SIMULACIÓN Y LA MÉTODO MONTE CARLO

SERIE DE WILEY EN PROBABILIDAD Y ESTADÍSTICA
Establecido por *Walter A. Shewhart y Samuel S. Wilks*

La Serie de Wiley en Probabilidad y Estadística está bien establecida y tiene autoridad. Cubre muchos temas de interés de investigación actual tanto en estadística pura como aplicada y teoría de la probabilidad. Escritos por destacados estadísticos e instituciones, los títulos abarcan tanto los desarrollos más avanzados en el campo como los métodos clásicos.

Como reflejo de la amplia gama de investigaciones actuales en estadística, la serie abarca estadísticas aplicadas, metodológicas y teóricas, que van desde aplicaciones y nuevas técnicas posibles gracias a los avances en la práctica computarizada hasta el tratamiento riguroso de los enfoques teóricos. Esta serie proporciona una lectura esencial e invaluable para todos los estadísticos, ya sea en la academia, la industria, el gobierno o la investigación.

Puede encontrar una lista completa de los títulos de esta serie enhttp://www.wiley.com/go/wsps

SIMULACIÓN Y LA MÉTODO MONTE CARLO

Tercera edicion

Reuven Y Rubinstein
Technión

dirk p kroese
Universidad de Queensland

WILEY

Publicado por John Wiley & Sons, Inc., Hoboken, Nueva Jersey.
Publicado simultáneamente en Canadá.

Datos de catalogación en publicación de la Biblioteca del Congreso:

Nombres: Rubinstein, Reuven Y. | Kroese, Dirk
P. Título: Simulación y el método Monte Carlo.
Descripción: Tercera edición / Reuven Rubinstein, Dirk P. Kroese. | hoboken, Nueva Jersey: John Wiley & Sons, Inc., [2017] | Serie: Serie de Wiley en probabilidad y estadística | Incluye referencias bibliográficas e indice. Identificadores: LCCN 2016020742 (imprimir) | LCCN 2016023293 (libro electrónico) | ISBN 9781118632161 (tela) | ISBN 9781118632208 (pdf) | ISBN 9781118632383 (epub)
Temas: LCSH: Método Monte Carlo. | Simulación por computadora digital. | Estadística matemática. | Muestreo (Estadísticas)
Clasificación: LCC QA298 .R8 2017 (imprimir) | LCC QA298 (libro electrónico) | DDC 518/.282--dc23
Registro LC disponible enhttps://lccn.loc.gov/2016020742

Impreso en los Estados Unidos de América.

10 9 8 7 6 5 4 3 2 1

A la memoria de

Reuven Y. Rubinstein (1938-2012)

Reuven Rubinstein falleció durante la redacción de esta tercera edición. Reuven fue uno de los pioneros de la simulación de Monte Carlo y permaneció a la vanguardia de la investigación en esta área hasta el final de su vida. En 2011 recibió el más alto honor otorgado por INFORMS Simulation Society: The Lifetime Professional Achievement Award, donde sus logros se resumen de la siguiente manera:

> El profesor Rubinstein ha sido una figura fundamental en la teoría y la práctica de la simulación tal como la conocemos hoy. Su carrera refleja un alto nivel de creatividad y contribución, con voluntad de explorar nuevas áreas y una asombrosa capacidad para sugerir nuevas vías de investigación sorprendentes e influir en el trabajo posterior.

Que su contagioso entusiasmo y curiosidad vivan a través de sus libros y de las muchas personas a las que inspiró.

dirk p kroese

CONTENIDO

PREFACIO

Desde la publicación en 2008 de la segunda edición de*Simulación y el Método Monte Carlo*, se han producido cambios significativos en el campo de la simulación Monte Carlo. Esta tercera edición ofrece una descripción completa y totalmente actualizada de los principales temas de la simulación Monte Carlo.

El libro se basa en un curso de pregrado sobre métodos de Monte Carlo impartido en el Instituto de Tecnología de Israel (Technion) y la Universidad de Queensland durante los últimos cinco años. Está dirigido a una amplia audiencia de estudiantes de ingeniería, ciencias físicas y de la vida, estadística, informática, matemáticas y simplemente cualquier persona interesada en utilizar la simulación de Monte Carlo en su estudio o trabajo. Nuestro objetivo es proporcionar una introducción accesible a los métodos modernos de Monte Carlo, centrándonos en los conceptos principales, al mismo tiempo que proporcionamos una base sólida para la resolución de problemas. Por esta razón, la mayoría de las ideas se presentan y explican a través de ejemplos concretos, algoritmos y experimentos.

Aunque suponemos que el lector tiene cierta formación matemática básica, como un curso elemental de probabilidad y estadística, en el capítulo 1 repasamos los conceptos básicos de probabilidad, procesos estocásticos, teoría de la información y optimización convexa.

En una simulación estocástica típica, la aleatoriedad se introduce en los modelos de simulación a través de variables aleatorias independientes distribuidas uniformemente. Estas variables aleatorias se utilizan luego como bloques de construcción para simular sistemas estocásticos más generales. El capítulo 2 trata de la generación de tales números aleatorios, variables aleatorias y procesos estocásticos.

Muchos sistemas complejos del mundo real se pueden modelar como sistemas de eventos discretos. Los ejemplos de sistemas de eventos discretos incluyen sistemas de tráfico, fabricación flexible

sistemas informáticos, sistemas de comunicación informática, sistemas de inventario, líneas de producción, sistemas de vida útil coherente, redes PERT y redes de flujo. El comportamiento de tales sistemas se identifica a través de una secuencia de "eventos" discretos que hacen que el sistema cambie de un "estado" a otro. Discutimos cómo modelar tales sistemas en una computadora en el Capítulo 3.

El Capítulo 4 trata el análisis estadístico de los datos de salida de los modelos de simulación estáticos y dinámicos. La principal diferencia es que los primeros no evolucionan en el tiempo mientras que los segundos sí. Para los modelos dinámicos, distinguimos entre simulaciones de horizonte finito y de estado estacionario. También se analizan dos métodos populares para estimar las medidas de rendimiento en estado estacionario: las medias por lotes y los métodos regenerativos.

El capítulo 5 trata de las técnicas de reducción de la varianza en la simulación de Monte Carlo, como números aleatorios antitéticos y comunes, variables aleatorias de control, Monte Carlo condicional, muestreo estratificado y muestreo de importancia. Usando el muestreo de importancia, a menudo se puede lograr una reducción sustancial (a veces dramática) de la varianza, en particular cuando se estiman probabilidades de eventos raros. Al tratar con el muestreo de importancia, presentamos dos enfoques alternativos, llamados el*minimización de la varianza*y el*entropía cruzada*métodos. Se presta especial atención a los algoritmos de muestreo de importancia en los que los caminos se generan de forma secuencial. Se obtienen mejoras adicionales de dichos algoritmos mediante el remuestreo de rutas exitosas, lo que da lugar a*remuestreo de importancia secuencial*algoritmos Ilustramos su uso a través de un ejemplo de filtrado no lineal. Además, este capítulo contiene dos nuevos métodos basados en muestreo de importancia, llamados*el método de razón de verosimilitud transformada*y*el método de selección*para la reducción de la varianza. El primero presenta una forma simple, conveniente y unificadora de construir estimadores de muestreo de importancia eficientes, mientras que el segundo asegura la reducción de la dimensionalidad de la densidad de muestreo de importancia. Esto se logra mediante la identificación (descarte) de los parámetros más importantes (cuello de botella) que se utilizarán en la distribución de muestreo de importancia. Como resultado, la precisión del estimador de muestreo de importancia aumenta sustancialmente.

El capítulo 6 da un tratamiento conciso del genérico*Cadena de Markov Montecarlo* (MCMC) método para*aproximadamente*generar muestras a partir de una distribución arbitraria. Analizamos el algoritmo clásico de Metropolis-Hastings y el muestreador de Gibbs. En el primero, se simula una cadena de Markov tal que su distribución estacionaria coincide con la distribución objetivo, mientras que en el segundo, la cadena de Markov subyacente se construye sobre la base de una secuencia de distribuciones condicionales. También tratamos las aplicaciones de MCMC en las estadísticas bayesianas y explicamos cómo se usa MCMC para tomar muestras de la distribución de Boltzmann para los modelos de Ising y Potts, que se usan ampliamente en la mecánica estadística. Además, mostramos cómo se usa MCMC en el método de recocido simulado para encontrar el mínimo global de una función multiextremal.*rodaja*y el*salto reversible*muestreadores

El Capítulo 7 trata sobre el análisis de sensibilidad y la optimización Monte Carlo de sistemas simulados. Debido a su complejidad, la evaluación del desempeño de los sistemas de eventos discretos generalmente se estudia mediante simulación, y la simulación a menudo se asocia con la estimación de la función de desempeño con respecto a algunos parámetros controlables. El análisis de sensibilidad se ocupa de evaluar las sensibilidades (gradientes, hessianas, etc.) de la función de rendimiento con respecto a los parámetros del sistema.

Esto proporciona una guía para las decisiones operativas y para seleccionar los parámetros del sistema que optimizan las medidas de rendimiento. La optimización de Monte Carlo se ocupa de resolver programas estocásticos, es decir, problemas de optimización donde la función objetivo y algunas de las restricciones son desconocidas y deben obtenerse mediante simulación. Nos ocupamos del análisis de sensibilidad y la optimización de modelos estáticos y dinámicos. Presentamos al célebre*función de puntuación*método para el análisis de sensibilidad, y dos métodos alternativos para la optimización de Monte Carlo, el llamado*aproximación estocástica*y *contraparte estocástica*métodos. En particular, en el último método, mostramos cómo usando un solo experimento de simulación uno puede aproximarse con bastante precisión a la verdadera solución óptima desconocida del programa determinista original.

El capítulo 8 trata de la*entropía cruzada*(CE), que fue introducido por el primer autor en 1997 como un algoritmo adaptativo para la estimación de eventos raros utilizando una técnica de minimización de entropía cruzada. Pronto se dio cuenta de que las ideas subyacentes tenían un rango de aplicación mucho más amplio que solo en la simulación de eventos raros: podrían adaptarse fácilmente para abordar problemas de optimización combinatoria y multiextremal bastante generales, incluidos muchos problemas asociados con algoritmos de aprendizaje y computación neuronal. Brindamos una introducción gradual al método CE y mostramos su elegancia y versatilidad. En particular, presentamos un algoritmo CE general para la estimación de probabilidades de eventos raros y luego lo modificamos ligeramente para resolver problemas de optimización combinatoria. Discutimos las aplicaciones del método CE a varios problemas de optimización combinatoria, como el problema del corte máximo y el problema del viajante de comercio, y proporciona resultados numéricos de apoyo sobre su eficacia. Debido a su versatilidad, manejabilidad y simplicidad, el método CE tiene potencialmente una amplia gama de aplicaciones, por ejemplo, en biología computacional, alineación de secuencias de ADN, teoría de grafos y programación. En los últimos 10 años se han escrito cientos de artículos sobre la teoría y las aplicaciones de la CE. Para más detalles ver el sitio En los últimos 10 años se han escrito cientos de artículos sobre la teoría y las aplicaciones de la CE. Para más detalles ver el sitio En los últimos 10 años se han escrito cientos de artículos sobre la teoría y las aplicaciones de la CE. Para más detalles ver el sitiowww.cemethod.org,nuestro libro *El método de entropía cruzada: un enfoque unificado para la optimización combinatoria, la simulación Monte-Carlo y el aprendizaje automático*(Springer, 2004), y busque en la wikipedia bajo “método de entropía cruzada”. El capítulo concluye con una discusión del programa de optimización de mínima entropía cruzada (MinxEnt).

El capítulo 9 presenta la*terrible*método, que utiliza un plan de muestreo secuencial para descomponer un problema "difícil" en una secuencia de problemas "fáciles". El método se diseñó originalmente para la simulación de eventos raros, pero se ha convertido en un algoritmo de "partículas MCMC" muy versátil que se puede utilizar para la estimación, optimización y muestreo de eventos raros. El capítulo presenta varios algoritmos de división para modelos de simulación dinámicos y estáticos, y demuestra cómo se pueden usar para (1) estimar probabilidades de eventos raros, (2) resolver problemas de conteo difíciles, (3) encontrar soluciones a problemas de optimización desafiantes y (4) muestra de distribuciones de probabilidad complicadas. El capítulo presenta una amplia variedad de estudios de casos y experimentos numéricos, lo que demuestra la eficacia del método.

Muchos problemas combinatorios se pueden formular en términos de buscar o contar el costo total de un árbol. El capítulo 10 presenta un nuevo Monte Carlo llamado*enumeración estocástica*(SE) que es muy adecuado para resolver este tipo de problemas mediante la generación de rutas aleatorias a través del árbol de forma paralela. El algoritmo SE puede verse como un método de muestreo de importancia secuencial en un "hiper-árbol" cuyos vértices son conjuntos de vértices del árbol original. Al combinar SE con algoritmos rápidos de decisión polinomial, mostramos cómo se puede usar para contar problemas #P-completos, como

como el número de asignaciones de satisfacibilidad, el número de rutas en una red general y el número de coincidencias perfectas en un gráfico. La utilidad del método se ilustra a través de una serie de ejemplos numéricos.

El apéndice presenta una variedad de temas complementarios, incluida una breve introducción a las familias exponenciales, el filtro de Kalman de tiempo discreto y el método de la raíz cuadrada de Cholesky. También se analiza la complejidad computacional de los algoritmos aleatorios. Al final de cada capítulo se proporciona una amplia gama de ejercicios.

Además de dos capítulos completamente nuevos (Capítulos 9 y 10), esta tercera edición ofrece actualizaciones sustanciales sobre una variedad de temas. El material sobre la generación de números aleatorios se ha revisado exhaustivamente al incluir generadores recursivos múltiples combinados de última generación y el Mersenne Twister. El material sobre la generación de procesos estocásticos se ha ampliado al incluir la simulación de procesos gaussianos, movimiento browniano y procesos de difusión. El capítulo de reducción de la varianza ahora incluye una discusión del novedoso método Monte Carlo multinivel. Nuestro tratamiento de la importancia secuencial se ha modificado significativamente al enfatizar la importancia del remuestreo de importancia. Esta adición también prepara para el enfoque de partículas MCMC en el nuevo capítulo de división. El capítulo de entropía cruzada se mejora aún más al agregar nuevos conocimientos sobre la degeneración de la razón de verosimilitud, lo que lleva al algoritmo CE mejorado de un solo nivel. Se han añadido veinticinco preguntas más, junto con sus soluciones en el*manual de soluciones en línea*que acompaña a este libro. Finalmente, para facilitar su implementación, la mayoría de los algoritmos han sido (re)escritos en pseudocódigo con control de flujo.

REUVENRUBINSTEIN YDFASTIDIARKROSA

Haifa y Brisbane

julio de 2016

EXPRESIONES DE GRATITUD

Agradecemos a todos los que contribuyeron a este libro. Robert Smith y Zelda Zabinski leyeron y proporcionaron sugerencias útiles sobre el Capítulo 6. Alex Shapiro dio una descripción detallada de la complejidad de los problemas de programación estocástica (Sección A.8.4). Pierre L'Ecuyer comunicó el contraejemplo en la Sección 9.10 y Zdravko Botev sugirió el ejemplo multinivel de Monte Carlo en la Sección 5.6. Jim Spall tuvo la amabilidad de proporcionar comentarios sobre la segunda edición, que hemos incorporado en la tercera edición. Josh Chan, Tom Taimre y Zdravko Botev ayudaron a corregir el nuevo material de este libro, lo cual es muy apreciado.

Estamos especialmente agradecidos a los muchos estudiantes de pregrado y posgrado del Technion y de la Universidad de Queensland que ayudaron a hacer posible este libro y cuyas valiosas ideas y experimentos fueron extremadamente alentadores y motivadores. Qibin Duan, Morgan Grant, Robert Salomone, Rohan Shah y Erli Wang leyeron el nuevo material, resolvieron los nuevos ejercicios y brindaron excelentes comentarios. Un agradecimiento especial a Slava Vaisman por su ayuda en muchos problemas computacionales que encontramos, y por las muchas discusiones fructíferas que tuvimos sobre los métodos de enumeración estocástica y división.

Este libro fue apoyado por el Australian Research Council*Centro de Excelencia para las Fronteras Matemáticas y Estadísticas*, con número de concesión CE140100049.

RYR, DPK

CAPÍTULO 1

PRELIMINARES

1.1 INTRODUCCIÓN

El propósito de este capítulo es revisar algunos hechos básicos de la probabilidad, la teoría de la información y la optimización. En particular, las secciones 1.2 a 1.11 resumen los puntos principales de la teoría de la probabilidad. Las secciones 1.12 a 1.14 describen varios procesos estocásticos fundamentales, como los procesos de Poisson, Markov y Gauss. Los elementos de la teoría de la información se dan en la Sección 1.15, y la Sección 1.16 concluye con un resumen de la teoría de la optimización convexa.

1.2 EXPERIMENTOS AL AZAR

La noción básica en la teoría de la probabilidad es la de un *experimento aleatorio*: un experimento cuyo resultado no se puede determinar de antemano. El ejemplo más fundamental es el experimento en el que se lanza una moneda justa varias veces. Por simplicidad, suponga que la moneda se lanza tres veces. los *espacio muestral*, denotado Ω, es el conjunto de todos los resultados posibles del experimento. En este caso Ω tiene ocho resultados posibles:

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\},$$

donde, por ejemplo, HTH significa que el primer lanzamiento es cara, el segundo cruz y el tercero cara.

Los subconjuntos del espacio muestral se llaman *eventos*. Por ejemplo, el evento A que el tercer lanzamiento es cara es

$$A=\{HHH, HTH, THH, TTH\}.$$

Decimos ese evento A *ocurre* si el resultado del experimento es uno de los elementos en A. Dado que los eventos son conjuntos, podemos aplicarles las operaciones habituales de conjuntos. Por ejemplo, el evento $A \cup B$, llamó al *Unión* de A y B, es el evento que A o B o ambos ocurren, y el evento $A \cap B$, llamó al *intersección* de A y B, es el evento que A y B ambos ocurren. Una notación similar es válida para uniones e intersecciones de más de dos eventos. El evento A_C, llamó al *complementar* de A, es el evento que A no se produce. dos eventos A y B que no tienen resultados en común, es decir, su intersección está vacía, se llaman *desarticular* eventos. El paso principal es especificar la probabilidad de cada evento.

