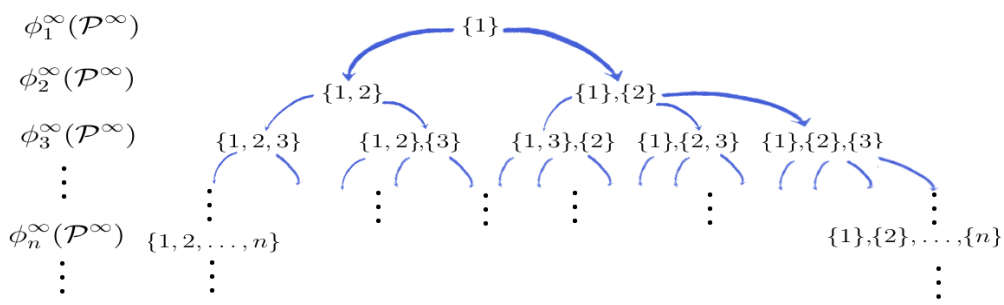

Figura 2-1: Gráfica del árbol de $\mathcal{P}^\infty$.

en $\mathcal{P}^\infty$. Esta métrica es típica de las estructuras de árbol.

Definición 8. *Llamaremos distancia (o métrica) de árbol a la función dada por:*

$$d(\pi, \pi') = \frac{1}{\text{máx}\,\{k : \phi_k(\pi) = \phi_k(\pi')\}},$$

cuando $\pi, \pi' \in \mathcal{P}^\infty$ son distintos y cero cuando son iguales.

Para visualizar a las bolas de la métrica de árbol volveremos a pensar en el árbol de la Figura 2-1. Pensando en este árbol, notamos que el tener una distancia entre dos particiones π y π' igual a $1/k$, significa que el ancestro común más reciente entre π y π' se encuentra en el nivel k. Utilizando este razonamiento, se observa que la bola con radio $\varepsilon > 0$ y centro en la partición π, $B_\varepsilon(\pi)$, consta de todos los descendientes de $\phi_k(\pi)$, donde k es el menor entero tal que $1/k < \varepsilon$.

Observación 5. *La topología generada por la métrica de árbol coincide con la topología usual de $\mathcal{P}^\infty$.*

Demostración. Los elementos de la subbase de la topología usual coinciden con las bolas de la métrica de árbol. □

Observación 6. *La subbase de la topología usual $\mathcal{S}$ es en realidad una base.*

Demostración. Ya que la subbase coincide con las bolas y las bolas son base de la topología generada por una métrica, entonces $\mathcal{S}$ es en realidad una base para la topología usual. □

Corolario 2. *El espacio $\mathcal{P}^\infty$ es métrico compacto (con la métrica de árbol).*

2.1.3. Los Coalescentes

Ya que nuestros espacios $\mathcal{P}^n$ y $\mathcal{P}^\infty$ son espacios métricos compactos (con las topologías discreta y de árbol respectivamente), el Teorema de Extensión de Kolmogórov nos permite definir procesos estocásticos en dichos espacios. Nosotros estamos interesados en los procesos coalescentes.

La coalescencia es la propiedad de dos o más materiales de unirse en un solo cuerpo. En probabilidad un coalescente es un proceso estocástico donde cada partícula está representada mediante un conjunto y la fusión de varias partículas está representada como la unión de conjuntos.

Definición 9. *Se le llama* ***coalescente*** *a un proceso estocástico* $\{\Pi_t\}_{t\geq 0}$*, con espacio de estados en* $\mathcal{P}^\infty$*, con trayectorias cadlag y tal que para cada* $t_1 < t_2$ Π_{t_1} *es un refinamiento de* Π_{t_2}*. En el caso en que el espacio de estados de* Π *sea* $\mathcal{P}^n$ *(en lugar de* $\mathcal{P}^\infty$*), se dirá que* Π *es un* $\mathbf{n}-$***coalescente****.*

La forma más usual de graficar un coalescente es mediante un dendograma. Cada linea del dendograma representa a una partícula presente en el tiempo t (ver Figura 2-2).

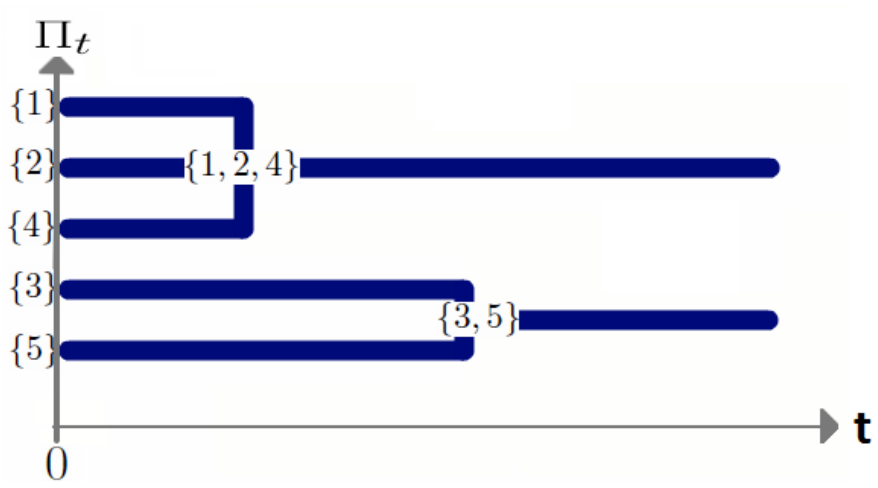

Figura 2-2: Gráfica de un 5-coalescente.

Nos interesa estudiar coalescentes que sean cadenas de Markov (para construcciones más precisas de dichos coalescentes ver [11]). Además queremos que todas

las partículas tengan el mismo comportamiento y por lo tanto la ley del proceso estocástico dependa solo de la cantidad de partículas presentes en cada tiempo. En otras palabras, necesitamos que la ley del proceso no dependa ni de la cardinalidad ni del contenido de cada partícula (bloque). Es en este sentido que se requiere una noción de intercambiabilidad en las particiones. Para definir esta intercambiabilidad, definiremos primero a la permutación de una partición.

Definición 10. *Dados $n \in \mathbb{N}$, $\pi \in \mathcal{P}^n$ y una permutación σ de (n), la* ***permutación*** *$\sigma(\pi)$ se define mediante:*

$$i \overset{\sigma(\pi)}{\sim} j \iff \sigma(i) \overset{\pi}{\sim} \sigma(j).$$

Ejemplo 5. *Para la partición $\pi \in \mathcal{P}^5$, dada mediante*

$$\pi = \{\{1,2\},\{3,4,5\}\},$$

y la permutación σ de (5), dada por

$$\sigma(5) = 1 \text{ y } \sigma(i) = i+1, \text{ para } i = 1,2,3,4,$$

la permutación $\sigma(\pi)$ está dada por

$$\sigma(\pi) = \{\{1,5\},\{2,3,4\}\}.$$

Ahora que tenemos definida una permutación en una partición, podemos dar una definición de intercambiabilidad en particiones (que resulta análoga a la Definición 1 del Capítulo 1).

Definición 11. *Sea $n \in \mathbb{N}$. Decimos que una* ***partición*** *aleatoria Π de (n) es* ***intercambiable*** *si conserva su ley bajo cualquier permutación. Es decir, si para cada permutación σ cumple la ecuación*

$$\Pi \overset{d}{=} \sigma(\Pi).$$

Decimos que una ***partición*** *aleatoria Π de $\mathbb{N}$ es* ***intercambiable*** *si cualquier*

proyección finita $\phi_n^\infty(\Pi)$ *es intercambiable.*

Por último, decimos que un ***coalescente*** $\{\Pi_t\}_{t\geq 0}$ *es* ***intercambiable*** *si* Π_t *es intercambiable para toda* t.

Un $\boldsymbol{\Xi}-$**coalescente** es un coalescente markoviano homogéneo intercambiable (ver [22]). En un Ξ-coalescente puede suceder que dos (o más) grupos de partículas se junten al mismo tiempo, formando dos (o más) partículas (una partícula por cada grupo). Un $\boldsymbol{\Lambda}-$**coalescente** es un Ξ-coalescente en el que solo se permite que un grupo de partículas se junte al mismo tiempo. En un Λ-coalescente puede suceder que un grupo de más de dos partículas se junte formando una sola partícula. Un **coalescente de Kingman** es un Λ-coalescente en el que solo se permite que se junten dos partículas para formar una (y por lo tanto no pueden juntarse grupos más grandes ni varios grupos al mismo tiempo). La Figura 2-3 nos muestra un ejemplo de cada uno de estos coalescentes.

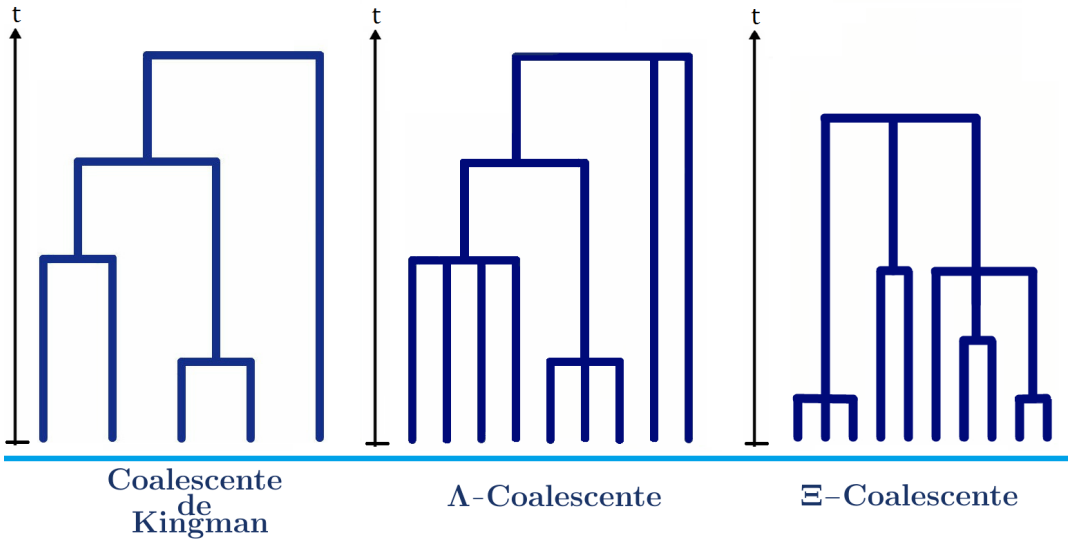

Figura 2-3: Ejemplos de coalescentes Kingman, Λ y Ξ.

En esta tesis solo nos interesaremos en los Λ-coalescentes, que en la literatura también suelen ser encontrados con el nombre de **coalescente con colisiones múltiples** o **coalescente simple**.

2.2. Λ-coalescentes

2.2.1. Caracterización

Para caracterizar una cadena de Markov homogénea en tiempo continuo con espacio de estados finito o numerable nos basta su generador infinitesimal. Gracias a la intercambibilidad de los Λ-coalescentes de (n) (para cada $n \in \mathbb{N}$), el generador infinitesimal queda determinado por un arreglo de tasas $\{\lambda_{b,k}\}_{2\leq l\leq b}$. Dichas tasas representan la intensidad con la que la cadena cambia de tener b partículas a tener $b-k+1$ partículas. Es decir, la intensidad con la que k partículas se fusionan en una sola. A menos que se especifique lo contrario, se acostumbra que todo Λ-coalescente comience con singuletes (bloques de un solo elemento). Además se acostumbra asignarle a $\lambda_{2,2}$ el valor 1, lo que equivaldría a "estandarizarlo").

Ejemplo 6. *Sea $\{\Pi_t\}_{t\geq 0}$ un Λ-coalescente de (n) con arreglo de tasas $\{\lambda_{b,k}\}_{2\leq k\leq b\leq n}$, entonces el tiempo T que dicho coalescente tarda en cambiar de estado por primera vez es una variable aleatoria con distribución exponencial de parámetro λ_T, donde*

$$\lambda_T = \lambda_{n,n} + \binom{n}{n-1}\lambda_{n,n-1} + \dots \binom{n}{2}\lambda_{n,2}.$$

Por el Teorema de extensión de Kolmogórov, para construir un Λ-coalescente de $\mathbb{N}$, necesitamos una propiedad de consistencia. Teniendo dicha consistencia podemos construir dicho coalescente como límite proyectivo de Λ-coalescentes de (n), $n \in \mathbb{N}$. Para cada $n \in \mathbb{N}$, la proyección ϕ^n_m del Λ-coalescente de (n) asociado al arreglo $\{\lambda_{b,k}\}_{2\leq k\leq b\leq n}$ debe de tener la misma ley que el Λ-coalescente de (m) asociado al arreglo $\{\lambda_{b,k}\}_{2\leq k\leq b\leq m}$.

La consistencia es equivalente a que al agregar partículas pueda generarse un proceso con el mismo arreglo de tasas. Por inducción bastará con que dicha consistencia se cumpla agregando una sola partícula. Esto se traduce en pedir que para cualesquiera b, k se cumple que $\lambda_{b,k} = \lambda_{b+1,k} + \lambda_{b+1,k+1}$ (ver Figura 2-4).

