Experimentación con un algoritmo de MCMC multiescala y autoajustable

Tesis para obtener el título de
Maestro en Ciencias de la Computación y Matemáticas Industriales,
CIMAT, que presenta:

Patricia Bautista Otero

Guanajuato, 10 de abril del 2007.

Directrores de tesis: **Dr. J Andrés Christen y Dr. Johan Van Horebeek**.

A mis cálidos compañeros: Víctor, Toño y Jair.
A mi madre, que nunca lo soñó.
A mi!.

Mi sincero agradecimiento y reconocimiento para: los doctores Andrés Christen Gracia y Johan Van Horebeek quienes dirigieron la tesis. Los doctores José Luis Marroquín Zaleta y Joaquín Ortega Sánchez por el trabajo de corrección. Al CIMAT por la valiosa oportunidad que me dió. Al CONACyT y CONCITEG por el apoyo económico brindado durante mis estudios.

Índice general

Índice de figuras

Índice de cuadros

Introducción

Uno de los intereses en la inferencia Bayesiana es el análisis de una distribución (inicial, posterior, predictiva) π, en particular el evaluar esperanzas $\int f(x)\pi(x)dx$ para varias $f(x)$. Cuando la evaluación directa no es posible, se puede recurrir a otras estrategias, como aproximaciones analíticas, integración numérica, y/o simulación de Monte Carlo. Esta última la más recomendada para casos de alta dimensionalidad. El objetivo de la simulación de Monte Carlo es extraer muestras de π para estimar con medias muestrales las medias poblacionales. Los métodos de Monte Carlo basados en cadenas de Markov (o MCMC por sus siglas en inglés) obtienen las muestras utilizando una cadena de Markov cuya distribución límite coincide con π. Esta estrategia es conveniente en muchos casos ya que la simulación directa es raramente posible.

La idea detrás de MCMC es muy sencilla. Se construye una cadena de Markov con distribución límite igual a π y se obtienen muestras de la cadena con las que se aproximan esperanzas bajo π. Si la cadena tiene ciertas propiedades y las esperanzas existen, entonces los promedios muestrales convergen a las esperanzas. La idea se debe a Metropolis [11], y fue extendida por Hastings [10]. Los métodos de MCMC han sido aplicados no sólo en la inferencia Bayesiana, sino también en áreas como la física estadística (donde de hecho tiene sus orígenes la simulación de Monte Carlo), la reconstrucción de imágenes y la optimización.

Para aplicar un método de MCMC en un problema particular se debe construir un muestreador. Hay dos ideas centrales para construirlo, la primera es *condicionar* para reducir la dimensión, la idea central del muestreador de Gibbs [2]. La segunda es la estrategia conocida como *propuesta y rechazo*, el corazón del algoritmo de Metropolis-Hastings [1]. Estas dos ideas no están en competencia, más bien se complementan. Condicionar es una estrategia

de divide y conquistarás, la clave del éxito de los métodos de MCMC en problemas de alta dimensionalidad.

Es importante aclarar que MCMC es una estrategia, no un algoritmo. Para diseñar un algoritmo para un problema específico se deben definir ciertas condiciones, tales como, cómo condicionar y cómo elegir distribuciones para las propuestas. Se pueden resolver estas cuestiones al elegir estrategias estándar. El condicionar se puede basar en distribuciones condicionales univariadas. Las propuestas pueden ser las de una caminata aleatoria. Usar estas alternativas básicas nos lleva a muestreadores que pueden ser aplicados a muchos problemas, sin embargo en muchos de ellos la eficiencia puede ser subóptima. El condicionamiento sin reparametrización puede llevar a correlaciones altas en las muestras y a un mezclado lento, y las propuestas de caminata aleatoria se mueven muy lentamente en espacios de estados de alta dimensión. Para muchos problemas las elecciones más simples funcionarán bien pero para otras no.

Por otro lado para obtener un muestreador de MCMC eficiente frecuentemente éste se debe construir a la medida del problema, lo cual en muchos casos implica un trabajo arduo.

Las formas de construir un algoritmo son ilimitadas, únicamente hay que considerar que se debe garantizar que el proceso estocástico asociado es Markoviano. Un conjunto de algoritmos muy conocido es el de los *muestreadores adaptables* [9], en el que se hace que el muestreador incluya información en cada iteración o cada cierto número de iteraciones de los datos muestreados anteriormente, es decir, el muestreador "aprende" de las simulaciones. El mecanismo de adaptación también es libre y actualmente hay muchas propuestas. Una de las más conocidas es la empleada en los *muestreadores adaptables direccionales*, ADS por sus siglas en inglés [8], en los que el objetivo es generar direcciones de muestreo que se adapten a la distribución objetivo. La dinámica incluye a n puntos que pertenecen al dominio de la distribución objetivo al que se le llama *conjunto actual*. En cada iteración se selecciona aleatoriamente a dos puntos de este conjunto, digamos x_i y x_j, y de alguna forma con ellos se obtiene un vector u, por ejemplo $u = x_i + x_j$, que será donde se ubicará a un nuevo punto, y, que sustituirá en el conjunto actual a x_i o a x_j. El punto nuevo que se elija dependerá de una v.a. definida sobre u. Observe que la distribución estacionaria para los algoritmos ADS es la distribución de n puntos muestreados independientemente de la densidad objetivo π. En un algoritmo conocido como *snooker* [8], la distribución de esta v.a. corresponde con la distribución condicional de la distribución ob-

jetivo en la dirección de u. En este caso la dificultad está en encontrar esta distribución condicional.

En este marco es más sencillo introducir al t-walk. Éste es un algoritmo que si tuviera que clasificarse dentro de alguna categoria estaría dentro de los algoritmos ADS, pero más en general se encuentra dentro de los *muestreadores adaptables de Metropolis* o AMS [9], pues está construido con base en el esquema de Metropolis-Hastings. En el caso del t-walk el conjunto actual se forma con sólo dos puntos x y x'. Las direcciones de las propuestas de cambio se generan de 5 formas diferentes que se van alternando aleatoriamente (pero no con igual probabilidad). Estas formas de generar direcciones se complementan y cumplen con las propiedades necesarias para asegurar la convergencia de la cadena de Markov asociada a la distribución objetivo. A diferencia del snooker, el t-walk sólo requiere poder evaluar la distribución objetivo, incluso sin estar normalizada, lo que justifica su característica de *genérico*. Entre lo conveniente del algoritmo está su invarianza ante cambios en la escala, así como su nivel de insensibilidad a la presencia de correlación alta. En lo práctico la ventaja del t-walk es el no requerir ajustar ningún parámetro para poder operarlo.

Dada la cantidad de muestreadores de MCMC que existen, el usuario para elegir uno debe considerar sus necesidades y disponibilidades. Algunas de las preguntas que se debe hacer son: ¿Qué tan rápida debe ser cada iteración del algoritmo?, ¿Qué diferencia entre la distribución objetivo y la distribución de mis muestras es posible aceptar?, y ¿Cuánto tiempo puedo invertir en el diseño?, de aquí la importancia de ubicar al t-walk dentro de alguna de las clases de algoritmos de MCMC, pues así es más sencillo valorar sus características. En general, mi apreciación es que es difícil comparar algoritmos entre sí pues simplemente están hechos a la medida de diferentes problemas.

Hasta este punto no se han mencionado las herramientas de cómputo existentes. Dentro de estas tal vez la más desarrollada y conocida es el BUGS (**B**ayesian Inference **U**sing **G**ibbs **S**ampling), (www.mrc-bsu.cam.ac.uk/bugs/) un software desarrollado desde hace más de 17 años en la Unidad de Bioestadística del Centro de Investigación Médica de la ciudad de Cambridge. Actualmente existen varias versiones del software, quiza la más conocida sea OpenBUGS un proyecto de la Universidad de Helsinki en Finlandia. El BUGS está diseñado para clases de problemas estándar. La lógica que sigue se basa en una lista de métodos de muestreo. Para poder definir la función objetivo (distribución posterior) se selecciona un modelo y *a priories* de una lista de distribuciones estándar. BUGS analiza este modelo y crea un MCMC

(comunmente kernels Gibbs) para simular de la distribución correspondiente. Dado que BUGS sólo incluye distribuciones estándar le es posible sacar ventaja de la estructura del modelo (por ejemplo funciones de verosimilitud continuas y diferenciables), de la estructura algebraica y de datos específicos del problema (modas, etc.) Esta información puede ser usada para decidir cómo combinar dimensiones o cómo elegir buenas propuestas para el siguiente estado. Como en BUGS en la mayoría de las herramientas computacionales de muestreo la limitante es la clase de problemas a la que están dirigidas, esto es, aquellos que incluyen únicamente distribuciones estándar. El t-walk es especialmente útil como una opción en este sentido, si se está trabajando con distribuciones no estándar.

La distribución de la tesis es como sigue: el primero y segundo capítulos tratan la parte teórica que da fundamento al t-walk: las cadenas de Markov y dentro de los métodos de MCMC, el algoritmo de Metropolis-Hastings, respectivamente. El tercer capítulo es dedicado al t-walk: su diseño, sus propiedades teóricas, los resultados de las simulaciones de distribuciones objetivo diferentes, y su implementación en C++. Después se incluyen las conclusiones, es decir, los resultados principales de la experimentación con el algoritmo. Con la intención de tener un texto autocontenido se agregó el Apendice A, que trata sobre los principios de la Teoría de la Medida a los que se hace referencia en la tesis.

Capítulo 1

Cadenas de Markov

Los métodos de Monte Carlo para simulación basados en cadenas de Markov permiten obtener una muestra de una densidad objetivo f sin simular directamente de ella. Esto puede resultar muy conveniente cuando f es muy complicada. La base de estos algoritmos es la construcción de una cadena de Markov con propiedades específicas que permitan garantizar que una muestra de la cadena pueda ser usada prácticamente como una muestra de f en casos en los que no se requiere la independencia de la muestra. Una interpretación sumamente sencilla de estas propiedades podría ser la siguiente: en principio, es necesario que la cadena visite todas la regiones del soporte de f (*irreducibilidad*), además que las visite un número infinito de veces (*recurrencia* y *aperiodicidad*). Dado que los resultados que permiten asegurar la equivalencia entre estas dos muestras son asintóticos es requerido también que las condiciones iniciales de la cadena, como es el punto de partida, "sean olvidadas" por la cadena (*ergodicidad*). Finalmente, la cadena debe construirse de modo que se asegure que la muestra que se obtiene es aproximadamente distribuida como f y no como una distribución "parecida" a f (*Ley de los Grandes Números para Cadenas de Markov*). En este capítulo se introducen los conceptos necesarios para definir estas propiedades, se dan las definiciones, y además se incluyen resultados con los que se prueba en los Capítulos 3 y 4 la convergencia del algoritmo de Metropolis-Hastings y la convergencia del t-walk, respectivamente.

1.1. Nociones básicas

DEFINICIÓN **1** *Una cadena de Markov es una sucesión de variables aleatorias* $X_0, X_1, \ldots, X_n, \ldots$, *que se denota por* (X_n), *en la que la distribución condicional de* X_n *dados* $x_0, x_1, \ldots, x_{n-1}$ *es igual a la distribución condicional de* X_n *dado únicamente* x_{n-1}, *para todo* n; *esto es,*

$$P(X_n \in A | x_0, x_1, \ldots, x_{n-1}) = P(X_n \in A | x_{n-1}).$$

En un sentido informal, esta propiedad puede interpretarse como que la cadena tiene memoria corta, pues la distribución condicional de X_n dada la historia de la cadena $x_0, x_1, \ldots, x_{n-1}$ sólo depende del estado más reciente x_{n-1}.

Asociada a cada cadena de Markov existe una función K conocida como el kernel de transición: $P(X_n \in A | x_{n-1}) = K(x_{n-1}, A)$.

DEFINICIÓN **2** *Una función* K *definida sobre* $\mathcal{X} \times \mathcal{B}(\mathcal{X})$[1] *es un kernel de transición si satisface:*

1. *para todo* x *en* $\mathcal{X}$, $K(x, .)$ *es una medida de probabilidad, y*

2. *para todo* $A \in \mathcal{B}(\mathcal{X})$, $K(., A)$ *es medible.*

Cuando $\mathcal{X}$ es discreto el kernel de transición es una matriz con elementos:

$$P_{ij} = P(X_n = x_j | X_{n-1} = x_i), \quad x_i, x_j \in \mathcal{X}.$$

La cadena se dice que es *homogénea* en el tiempo, o simplemente homogénea, si la distribución de $(X_{n_1}, X_{n_2}, \ldots, X_{n_k})$ dado x_{n_0} es igual a la distribución de $(X_{n_1 - n_0}, X_{n_2 - n_0}, \ldots, X_{n_k - n_0})$ dado x_0, para toda sucesión $n_0 \leq n_1 \leq n_2 \leq \ldots \leq n_k$ y toda $k \geq 1$. Una interpretación de esto es decir que la cadena es independiente del tiempo. En lo sucesivo se asumirá que las cadenas de las que se hable tienen esta propiedad.

[1] $\mathcal{B}(\mathcal{X})$ es una σ-álgebra de subconjuntos de X

Considerando $K(x, A) = K^1(x, A)$, el kernel de transición de n pasos se define como $(n > 1)$

$$K^n(x, A) = \int_{\mathcal{X}} K^{n-1}(y, A)K^1(x, dy),$$

de donde se derivan las ecuaciones de Chapman-Kolmogorov.

LEMA **1 Ecuaciones de Chapman-Kolmogorov** *Para toda* m, $n \in \mathbb{N}$, $x \in \mathcal{X}$ *y* $A \in \mathcal{B}(\mathcal{X})$,

$$K^{n+m}(x, A) = \int_{\mathcal{X}} K^m(y, A)K^n(x, dy).$$

La idea intuitiva es que para llegar a A en $n + m$ pasos partiendo de x es necesario pasar por algún y en el paso n.

PROPOSICIÓN **1 Propiedad de Markov** *Para toda distribución inicial* μ *y toda muestra* $(X_0, X_1, \ldots, X_k)$,

$$\mathbb{E}_\mu(h(X_{k+1}, X_{k+2}, \ldots)|x_0, x_1, \ldots, x_k) = \mathbb{E}_{x_k}(h(X_1, X_2, \ldots)),$$

con h *cualquier función de medida positiva.*

Esta propiedad se puede entender como que para cada tiempo k, condicionando sobre X_k, la cadena después del tiempo k inicia otra vez partiendo de x_k.

Las siguientes cantidades permiten conocer sobre las trayectorias que podría seguir una cadena de Markov.

DEFINICIÓN **3** *Considerar* $A \in \mathcal{B}(\mathcal{X})$,

1. $\tau_A = inf\{n \geq 1; X_n \in A\}$*; el primer momento en que la cadena* (X_n) *entra al conjunto* A,

2. $\eta_A = \sum_{n=1}^{\infty} 1_A(X_n)$*; el número de visitas que hace la cadena a* A*.*

De gran importancia serán en lo que sigue las cantidades $P_x(\tau_A < \infty)$ y $\mathbb{E}_x(\eta_A)$: la probabilidad de visitar A en un tiempo finito, y el número promedio de visitas de (X_n) a A, respectivamente.

Cuando (X_n) nunca entra a A se asume que $\tau_A = \infty$. τ_A es un ejemplo común de un *tiempo de paro*. Más generalmente, una v.a. $T : \Omega \to \{0, 1, 2, \ldots\} \cup \{\infty\}$ es llamada un tiempo de paro si el evento $\{T = n\}$ depende únicamente de $X_0, X_1, \ldots, X_n$ para $n = 0, 1, 2, \ldots$. Intuitivamente si se observa a la cadena se puede saber el tiempo en el que T ocurre.

PROPOSICIÓN **2 Propiedad fuerte de Markov** *Para toda distribución inicial* μ *y todo tiempo de paro* T *finito, casi seguramente,*

$$\mathbb{E}_\mu(h(X_{T+1}, X_{T+2}, \ldots)|x_1, x_2, \ldots, x_T) = \mathbb{E}_{x_T}(h(X_1, X_2, \ldots)),$$

siempre que las esperanzas existan.