Definición 1.2.1 (Probabilidad) A *probabilidad* PAGS es una regla que asigna un número 0 - PAGS(A) - 1 para cada evento A, tal que P(Ω) =1, y tal que para cualquier sucesión $A_1, A_2, \ldots$ de eventos disjuntos

$$\text{PAGS}\left(\bigcup_i A_i\right) = \sum_i \text{PAGS}(A_i). \tag{1.1}$$

La ecuación (1.1) se conoce como la *regla de la suma* de probabilidad Afirma que si un evento puede ocurrir de varias maneras diferentes, pero no simultáneamente, la probabilidad de ese evento es simplemente la suma de las probabilidades de los eventos que lo componen.

Para el experimento del lanzamiento justo de una moneda, la probabilidad de cualquier evento se da fácilmente. Es decir, debido a que la moneda es justa, cada uno de los ocho resultados posibles es igualmente probable, de modo que PAGS($\{HHH\}$) = $\cdots$ = PAGS($\{TTT\}$) = 1/8. Desde cualquier evento A es la unión de los eventos "elementales" $\{HHH\}, \ldots, \{TTT\}$, la regla de la suma implica que

$$\text{PAGS}(A) = \frac{|A|}{|\Omega|} \tag{1.2}$$

dónde $|A|$ denota el número de resultados en A y $|\Omega|=8$. En términos más generales, si un experimento aleatorio tiene un número finito de resultados igualmente probables, la probabilidad es siempre de la forma (1.2). En ese caso, el cálculo de probabilidades se reduce a contar.

1.3 PROBABILIDAD CONDICIONAL E INDEPENDENCIA

¿Cómo cambian las probabilidades cuando sabemos que algún evento $B \subset \Omega$ ¿ha ocurrido? Dado que el resultado está en B, el evento A ocurrirá si y sólo si $A \cap B$ ocurre, y la probabilidad relativa de A ocurriendo es por lo tanto PAGS($A \cap B$)/PAGS(B). Esto conduce a la definición de la *la probabilidad condicional* de A dado B:

$$\text{PAGS}(A|B) = \frac{\text{PAGS}(A \cap B)}{\text{PAGS}(B)}. \tag{1.3}$$

Por ejemplo, supongamos que lanzamos una moneda justa tres veces. Dejar B Sea el evento de que el número total de caras sea dos. La probabilidad condicional del evento. A que el primer lanzamiento es cara, dado que B ocurre, es (2/8)/(3/8) = 2/3.

Reescribiendo (1.3) e intercambiando el papel de A y B da la relacion $P(A \cap B) = P(A)P(B|A)$. Esto se puede generalizar fácilmente a la *regla del producto* de probabilidad, que establece que para cualquier secuencia de eventos $A_1, A_2, \ldots, A_n$,

$$P(A_1 \cdots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\cdots P(A_n|A_1 \cdots A_{n-1}), \tag{1.4}$$

usando la abreviatura $A_1A_2 \cdots A_k \equiv A_1 \cap A_2 \cap \cdots \cap A_k$.

Suponer que $B_1, B_2, \ldots, B_n$ es un *dividir* de Ω. Eso es, $B_1, B_2, \ldots, B_n$ son disjuntos y su unión es Ω. Entonces, por la regla de la suma, $P(A) = \sum_{i=1}^{n} P(A \cap B_i)$ y por lo tanto, por la definición de probabilidad condicional, tenemos la *ley de probabilidad total*:

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i). \tag{1.5}$$

Combinando esto con la definición de probabilidad condicional da *regla de bayes*:

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}. \tag{1.6}$$

La independencia es de crucial importancia en probabilidad y estadística. Flojamente hablando, modela la falta de información entre eventos. dos eventos A y B se dice que son *independiente* si el conocimiento que B ha ocurrido no cambia la probabilidad de que A ocurre. Eso es, A, B independiente $\Leftrightarrow P(A|B) = P(A)$. Ya que $P(A|B) = P(A \cap B)/P(B)$, una definición alternativa de independencia es

$$A, B \text{ independiente} \Leftrightarrow P(A \cap B) = P(A)P(B).$$

Esta definición cubre el caso en que $B = \varnothing$ (conjunto vacio). Podemos extender esta definición a arbitrariamente muchos eventos.

Definición 1.3.1 (Independencia) Los eventos $A_1, A_2, \ldots,$ se dice que son *independiente* si por alguna k y cualquier elección de índices distintos $i_1, \ldots, i_k$,

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P_{i_k}P(A_{i_1})P(A_{i_2})\cdots P(A_{i_k})$$

Observación 1.3.1 En la mayoría de los casos, la independencia de los eventos es una suposición del modelo. Es decir, suponemos que existe un P tal que ciertos eventos son independientes.

EJEMPLO 1.1

Lanzamos una moneda sesgada *norte* veces. Dejar *pags* ser la probabilidad de cara (para una moneda justa *pags* = 1/2). Dejar A_i denota el evento de que el i-th tirada da cara, $i = 1, \ldots,$ *norte*. Después P debe ser tal que los eventos $A_1, \ldots, A_n$ son independientes y $P(A_i) =$ *pags* para todos i. Estas dos reglas especifican completamente P. Por ejemplo, la probabilidad de que la primera k los tiros son cara y el ultimo $n-k$ son colas es

$$\begin{aligned} P(A_1 \cdots A_k A_{C_{k+1}} \cdots A_{C_{norte}}) &= P(A_1)\cdots P(A_k)P(A_{C_{k+1}})\cdots P(A_{C_{norte}}) \\ &= pags^k(1-pags)^{n-k}. \end{aligned}$$

1.4 VARIABLES ALEATORIAS Y DISTRIBUCIONES DE PROBABILIDAD

Especificar un modelo para un experimento aleatorio a través de una descripción completa de Ω y PAGSno siempre es conveniente o necesario. En la práctica, sólo estamos interesados en ciertas observaciones (es decir, medidas numéricas) en el experimento. Los incorporamos a nuestro proceso de modelado a través de la introducción de *variables aleatorias*, generalmente indicado con letras mayúsculas de la última parte del alfabeto (p. ej., X, X_1, X_2, . . . , Y, Z).

EJEMPLO 1.2

Lanzamos una moneda sesgada *norte* veces, con *pags* la probabilidad de cara. Supongamos que solo estamos interesados en el número de cabezas, digamos X. Tenga en cuenta que X puede tomar cualquiera de los valores de $\{0,1, \ldots n\}$. los *Distribución de probabilidad* de X es dado por el *fórmula binomial*

$$\text{PAGS}(X=k) = \binom{norte}{k} pags^k(1-pags)^{n-k},\ k=0,1, \ldots, norte \tag{1.7}$$

Es decir, por el Ejemplo 1.1, cada evento elemental $\{HTH\cdots T\}$ con e(xa)ctamente k cabezas y $norte - k$ cruz tiene probabilidad $pags^k(1-pags)^{n-k}$, y hay $\binom{norte}{k}$ eventos. tal

La distribución de probabilidad de una variable aleatoria general X—identificando tales probabilidades como PAGS($X=x$), PAGS($a<X<b$), y así sucesivamente, está completamente especificado por el *función de distribución acumulativa* (cdf), definido por

$$F(x) = \text{PAGS}(X<x),\ x \in \mathrm{R}.$$

Una variable aleatoria X se dice que tiene un *discreto* distribución si, fo $\sum$ r algo finito o conjunto contable de valores $x_1, x_2, \ldots$, PAGS($X=x_i$) $>0, i=1,2, \ldots$ y $\sum_i$ PAGS($X=x_i$) = 1. La función $F(x)$ = PAGS($X=x$) se llama el *función de probabilidad* (pmf) de X—pero véase la observación 1.4.1.

EJEMPLO 1.3

Lance dos dados justos y deje *METRO* ser el mayor valor nominal que muestra. El pmf de *METRO* es dado por

metro	1	2	3	4	5	6	$\sum$
F(*metro*)	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$	1

Por ejemplo, para obtener *METRO*=3, ya sea (1,3),(2,3),(3,3),(3,2 o 3,1) tiene que ser lanzado, cada uno de los cuales sucede con probabilidad 1/36.

Una variable aleatoria X se dice que tiene un *continuo* distribución si existe una función positiva F con integral total 1, tal que para todo un, b,

$$\text{PAGS}(a<X<b) = \int_a^b F(tu)\, \mathrm{d}tu \tag{1.8}$$

La función f se llama el *función de densidad de probabilidad* (pdf) de X. Tenga en cuenta que en el caso continuo el cdf está dado por

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)\,dt,$$

y f es la derivada de F. podemos interpretar $f(x)$ como la "densidad" de probabilidad en $X=x$ en el sentido de que

$$P(x \le X \le x+h) = \int_{x}^{x+h} f(t)\,dt \approx h f(x).$$

Observación 1.4.1 (Densidad de probabilidad) Tenga en cuenta que hemos utilizado deliberadamente el *mismo* símbolo, f, tanto para pmf como para pdf. Esto se debe a que la pmf y la pdf juegan papeles muy similares y pueden, en una teoría de la probabilidad más avanzada, ser vistas como instancias particulares de la noción general de *densidad de probabilidad*. Para enfatizar este punto de vista, llamaremos f en *ambas cosas* el caso discreto y continuo la pdf o (probabilidad) densidad (función).

1.5 ALGUNAS DISTRIBUCIONES IMPORTANTES

Las tablas 1.1 y 1.2 enumeran una serie de importantes distribuciones continuas y discretas. Usaremos la notación $X \sim f$, $X \sim F$, o $X \sim$ Distrito para significar que x tiene un pdf f, un CDF F o una distribución Dist. A veces escribimos F_X en vez de F para enfatizar que el pdf se refiere a t∫la variable aleatoria X. Note que en la Tabla 1.1, Γ es la función gamma: $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\,dx$, $\alpha > 0$.

Tabla 1.1: Distribuciones continuas comúnmente utilizadas.

Nombre	Notación	$f(x)$	$X \in$	Parámetros
Uniforme	$U[\alpha, \beta]$	$\frac{1}{\beta - \alpha}$	$[\alpha, \beta]$	$\alpha < \beta$
Normal	$N(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	$\mathbb{R}$	$\sigma > 0$, $m \in \mathbb{R}$
Gama	$\text{Gama}(\alpha, \lambda)$	$\frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}$	$\mathbb{R}_+$	$\alpha, \lambda > 0$
Exponencial	$\text{Exp}(\lambda)$	$\lambda e^{-\lambda x}$	$\mathbb{R}_+$	$\lambda > 0$
Beta	$\text{Beta}(\alpha, \beta)$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$[0,1]$	$\alpha, \beta > 0$
Weibull	$\text{Weib}(\alpha, \lambda)$	$\alpha\lambda(\lambda x)^{\alpha-1} e^{-(\lambda x)^\alpha}$	$\mathbb{R}_+$	$\alpha, \lambda > 0$
Pareto	$\text{Pareto}(\alpha, \lambda)$	$\alpha\lambda(1+\lambda x)^{-(\alpha+1)}$	$\mathbb{R}_+$	$\alpha, \lambda > 0$

Tabla 1.2: Distribuciones discretas comúnmente utilizadas.

Nombre	Notación	$F(x)$	$X\in$	Parámetros
Bernoulli	Ber(p)	$p^x(1-p)^{1-x}$	$\{0,1\}$	$0<p<1$
Binomio	Compartimiento(n, p)	$\binom{n}{x}p^x(1-p)^{n-x}$	$\{0,1,\ldots n\}$	$0<p<1$, $n\in\{1,2,\ldots\}$
Discreto uniforme	DU$\{1,\ldots n\}$	$\frac{1}{n}$	$\{1,\ldots n\}$	$n\in\{1,2,\ldots\}$
Geométrico	GRAMO(p)	$p(1-p)^{x-1}$	$\{1,2,\ldots\}$	$0<p<1$
veneno	Poi(λ)	$e^{-\lambda}\frac{\lambda^x}{x!}$	n	$\lambda>0$

1.6 EXPECTATIVA

A menudo es útil considerar diferentes tipos de características numéricas de una variable aleatoria. Una de esas cantidades es la expectativa, que mide el valor medio de la distribución.

Definición 1.6.1 (Expectativa) Dejar X ser una variable aleatoria con pdf F. los *expectativa* (o valor esperado o media) de X, denotado por $E[X]$ (o algunas veces m), es definido por

$$E[X] = \begin{cases} \sum_x x f(x) & \text{caso discreto,} \\ \int_{-\infty}^{\infty} x f(x)\,dx & \text{caso continuo.} \end{cases}$$

Si X es una variable aleatoria, entonces una función de X, como X^2 o pecado(X), es de nuevo una variable aleatoria. Además, el valor esperado de una función de X es simplemente un promedio ponderado de los posibles valores que puede tomar esta función. Es decir, para cualquier función real h

$$E[h(X)] = \begin{cases} \sum_x h(x) f(x) & \text{caso discreto,} \\ \int_{-\infty}^{\infty} h(x) f(x)\,dx & \text{caso continuo.} \end{cases}$$

Otra cantidad útil es la varianza, que mide la extensión o dispersión de la distribución.

Definición 1.6.2 (Varianza) los *diferencia* de una variable aleatoria X, denotada por Var(X) (o algunas veces σ^2), es definido por

$$\mathrm{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2.$$

La raíz cuadrada de la varianza se llama *Desviación Estándar*. La tabla 1.3 enumera las expectativas y las varianzas de algunas distribuciones conocidas.

Tabla 1.3: Expectativas y varianzas para algunas distribuciones conocidas.

Dist.	MI[X]	Var(X)	Dist.	MI[X]	Var(X)
Compartimiento($norte$, pag)	$notario\ público\ notario\ público(1-pags)$		Gama(α, λ)	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
GRAMO($pags$)	$\frac{1}{pags}$	$\frac{1-pags}{pags^2}$	NORTE(μ, σ^2)	m	σ^2
Poi(λ)	λ	λ	Beta(α, β)	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(1+\alpha+\beta)}$
tu(α, β)	$\frac{\alpha+\beta}{2}$	$\frac{(\beta-\alpha)^2}{12}$	Weib(α, λ)	$\frac{\Gamma(1/\alpha)}{\alpha\lambda}$	$\frac{2\Gamma(2/\alpha)}{\alpha} - \left(\frac{\Gamma(1/\alpha)}{\alpha\lambda}\right)^2$
Exp(λ)	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$			

La media y la varianza no dan, en general, suficiente información para especificar completamente la distribución de una variable aleatoria. Sin embargo, pueden proporcionar límites útiles. Discutimos dos de tales límites. Suponer$X$solo puede tomar valores no negativos y tiene pdf$F$. Para cualquier$X>0$, podemos escribir

$$\text{MI}[X] = \int_0^X t.f.(t)\,dt + \int_X^\infty t.f.(t)\,dt - \int_X^\infty t.f.(t)\,dt$$
$$- \int_X^\infty xf(t)\,dt = X\text{PAGS}(X-X),$$

de donde se sigue la*Desigualdad de Markov*:si$X$-0, entonces para todos$X>0$,

$$\text{PAGS}(X-X) - \frac{\text{MI}[X]}{X}\ . \tag{1.9}$$

Si también conocemos la varianza de una variable aleatoria, podemos dar un límite más estrecho. Es decir, para cualquier variable aleatoria$X$con media$m$y varianzaσ^2, tenemos

$$\text{PAGS}(|X-\mu|-X) - \frac{\sigma^2}{X^2}\ . \tag{1.10}$$

Esto se llama el*Desigualdad de Chebyshev*. La demostración es la siguiente: Sea$D^2=(X-\mu)^2$; luego, por la desigualdad de Markov (1.9) y la definición de la varianza,

$$\text{PAGS}(D^2-X^2) - \frac{\sigma^2}{X^2}\ .$$

Además, tenga en cuenta que el evento$\{D^2-X^2\}$es equivalente al evento$\{|X-\mu|-X\}$, de modo que se sigue (1.10).

1.7 DISTRIBUCIONES CONJUNTAS

A menudo, un experimento aleatorio se describe mediante más de una variable aleatoria. La teoría para múltiples variables aleatorias es similar a la de una sola variable aleatoria.

Dejar $X_1, \ldots, X_{norte}$ ser variables aleatorias que describen algún experimento aleatorio. Podemos acumularlos en un *vector aleatorio* $\mathbf{X} = (X_1, \ldots, X_{norte})$. Más generalmente, una colección $\{X_t, t \in T\}$ de variables aleatorias se llama *Proceso estocástico*. El conjunto T se llama el *conjunto de parámetros* o *conjunto de índices* del proceso Puede ser discreto (por ejemplo, norteo $\{1, \ldots, 10\}$) o continuo (p. ej., $\mathbb{R}_+ = [0, \infty)$ o $[1,10]$). El conjunto de valores posibles para el proceso estocástico se denomina *espacio de Estados*.

La distribución conjunta de $X_1, \ldots, X_{norte}$ es especificado por el *cdf conjunto*

$$F(X_1, \ldots, X_{norte}) = \text{PAGS}(X_1 \le x_1, \ldots, X_{norte} \le x_{norte}).$$

los *pdf conjunto* F viene dada, en el caso discreto, por $F(X_1, \ldots, X_{norte}) = \text{PAGS}(X_1 = x_1, \ldots, X_{norte} = x_{norte})$, y en el caso continuo F es tal que

$$\text{PAGS}(\mathbf{X} \in B) = \int_B F(x_1, \ldots, x_{norte})\, dx_1 \ldots dx_{norte}$$

para cualquier región (medible) B en $\mathbb{R}_{norte}$. Los pdf marginales se pueden recuperar del pdf conjunto por integración o suma. Por ejemplo, en el caso de un vector aleatorio continuo (X, Y) con pdf conjunto F, el pdf F_X de X se encuentra como

$$F_X(x) = \int F(x, y)\, dy .$$

Suponer que X y Y son ambos discretos o ambos continuos, con pdf conjunto F, y supongamos que $F_X(x) > 0$. Entonces el *pdf condicional* de Y dado $X = x$ es dado por

$$F_{Y|X}(y \mid x) = \frac{F(x, y)}{F_X(x)} \quad \text{para todos } y .$$

El correspondiente *expectativa condicional* es (en el caso continuo)

$$\text{MI}[Y \mid X = x] = \int y\, f_{Y|X}(y \mid x)\, dy .$$

Tenga en cuenta que $\text{MI}[Y \mid X = x]$ es una función de x, decir $h(x)$. La variable aleatoria correspondiente $h(X)$ se escribe como $\text{MI}[Y \mid X]$. Se puede demostrar (ver, por ejemplo, [3]) que su expectativa es simplemente la expectativa de Y, eso es,

$$\text{E}[\text{E}[Y \mid X]] = \text{MI}[Y]. \tag{1.11}$$

Cuando la distribución condicional de Y dado X es idéntico al de Y, X y Y se dice que son independientes. Más precisamente:

Definición 1.7.1 (Variables aleatorias independientes) Las variables aleatorias $X_1, \ldots, X_{norte}$ son llamados *independientes* si para todos los eventos $\{X_i \in A_i\}$ con $A_i \subset \mathbb{R}$, $i = 1, \ldots, norte$,

$$\text{PAGS}(X_1 \in A_1, \ldots, X_{norte} \in A_{norte}) = \text{PAGS}(X_1 \in A_1) \cdots \text{PAGS}(X_{norte} \in A_{norte}).$$

Una consecuencia directa de la definición anterior de independencia es que las variables aleatorias $X_1, \ldots, X_{norte}$ con pdf conjunto F (discretas o continuas) son independientes si y sólo si

$$F(x_1, \ldots, x_{norte}) = F_{X_1}(x_1) \cdots F_{X_{norte}}(x_{norte}) \tag{1.12}$$

para todos $x_1, \ldots, x_{norte}$, dónde $\{F_{X_i}\}$ son los pdf marginales.