Definición 12. *Un **arreglo de tasas** $\{\lambda_{b,k}\}_{2\leq k\leq b}$, con $\lambda_{2,2} = 1$, se llama **consis-***

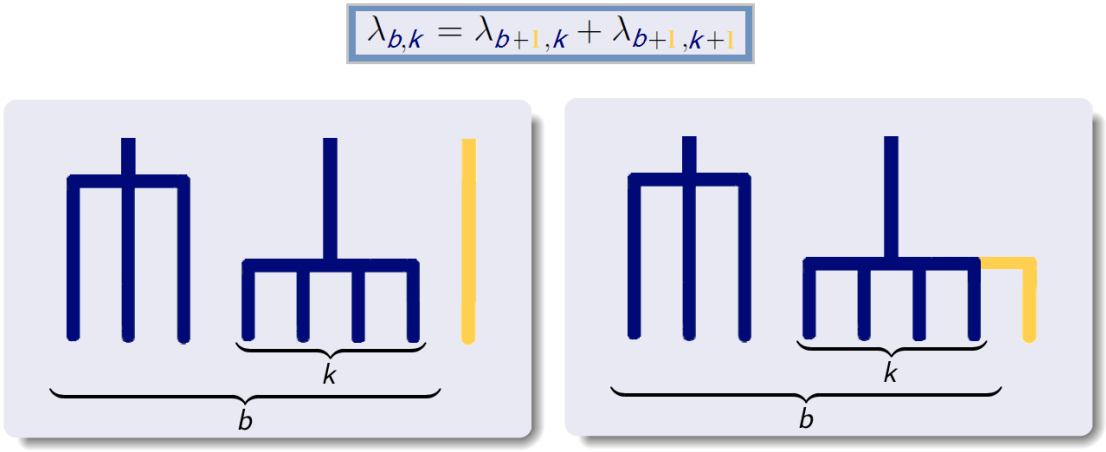

Figura 2-4: En este ejemplo "la rama" de color amarillo representa a la partícula que se agrega al grupo de partículas. La suma de las tasas representa a las dos posibilidades que tiene la partícula nueva (de unirse o no unirse al grupo).

tente *si para cualesquiera* $2 \leq k \leq b$ *se cumple la ecuación*

$$\lambda_{b,k} = \lambda_{b+1,k} + \lambda_{b+1,k+1}. \tag{2.2}$$

Teorema 6. *Para cada arreglo de tasas* $\{\lambda_{b,k}\}_{2\leq k\leq b<\infty}$ *consistente existe una única medida de probabilidad* Λ *con soporte contenido en* $[0,1]$, *tal que:*

$$\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k}\Lambda(dx), \qquad 2 \leq k \leq b. \tag{2.3}$$

Análogamente, para cada medida de probabilidad concentrada en $[0,1]$, *el arreglo de tasas definido mediante* (2.3) *cumple la propiedad de consistencia.*

Demostración. Es claro que para toda medida de probabilidad Λ concentrada en $[0,1]$ el arreglo de tasas definido mediante (2.3) cumple la propiedad de consistencia.

Supongamos ahora que tenemos un arreglo de tasas $\{\lambda_{b,k}\}_{2\leq k\leq b}$ consistente. Para probar la existencia de la medida Λ utilizaremos la prueba del Teorema de De Finetti para sucesiones binarias (1.4).

Sean

$$p_{i,n} = \lambda_{n+2,i+2}, \text{ para } 0 \leq i \leq n.$$

De la propiedad de consistencia del arreglo de tasas tenemos que

$$p_{i,n} = p_{i,n+1} + p_{i+1,n+1}.$$

De aquí reconocemos la relación con la ecuación (1.5). Además $p_{0,0} = \lambda_{2,2} = 1$. Entonces, siguiendo la prueba del Teorema de De Finetti existe una única medida Λ en $[0,1]$ tal que

$$p_{1,n} = \int_0^1 x^i (1-x)^{(n-i)} \Lambda(dx).$$

Por lo tanto

$$\lambda_{b,k} = p_{k-2,b-2} = \int_0^1 x^{k-2} (1-x)^{b-k} \Lambda(dx).$$

□

Hasta ahora todo el capítulo ha tenido como objetivo construir a los Λ-coalescentes. Condensaremos este trabajo mediante el siguiente teorema.

Teorema 7. *La ley de todo Λ-coalescente está determinada de manera única mediante una medida de probabilidad Λ concentrada en $[0,1]$.*

2.2.2. Algunos ejemplos de Λ-coalescentes

En esta subsección presentaremos algunos ejemplos de Λ-coalescentes, de acuerdo a la medida Λ utilizada en cada caso. Los ejemplos presentados en esta subsección son importantes por sus aplicaciones a genética de poblaciones.

Ejemplo 7. *En la subsección 2.1.3, se define al* ***coalescente de Kingman*** *como el Λ-coalescente que solo permite coalescencia de dos individuos. Ya que la medida Λ se encuentra estandarizada, esto significa que las tasas de coalescencia son de la forma $\lambda_{b,k} = \mathbb{1}_{\{k=2\}}$. Observamos entonces que estas tasas se obtienen cuando Λ es la medida de Dirac en cero δ_0 (medida que asigna el valor 1 a todo conjunto que contiene al 0 y asigna el valor 0 a todo conjunto que no lo contiene) . Observamos además que el tiempo que dura un coalescente de Kingman finito teniendo b partículas antes de colisionar es una variable aleatoria exponencial de parámetro $\binom{b}{2}$.*

En la subsección 2.3.1 se retomará este coalescente para mencionar una aplicación a genética de poblaciones.

Ejemplo 8. *El coalescente asociado a la medida uniforme en* $[0,1]$ *se llama* ***coalescente de Bolthausen-Sznitman****. Sus tasas de coalescencia se pueden calcular explícitamente usando la función Beta:*

$$\begin{aligned}\lambda_{b,k} =& \int_0^1 x^{k-2}(1-x)^{b-k}dx \\ =& B(k-1, b-k+1) \\ =& \frac{(k-2)!(b-k)!}{(b-1)!},\end{aligned}$$

donde B *representa a la función Beta. Originalmente este coalescente surge con relación a un concepto de física llamado spin-glass (ver [6]). Sin embargo, actualmente se puede utilizar para modelar poblaciones con fuerte selección natural (ver [9] y [23]).*

Ejemplo 9. *Otro ejemplo de* Λ *coalescente importante es con cuando tomamos* Λ *con distribución* $Beta(2-\alpha, \alpha)$*, para* $0 < \alpha < 2$*. Es decir*

$$\Lambda(dx) = \frac{\mathbb{1}_{[0,1]}(x)}{\Gamma(2-\alpha)\Gamma(\alpha)} x^{1-\alpha}(1-x)^{\alpha-1}dx.$$

Este coalescente se llama ***Beta coalescente****. Se hablará un poco más de este coalescente en la subsección 2.3.2.*

Ejemplo 10. *Solo como dato curioso, en el* Λ*-coalescente asociado a la delta de Dirac en uno* δ_1*, se tiene simplemente un tiempo exponencial de media 1 en el que todas las partículas coalescen de manera simultánea.*

Es fácil ver que el coalescente de Bolthausen-Sznitman es el caso particular del Beta coalescente con $\alpha = 1$. Además, Schweinsberg prueba que, cuando α se acerca a 2, el Beta coalescente converge débilmente a un coalescente de Kingman [21].

En la figura 2-5 se aparecen las simulaciones de varios Beta coalescentes, realizadas por G. Kersting. El coalescente de Kingman corresponde al "Beta coalescente" con $\alpha = 2$, que se puede ver a la izquierda. Y el coalescente de Bolthausen-Sznitman

corresponde al Beta coalescente con $\alpha = 1$. Observamos que a medida que $\alpha \to 0$ las coaliciones suelen abarcar mayor número de partículas. En el tercer capítulo de la tesis se utilizaran estos coalescentes para representar los árboles genealógicos de ciertas poblaciones. Intuitivamente podemos pensar que a mayor varianza en la tasa de reproducción de la población, será más apropiado ajustar un Beta coalescente con parámetro α más cercano a cero.

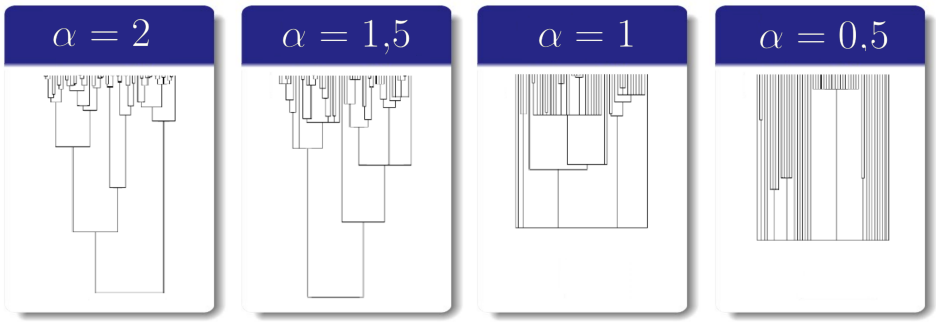

Figura 2-5: **Realización de Beta coalescentes**. En esta figura se muestran las simulaciones realizadas por G. Kersing de Beta Coalescentes que comienzan con 50 partículas. De izquierda a derecha los parámetros α de estos coalescentes son respectivamente 2, 1.5, 1 y 0.5.

2.3. Los Coalescentes como Límite de Modelos de Poblaciones

La **genética de poblaciones** es la rama de la genética cuyo objetivo es describir la distribución de las variaciones genéticas para explicar los fenómenos evolutivos. Para realizar este estudio, se define una **población** como un grupo de individuos de la misma especie y se reproducen entre ellos. Durante el resto de la sección se mencionarán algunas importantes aplicaciones de los Λ-coalescentes a modelos genéticos de poblaciones.

2.3.1. El modelo Wright-Fisher y el coalescente de Kingman

Uno de los modelos probabilistas más conocidos en genética de poblaciones es el modelo **Wright-Fisher**. Este modelo representa una población evolucionando en tiempo discreto. En el modelo Wright-Fisher se supone que la reproducción es asexual (no hay recombinación genética). Además se tiene la hipótesis de que no hay saltos de generaciones, lo que significa que un individuo de la generación $t \in \mathbb{Z}$ tiene hijos en la generación $t+1$. Para describir a este modelo imaginemos que en cada generación t hay un número constante de N individuos. Luego cada individuo de la generación $t+1$ escoge al azar y de manera independiente a su padre de entre los individuos de la generación t. Este proceso se repite en cada generación de manera independiente. Ahora fijamos una generación, la cual representará el presente. Luego tomamos una muestra de individuos de tamaño $n \leq N$, de dicha generación y rescatamos solo a dicha muestra y a todos sus ancestros, hasta llegar al ancestro común (ver Figura 2-6). Ahora diremos que dos individuos i y j, de entre los n individuos de la muestra,

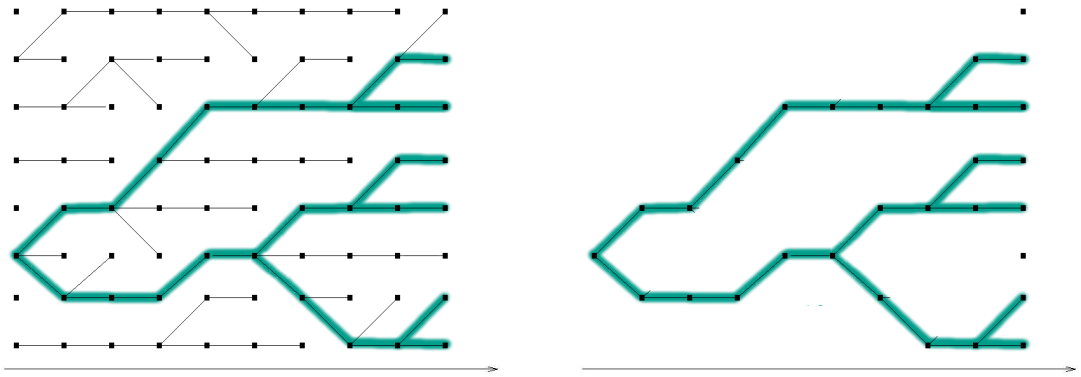

Figura 2-6: **Realización del modelo de Wright-Fisher**. En la figura de izquierda se encuentran las generaciones de la muestra avanzando en el tiempo (de izquierda a derecha). Las líneas conectan a cada individuo con sus descendientes y ancestros. En la figura de la derecha está la misma realización del modelo, pero se han eliminado a todos los individuos del pasado que no tienen descendencia en la muestra.

están relacionados r generaciones antes, si tienen un antepasado común r generaciones antes. Denotaremos a esta relación de equivalencia mediante $i \overset{r}{\sim} j$.

Ahora podemos calcular la probabilidad de que individuos de la muestra sean o no sean hermanos. Por ejemplo, la probabilidad de que dos individuos sean hermanos es $\mathbb{P}(1 \overset{1}{\sim} 2) = \frac{1}{N}$. En general, la probabilidad de que m de los N individuos fijos de la muestra sean hermanos es $\left(\frac{1}{N}\right)^{m-1}$.

Por otro lado la probabilidad de que los individuos 1 y 2 sean hermanos entre sí y que los individuos 3 y 4 también sean hermanos entre sí, pero que 1 y 3 no lo sean es $\mathbb{P}(1 \overset{1}{\sim} 2 \overset{1}{\nsim} 3 \overset{1}{\sim} 4) = \frac{N-1}{N^3}$.