Esto es, la propiedad de Markov se cumple en tiempos de paro.

1.2. Irreducibilidad

La propiedad de irreducibilidad permite asegurar que todas las secciones del espacio $\mathcal{X}$ sean visitadas por la cadena (X_n) sin importar el punto de partida. Esta característica resulta claramente importante en los métodos de MCMC. Considere uno de estos métodos donde se quiere simular de una distribución específica, es importante que todo el soporte de la función esté representado por la muestra generada.

En el caso en que $\mathcal{X}$ es discreto, la irreducibilidad se verifica si ocurre que $P_x(\tau_y < \infty) > 0$ para todo y y x en $\mathcal{X}$. Según [1] frecuentemente ocurre que $P_x(\tau_y < \infty)$ es uniformemente igual a cero, por lo que para que la irreducibilidad esté bien definida se requiere introducir una medida auxiliar ϕ.

DEFINICIÓN **4** *Dada una medida ϕ, la cadena de Markov (X_n) con kernel de transición K, es ϕ-irreducible si para todo $x \in \mathcal{X}$ y todo $A \in \mathcal{B}(\mathcal{X})$ con $\phi(A) > 0$, existe $n \in \mathbb{N}$ tal que $K^n(x, A) > 0$. La cadena se dice fuertemente ϕ-irreducible si $n = 1$ para todo A medible.*

TEOREMA **1** *La cadena (X_n) es ϕ-irreducible si, y sólo si, para todo $x \in \mathcal{X}$ y todo $A \in \mathcal{B}(\mathcal{X})$ con $\phi(A) > 0$, se cumple alguna de las siguientes condiciones:*

1. *existe $n \in \mathbb{N}$ tal que $K^n(x, A) > 0$;*
2. $\mathbb{E}_x(\eta_A) > 0$.

Es importante señalar que la irreducibilidad es una propiedad intrínseca a la cadena de modo que no depende de la medida ϕ.

1.2.1. Átomos y conjuntos pequeños

La noción de átomo fue introducida por Nummelin [13] en 1978. El objetivo fue tener para el caso de espacios de estados generales un equivalente al caso discreto en el que existen puntos con masa positiva. Partiendo de este concepto es posible desarrollar la teoría de cadenas de Markov en espacios de estados generales en completa analogía al caso de espacios de estados contables.

DEFINICIÓN **5** *Una cadena de Markov (X_n) tiene un átomo $\alpha \in \mathcal{B}(\mathcal{X})$ si existe una medida $\nu > 0$ sobre $\mathcal{B}(\mathcal{X})$ tal que*

$$K(x, A) = \nu(A), \ \forall x \in \alpha, \ \forall A \in \mathcal{B}(\mathcal{X}).$$

Si (X_n) es ϕ-irreducible y $\phi(\alpha) > 0$ se llama al átomo α accesible.

Intuitivamente un átomo es un conjunto con probabilidad constante. Un punto siempre es un átomo. Claramente, cuando $\mathcal{X}$ es contable y la cadena es irreducible, cada punto es un átomo accesible. En espacios de estados generales los átomos no son tan frecuentes, sin embargo, cuando (X_n)

es ϕ-irreducible es posible construir artificialmente conjuntos con estructura atómica. Este resultado, que es en palabras de Meyn y Tweedie [12] posiblemente la mayor innovación en el análisis de cadenas de Markov en las últimas décadas, fue descubierto de formas diferentes y casi simultáneamente por Nummelin, y Athreya y Ney. Si el lector está interesado en conocer el procedimiento de construcción de un pseudo-átomo se puede referir a [12] páginas 102 a la 106. La esencia del procedimiento es una partición probabilística del espacio de estados de forma que los átomos para la cadena partida sean objetos naturales. Para esta construcción es necesario considerar conjuntos que satisfacen la llamada *condición de minorización* que se enuncia a continuación.

Para algún $\delta > 0$, algún $C \in \mathcal{B}(\mathcal{X})$ y alguna medida de probabilidad ν con $\nu(C^c) = 0$ y $\nu(C) = 1$

$$K(x, A) \geq \delta 1_C(x)\nu(A), \quad A \in \mathcal{B}(\mathcal{X}),\ x \in \mathcal{X}.$$

Esto asegura que la cadena tiene probabilidad acotada uniformemente por abajo para toda $x \in C$. Esta condición lleva a la siguiente noción.

Definición **6** *Un conjunto C es pequeño si existe $m > 0$ y una medida no trivial ν_m sobre $\mathcal{B}(\mathcal{X})$ tal que*

$$K^m(x, A) \geq \nu_m(A), \quad \forall\ x \in C, \forall A \in \mathcal{B}(\mathcal{X}).$$

En este caso al conjunto C se le llama ν_m-pequeño.

En [12] (p. 109) se prueba que para una cadena (X_n) ϕ-irreducible siempre existe un conjunto pequeño de medida positiva.

El siguiente resultado dá la conexión entre conjuntos pequeños e irreducibilidad.

Teorema **2** *Sea (X_n) una cadena ϕ-irreducible. Para todo conjunto $A \in \mathcal{B}(\mathcal{X})$ tal que $\phi(A) > 0$, existe $m \in \mathbb{N}$ y un conjunto pequeño $C \subset A$, tal que $\nu_m(C) > 0$. Más aún, $\mathcal{X}$ puede ser descompuesto en una partición numerable de conjuntos pequeños.*

Es claro que los conjuntos pequeños son más sencillos de encontrar que los átomos. De hecho Meyn y Tweedie muestran que para cadenas de Markov suficientemente regulares (en el sentido topológico) todo conjunto compacto es un conjunto pequeño.

1.3. Aperiodicidad

El concepto de período de una cadena de Markov, como varios otros, es más sencillo introducirlo cuando el espacio de estados es contable, esto se hace a continuación.
El *período* de un estado $x \in \mathcal{X}$ se define como

$$d(x) = m.c.d.\{n \geq 1 : K^n(x, x) > 0\}.$$

Esta definición no implica que $K^{md(x)}(x, x) > 0$ para toda m. Lo que se puede decir es que $K^n(x, x) = 0$ a menos que $n = md(x)$ para alguna m. En el caso de una cadena (X_n) irreducible el período es el mismo para todos los estados lo cual se muestra fácilmente vía las ecuaciones de Chapman-Kolmogorov.

DEFINICIÓN **7** *Sea (X_n) una cadena de Markov irreducible sobre un espacio de estados $\mathcal{X}$ contable, (X_n) es llamada*

- *aperiódica, si $d(x) = 1$, para toda $x \in \mathcal{X}$;*
- *fuertemente aperiódica, si $K(x, x) > 0$ para alguna $x \in \mathcal{X}$.*

Para un espacio de estados general la analogía se establece con base a la definición de conjunto pequeño.

Considérese C un conjunto ν_M-pequeño, sin pérdida de generalidad se asume que $\nu_M(C) > 0$ (esto se encuentra justificado en [12] p. 108). Entonces $K^M(x, .) \geq \nu_M(.)$, como $\nu_M(C) > 0$ existe una probabilidad positiva de regresar a C al tiempo M.

Sea

$$E_C = \{n \geq 1 : \exists\, \delta_n > 0, \text{ tal que C es pequeño para } \nu_n > \delta_n \nu_M\}$$

el conjunto de tiempos para los que C es un conjunto pequeño con medida de minorización proporcional a ν_M. En particular E_C es el conjunto de tiempos para los que hay una probabilidad positiva de regresar a C.

El período para un conjunto pequeño C se define como $d(C) = m.c.d(E_C)$. Es posible mostrar que debido a que E_C es cerrado bajo la suma C es $\nu_{nd(C)}$-pequeño para toda n lo suficientemente grande.

En el caso de una cadena de Markov (X_n) ϕ-irreducible, el período de la cadena, d, es el más grande de todos los períodos $d(C)$. De aquí es claro que no depende de ningún conjunto pequeño particular.

DEFINICIÓN **8** *Sea (X_n) una cadena de Markov ϕ-irreducible con espacio de estados $\mathcal{X}$, la cadena es llamada*

- *aperiódica, cuando $d = 1$;*
- *fuertemente aperiódica, en caso de existir un conjunto ν_1-pequeño A, tal que $\nu_1(A) > 0$.*

1.4. Transitoriedad y recurrencia

En el caso de los métodos de MCMC, una cadena de Markov (X_n) debe tener propiedades fuertes de estabilidad. La irreducibilidad garantiza que todas las secciones del espacio de estados sean visitadas, pero es importante que sean visitadas un número infinito de veces, esto es, el regreso seguro. A esta propiedad se le llama *recurrencia*.

DEFINICIÓN **9** *En un espacio de estados finito $\mathcal{X}$, un estado $\omega \in \mathcal{X}$ es transitorio si el número promedio de visitas a ω, $\mathbb{E}_\omega(\eta_\omega)$, es finito y recurrente si $\mathbb{E}_\omega(\eta_\omega) = \infty$.*

Para el caso general se tiene la definición siguiente.

DEFINICIÓN **10** *Un conjunto A es recurrente si $\mathbb{E}_x(\eta_A) = \infty$ para todo $x \in A$. El conjunto es uniformemente transitorio si existe una constante M tal*

que $\mathbb{E}_x(\eta_A) < M$ *para todo* $x \in A$. *A es transitorio si existe una cubierta numerable de conjuntos uniformemente transitorios* B_j *en* $\mathcal{X}$, *tales que*

$$A = \bigcup_j B_j.$$

TEOREMA **3** *Sea* (X_n) *una cadena de Markov irreducible con un átomo* α *de medida positiva.*

1. *Si* α *es recurrente, entonces todo* $A \in \mathcal{B}(\mathcal{X})$ *con* $\phi(A) > 0$, *es recurrente.*

2. *Si* α *es transitorio,* $\mathcal{X}$ *es transitorio.*

DEFINICIÓN **11** *Una cadena de Markov* (X_n) *es recurrente si*

1. *existe una medida* ϕ *tal que* (X_n) *es* ϕ*-irreducible y*

2. *para todo* $A \in \mathcal{B}(\mathcal{X})$ *tal que* $\phi(A) > 0$, $\mathbb{E}_x(\eta_A) = \infty$ *para todo* $x \in A$.

La cadena es transitoria si es ϕ-irreducible y $\mathcal{X}$ es transitorio.

TEOREMA **4** *Una cadena* ϕ*-irreducible es recurrente o es transitoria.*

Del Teorema 3 y de la Definición 11 es claro que si (X_n) es ϕ-irreducible entonces se puede determinar si es transitoria o recurrente con la inspección de un átomo α cuya medida sea positiva, es decir, no es necesario explorar todo el espacio de estados.

El siguiente resultado es un criterio más sencillo para determinar la recurrencia de una cadena de Markov pues está en términos de un conjunto pequeño C, y de la probabilidad de que el primer retorno a C ocurra en un tiempo finito, $P_x(\tau_c < \infty)$.

PROPOSICIÓN **3** *Una cadena* (X_n) ϕ*-irreducible es recurrente si existe un conjunto pequeño* C *con* $\phi(C) > 0$, *tal que* $P_x(\tau_c < \infty) = 1$ *para todo* $x \in C$.

1.4.1. Harris-Recurrencia

La noción de *Harris-recurrencia* fue introducida por Harris en 1956. Como se puede sospechar esta propiedad es más fuerte que la sola propiedad de recurrencia. En este caso no basta con que el número promedio de retornos a cada conjunto de medida positiva sea infinito, ahora es necesario que el número de retornos sea infinito con probabilidad 1.

DEFINICIÓN **12** *Un conjunto A es Harris-recurrente si $P_x(\eta_A = \infty) = 1$ para todo $x \in A$. La cadena (X_n) es Harris-recurrente si existe ϕ una medida bajo la cual (X_n) es ϕ-irreducible, y todo $A \in \mathcal{B}(\mathcal{X})$ con $\phi(A) > 0$ es Harris-recurrente.*

La siguiente proposición expresa la Harris-recurrencia como una condición sobre $P_x(\tau_A < \infty)$.

PROPOSICIÓN **4** *Si para todo $A \in \mathcal{B}(\mathcal{X})$, $P_x(\tau_A < \infty) = 1$ para todo $x \in A$, entonces $P_x(\eta_A = \infty) = 1$ para todo $x \in A$, y (X_n) es Harris-recurrente.*

Observe que la propiedad de Harris-recurrencia no es necesaria cuando $\mathcal{X}$ es finito o contable, pues en este caso es posible mostrar que $\mathbb{E}_x(\eta_x) = \infty$ si y sólo si $P_x(\tau_x < \infty) = 1$ para toda $x \in \mathcal{X}$.

En este caso un resultado similar al mencionado en la Proposición 3 es el siguiente. Sólo obsérvese que ahora se requiere que $P_x(\tau_C < \infty) = 1$ para todo $x \in \mathcal{X}$, no sólo en C.

TEOREMA **5** *Si (X_n) es una cadena de Markov ϕ-irreducible con un conjunto pequeño C tal que $P_x(\tau_C < \infty) = 1$ para toda $x \in \mathcal{X}$, entonces (X_n) es Harris-recurrente.*

En [3] se menciona que en la mayoría de los algoritmos de MCMC se satisface la Harris-recurrencia.

1.5. Medidas invariantes

DEFINICIÓN **13** *Una medida σ-finita π es invariante para el kernel de transición $K(\cdot,\cdot)$ (y para la cadena asociada) si*

$$\phi(B) = \int_{\mathcal{X}} K(x,B)\pi(dx), \quad \forall B \in \mathcal{B}(\mathcal{X}).$$

Cuando existe una medida de probabilidad invariante para una cadena irreducible se dice que la cadena es *positiva*.

A la distribución invariante π se le llama *estacionaria* cuando es una medida de probabilidad, pues significa que si $X_0 \sim \pi$ entonces $X_n \sim \pi$, es decir, la cadena es estacionaria en distribución. No es difícil mostrar que cuando la cadena es irreducible y tiene una medida invariante σ-finita ésta es única.

PROPOSICIÓN **5** *Si la cadena (X_n) es positiva, es recurrente.*

De la proposición anterior es claro que ya no es necesario decir Harris-recurrente positiva, ahora es suficiente decir Harris positiva. Es importante señalar que los métodos de MCMC por construcción están asociados a una cadena positiva.

1.6. Ergodicidad y convergencia

Considerando a la cadena (X_n) como evolucionando en el tiempo, es importante preguntarnos ¿a dónde converge?, es decir, cuál será la distribución de X_n para n muy grande. Una candidata natural es la distribución invariante π. A continuación se dan las condiciones que se deben cumplir en el caso numerable y en el caso general para que π coincida con la distribución límite de la cadena.

TEOREMA **6** *Para una cadena de Markov recurrente positiva y aperiódica sobre un espacio contable, para todo estado inicial x*

$$\lim_{n\to\infty} ||K^n(x,.) - \pi||_{TV} = 0.$$

Es decir, K^n converge asintóticamente a π con respecto a la norma de variación total[2] (TV por sus siglas en inglés) sin importar cuál sea el punto inicial x.

Para el caso general el resultado que se tiene es la convergencia digamos en promedio de K^n a π para cualquier distribución inicial μ.

TEOREMA **7** *Si (X_n) es Harris positiva y aperiódica, entonces*

$$\lim_{n\to\infty} || \int K^n(x,.)\mu(dx) - \pi||_{TV} = 0$$

para toda distribución inicial μ.

TEOREMA **8** *Sea (X_n) positiva, recurrente, y aperiódica.*

1. *Si $\mathbb{E}^\pi(|h(X)|) = \infty$, entonces $\mathbb{E}_x(|h(X_n)|) \to \infty$ para toda x.*

2. *Si $\int |h(x)|\pi(dx) < \infty$, entonces*

$$\lim_{n\to\infty} \sup_{|m(x)|\leq|h(x)|} |\mathbb{E}_y(m(X_n)) - \mathbb{E}^\pi(m(X))| = 0$$

en todo conjunto pequeño C tal que

$$\sup_{y\in C} \mathbb{E}_y \left(\sum_{t=0}^{\tau_C - 1} h(X_n) \right) < \infty.$$

1.7. Teoremas Límite

Los resultados mencionados hasta este punto permiten justificar los algoritmos de simulación, sin embargo, no dan información directa sobre la única observación disponible de P_x^n, x_n. Con estos resultados se pueden determinar las propiedades probabilísticas del comportamiento promedio de la cadena en un instante fijo, pero estas propiedades no aportan información para el

[2] $||\mu_1 - \mu_2||_{TV} = \sup_A |\mu_1(A) - \mu_2(A)|$.

control de la convergencia de una simulación. Lo que es relevante en este caso son las propiedades de la realización (x_n) de la cadena.