EJEMPLO 1.4 Secuencia de Bernoulli

Considere el experimento en el que lanzamos una moneda sesgada *norte* veces, con probabilidad *pags* de cabezas Podemos modelar este experimento de la siguiente manera. Para $i=1, \ldots, norte$, dejar X_i ser el resultado de la i-th lanzamiento: $\{X_i=1\}$ significa cara (o éxito), $\{X_i=0\}$ significa cruz (o fracaso). También, deja

$$\text{PAGS}(X_i=1) = pags = 1 - \text{PAGS}(X_i=0),\ i=1,2, \ldots, norte$$

Por último, suponga que $X_1, \ldots, X_{norte}$ son independientes La secuencia $\{X_i,\ i=1,2, \ldots\}$ se llama un *secuencia de Bernoulli* o *proceso de Bernoulli* con probabilidad de éxito *pags*. Dejar $X=X_1+\cdots+X_{norte}$ Sea el número total de éxitos en *norte* juicios (lanzamientos de ét acuña). Denotamos por B el conjunto (fa)ll vectores binarios $\mathbf{x}=(X_1, \ldots, X_{norte})$ tal que $\sum_{i=1}^{norte} X_i=k$. Tenga en cuenta que B posee $\binom{norte}{k}$ elementos. ahora tenemos

$$\begin{aligned}\text{PAGS}(X=k) &= \sum_{\mathbf{x}\in B} \text{PAGS}(X_1=X_1, \ldots, X_{norte}=X_{norte}) \\ &= \sum_{\mathbf{x}\in B} \text{PAGS}(X_1=X_1)\cdots\text{PAGS}(X_{norte}=X_{norte}) = \sum_{\mathbf{x}\in B} pags^k(1-pags)^{n-k} \\ &= \binom{norte}{k} pags^k(1-pags)^{n-k}.\end{aligned}$$

En otras palabras, $X \sim$ Compartimiento($norte$, pag). Compare esto con el Ejemplo 1.2.

Observación 1.7.1 Un *infinito* secuencia $X_1, X_2, \ldots$ de variables aleatorias se llama independiente si para cualquier elección finita de parámetros $i_1, i_2, \ldots, i_{norte}$ (ninguno de ellos es igual) las variables aleatorias $X_{i_1}, \ldots, X_{i_{norte}}$ son independientes Muchos modelos probabilísticos involucran variables aleatorias $X_1, X_2, \ldots$ que son *independientes e idénticamente distribuidas*, abreviado como *iid*. Usaremos esta abreviatura a lo largo de este libro.

Similar al caso unidimensional, el valor esperado de cualquier función de valor real h de $X_1, \ldots, X_{norte}$ es un promedio ponderado de todos los valores que puede tomar esta función. Específicamente, en el caso continuo,

$$\text{MI}[h(X_1, \ldots, X_{norte})] = \int\cdots\int h(X_1, \ldots, X_{norte}) f(X_1, \ldots, X_{norte})\, dX_1 \ldots dX_{norte}.$$

Como consecuencia directa de las definiciones de expectativa e independencia, tenemos

$$\text{MI}[a+b_1X_1+b_2X_2+\cdots+b_{norte}X_{norte}] = a+b_1m_1+\cdots+b_{norte}m_{norte} \tag{1.13}$$

para cualquier secuencia de variables aleatorias $X_1, X_2, \ldots, X_{norte}$ con expectativas $m_1, m_2, \ldots, m_{norte}$, dónde $a, b_1, b_2, \ldots, b_{norte}$ son constantes. Del mismo modo, para *independiente* variables aleatorias, tenemos

$$\text{MI}[X_1X_2\cdots X_{norte}] = m_1m_2\cdots\mu_{norte}.$$

los *covarianza* de dos variables aleatorias X y Y con expectativas $\text{MI}[X] = m_X$ y $\text{MI}[Y] = m_Y$, respectivamente, se define como

$$\text{cov}(X, Y) = \text{MI}[(X - \mu_X)(Y - \mu_Y)].$$

Esta es una medida de la cantidad de dependencia lineal entre las variables. Una versión escalada de la covarianza viene dada por la *coeficiente de correlación*,

$$(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y},$$

dónde $\sigma_X^2 = \text{Var}(X)$ y $\sigma_{Y}^2 = \text{Var}(Y)$. Se puede demostrar que la correlación coeficiente siempre se encuentra entre -1 y 1; vea el Problema 1.13.

Para facilitar la referencia, la tabla 1.4 enumera algunas propiedades importantes de la varianza y la covarianza. Las pruebas se derivan directamente de las definiciones de covarianza y varianza y las propiedades de la expectativa.

Tabla 1.4: Propiedades de varianza y covarianza.

1	$\text{Var}(X) = \text{MI}[X^2] - (\text{MI}[X])^2$
2	$\text{Var}(\textit{hacha} + b) = a^2 \text{Var}(X)$
3	$\text{cov}(X, Y) = \text{MI}[XY] - \text{MI}[X]\text{MI}[Y]$
4	$\text{cov}(X, Y) = \text{cov}(Y, X)$
5	$\text{cov}(\textit{hacha} + \textit{por } Y, Z) = a\,\text{cov}(X,Z) + b\,\text{cov}(Y, Z)$
6	$\text{cov}(X, X) = \text{Var}(X)$
7	$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{cov}(X, Y)$
8	X y Y independiente $\Rightarrow \text{cov}(X, Y) = 0$

Como consecuencia de las propiedades 2 y 7, para cualquier secuencia de *independiente* aleatorio Variables $X_1, \ldots, X_{norte}$ con variaciones $\sigma_1^2, \ldots, \sigma_{norte}^2$,

$$\text{Var}(a + b_1 X_1 + b_2 X_2 + \cdots + b_{norte} X_{norte}) = b_1^2 \sigma_1^2 + \cdots + b_{norte}^2 \sigma_{norte}^2 \tag{1.14}$$

para cualquier elección de constantes a y $b_1, \ldots, b_{norte}$.

Para vectores aleatorios, como $\mathbf{X} = (X_1, \ldots, X_{norte})$, es conveniente escribir las expectativas y covarianzas en notación vectorial.

Definición 1.7.2 (Vector de expectativa y matriz de covarianza) Para cualquier vector aleatorio $\mathbf{X}$, definimos el *expectativa de vectores* como vector de expectativas

$$\boldsymbol{m} = (m_1, \ldots, m_{norte}) = (\text{MI}[X_1], \ldots, \text{MI}[X_{norte}]).$$

los *Matriz de covarianza* Σ se define como la matriz cuya (yo, j)-ésimo elemento es

$$\text{cov}(X_i, X_j) = \text{MI}[(X_i - m_i)(X_j - m_j)].$$

Si definimos la expectativa de un vector (matriz) como el vector (matriz) de las expectativas, entonces podemos escribir

$$\boldsymbol{m} = \text{MI}[\mathbf{X}]$$

y

$$\Sigma = \mathrm{MI}[(\mathbf{X} - \boldsymbol{m})(\mathbf{X} - \boldsymbol{m})^{t}].$$

Tenga en cuenta que $\boldsymbol{m}$ y Σ asumen el mismo papel que m y σ^2 en el caso unidimensional.

Observación 1.7.2 Tenga en cuenta que cualquier matriz de covarianza Σ es *simétrico*. De hecho (vea el Problema 1.16), es *semidefinido positivo*, es decir, para cualquier vector (columna) $\mathbf{u}$,

$$\mathbf{u}^{t}\Sigma\mathbf{u} \geq 0.$$

1.8 FUNCIONES DE VARIABLES ALEATORIAS

Suponer que $X_1, \ldots, X_n$ son medidas de un experimento aleatorio. A menudo sólo estamos interesados en ciertos *funciones* de las medidas en lugar de las medidas individuales. Aquí hay unos ejemplos.

EJEMPLO 1.5

Dejar X ser una variable aleatoria continua con pdf f_X y deja $Z = aX + b$, donde $a \neq 0$. Deseamos determinar el pdf f_Z de Z. Suponer que $a > 0$. Tenemos para cualquier z

$$F_Z(z) = P(Z \leq z) = P\left(X \leq (z - b)/a\right) = F_X\left((z - b)/a\right).$$

Diferenciando esto con respecto a z da $f_Z(z) = f_X\left((z - b)/a\right)/a$. Para $a < 0$ obtenemos de manera similar $F_Z(z) = F_X\left((z - b)/a\right)/(-a)$. Así, en general,

$$f_Z(z) = \frac{1}{|a|} f_X\left(\frac{z - b}{a}\right). \tag{1.15}$$

EJEMPLO 1.6

Generalizando el ejemplo anterior, supongamos que $Z = g(X)$ para alguna función monótonamente creciente g. Para encontrar el pdf de Z de eso de X nosotros primero escribe

$$F_Z(z) = P(Z \leq z) = P\left(X \leq g^{-1}(z)\right) = F_X\left(g^{-1}(z)\right),$$

dónde g^{-1} es el inverso de g. Diferenciar con respecto a z ahora da

$$f_Z(z) = f_X\left(g^{-1}(z)\right) \frac{\mathrm{d}}{\mathrm{d}z} g^{-1}(z) = \frac{f_X\left(g^{-1}(z)\right)}{g'\left(g^{-1}(z)\right)}. \tag{1.16}$$

Para funciones monótonamente decrecientes, $\frac{\mathrm{d}}{\mathrm{d}z} g^{-1}(z)$ en la primera ecuación necesita se reemplaza por su valor negativo.

EJEMPLO 1.7 Estadísticas de pedidos

Dejar $X_1, \ldots, X_n$ ser una secuencia iid de variables aleatorias con pdf común f y CDF F. En muchas aplicaciones uno está interesado en la distribución de la

estadísticas de pedidos $X_{(1)}, X_{(2)}, \ldots, X_{(norte)}$, dónde $X_{(1)}$ es el mas pequeño de los $\{X_i, i = 1, \ldots n\}$, $X_{(2)}$ es el segundo más pequeño, y así sucesivamente. el CDF de $X_{(norte)}$ sigue desde

$$\text{PAGS}(X_{(norte)} - X) = \text{PAGS}(X_1 - X, \ldots, X_{norte} - X) = \prod_{i=1}^{norte} \text{PAGS}(X_i - X) = (F(X))_{norte}.$$

Similarmente,

$$\text{PAGS}(X_{(1)} > X) = \text{PAGS}(X_1 > X, \ldots, X_{norte} > X) = \prod_{i=1}^{norte} \text{PAGS}(X_i > X) = (1 - F(X))_{norte}.$$

Además, debido a que todos los pedidos de $X_1, \ldots, X_{norte}$ son igualmente probables, se deduce que el pdf conjunto de la muestra ordenada es, en la cuña $\{(X_1, \ldots, X_{norte}) : X_1 < X_2 < \cdots < X_{norte}\}$, simplemente *norte*! veces la densidad conjunta de la muestra desordenada y cero en el resto.

1.8.1 Transformaciones lineales

Dejar $\mathbf{X} = (X_1, \ldots, X_{norte})$ sea un vector columna en R_{norte} y A un *metro*×*norte* matriz. el mapeo $\mathbf{X} \to \mathbf{z}$, con $\mathbf{z} = A\mathbf{X}$, se llama un *transformación lineal*. Ahora considere un *aleatorio* vector $\mathbf{X} = (X_1, \ldots, X_{norte})$, y deja

$$\mathbf{Z} = A\mathbf{X}.$$

Después $\mathbf{Z}$ es un vector aleatorio en R_{metro}. En principio, si conocemos la distribución conjunta de $\mathbf{X}$, entonces podemos derivar la distribución conjunta de $\mathbf{Z}$. Veamos primero cómo se transforman el vector de expectativas y la matriz de covarianza.

Teorema 1.8.1 *Si* $\mathbf{X}$ *tiene un vector de expectativa* $\boldsymbol{m}_x$ *y matriz de covarianza* Σ_x*, entonces el vector de expectativa y la matriz de covarianza de* $\mathbf{Z} = A\mathbf{X}$ *son dados por*

$$\boldsymbol{m}_z = A\boldsymbol{m}_x \tag{1.17}$$

y

$$\Sigma_z = A\Sigma_x A \quad . \tag{1.18}$$

Prueba: Tenemos $\boldsymbol{m}_z = \text{MI}[\mathbf{Z}] = \text{MI}[A\mathbf{X}] = A\text{MI}[\mathbf{X}] = A\boldsymbol{m}_x$ y

$$\begin{aligned} \Sigma_z &= \text{MI}[(\mathbf{Z} - \boldsymbol{m}_z)(\mathbf{Z} - \boldsymbol{m}_z)^{\top}] = \text{MI}[A(\mathbf{X} - \boldsymbol{m}_x)(A(\mathbf{X} - \boldsymbol{m}_x))^{\top} A\text{MI}[(\mathbf{X} - \boldsymbol{m}_x)(\mathbf{X} - \boldsymbol{m}_x)^{\top}] \\ &=)\,]\,\textit{Automóvil club británico}\,\Sigma_x\,\textit{UNA}\,. \\ &= \end{aligned}$$

-

Suponer que A es un invertible *norte*×*norte* matriz. Si $\mathbf{X}$ tiene una densidad conjunta f_x, ¿cuál es la densidad conjunta f_z de $\mathbf{Z}$? Considere la figura 1.1. Para cualquier fijo $\mathbf{X}$, dejar $\mathbf{z} = A\mathbf{X}$. Por eso, $\mathbf{X} = A^{-1}\mathbf{z}$. Considera el *norte*-cubo dimensional $C = [z_1, z_1 + h] \times \cdots \times [z_{norte}, z_{norte} + h]$. Dejar D ser la imagen de C por debajo A^{-1}, es decir, el paralelepípedo de todos los puntos $\mathbf{X}$ tal que $A\mathbf{X} \in C$. Después,

$$\text{PAGS}(\mathbf{Z} \in C) \approx h_{norte} f_z(\mathbf{z}).$$

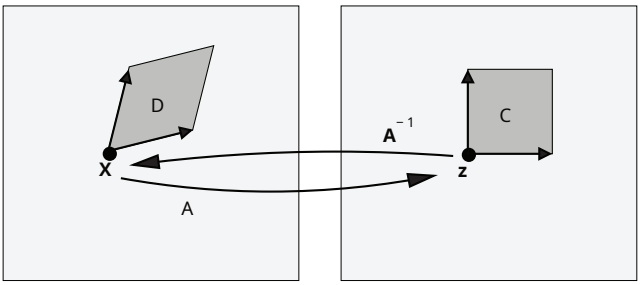


Figura 1.1: Transformación lineal.

Ahora recuerde del álgebra lineal (p. ej., [5]) que cualquier matriz B transforma linealmente un *norte*-rectángulo dimensional con volumen V en una *norte*-paralelepípedo dimensional con volumen $V\,|B|$, dónde $|B| = |\det(B)|$. De este modo,

$$\text{PAGS}(\mathbf{Z}\in C) = \text{PAGS}(\mathbf{X}\in D) \approx h_{norte}\,|U n^{-1}|\,f_{\mathbf{X}}(\mathbf{X}) = h_{norte}\,|A|^{-1} f_{\mathbf{X}}(\mathbf{X}).$$

Alquiler h vamos a 0, obtenemos

$$F_{\mathbf{Z}}(\mathbf{z}) = \frac{F_{\mathbf{X}}(A^{-1}\mathbf{z})}{|A|}, \quad \mathbf{z}\in \mathrm{R}_{norte}. \tag{1.19}$$

1.8.2 Transformaciones generales

Podemos aplicar un razonamiento similar al anterior para tratar con transformaciones generales $\mathbf{X} \to \boldsymbol{gramo}(\mathbf{X})$, escrito como

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_{norte} \end{pmatrix} \to \begin{pmatrix} gramo_1(\mathbf{X}) \\ gramo_2(\mathbf{X}) \\ \dots \\ gramo_{norte}(\mathbf{X}) \end{pmatrix}.$$

por un fijo $\mathbf{X}$, dejar $\mathbf{z} = \boldsymbol{gramo}(\mathbf{X})$. Suponer que $\boldsymbol{gramo}$ es invertible; por eso $\mathbf{X} = \boldsymbol{gramo}^{-1}(\mathbf{z})$. Cualquier infinitesimal *norte*-rectángulo dimensional en $\mathbf{X}$ con volumen V se transforma en un *norte* paralelepípedo dimensional en $\mathbf{z}$ con volumen $V\,|J_{\mathbf{X}}(\boldsymbol{gramo})|$, dónde $j_{\mathbf{X}}(\boldsymbol{gramo})$ es el *matriz de jacobia* $\mathbf{X}$ de la transformación $\boldsymbol{gramo}$, eso es,

$$J_{\mathbf{X}}(\boldsymbol{gramo}) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \dots & \frac{\partial g_1}{\partial x_{norte}} \\ \dots & \dots & \vdots \\ \frac{\partial g_{norte}}{\partial x_1} & \dots & \frac{\partial g_{norte}}{\partial x_{norte}} \end{pmatrix}.$$

Ahora considere un vector de columna aleatoria $\mathbf{Z} = \boldsymbol{gramo}(\mathbf{X})$. Dejar C ser un pequeño cubo alrededor $\mathbf{z}$ con volumen h_{norte}. Dejar D ser la imagen de C por debajo $\boldsymbol{gramo}^{-1}$. Entonces, como en el caso lineal,

$$\text{PAGS}(\mathbf{Z}\in C) \approx h_{norte} F_{\mathbf{Z}}(\mathbf{z}) \approx h_{norte}\,|J_{\mathbf{Z}}(\boldsymbol{gramo}^{-1})|\,f_{\mathbf{X}}(\mathbf{X}).$$

Por lo tanto tenemos la regla de transformación

$$F_{\mathbf{Z}}(\mathbf{z}) = F_{\mathbf{X}}(\boldsymbol{gramo}^{-1}(\mathbf{z}))\,|J_{\mathbf{Z}}(\boldsymbol{gramo}^{-1})|, \mathbf{z}\in \mathrm{R}_{norte}. \tag{1.20}$$

(Nota: $|J_{\mathbf{Z}}(\boldsymbol{gramo}^{-1})| = 1/|J_{\mathbf{X}}(\boldsymbol{gramo})|$.)