El orden de la probabilidad de tener solo dos hermanos es mayor que el de tener tres o más. Por lo que, para N muy grande, si reescalamos el tiempo de tal manera que se vean las colisiones de dos linajes, no se verán colisiones de tres o más y tampoco se verán colisiones simultaneas. Esto quiere decir que en el límite el árbol genealógico se vería como un coalescente binario.

Ya que implícitamente hemos estado trabajando con intercambiabilidad en los individuos, veamos el tiempo de colisión de los linajes de los individuos 1 y 2, al cual llamaremos T_N. Claramente, la variable T_N es distribuida geométricamente con parámetro $1/N$. Por lo que tenemos $\mathbb{P}(T_N > r) = (1 - \frac{1}{N})^r$. Ahora volvemos a reescalar el tiempo de tal manera que $r = \lfloor Nt \rfloor$, para $t \geq 0$. Por lo tanto en el límite tenemos a la distribución de la variable aleatoria exponencial de parámetro 1: $\mathbb{P}(T_N > \lfloor Nt \rfloor) \sim e^{-t}$. Por lo que las genealogías de una muestra de tamaño n en un modelo Wright-Fisher se comportan como un $n-$coalescente de Kingman para poblaciones grandes (ver referencia [16]).

Para más información y detalles sobre el modelo Wright-Fisher y el coalescente de Kingman se pueden consultar las notas de Durrett [10].

2.3.2. El modelo de Cannings y el Ξ-Coalescente

Entre 1974 y 1975 Cannings introduce otro modelo para ciertos tipos de poblaciones. Möhle y Sagitov probaron que de la misma manera que las genealogías del modelo de Wright-Fisher se comportan asintóticamente como un coalescente de Kingman, las genealogías del modelo de Cannings se comportan asintóticamente como un Ξ-coalescente [18].

Para describir el modelo pensemos en una población que se mantiene constante en N individuos. Consideramos entonces al vector aleatorio de descendencia $(v_1(t), \ldots, v_N(t))$ que nos da el número de hijos que tiene cada uno de los N individuos en la generación t. Para cada $t \geq 0$, los vectores $(v_1(t), \ldots, v_N(t))$ son independientes e idénticamente distribuidos. Supondremos además que el vector de descendencia es intercambiable (no se hace distinción entre los individuos). Ya que la población se mantiene constante, entonces $\sum_{i=1}^{N} v_i(t) = N$, para cada t.

Observación 7. *El modelo Wright-Fisher es un caso particular del modelo de Cannanings en el que el vector de descendencia se distribuye como una multinomial de N ensayos y pesos iguales $1/N$.*

Por intercambiabilid, para este modelo que la probabilidad de que dos individuos tomados al azar de la misma generación sean hermanos es $c_N^{(2)} = \frac{\mathbb{E}[v_1(v_1-1)]}{N-1}$. Además la probabilidad de que tres individuos sean hermanos es $c_N^{(3)} = \frac{\mathbb{E}[v_1(v_1-1)(v_1-2)]}{(N-1)(N-2)}$. De manera general tenemos

$$c_N^{(n)} = \mathbb{E}[v_1(v_1-1)\ldots(v_1-n+1)]\frac{n!}{(N-1)!}.$$

Observamos que el orden de $c_N^{(n)}$ está relacionado con el orden del n-ésimo momento de v_1.

En el modelo Wright-Fisher el orden de $c_N^{(2)}$ es mayor que el de cualquier otro valor de $c_N^{(n)}$, por lo que solo se permiten colisiones de 2 partículas. Sin embargo, en el modelo general de Cannanings podría suceder que $c_N^{(2)}$ y $c_N^{(3)}$ sean del mismo orden generando así las colecciones múltiples, como veremos en la siguiente sección.

2.3.3. El proceso Galton-Watson Supercrítico y los Coalescentes

Un ejemplo particular del modelo de Cannings se obtiene mediante procesos de Galton-Watson supercríticos.

Un **proceso de Galton-Watson** es una cadena de Markov que modela la reproducción de individuos. Para describir a este proceso supongamos que a cada generación t hay una cantidad $N(t)$ de individuos, cada uno de los cuales tiene $v_1(t), \ldots, v_{N(t)}(t)$ hijos respectivamente, donde las $v_i(t)$ son independientes e idénticamente distribuidas. En dicha cadena además se pide que el comportamiento de cada generación sea independiente. Además se dice que un proceso de Galton-Watson es supercrítico si $\mathbb{E}[v_1(t)] > 1$.

Consideremos a un proceso de Galton-Watson supercrítico que comienza con N individuos y tal que $\mathbb{P}(v_1(t) = 0) = 0$. Luego en cada generación se seleccionarán al azar solo N individuos y solo a estos N individuos se les permite reproducirse, manteniendo así una población constante. Este modelo es equivalente al modelo de Cannanings, con la ventaja que podemos trabajar directamente sobre la distribución de $v_1(t)$. Este modelo podría aparecer por ejemplo en el maíz, donde cada planta tiene muchas semillas, pero cada año se escoge al azar una pequeña cantidad de semillas para sembrarlas el año siguiente.

El modelo presentado por Schweinsberg, en [21], es un caso particular al modelo de Cannanings, pero con la hipótesis adicional de la existencia de ciertas $a > 0$ y $C > 0$ tales que $\mathbb{P}(v_1(t) \geq k) \sim Ck^{-a}$, para cada k. De acuerdo al valor de a, las genealogías del modelo de Schweinsberg pueden converger a distintos coalescentes, cuando el tamaño de N tiende a infinito:

- Cuando $a \geq 2$ el modelo converge a un coalescente de de Kingman.
- Cuando $1 \leq a < 2$ el modelo converge a un Beta coalescente de parámetro a, introducido en el Ejemplo 9 de la subsección 2.2.2.
- Cuando $0 < a < 1$ el proceso converge a cierto Ξ-coalescente.

Capítulo 3

El Espectro de Frecuencia de Sitio

Un concepto importante en genética de poblaciones es el espectro de Frecuencia de Sitio. En este capítulo se define dicho concepto y se muestra como puede utilizarse para hacer inferencia sobre la genealogía de una muestra de una población de individuos.

3.1. Alineamiento de Secuencias y las Mutaciones

Supongamos que tenemos fragmentos de la secuencia de ADN de varios individuos y nos interesa compararlas. Lo primero que debemos hacer es buscar secciones similares para encontrar que bases nitrogenadas que resultan análogas y en base a esto encontrar las mutaciones. Al proceso de buscar secciones similares se le llama **alineamiento de secuencias**.

Ejemplo 11. *Supongamos que tenemos los siguientes fragmentos de la secuencia de ADN de tres individuos*

- *TATCAATCGATT*
- *CGATCGAGTAGCAT*
- *TTCGATCGATTAG.*

Entonces al alinear las secuencias obtendríamos el siguiente arreglo:

$$\begin{array}{cccccccccccccccc} T & A & T & C & A & A & T & C & G & A & T & T & & & & \\ & & & C & G & A & T & C & G & A & C & T & A & G & C & A & T \\ & T & T & C & G & A & T & C & G & A & T & T & A & G & & \end{array}.$$

En el arreglo anterior el color azul corresponde a los fragmentos de secuencia iguales y el color gris a las diferencias generadas por mutaciones.

3.2. El Espectro de Frecuencia de Sitio

Uno de los objetivos de comparar las secuencias de ADN de los individuos es inferir sobre la frecuencia con que se están produciendo mutaciones. Existen muchos tipos de mutaciones y, de acuerdo al tipo de mutación, estas pueden surgir con mayor o menor frecuencia. Sin embargo, para facilitar el estudio, supondremos que las mutaciones son simplemente cambios en los que una base nitrogenada se cambia por otra. Además supondremos que cualquier posible cambio (entre cualquier posible base por cualquier otra distinta) sucede con la misma frecuencia. Estas hipótesis son poco realistas. Sin embargo, son muy útiles para estructurar un modelo general. Teniendo un modelo general como base, es posible estudiar con más detalle los tipos de mutación de interés.

A lo largo de su cadena de ADN un organismo simple, como una bacteria, puede tener 500 pares de bases nitrogenadas y un organismo más complejo, como los humanos, puede tener cantidades de pares del orden de 10^9, o incluso más. Por lo tanto, de acuerdo a la hipótesis de que las mutaciones suceden con la misma frecuencia en cada base, la probabilidad de que dos mutaciones ocurran sobre la misma base nitrogenada se desprecia. A este modelo se le llama se le llama **modelo de infinitos sitios**.

Definición 13. *Supongamos que se tiene una muestra de secuencias de ADN de n individuos. Entonces se dice que una* ***mutación*** *es* ***de tamaño*** *i, para $1 \leq i \leq n$, si dicha mutación afecta o fue heredada por i de los n individuos muestreados.*

Ejemplo 12. *Consideremos de nuevo a las secuencias de ADN de los tres individuos del Ejemplo 11, que son CAATCGATT, CGATCGACT y CGATCGATT.*

Podemos observar dos mutaciones:

a) Ya sea que el primer individuo mutó una G por una A o que los otros dos individuos tienen un ancestro en común que mutó la A por la G.

b) Ya sea que el segundo individuo mutó una T por una C o que los otros dos individuos tienen un ancestro en común que mutó la C por la T.

Entonces tanto en el inciso a) como en el inciso b) no sabemos si el tamaño de estas mutaciones es 1 o es 2.

En el ejemplo anterior se ilustra la dificultad para distinguir a una mutación que afecta a i individuos de una que afecta a $n - i$. Para resolver este problema se requiere tener acceso a la secuencia de ADN original. Generalmente no se puede tener acceso a la secuencia de ADN original pero se tiene acceso a la secuencia de ADN de un individuo de la misma especie pero ajeno a la muestra. A la secuencia de dicho individuo se le llama **salvaje**. El individuo con secuencia salvaje generalmente es un individuo extraído de una población lejana. La lejanía de la población nos hace suponer que no tiene ancestros comunes con la muestra que se desea estudiar. Además se supone que la tasa de que este individuo haya mutado en las bases que son de nuestro interés es despreciable. En este sentido podemos pensar que el individuo salvaje tiene la secuencia original.

Ejemplo 13. *Supongamos ahora que en el Ejemplo 12 además tenemos acceso a la secuencia salvaje CGATCGACT. En esta secuencia las bases coloreadas representan mutaciones en la muestra del Ejemplo 12 (recordemos que se está suponiendo que la secuencia salvaje no tiene mutaciones). Entonces la mutación asociada al color azul es de tamaño 1, la mutación asociada al color rojo es de tamaño 2.*

El conteo del ejemplo anterior nos representa a un vector que se denomina "Espectro de Frecuencia de Sitio".

Definición 14. *Dadas las secuencias de ADN de n individuos y el acceso a una secuencia salvaje extra, se le llama* ***Espectro de Frecuencia de Sitio (EFS)*** *al vector*

$$\xi^{(\mathbf{n})} := \left(\xi_{\mathbf{1}}^{(\mathbf{n})}, \ldots, \xi_{\mathbf{n-1}}^{(\mathbf{n})}\right)$$

tal que $\xi_i^{(n)}$ es el número de mutaciones observadas de tamaño i.

3.2.1. Comportamiento asintótico

Idealmente debería bastar con conocer la tasa de mutación θ y la medida Λ asociada a un $\Lambda-$coalescente para poder deducir la ley conjunta del EFS de una población regida bajo este modelo. Sin embargo, no hay un camino claro por el cual se pueda obtener dicha ley. La propuesta de Limic y los hermanos Berestycki para este problema fue la deducción del comportamiento asintótico del EFS para una familia de $\Lambda-$coalescentes entre los cuales está el Beta coalescente [5]. Dichos coalescentes son los asociados a medidas Λ con la propiedad llamada fuerte $\alpha-$regular variación en cero.

Definición 15. *Se dice que una medida Λ* ***tiene una fuerte $\alpha-$regular variación en cero****, si existe un $\alpha \in (1,2)$ y un $A > 0$ tal que*

$$\Lambda(dx) = f(x)dx, \text{ donde } f(x) \sim Ax^{1-\alpha} \text{ cuando } x \to 0.$$

Observación 8. *La medida Beta$(2-\alpha, \alpha)$, asociada a un Beta coalescente tiene una fuerte $\alpha-$regular variación en cero, cuando $\alpha \in (1,2)$.*

En términos del EFS, el teorema de Limic y los hermanos Berestycki nos dice lo siguiente.

Teorema 8. *Sea Λ una medida asociada a un $\Lambda-$coalescente que cumple la fuerte $\alpha-$regular variación en cero, para algunos $\alpha \in (1,2)$ y $A > 0$. Entonces el EFS asociado a una muestra de tamaño n y con tasa de mutación $\theta/2$ tiene el siguiente*

comportamiento asintótico.

$$\lim_{n\to\infty} \frac{\xi_j^{(n)}}{n^{2-\alpha}} = \frac{\theta}{2} C_{A,\alpha} \frac{(2-\alpha)\Gamma(i+\alpha-2)}{i!\Gamma(\alpha-1)},$$

casi seguramente, para

$$C_{A,\alpha} = \frac{\alpha(\alpha-1)}{A\Gamma(2-\alpha)(2-\alpha)}.$$

El teorema anterior debería ser de utilidad para muestras grandes, sin embargo se desconoce cómo obtener la tasa de convergencia.

3.3. Los momentos del EFS para el Λ-coalescente

La ventaja de utilizar la estructura de coalescente, en vez de la estructura de árbol, en un árbol genealógico es que el coalescente tiene además la longitud de las ramas.