Los problemas por los que no es posible aplicar directamente los teoremas límite clásicos a la muestra $(X_1, X_2, \ldots, X_n)$ son: el primero es por la dependencia entre cualesquiera dos observaciones sucesivas, esto es, la propiedad Markoviana, y el segundo es por la no estacionariedad de la sucesión (ya que la distribución de X_0 puede ser diferente de π).

A continuación se da un resultado de convergencia equivalente a la Ley de los Grandes Números que es conocido como Teorema de Ergodicidad.

1.7.1. Teorema de Ergodicidad

Dada una muestra $X_1, X_2, \ldots, X_n$ de una cadena de Markov, en lo que sigue se estudia el comportamiento límite de las sumas parciales

$$S_n(h) = \frac{1}{n} \sum_{i=1}^{n} h(X_i)$$

cuando $n \to \infty$.

En principio considérese la definición de función armónica. Como se verá pronto las funciones armónicas estan estrechamente relacionadas con la recurrencia Harris de una cadena.

DEFINICIÓN **14** *Una función medible h es armónica para la cadena de Markov (X_n) si*

$$\mathbb{E}[h(X_{n+1}|x_n)] = h(x_n).$$

PROPOSICIÓN **6** *Para una cadena de Markov positiva, si las funciones constantes son las únicas funciones armónicas acotadas, entonces la cadena es recurrente Harris.*

La demostración de esta proposición no se da con todo detalle, pero se da un esbozo en el que hay un resultado que ayuda a comprender la importancia de las funciones armónicas en la versión de la Ley de los Grandes Números para cadenas de Markov.

Considérese la probabilidad de un número infinito de visitas $Q(x, A) = P_x(\eta_A = \infty)$ como una función de x, $h(x)$. Ocurre que $h(x)$ es una función armónica, ya que

$$\mathbb{E}_{x_n}(h(X_{n+1})) = \mathbb{E}_{x_n}(P_{x_{n+1}}(\eta_A = \infty)) = P_{x_n}(\eta_A = \infty) = h(x_n).$$

De aquí que $Q(x, A)$ sea constante en x. En [1] se argumenta que $Q(x, A)$ sigue una ley $0 - 1$ y se prueba que para toda x, $Q(x, A) = 1$, de donde se establece que la cadena es Harris.

De la Proposición 6 se tiene un resultado que puede ser considerado como una propiedad de continuidad, esto es: por inducción se tiene que una función armónica h satisface $h(x) = \mathbb{E}_x(h(X_n))$, por el Teorema 8 $h(x)$ es casi seguramente igual a $\mathbb{E}^\pi(h(X))$, lo que significa que es constante casi en cualquier parte. En el caso de una cadena Harris, la Proposición 6 establece que esto implica que $h(x)$ es constante en todas partes.

La Proposición 6 permite asegurar la recurrencia Harris de la cadena de Markov asociada a algunos algoritmos de MCMC, en particular al algoritmo de Metropolis-Hastings.

Como se mencionó antes las funciones armónicas permiten caracterizar la recurrencia Harris, esto es claro considerando que el recíproco de la Proposición 6 es cierto.

LEMA **2** *Para las cadenas de Markov recurrentes Harris, las funciones constantes son las únicas funciones armónicas acotadas.*

Una consecuencia del Lema 2 es que si se tiene una cadena de Markov recurrente Harris con distribución estacionaria π, y si ocurre que $S_n(h)$ converge μ_0-casi seguramente a

$$\int_{\mathcal{X}} h(x)\pi(dx),$$

para μ_0 alguna distribución inicial, entonces esta convergencia ocurre para cualquier μ distribución inicial. Aún más, la probabilidad

$$P_x(S_n(h) \to \mathbb{E}^\pi(h))$$

es armónica, i.e. no depende de x. Esto muestra que la recurrencia Harris significa estabilidad fuerte, ya que la convergencia casi segura se reemplaza por convergencia en todo punto.

El siguiente resultado conocido como la Ley de los Grandes Números para cadenas de Markov o Teorema de Ergodicidad garantiza la convergencia de $S_n(h)$.

TEOREMA **9 Teorema de Ergodicidad** *Si (X_n) tiene una medida invariante σ-finita π, las siguientes dos afirmaciones son equivalentes:*

1. *Si $f, g \in L^1(\pi)$ con $\int g(x)d\pi(x) \neq 0$, entonces*
$$\lim_{n\to\infty} \frac{S_n(f)}{S_n(g)} = \frac{\int f(x)\pi(x)}{\int g(x)\pi(x)}.$$
2. *La cadena de Markov (X_n) es recurrente Harris.*

Demostración Si (1) se cumple entonces considérese a f como la función indicadora de cualquier conjunto A de medida finita y positiva, y a g como cualquier función cuya integral es finita y positiva, en [1] se menciona que esto implica lo siguiente:
$$P_x(\eta_A = \infty) = 1,$$
para toda x, con lo cual se muestra la recurrencia Harris de la cadena.

Para mostrar que (2) implica (1) se supone que existe un átomo α y se definen $\tau_\alpha(k)$ como el tiempo de la $(k+1)$-ésima visita a α, y l_n como el número de visitas a α que han ocurrido al tiempo n. En la Figura 1.1 se muestran esquemáticamente estas cantidades.
Además se definen
$$S_j(f) = \sum_{i=\tau_\alpha(j)+1}^{\tau_\alpha(j+1)} f(x_i),$$
para $j \geq 0$. Por la propiedad fuerte de Markov las variables aleatorias $\{S_j(f) : j \geq 0\}$ son idependientes e idénticamente distribuídas con esperanza común
$$\mathbb{E}_\alpha(S_1(f)) = \mathbb{E}_\alpha(\sum_{i=1}^{\tau_\alpha} f(x_i)) = \int f(x)d\pi(x).$$

Considérese la siguiente serie de desigualdades (ver Figura 1.1)
$$\sum_{i=\tau_\alpha(0)+1}^{\tau_\alpha(l_n)} f(x_i) \leq \sum_{i=1}^{n} f(x_i) \leq \sum_{i=1}^{\tau_\alpha(l_n+1)} f(x_i). \tag{1.1}$$

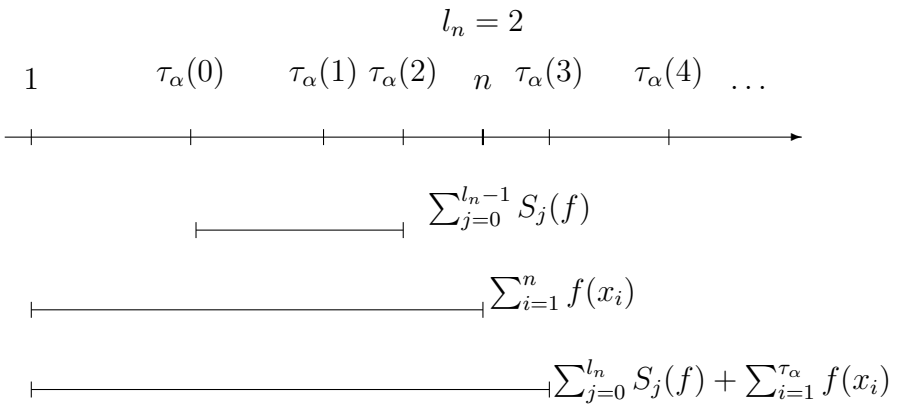

Figura 1.1: Definición de $\tau_\alpha(k)$ y l_N.

Ahora se expresa (1.1) en términos de los $S_j(f)$, esto es:

$$\sum_{j=0}^{l_n-1} S_j(f) \leq \sum_{i=1}^{n} f(x_i) \leq \sum_{j=0}^{l_n} S_j(f) + \sum_{i=1}^{\tau_\alpha(0)} f(x_i).$$

Ya que la misma relación se cumple con g tenemos

$$\frac{\sum_{i=1}^{n} f(x_i)}{\sum_{i=1}^{n} g(x_i)} \leq \frac{l_n}{l_n - 1} \frac{\left(\sum_{j=0}^{l_n} S_j(f) + \sum_{k=1}^{\tau_\alpha} f(x_k)\right)/l_n}{\sum_{j=0}^{l_n-1} S_j(g)/(l_n - 1)}.$$

Aplicando la Ley de los Grandes Números para variables aleatorias i.i.d. se tiene que

$$\frac{1}{l_n} \sum_{j=0}^{l_n} S_j(f) \to \mathbb{E}(S_1(f)) = \int f(x) d\pi(x),$$

y lo mismo ocurre para g. Entonces

$$\limsup_{n\to\infty} \frac{\sum_{i=1}^{n} f(x_i)}{\sum_{i=1}^{n} g(x_i)} \leq \frac{\int f(x) d\pi(x)}{\int g(x) d\pi(x)}$$

e intercambiando f por g se obtiene

$$\liminf_{n\to\infty} \frac{\sum_{i=1}^{n} f(x_i)}{\sum_{i=1}^{n} g(x_i)} \geq \frac{\int f(x)d\pi(x)}{\int g(x)d\pi(x)}$$

con lo que se tiene la prueba.

Capítulo 2

El algoritmo de Metropolis-Hastings

2.1. Métodos de Monte Carlo basados en cadenas de Markov

Dada una densidad objetivo f, los métodos de Monte Carlo basados en cadenas de Markov permiten obtener una muestra aproximadamente distribuida como f con la ventaja importante de no tener que simular directamente de f. La estrategia consiste en construir una cadena de Markov ergódica con densidad invariante f.

Dado un valor inicial x_0, se construye una cadena de Markov (X_t) a través de una densidad condicional $q(y|x)$ cuya densidad invariante es f. Esto asegura la convergencia en distribución de (X_t) a una v.a. con distribución f. De aquí para una t suficientemente grande $X_t, X_{t+1}, \ldots$ puede tomarse como una muestra **dependiente** de f.

Definición **15** *Un método de Monte Carlo basado en Cadenas de Markov (MCMC) usado para simular de una distribución f es cualquier método que*

genere una cadena de Markov ergódica (X_t) *con distribución estacionaria* f*.*

En comparación con otras técnicas para simulación esta estrategia puede parecer no óptima pues depende de resultados asintóticos, sin embargo su importancia está en su generalidad como se mostrará pronto.

Siempre que no haya lugar a confusión se hará referencia a los métodos de MCMC para simulación como métodos de MCMC.

Cuando la condición de independencia de la muestra no es necesaria, se puede usar una sucesión (X_t) obtenida através de una cadena de Markov asociada con algún método de MCMC de manera similar que una muestra i.i.d. Una situación de este tipo es cuando se busca examinar las propiedades de la densidad objetivo, como por ejemplo, evaluar $E_f[h(X)]$. En este caso el Teorema de Ergodicidad garantiza la convergencia del promedio empírico

$$\hat{A} = \frac{1}{T}\sum_{t=1}^{T} h(X_t) \tag{2.1}$$

a $E_f[h(X)]$.

De la definición anterior es claro que es posible proponer un número infinito de métodos de MCMC. Sin embargo, el método conocido como el algoritmo de Metropolis-Hastings (M-H) [10] [11] tiene ventajas importantes con respecto a otros métodos de MCMC. La ventaja principal es su universalidad, pues las restricciones que la densidad objetivo debe satisfacer son mínimas. Otra ventaja es su flexibilidad ya que es posible implementar este procedimiento de diversas maneras.

2.2. El algoritmo de Metropolis-Hastings

2.2.1. Definición

El algoritmo de M-H inicia con una densidad objetivo f. De algún modo se debe elegir una densidad condicional $q(y|x)$, definida con respecto a la medida f. Para que el algoritmo pueda ser implementado en la práctica es necesario que no sea difícil simular de $q(.|x)$, y que ésta sea conocida analíticamente excepto posiblemente por una constante independiente de x, o que sea simétrica; es decir, que $q(y|x) = q(x|y)$.

El procedimiento de M-H que está asociado a la densidad f y a la densidad condicional q genera la cadena de Markov (X_t) con la siguiente transición:

ALGORITMO **1** *-Metropolis-Hastings-*

Dado x_t,

1. *Genere* $Y_t \sim q(y|x_t)$.

2. *Asigne*

$$X_{t+1} = \begin{cases} Y_t & \textit{con probabilidad } \rho(x_t, Y_t), \\ x_t & \textit{con probabilidad } 1 - \rho(x_t, Y_t), \end{cases} \tag{2.2}$$

donde

$$\rho(x, y) = min \left\{ \frac{f(y)}{f(x)} \frac{q(x|y)}{q(y|x)}, 1 \right\}.$$

A la distribución q se le conoce como la *distribución instrumental* o como la *distribución de propuestas*.

Observe que la propuesta y_t siempre se acepta cuando el cociente $f(y_t)/q(y_t|x_t)$ es mayor que el cociente anterior $f(x_t)/q(x_t|y_t)$. Pero también es posible aceptar propuestas con las que el cociente correspondiente disminuye aunque ésto ocurre menos seguido. En el caso en que $q(.|x)$ es simétrica la probabilidad ρ únicamente depende del cociente de las densidades, esto es, de $f(y_t)/f(x_t)$.

2.2.2. El kernel de transición

El kernel de transición $K(x, y)$ de la cadena de M-H esta dado por

$$K(x, y) = \begin{cases} q(y|x)\rho(x, y), & \text{si } y \neq x \\ q(y|x)\rho(x, y) + (1 - r(x)), & \text{si } y = x, \end{cases} \tag{2.3}$$

con $r(x) = \int q(x, y)\rho(x, y)dy$.

Definición **16** *Una cadena de Markov con kernel de transición K satisface la ecuación de balance detallado si existe una función f con la que se cumple que*

$$K(x,y)f(x) = K(y,x)f(y)$$

para todo (x,y).

No es difícil mostrar que la cadena de Markov de M-H satisface la ecuación de balance detallado con f (la densidad objetivo). Primero para cualquiera que sea la relación entre x y y ($y \neq x$ o $y = x$) se cumple que si $\rho(x,y) < 1$ entonces necesariamente $\rho(y,x) = 1$, de donde se tiene

$$\rho(x,y)q(y|x)f(x) = \rho(y,x)q(x|y)f(y). \tag{2.4}$$

Ahora considerando el caso $y = x$ y la relación anterior, se sigue que

$$(1 - r(x))f(x) = (1 - r(y))f(y). \tag{2.5}$$

Con (2.4) y con la suma de (2.4) y (2.5) se establece el cumplimiento de la ecuación de balance detallado para $y \neq x$ y $y = x$, respectivamente.

Teorema **10** *Suponga que una cadena de Markov con función de transición K satisface la ecuación de balance detallado con f una función de densidad de probabilidad. Entonces la densidad f es la densidad invariante de la cadena.*

Demostración. Usando la condición de balance detallado se tiene que para todo conjunto medible B,

$$\begin{aligned}
\int_{\mathcal{Y}} K(y,B)f(y)dy &= \int_{\mathcal{Y}}\int_{B} K(y,x)f(y)dxdy \\
&= \int_{\mathcal{Y}}\int_{B} K(x,y)f(x)dxdy \\
&= \int_{B} f(x)dx,
\end{aligned}$$

ya que $\int K(x,y)dy = 1$. Esto prueba que f es la densidad invariante.$\star$
Con lo anterior se muestra el siguiente resultado para la cadena de M-H.

TEOREMA **11** *Para cualquier distribución condicional q, cuyo soporte debe incluir a $\mathcal{E}$ (el soporte de f) f es la distribución estacionaria de la cadena (X_t) generada por el algoritmo [1].*

Demostración. Se sabe que la cadena (X_t) satisface la ecuación de balance detallado y por tanto que f es la densidad invariante. Dado que f es una medida de probabilidad entonces f es una distribución estacionaria de la cadena. $\star$

Este teorema establece la generalidad del algoritmo pues se cumple casi para cualquier distribución condicional q.