Observación 1.8.1 En la mayoría de las transformaciones de coordenadas, es $\boldsymbol{gramo}^{-1}$ que se da, es decir, una expresión para $\mathbf{X}$ como una función de $\mathbf{z}$ más bien que $\boldsymbol{gramo}$.

1.9 TRANSFORMACIONES

Muchos cálculos y manipulaciones que implican distribuciones de probabilidad se facilitan mediante el uso de transformadas. Dos ejemplos típicos son los *función generadora de probabilidad* de una variable aleatoria de valor entero positivo *norte*, definido por

$$G(z) = \mathrm{MI}[z^{norte}] = \sum_{k=0}^{\infty} z^k \mathrm{PAGS}(norte = k), \quad |z| \le 1,$$

y el *Transformada de Laplace* de una variable aleatoria positiva X definido, por $s \ge 0$, por

$$L(s) = \mathrm{MI}[\mathrm{mi}^{-sX}] = \begin{cases} \sum_x \mathrm{mi}^{-sx} F(x) & \text{di caso secreto,} \\ \int_0^{\infty} \mathrm{mi}^{-sx} F(x)\, dx & \text{caso continuo.} \end{cases}$$

Todas las transformaciones comparten una importante *propiedad de unicidad*: dos distribuciones son iguales si y solo si sus respectivas transformadas son iguales.

EJEMPLO 1.8

Dejar $METRO \sim \mathrm{Poi}(m)$; entonces su función generadora de probabilidad está dada por

$$G(z) = \sum_{k=0}^{\infty} z^k \mathrm{mi}^{-\mu} \frac{m^k}{k!} = \mathrm{mi}^{-\mu} \sum_{k=0}^{\infty} \frac{(z\mu)^k}{k!} = \mathrm{mi}^{-\mu} \mathrm{mi}^{z\mu} = \mathrm{mi}^{-\mu(1-z)}. \tag{1.21}$$

Ahora deja $norte \sim \mathrm{Poi}(v)$ independientemente de $METRO$. Entonces la función generadora de probabilidad de $METRO + norte$ es dado por

$$\mathrm{MI}[z^{METRO+norte}] = \mathrm{MI}[z^{METRO}]\,\mathrm{MI}[z^{norte}] = \mathrm{mi}^{-\mu(1-z)} \mathrm{mi}^{-v(1-z)} = \mathrm{mi}^{-(m+v)(1-z)}.$$

Así, por la propiedad de unicidad, $METRO + norte \sim \mathrm{Poi}(m+v)$.

EJEMPLO 1.9

La transformada de Laplace de $X \sim \mathrm{Gama}(\alpha, \lambda)$ es dado por

$$\begin{aligned} \mathrm{MI}[\mathrm{mi}^{-sX}] &= \int_0^{\infty} \frac{\mathrm{mi}^{-\lambda x} \lambda^{\alpha} x^{\alpha-1}}{\Gamma(\alpha)} \mathrm{mi}^{-sx}\, dx \\ &= \left(\frac{\lambda}{\lambda+s}\right)^{\alpha} \int_0^{\infty} \frac{\mathrm{mi}^{-(\lambda+s)x} (\lambda+s)^{\alpha} x^{\alpha-1}}{\Gamma(\alpha)}\, dx \\ &= \left(\frac{\lambda}{\lambda+s}\right)^{\alpha}. \end{aligned}$$

Como caso especial, la transformada de Laplace de la $\mathrm{Exp}(\lambda)$ la distribución está dada por $\lambda/(\lambda+s)$. Ahora deja $X_1, \ldots, X_{norte}$ ser iid $\mathrm{Exp}(\lambda)$ variables aleatorias. La transformada de Laplace de $S_{norte} = X_1 + \cdots + X_{norte}$ es

$$\mathrm{MI}[\mathrm{mi}^{-sS_{norte}}] = \mathrm{MI}[\mathrm{mi}^{-sX_1} \cdots \mathrm{mi}^{-sX_{norte}}] = \mathrm{MI}[\mathrm{mi}^{-sX_1}] \cdots \mathrm{MI}[\mathrm{mi}^{-sX_{norte}}] = \left(\frac{\lambda}{\lambda+s}\right)^{norte},$$

lo que demuestra que $S_{norte} \sim \mathrm{Gama}(norte, \lambda)$.

1.10 VARIABLES ALEATORIAS NORMALES CONJUNTAS

Es útil ver las variables aleatorias normalmente distribuidas como simples transformaciones de *normal estándar*—eso es, NORTE(0,1)-distribuido — variables aleatorias. En particular, deja X~NORTE(0,1). Después X tiene densidad f_X dada por

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

Ahora considere la transformación $Z = \mu + \sigma X$. Entonces, por (1.15), Z tiene densidad

$$f_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right).$$

En otras palabras, Z~NORTE(μ, σ^2). También podemos expresar esto de la siguiente manera: si Z~NORTE(μ, σ^2), después ($Z - \mu$)/σ~NORTE(0,1). Este procedimiento se llama *Estandarización*.

Ahora generalizamos esto a *norte* dimensiones. Dejar $X_1, \ldots, X_{norte}$ ser variables aleatorias normales estándar e independientes. El pdf conjunto de $\mathbf{X} = (X_1, \ldots, X_{norte})$ está dado por

$$f_X(\mathbf{x}) = (2\pi)^{-n/2} \exp\left(-\tfrac{1}{2}\mathbf{x}\cdot\mathbf{x}\right), \mathbf{x} \in \mathbb{R}^{norte}. \tag{1.22}$$

Considera el *afín* transformación (es decir, una transformación lineal más un vector constante)

$$\mathbf{Z} = \boldsymbol{\mu} + B\mathbf{X} \tag{1.23}$$

para algunos *metro*×*norte* matriz B. Nótese que, por el Teorema 1.8.1, $\mathbf{Z}$ tiene expectativa vector $\boldsymbol{\mu}$ y matriz de covarianza $\Sigma = BB'$. Cualquier vector aleatorio de la forma (1.23) se dice que tiene un *conjuntamente normales* o *normal multivariado* distribución. Nosotros escribimos $\mathbf{Z}$~NORTE($\boldsymbol{\mu}$,Σ). Suponer que B es un invertible *norte*×*norte* matriz. $\mathbf{Y} = \mathbf{Z} - \boldsymbol{\mu}$ tiene entonces, por (1.19), la densidad de

$$f_Y(\mathbf{y}) = \frac{1}{\sqrt{|B|}\sqrt{(2\pi)^{norte}}} \exp\left(-\tfrac{1}{2}(B^{-1}\mathbf{y})\cdot B^{-1}\mathbf{y}\right) = \frac{1}{\sqrt{|B|}\sqrt{(2\pi)^{norte}}} \exp\left(-\tfrac{1}{2}\mathbf{y}\cdot(B^{-1})'B^{-1}\mathbf{y}\right).$$

Tenemos $|B| = \sqrt{|\Sigma|}\sqrt{|I|}$ $(B^{-1})'B^{-1} = (BB')^{-1} = \Sigma^{-1}$, de modo que

$$f_Y(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^{norte}|\Sigma|}} \exp\left(-\tfrac{1}{2}\mathbf{y}\cdot\Sigma^{-1}\mathbf{y}\right).$$

Porque $\mathbf{Z}$ se obtiene de $\mathbf{Y}$ simplemente agregando un vector constante $\boldsymbol{\mu}$, tenemos $f_Z(\mathbf{z}) = f_Y(\mathbf{z} - \boldsymbol{\mu})$, y por lo tanto

$$f_Z(\mathbf{z}) = \frac{1}{\sqrt{(2\pi)^{norte}|\Sigma|}} \exp\left(-\tfrac{1}{2}(\mathbf{z}-\boldsymbol{\mu})\cdot\Sigma^{-1}(\mathbf{z}-\boldsymbol{\mu})\right), \mathbf{z} \in \mathbb{R}^{norte}. \tag{1.24}$$

Tenga en cuenta que esta fórmula es muy similar a la del caso unidimensional.

Por el contrario, dada una matriz de covarianza $\Sigma = (\sigma_{ij})$, existe una única matriz triangular inferior

$$B = \begin{pmatrix} b_{11} & 0 & \cdots & 0 \\ b_{21} & b_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ b_{norte1} & b_{norte2} & \cdots & b_{nn} \end{pmatrix} \tag{1.25}$$

tal que $\Sigma = BB'$. Esta matriz se puede obtener de manera eficiente a través de la *Método de la raíz cuadrada de Cholesky*; véase la Sección A.1 del Apéndice.

1.11 TEOREMAS DEL LÍMITE

Discutimos brevemente dos de los principales resultados en probabilidad: la ley de los grandes números y el teorema del límite central. Ambos están asociados con sumas de variables aleatorias independientes.

Dejar $X_1, X_2, \ldots$ ser iid variables aleatorias con expectativa m y varianza σ^2. Para cada $norte$, dejar $S_{norte} = X_1 + \cdots + X_{norte}$. Ya que $X_1, X_2, \ldots$ son iid, tenemos MI$[S_{norte}] = norte$ MI$[X_1] = n\mu$ y Var$(S_{norte}) = norte$ Var$(X_1) = n\sigma^2$.

La ley de los grandes números establece que $S_{norte}/norte$ esta cerca de m para grande $norte$. Aquí está la declaración más precisa.

Teorema 1.11.1 (Ley fuerte de los grandes números) *Si $X_1, \ldots, X_{norte}$ son iid con expectativa μ, entonces*

$$\text{PAGS}\left(\lim_{norte\to\infty} \frac{S_{norte}}{norte} = m\right) = 1.$$

El teorema del límite central describe la distribución límite de S_{norte} (o $S_{norte}/norte$), y se aplica tanto a variables aleatorias continuas como discretas. En términos generales, establece que la suma aleatoria S_{norte} tiene una distribución que es aproximadamente normal, cuando $norte$ es largo. La declaración más precisa se da a continuación.

Teorema 1.11.2 (Teorema del límite central) *Si $X_1, \ldots, X_{norte}$ son iid con expectativa μ y varianza $\sigma^2 < \infty$, entonces para todos $x \in \mathbb{R}$,*

$$\lim_{norte\to\infty} \text{PAGS}\left(\frac{S_{norte} - norte\,m}{\sigma\sqrt{norte}} \le x\right) = \Phi(x),$$

dónde Φ es la cdf de la distribución normal estándar.

En otras palabras, S_{norte} tiene una distribución que es aproximadamente normal, con expectativa $n\mu$ y varianza $n\sigma^2$. Para ver el teorema del límite central en acción, considere la figura 1.2. La parte izquierda muestra los pdfs de $S_1, \ldots, S_4$ para el caso en que el $\{X_i\}$ tener untu[0,1] distribución. La parte derecha muestra lo mismo para el Exp(1) distribución. Vemos claramente la convergencia a una curva en forma de campana, característica de la distribución normal.

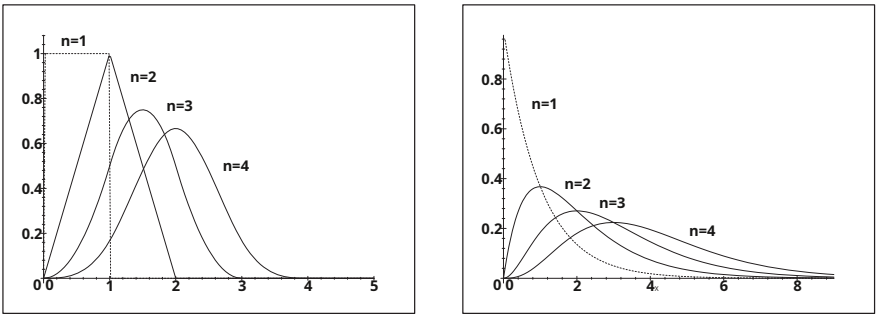


Figura 1.2: Ilustración del teorema del límite central para (izquierda) la distribución uniforme y (derecha) la distribución exponencial.

Una consecuencia directa del teorema del límite central y del hecho de que unCompartimiento(*norte*, *pag*) variable aleatoria*X*puede verse como la suma de*norte*iidBer(*pags*) variables aleatorias, $X = X_1 + \cdots + X_{norte}$, es que para grandes*norte*

$$\text{PAGS}(X - k) \approx \text{PAGS}(Y - k), \tag{1.26}$$

con$Y \sim \text{NORTE}(np, np(1 - pags))$. Como regla general, este*aproximación normal a la distribución binomial*es exacto si ambos*notario público*y*norte*$(1 - pags)$ son mayores que 5.

También hay un teorema del límite central para vectores aleatorios. La versión multidimensional es la siguiente: Sea$\mathbf{X}_1, \ldots, \mathbf{X}_{norte}$ser iid vectores aleatorios con vector de expectativa $\boldsymbol{m}$y matriz de covarianza Σ. Entonces para grandes*norte*el vector aleatorio$\mathbf{X}_1 + \cdots + \mathbf{X}_{norte}$tiene aproximadamente una distribución normal multivariante con vector de expectativa *norte*$\boldsymbol{m}$y matriz de covarianza*norte*Σ.

1.12 PROCESOS VENENOSOS

El proceso de Poisson se usa para modelar ciertos tipos de llegadas o patrones. Imagine, por ejemplo, un telescopio que pueda detectar fotones individuales de una galaxia lejana. Los fotones llegan en momentos aleatorios. $T_1, T_2, \ldots$. Dejar$norte_t$denote el número de llegadas en el intervalo de tiempo $[0, t]$, eso es,$norte_t = \text{sorber}\{k : T_k \le t\}$. Tenga en cuenta que el número de llegadas en un intervalo$yo = (un, b]$ es dado por$norte_b - norte_a$. También lo denotaremos por$norte(un, b]$. Una ruta de muestra del proceso de conteo de llegadas$\{norte_t, t \ge 0\}$se da en la figura 1.3.

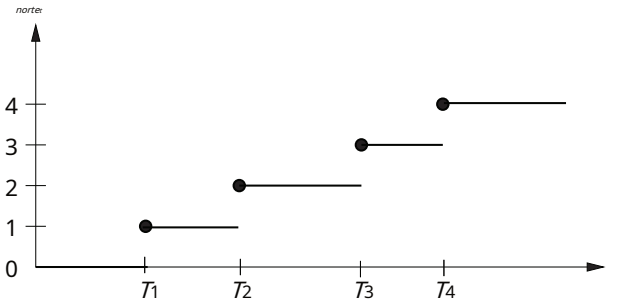


Figura 1.3: Una ruta de muestra del proceso de conteo de llegadas$\{norte_t, t \ge 0\}$.

Para este proceso de llegada en particular, se supondría que el número de llegadas en un intervalo (*un, b*) es independiente del número de llegadas en el intervalo (*discos compactos*) cuando los dos intervalos no se cortan. Tales consideraciones conducen a la siguiente definición:

Definición 1.12.1 (Proceso de Poisson) Un proceso de conteo de llegadas$norte = \{NORTE_t\}$se llama un *proceso de envenenamiento*con*tasa* $\lambda > 0$ si

(a) Los números de puntos en intervalos que no se superponen son independientes.

(b) El número de puntos en el intervalo*yo*tiene una distribución de Poisson con media $\lambda \times$ longitud(*yo*).

Combinando (a) y (b) vemos que el número de llegadas en cualquier intervalo pequeño $(t, t+h]$ es independiente del proceso de llegada hasta el momento t y tiene un Poi(λh) distribución. En particular, la probabilidad condicional de que ocurra exactamente una llegada durante el intervalo de tiempo $(t, t+h]$ es PAGS$(norte(t, t+h] = 1 \mid N_t) = mi^{-\lambda h} \lambda h \approx \lambda h$. De manera similar, la probabilidad de que no haya llegadas es de aproximadamente $1-\lambda h$ Para pequeños h. En otras palabras, λ es el *Velocidad* en el que se producen las llegadas. Nótese también que desde $norte_t \sim$ Poi(λt), el número esperado de llegadas en $[0, t]$ es λt, eso es, MI$[norte_t] = \lambda t$. En la Definición 1.12.1 *norte* se ve como una medida de conteo aleatorio, donde *norte*(yo) cuenta el número aleatorio de llegadas en el conjunto yo.

Una relación importante entre $norte_t$ y T_{norte} es

$$\{NORTE_t \text{-} norte\} = \{T_{norte} \text{-} t\}. \tag{1.27}$$

En otras palabras, el número de llegadas en $[0, t]$ Por lo menos $norte$ si y solo si el $norte$-la llegada se produce a la hora o antes t. Como consecuencia, tenemos

$$\begin{aligned} \text{PAGS}(T_{norte} \text{-} t) &= \text{PAGS}(norte_t \text{-} norte) = 1 - \sum_{k=0}^{n-1} \text{PAGS}(norte_t = k) \\ &= 1 - \sum_{k=0}^{n-1} mi^{-\lambda t} \frac{(\lambda t)^k}{k!}, \end{aligned}$$

que corresponde exactamente a la cdf de la Gama$(norte, \lambda)$ distribución; vea el problema 1.17. De este modo

$$T_{norte} \sim \text{Gama}(norte, \lambda). \tag{1.28}$$

Por lo tanto, cada T_{norte} tiene la misma distribución que la suma de $norte$ independiente Exp(λ)-variables aleatorias distribuidas. Esto se corresponde con la segunda caracterización importante de un proceso de Poisson:

> *Un proceso de conteo de llegadas $\{norte_t\}$ es un proceso de Poisson con tasa λ si y solo si los tiempos entre llegadas $A_1 = T_1$, $A_2 = T_2 - T_1$, . . . son independientes y* Exp(λ) *-Variables aleatorias distribuidas.*

Los procesos de Poisson y Bernoulli son similares, y se puede aprender mucho sobre los procesos de Poisson a través de lo siguiente *Aproximación de Bernoulli*. Dejar $norte = \{NORTE_t\}$ sea un proceso de Poisson con parámetro λ. Dividimos el eje de tiempo en pequeños intervalos de tiempo $[0, h), [h, 2h), \ldots$ y cuente cuántas llegadas ocurren en cada intervalo. Tenga en cuenta que el número de llegadas en cualquier pequeño intervalo de tiempo de longitud h es, con alta probabilidad, 1 (con probabilidad λh mi$^{-\lambda h} \approx \lambda h$) o 0 (con probabilidad $e^{-\lambda h} \approx 1 - \lambda h$). A continuación, defina $X = \{X_{norte}\}$ ser un proceso de Bernoulli con parámetro de éxito $pags = \lambda h$. Poner $Y_0 = 0$ y deja $Y_{norte} = X_1 + \cdots + X_{norte}$ Sea el número total de éxitos en $norte$ juicios $Y = \{Y_{norte}\}$ se llama el *Aproximación de Bernoulli a norte*. podemos ver $norte$ como un caso límite de Y a medida que disminuimos h.