Definición 16. *Dado un proceso coalescente $\{\Pi_t\}_{t\geq 0}$, se le llama **rama** a un elemento $A \in \Pi_t$, para algún t. Además al tiempo que dura A desde que se forma hasta que colisiona con algún otro elemento de Π_t se le llama **longitud de la rama**.*

Podemos pensar a las ramas como individuos del árbol genealógico. De esta forma, tiene sentido pensar que entre más larga sea una rama mayor probabilidad de mutaciones habrá. Además, a todos los conjuntos que en algún momento colisionaron podemos pensarlos como descendientes de una rama. De esta forma, las mutaciones que sucedan en una rama serán heredadas a todos los descendientes.

Definición 17. *Se le llama **tamaño de una rama** a la cardinalidad de esta.*

Observación 9. *El tamaño de una mutación coincide el tamaño de la rama en la cual sucede.*

Ejemplo 14. *En la figura 3-1 tenemos un coalescente genealógico. A la cabeza se encuentra la cadena salvaje. Las mutaciones coloreadas en azul, verde y morado afectan a un solo individuo, por lo que $\xi_1^{(n)} = 3$. Las mutaciones coloreadas en amarillo y café son las afectan a tres individuos, por lo que $\xi_3^{(n)} = 2$. Por último tenemos una*

mutación roja que afecta a 7 individuos, por lo que $\xi_7^{(n)} = 1$. *Entonces el EFS es* $\xi_i^{(n)} = (3, 0, 2, 0, 0, 0, 1, 0)$.

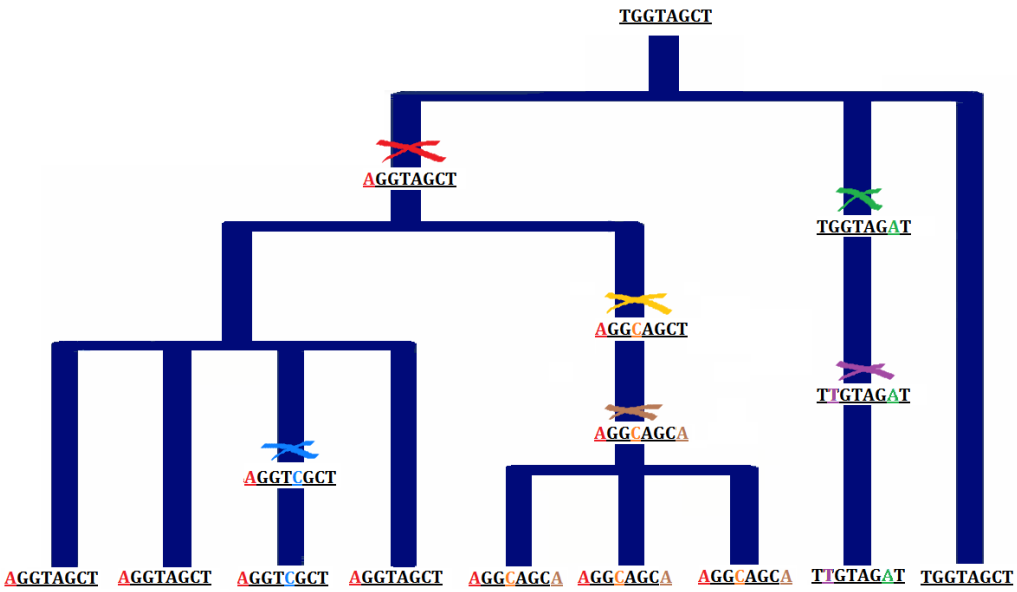

Figura 3-1: El EFS y el coalescente genealógico.

3.3.1. Los momentos del EFS

Se tiene una muestra de cadenas de ADN y se desea inferir el coalescente genealógico. Como el problema en sí es muy complejo, resulta conveniente hacer hipótesis extra sobre dicho coalescente. La primera hipótesis razonable es la intercambiabilidad, ya que generalmente se observan mutaciones neutras, es decir sin ventaja selectiva. Otra hipótesis razonable es la propiedad de Markov. La propiedad de Markov es una hipótesis razonable, ya que no tenemos información sobre el pasado de los individuos de dicha muestra. Por lo anterior, dependiendo de la forma de reproducción de la especie puede ser útil suponer un coalescente de Kingman, un Λ-coalescente o un Ξ-coalescente. En la práctica, por dificultades de inferencia, los Ξ-coalescentes no son tan utilizados. Por lo que en adelante supondremos un Λ-coalescente.

Supongamos por un momento que tenemos acceso al coalescente genealógico, pero queremos hacer inferencia sobre las mutaciones que suceden en las ramas. Lo más natural es suponer que las mutaciones se distribuyen uniformemente sobre toda la

longitud de las ramas. Es decir, condicionando sobre el coalescente genealógico, se supondrá que las mutaciones se distribuyen de acuerdo a un Proceso de Poisson Puntual sobre dichas ramas. Generalmente se supone una tasa de $\theta/2$, donde θ es un parámetro sobre el cual también se desea hacer inferencia. En resumen, para inferir la genealogía de una población de tamaño n, se supone un Λ-coalescente, sobre el cual suceden mutaciones de acuerdo a un Proceso de Poisson Puntual de parámetro $\theta/2$.

Utópicamente, conocer la medida Λ y el parámetro θ debería ser suficiente para deducir una distribución del EFS, que en este caso sería un vector aleatorio. Sin embargo, deducir la distribución conjunta de EFS es muy complicado. Por lo que solo se deduce su vector de esperanzas y su matriz de covarianzas como se estudiará en el resto de este capítulo.

3.3.2. La intuición de Fu

En 1994 Fu [13] publicó fórmulas del vector de medias y matriz de covarianzas para el EFS del caso del coalescente de Kingman. Dichas fórmulas dependen solo de la tasa de mutación $\theta/2$. Para explicar de manera muy general la idea de Fu y más adelante la de Birkner, Blath y Eldon [4], haremos uso de la siguiente cadena de Markov asociada a un Λ-coalescente de (n).

Definición 18. *Sea* $\{\Pi_t\}_{t\geq 0}$, *entonces el* ***proceso de conteo de bloques*** $\{Y_t\}_{t\geq 0}$ *es la cadena de Markov que nos dice la cantidad de bloques o partículas que tiene* Π *al tiempo* t, *es decir*

$$Y_t = \#\Pi_t.$$

Además diremos que Π *está en el* ***estado*** k *cuando* Y *se encuentre en dicho estado.*

Ejemplo 15. *Supongamos que tenemos un 5-coalescente de Kingman. Entonces cualquiera de los vectores (1,1,3), (1,3,1), (3,1,1),(1,2,2), (2,1,2), (2,2,1) podría representar al vector de tamaños de las ramas en el estado 3. Supongamos que el vector que ocurrió es el vector (1,2,2) como en la Figura 3-2. Entonces cualquier mutación que ocurra en la primera rama será de tamaño 1 y cualquier mutación que ocurra en las otras ramas será de tamaño 2.*

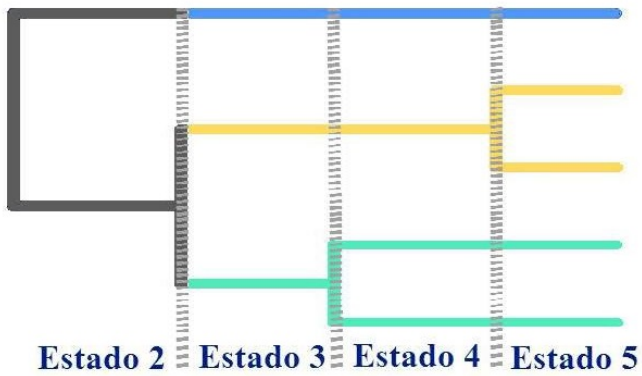

Figura 3-2: El vector de hojas para las ramas del estado 3 es **(1,2,2).**

En el n-coalescente de Kingman, cada posible par de bloques tiene una tasa de 1 de colisionar. Por lo que Y solo puede cambiar si se encuentra en un estado mayor o igual que 2. Y se mantiene en un estado durante un tiempo exponencial de parámetro $\binom{Y_t}{2}$, para luego cambiar a $Y_t - 1$.

La idea de Fu es utilizar el hecho de que en un n-coalescente de Kingman (y en cualquier n-coalescente intercambiable), cada una de las hojas tiene igual probabilidad de ser descendiente de cada una de las ramas del estado k, citando [15]. De esta manera reparte las hojas en las ramas utilizando una distribución multinomial. La repartición de las hojas corresponde a la ley de los tamaños de cada rama. En base a esto es posible obtener el vector esperanza para el EFS en términos de θ. La fórmula obtenida por Fu fue

$$\mathbb{E}(\xi_i^{(n)}) = \frac{1}{i}\theta.$$

En el mismo artículo [13], Fu encuentra también a la matriz de covarianzas del EFS.

3.3.3. La esperanza del EFS

Birkner, Blath y Eldon retoman la idea de Fu para encontrar los parámetros del EFS (vector de esperanza y matriz de covarianzas) para el caso del Λ-coalescente de (n) [4].La expresión obtenida para dichos parámetros queda solo en términos de unas complicadas recursiones. Sin embargo, dichas recursiones pueden ser muy útiles para obtener dichos parámetros con herramientas de cómputo.

En el caso del coalescente de Kingman, Fu aprovecha que la cadena asociada Y_t cambia de un estado k a un estado $k-1$ directamente, lo cual no se cumple para otros Λ-coalescentes. Para generalizar la idea de Fu al caso de los Λ-coalescentes, habrá que condicionar sobre el evento de que la cadena pase por cada estado k. Este condicionamiento aparecerá en el Lema 4, pero antes recordaremos e introduciremos un poco más de notación.

Para contar el tamaño de cada rama del estado k, recordemos que $\Pi_t^1, \Pi_t^2, \ldots,$ denotan los bloques ordenados de un coalescente Π al tiempo t, como en la Definición 5. En particular $1 \in \Pi_t^1$, para cada t. Además utilizaremos la siguiente notación.

Denotaremos como $\mathbb{P}_n$ a la medida de probabilidad asociada a un Λ-coalescente de (n). En este contexto $\mathbb{E}_n$ representará la esperanza o integral de Lebesgue con respecto a $\mathbb{P}_n$. También denotaremos como T_1 al tiempo de la primera coalescencia. Además $g(n,m)$ denotará el tiempo esperado que el coalescente dura en el estado m. Es decir

$$g(n,m) = \mathbb{E}_n \left[\int_0^\infty \mathbb{1}_{(Y_s=m)} ds \right].$$

Ya que T_1 es un tiempo de paro, entonces por propiedad de Markov Fuerte $\{\Pi_{t+T_1}\}_{t\geq 0}$ se comporta como un coalescente con las mismas tasas, solo que no comienza con singuletes. Esto nos permite obtener muchas recursiones que hacen posible encontrar numéricamente muchos valores de interés para el estudio e inferencia del EFS. Por ejemplo, el valor numérico de $g(n,m)$ puede obtenerse mediante una sencilla recursión. Para mencionar esta y otras recursiones, denotaremos mediante $p_{a,b}$ a la probabilidad de que Y pase de cierto estado a a cierto estado b.

Observamos que una recursión que hace posible el cálculo de $g(n,m)$ está dada mediante

$$g(n,k) = \sum_{n'=k}^{n-1} p_{n,n'} g(n',k),$$

con la condición inicial $g(k,k) = \frac{1}{q_k}$, donde $q_k = \sum_{i=2}^{k} \binom{k}{i} \lambda_{k,i}$ es la tasa en la que Y sale del estado k.

Para dar algunas otras recursiones importantes utilizaremos la siguiente notación

de eventos:

$$\begin{aligned} A_k &:= \text{ Existe } t > 0 \text{ tal que } \#\Pi_t = k \\ B_{n'} &:= \#\Pi_{T_1} = n'. \end{aligned}$$

Otra recursión útil es para calcular $\mathbb{P}(A_k)$ está dada mediante

$$\mathbb{P}_n(A_k) = \sum_{n'=k}^{n-1} p_{n,n'} \mathbb{P}_{n'}(A_k),$$

con la condición inicial $\mathbb{P}_k(A_k) = 1$. En particular

$$\mathbb{P}_n(A_k) = \frac{g(n,k)}{g(k,k)}.$$

Lema 4. *El condicionamiento sobre el primer paso sujeto a que la cadena Y pase por el estado k de un Λ-coalescente de (n) es de la forma:*

$$\mathbb{P}_n(B_{n'}|A_k) = p_{n,n'} \frac{g(n',k)}{g(n,k)}.$$

Demostración. Tenemos:

$$\begin{aligned} \mathbb{P}_n(B_{n'}|A_k) &= \frac{\mathbb{P}_n(B_{n'} \cap A_k)}{\mathbb{P}_n(A_k)} \\ &= \frac{\mathbb{P}_n(B_{n'})\mathbb{P}_n(A_k|B_{n'})}{\mathbb{P}_n(A_k)} \\ &= \frac{p_{n,n'}\mathbb{P}_{n'}(A_k)}{\mathbb{P}_n(A_k)} \\ &= p_{n,n'} \frac{g(n',k)}{g(n,k)}. \end{aligned}$$

□

Con intención de generalizar la idea de Fu para un n-coalescente de Kingman, se define el valor $p^{(n)}[k,b]$. Este valor se define como la probabilidad de que, condicionalmente a que la cadena pase por el estado k, una de estas ramas, escogida al azar (por ejemplo la rama que contiene al 1, gracias a la intercambiabilidad), sea de tamaño b (ver Figura 3-3).