Hasta aquí se ha probado que f es una distribución estacionaria para la cadena de Markov de M-H, ahora se enuncian las condiciones se deben cumplir para que la cadena sea aperiódica, irreducible y recurrente positiva, propiedades con las que se garantiza la Ergodicidad.

2.2.3. Propiedades de convergencia

Irreducibilidad y aperiodicidad. Roberts y Smith [14] mostraron que si q es f-irreducible y aperiódica, y si $\rho(x,y) > 0$, para todo x y todo y, entonces la cadena de M-H es f-irreducible y aperiódica.

Recurrencia positiva. La condición que se refiere a la recurrencia se tiene por el razonamiento siguiente: si la cadena es f-irreducible y aperiódica, entonces como f es su medida invariante, la cadena es positiva, lo cual implica que la cadena es recurrente.

Ahora es posible formular un resultado fuerte para la cadena de M-H. La prueba puede verse en [1]

LEMA **3** *Si la cadena (X_t) de M-H es f-irreducible entonces es recurrente Harris.*

Con lo anterior se tiene el siguiente resultado de convergencia tomado de [1].

TEOREMA **12** *Suponer que la cadena de Markov de M-H (X_t) es f-irreducible.*

1. *Si $h \in L^1(f)$, entonces*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} h(X_t) = \int h(x)f(x)dx \ \ a.e.\ f.$$

2. *Si además, (X_t) es aperiódica, entonces*

$$\lim_{n\to\infty} || \int K^n(x,.)\mu(dx) - f||_{TV} = 0$$

para toda distribución inicial μ, donde $K^n(x,.)$ es el kernel para n transiciones.

Demostración. Si (X_t) es f-irreducible, por el Lema 3 es recurrente Harris, por tanto la parte 1 se sigue del Teorema de Ergodicidad 9. La parte 2 es una consecuencia del Teorema 6.$\star$

La primera parte del teorema anterior asegura que es posible utilizar una cadena de Markov de forma similar que una muestra i.i.d cuando se quiere aproximar $E_f[h(x)]$. Y la segunda establece la convergencia asintótica de la distribución de X_t a f sin importar el valor inicial de la cadena, esto es, la ergodicidad de la cadena de M-H.

2.3. Dos ejemplos del algoritmo de M-H

A continuación se describen brevemente dos implementaciones del algoritmo de M-H que son muy conocidas y muy sencillas.

2.3.1. La propuesta independiente

En este método la propuesta Y_t es independiente del valor actual de la cadena de Markov x_t; esto es, $Y_t \sim g(y)$, con g la distribución instrumental. En este caso la probabilidad de aceptación de la propuesta es

$$min\left\{\frac{f(Y_t)g(x_t)}{f(x_t)g(Y_t)}, 1\right\}.$$

La convergencia de la cadena (X_t) depende de las propiedades de g, en el sentido de que (X_t) es irreducible y aperiodica, y por lo tanto ergódica, si y sólo si g es positiva casi en cualquier parte del soporte de f.

2.3.2. La caminata aleatoria

En esta implementación la propuesta Y_t depende del valor actual de la cadena x_t de la siguiente forma

$$Y_t = X_t + \epsilon_t,$$

donde ϵ_t es una perturbación aleatoria con distribución g, independiente de X_t. En este enfoque q es de la forma $g(y-x)$ y debe ser tal que saltos grandes ocurran con probabilidad positiva. Es claro que la convergencia se tiene para esta implementación.
Ahora se describe un tipo de algoritmos de M-H de caracteristicas distintivas.

2.4. Saltos reversibles

Los algoritmos de Metropolis-Hastings son usados para simular de una distribución definida sobre un espacio de dimensión fija, pero cuando la dimensión es una de las variables de las que se quiere simular estos procedimientos no pueden ser aplicados. Este es el caso cuando se trata de seleccionar un modelo de entre varios para ajustarlo a un conjunto de datos, pues diferentes modelos tienen diferentes espacios parametrales. Un ejemplo claro es la determinación del número de componentes en un modelo de mezcla. Los algoritmos de M-H con saltos reversibles permiten simular de una distribución objetivo sobre un espacio de dimensión variable. En el diseño del t-walk se hace uso de un caso particular de estos procedimientos: el caso cuando

sólo se tiene un modelo pero las distribuciones objetivo y de propuestas de cambio están definidas sobre espacios de dimensión diferente. En lo sucesivo se trata esta situación. Los algoritmos de MCMC con saltos reversibles son introducidos en [6] y [7].

2.4.1. Especificación del algoritmo

El objetivo de los algoritmos de M-H con saltos reversibles es el mismo que el de los algoritmos de M-H: construir una cadena de Markov reversible y ergódica con distribución invariante f. La diferencia está en la forma en la que se hacen las propuestas de cambio.

Suponga que x_t es el valor del estado actual X_t de la cadena. Se genera una propuesta Y_t aplicando un mapeo determinístico al estado x_t y a una componente aleatoria U. Esto se puede expresar como $Y_t = g_1(x_t, U)$, con U un vector aleatorio en R^m con densidad $q(x_t, .)$ definida sobre R^m, y $g_1 : R^{n+m} \to R^n$. La propuesta se acepta con una probabilidad $\rho(x_t, Y_t)$ definida más adelante.

Dada la diferencia en la dimensión de los espacios se hacen necesarias varias restricciones. En principio, al considerar el cambio del estado x_t al estado $x_{t+1} = g_1(x_t, u)$ y el cambio en reversa de x_{t+1} a $x_t = g_{1r}(x_{t+1}, u')$ (con $g_{1r} : R^{n+m} \to R^n$) es claro que las variables aleatorias u y u' deben ser de la misma dimensión. Además es necesario tener funciones $g_2 : R^{n+m} \to R^m$ y $g_{2r} : R^{n+m} \to R^m$, tales que el mapeo dado por

$$(x_{t+1}, u') = g(x_t, u) = (g_1(x_t, u), g_2(x_t, u))$$

es uno a uno con

$$(x_t, u) = g^{-1}(x_{t+1}, u') = (g_{1r}(x_{t+1}, u'), g_{2r}(x_{t+1}, u')),$$

y diferenciable.

Con los supuestos anteriores el algoritmo de M-H con saltos reversibles puede resumirse como sigue.

ALGORITMO **2** *-Metropolis-Hastings con saltos reversibles-*

Dado x_t,

1. *Genere* $u \sim q(x_t, .)$.

2. *Calcule* $Y_t = g_1(x_t, u)$ *y* $u' = g_2(x_t, u)$.

   $Y_t = g_1(x_t, u)$.

3. *Asigne*

$$X_{t+1} = \begin{cases} Y_t & \textit{con probabilidad } \rho(x_t, Y_t), \\ x_t & \textit{con probabilidad } 1 - \rho(x_t, Y_t), \end{cases} \tag{2.6}$$

   donde

$$\rho(x, y) = min \left\{ \frac{f(y)}{f(x)} \frac{q(y, u')}{q(x, u)} \left| \frac{\partial g(x, u)}{\partial x \partial u} \right|, 1 \right\}.$$

Ejemplo **1**

Como antes sea f una densidad sobre R^n. Un algoritmo de M-H de caminata aleatoria para simular de f se obtiene con $g(x_t, U) = (x_t + U, -U)$ donde U es generada de una densidad $q(.)$ sobre R^n. En este caso el determinante del Jacobiano es 1 por lo tanto la probabilidad de aceptar una propuesta Y_t dado x_t es

$$\rho(x_t, Y_t) = min \left\{ \frac{f(Y_t)}{f(x_t)} \frac{q(U')}{q(u)}, 1 \right\}.$$

2.5. Mezcla de kerneles

Definición **17** *Sean* $K_1, K_2, \ldots, K_n$ *kerneles de transición cada uno con distribución estacionaria* f, *y* $(w_1, w_2, \ldots, w_n)$ *una distribución de probabilidad. Una mezcla de estos kerneles corresponde a*

$$\hat{K} = w_1 K_1 + w_2 K_2 + \ldots + w_n K_n. \tag{2.7}$$

Las propiedades del kernel $\hat{K}$ se heredan de los kerneles $K_j,\ (j = 1, \ldots, n)$. La distribución invariante de $\hat{K}$ también es f. La irreducibilidad y aperiodicidad se garantizan si se tiene al menos un kernel K_j irreducible y aperiódico.

La importancia de la mezcla está en poder construir un kernel K con propiedades fuertes de estabilidad con base en otros kerneles K_j con menos propiedades y que comúnmente son más fáciles de definir.

PROPOSICIÓN **7** *Si K_1 y K_2 son dos kerneles con la misma distribución estacionaria f y si K_1 es el kernel de una cadena de Markov ergódica (X_t), el kernel mezcla*

$$\hat{K} = wK_1 + (1 - w)K_2,\ (0 < w < 1)$$

también es el kernel de (X_t).

Capítulo 3

El algoritmo t-walk

3.1. Introducción

El algoritmo t-walk podría ser visto como un tipo especial de AMS (muestreador adaptable de Metropolis) cuya distinción está en la propuesta de salto. Las ventajas del algoritmo son dos: la primera es que supera los problemas comunes de sensibilidad a la escala y a la correlación, y la segunda es que no requiere configuración alguna de parámetros por el usuario final.

En este capítulo se detalla la estructura del t-walk, así como las ventajas que ésta representa. En la primera sección se señala el lugar del t-walk en el conjunto de algoritmos de MCMC adaptable. En la segunda se explica con detalle su diseño, se muestra su convergencia, y se prueba la invarianza ante cambios en la escala y en el punto de referencia. En la tercera parte se incluyen varios ejemplos clasificados por la dimensión del dominio de la distribución objetivo. En la última sección se dan los detalles de la implementación del algoritmo en el lenguaje C++.

3.2. Generalidades del t-walk

El algoritmo t-walk no es precisamente una especialización del AMS pero tiene una estructura parecida. Lo que lo distingue de otros algoritmos, co-

mo el snooker (mencionado en la introducción), es la forma en que se hace la propuesta de cambio. Ésta es una mezcla de cuatro propuestas que son complementarias. La primera es similar a las propuestas que se hacen en los algoritmos de muestreo adaptable direccional, ADS, (también mencionado en la introducción de la tesis). Las otras, especialmente la segunda, usan un "tamaño de paso" como $||x_n - x'_n||$ para generar una caminata aleatoria. Esta caminata tendrá un tamaño de paso adaptado en el espacio de estados original (recordando que en AMS se construye la cadena de Markov en un espacio de dimensión mayor). En el caso del t-walk el conjunto actual incluye únicamente dos puntos, de los cuales, como ocurre en AMS, se selecciona uno de ellos para ser actualizado; la actualización se hace con base en el conjunto actual, es decir con base en ambos puntos. En el algoritmo se han establecido algunos parámetros de modo que la eficiencia del mismo sea óptima, pero no se hace necesario ningún otro ajuste adicional.

La aplicación del t-walk está limitada a densidades continuas; por eficiencia es recomendable para casos de dimensión pequeña a moderada. La ventaja fundamental del algoritmo es que supera los problemas comunes de sensibilidad a la escala y a la estructura de correlación de la densidad objetivo, de aquí que sea útil como una herramienta en análisis exploratorios de distribuciones complicadas o como un simulador genérico.

El diseño del t-walk se debe a Christen y Fox [4]. La descripción siguiente se basa principalmente en esta referencia.

3.3. El diseño

Para una distribución objetivo $\pi(x), x \in \mathcal{X}$ ($\mathcal{X}$ es de dimensión n y es un subconjunto de $\Re^n$), se forma la nueva distribución objetivo $f(x, x') = \pi(x)\pi(x')$ en el espacio producto correspondiente $\mathcal{X} \times \mathcal{X}$. La distribución instrumental se denota por

$$q_h\{(y, y')|(x, x')\},$$

donde $h(x, x')$ es una variable aleatoria necesaria para formar la propuesta. Se consideran dos casos de cambio:

$$(y, y') = \begin{cases} (x, h(x', x)), & \text{con probabilidad 0.5} \\ (h(x, x'), x'), & \text{con probabilidad 0.5.} \end{cases} \tag{3.1}$$

En adelante se denotará $h(x, x')$ por h y $h(x', x)$ por h'.

Siguiendo un esquema de Metropolis-Hastings se calcula el cociente de aceptación dado por

$$\frac{f(y, y')q_h\{(x, x')|(y, y')\}}{f(x, x')q_h\{(y, y')|(x, x')\}}. \tag{3.2}$$

Denotando la función de densidad de $h(x, x')$ por $g(.|x, x')$, el cociente para el primer caso en (3.1) es igual a

$$\frac{\pi(y')g(x'|y', x)}{\pi(x')g(y'|x', x)} \tag{3.3}$$

considerando que $y = x$. Para el segundo caso, donde $y' = x'$, el cociente de aceptación es

$$\frac{\pi(y)g(x|y, x')}{\pi(x)g(y|x, x')}. \tag{3.4}$$

A continuación se describen las cuatro elecciones para h que fueron seleccionadas por los autores. Éstas, según se menciona en [4], dan lugar a una velocidad de mezclado buena para un rango amplio de densidades.

1. La travesía (*the traverse step*):

$$h_1(x, x') = x' + \beta(x' - x), \tag{3.5}$$

donde $\beta \in \Re^+$ es una variable aleatoria con densidad $\phi_1(.)$ (Figura 3.1).

La propuesta es casi simétrica en dimensiones bajas, lo cual fue obtenido al buscar satisfacer $\phi_1(1/\beta) = \phi_1(\beta)$ como se muestra enseguida.

A continuación se calcula el cociente de aceptación de la propuesta. Considere el caso 2 de (3.1). Dada la forma de la propuesta (3.5) el cociente puede calcularse mediante (3.4), pero también puede calcularse usando

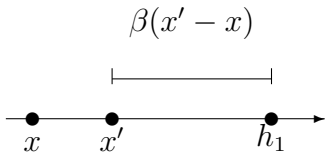

Figura 3.1: La travesía

$$\frac{\pi(y)\phi_1(\beta')}{\pi(x)\phi_1(\beta)}\left|\frac{\partial g(x,\beta)}{\partial x \partial \beta}\right| \tag{3.6}$$

vía la metodología del salto reversible vista en 2.4. Para h_1 y h_2 se evaluará el cociente de aceptación usando (3.6).

En el caso de estudio se ha elegido a x para ser actualizado mientras que x' permanece sin cambio, esto es, $y' = x'$. La propuesta de cambio está dada por $y = x' + \beta(x' - x)$ y el movimiento de reversa es $x = x' + \beta'(x' - y)$ con $\beta' = 1/\beta$. De donde $(y, \beta') = g(x, \beta) = (x' + \beta(x - x'), 1/\beta)$. El determinante de la matriz Jacobiana es

$$\begin{vmatrix} \frac{\partial y}{\partial x} & \frac{\partial y}{\partial \beta} \\ \frac{\partial \beta'}{\partial x} & \frac{\partial \beta'}{\partial \beta} \end{vmatrix} = \begin{vmatrix} -\beta I_n & x' - x \\ 0, \cdots, 0 & -\beta^{-2} \end{vmatrix} = \beta^{n-2}.$$

donde I_n es la matriz identidad de $n \times n$.

Ya que $\phi_1(1/\beta) = \phi_1(\beta)$ el cociente de aceptación se reduce a $\frac{\pi(y)}{\pi(x)}\beta^{n-2}$. Siguiendo un razonamiento similar el cociente para el caso 1 de (3.1) es $\frac{\pi(y')}{\pi(x')}\beta^{n-2}$. Dado que el cociente de M-H es el cociente de las densidades objetivo multiplicado por un término que depende de la dimensión se dice que la propuesta es casi simétrica. En el caso en que $n = 2$ es claro que la propuesta es simétrica.