Como ejemplo de la utilidad de esta interpretación, ahora demostramos que la propiedad de Poisson (b) en la Definición 1.12.1 se sigue básicamente de la *independencia* supuesto (a). Para pequeños h, $norte_t$ debe tener aproximadamente la misma distribución

como Y_{norte}, dónde $norte$ es la parte entera de t/h (nosotros escribimos $norte = t/h$). Por eso,

$$\begin{aligned} \text{PAGS}(norte_t = k) &\approx \text{PAGS}(Y_{norte} = k) \\ &= \binom{norte}{k} (\lambda h)^k (1 - (\lambda h))^{n-k} \\ &\approx \binom{norte}{k} (\lambda t/n)^k (1 - (\lambda t/n))^{n-k} \\ &\approx \text{mi}^{-\lambda t} \frac{(\lambda t)^k}{k!} . \end{aligned} \tag{1.29}$$

La ecuación (1.29) se deriva de la aproximación de Poisson a la distribución binomial; vea el problema 1.22.

Otra aplicación de la aproximación de Bernoulli es la siguiente. Para el proceso de Bernoulli, dado que el número total de éxitos es k, las posiciones de los k los éxitos se distribuyen uniformemente en los puntos $1, \ldots, norte$. La propiedad correspondiente para el proceso de Poisson $norte$ es eso dado $norte_t = norte$, los tiempos de llegada $T_1, \ldots, T_{norte}$ se distribuyen según las estadísticas del pedido $X_{(1)}, \ldots, X_{(norte)}$, dónde $X_1, \ldots, X_{norte}$ son iidtu$[0, t]$.

1.13 PROCESOS DE MARKOV

Los procesos de Markov son procesos estocásticos cuyos futuros son condicionalmente independientes de sus pasados dados sus valores presentes. Más formalmente, un proceso estocástico $\{X_t, t \in T\}$, con $T \subseteq$ R, se llama un *proceso de Markov* si por cada $s > 0$ y t,

$$(X_{t+s} / X_{tu}, tu - t) \sim (X_{t+s} / X_t). \tag{1.30}$$

En otras palabras, la distribución condicional de la variable futura X_{t+s}, dado todo el pasado del proceso $\{X_{tu}, tu - t\}$, es lo mismo que la distribución condicional de X_{t+s} dado solo el presente X_t. Es decir, para predecir estados futuros, solo necesitamos conocer el presente. La propiedad (1.30) se llama *Propiedad de Markov*.

Dependiendo del conjunto de índices T y espacio de estado mi (el conjunto de todos los valores $\{X_t\}$ puede tomar), los procesos de Markov vienen en muchas formas diferentes. Un proceso de Markov con un conjunto de índices discretos se llama *cadena de Markov*. Un proceso de Markov con un espacio de estado discreto y un conjunto de índices continuo (como RoR$_+$) se llama *Proceso de salto de Markov*.

1.13.1 Cadenas de Markov

Considere una cadena de Markov $X = \{X_t, t \in \text{norte}\}$ con un espacio de estado discreto (es decir, contable) mi. En este caso la propiedad de Markov (1.30) es

$$\text{PAGS}(X_{t+1} = X_{t+1} / X_0 = X_0, \ldots, X_t = X_t) = \text{PAGS}(X_{t+1} = X_{t+1} / X_t = X_t) \tag{1.31}$$

para todos $X_0, \ldots, X_{t+1}, \in mi$ y $t \in$ NORTE. Nos restringimos a las cadenas de Markov para las cuales las probabilidades condicionales

$$\text{PAGS}(X_{t+1} = j / X_t = i), \; yo, j \in mi \tag{1.32}$$

son independientes del tiempo t. Tales cadenas se llaman *homogéneo en el tiempo*. Las probabilidades en (1.32) se llaman *(un paso) probabilidades de transición* de X. La distribución de X_0 se llama el *distribución inicial* de la cadena de Markov. Las probabilidades de transición de un paso y la distribución inicial especifican completamente la distribución de X. A saber, tenemos por la regla del producto (1.4) y la propiedad de Markov (1.30),

$$\begin{aligned}\text{PAGS}(X_0=x_0, \ldots, X_t=x_t) &= \text{PAGS}(X_0=x_0)\text{PAGS}(X_1=x_1 \mid X_0=x_0)\cdots\text{PAGS}(X_t=x_t \mid X_0=x_0, \ldots X_{t-1}=x_{t-1}) \\ &= \text{PAGS}(X_0=x_0)\text{PAGS}(X_1=x_1 \mid X_0=x_0)\cdots\text{PAGS}(X_t=x_t \mid X_{t-1}=x_{t-1}).\end{aligned}$$

Ya que mi es contable, podemos organizar las probabilidades de transición de un paso en una matriz. Esta matriz se llama (un paso) *matriz de transición* de X. Normalmente denotamos por $PAGS$. por ejemplo, cuando $mi = \{0,1,2, \ldots\}$, la matriz de transición $PAGS$ tiene la forma

$$PAGS = \begin{pmatrix} pags_{00} & pags_{01} & pags_{02} & \ldots \\ pags_{10} & pags_{11} & pags_{12} & \ldots \\ pags_{20} & pags_{21} & pags_{22} & \ldots \\ \vdots & & \vdots & \end{pmatrix}.$$

Tenga en cuenta que los elementos en cada fila son positivos y suman la unidad.

Otra forma conveniente de describir una cadena de Markov X es a través de su *gráfico de transición*. Los estados se indican mediante los nodos del grafo y un signo estrictamente positivo (>0) probabilidad de transición $pags_{yo}$ del estado i a j se indica con una flecha desde i a j con peso $pags_{yo}$.

■ EJEMPLO 1.10 Paseo aleatorio sobre los enteros

Dejar $pags$ ser un número entre 0 y 1. La cadena de Markov X con espacio de estado Z y matriz de transición $PAGS$ definido por

$$PAGS(yo, yo+1) = pag \quad pag(yo, yo-1) = q = 1 - pags, \qquad \text{para todos } i \in Z$$

se llama un *paseo aleatorio en los enteros*. Dejar X empezar en 0; de este modo, PAGS(X_0=0) = 1. El gráfico de transición correspondiente se muestra en la Figura 1.4. Comenzando en 0, la cadena toma pasos subsiguientes hacia la derecha con probabilidad $pags$ y a la izquierda con probabilidad q.

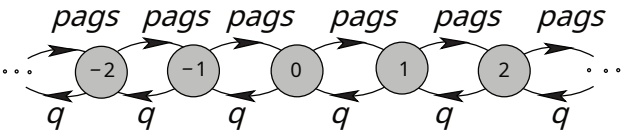


Figura 1.4: Gráfico de transición para un paseo aleatorio enz

Mostramos a continuación cómo calcular la probabilidad de que, partiendo del estado i en algún tiempo (discreto) t, estamos en j en tiempo (discreto) $t+s$, es decir, la probabilidad PAGS($X_{t+s}=j \mid X_t=i$). Para mayor claridad, supongamos que $mi = \{1,2, \ldots, m\}$ para algunos fijos $metro$, de modo que $PAGS$ es un $metro \times metro$ matriz. Para $t=0,1,2, \ldots$, define el vector fila

$$\boldsymbol{\pi}_{(t)} = (\text{PAGS}(X_t=1), \ldots, \text{PAGS}(X_t=metro)).$$

Nosotros llamamos $\boldsymbol{\pi}_{(t)}$ la *vectores de distribucion*, o simplemente el *distribución*, de X en el momento t y $\boldsymbol{\pi}_{(0)}$ la *distribución inicial* de X. El siguiente resultado muestra que la t Las probabilidades de paso se pueden encontrar simplemente mediante la multiplicación de matrices.

Teorema 1.13.1 *La distribución de X en el momento t es dado por*

$$\boldsymbol{\pi}_{(t)} = \boldsymbol{\pi}_{(0)} PAGS^t \tag{1.33}$$

para todos t=0,1, ... (Aquí $PAGS^0$ denota la matriz identidad.)

Prueba: La prueba es por inducción. La igualdad (1.33) se cumple para t=0 por definición. Supongamos que esta igualdad es cierta para algunos t=0,1, ... Tenemos

$$\text{PAGS}(X_{t+1}=k) = \sum_{i=1}^{\text{metro}} \text{PAGS}(X_{t+1}=k \mid X_t=i)\text{PAGS}(X_t=i).$$

Pero se supone que (1.33) es cierta para t, asi que PAGS(X_t=i) es el i-ésimo elemento de $\boldsymbol{\pi}_{(0)} PAGS^t$. Es más, PAGS($X_{t+1}$=$k$ | X_t=i) es el (*yo k*)-ésimo elemento de *PAGS*. Por lo tanto, por cada k,

$$\sum_{i=1}^{\text{metro}} \text{PAGS}(X_{t+1}=k \mid X_t=i)\text{PAGS}(X_t=i) = \sum_{i=1}^{\text{metro}} PAGS(yo\ k)(\boldsymbol{\pi}_{(0)} PAGS^t)(i),$$

que es solo el k-ésimo elemento de $\boldsymbol{\pi}_{(0)} PAGS^{t+1}$. Esto completa el paso de inducción y, por lo tanto, se prueba el teorema. -

Tomando $\boldsymbol{\pi}_{(0)}$ como el i-ésimo vector unitario, $\mathbf{mi}_i$, la t Las probabilidades de transición de paso se pueden encontrar como PAGS(X_t=j | X_0=i) = ($\mathbf{mi}_i PAGS^t$)(j) = $PAGS^t$(*yo, j*), Cuál es el (*yo, j*)-ésimo elemento de matriz $PAGS^t$. Así, para encontrar la t probabilidades de transición de paso, solo tenemos que calcular el t -ésima potencia de *PAGS*.

1.13.2 Clasificación de Estados

Dejar X ser una cadena de Markov con espacio de estado discreto mi y matriz de transición *PAGS*. Podemos caracterizar las relaciones entre estados de la siguiente manera: Si los estados i y j son tales que $PAGS^t$(*yo, j*)>0 para algunos t-0, decimos que i *lleva a* j y escribe $i \to j$. Nosotros decimos eso i y j *comunicar* si $i \to j$ y $j \to i$, y escribe $i \leftrightarrow j$. Usando la relación "$\leftrightarrow$", podemos dividir mi dentro *clases de equivalencia* tal que todos los estados en una clase de equivalencia se comunican entre sí pero no con ningún estado fuera de esa clase. Si solo hay una clase equivalente, es decir, si $\sum_{j \in A} PAGS(yo, j) = 1$ para todos $i \in A$, entonces se dice que la cadena de Markov ser *irreducible*. Si un conjunto de estados A es tal que después A se llama un *cerrado* establecer. Un estado i se llama un *absorbente* indicar si $\{i\}$ está cerrado. Por ejemplo, en el gráfico de transición representado en la Figura 1.5, las clases de equivalencia son $\{1,2\}$, $\{3\}$, y $\{4,5\}$. Clase $\{1,2\}$ es el único conjunto cerrado: la cadena de Markov no puede escapar de él. Si faltara el estado 1, el estado 2 sería absorbente. En el ejemplo 1.10, la cadena de Markov es irreducible ya que todos los estados se comunican.

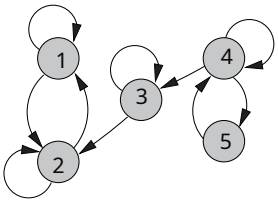


Figura 1.5: Un gráfico de transición con tres clases de equivalencia.

Otra clasificación de estados se obtiene observando el sistema desde un punto de vista local. En particular, deja T_i indicar el momento en que la cadena visita por primera vez el estado j, o primero vuelve a j si empezó ahí, y deja $norte_{ij}$ denote el número total de visitas a j desde el tiempo 0 en adelante. Nosotros escribimos $\text{PAGS}_i(A)$ por $\text{PAGS}(A \mid X_0 = i)$ para cualquier evento A. Denotamos el operador de expectativa correspondiente por mi_i. Estado j se llama un *recurrente* indicar si $\text{PAGS}_j(T < \infty) = 1$; de lo contrario, j se llama *transitorio*. Un estado recurrente se llama *positivo recurrente* si $mi_j[T] < \infty$; de lo contrario, se llama *nulo recurrente*. Finalmente, se dice que un estado es *periódico, con periodo* δ, si d - 2 es el entero más grande para el cual $\text{PAGS}_j(T = n\delta$ para algunos $norte$-1) = 1; de lo contrario, se llama *aperiódico*. Por ejemplo, en la figura 1.5, los estados 1 y 2 son recurrentes y los otros estados son transitorios. Todos estos estados son aperiódicos. Los estados de la caminata aleatoria del ejemplo 1.10 son periódicos con período 2.

Se puede demostrar que la recurrencia y la transitoriedad son propiedades de clase. En particular, si $i \leftrightarrow j$, después i recurrente (transitorio) $\Leftrightarrow j$ recurrente (transitorio). Por lo tanto, en una cadena de Markov irreducible, que un estado sea recurrente implica que todos los demás estados también son recurrentes. Y si un estado es transitorio, también lo son todos los demás.

1.13.3 Comportamiento limitante

El comportamiento límite o de "estado estacionario" de las cadenas de Markov como $t \to \infty$ es de considerable interés e importancia, y este tipo de comportamiento es a menudo más simple de describir y analizar que el comportamiento "transitorio" de la cadena para fijos t. Se puede demostrar (ver, por ejemplo, [3]) que en una cadena de Markov irreducible y aperiódica con matriz de transición $PAGS$ la t Las probabilidades de paso convergen a una constante que no depende del estado inicial. Más específicamente,

$$\lim_{t \to \infty} PAGS_{ij}(t) = \pi_j \tag{1.34}$$

por algún número $0 \le \pi_j \le 1$. Además, $\pi_j > 0$ si j es recurrente positivo y $\pi_j = 0$ de lo contrario. La razón intuitiva detrás de este resultado es que el proceso "olvida" dónde estaba inicialmente si se prolonga lo suficiente. Esto es cierto tanto para las cadenas de Markov finitas como para las numerables infinitas. Los números $\{\pi_j,\ j \in mi\}$ $\sum$ formar el *limitando distribución* de la cadena de Markov, siempre que $\pi_j > 0$ y $_j \pi_j = 1$. Nota que estas condiciones no siempre se cumplen: claramente no se cumplen si la cadena de Markov es transitoria, y pueden no cumplirse si la cadena de Markov es recurrente (es decir, cuando los estados son nulos-recurrentes). El siguiente teorema da un método para obtener distribuciones límite. Aquí suponemos por simplicidad que $mi = \{0,1,2, \ldots\}$. La distribución límite se identifica con el vector fila $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots)$.

Teorema 1.13.2 *Para una cadena de Markov irreducible y aperiódica con matriz de transición PAGS, si la distribución límite* $\boldsymbol{\pi}$ *existe, entonces está determinada únicamente por la solución de*

$$\boldsymbol{\pi} = \boldsymbol{\pi} PAGS, \tag{1.35}$$

con $\pi_j > 0$ *y* $\sum_j \pi_j = 1$. *Por el contrario, si existe un vector fila positivo* $\boldsymbol{\pi}$ *satisfactorio* (1.35) *y sumando 1, entonces* $\boldsymbol{\pi}$ *es la distribución límite de la cadena de Markov. Además, en ese caso,* $\pi_j > 0$ *para todos* j *y todos los estados son recurrentes positivos.*

Prueba. (Sketch) Para el caso en que mi es finito, el resultado es simplemente una consecuencia de (1.33). Es decir, con $\boldsymbol{\pi}(0)$ siendo el i-ésimo vector unitario, tenemos

$$PAGS_{t+1}(yo, j) = \boldsymbol{\pi}(0) PAGS^t PAGS(j) = \sum_{k \in mi} PAGS^t(yo\ k) PAGS(kj).$$

Alquiler $t \to \infty$, obtenemos (1.35) de (1.34), siempre que podamos cambiar el orden del límite y el $\sum$ suma. Para mostrar la unicidad, suponga que otro vector $\mathbf{y}$, con $y_j > 0$ y $\sum_j y_j = 1$, satisface $\mathbf{y} = \mathbf{y} PAGS$. Entonces es fácil mostrar por inducción que $\mathbf{y} = \mathbf{y} PAGS^t$, para cada t. Por lo tanto, dejar $t \to \infty$, obtenemos para cada j

$$y_j = \sum_i y_i \pi_j = \pi_j,$$

desde el $\{y_j\}$ suma a la unidad. Omitimos la demostración del enunciado inverso.-

■ EJEMPLO 1.11 Paseo aleatorio sobre los enteros positivos

Esta es una caminata aleatoria ligeramente diferente a la del Ejemplo 1.10. Dejar X ser un paseo al azar en $mi = \{0,1,2, \ldots\}$ con matriz de transición

$$PAGS = \begin{pmatrix} q & pags & 0 & \ldots \\ q & 0 & pags & 0 & \ldots \\ 0 & q & 0 & pags & 0 & \ldots \\ \ldots & \ddots & \ldots & \ddots & \ldots & \ldots \end{pmatrix},$$

donde $0 < pag < 1$ y $q = 1 - pags$. X_t podría representar, por ejemplo, el número de clientes que están esperando en una cola a la vez t.