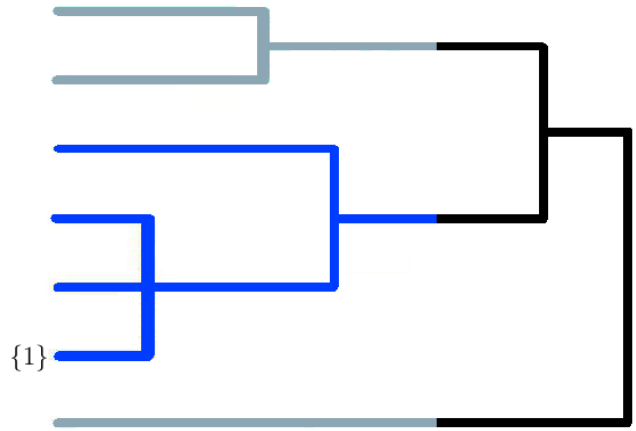

Figura 3-3: En esta figura se muestra la ocurrencia de un evento cuya probabilidad es $p^{(6)}[3,4]$. La rama azul contiene al uno y es de tamaño $b = 4$. Las ramas de azul y gris corresponden a las ramas del estado $k = 3$.

Para obtener la esperanza del EFS, Birkner, Blath y Eldon obtienen la siguiente recursión para $p^{(n)}[k,b]$, cuando $1 < k \leq n$.

$$p^{(n)}[k,b] = \sum_{n'=k}^{n-1} p_{n,n'} \frac{g(n',k)}{g(n,k)} \Bigg(\mathbb{1}_{b>n-n'} \frac{b-(n-n')}{n'} p^{(n')}[k, b-(n-n')] + \mathbb{1}_{b<n'} \frac{n'-b}{n'} p^{(n')}[k,b] \Bigg). \tag{3.1}$$

Para que los valores de $p^{(n)}[k,b]$ se puedan computar, además de la recursión es necesario considerar las siguientes condiciones de frontera

$$p^{(n)}[n,b] = \delta_{1,b} \quad \text{y} \quad p^{(n)}[k,b] = 0 \text{ cuando } b > n-(k-1).$$

Para obtener dicha recursión primero se condiciona sobre el valor que toma Y después del primer salto, luego sobre si dicha primer coalescencia involucra o no a elementos contenidos en la rama fija (la que contiene a 1 en el estado k). Para demostrar con más detalle la recursión necesitaremos un poco más de notación.

Denotamos mediante a C al conjunto formado en la primera coalescencia. Es decir C es el conjunto tal que $C \in \Pi_{T_1}$ y $\#C > 1$. Denotaremos además al evento $A_{k,b}$ como el evento de que en algún momento el coalescente se encuentra en el estado k y la rama fija, la que contiene a 1 en dicho estado. Además denotamos al evento $C_{k,b}$

como el evento en el que se cumple $A_{k,b}$ y además los elementos de C son también elementos del bloque conteniendo al 1, cuando el coalescente está en el estado k (ver Figura 3-4). Formalmente tenemos los eventos se definen como a continuación:

$$\begin{aligned} A_{k,b} := &\quad \text{Existe } t > 0 \text{ tal que } \#\Pi_t = k \text{ y } \#\Pi_t^1 = b, \\ C_{k,b} := &\quad \text{Existe } t > 0 \text{ tal que } \#\Pi_t = k, \#\Pi_t^1 = b \text{ y } C \subset \#\Pi_t^1. \end{aligned}$$

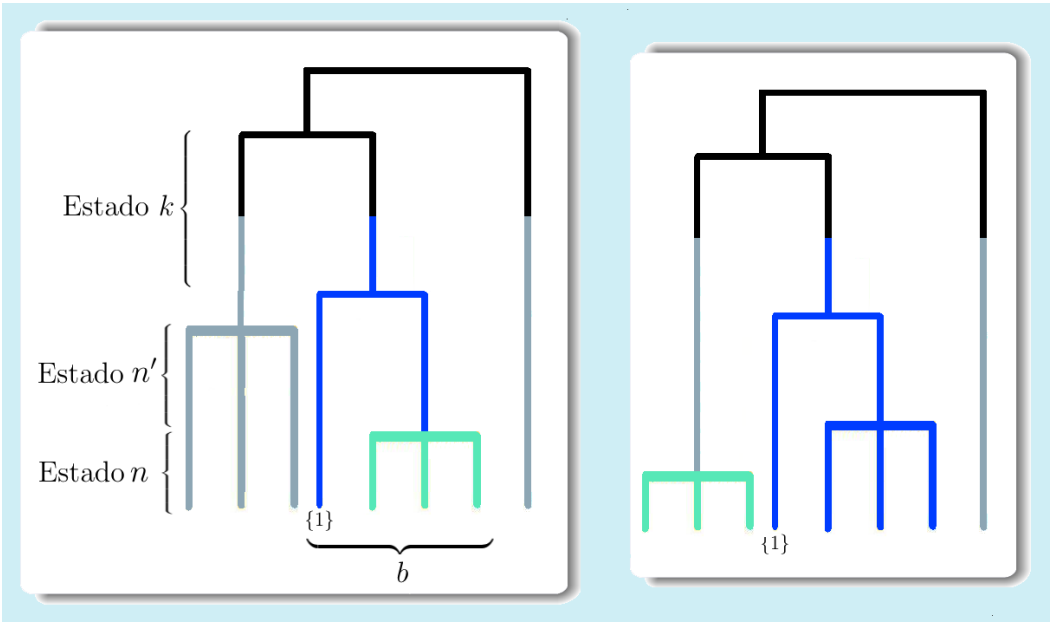

Figura 3-4: En ambas figuras estamos viendo una ocurrencia de un evento $A_{3,4}$. El color azul marino corresponde a la primera rama del estado k y el color verde corresponde a la primera coalescencia. En la figura de la izquierda la rama fija contiene a la primera coalescencia, es decir se cumple $C_{3,4}$. Sin embargo en la figura de la derecha no, es decir se cumple $C_{3,4}^C$.

Entonces, para probar la recursión (3.1) basta con condicionar sobre el valor de Y_{T_1}, luego sobre la posible ocurrencia de $C_{k,b}$ y finalizar aplicando la propiedad de

Markov fuerte sobre el tiempo de paro T_1. Formalmente tenemos :

$$\begin{aligned}
p^{(n)}[k,b] =& \mathbb{P}_n(A_{k,b}|A_k) \\
=& \sum_{n'=k}^{n-1} \mathbb{P}_n(A_{k,b}|A_k \cap B_{n'})\mathbb{P}_n(B_{n'}|A_k) \\
=& \sum_{n'=k}^{n-1} \mathbb{P}_n(B_{n'}|A_k)\Big[\mathbb{P}_n(A_{k,b}|A_k \cap B_{n'} \cap C_{k,b})\mathbb{P}(C_{k,b}|A_k \cap B_{n'}) \\
& + \mathbb{P}_n(A_{k,b}|A_k \cap B_{n'} \cap C^C_{k,b})\mathbb{P}(C^C_{k,b}|A_k \cap B_{n'})\Big] \\
=& \sum_{n'=k}^{n-1} \mathbb{P}_n(B_{n'}|A_k)\Big[\mathbb{1}_{b>n-n'}\frac{b-(n-n')}{n'}p^{(n')}[k,b-(n-n')] \\
& + \mathbb{1}_{b<n'}\frac{n'-b}{n'}p^{(n')}[k,b]\Big] \\
=& \sum_{n'=k}^{n-1} p_{n,n'}\frac{g(n',k)}{g(n,k)}\Big[\mathbb{1}_{b>n-n'}\frac{b-(n-n')}{n'}p^{(n')}[k,b-(n-n')] \\
& + \mathbb{1}_{b<n'}\frac{n'-b}{n'}p^{(n')}[k,b]\Big], \text{ esto por Lema 4.}
\end{aligned}$$

Ahora que tenemos demostrada la recursión (3.1) podemos encontrar una fórmula para el vector esperanza del EFS, como se hará en el siguiente teorema.

Teorema 9. *El vector esperanza del EFS para un Λ-coalescente de (n) está dado mediante*

$$\mathbb{E}\left[\xi_i^{(n)}\right] = \frac{\theta}{2}\sum_{k=2}^{n-i+1} p^{(n)}[k,i]kg(n,k).$$

Demostración. Denotaremos mediante $\psi_n(b)$ a la longitud total de todas las ramas de tamaño b. Además denotaremos mediante $\psi_n(b)[l,k]$ a la longitud de la l-ésima rama del estado k cuando dicha rama es de tamaño b. Si la cadena Y no toca el estado k o si la l-ésima rama no es de tamaño b, entonces $\psi_n(b)[l,k]$ simplemente toma el valor 0. Debido a que las mutaciones ocurren sobre el coalescente según un Proceso Puntual de Poisson de parámetro $\theta/2$, entonces tenemos:

$$
\begin{aligned}
\mathbb{E}\left[\xi_i^{(n)}\right] &= \frac{\theta}{2}\mathbb{E}\left[\psi_n(i)\right] \\
&= \frac{\theta}{2}\sum_{k=2}^{n-i+1}\sum_{l=1}^{k}\mathbb{E}\left[\mathbb{1}_{A_{k,i}}\psi_n(i)[l,k]\right] \\
&= \frac{\theta}{2}\sum_{k=2}^{n-i+1} k\mathbb{E}\left[\mathbb{1}_{A_{k,i}}\psi_n(i)[1,k]\right] \\
&= \frac{\theta}{2}\sum_{k=2}^{n-i+1} k\mathbb{E}_n\left[\mathbb{1}_{A_{k,i}}\int_0^\infty \mathbb{1}_{\{Y_s=k\}}ds\right] \\
&= \frac{\theta}{2}\sum_{k=2}^{n-i+1} k\mathbb{E}_n\left[\mathbb{E}_n\left[\mathbb{1}_{A_{k,i}}\int_0^\infty \mathbb{1}_{\{Y_s=k\}}ds\middle| A_k\right]\right] \\
&= \frac{\theta}{2}\sum_{k=2}^{n-i+1} k\mathbb{E}_n\left[\mathbb{E}_n\left[\mathbb{1}_{A_{k,i}}\middle| A_k\right]\mathbb{E}_n\left[\int_0^\infty \mathbb{1}_{\{Y_s=k\}}ds\middle| A_k\right]\right] \\
&= \frac{\theta}{2}\sum_{k=2}^{n-i+1} k\mathbb{E}_n\left[p^{(n)}[k,i]\mathbb{E}_n\left[\int_0^\infty \mathbb{1}_{\{Y_s=k\}}ds\middle| A_k\right]\right] \\
&= \frac{\theta}{2}\sum_{k=2}^{n-i+1} p^{(n)}[k,i]k\mathbb{E}_n\left[\mathbb{E}_n\left[\int_0^\infty \mathbb{1}_{\{Y_s=k\}}ds\middle| A_k\right]\right] \\
&= \frac{\theta}{2}\sum_{k=2}^{n-i+1} p^{(n)}[k,i]k\mathbb{E}_n\left[\int_0^\infty \mathbb{1}_{\{Y_s=k\}}ds\right] \\
&= \frac{\theta}{2}\sum_{k=2}^{n-i+1} p^{(n)}[k,i]kg(n,k).
\end{aligned}
$$

□

Utilizando el Teorema 9 y la recursión (3.1) es posible computar el vector de esperanzas.

El caso particular en el caso del coalescente de Kingman $g(n,k) = \frac{(k-2)!2}{k!}$ y $p^{(n)}[k,i] = \binom{n-k}{i-1}\frac{1}{k^{i-1}}\left(\frac{k-1}{k}\right)^{n-2k}$. Por lo que se puede calcular de manera explícita al vector de esperanzas, obteniendo

$$\mathbb{E}\left[\xi_i^{(n)}\right] = \frac{\theta}{i}.$$

3.3.4. Las covarianzas del EFS

Para calcular el vector de covarianzas del EFS, Birkner, Blath y Eldon mencionan otras tres recursiones con sus respectivas condiciones de frontera.

Para la siguiente recursión definimos para $2 \leq k \leq n$ a $p_{eq}^{(n)}[k; i, j]$ como la probabilidad de que, condicionado a que en algún momento Y está en el estado k, el primer bloque del estado k es de tamaño i y el segundo es de tamaño j (Ver Figura 3-5).

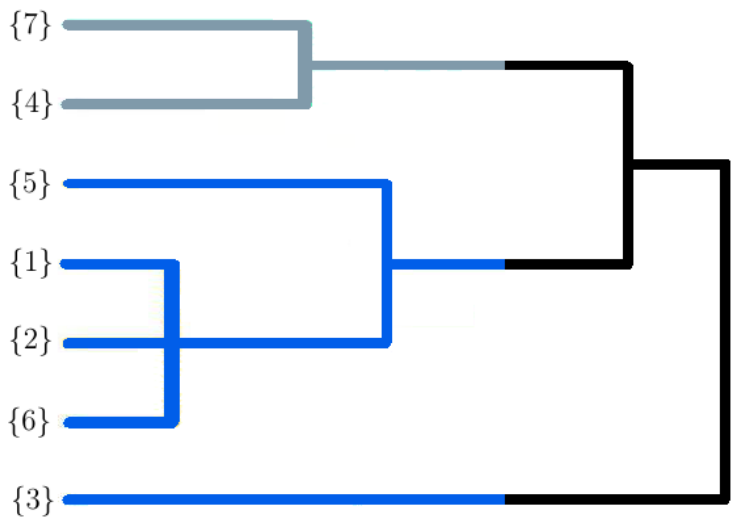

Figura 3-5: En esta figura se puede observar que, cuando Y está en el estado 3, se tienen los bloques $\{1, 2, 5, 6\}, \{3\}, \{4, 7\}$. Por lo que esta figura se representa a un evento con probabilidad $p_{eq}^{(n)}[3; 4, 1]$. Recordemos que los bloques están ordenados en orden creciente del mínimo de cada bloque.