Hasta aquí se ha mantenido la hipótesis de que ϕ_1 satisface $\phi_1(1/\beta) = \phi_1(\beta)$. Una densidad de esta clase se puede obtener con base a una densidad $\psi(.)$ sobre $\Re^+$ y definiendo $\phi_1(\beta) = K\psi(\beta^{-1} - 1)I_{(0,1]}(\beta) + \psi(\beta - 1)I_{(1,\infty)}(\beta)$, con K la constante de normalización. Un resultado

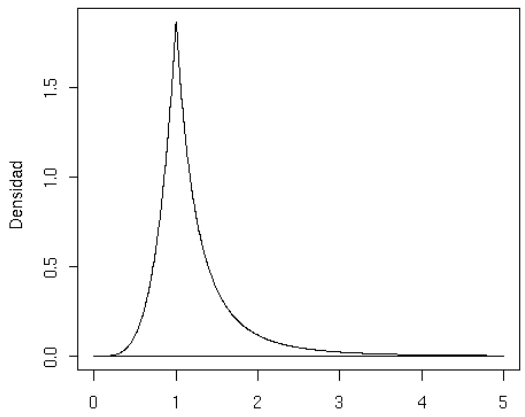

Figura 3.2: $\phi_1(\beta)$ con $a = 4$ dando $P(\beta < 2) \approx 0.92$

conveniente y sencillo se obtiene con $\psi(y) = (a-1)(y+1)^{-a}$, para cualquier $a > 1$, esto es:

$$\phi_1(\beta) = \frac{a-1}{2a}\left\{(a+1)\beta^a I_{(0,1]}(\beta)\right\} + \frac{a+1}{2a}\left\{(a-1)\beta^{-a} I_{(1,\infty)}(\beta)\right\}.$$

Debido a que la longitud de paso deseable es una cantidad alrededor de la longitud de $||x-x'||$, tomando $a = 4$ se consiguió que $P(\beta < 2) \approx 0.9$. La gráfica de $\phi_1(\beta)$ con $a = 4$ se presenta en la Figura 3.2.

Para simular de $\phi_1(\beta)$ se puede usar el siguiente algoritmo

$$\beta = \begin{cases} u^{1/(a+1)}, & \text{con probabilidad } \frac{a-1}{2a} \\ u^{1/(1-a)}, & \text{con probabilidad } \frac{a+1}{2a}, \end{cases}$$

donde $u \sim U(0,1)$.

2. La caminata (*the walk step*):

$$h_2(x, x')_j = x_j + (x_j - x'_j)z_j,$$

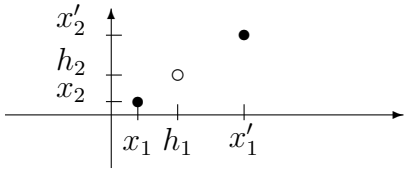

Figura 3.3: La caminata.

para $j = 1, 2, \ldots, n$, donde $z_j \in \Re$ son variables aleatorias i.i.d con densidad $\phi_2(.)$ (Figura 3.3). Al calcular el cociente de aceptación según el enfoque del salto reversible se encuentra, de forma similar a la travesía, que para que la propuesta sea simétrica es necesario que:

$$\phi_2\left(\frac{-z}{1+z}\right) = (1+z)\phi_2(z).$$

Lo cual se obtiene al tomar

$$\phi_2(z) = \begin{cases} \frac{1}{k\sqrt{1+z}}, & z \in [\frac{-a}{1+a}, a] \\ 0, & \text{en otro caso,} \end{cases} \tag{3.7}$$

para $a > 0$, con constante de normalización $k = 2(\sqrt{(1+a)} - 1/\sqrt{(1+a)})$. En la implementación del t-walk $a = 1/2$. En la Figura 3.4 se muestra la gráfica de ϕ_2.

Debido a que el cociente de M-H en el primer y segundo caso es igual a 1, el cociente de aceptación resultante es simplemente el cociente de las densidades objetivo.

Es sencillo simular de esta densidad usando la inversa de la distribución acumulativa, esto es,

$$z = \frac{a}{1+a}(-1 + 2u + au^2)$$

con $u \sim U(0, 1)$.

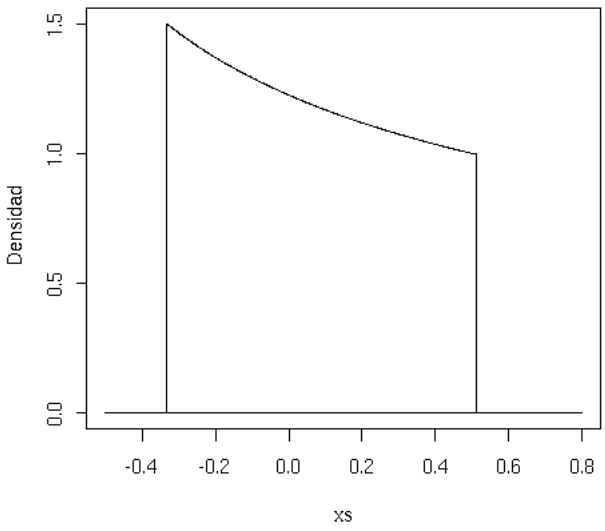

Figura 3.4: $\phi_2(z)$ con $a = 1/2$.

Las propuestas 3 y 4 se pueden considerar como perturbaciones que se hacen eventualmente para dar movilidad a la cadena.

3. El salto (*the hop step*):

$$h_3(x, x') = (x_j + z_j\sigma(x, x')/3),$$

con $z_j \sim N(0, 1)$, donde $\sigma(x, x') = max_{j=1,\ldots,n}|x_j - x'_j|$.

Para esta propuesta

$$g_{h_3}(h|x, x') = \frac{(2\pi)^{-n/2}3^n}{\sigma(x, x')^n} \exp\{\frac{-9}{2\sigma(x, x')^2} \sum_{j=1}^{n}(h_j - x_j)^2\}.$$

Observe que el movimiento se centra en x.

4. La explosión (*the blow step*):

$$h_4(x, x') = (x'_j + \sigma(x, x')z_j),$$

con $z_j \sim N(0, 1)$. Entonces se tiene

$$g_{h_4}(h|x, x') = \frac{(2\pi)^{-n/2}}{\sigma(x, x')^n} \exp\{\frac{-1}{2\sigma(x, x')^2} \sum_{j=1}^{n}(h_j - x'_j)^2\}.$$

Observe que a diferencia de la caminata y del salto pequeño, este movimiento se centra en x'.

3.3.1. Convergencia

Debido a que el t-walk fue construido como un algoritmo de M-H en el espacio producto, su convergencia se tiene bajo las condiciones usuales como se muestra enseguida.

Sea $K_i(.,.)$ el kernel de transición de M-H correspondiente para la propuesta q_{h_i}, donde $i = 0, 1, \ldots, k(k = 4)$. Para el caso $i = 0$ se define $K_0(x, y) = \delta_x(y)$, el cual, al igual que cada K_i, satisface la ecuación de balance detallado con f. Ahora se forma el kernel de transición:

$$K\{(x,x'),(y,y')\} = \sum_{i=0}^{4} w_i K_i\{(y,y')|(x,x')\},$$

donde $\sum_{i=1}^{4} w_i = 1$, dando como consecuencia que la combinación lineal aleatoria de kerneles también satisfaga la condición de balance detallado con f. Asumiendo además que K es f-irreducible (observe que los movimientos 3 y 4 garantizan la irreducibilidad), entonces f es la distribución límite de K (la aperiodicidad fuerte se asegura con el kernel K_0).

Las probabilidades de la mezcla w_0, w_1, w_2, w_3 y w_4 se fijaron en: 0.0008, 0.4914, 0.4914, 0.0082, 0.0082, respectivamente. Ya que con estos valores se minimiza el tiempo integrado de autocorrelación de la cadena de Markov.[1]

$$(1-0.0082)K_1 + 0.0082K4$$
$$(1-0.0082)K_1 + 0.0082K3$$
$$0.5K_1 + 0.5K_2$$
$$0.4918K_1 + 0.4918K_2 + 0.0082K_3 + 0.0082K4$$

En las 4 gráficas de la Figura 3.5 se muestra el comportamiento de los kerneles $K_1, \ldots, K_4$. En cada una de ellas se presenta una muestra de 4000 puntos del círculo unitario (todos son igualmente probables). En la gráfica superior izquierda se ha usado la mezcla $(1-0.0082)K_1 + 0.0082K_4$. El kernel K_1 genera puntos en la recta que une a x con x', y el kernel K_4 perturba la recta cambiandole la orientación. En la gráfica superior derecha se usó la mezcla $(1-0.0082)K_1 + 0.0082K_3$. En este caso la perturbación que produce el kernel K_3 es menor. En la gráfica inferior izquierda la mezcla usada fue $0.5K_1 + 0.5K_2$. Como vemos se cubre una mayor parte del círculo con estos

[1]Cuando se mide el error del estimador (2.1) se consideran las autocorrelaciones de la cadena, esto es: $Var(\hat{A}) \approx \tau(Var_f(V(X))/L)$. Con $Var_f(V(X))$ la varianza común de $V(X_t)$ y $V(X_{t+s})$ considerando $t \to \infty$ y asumiendo que $X_0 \sim f$. τ es el *tiempo integrado de autocorrelación* y es una cantidad que depende de las autocorrelaciones entre X_t y X_{t+s} para $s = 1, 2, \ldots, L$ y $t = 1, 2, \ldots,$. Para una muestra no correlacionada $\tau = 1$. Minimizar τ es minimizar las autocorrelaciones de la cadena. También puede verse como hacer que el error de (2.1) y el error del mismo estimador pero usando una muestra independiente sean lo más parecidos posible.

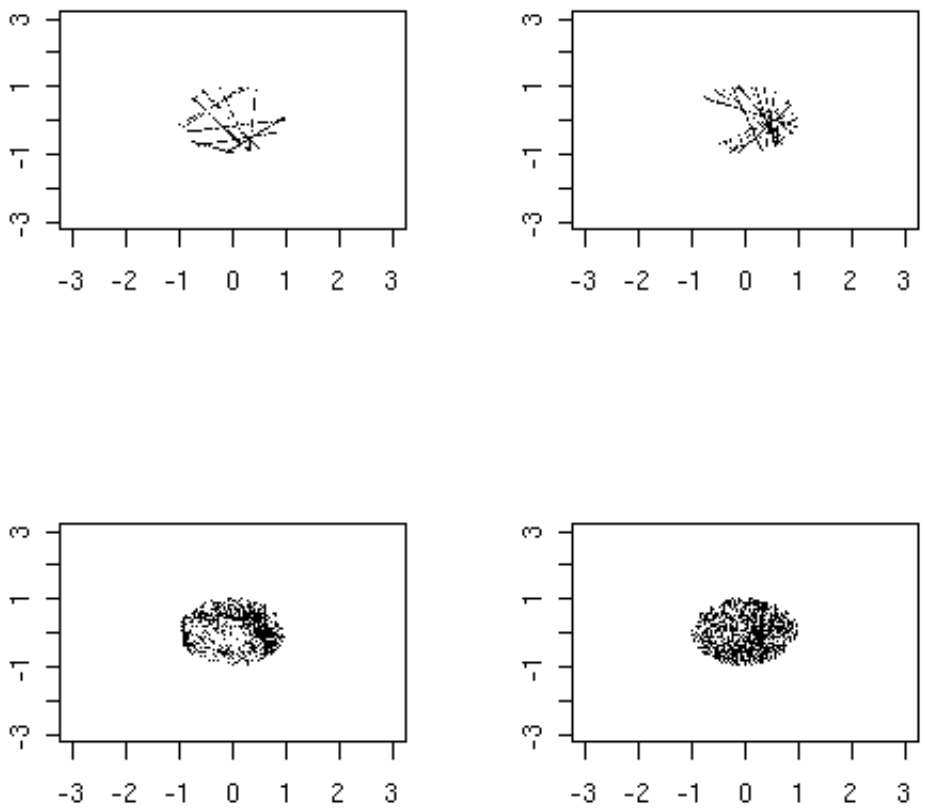

Figura 3.5: Muestras de 4000 puntos del círculo unitario. Por filas y de izquierda a derecha la simulación usando la mezcla: $(1-0.0082)K_1+0.0082K4$, $(1-0.0082)K_1+0.0082K3$, $0.5K_1+0.5K_2$ y $0.4918K_1+0.4918K_2+0.0082K_3+0.0082K4$.

dos kerneles. En la última gráfica, la inferior derecha, se incluyeron los 4 kerneles en la mezcla. Las probabilidades de la mezcla son prácticamente las que minimizan el tiempo de mezclado de la cadena: $0.4918K_1 + 0.4918K_2 + 0.0082K_3 + 0.0082K_4$. El espacio se cubre mejor en este caso.

3.3.2. Propiedades

El teorema siguiente, extraído de la referencia citada [4], establece que el algoritmo t-walk es invariante ante cambios en la escala y en el punto de referencia.

TEOREMA **13** *Dada una transformación del espacio $\mathcal{X}$, $\phi(x) = ax+b$, donde $a \in \Re$, $a \neq 0$ y $b \in \Re^n$, que genera la distribución objetivo nueva $\lambda(z) = |a^{-n}|\pi(\phi^{-1}(z))$, es posible generar una realización del t-walk ya sea aplicando el kernel con λ como distribución objetivo, con valores iniciales z_0, z_0', o aplicando el kernel a π, con valores iniciales $\phi^{-1}(z_0)$, $\phi^{-1}(z_0')$, y entonces transformar la cadena resultante con ϕ.*

Prueba: Sea $V_0 = (\phi^{-1}(z_0), \phi^{-1}(z_0'))$ y $W_1 \in \phi(\mathcal{X}) \times \phi(\mathcal{X})$. Es posible mostrar con cálculos sencillos que $|a^{-n}|q_{h_j}(\phi^{-1}(W_1)|V_0) = q_{h_j}(W_1|\phi(V_0))$ para $j = 1, 2, 3, 4$. Usando esto es fácil ver que las probabilidades de aceptación de M-H, usando π y λ, satisfacen $\rho^{\pi}_{h_j}(V_0, \phi^{-1}(W_1)) = \rho^{\lambda}_{h_j}(\phi(V_0), W_1)$. Entonces es claro que $\rho^{\pi}_{h_j}(V_0, \phi^{-1}(W_1))|a^{-n}|q_{h_j}(\phi^{-1}(W_1)|V_0) = \rho^{\lambda}_{h_j}(\phi(V_0), W_1)q_{h_j}(W_1|\phi(V_0))$. Este hecho aunado a que la probabilidad de no saltar en cualquiera de los casos es la misma, $1-r^{\lambda}_{h_j}(\phi(V_0)) = 1-r^{\pi}_{h_j}(V_0)$, establece el resultado deseado. $\star$

Este teorema establece que aplicar el kernel del t-walk con π y transformarlo con ϕ tiene densidad $|a^{-n}|q_{h_j}(\phi^{-1}(W_1)|V_0) + \delta_{V_0}(W_1)(1 - r^{\lambda}_{h_j}[\phi(V_0)])$ lo cual es igual a $K_\lambda(\phi(V_0), W_1)$. Es inmediato que esto se mantiene para n pasos del t-walk y de aquí que, para cualquier conjunto B (del espacio transformado) $K^n_\pi(V_0, \phi^{-1}(B)) = K^n_\lambda(\phi(V_0), B)$ y ya que también $f_\pi(\phi^{-1}(B)) = f_\lambda(B)$ se tiene que

$$||K^n_\pi(V_0, .) - f_\pi(.)||_{TV} = ||K^n_\lambda(\phi(V_0), .) - f_\lambda(.)||_{TV},$$

donde $f_\pi(A) = \int_A \pi(dx)\pi(dx')$ y $f_\lambda(B) = \int_B \lambda(dx)\lambda(dx')$.

Lo anterior prueba una característica importante del t-walk; su eficiencia (velocidad de convergencia, autocorrelación, etc.) se mantiene aún con un cambio en la escala y/o posición, como se hace con ϕ. Y se puede ir más allá, si el t-walk se limita a los movimientos 1 y 2, el teorema es válido para un cambio en escala más general, es decir, cuando $a = diag(a_j)$, una matriz diagonal, $a_j \in \Re$, $a_j \neq 0$.