Todos los estados pueden alcanzarse entre sí, por lo que la cadena es irreductible y cada estado es recurrente o transitorio. La ecuacion $\boldsymbol{\pi} = \boldsymbol{\pi} PAGS$ se convierte

$$\begin{aligned} \pi_0 &= q\,\pi_0 + q\,\pi_1, \\ \pi_1 &= pag\,\pi_0 + q\,\pi_2, \\ \pi_2 &= pag\,\pi_1 + q\,\pi_3, \\ \pi_3 &= pag\,\pi_2 + q\,\pi_4, \end{aligned}$$

y así. Podemos resolver este conjunto de ecuaciones secuencialmente. si dejamos $r = p/q$, entonces podemos expresar la $\pi_1, \pi_2, \ldots$ en términos de π_0 y r como

$$\pi_j = r^j \pi_0, \ j = 0,1,2, \ldots.$$

Si $p < q$, después $r < 1$ una $\sum_{j=0}^{\infty} \pi_j = \pi_0/(1-r)$, y al elegir $\pi_0 = 1-r$, $\pi_j = 1$. Por lo tanto, podemos hacer la suma para $r < 1$, hemos encontrado la limitante distribución $\boldsymbol{\pi} = (1-r)(1, r, r^2, r^3, \ldots)$ para este Markov cha∑ Los estados son en, y todo el por lo tanto positivos recurrentes. Sin embargo cuando $p = q$, π_j es 0 o infinito, y por lo tanto todos los estados son nulos-recurrentes o transitorios. (Se puede demostrar que sólo el caso $p = q$ conduce a estados recurrentes nulos.)

Dejar X sea una cadena de Markov con distribución limitante $\boldsymbol{\pi}$. Suponer $\boldsymbol{\pi}^{(0)} = \boldsymbol{\pi}$. Entonces, combinando (1.33) y (1.35), tenemos $\boldsymbol{\pi}^{(t)} = \boldsymbol{\pi}$. Por tanto, si la distribución inicial de la cadena de Markov es igual a la distribución límite, entonces la distribución de X_t es igual para todos t (y está dada por esta distribución límite). De hecho, no es difícil demostrar que para cualquier k la distribución de $X_k, X_{k+1}, X_{k+2}, \ldots$ es el mismo que el de $X_0, X_1, \ldots$. En otras palabras, cuando $\boldsymbol{\pi}^{(0)} = \boldsymbol{\pi}$, la cadena de Markov es un proceso estocástico estacionario. Más formalmente, un proceso estocástico $\{X_t, t \in \text{norte}\}$ se llama *estacionario* si, para cualquier positivo $\tau, t_1, \ldots, t_{norte}$, el vector $(X_t, \ldots, X_t)$ tiene la misma distribución que $(X_{t_1+\tau}, \ldots, X_{t_{norte}+\tau})$. Definiciones similares son válidas cuando el conjunto de índices es Z, R+, or Por esta razón cualquier distribución $\boldsymbol{\pi}$ para lo cual (1.35) se cumple se llama *estación*∑ *distribución ry*.

Señalando que $_j p_{ags_{yo}} = 1$, podemos reescribir (1.35) como el sistema de ecuaciones

$$\sum_j \pi_i p_{ags_{yo}} = \sum_j \pi_j p_{ags_{ji}} \text{ para todos } i \in mi \tag{1.36}$$

Estos se llaman los *ecuaciones de equilibrio global*. Podemos interpretar (1.35) como el enunciado de que el "flujo de probabilidad" de i se equilibra con el flujo de probabilidad en i. Una generalización importante, que se deriva directamente de (1.36), establece que el mismo equilibrio de los flujos de probabilidad se cumple para un conjunto arbitrario A. Es decir, para cada conjunto A de estados que tenemos

$$\sum_{i \in A} \sum_{j \in A^c} \pi_i p_{ags_{yo}} = \sum_{i \in A} \sum_{j \in A^c} \pi_j p_{ags_{ji}}. \tag{1.37}$$

1.13.4 Reversibilidad

La reversibilidad es una noción importante en la teoría de Markov y procesos más generales. Un proceso estocástico estacionario $\{X_t\}$ con juego de índices Zo R se ha dicho *reversible* si, para cualquier entero positivo *norte* y para todos $t_1, \ldots, t_{norte}$, el vector $(X_t, \ldots, X_t)$ tiene la misma distribución que $(X_{-t_1}, \ldots, X_{-t_{norte}})$. Una forma de visualizar esto es imaginar que hemos tomado un video del proceso estocástico, que podemos ejecutar hacia adelante y hacia atrás. Si no podemos determinar si el video avanza o retrocede, el proceso es reversible. El resultado principal para las cadenas de Markov reversibles es que un proceso de Markov estacionario es reversible si y solo si existe una colección de números positivos $\{\pi_i, i \in mi\}$ sumando a la unidad que satisfacen el *ecuaciones de equilibrio detalladas (o locales)*

$$\pi_i p_{ags_{yo}} = \pi_j p_{ags_{ji}}, yo, j \in mi \tag{1.38}$$

Siempre que tal colección $\{\pi_j\}$ existe, es la distribución estacionaria del proceso.

Una buena manera de pensar en las ecuaciones de equilibrio detalladas es que equilibran el flujo de probabilidad del estado i a estado j con eso del estado j a estado i. Contraste

esto con las ecuaciones de equilibrio (1.36), que equilibran el flujo de probabilidad fuera del estado *i* con eso en estado *i*.

criterio de Kolmogorov es un criterio simple para la reversibilidad basado en las probabilidades de transición. Establece que un proceso de Markov estacionario es reversible si y solo si sus tasas de transición satisfacen

$$pags(i_1, i_2)pags(i_2, i_3)\ldots pags(i_{n-1}, i_{norte})pags(i_{norte}, i_1) = pags(i_1, i_{norte})pags(i_{norte}, i_{n-1})\ldots pags(i_2, i_1) \quad (1.39)$$

para todos los bucles finitos de estados $i_1, \ldots, i_{norte}, i_1$. (Para mayor claridad, hemos utilizado la notación $pags(yo, j)$ más bien que $pags_{yo}$ para las probabilidades de transición.) La idea es bastante intuitiva: si es más probable que el proceso en el tiempo hacia adelante atraviese un cierto ciclo cerrado en una dirección que en la dirección opuesta, entonces en el tiempo hacia atrás exhibirá el comportamiento opuesto y, por lo tanto, tener un criterio para detectar la dirección del tiempo. Si tal comportamiento de "bucle" no ocurre, el proceso debe ser reversible.

1.13.5 Procesos de salto de Markov

A *Proceso de salto de Markov* $X=\{X_t, t\text{-}0\}$ puede verse como una generalización en tiempo continuo de una cadena de Markov y también de un proceso de Poisson. La propiedad de Markov (1.30) ahora dice

$$\text{PAGS}(X_{t+s}=X_{t+s}/X_{tu}=X_{tu},\ tu\text{-}t) = \text{PAGS}(X_{t+s}=X_{t+s}/X_t=X_t). \quad (1.40)$$

Como en el caso de la cadena de Markov, por lo general se supone que el proceso es *homogéneo en el tiempo*, eso es, $\text{PAGS}(X_{t+s}=j\ /X_t=i)$ no depende de t. Denote esta probabilidad por $PAGS_s(yo, j)$. Una cantidad importante es la *tasa de transición* q_{yo} del estado i a j, definido para $i=j$ como

$$q_{yo}=\lim_{t\downarrow 0}\ \frac{PAGS(yo, j)}{t}\ .$$

La suma de las tasas fuera del estado i se denota por q_i. Una ruta de muestra típica de X se muestra en la Figura 1.6. El proceso salta a veces. $T_1, T_2, \ldots$ a los estados $Y_1, Y_2, \ldots$, permaneciendo cierto tiempo en cada estado.

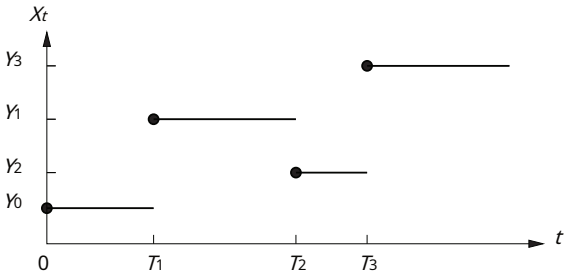


Figura 1.6: Una ruta de muestra de un proceso de salto de Markov $\{X_t, t\text{-}0\}$.

Más precisamente, un proceso de salto de Markov X se comporta (bajo condiciones de regularidad adecuadas; ver [3]) de la siguiente manera:

1. Dado su pasado, la probabilidad de que X salta de su estado actual i a estado j es $k_{yo}=q_{yo}/q_i$.

2. La cantidad de tiempo que X gasta en el estado j tiene una distribución exponencial con media $1/q_j$, independiente de su historia pasada.

La primera afirmación implica que el proceso $\{Y_{norte}\}$ es de hecho una cadena de Markov, con matriz de transición $k = (k_{yo})$.

Una forma conveniente de describir un proceso de salto de Markov es a través de su *gráfico de tasa de transición*. Esto es similar a un gráfico de transición para las cadenas de Markov. Los estados están representados por los nodos del gráfico y una tasa de transición del estado i a j se indica con una flecha desde i a j con peso q_{yo}.

EJEMPLO 1.12 Proceso de nacimiento y muerte

A *proceso de nacimiento y muerte* es un proceso de salto de Markov con un gráfico de tasa de transición de la forma dada en la Figura 1.7. Imagina eso X_t representa el número total de individuos en una población en el tiempo t. Los saltos a la derecha corresponden a nacimientos, y los saltos a la izquierda a defunciones. los *tasas de natalidad* $\{b_i\}$ y el *tasas de mortalidad* $\{d_i\}$ puede diferir de un estado a otro. Muchas aplicaciones de las cadenas de Markov involucran procesos de este tipo. Tenga en cuenta que el proceso salta de un estado a

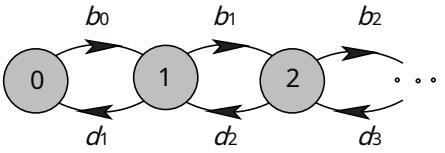


Figura 1.7: El gráfico de tasa de transición de un proceso de nacimiento y muerte.

el siguiente según una cadena de Markov con probabilidades de transición $k_{0,1} = 1$, $k_{yo,\, yo+1} = b_i/(b_i + d_i)$, y $k_{yo,\, yo-1} = d_i/(b_i + d_i)$, $i = 1, 2, \ldots$. Además, gasta un Exp(b_0) cantidad de tiempo en el estado 0 y Exp($b_i + d_i$) en los demás estados.

Comportamiento limitante Ahora formulamos los análogos en tiempo continuo de (1.34) y el Teorema 1.13.2. La irreductibilidad y la recurrencia para los procesos de salto de Markov se definen de la misma manera que para las cadenas de Markov. Por simplicidad, supongamos que $mi = \{1, 2, \ldots\}$. Si X es un proceso de salto de Markov recurrente e irreducible, entonces independientemente de i,

$$\lim_{t\to\infty} \text{PAGS}(X_t = j \mid X_0 = i) = \pi_j \tag{1.41}$$

por algún número $\pi_j > 0$. Además, $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots)$ es la solución a

$$\sum_{j=i} \pi_i q_{yo} = \sum_{j=i} \pi_j q_{ji}, \text{para todos } i = 1, \ldots, \textit{metro} \tag{1.42}$$

con $\sum_j \pi_j = 1$, si tal solución existe, en cuyo caso todos los estados son positivos recurrente. Si tal solución no existe, todos π_j son 0

Como en el caso de la cadena de Markov, $\{\pi_j\}$ se llama el *limitar la distribución* o $\sum$ F X y es generalmente identificado con el vector de fila $\boldsymbol{\pi}$. Alguna solución $\boldsymbol{\pi}$ de (1.42) con $\sum_j \pi_j = 1$ se llama un *distribución estacionaria*, ya que tomarlo como la distribución inicial del proceso de salto de Markov hace que el proceso sea estacionario.

Las ecuaciones (1.42) se llaman de nuevo *ecuaciones de equilibrio global* y se generalizan fácilmente a (1.37), reemplazando las probabilidades de transición con tasas de transición. Más importante aún, si el proceso es reversible, entonces, al igual que con las cadenas de Markov, la distribución estacionaria se puede encontrar a partir de la *ecuaciones de balance local*:

$$\pi_i q_{yo} = \pi_j q_{ji}, yo, j \in mi \tag{1.43}$$

La reversibilidad se puede verificar fácilmente comprobando que no se produzca bucle, es decir, mediante el criterio de Kolmogorov (1.39), reemplazando las probabilidades *pags* con tarifas *q*.

EJEMPLO 1.13 *M/M/*1 cola

Considere una instalación de servicio donde los clientes llegan en ciertos momentos aleatorios y son atendidos por un solo servidor. Los clientes que llegan y encuentran el servidor ocupado esperan en la cola. Los clientes son atendidos en el orden en que llegan. Los tiempos entre llegadas son variables aleatorias exponenciales con tasas λ, y los tiempos de servicio de los clientes son variables aleatorias exponenciales iid con tasas m. Por último, los tiempos de servicio son independientes de los tiempos entre llegadas. Dejar X_t sea el número de clientes en el sistema en ese momento t. Por la propiedad sin memoria de la distribución exponencial (vea el problema 1.7), no es difícil ver que $X = \{X_t, t$-0 $\}$ es un proceso de salto de Markov y, de hecho, un proceso de nacimiento y muerte con tasas de natalidad $b_i = \lambda, i = 0,1,2, \ldots$ y tasas de mortalidad $d_i = mi, i = 1,2, \ldots$.

Resolver las ecuaciones de balance global (o, más fácilmente, las ecuaciones de balance local, ya que X es reversible), vemos que X_t tiene una distribución límite dada por

$$\lim_{t\to\infty} \text{PAGS}(X_t = norte) = (1 -)_{norte},\ norte = 0,1,2, \ldots, \tag{1.44}$$

siempre que $= \lambda/\mu < 1$. Esto significa que el tiempo de servicio esperado debe ser menor que el tiempo entre llegadas esperado para que exista una distribución límite. En ese caso, la distribución límite es también la distribución estacionaria. En particular, si X_0 se distribuye de acuerdo con (1.44), entonces X_t tiene la misma distribucion para todos $t > 0$.

1.14 PROCESOS GAUSSIANOS

La distribución normal también se llama distribución *gaussiano* distribución. Los procesos gaussianos son generalizaciones de vectores aleatorios normales multivariados (discutidos en la Sección 1.10). Específicamente, un proceso estocástico $\{X_t, t \in T\}$ se ha dicho *gaussianos* si todas sus distribuciones de dimensión finita son gaussianas. Es decir, si para cualquier elección de *norte* y $t_1, \ldots, t_{norte} \in T$, sostiene que

$$(X_{t_1}, \ldots, X_{t_{norte}}) \sim \text{NORTE}(\boldsymbol{m}, \Sigma) \tag{1.45}$$

para algún vector de expectativa $\boldsymbol{m}$ y matriz de covarianza Σ (ambos dependen de la elección de $t_1, \ldots, t_{norte}$). combinación Equivalentemente, $\{X_t, t \in T\}$ es gaussiana si alguna lineal $\sum_{i=1}^{norte} b_i X_{t_i}$ tiene una distribución normal. Tenga en cuenta que un proceso gaussiano está completamente determinada por su *función de expectativa* $\mu_t = \text{MI}[X_t], t \in T$, y *función de covarianza* $\Sigma_{s\,t} = \text{cov}(X_s, X_t), S\ t \in T$.

EJEMPLO 1.14 Proceso de Wiener (movimiento browniano)

El proceso gaussiano por excelencia es el *proceso de salchicha* o (estándar) *movimiento browniano*. Puede verse como una versión continua de un proceso de caminata aleatoria. La Figura 1.8 da una ruta de muestra típica. El proceso de Wiener juega un papel central en la probabilidad y constituye la base de muchos otros procesos estocásticos.

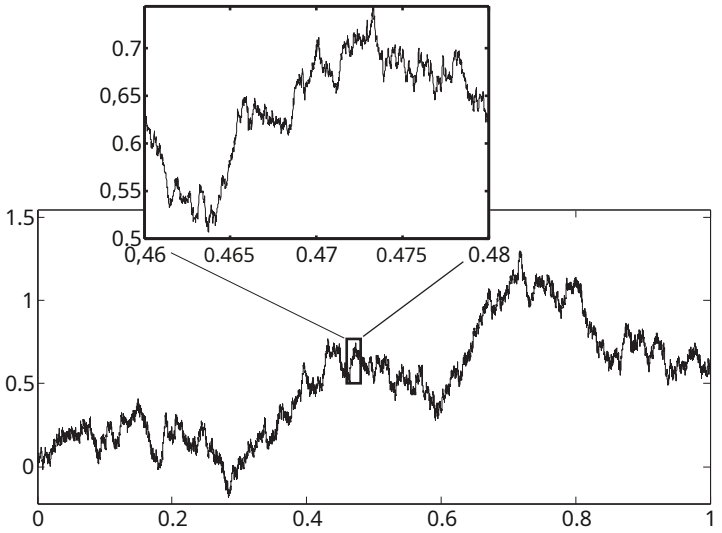


Figura 1.8: Una ruta de muestra del proceso de Wiener. El recuadro muestra una ampliación de la ruta durante un pequeño intervalo de tiempo.

El proceso de Wiener se puede definir como un proceso gaussiano $\{X_t, t \geq 0\}$ con función de expectativa $m_t = 0$ para todos t y función de covarianza $\Sigma_{s,t} = s$ para $0 \leq s \leq t$. El proceso de Wiener tiene muchas propiedades fascinantes (p. ej., [11]). Por ejemplo, es un proceso de Markov (es decir, satisface la propiedad de Markov (1.30)) con rutas de muestra continuas que son *en ninguna parte diferenciable*. Además, los incrementos $X_t - X_s$ en intervalos [*calle*] son independientes y normalmente distribuidos. En concreto, para cualquier $t_1 < t_2 \leq t_3 < t_4$,

$$X_{t_4} - X_{t_3} \quad \text{y} \quad X_{t_2} - X_{t_1}$$

son variables aleatorias independientes, y para todos $t \geq s \geq 0$,

$$X_t - X_s \sim \text{NORTE}(0, t - s).$$

Esto conduce a un procedimiento de simulación simple para los procesos de Wiener, que se analiza en la Sección 2.8.