La recursión de Birkner, Blath y Eldon para $p_{eq}^{(n)}[k; i, j]$ es la siguiente

$$\begin{aligned} p_{eq}[k; i, j] = \sum_{m=k}^{n-1} p_{n,m} \frac{g(m,k)}{g(n,k)} \Bigg[& \frac{i-(n-m)}{m} p_{eq}^{(m)}[k; i-(n-m), j] \mathbb{1}_{(i>n-m)} \\ & + \frac{j-(n-m)}{m} p_{eq}^{(m)}[k; i, j-(n-m)] \mathbb{1}_{(j>n-m)} \\ & + \frac{m-i-j}{m} p_{eq}^{(m)}[k; i, j] \mathbb{1}_{(i+j>m)} \Bigg], \end{aligned} \tag{3.2}$$

con la condición inicial $p_{eq}^{(n)}[n; i, j] = \mathbb{1}_{(i=j=1)}$.

La prueba de está recursión es análoga a la prueba de la recursión (3.1). Se condiciona sobre el valor que tome Y, después del primer salto. Luego se condiciona sobre

si dicha coalescencia es descendiente de la primera rama del estado k, si la primera coalescencia es descendiente de la segunda rama o si no no es descendiente de ninguna de estas dos.

Para la siguiente recursión tomamos ahora $2 \leq k < l \leq n$. Entonces $p_{un}^{(n)}[k, i; l, j]$ denota a la probabilidad, condicionada a que un Λ-coalescente de (n) pase por el estado k en algún momento y por el estado l en algún otro momento, además la primera de las k ramas es de orden i y una rama fija, elegida al azar de las l ramas sea de tamaño j y que además esta rama elegida al azar no sea descendiente de la primera rama de k (Ver Figura 3-6).

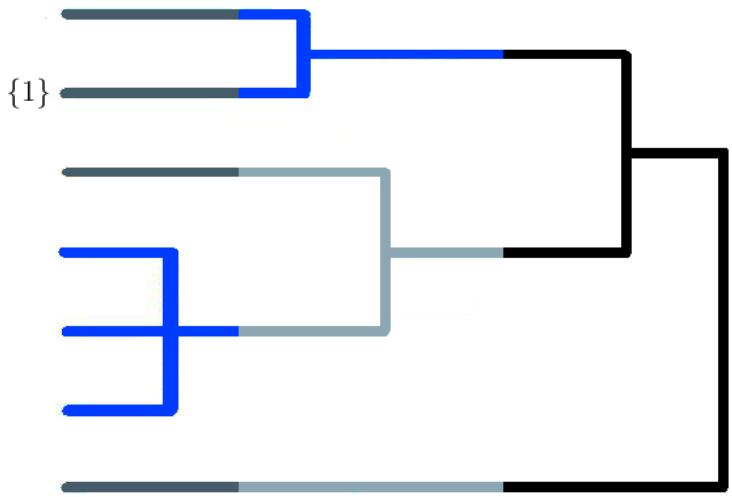

Figura 3-6: En esta figura se puede observar un evento de probabilidad $p_{un}^{(n)}[3, 2; 5, 3]$. El color azul representa a las ramas escogidas.

Entonces tenemos la siguiente recursión para $2 \leq k < l \leq n$

$$\begin{aligned} p_{un}^{(n)}[k, i; l, j] = & \sum_{m=l}^{n-1} p_{n,m} \frac{g(m,l)}{g(n,l)} \Bigg[\frac{i-(n-m)}{m} p_{un}^{(m)}[k, i-(n-m); l, j] \mathbb{1}_{(i>n-m)} \\ & + \frac{j-(n-m)}{m} p_{un}^{(m)}[k, i; j-(n-m)] \mathbb{1}_{(j>n-m)} \\ & + \frac{m-i-j}{m} p_{un}^{(m)}[k, i; l, j] \mathbb{1}_{(m>i+j)}, \end{aligned}$$

con las condiciones de frontera $p_{un}^{(n)}[k, i; n, j] = \mathbb{1}_{(j=1)} p^{(n)}[k, i] \frac{n-i}{n}$. La prueba de esta recursión es también análoga a la de la recursión (3.1).

Para la última recursión tomamos de nuevo $2 \leq k < l \leq n$. Entonces $p_{ne}^{(n)}[k,i;l,j]$ denota a la probabilidad, condicionada a que un Λ-coalscente de (n) pase por el estado k en algún momento y por el estado l en algún otro momento, además la primera de las k ramas es de orden i y una rama fija, elegida al azar de las l ramas sea de tamaño j (Ver Figura 3-6) y que además esta rama elegida al azar resulte ser descendiente de la primera rama de k (Ver Figura 3-7).

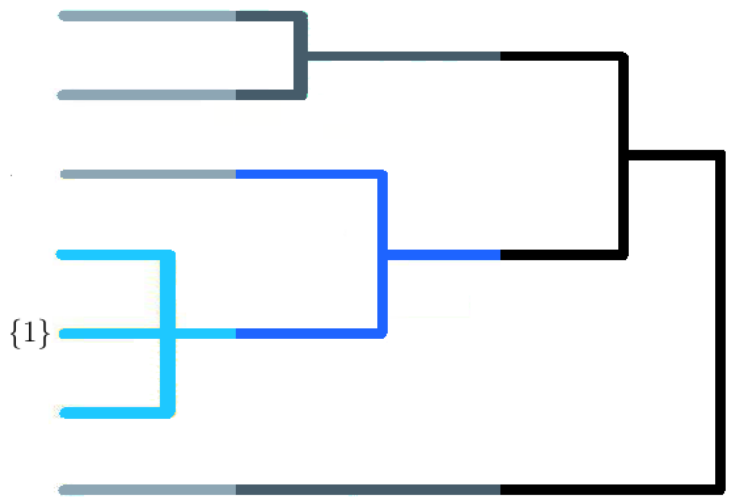

Figura 3-7: En esta figura se puede observar un evento de probabilidad $p_{ne}^{(n)}[3,4;5,3]$.

De nuevo condicionando sobre el primer valor de Y y luego sobre el lugar en que ocurrió la primera coalescencia tenemos la siguiente recursión para $2 \leq k < l \leq n$:

$$\begin{aligned} p_{ne}^{(ne)}[k,i;l,j] = \sum_{m=l}^{n-1} p_{n,m} \frac{g(m,l)}{g(n,l)} \Bigg[& \frac{i-j-(n-m)}{m} p_{ne}^{(m)}[k,i-(n-m);l,j] \mathbb{1}_{(i-j>n-m)} \\ & + \frac{j-(n-m)}{m} p_{ne}^{(m)}[k,i-(n-m);l,j-(n-m)] \mathbb{1}_{(j>n-m)} \\ & + \frac{m-i}{m} p_{ne}^{(m)}[k,i;l,j] \mathbb{1}_{(m>i)}, \end{aligned}$$

con las condiciones de frontera $p_{ne}^{(n)}(k,i;n,j) = \mathbb{1}_{(j=1)} p^{(n)}[k,i] \frac{i}{n}$, cuando $2 \leq k < n$ y $1 \leq i < n$.

Observación 10. *Los subíndices eq, un, y ne provienen respectivamente del inglés " equal," "unnested" y "nested".*

El siguiente teorema, obtenido por Birkner, Blath y Eldon, nos da una forma de computar la matriz de covarianzas del EFS.

Teorema 10. *Dado* $\{\Pi_t\}_{t\geq 0}$, *un* Λ*-coalescente de* (n), *cuando* i, j *sean tales que* $1 \leq, i, j < n$ *y* $2 \leq i + j \leq n$, *podremos computar a la covarianza entre la* i*-ésima y la* j*-ésima entrada del EFS mediante*

$$\begin{aligned}\mathbb{E}\left[\xi_i^{(n)}\xi_j^{(n)}\right] =& \frac{\theta}{4}\sum_{k=2}^{n} k(k-1)p_{eq}^{(n)}[k;i,j]\frac{g(n,k)}{g(k,k)}\frac{2}{(-q_k)^2} \\ &+ \mathbb{1}_{(i=j)}\sum_{k=2}^{n} kp^{(n)}[k,i]\frac{g(n,k)}{g(k,k)}\left(\frac{\theta}{2}\frac{1}{q_k}+\frac{\theta^2}{4}\frac{2}{q_k^2}\right) \\ &+ \sum_{k=3}^{n}\sum_{l=2}^{k-1} kk'\frac{p_{un}^{(n)}[k,i;l,j]+p_{ne}^{(n)}[k,i;l,j]+p_{un}^{(n)}[k,j;l,i]+p_{ne}^{(n)}[k,j;l,i]}{q_k q_l} \\ &\cdot \frac{g(n,l)}{g(l,l)}\frac{g(l,k)}{g(k,k)},\end{aligned}$$

donde $q_k = \sum_{i=2}^{k}\binom{k}{i}\lambda_{k,i}$ *es la tasa con la cual la cadena* Y *sale del estado* k.

Demostración. Para la prueba de este teorema se introducirá un poco más de notación. Para $2 \leq k \leq n, l \in [k]$, se define $M_{k,l}^{(n)}$ como el número de mutaciones que ocurren en el l-ésismo eje, cuando Y está en el estado k. Además denotaremos a los tamaños de las ramas $L_{k,l}^{(n)}$, mediante:

$$L_{k,l}^{(n)} = \begin{cases} \#\Pi_t^l, & \text{si } t \text{ es tal que } \#\Pi_t = k, \\ 0, & \text{sino existe } t \text{ tal que } \#\Pi_t = k. \end{cases}$$

Entonces tenemos

$$
\begin{aligned}
\mathbb{E}\left[\xi_i^{(n)}\xi_j^{(n)}\right] &= \sum_{k=2}^{n}\sum_{l=1}^{k}\sum_{k'=2}^{n}\sum_{l'=1}^{k'}\mathbb{P}\left\{L_{k,l}^{(n)}=i, L_{k',l'}^{(n)}=j\right\}\mathbb{E}\left[M_{k,l}^{(n)}M_{k',l'}^{(n)}\right] \\
&= \sum_{k=2}^{n} k(k-1)\mathbb{P}\left\{L_{k,1}^{(n)}=i, L_{k,2}^{(n)}=j\right\}\mathbb{E}\left[M_{k,1}^{(n)}M_{k,2}^{(n)}\right] \\
&\quad + \mathbb{1}_{(i=j)}\sum_{k=2}^{n} k\mathbb{P}\left\{L_{k,1}^{(n)}=i\right\}\mathbb{E}\left[\left(M_{k,1}^{(n)}\right)^2\right] \\
&\quad + 2\sum_{k=3}^{n}\sum_{k'=2}^{k-1} kk'\mathbb{P}\left\{L_{k,1}^{(n)}=i, L_{k',1}^{(n)}=j\right\}\mathbb{E}\left[M_{k,1}^{(n)}M_{k',1}^{(n)}\right].
\end{aligned}
$$

Para demostrar el teorema solo tenemos que calcular las distribuciones conjuntas de los tamaños de las ramas y las esperanzas de los productos de las mutaciones.

Por intercambiabilidad, la ley del tamaño $L_{k,l}^{(n)}$ no depende de l. Por lo tanto tenemos

$$
\begin{aligned}
\mathbb{P}\left\{L_{k,l}^{(n)}=i\right\} &= \mathbb{P}_n(A_{k,i}) \\
&= \mathbb{P}_n(A_{k,i}|A_k)\mathbb{P}_n(A_k) \\
&= p^{(n)}[k,i]\frac{g(n,k)}{g(k,k)}.
\end{aligned}
$$

De nuevo por intercambiabilidad, para los tamaños conjuntos tenemos

$$
\begin{aligned}
\mathbb{P}\left\{L_{k,l}^{(n)}=i, L_{k',l'}^{(n)}=j\right\} &= \mathbb{P}\left\{L_{k,1}^{(n)}=i, L_{k',1}^{(n)}=j\right\} \\
&= \left(p_{un}^{(n)}[k,i;k',j] + p_{ne}^{(n)}[k,i;k',j]\right)\frac{g(n,k')g(k',k)}{g(k',k')g(k,k)}.
\end{aligned}
$$

Para las esperanzas de las mutaciones se condiciona sobre el tiempo en que la cadena dura en un estado k. Por lo que será necesario denotar mediante q_k a la tasa con la que la cadena Y se mantiene en cada posible estado k.

Para calcular $\mathbb{E}\left[M_{k,1}^{(n)}M_{k,2}^{(n)}\right]$, recordamos que condicionado a la longitud de las ramas, las mutaciones se rigen bajo un Proceso de Poisson Puntual de parámetro $\frac{\theta}{2}$.