3.4. Experimentación con el algoritmo

En esta sección se analizan varias simulaciones de diferentes distribuciones realizadas con el t-walk. Estas distribuciones se han clasificado por su dimensión. En la primera parte se analizan cuatro bidimensionales, mientras que en la segunda las distribuciones están en dimensiones mayores. Estas distribuciones fueron elegidas de modo que se mostrará el desempeño del algoritmo ante casos típicamente complicados para los métodos genéricos de MCMC.

3.4.1. Experimentos bidimensionales

Normal bivariada correlacionada

Este primer caso se refiere a una función de densidad normal bivariada con vector de medias y vector de varianzas igual a (-12,12) y (4,9), respectivamente. Se consideraron valores para la correlación, ρ, entre 0.2 y 0.95. En la Figura 3.6 se muestra la trayectoria de una simulación para $\rho = 0.95$. El número de simulaciones fue $5,000$, y los puntos iniciales fueron $x_0 = (0, 0)$ y $x_1 = (1, 1)$.

Los resultados de la simulación fueron buenos pues de la Figura 3.6 se observa que el algoritmo recorrió todo el dominio, y para verificarlo se calculó el error cuadrático medio (ECM) del estimador de máxima verosimilitud del vector de parámetros. Lo que se encontró coincide con la intuición, esto es, a mayor correlación se requiere de un número mayor de simulaciones para mantener la precisión de las estimaciones. Sin embargo, el ECM siempre tuvo valores razonablemente buenos considerando que el t-walk es un algoritmo genérico.

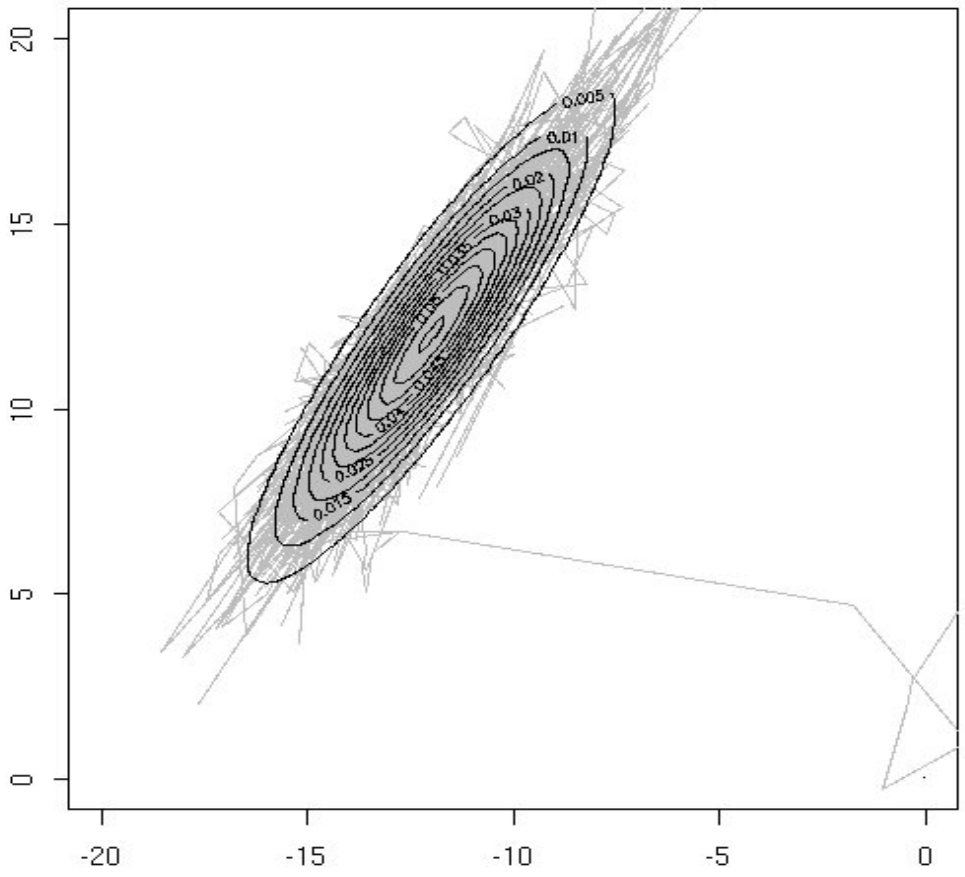

Figura 3.6: Trayectoria de una simulación hecha con el t-walk de una función de densidad normal bivariada con correlación igual a 0.95. El número de simulaciones fue $5,000$, y los puntos iniciales fueron $x_0 = (0, 0)$ y $x_1 = (1, 1)$. El cociente de aceptación fue del $52\,\%$ en este caso.

Otro dato importante sobre las simulaciones es que el valor del cociente de aceptación en promedio fue del 45 %, un valor alto.

Cambios en la escala

El caso que ahora se aborda trata sobre una función bimodal con la que se muestra la invarianza del t-walk ante cambios en la escala. La función esta definida por la expresión siguiente

$$h(x) = Kexp\left\{-\tau\left(\sum_{i=1}^{2}(x_i - m_{1i})^2\right)\left(\sum_{i=1}^{2}(x_i - m_{2i})^2\right)\right\}, \tag{3.8}$$

donde m_1 y m_2 son las modas, $\tau(> 0)$ es un parámetro de escala, y K es la constante de normalización. En la Figura 3.7 se muestran las trayectorias de cuatro simulaciones en las que la diferencia fue el valor de τ: de arriba hacia abajo y de izquierda a derecha, $\tau = 1000, 0.001, 0.01, 0.1$.

Para los diferentes valores de τ los resultados fueron igualmente buenos. El algoritmo se movió a lo largo del dominio, de una moda a la otra. El cociente de aceptación siempre fue alto, de 40 a 50 %.

Sólo para verificar que los valores simulados se distribuyen como en (3.8) se calcularon estimadores de las medias. Los errores absolutos de las estimaciones en los cuatro casos ($\tau = 0.001, 0.01, 0.1, 1000$) fueron del mismo orden, y siempre muy pequeños.

Función de Rosenbrock

La función de Rosenbrock resulta un ejemplo interesante debido a su forma (Figura 3.8): curveada, altamente correlacionada con extremos muy delgados y parte central apenas un poco más gruesa, donde se localiza la moda. En la Figura 3.8 se muestra una simulación de ella tomando 100,000 iteraciones. La tasa de aceptación siempre mayor al 40 %.

Modas contrastantes

En este caso la distribución objetivo es una mezcla de dos normales bivariadas de características contrastantes. Las diferencias fundamentales entre

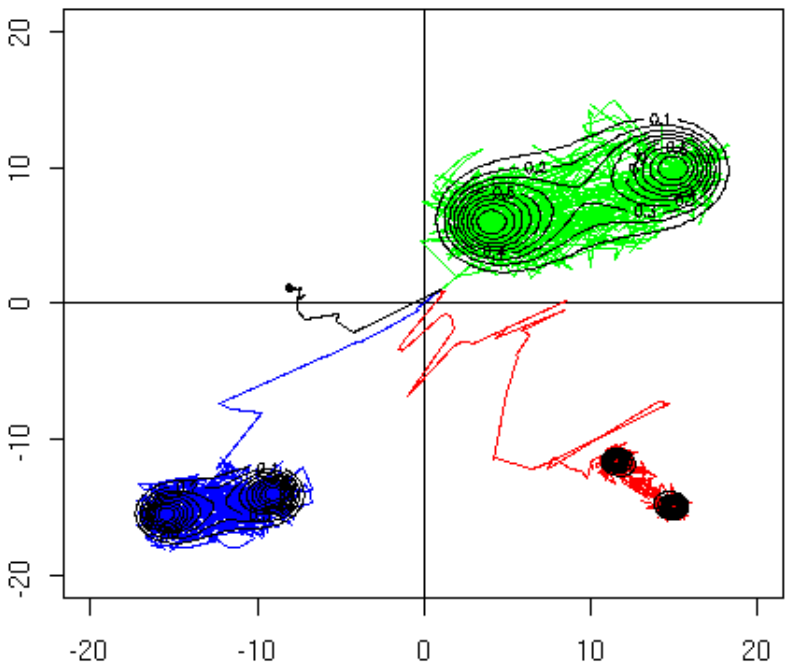

Figura 3.7: En las cuatro gráficas se muestran las trayectorias de una simulación del t-walk para la densidad objetivo en (3.8). La diferencia entre gráficas es la escala. En la gráfica superior izquierda el valor de τ es de 1000, en la que le sigue a la derecha es de 0.001, en la de abajo a la izquierda es de 0.01 y en la última τ es 0.1. Aún cuando la escala varia drásticamente la eficiencia del algoritmo es prácticamente la misma. El cociente de aceptación en todos los casos resultó estar entre el 40 y 50 %. El número de iteraciones fue $5,000$.

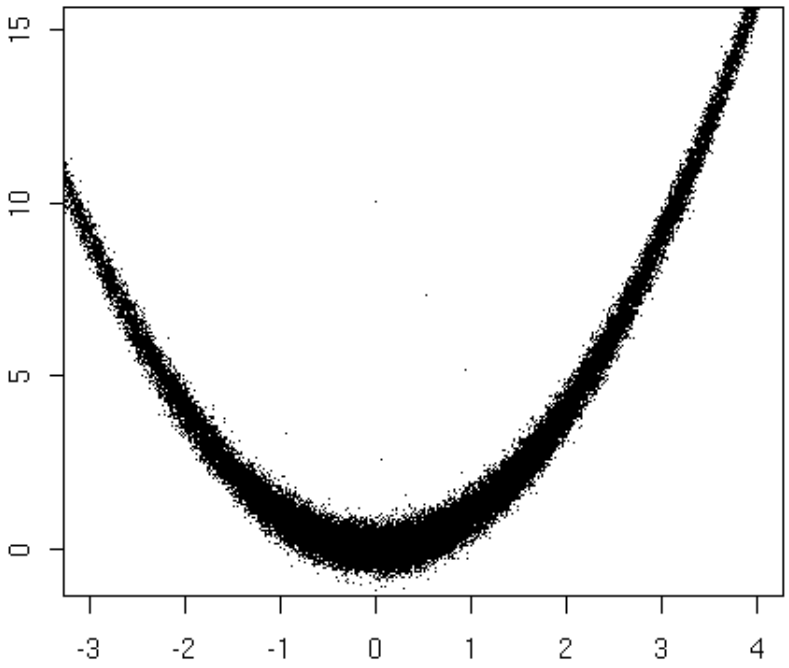

Figura 3.8: Una simulación de la conocida función de Rosenbrock. El número de iteraciones fue de $100,000$

las dos componentes de la mezcla son la estructura de correlación y la escala. Para la más pequeña la correlación es de 0.1 mientras que para la otra es de 0.8. La probabilidad de la primera componente es 0.7 y para la segunda es 1-0.7. En la Figura 3.9 se muestra el resultado de una simulación de $100,000$ iteraciones. La tasa de aceptación muy alta, alrededor de 45 %.

Como antes, se calcularon estimadores para los parámetros de la mezcla con lo que se corroboró que las desviaciones son mínimas y poco significativas.

3.4.2. Experimentos en dimensiones mayores

Fallas en bombas de agua

Este ejemplo, usado ampliamente por estadísticos bayesianos ([9]), trata sobre un modelo que describe fallas múltiples de 10 bombas de agua en una planta nuclear, los datos se muestran en la Cuadro 3.1. Para la bomba i, la tasa de falla se denota por θ_i y la longitud del tiempo de operación (en cientos de horas) se denota por t_i.

Condicionando sobre θ_i, el número de fallas X_i se asume que sigue una distribución Poisson, $X_i|\theta_i \sim \text{Poisson}(\eta_i)$, $i = 1, \ldots, 10$, donde $\eta_i = \theta_i t_i$ y X_i es independiente de X_j para $i \neq j$.

i	1	2	3	4	5	6	7	8	9	10
t_i	94.32	15.72	62.88	12.76	5.24	31.44	1.05	1.05	2.09	10.48
x_i	5	1	5	14	3	19	1	1	4	22

Cuadro 3.1: Número de fallas y longitud del periodo de observación de 10 bombas en una planta nuclear.

Las distribuciones apriori que se asumen son:

para las tasas de falla, condicionando sobre α y β, distribuciones gamma independientes,

$$\theta_i|\alpha, \beta \sim Gamma(\alpha, \beta),$$

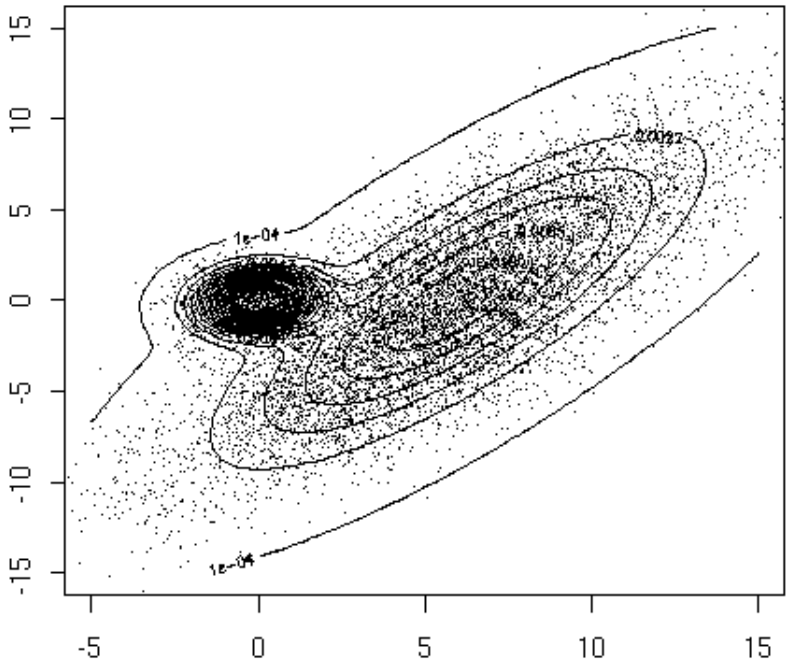

Figura 3.9: Una simulación de una mezcla de normales bivariadas: moda más baja con peso 0.7, $\mu_1 = 6$, $\sigma_1 = 4$, $\mu_2 = 0$, $\sigma_2 = 5$, $\rho = 0.8$, moda alta con peso 0.3, $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 0$, $\sigma_2 = 1$, $\rho = 0.1$. Se tomaron 100000 iteraciones con una tasa de aceptación de alrededor del 45 %.

θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}
0.059	0.102	0.089	0.116	0.604	0.609	0.893	0.881	1.584	1.992

Cuadro 3.2: Estimación de las medias posteriores en el ejemplo de las fallas en bombas de agua.

mientras que para los hiperparámetros α y β se tiene,

$$\alpha \sim Exp(\lambda_1),$$

$$\beta \sim Gamma(\lambda_2, \lambda_3),$$

donde $\lambda_1 = 1$, $\lambda_2 = 0.1$, y $\lambda_3 = 1$.

Las estimaciones de las medias posteriores, obtenidas después de $500,000$ iteraciones del t-walk, se muestran en el Cuadro 3.2.

Las estimaciones de las medias posteriores para α y β fueron 0.7 y 0.92 respectivamente. Los resultados obtenidos coinciden con los reportados en la literatura, [9].

Ensayos de dilución

Los ensayos de dilución son utilizados comunmente para estimar la concentración de algún compuesto en una muestra biológica. El procedimiento consiste en diluir una muestra varias veces y en cada dilución hacer una medición óptica y automática de un cambio de color. La razón de tener diluciones seriales porque en concentraciones muy bajas o muy altas el cambio de color es imperceptible, entonces con varias diluciones se obtienen varias medidas de diferente precisión. En esta situación un análisis de verosimilitud puede permitir combinar la información de estas medidas apropiadamente.

Los ensayos se realizan en una paleta que tiene varios contenedores. En cada contenedor se deposita una muestra o una dilución de una muestra. Hay dos tipos de muestras: las *desconocidas* y las estándar, las primeras son aquellas en las que se quiere medir la concentración del compuesto y sus diluciones, y las segundas son aquellas en las que la concentración es conocida.