1.15 INFORMACIÓN

En esta sección discutimos brevemente varias medidas de información en un experimento aleatorio. Supongamos que describimos las medidas en un experimento aleatorio a través de

un vector aleatorio$\mathbf{X} = (X_1, \ldots, X_n)$ con pdf F. Entonces toda la información sobre el experimento (todo nuestro conocimiento probabilístico) obviamente está contenida en el pdf F. Sin embargo, en la mayoría de los casos nos gustaría caracterizar nuestra información sobre los experimentos con solo unos pocos números clave, como el *expectativa* y el *Matriz de covarianza* de $\mathbf{X}$, que proporcionan información sobre las medidas medias y la variabilidad de las medidas, respectivamente. Otra medida informativa proviene de la teoría de la codificación y las comunicaciones, donde el *Entropía de Shannon* caracteriza el número promedio de bits necesarios para transmitir un mensaje $\mathbf{X}$ a través de un canal de comunicación (binario). Otro enfoque más de la información se puede encontrar en las estadísticas. Específicamente, en la teoría de la estimación puntual, la función de densidad de probabilidad F depende de un vector de parámetros $\boldsymbol{\theta}$. La pregunta es qué tan bien $\boldsymbol{\theta}$ se puede estimar a través de un resultado de $\mathbf{X}$—en otras palabras, cuánta información sobre $\boldsymbol{\theta}$ está contenido en los "datos" $\mathbf{X}$. Diversas medidas para este tipo de información están asociadas a la *máxima verosimilitud*, la *puntaje*, y el *(Fisher) matriz de información*. Finalmente, la cantidad de información en un experimento aleatorio a menudo se puede cuantificar a través de un *distancia* concepto, como el *Kullback–Leibler* "distancia" (divergencia), también llamada *entropía cruzada*.

1.15.1 Entropía de Shannon

Una de las medidas de incertidumbre más celebradas en la teoría de la información es la *Entropía de Shannon*, o simplemente *entropía*. Una buena referencia es [4], donde la entropía de una variable aleatoria discreta X con densidad F Se define como

$$\text{mi}\left[\text{Iniciar sesión}_2 F(X)\right] = -\text{mi}\left[\text{Iniciar sesión}_2 f(X)\right] = -\sum_X F(X)\,\text{Iniciar sesión}_2 F(X).$$

Aquí X se interpreta como un carácter aleatorio de un alfabeto $\mathcal{X}$, tal que $X = x$ con probabilidad $F(x)$. Usaremos la convención $0 \ln 0 = 0$.

Se puede demostrar que la forma más eficiente de transmitir caracteres muestreados de F sobre un canal binario es codificarlos de tal manera que el número de bits necesarios transmitir X es igual a $\log_2(1/F(X))$. Resulta que $-\sum_x F(x)\,\text{Iniciar sesión}_2 F(x)$ es la longitud de bit esperada necesaria para enviar un carácter aleatorio $X \sim F$; ver [4].

Un enfoque más general, que incluye variables aleatorias continuas, es definir la entropía de una variable aleatoria X con densidad F por

$$\mathrm{H}(X) = -\mathrm{MI}[\text{en}\, F(X)] = \begin{cases} -\sum F(x)\,\text{en}\, F(x) & \text{caso discreto,} \\ -\int F(x)\,\text{en}\, F(x)\,\mathrm{d}x & \text{caso continuo.} \end{cases} \tag{1.46}$$

La definición (1.46) se puede extender fácilmente a vectores aleatorios $\mathbf{X}$ como (en el caso continuo)

$$\mathrm{H}(\mathbf{X}) = -\mathrm{MI}[\text{en}\, F(\mathbf{X})] = -\int F(\mathbf{x})\,\text{en}\, F(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{1.47}$$

$\mathrm{H}(\mathbf{X})$ es a menudo llamado el *articulación* entropía de las variables aleatorias $X_1, \ldots, X_n$, y también se escribe como $\mathrm{H}(X_1, \ldots, X_n)$. En el caso continuo, $\mathrm{H}(\mathbf{X})$ es frecuentemente referido como el *entropía diferencial* distinguirlo del caso discreto.

EJEMPLO 1.15

Dejar X tener un Ber($pags$) distribución para algunos 0 - $pags$ -1. La densidad F de X es dado por $F(1)$ =PAGS(X=1) = $pags$ y $F(0)$ =PAGS(X=0) = 1 − $pags$ de modo que la entropía de X es

$$H(X) = -pags \text{ en } pag - (1-pags) \text{ en } (1-pags).$$

La gráfica de la entropía en función de $pags$ se representa en la figura 1.9. Tenga en cuenta que la entropía es máxima para $pags$=1/2, que da la densidad "uniforme" en {0 ,1}.

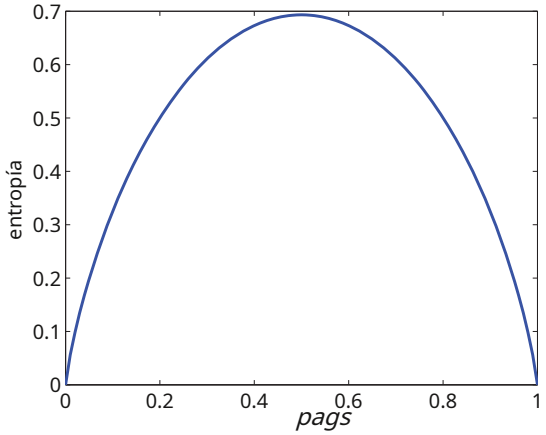


Figura 1.9: La entropía para el Ber($pags$) distribución en función de $pags$.

A continuación, considere una secuencia $X_1, \ldots, X_{norte}$ de iid Ber($pags$) variables aleatorias. Dejar $\mathbf{X} = (X_1, \ldots, X_{norte})$. La densidad de $\mathbf{X}$, decir $gramo$, es simplemente el producto de las densidades de los X_i, de modo que

$$H(\mathbf{X}) = -\text{mi } [\text{en } gramo(\mathbf{X})] = -\text{mien } \left[\prod_{i=1}^{norte} F(X_i) \right] = \sum_{i=1}^{norte} -\text{mi } [\text{en } F(X_i)] = norte H(X).$$

las propiedades de H($\mathbf{X}$) en el caso continuo son algo diferentes de las del caso discreto. En particular:

1. La entropía diferencial puede ser negativa, mientras que la entropía discreta siempre es positiva.

2. La entropía discreta es insensible a las transformaciones invertibles, mientras que la entropía diferencial no lo es. Específicamente, si $\mathbf{X}$ es discreto, $\mathbf{Y} = gramo(\mathbf{X})$, y $gramo$ es una aplicación invertible, entonces H($\mathbf{X}$) =H($\mathbf{Y}$) porque $F_Y(\mathbf{y}) = F_X(gramo^{-1}(\mathbf{y}))$. Sin embargo, en el caso continuo, tenemos un factor adicional debido al jacobiano de la transformación.

No es difícil ver que de cualquier densidad F, la que da la máxima entropía es la densidad uniforme en X. Eso es,

$$H(\mathbf{X}) \text{ es máximo} \Leftrightarrow F(\mathbf{X}) = \frac{1}{|X|} \text{ (constante).} \tag{1.48}$$

Para dos vectores aleatorios**X**y**Y**con pdf conjunto*F*, definimos el*entropía condicional*de**Y**dado**X**como

$$H(\mathbf{Y}/\mathbf{X}) = -\text{mien}\left[\frac{F(\mathbf{X},\mathbf{Y})}{F_{\mathbf{X}}(\mathbf{X})}\right] = H(\mathbf{X},\mathbf{Y}) - H(\mathbf{X}), \tag{1.49}$$

dónde$F_{\mathbf{X}}$es el pdf de**X**y$F_{(\mathbf{X},\mathbf{y})}$ $F_{\mathbf{X}}(\mathbf{x})$es la densidad condicional de**Y**(a**y**), dado **X**=**X**. Resulta que

$$H(\mathbf{X},\mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y}/\mathbf{X}) = H(\mathbf{Y}) + H(\mathbf{X}/\mathbf{Y}). \tag{1.50}$$

Es razonable exigir que cualquier medida aditiva sensata que describa la cantidad promedio de incertidumbre satisfaga al menos (1.50) y (1.48). De ello se deduce que la densidad uniforme lleva la menor cantidad de información, y la entropía (cantidad media de incertidumbre) de (**X**,**Y**) es igual a la suma de la entropía de**X** y la cantidad de entropía en**Y**después de la información en**X**ha sido contabilizado. Se argumenta en [10] que cualquier concepto de entropía que incluya las propiedades generales (1.48) y (1.50) debe conducir a la definición (1.47).

los*información mutua*de**X**y**Y**Se define como

$$\text{METRO}(\mathbf{X},\mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y}) - H(\mathbf{X},\mathbf{Y}), \tag{1.51}$$

que, como su nombre indica, puede interpretarse como la cantidad de información compartida por**X**y**Y**. Una expresión alternativa, que se sigue de (1.50) y (1.51), es

$$\text{METRO}(\mathbf{X},\mathbf{Y}) = H(\mathbf{X}) - H(\mathbf{X}/\mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y}/\mathbf{X}), \tag{1.52}$$

que puede interpretarse como la reducción de la incertidumbre de una variable aleatoria debido al conocimiento de la otra. No es difícil demostrar que la información mutua siempre es positiva. También está relacionado con el concepto de entropía cruzada, que sigue.

1.15.2 Entropía cruzada de Kullback-Leibler

Dejar*gramo*y*h*ser dos densidades en*X*.La entropía cruzada de Kullback-Leibler entre *gramo*y*h*(comparar con (1.47)) se define (en el caso continuo) como

$$D(g,h) = \text{mi}_{gramo}\left[\text{en}\frac{gramo(\mathbf{X})}{h(\mathbf{X})}\right] = \int_{gramo(\mathbf{X})\,\text{en}\,gramo(\mathbf{X})\,d\mathbf{X}} - \int_{gramo(\mathbf{X})\,\text{en}\,h(\mathbf{X})\,d\mathbf{X}}. \tag{1.53}$$

D(*g, h*) también se llama*Divergencia Kullback-Leibler*, la*entropía cruzada*, y el *entropía relativa*. Si no se indica lo contrario, llamaremosD(*g, h*) la*entropía cruzada* (CE) entre*gramo*y*h*. Darse cuenta deD(*g, h*) no es una distancia entre*gramo*y*h*en el sentido formal, ya que en generalD(*g, h*) =D(*h, g*). Sin embargo, a menudo es útil pensar enD(*g, h*) como una distancia porque

$$D(g,h) - 0$$

y$D(g, h) = 0$ si y solo si$gramo(X) = h(X)$. Esto se sigue de la desigualdad de Jensen (siφes una función convexa, tal como$-$entonces$MI[\varphi(X)] \le \varphi(MI[X])$. A saber

$$D(g, h) = mi_{gramo}-en\left[\frac{h(\mathbf{X})}{gramo(\mathbf{X})}\right] \le -en\left\{\left[\frac{h(\mathbf{X})}{mi_{gramo}\, gramo(\mathbf{X})}\right]\right\} = -en\, 1 = 0.$$

Se puede ver fácilmente que la información mutuaMETRO($\mathbf{X},\mathbf{Y}$) de vectores$\mathbf{X}$y $\mathbf{Y}$definido en (1.51) está relacionado con el CE de la siguiente manera:

$$METRO(\mathbf{X},\mathbf{Y}) = D(f, f_X f_Y) = mi_f en\left[\frac{f(\mathbf{X},\mathbf{Y})}{f_X(\mathbf{X}) f_Y(\mathbf{Y})}\right],$$

dónde$f$es el pdf (conjunto) de ($\mathbf{X},\mathbf{Y}$) yf_Xy$f_Y$son los pdf (marginales) de$\mathbf{X}$ y$\mathbf{Y}$, respectivamente. En otras palabras, la información mutua se puede ver como el CE que mide la distancia entre el pdf conjunto$f$de$\mathbf{X}$y$\mathbf{Y}$y el producto de sus pdf marginalesf_Xyf_Y, es decir, bajo el supuesto de que los vectores $\mathbf{X}$y$\mathbf{Y}$son *independiente*.

1.15.3 Estimador de máxima verosimilitud y función de puntuación

Introducimos aquí la noción de*función de puntuación*(SF) a través de la clásica*estimador de máxima verosimilitud*. Considere un vector aleatorio$\mathbf{X} = (X_1, \ldots, X_{norte})$ que se distribuye de acuerdo con un pdf fijo$f(\cdot;\boldsymbol{\theta})$ con parámetro desconocido (vector)$\boldsymbol{\theta}\in\Theta$. Digamos que queremos estimar$\boldsymbol{\theta}$sobre la base de un resultado dado$\mathbf{X}$(los datos) de$\mathbf{X}$. Para una dada$\mathbf{X}$, la función$L(\boldsymbol{\theta};\mathbf{X}) = f(\mathbf{X};\boldsymbol{\theta})$ se llama el*función de probabilidad*. Tenga en cuenta que$L$es una función de$\boldsymbol{\theta}$para un parámetro fijo$\mathbf{X}$, mientras que para el pdf$f$es al revés. La máxima probabilidad*estimar*$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{X})$ de$\boldsymbol{\theta}$Se define como

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}} L(\boldsymbol{\theta};\mathbf{X}). \tag{1.54}$$

Como la función ln es monótona creciente, también tenemos

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{entrada argmax}} L(\boldsymbol{\theta};\mathbf{X}). \tag{1.55}$$

la variable aleatoria$\hat{\boldsymbol{\theta}}(\mathbf{X})$ con$\mathbf{X}\sim f(\cdot;\boldsymbol{\theta})$ es la probabilidad máxima correspondiente capucha*estimador*, que se escribe de nuevo como$\hat{\boldsymbol{\theta}}$. Tenga en cuenta que a menudo el da$\prod_{i=1}^{norte} f_1(X_i;\boldsymbol{\theta})$ ejército de reserva$X_1, \ldots, X_{norte}$ formar una muestra aleatoria de algún pdf$f_1(\cdot;\boldsymbol{\theta})$, en ese caso$f(\mathbf{X};\boldsymbol{\theta}) = $ y

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}} \sum_{i=1}^{norte} en\, f_1(X_i;\boldsymbol{\theta}). \tag{1.56}$$

Si$L(\boldsymbol{\theta};\mathbf{X})$ es una función cóncava continuamente diferenciable con respecto a$\boldsymbol{\theta}$ y el máximo se alcanza en el interior de Θ, entonces podemos encontrar el estimador de máxima verosimilitud de$\boldsymbol{\theta}$resolviendo

$$\nabla_{\boldsymbol{\theta}} en\, L(\boldsymbol{\theta};\mathbf{X}) = \mathbf{0}.$$

La función$S(\cdot;\mathbf{X})$ definido por

$$S(\boldsymbol{\theta};\mathbf{X}) = \nabla_{\boldsymbol{\theta}} en\, L(\boldsymbol{\theta};\mathbf{X}) = \frac{\nabla_{\boldsymbol{\theta}} f(\mathbf{X};\boldsymbol{\theta})}{f(\mathbf{X};\boldsymbol{\theta})} \tag{1.57}$$

se llama el*función de puntuación*. Para la familia exponencial (A.9) es fácil ver que

$$S(\boldsymbol{\theta};\mathbf{X}) = \frac{\nabla C(\boldsymbol{\theta})}{C(\boldsymbol{\theta})} + \mathbf{t}(\mathbf{X}). \tag{1.58}$$

los*vector aleatorio*$S(\boldsymbol{\theta}) = S(\boldsymbol{\theta};\mathbf{X})$ con$\mathbf{X} \sim F(\cdot;\boldsymbol{\theta})$ se llama el*(eficiente) puntuación*. La puntuación esperada siempre es igual al vector cero, es decir,

$$\mathrm{mi}_{\boldsymbol{\theta}}[S(\boldsymbol{\theta})] = \int \nabla_{\boldsymbol{\theta}} F(\mathbf{X};\boldsymbol{\theta})\, m(\mathrm{d}\mathbf{X}) = \nabla_{\boldsymbol{\theta}} \int F(\mathbf{X};\theta)\, m(\mathrm{d}\mathbf{X}) = \nabla_{\boldsymbol{\theta}} 1 = \mathbf{0},$$

donde el intercambio de diferenciación e integración se justifica a través del teorema de convergencia acotada.

1.15.4 Información del pescador

La matriz de covarianza$\mathrm{YO}(\boldsymbol{\theta})$ de la partitura$S(\boldsymbol{\theta})$ se llama el*Matriz de información de Fisher*. Dado que la puntuación esperada siempre es$\mathbf{0}$, tenemos

$$\mathrm{YO}(\boldsymbol{\theta}) = \mathrm{mi}_{\boldsymbol{\theta}}\left[S(\boldsymbol{\theta})S(\boldsymbol{\theta})^{\mathsf{T}}\right]. \tag{1.59}$$

En el caso unidimensional, tenemos entonces

$$\mathrm{YO}(\theta) = \mathrm{mi}_{\theta}\left[\left(\frac{\partial \operatorname{en} F(X;\theta)}{\partial \theta}\right)^2\right].$$

Porque

$$\frac{\partial^2}{\partial \theta^2} \operatorname{en} F(X;\theta) = \frac{\frac{\partial^2}{\partial \theta^2} F(X;\theta)}{F(X;\theta)} - \left(\frac{\frac{\partial}{\partial \theta} F(X;\theta)}{F(X;\theta)}\right)^2,$$

vemos que (en condiciones de regularidad directa) la información de Fisher también está dada por

$$\mathrm{YO}(\theta) = -\mathrm{mi}_{\theta}\left[\frac{\partial^2 \operatorname{en} F(X;\theta)}{\partial \theta^2}\right].$$

En el caso multidimensional tenemos de manera similar

$$\mathrm{YO}(\boldsymbol{\theta}) = -\mathrm{mi}_{\boldsymbol{\theta}}[\nabla S(\boldsymbol{\theta})] = -\mathrm{mi}_{\boldsymbol{\theta}}\left[\nabla^2 \operatorname{en} F(\mathbf{X};\boldsymbol{\theta})\right], \tag{1.60}$$

dónde$\nabla^2 \operatorname{en} F(\mathbf{X};\boldsymbol{\theta})$ denota el*arpillera*de $\ln F(\mathbf{X};\boldsymbol{\theta})$, es decir, la matriz (aleatoria)

$$\left(\frac{\partial^2 \operatorname{en} F(\mathbf{X};\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right).$$

La importancia de la información de Fisher en las estadísticas es corroborada por el famoso *Desigualdad de Cramér-Rao*, que (de forma simplificada) establece que la varianza de cualquier estimador insesgado$Z$de$gramo(\boldsymbol{\theta})$ está acotado desde abajo a través de

$$\mathrm{Var}(Z) \geq (\nabla gramo(\boldsymbol{\theta}))^{\mathsf{T}} \mathrm{yo}^{-1}(\boldsymbol{\theta}) \nabla gramo(\boldsymbol{\theta}). \tag{1.61}$$

Para más detalles, véase [12].