Por lo tanto, tenemos:

$$\begin{aligned}
\mathbb{E}\left[M_{k,1}^{(n)}M_{k,2}^{(n)}\right] &= \int_0^\infty \mathbb{E}\left[M_{k,1}^{(n)}M_{k,2}^{(n)}\middle|\left\{\int_0^\infty \mathbb{1}_{Y_s=k}ds = t\right\}\right] q_k e^{-q_k t}dt \\
&= \int_0^\infty \left(\mathbb{E}\left[M_{k,1}^{(n)}\middle|\left\{\int_0^\infty \mathbb{1}_{Y_s=k}ds = t\right\}\right]\right)^2 q_k e^{-q_k t}dt \\
&= \int_0^\infty \left(\frac{\theta}{2}t\right)^2 q_k e^{-q_k t}dt \\
&= \frac{\theta^2}{4}\frac{2}{q_k^2}.
\end{aligned}$$

La tercera igualdad se debe a que, condicionado al evento $\left\{\int_0^\infty \mathbb{1}_{Y_s=k}ds = t\right\}$, la variable aleatoria $M_{k,1}^{(n)}$ tiene ley Poisson($\frac{\theta}{2}t$).

Finalmente, para $\mathbb{E}\left[\left(M_{k,1}^{(n)}\right)^2\right]$ tenemos:

$$\begin{aligned}
\mathbb{E}\left[\left(M_{k,2}^{(n)}\right)^2\right] &= \int_0^\infty \mathbb{E}\left[\left(M_{k,l}^{(n)}\right)^2\middle|\left\{\int_0^\infty \mathbb{1}_{Y_s=k}ds = t\right\}\right] q_k e^{-q_k^t}dt \\
&= \int_0^\infty \left[\left(\frac{\theta}{2}t\right)^2 + \frac{\theta}{2}t\right] q_k e^{-q_k t}dt \\
&= \frac{\theta}{2}\frac{1}{q_k} + \frac{\theta^2}{4}\frac{2}{q_k^2}.
\end{aligned}$$

Juntando todos los términos calculados se obtiene el teorema. □

En el caso particular del coalescente de Kingman tenemos que

$$Var\left[\xi_i^{(n)}\right] = \frac{\theta}{i} + \sigma_{ii}\theta^2$$

y para $i < j$

$$Cov\left[\xi_i^{(n)}, \xi_j^{(n)}\right] = \sigma_{ij}\theta^2,$$

donde

$$\sigma_{ii} = \begin{cases} \beta_n(i+1), & \text{si } i < n/2, \\ 2\frac{a_n - a_i}{n-i} - i^{-2}, & \text{si } i = n/2, \\ \beta_n(i) - i^{-2}, & \text{si } i > n/2, \end{cases}$$

$$\sigma_{ij} = \begin{cases} \frac{\beta_n(j+1)-\beta_n(j)}{2}, & \text{si } i+j<n, \\ \frac{a_n-a_i}{n-i} + \frac{a_n-a_j}{n-j} - \frac{\beta_n(i+1)+\beta_n(j)}{2} - \frac{1}{ij}, & \text{si } i+j=n, \\ \frac{\beta_n(i)-\beta_n(i+1)}{2} - \frac{1}{ij}, & \text{si } i+j>n, \end{cases}$$

$$a_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n-1}$$

y

$$\beta_n(i) = \frac{2n}{(n-i+1)(n-i)}(a_{n+1} - a_i) - \frac{2}{n-i}.$$

3.4. Inferencia sobre los parámetros del modelo

Hasta ahora hemos hablado de que para la genealogía de ciertas poblaciones resulta apropiado ajustar modelos basados en $\Lambda-$coalescentes. En esta sección se hablará de como ajustar parámetros a dichos modelos.

Una posible idea sería basar nuestros modelos en un Beta coalescente y comparar el comportamiento observado con el comportamiento asintótico esperado para una rejilla de parámetros, de acuerdo al Teorema 8. El problema es que no se conoce la tasa de convergencia y por lo tanto es difícil saber que tan apropiado es utilizar este teorema para modelar datos reales. Birkner, Blath y Eldon [4] realizaron algunas pruebas para comparar los resultados asintóticos con simulaciones. Ellos observaron que incluso para valores tan grandes como $n = 10^4$ no se obtiene un buen ajuste para $\alpha < 1{,}5$ La correspondencia mejora mucho cuando α aumenta llegando a buenas correspondencias para $\alpha = 1{,}5$ y $n = 10^3$. En la Figura 3-8 de Birkner Balth y Eldon comparan los resultadas obtenidos con los del Teorema 8 para distintos valores de α y muestras de tamaño $n = 100$ y $n = 250$. Podemos observar que en ambas muestras el ajuste para $\alpha = 1{,}05$ es muy malo, pero para los demás valores de α el ajuste es bueno.

Al observar la Figura 3-8 uno se puede preguntar cómo ajustar modelos a cuando se tienen muestras de tamaño menor a 100 y valores α cercanos a uno. La propuesta

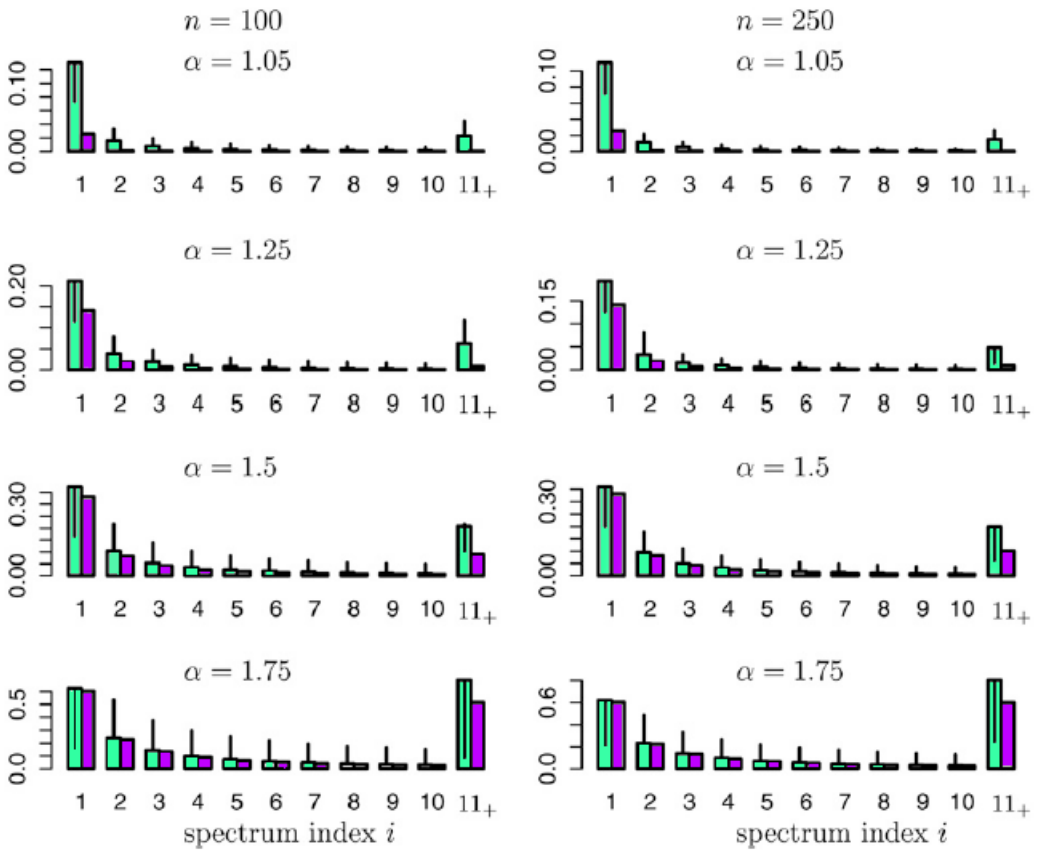

Figura 3-8: Comparación de los resultados simulados para Beta coalescentes de distintos parámetros y distintos tamaños de muestra, con comportamiento asintótico esperado de acuerdo al Teorema 8. Las barras rosas representan los resultados obtenidos mediante dicho teorema, mientras que las barras verdes representan los resultados obtenidos mediante las simulaciones. Los bigotitos de las barras verdes representan la desviación estándar. En todos los casos se consideró la tasa de mutación es $\theta = 1$. La columna de la izquierda representa las muestras de tamaño 100 y la de la derecha las muestras de tamaño 250. De arriba hacia abajo los renglones representan respectivamente los valores de $\alpha = 1{,}05$, $\alpha = 1{,}25$, $\alpha = 1{,}5$ y $\alpha = 1{,}75$.

de Birkner, Blath y Eldon [4] es utilizar las varianzas y convarianzas esperadas y compararlas con los datos obtenidos en un método que llaman la "distancia ℓ^2".

3.4.1. La distancia ℓ^2

En esta subsección se presentan las ideas de Birkner, Blath y Eldon [4] para estimar los parámetros de un modelo para el linaje de una población basado en $\Lambda-$coalescentes. Sus ideas utilizan los momentos presentados en la sección 3.3. En dicha sección se muestra como obtener numéricamente algunos momentos del EFS. Sin embargo, la gran cantidad de recursiones necesarias hace que sea computacionalmente muy complicado obtener dichos valores para muestras de tamaños $n \geq 100$.

La manera más sencilla de encontrar un parámetro α para ajustar un Beta coalescente a la genealogía de una muestra de cierta población podría ser comparar los valores esperados del EFS con los valores observados y utilizar el parámetro que minimice la suma de cuadrados (Ecuación (3.3)).

Para no estimar por ahora a la tasa de mutación θ podemos reescalar al EFS mediante

$$\zeta_i^{(n)} := \frac{\xi_i^{(n)}}{\sum_{k=1}^{n-1} \xi_k^{(n)}}.$$

Lo mismo hacemos para los valores esperados y obtenemos

$$r_i^{(n)} = \frac{\mathbb{E}\left[\xi_i^{(n)}\right]}{\sum_{k=1}^{n-1} \mathbb{E}\left[\xi_k^{(n)}\right]}.$$

La comparación de estos valores reescalados nos da la suma de cuadrados o **distancia** l^2, mostrada en la siguiente ecuación:

$$l^2 := \sqrt{\sum_{k=1}^{n-1} (\zeta_i - r_i)^2}. \tag{3.3}$$

Una vez estimado el parámetro α es posible estimar la tasa de mutación. Por ejemplo, el Teorema 9 nos podría ayudar a obtener un estimador $\widehat{\theta}_i$ para cada entrada

i del EFS. Otra idea podría volver a utilizar la distancia l^2, pero esta vez sin reescalar. De esta manera, esta vez se obtendría un estimador $\widehat{\theta}$, puesto que α ya habría sido estimado con la distancia l^2 reescalada. Intuitivamente este último estimador $\widehat{\theta}$ parece más apropiado que los estimadores $\widehat{\theta}_i$. Sin embargo, dada la dificultad del tema, en esta tesis no se encuentran propiedades de los estimadores, por lo que no se presenta ningún resultado formal que pueda compararlos. Además, por el mismo motivo, en esta tesis no se presenta ningún intervalo de confianza para ningún parámetro.

Observación 11. *El método de la distancia l^2 puede resultar útil para ajustar parámetros a modelos basados en familias de $\Lambda-$coalescentes, indexadas por otros parámetros, que no necesariamente sean Beta coalescentes.*

Uno podría pensar que sería más apropiado estimar los parámetros mediante algún método de verosimilitud para obtener propiedades importantes del estimador y además un intervalo de confianza. Sin embargo, para tener una función de verosimilitud contienendo , deberíamos tener una manera de expresar la probabilidad de obtener el EFS observado en términos de los parámetros del coalescente. Debido a que es muy complicado hacer explícita la dependencia que tiene dicha probabilidad sobre los parámetros, Birkner, Blath y Eldon sugieren interpretar al EFS como una observación con ley multinomial y en base a esto proponen una **pseudoverosimilitud.** Sin embargo, como ellos mismos dicen, este método está basado en los valores numéricos obtenidos en la sección 3.3. Por lo que la información de Fisher no puede ser obtenida fácilmente y no es clara la validez de dichos intervalos de confianza.

3.4.2. Análisis de datos del ADNm para el bacalao del Atlántico

El bacalao del Atántico se conoce por su alta varianza en la tasa de fecundidad, por lo que se observan genealogías poco profundas. Esto nos sugiere que utilizar modelos basados en el coalescente de Kingman podría no ser apropiado. Es por esto que Birkner, Blath y Eldon [4] analizan los datos de Árnason, Kristinsson, Pálsson,

Petersen y Sigurgíslason ([2] y [3]) sobre el bacalao del Atlántico Norte, utilizando un Beta coalescente.

En la Figura 3-8 Birkner, Blath y Eldon presentan una comparación entre el EFS observado y el EFS esperado utilizando un modelo basado en un coalescente de Kingman y un modelo basado en un Beta coalescente. Se puede observar que el Beta coalescente tiene mucho mejor ajuste.