En el Cuadro 3.3 se presenta un diseño de una paleta con 96 contenedores; las primeras dos columnas se refieren a dos muestras estándar cuya concentración es de 0.64 y que han sido diluidas 6 veces (valores de dilución: 1/2, 1/4, 1/8, 1/16, 1/32 y 1/64). En el último renglón de estas columnas se tiene una muestra en la que la concentración se ha reducido a cero. Las 10 columnas restantes se refieren a 10 muestras desconocidas y a 3 diluciones (valores de dilución: 1/3, 1/9 y 1/27) realizadas dos veces. Los valores de dilución para las muestras desconocidas son más espaciados que para las estándar ya que se quiere cubrir un rango más amplio de concentraciones. Este ejemplo trata sobre un estudio de concentraciones de un compuesto que causa alergia a personas con asma, descrito en [5]. El Cuadro 3.3 y el Cuadro 3.4 se refieren a este experimento. En la parte izquierda del Cuadro 3.4 se dan los resultados de las mediciones del cambio de color y para las muestras estándar y sus diluciones. Estas mediciones empiezan con valores por encima de 100 y decrecen con la dilución hasta 14 para los compuestos inertes. Los valores de la concentración (primera columna) se obtienen simplemente multiplicando la concentración inicial de 0.64 por los valores de dilución señalados arriba. En la parte derecha del mismo cuadro, los datos que se tienen son únicamente las mediciones de cambio de color para las muestras desconocidas 1 y 2 y sus diluciones.

La definición del modelo, la configuración de los parámetros y las distribuciones apriori son las mismas que en [5] y se detallan a continuación.

Los parámetros de interés son las concentraciones de las muestras desconocidas; éstas serán denotadas por $\theta_1, \ldots, \theta_{10}$. La concentración de la muestra estándar será etiquetada como θ_0. Se usará la notación x_i para la concentración en el contenedor i y y_i para la medida de cambio de color correspondiente, con $i = 1, \ldots, 96$ para este caso.

El modelo se da en partes: primero un modelo paramétrico para la intensidad esperada del color para una concentración dada, después los errores de medición y aquellos introducidos durante el proceso de preparación de las diluciones, y finalmente las distribuciones apriori para todos los parámetros.

Se usa el modelo común en estos casos para la lectura óptica dada la concentración x:

Est	Est	Desc1	Desc2	Desc3	Desc4	Desc5	Desc6	Desc7	Desc8	Desc9	Desc10
1	1	1	1	1	1	1	1	1	1	1	1
1/2	1/2	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3
1/4	1/4	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9
1/8	1/8	1/27	1/27	1/27	1/27	1/27	1/27	1/27	1/27	1/27	1/27
1/16	1/16	1	1	1	1	1	1	1	1	1	1
1/32	1/32	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3
1/64	1/64	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9
0	0	1/27	1/27	1/27	1/27	1/27	1/27	1/27	1/27	1/27	1/27

Cuadro 3.3: Diseño de un ensayo de dilución con una paleta con 96 contenedores. Las primeras dos columnas son las diluciones de dos muestras estándar (concentraciones conocidas) y las restantes son las diluciones de 10 muestras desconocidas (concentraciones desconocidas). El objetivo del experimento es estimar las concentraciones desconocidas usando las muestras estándar para calibrar las estimaciones.

$$E(y|x,\beta) = g(x,\beta) = \beta_1 + \frac{\beta_2}{1+(x/\beta_3)^{-\beta_4}}, \tag{3.9}$$

donde β_1 es la intensidad del color cuando la concentración es cero, β_2 es el crecimiento en la saturación, β_3 es la concentración en la que el gradiente de la curva cambia, y β_4 es la tasa a la que ocurre la saturación. Todos los parámetros toman valores no negativos. Este modelo ajusta a los datos muy bien como puede ser verificado en las Figuras 3.10 y 3.11. Los errores de medición se asumen normalmente distribuidos con varianzas diferentes:

$$y_i \sim N\left(g(x_i,\beta), \left(\frac{g(x_i,\beta)}{A}\right)^{2\alpha}\sigma_y^2\right),$$

Datos de las dos muestras estándar

Conc.	Dilución	y
0.64	1	101.8
0.64	1	121.4
0.32	1/2	105.2
0.32	1/2	114.1
0.16	1/4	92.7
0.16	1/4	93.3
0.08	1/8	72.4
0.08	1/8	61.1
0.04	1/16	57.6
0.04	1/16	50.0
0.02	1/32	38.5
0.02	1/32	35.1
0.01	1/64	26.6
0.01	1/64	25.0
0	0	14.7
0	0	14.2

Datos de la primer muestra desconocida

Dilución	y
1	43.6
1	38.1
1/3	19.6
1/3	19.4
1/9	15.8
1/9	15.2
1/27	13.1
1/27	14.6

Cuadro 3.4: Mediciones del cambio de color y para las dos muestras estándar y sus diluciones y para la muestra desconocida 1 y sus diluciones. Los datos de las muestras estándar son usados para estimar una curva de calibración con la que se estiman las concentraciones de las muestras desconocidas.

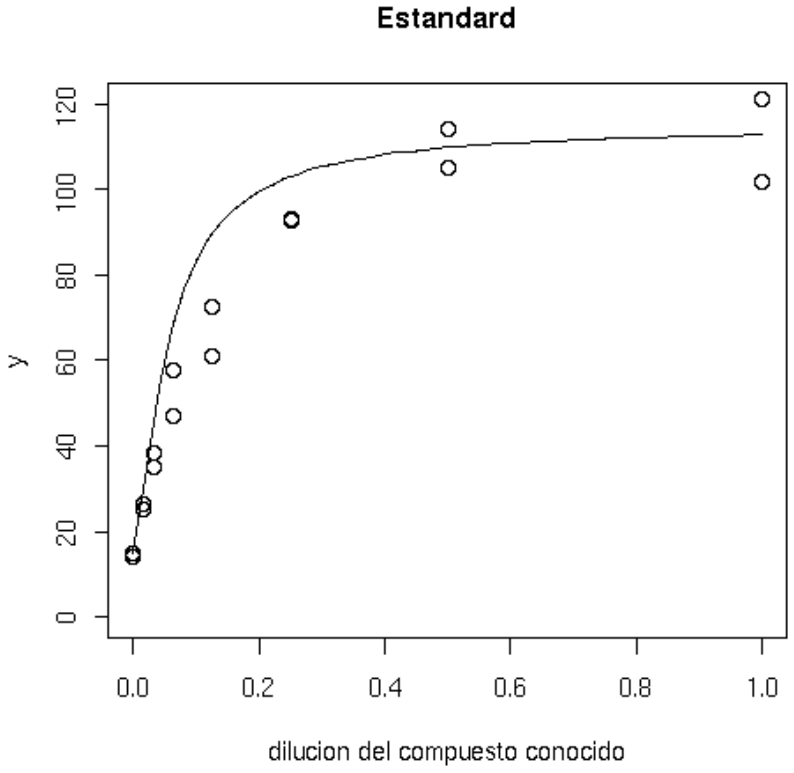

Figura 3.10: Datos de un ensayo de dilución. Curva de calibración.

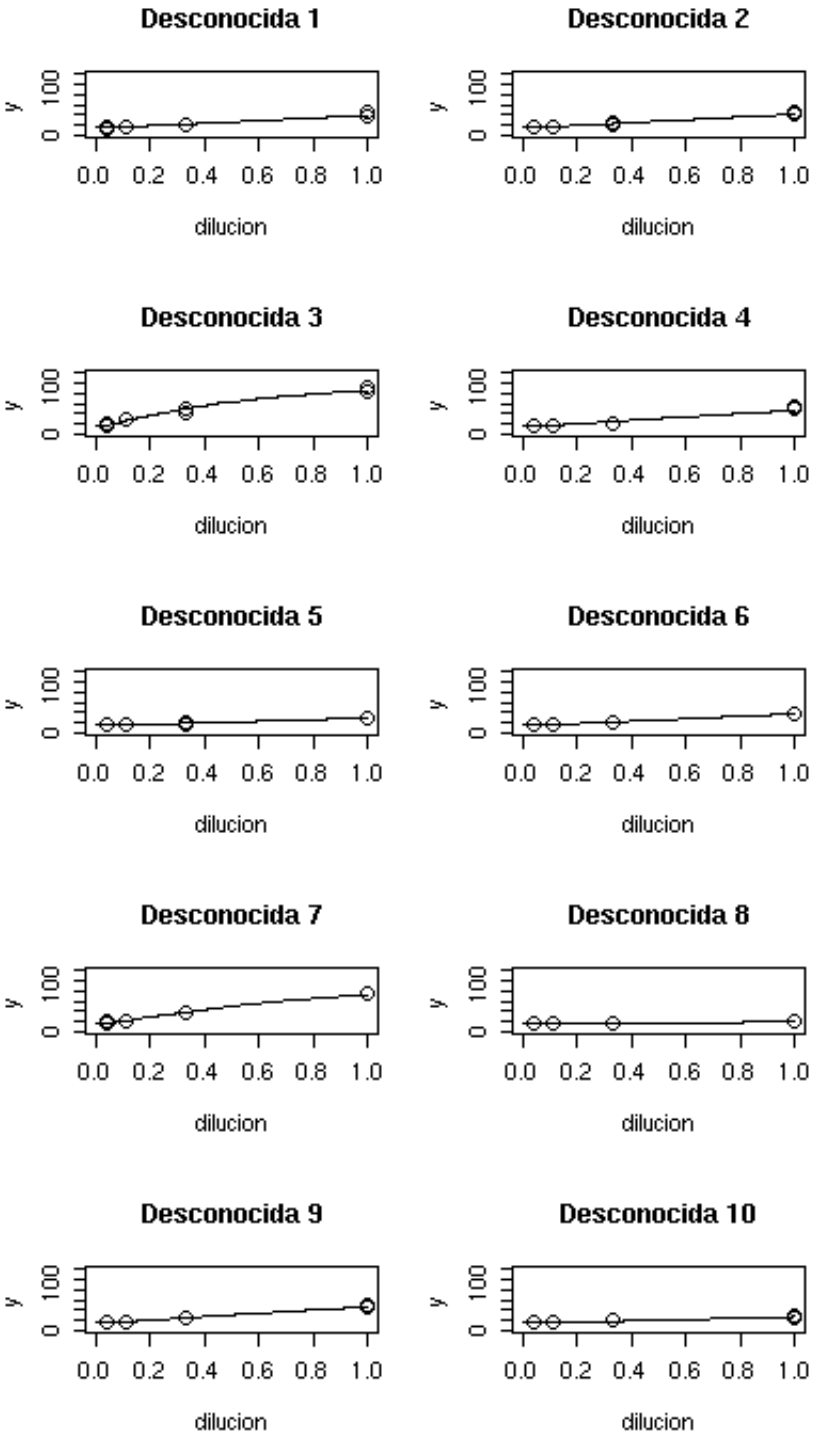

Figura 3.11: Estimación de las concentraciones para las 10 muestras desconocidas.

donde $0 \leq \alpha \leq 1$ y A es una constante cuyo valor fue fijado en 30. Para más detalles sobre estos parámetros consultar el artículo referido arriba.

En el proceso se introduce un error al diluir cada muestra estándar por primera vez. Este error es modelado como normal en escala logarítmica. Para cada muestra que es sujeta a una dilución inicial, se define θ como la concentración verdadera de la muestra sin diluir, d^{ini} como la dilución inicial (conocida), y x^{ini} como la concentración (desconocida) de la dilución inicial, con

$$log(x^{ini}) \sim N(log(d^{ini}.\theta), (\sigma^{ini})^2),$$

Para el resto de las diluciones simplemente se toma

$$x_i = d_i.x^{ini}_{j(i)},$$

donde $j(i)$ es la muestra $(0, 1, 2, \ldots, 10)$ correspondiente a la observación i, y d_i es la dilución de la observación i relativa a la dilución inicial.

Se asignan las apriori no informativas:

$$\begin{aligned} log(\beta_k) &\sim N(0, \infty), k = 1, \ldots, 4, \\ \sigma_y &\sim U(0, \infty), \\ \alpha &\sim U(0, 1), \\ log(\theta_j) &\sim N(0, \infty), j = 1, \ldots, 10. \end{aligned}$$

El párametro σ_y, la escala del error de dilución inicial, fue fijado en el valor 0.02 como se hace en [5]; este valor según se reporta fue calculado con base a análisis previos que incluyeron varias diluciones iniciales de las muestras estándar.

Las concentraciones estimadas después de $500,000$ iteraciones del t-walk se muestran en la Figura 3.12. Las estimaciones de la media posterior para los

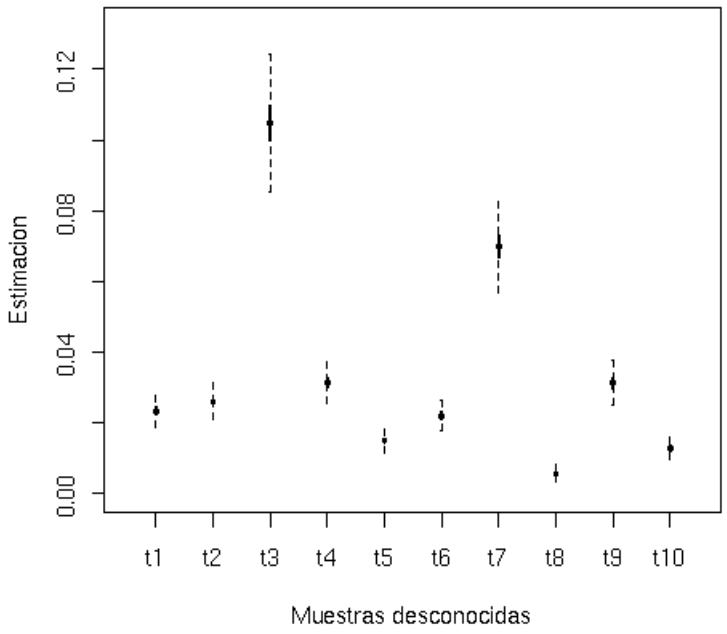

Figura 3.12: Medias y varianzas posteriores para las 10 concentraciones desconocidas.

parámetros de calibración de la curva, la varianza σ_y y α fueron: $\beta_1 = 14.64$, $\beta_2 = 102.8$, $\beta_3 = 0.06$, $\beta_4 = 1.3$, $\sigma_y = 2.74$ y $\alpha = 0.895$.

Estos valores están de acuerdo a lo publicado en [5]. Esto se confirma con las Figuras 3.10 y 3.11. En la primera se muestra la gráfica de (3.9) tomando beta como la estimación de su media posterior. Los puntos son las lecturas ópticas y para las muestras conocidas y sus diluciones. En las gráficas de la Figura 3.11 lo que se muestra es la misma función pero haciendo $x = x_j^{init} d$ para cada $j = 1, ..., 10$.

La ventaja de tener modelos que ya han sido referenciados es que se tienen puntos de comparación. En [5] se menciona que las simulaciones se hicieron usando BUGS, con 2 cadenas paralelas, con $50,000$ iteraciones cada una. Como las distribuciones en el modelo son comunes no fue rebuscada la

configuracion, lo cual no quiere decir que haya sido fácil. Aún cuando el t-walk es muy eficiente en tiempos, el BUGS fue mucho más. Esto no sorprende pues como se ha dicho ya el BUGS es tan elaborado que incluso es capaz de reconocer cuando el muestreo puede ser exacto. En cuanto a complejidad de programación del problema en el t-walk, pienso que fue del mismo orden que la programacion en BUGS, pero la diferencia la podría hacer el hecho de que C++ es un lenguaje más popular que la sintaxis y trucos del BUGS.

3.5. Implementación del t-walk

El t-walk fue implementado en el lenguaje C++ bajo el sistema operativo Linux y fue llamado *cpptwalk*. Las rutinas de números aleatorios fueron extraídas de la biblioteca gsl (GNU Scientific Library) y se adaptaron a la implementación, de modo que para ejecutar el programa no es necesario instalar ninguna biblioteca no estándar de C++. El diagrama de clases se muestra en la Figura 3.13.