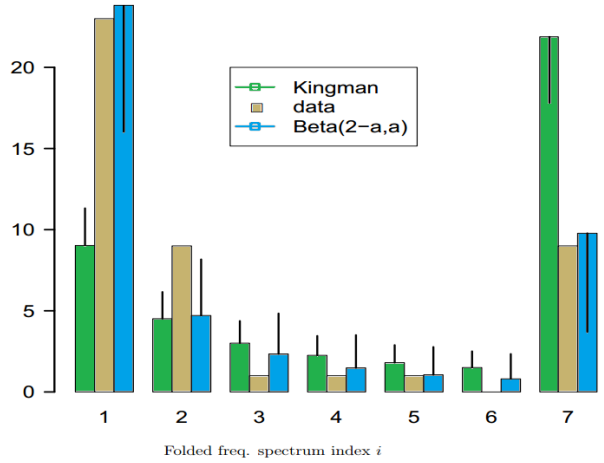

Figura 3-9: Las barras café representa el EFS observado, las barras verdes representan al comportamiento del EFS esperado, ajustando un coalescente de Kingman y las barras azules representan al comportamiento del EFS esperado, ajustando un Beta coalescente.

Observación 12. *En los datos no se contaba con acceso a un individuo salvaje (ver sección 3.2). Por lo que, para una muestra de tamaño n, las mutaciones de tamaño $n-i$ no se pueden diferenciar de tamaño i y son contadas junto con las de tamaño i. Es por eso que en vez de hablar del Espectro de Frecuencia de Sitio, ellos hablan del* ***"Espectro de Frecuencia de Sitio Doblado"****.*

Conclusiones

La ley de un $\Lambda-$coalescente queda determinada por una medida Λ de probabilidad en [0,1]. Dicha medida corresponde con la medida de De Finetti que resulta de la propiedad de intercambiabilidad de dichos coalescentes.

Los Beta coalescentes son una familia importante de $\Lambda-$coalescentes cuya medida Λ se parametriza por un $\alpha \in (0, 2)$.

Podemos representar al comportamiento genealógico de una población con " selección neutra " mediante un Beta coalescente, sobre el cual pueden suceder mutaciones que se heredan a todo el linaje. Una manera de estudiar la genealogía de una muestra de dicha población es observar un importante estadístico, llamado Espectro de Frecuencia de Sitio (EFS).

Obtener relaciones explícitas entre los parámetros de un $\Lambda-$coalescente y el la ley conjunta del EFS sigue siendo una importante área de investigación. Sin embargo, se conoce el comportamiento asintótico del EFS para poblaciones grandes asociadas a Beta coalescentes de parámetro $\alpha \in (1, 2)$. Para poblaciones pequeñas es posible obtener numéricamente la matriz de covarianzas del EFS asociado a cualquier $\Lambda-$coalescente, en términos de sus parámetros.

Comprar características entre el EFS observado en una población y el comportamiento esperado del ESF asociado a una familia de $\Lambda-$coalescentes indexadas por parámetros es la clave para inferir dichos parámetros. Es en este sentido es de utilidad conocer el comportamiento asintótico de dicha familia o sus momentos.

Apéndice A

Elementos de topología

En este apéndice se presentan algunos resultados y definiciones básicos de topología sin demostración. Estos resultados y definiciones han sido tomados del libro de Munkres [19]. El objetivo de este apéndice es presentar un repaso de las herramientas de topología utilizadas en la Sección 2.1.1. El teorema de Tychonoff es el resultado más importante de este apéndice.

Definición 19. *Un* ***espacio topológico*** *es un conjunto X que posse una colección $\mathcal{F}$ de subconjuntos de X, que cumple las siguientes propiedades:*

a) $\emptyset \in \mathcal{F}$,

b) La unión de cualquier subcolección de $\mathcal{F}$ es elemento de $\mathcal{F}$,

c) La intersección de cualquier subcolección finita de $\mathcal{F}$ es elemento de $\mathcal{F}$.

En este caso, a $\mathcal{F}$ se llama ***topología*** *de X y a cada elemento de $\mathcal{F}$ se le llama conjunto* ***abierto****.*

Definición 20. *Se le llama* ***topología discreta*** *a la topología de un espacio en la que todos sus subconjuntos son abiertos.*

Definición 21. *Sea X espacio con una topología $\mathcal{F}$. Sea $A \subset X$, con $A \neq \emptyset$. Entonces, a la colección de subconjuntos $\mathcal{G}$ definida mediante*

$$\mathcal{G} = \{A \cap F : F \in \mathcal{F}\},$$

se le llama ***topología de subconjunto*** *o topología heredada de X.*

Proposición 3. *La topología de subconjunto es efectivamente una topología.*

Definición 22. *Dado un espacio topologíco X se le llama* ***conjunto cerrado*** *a cualquier subconjunto $F \subset X$ tal que su complemento F^C es un conjunto abierto.*

Definición 23. *Dado un espacio X con una topología $\mathcal{F}$, se le llama* ***cubierta abierta*** *cualquier subconjunto de $\mathcal{F}$ tal que su unión forma a todo el espacio X.*

Definición 24. *Dado un espacio topológico X con una cubierta abierta $\mathcal{G}$, se le llama* ***subcubierta*** *abierta de $\mathcal{G}$ (o simplemente subcubierta) a un subconjunto de $\mathcal{G}$ que es a su vez cubierta abierta de X.*

Definición 25. *Dado un espacio topológico X, se dice que un subconjunto $C \subset X$ es compacto si toda cubierta abierta de C tiene una subcubierta finita.*

Proposición 4. *Un subconjunto cerrado de un espacio topológico compacto es compacto.*

Definición 26. *Para cualquier conjunto X, se llama* ***base de una topología*** *en X a cualquier colección $\mathcal{B}$ de subconjuntos de X que cumple las siguientes propiedades:*

a) Para cada $x \in X$ existe al menos un elemento $B \in \mathcal{B}$, tal que $x \in B$,

b) Si $x \in A \cap B$ para $A, B \in \mathcal{B}$, entonces al menos existe un $C \in \mathcal{B}$ tal que $x \in C$ y $C \subset A \cap B$,

Definición 27. *Dado un espacio X y una base $\mathcal{B}$ de X. Se le llama la* ***topología generada*** *por $\mathcal{B}$ al conjunto $\mathcal{F}$ de todas las uniones de elementos de $\mathcal{B}$.*

Proposición 5. *La topología generada por una base es efectivamente una topología.*

Definición 28. *Un* ***espacio métrico*** *es un conjunto X con una operación $d : X \times X \to \mathbb{R}$, llamada* ***distancia*** *que cumple las siguientes condiciones para cualesquiera elementos $x, y, x \in X$*

a) $d(x, x) \geq 0$,

b) $d(x,y) = 0$ *si y solo si* $x = y$,

c) $d(x,y) = d(y,x)$,

d) $d(x,z) \leq d(x,y) + d(y,z)$.

A los elementos de X *se les suele llamar* ***puntos****.*

Definición 29. *Se le llama* ***espacio métrico discreto*** *al espacio métrico con distancia*

$$d(x,y) = \begin{cases} 1, & six \neq y, \\ 0, & six = y. \end{cases}$$

Proposición 6. *La topología generada por la métrica discreta coincide con la topología discreta.*

Definición 30. *Dados un espacio métrico* X*, un punto* $x \in X$ *y un valor* $\varepsilon > 0$*, se le llama bola abierta con centro en* x *y radio* ϵ *al subconjunto de* $B_\epsilon(x) \subset X$ *dado por*

$$B_\epsilon(x) := \{y \in X : d(x,y) < \epsilon\}.$$

Proposición 7. *El conjunto de bolas abiertas de un espacio métrico es base de una topología.*

Definición 31. *A la topología generada por las bolas abiertas de un espacio métrico se le suele conocer como topología de un espacio métrico o topología generada por la métrica.*

Definición 32. *Para cualquier conjunto* X*, se llama* ***subbase de una topología*** *en* X *a cualquier colección* $\mathcal{S}$ *de subconjuntos de* X *tal que su unión es* X*.*

Definición 33. *Dado un espacio* X *y una subbase* $\mathcal{S}$ *de* X*. Se le llama la* ***topología generada*** *por* $\mathcal{S}$ *a la colección de todas las uniones de intersecciones finitas de elementos de* $\mathcal{S}$*.*

Observación 13. *La topología generada por una subbase es efectivamente una topología.*

Notación 5. *Dada una familia indexada de conjuntos* $\{A_\alpha\}_{\alpha\in J}$, $\prod_{\alpha\in J} A_\alpha$ *denota al espacio producto de todas las* J*-tuplas* $(x_\alpha)_{\alpha\in J}$.

Definición 34. *Dado un espacio producto* $\prod_{\alpha\in J} A_\alpha$, *para cada* $\beta \in J$ *definimos a la función* ***proyección*** $\phi_\beta : \prod_{\alpha\in J} \to A_\beta$ *mediante*

$$\phi_\beta((x_\alpha)_{\alpha\in J}) = x_\beta.$$

Definición 35. *Dado un espacio producto* $\prod_{\alpha\in J} A_\alpha$, *consideramos a la colección* $S_\beta = \{\phi_\beta^{-1}(U_\beta) : U_\beta$ *es abierto de* $A_\beta\}$, *donde* ϕ_β *representa a la proyección dada en la Definición 34. Entonces La unión*

$$\mathcal{S} = \bigcup_{\beta\in J} S_\beta.$$

es claramente una subbase y a la topología generada por dicha subbase se le llama ***topología producto*** *o topología usual del espacio producto. En caso de no mencionarse una topología especifica en un producto de espacios topológicos, se estará considerando a dicho espacio dotado de la topología producto.*

Teorema 11 (Teorema de Tychonoff)**.** *El producto de espacios topológicos compactos es compacto.*

Referencias

[1] Aldous, D. & Ibragimov I. & Jacod J. (1985) Exchangeability and Related Topics. *Ecole d'Ete de Probabilites de Saint-Flour XIII, 1983.* Serie: Lecture Notes in Mathematics. New York, New York, U.S.A.:Springer.

[2] Árnason, E. (2004) Mitochondrial Cytochrome b DNA Variation in the High-Fecundity Atlantic Cod: Trans-Atlantic Clines and Shallow Gene Genealogy. *Genetics Society of America,* 166, 1871-1885.

[3] Árnason, E. & Kristinsson, K.& Pálsson,S. & Petersen, P.& Sigurgíslason, H.(2000) Mitochondrial cytochrome sequence variation of Atlantic cod from Iceland Adn Greenland. *Journal of Fish Biology,* 56, 409-430.

[4] Birkner, M. & Blath, J. & Eldon, B. (2013) Statistical Properties of Site-Frequency Spectrum Associated with Λ Coalescents. *Genetics Soc America,* 195, 1037-1053.

[5] Berestycki, J. & Berestycki, N. & Limic, V. (2014) Asymptotic sampling formulae for $\Lambda-$coalescents. *Annales de I'Institut Henri Poincaré-Probabilités et Statisques,* 50, 715-731.

[6] Bolthausen, E. & Sznitman, A.(1998) On Ruelle's Probability Cascades and an Abstract Cavity Method. *Communications in Mathematical Physics,* 197, 247-276.

[7] Carroll, S. & Doebley, J. & Griffiths, A. & Wessler, S. (2011) *Introduction to Genetic Analysis.* New York, New York, U.S.A.: W. H. Freeman. Tenth Edition.

[8] Cummings, M. & Klung, W. & Spencer, C. & Palladino, M. (2008) *Concepts of Genetics.* San Francisco, California, U.S.A.: Pearson Custom Publishing. Ninth Edition.

[9] Desai, M. & Fisher, D. & Walczak, A. (2013) Genetic Diversity and the Structure of Genealogies in Rapidly Adapting Poplations. *Genetics Society of America,* 193, 565-585.

[10] Durrett, R. (2008) *Probability Models for DNA Sequence Evolution.* Serie: Probability and Its Applications. New York, New York, U.S.A.:Springer. Second Edition.

[11] Evans, S. & Pitman, J. (1998) Construction of Markovian Coalescents. *Annales Inst H Poincaré,* 34, 339–383.

[12] Feller, W. (1966) *An Introduction to Probability Theory and Its Applications. Volume II.* New York, New York, U.S.A.: John Wiley & Sons Inc.

[13] Fu, Y. (1995) Statistical Properties of Segregating Sites. *Theoretical Population Biology,* 48, 172-197.

[14] Hjort, N. & Holmes, C. & Müller, P. & Walker, S. (2010) *Bayesian Nonparametrics.* Serie: Cambridge Series in Statistical and Probabilistic Mathematics. New York, New York, U.S.A.:Cambridge University Press.

[15] Johnson, N. & Kotz, S. (1977) *Urn Models and Their Application.* New York, New York, U.S.A.: John Wiley & Sons Inc.

[16] Kingman, J. (1982) On the Genealogy of Large Populations. *Journal of Applied Probability,* 19, 27-43.

[17] Li, Z. (2012) Continuos-State Branching Processes, *Beijing Normal University.*

[18] Möhle, M. & Sagitov, S. (2001) A Classification of Coalescent Processes for Haploid Exchangeable Population Models. *The Annals of Probability,* 29, 1547-1562.

[19] Munkres, J. (2000) *Topology.* Upper Saddle River, New Jersey, U.S.A.: Prentice Hall, Inc. Second Edition.

[20] Pitman, J. (1999) Coalescents with Multiple Collisions. *The Annals of Probability,* 27, 1870-1902.

[21] Schweinsberg, J. (2003) Coalescent Processes Obtained from Supercritical Galton-Watson Processes. *Stochastic Processes and Their Applications,* 106, 107–139.

[22] Schweinsberg, J. (2000) Coalescentes with Simultaneous Multiple Collisions. *Electronic Journal of Probability,* 5, 1-50.

[23] Schweinsberg J.(2015) Rigorous Results for a Population Model with Selection II: Genealogy of the Population, *University of California at San Diego.*