La clase principal es `twalk`, sus atributos más importantes son: `Obj`, una instancia de la clase `ObjFcn`, {`x`, `xp`}, el par de puntos usados para generar las direcciones con las que se atraviesa el espacio de estados, y `K0`, `K1`, `K2`, `K3`, `K4`, los kerneles usados para sugerir propuestas de salto. Los métodos centrales de la clase son dos: el primero se llama `selectKernel` y su función es seleccionar uno de los 5 kerneles de trancisión en cada iteración de MCMC, y el segundo es `simulation`, donde se implementa el algoritmo.

Cualquier clase que represente una densidad objetivo debe ser una especialización de la clase abstracta `ObjFcn`. El único atributo de esta clase es `dim` que indica la dimensión del dominio de la densidad. Los métodos abstractos son: `inSupport(vector x)` usado para verificar cuando `x` esta en el soporte, y el método `eval(vector x)`, usado para evaluar la distribución objetivo en `x`.

La clase `kernel` también es una clase abstracta y sus clases derivadas son `K0`, `K1`, `K2`, `K3` y `K4`. Cada una de ellas tiene una definición propia de los métodos `simh` y `GU`, los cuales regresan una propuesta de salto y la probabilidad de una propuesta dada, respectivamente.

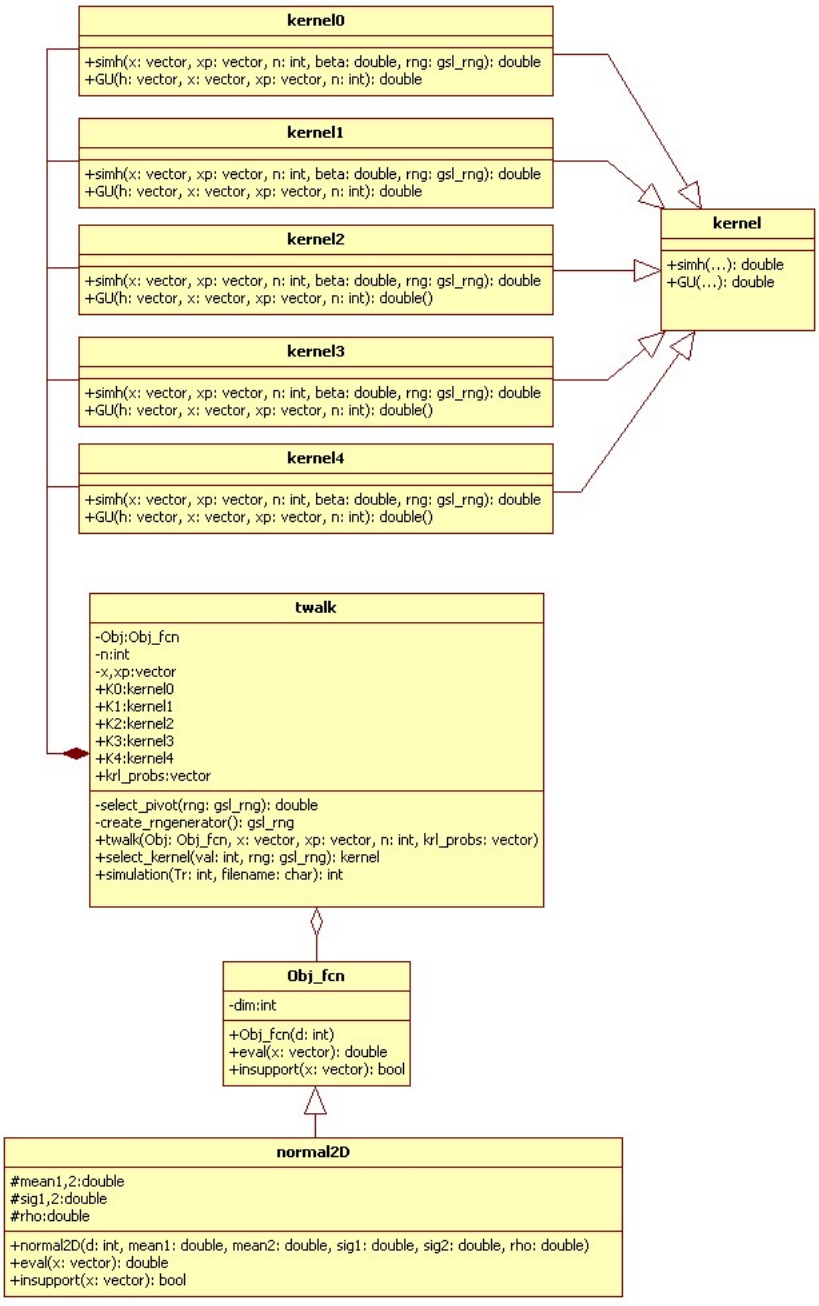

Figura 3.13: Diseño de clases del cpptwalk.

3.5.1. Ejemplo de la definición de una distribución de usuario

A continuación se incluye el código en C++ para definir a la distribución normal bivariada.

```
class normal2D: public obj_fcn {
    protected:
         double sig1; double sig2;
         double mean1; double mean2;
         double rho;
    public:
         normal2D(double, double, double, double, double);
         ~normal2D() {};
         virtual void show_descrip() const;
         virtual int insupport(double *) const;
         virtual double eval(double *) const;
};

normal2D::normal2D(double s1, double s2, double m1, double m2, double r):
                        obj_fcn(2) {
    sig1 = s1; sig2 = s2; mean1 = m1; mean2 = m2; rho = r;
}

void normal2D::show_descrip() const {
    cout << endl << "Bibariate normal density";
```

```
    cout << endl << "sig = ("  << sig1 << "," << sig2 << ")";
    cout << endl << "mean = (" << mean1 << ","  << mean2 << ")";
    cout << endl << "rho = " << rho;
}

//Evalua cuando x esta en el soporte de la densidad objetivo
int normal2D::inSupport(double *x) const {
    return 1; // The support is (-inf, inf).
}

//Evaluacion de la densidad en x
double normal2D::eval(double *x) const {
    double z1, z2, r2, mnLogConst;
    r2 = (1.0-rho*rho);
    mnLogConst = log(2.0*PI)+log(sig1)+log(sig2)+0.5*log(r2);
    z1 = (x[0]-mean1)/sig1;
    z2 = (x[1]-mean2)/sig2;
    return (mnLogConst + (0.5/r2)*(pow(z1,2.0)-2.0*rho*z1*z2+pow(z2,2.0)));
}
```

La documentación detallada se encuentra junto con el código fuente en `www.cimat.mx/~bautista`.

Capítulo 4

Discusión y conclusiones

Este capítulo pretende dar respuesta a preguntas clave con las que se resumen apropiadamente mis resultados de la experimentación con el algoritmo. Las interrogantes a las que respondo incluyen a las siguientes: ¿Cuál fue la eficiencia del algoritmo con las distribuciones objetivo seleccionadas para los experimentos?, con base en lo anterior, ¿Cuál podría ser el rendimiento del t-walk con distribuciones diferentes?, ¿Cuáles son las límitantes, y cómo se podrían resolver?, y finalmente, comparado con otros algoritmos ¿cuál es su aportación?.

Las distribuciones objetivo con las que se hicieron los experimentos fueron elegidas de forma que se pudiera mostrar el rendimiento del t-walk ante casos difíciles para un muestreador de su tipo, genérico y adaptable. La propiedad de adaptabilidad quedó probada con los ejemplos en los que las distribuciones objetivo tenian formas difíciles, como la de Rosenbrock. Con los casos bimodales se mostró que aún cuando el algoritmo se adapta a la estructura local de la distribución es capaz de moverse entre modas. Además los resultados del ejemplo 2 permitieron corroborar la propiedad de invarianza ante cambios en la escala. De las gráficas mostradas y del cálculo de estimadores se confirma que el simulador obtiene muestras efectivamente de la distribución objetivo y no de una muy parecida a ella.

La conjetura natural es que en general la eficiencia del algoritmo (número de simulaciones, tiempo de cómputo, cociente de aceptación) depende de la forma de la distribución objetivo así como de la dimensión de su dominio. Para fijar ideas, en el ejemplo de la normal bivariada correlacionada, cuando el

coeficiente de correlación se aumentó de 0.2 a 0.95 fue necesario incrementar el número de iteraciones en un factor de 10 para así mantener la precisión en los estimadores de los parámetros de la distribución. En los casos bivariados el cociente de aceptación siempre fue alto, por encima del 40 %. En el ejemplo de los ensayos de dilución este indicador fue en promedio del 12 % pero la dimensión del espacio fue 16.

Aparte de los casos estudiados en esta tesis, el t-walk ha sido aplicado exitosamente en otros problemas. En un problema de confiabilidad relativo a tiempos de fallas censurados y también en un problema de radioastronomía. Actualmente Andres Christen y Colin Fox están trabajando entre otras varias aplicaciones del t-walk, en estadística espacial en inferencia con procesos gaussianos y en procesamiento de imágenes, respectivamente.

Hasta este punto se han revisado las características del algoritmo en sí, ahora se aborda lo referente a su implementación. El t-walk está disponible como biblioteca para R, Matlab y C++, lo cual lo pone en ventaja ante otras aplicaciones similares pues de esta forma se facilita su inclusión y ejecución en y desde otros programas escritos en alguno de estos lenguajes. Considerando que el muestreador es genérico, es necesario que el usuario codifique su distribución objetivo, en lugar de seleccionarla de una lista como se hace en BUGS por ejemplo. Sin embargo pienso que programar en C++, R o Matlab es más sencillo que aprender una sintaxis y semántica poco popular en el campo científico como la de BUGS.

Como mencioné en la introducción de la tesis, para evaluar el desempeño del t-walk es importante saber qué lugar ocupa dentro de los algoritmos de MCMC. Es un algoritmo adaptable, del tipo de los AMS (ver Introducción), pero además es genérico, de modo que no resulta clara la comparación de éste con algoritmos como el muestreador de Gibbs, o el algoritmo de Metropolis de propuesta independiente, o incluso con el algoritmo snooker. No es correcto comparar algoritmos hechos a la medida con algoritmos genéricos, ni algoritmos adaptables con estándares. Aún cuando el t-walk puede aplicarse a todo tipo de distribuciones continuas, desde luego no tiene mucho sentido utilizarlo con distribuciones estándar para las que hay alternativas más eficientes, por ejemplo, una normal bivariada correlacionada, distribución para la que incluso se puede muestrear exactamente. El ejemplo usado de esta distribución es solo como ilustración del funcionamiento del t-walk con correlaciones altas. El t-walk es especialmente útil en aplicaciones en donde las distribuciones no son comunes, y en casos de modelos complicados puede ser usado como una herramienta para análisis exploratorios.

Actualmente el t-walk tiene como limitantes la dimensión del espacio, para mantenerlo eficiente, y la continuidad de la distribución, por diseño. Para extenderlo entonces se debe buscar alguna forma de condicionamiento genérico para manejar el problema de la dimensión, así como buscar alguna estrategia de muestreo eficiente en espacios discretos. En este último sentido, yo intenté algo (discretizar los kérneles que ahora se utilizan) pero las tasas de aceptación que se obtuvieron indican poca eficacia. Se requerirá un método más complicado y más *ad hoc* para generalizar el t-walk al uso también de distribuciones discretas.

Bibliografía

[1] Casella, G. y Robert, C.P. (1999). *Monte Carlo Statistical Methods*. Ed. Springer.

[2] Casella, G. (1992). *"Explaining the Gibbs sampler"*, *The American Statistician.*, **46**, 167-174.

[3] Chan, K.S. y Geyer, C.J. (1994). *Discussion of "Markov chains for exploring posterior distribution"*, *Ann. Statist.*, **22**, 1747-1758.

[4] Christen, J.A. y Fox, C. (2005). "A self-adjusting multi-scale MCMC algorithm", por aparecer.

[5] Gelman, A. y Chew, G.L. (2004). "Bayesian analysis of serial dilution assays", *Biometrics*, **60**, 407-417.

[6] Green, P.J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination", *Biometrika*, **82**, 711-732.

[7] Green, P.J. y Richardson, S. (1997). "On Bayesian analysis of mixutes with unknown number of components", *J. R. Statist. Soc. B*, **59**, 731-792.

[8] Gilks, W.R. Roberts, G.O. y George, E.I. (1994). "Adaptive direction sampling", *The Statistician*, **43**, 179–189.

[9] Gilks, W.R. y Roberts, G.O. (1996). "Strategies for improving MCMC", en: *Markov Chain Monte Carlo in Practice*, Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. Eds. Chapman and Hall: London, 89–114.

[10] Hastings, W.K. (1970). "Monte Carlo sampling methods using Markov chains and their applications", *Biometrika*, **57**, 97-109.

[11] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. y Teller, E. (1953). "Equations of state calculations by fast computing machine", *J. Chem. Phys.*, **21**, 1087-1091.

[12] Meyn, S.P. y Tweedie, R.L. (1993). *Markov chains and stochastic stability*. Ed. Springer-Verlag, New York.

[13] Nummelin, E. (1978), "A splitting technique for Harris recurrent chains", *Zeit. Warsch. Verv. Gebiete*, **43**, 309-318.

[14] Roberts, G.O. y Smith, A.F. (1994), "Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms", *Stochastic Processes and their Applications*, **49**, 207-16.

[15] R. Srinivasan. (2002). *Importance Sampling*. Ed. Springer-Verlag, New York.

Apéndice A

Principios básicos de teoría de la medida

A.1. Espacios medibles y sigmas álgebras

Un espacio medible es una pareja $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ con

- $\mathcal{X}$: un conjunto abstracto de puntos;
- $\mathcal{B}(\mathcal{X})$: una σ-álgebra de subconjuntos de $\mathcal{X}$; esto es,
  1. $\mathcal{X} \in \mathcal{B}(\mathcal{X})$;
  2. si $A \in \mathcal{B}(\mathcal{X})$ entonces $A^c \in \mathcal{B}(\mathcal{X})$
  3. si $A_k \in \mathcal{B}(\mathcal{X})$ para $k = 1, 2, \ldots$ entonces $\cup_{k=1}^{\infty} A_k \in \mathcal{B}(\mathcal{X})$.

Una σ-álgebra $\mathcal{B}$ es generada por una colección de conjuntos $\mathcal{A}$ en $\mathcal{B}$ si $\mathcal{B}$ es la σ-álgebra más pequeña que contiene a los conjuntos $\mathcal{A}$, lo cual se denota por $\mathcal{B} = \sigma(\mathcal{A})$; $\mathcal{B}$ se dice generada numerablemente si la colección de conjuntos $\mathcal{A}$ es numerable. En la línea real $\Re := (-\infty, \infty)$ la σ-álgebra de Borel $\mathcal{B}(\Re)$ puede ser generada por la colección contable de conjuntos $\mathcal{A} =$ (a,b], donde a, b son racionales.

En lo que sigue se asume que las σ-álgebras de las que se hace referencia son generadas numerablemente.

Si $(\mathcal{X}_1, \mathcal{B}(\mathcal{X}_1))$ y $(\mathcal{X}_2, \mathcal{B}(\mathcal{X}_2))$ son dos espacios medibles, entonces un mapeo $h : \mathcal{X}_1 \to \mathcal{X}_2$ es llamado una *función medible* si

$$h^{-1}\{B\} := \{x : h(x) \in \mathcal{B}\} \in \mathcal{B}(\mathcal{X}_1)$$

para todo $B \in \mathcal{B}(\mathcal{X}_2)$.

A.2. Medidas

Una medida con signo μ sobre el espacio $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ es una función de $\mathcal{B}(\mathcal{X})$ a $(-\infty, \infty)$ contablemente aditiva, es decir, que si $A_k \in \mathcal{B}(\mathcal{X})$ para $k = 1, 2, \ldots$, y $A_i \cap A_j = \Phi$, $i \neq j$, entonces se tiene

$$\mu\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

Se dice que μ es positiva si $\mu(A) \geq 0$ para todo A. La medida μ es llamada una medida de probabilidad si es positiva y $\mu(\mathcal{X}) = 1$.

Una medida positiva μ es σ-finita si existe una colección contable de conjuntos A_k tales que $\cup_{k=1}^{\infty} A_k = \mathcal{X}$ y $\mu(A_k) < \infty$ para todo k.