1.16 OPTIMIZACIÓN CONVEXA Y DUALIDAD

Dejar $F(X)$, $X \in \mathrm{R}$, sea una función de valor real con derivadas continuas, también llamada función C_1 función. El enfoque estándar para minimizar $F(X)$ es resolver la ecuación

$$F'(X) = 0. \tag{1.62}$$

Las soluciones de (1.62) se llaman *puntos estacionarios*. Si, además, la función tiene segundas derivadas continuas (la llamada C_2 función), la condición

$$F''(X_*) > 0 \tag{1.63}$$

asegura que un punto estacionario X_* es un *minimizador local*, eso es, $F(X_*) < F(X)$ para todos X en un barrio bastante pequeño de X_*.

Para C_1 función en R^{norte}, (1.62) se generaliza a

$$\nabla F(\mathbf{X}) \equiv \begin{pmatrix} \frac{\partial f(\mathbf{X})}{\partial x_1} \\ \dots \\ \frac{\partial f(\mathbf{X})}{\partial x_{norte}} \end{pmatrix} = \mathbf{0}, \tag{1.64}$$

dónde $\nabla F(\mathbf{X})$ es el *degradado* de F a $\mathbf{X}$. Del mismo modo, un punto estacionario $\mathbf{X}_*$ es un minimizador local de F Si el *matriz Hessiana* (o simplemente *arpillera*) a $\mathbf{X}_*$,

$$\nabla^2 F(\mathbf{X}_*) \equiv \begin{pmatrix} \frac{\partial^2 F(\mathbf{X}_*)}{\partial x_1^2} & \dots & \frac{\partial^2 F(\mathbf{X}_*)}{\partial x_1 \partial x_{norte}} \\ \dots & \dots & \vdots \\ \frac{\partial^2 F(\mathbf{X}_*)}{\partial x_1 \partial x_{norte}} & \dots & \frac{\partial^2 F(\mathbf{X}_*)}{\partial x_{norte}^2} \end{pmatrix}, \tag{1.65}$$

es *positivo definitivo*, eso es, $\mathbf{X}^T [\nabla^2 F(\mathbf{X}_*)] \mathbf{X} > 0$ para todos $\mathbf{X} = \mathbf{0}$.

La situación se puede generalizar aún más introduciendo *restricciones*. Un problema general de optimización con restricciones se puede escribir como

$$\min_{\mathbf{X} \in \mathrm{R}^{norte}} \quad F(\mathbf{X}) \tag{1.66}$$

$$\text{sujeto a:} \quad h_i(\mathbf{X}) = 0, \quad i = 1, \dots, m, \tag{1.67}$$

$$g_{yo}(\mathbf{X}) \le 0, \quad yo = 1, \dots, k \tag{1.68}$$

Aquí f, g_i, y h_i se dan funciones, $F(\mathbf{X})$ se llama el *función objetiva*, y $h_i(\mathbf{X}) = 0$ y $g_{yo}(\mathbf{X}) \le 0$ representa el *igualdad y desigualdad* restricciones, respectivamente.

La región del dominio donde está definida la función objetivo y donde se satisfacen todas las restricciones se llama *región factible*. A *solución mundial* al problema de optimización es un punto $\mathbf{X}_* \in \mathrm{R}^{norte}$ tal que no existe otro punto $\mathbf{X} \in \mathrm{R}^{norte}$ para cual $F(\mathbf{X}) < F(\mathbf{X}_*)$. Los nombres alternativos son *minimizador global* y *mínimo global*, aunque este último podría confundirse con el valor mínimo de la función. Del mismo modo, para un *local* solución/minimizador, la condición $F(\mathbf{X}) < F(\mathbf{X}_*)$ solo necesita mantenerse en algún vecindario de $\mathbf{X}_*$.

Dentro de esta formulación se encuentran muchos de los problemas de optimización tradicionales. Un problema de optimización en el que la función objetivo y las restricciones de igualdad y desigualdad son funciones lineales se denomina *programa lineal*. Una optimización

problema en el que la función objetivo es cuadrática, mientras que las restricciones son funciones lineales se llama un *programa cuadrático*. La convexidad juega un papel importante en muchos problemas prácticos de optimización.

Definición 1.16.1 (Conjunto convexo) Un conjunto $X \in \mathrm{R}^{norte}$ se llama *convexo* si por todo $\mathbf{X}, \mathbf{y} \in X$ y $\theta \in (0,1)$, el punto $(\theta\mathbf{X} + (1-\theta)\mathbf{y}) \in X$.

Definición 1.16.2 (Función convexa) Una función $F(\mathbf{X})$ en un conjunto convexo X se llama *convexo* si por todo $\mathbf{X}, \mathbf{y} \in X$ y $\theta \in (0,1)$,

$$F(\theta\mathbf{X} + (1-\theta)\mathbf{y}) \leq \theta F(\mathbf{X}) + (1-\theta)F(\mathbf{y}). \tag{1.69}$$

Si se cumple una desigualdad estricta en (1.69), se dice que la función es *estrictamente convexo*. Si una función F es (estrictamente) convexo, entonces $-F$ se dice que es (estrictamente) *cóncavo*. Asumiendo X es un conjunto abierto, convexidad para $F \in C^1$ es equivalente a

$$F(\mathbf{y}) - F(\mathbf{X}) + (\mathbf{y}-\mathbf{X})\nabla F(\mathbf{X}) \qquad \text{para todos } \mathbf{X}, \mathbf{y} \in X\,.$$

Además, por $F \in C^2$, la convexidad es equivalente a que la matriz hessiana sea semidefinida positiva para todo $\mathbf{X} \in X$:

$$\mathbf{y}^{T}\left[\nabla^2 F(\mathbf{X})\right]\mathbf{y} \geq 0, \text{para todos } \mathbf{y} \in \mathrm{R}^{norte}.$$

Se dice que el problema (1.66) es un *problema de programacion convexa* si

1. la función objetivo F es convexo,

2. las funciones de restricción de desigualdad $\{gramo_i(\mathbf{X})\}$ son convexos y

3. las funciones de restricción de igualdad $\{h_i(\mathbf{X})\}$ son *afín*, es decir, de la forma $\boldsymbol{a}_i\mathbf{X} - b_i$.

Tenga en cuenta que el último requisito se deriva del hecho de que una restricción de igualdad $h_i(\mathbf{X}) = 0$ puede verse como una combinación de las restricciones de desigualdad $h_i(\mathbf{X}) \leq 0$ y $-h_i(\mathbf{X}) \leq 0$, de modo que ambos h_i y $-h_i$ necesita ser convexo. Tanto el programa lineal como el cuadrático (con matriz definida positiva C) son convexos.

1.16.1 Método Lagrangiano

Los componentes principales del método de Lagrange son los multiplicadores de Lagrange y la función de Lagrange. El método fue desarrollado por Lagrange en 1797 para el problema de optimización (1.66) con restricciones de igualdad (1.67). En 1951 Kuhn y Tucker extendió el método de Lagrange a las restricciones de desigualdad.

Definición 1.16.3 (Función de Lagrange) Dado un problema de optimización (1.66) que contiene solo restricciones de igualdad $h_i(\mathbf{X}) = 0$, $i = 1, \ldots, metro$, la *Función de Lagrange*, o *Lagrangiano*, Se define como

$$L(\mathbf{X}, \boldsymbol{\beta}) = F(\mathbf{X}) + \sum_{i} \beta_i h_i(\mathbf{X}),$$

donde los coeficientes $\{\beta_i\}$ se llaman los *Multiplicadores de Lagrange*.

Una condición necesaria para un punto.$\mathbf{X}_*$ser un minimizador local de$F(\mathbf{X})$ sujeto a las restricciones de igualdad$h(\mathbf{X}) = 0,\ i=1, \ldots, m$, es

$$\nabla_{\mathbf{x}} L(\mathbf{X}_*, \boldsymbol{\beta}_*) = \mathbf{0},\ \nabla_{\boldsymbol{\beta}} L(\mathbf{X}_*, \boldsymbol{\beta}_*) = \mathbf{0},$$

por algún valor$\boldsymbol{\beta}_*$. Las condiciones anteriores también son suficientes si$L(\mathbf{X}, \boldsymbol{\beta}_*)$ es una función convexa de$\mathbf{X}$.

EJEMPLO 1.16 Distribución máxima de entropía

Dejar$p = \{p_i,\ i=1, \ldots n\}$Sea una distribución de probabilidad. Considere el siguiente programa, que maximiza la entropía (Shannon):

$$\max_{\mathbf{p}} -\sum_{i=1}^{n} p_i \text{en} p_i$$
$$\text{sujeto a: } \sum_{i=1}^{n} p_i = 1.$$

El lagrangiano es

$$L(\mathbf{p}, \beta) = \sum_{i=1}^{n} p_i \text{en} p_i + \beta \left(\sum_{i=1}^{n} p_i - 1 \right)$$

sobre el dominio$\{(\mathbf{p}, \beta) : p_i \geq 0,\ i=1, \ldots, n, \beta \in \mathbb{R}\}$. La solución óptima$\mathbf{p}_*$ del problema es la distribución uniforme, es decir,$\mathbf{p}_* = (1/n, \ldots, 1/n)$; vea el problema 1.35.

Definición 1.16.4 (Función de Lagrange generalizada) Dado el problema de optimización original (1.66), que contiene las restricciones de igualdad y desigualdad, el*función de Lagrange generalizada*, o simplemente*Lagrangiano*, Se define como

$$L(\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = F(\mathbf{X}) + \sum_{i=1}^{k} \alpha_i g_i(\mathbf{X}) + \sum_{i=1}^{m} \beta_i h_i(\mathbf{X}).$$

Una condición necesaria para un punto.$\mathbf{X}_*$ser un minimizador local de$F(\mathbf{X})$ en el problema de optimización (1.66) es la existencia de un$\boldsymbol{\alpha}_*$y$\boldsymbol{\beta}_*$tal que

$$\nabla_{\mathbf{x}} L(\mathbf{X}_*, \boldsymbol{\alpha}_*, \boldsymbol{\beta}_*) = \mathbf{0},\ \nabla_{\boldsymbol{\beta}} L(\mathbf{X}_*, \boldsymbol{\alpha}_*, \boldsymbol{\beta}_*) = \mathbf{0},$$
$$g_i(\mathbf{X}_*) \leq 0, \quad i=1, \ldots, k,$$
$$\alpha_i \geq 0, \quad i=1, \ldots, k,$$
$$\alpha_i g_i(\mathbf{X}_*) = 0, \quad i=1, \ldots, k$$

Estas ecuaciones generalmente se denominan*Condiciones de Karush-Kuhn-Tucker (KKT)*. Para*convexo*programas tenemos los siguientes resultados importantes:

1. Cada solución local$\mathbf{X}_*$a un problema de programación convexo es una solución global y el conjunto de soluciones globales es convexo. Si, además, la función objetivo es estrictamente convexa, entonces cualquier solución global es única.

2. Para un problema de programación estrictamente convexa con C^1 funciones objetivo y de restricción, las condiciones KKT son necesarias y suficientes para una solución global única.

1.16.2 Dualidad

El objetivo de la dualidad es proporcionar una formulación alternativa de un problema de optimización que a menudo es computacionalmente más eficiente o tiene algún significado teórico (ver [7], página 219). El problema original (1.66) se conoce como el *primitivo* problema, mientras que el problema reformulado, basado en los multiplicadores de Lagrange, se conoce como el *doble* problema. La teoría de la dualidad es más relevante para los problemas de optimización convexa. Es bien sabido que si el problema de optimización primal es (estrictamente) convexo, entonces el problema dual es (estrictamente) cóncavo y tiene una solución (única) a partir de la cual se puede deducir la solución primal óptima (única).

Definición 1.16.5 (Programa dual de Lagrange) los *Programa dual de Lagrange* del programa primal (1.66), es

$$\begin{aligned} &\underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\text{máximo}} && L_*(\boldsymbol{\alpha},\boldsymbol{\beta}) \\ &\text{sujeto a:} && \boldsymbol{\alpha} \geq 0, \end{aligned}$$

dónde L_* es el *Doble función de Lagrange*:

$$L_*(\boldsymbol{\alpha},\boldsymbol{\beta}) = \inf_{\mathbf{X}\in X} L(\mathbf{X},\boldsymbol{\alpha},\boldsymbol{\beta}). \tag{1.70}$$

No es difícil ver que si F_* es el valor mínimo del problema primal, entonces $L_*(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leq F_*$ para cualquier $\boldsymbol{\alpha} \geq 0$ y cualquiera $\boldsymbol{\beta}$. Esta propiedad se llama *dualidad débil*. El programa dual lagrangiano determina así el mejor límite inferior de F_*. Si d_* es el valor óptimo para el problema dual, entonces $d_* \leq F_*$. La diferencia $F_* - d_*$ se llama el *brecha de dualidad*.

La brecha de dualidad es extremadamente útil para proporcionar límites inferiores para las soluciones de problemas primarios que pueden ser imposibles de resolver directamente. Es importante tener en cuenta que para los problemas con restricciones lineales, si el primario no es factible (no tiene una solución que satisfaga las restricciones), entonces el dual es no factible o ilimitado. Por el contrario, si el dual es inviable, entonces el primal no tiene solución. De crucial importancia es la *fuerte dualidad* teorema, que establece que para programas convexos (1.66) con funciones lineales restringidas *h* y *g* ramo la brecha de dualidad es cero, y cualquier $\mathbf{X}_*$ y $(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*)$ que satisfacen las condiciones KKT son soluciones (globales) para los programas primal y dual, respectivamente. En particular, esto es válido para los programas cuadráticos lineales y convexos (tenga en cuenta que no todos los programas cuadráticos son convexos).

Para un programa primal convexo con C^1 funciones objetivo y de restricción, la función dual lagrangiana (1.70) se puede obtener simplemente estableciendo el gradiente (con respecto a $\mathbf{X}$) del lagrangiano $L(\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ a cero. Se puede simplificar aún más el programa dual sustituyendo en el Lagrangiano las relaciones entre las variables así obtenidas.

EJEMPLO 1.17 Problema de Programación Lineal

Considere el siguiente problema de programación lineal:

$$\begin{aligned} \min_{\mathbf{x}} \quad & \mathbf{cx} \\ \text{sujeto a:} \quad & A\mathbf{X}\text{-}\mathbf{b}. \end{aligned}$$

El lagrangiano es $L(\mathbf{X},\boldsymbol{\alpha}) = \mathbf{cx} - \boldsymbol{\alpha}(A\mathbf{X} - \mathbf{b})$. La función dual de Lagrange es el mínimo de L general $\mathbf{X}$; de este modo

$$L_*(\boldsymbol{\alpha}) = \begin{cases} \mathbf{b}\boldsymbol{\alpha} & \text{si } A\boldsymbol{\alpha} = \mathbf{C}, \\ -\infty & \text{de lo contrario,} \end{cases}$$

para que el programa dual de Lagrange se convierta en

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \mathbf{b} \quad \boldsymbol{\alpha} \\ \text{sujeto a:} \quad & A \quad \boldsymbol{\alpha} = \mathbf{C}, \\ & \boldsymbol{\alpha}\text{-}\mathbf{0}. \end{aligned}$$

Un hecho interesante a tener en cuenta aquí es que para el problema de programación lineal, el dual del problema dual siempre devuelve el problema primal.

EJEMPLO 1.18 Problema de programación cuadrática

Considere el siguiente problema de programación cuadrática:

$$\begin{aligned} \min_{\mathbf{x}} \quad & \frac{1}{2}\mathbf{X}C\mathbf{X} \\ \text{sujeto a:} \quad & C\mathbf{X}\text{-}\mathbf{b}, \end{aligned}$$

donde el $norte \times norte$ matriz C se supone que es definida positiva (para un problema general de programación cuadrática, la matriz C siempre se puede suponer que es simétrica, pero no es necesariamente definida positiva). El lagrangiano es $L(\mathbf{X},\boldsymbol{\alpha}) = \frac{1}{2}\mathbf{X}C\mathbf{X} - \boldsymbol{\alpha}(C\mathbf{X} - \mathbf{b})$. Podemos minimizar esto tomando su gradiente con respecto a $\mathbf{X}$ y ponerlo a cero. Esto da $C\mathbf{X} - C\boldsymbol{\alpha} = C(\mathbf{X} - \boldsymbol{\alpha}) = \mathbf{0}$. La certeza positiva de C implica que $\mathbf{X} = \boldsymbol{\alpha}$. La maximización de la Lagrangiano ahora se reduce a maximizar $L(\boldsymbol{\alpha},\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}C\boldsymbol{\alpha} - \boldsymbol{\alpha}(C\boldsymbol{\alpha} - \mathbf{b}) = -\frac{1}{2}\boldsymbol{\alpha}C\boldsymbol{\alpha} + \boldsymbol{\alpha}\mathbf{b}$ sujeto a $\boldsymbol{\alpha}\text{-}\mathbf{0}$. Por lo tanto, podemos escribir el problema dual como

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & -\frac{1}{2}\boldsymbol{\alpha}C\boldsymbol{\alpha} + \boldsymbol{\alpha}\mathbf{b} \\ \text{sujeto a:} \quad & \boldsymbol{\alpha}\text{-}\mathbf{0}. \end{aligned}$$

Observe que el problema dual involucra solo restricciones simples de no negatividad.

Ahora supongamos que nos dan la factorización de Cholesky $C = cama\ y\ desayuno$. Resulta (vea el problema 1.36) que el dual de Lagrange del problema dual anterior se puede escribir como

$$\begin{aligned} \min_{\boldsymbol{m}} \quad & \frac{1}{2}\boldsymbol{\mu}\boldsymbol{\mu} \\ \text{sujeto a:} \quad & B\boldsymbol{m}\text{-}\mathbf{b}, \end{aligned} \tag{1.71}$$

con $\boldsymbol{m} = B\boldsymbol{\alpha}$. Este es un llamado *menor distancia* problema, el cual, siempre que conozcamos la factorización de Cholesky de C, es más fácil de resolver que el problema original de programación cuadrática.

Un último ejemplo de dualidad lo proporciona el ampliamente utilizado *método de mínima entropía cruzada* [9].

EJEMPLO 1.19 Método de entropía cruzada mínima (MinxEnt)

Dejar $\mathbf{X}$ ser una variable aleatoria discreta (o vector) tomando valores $\mathbf{X}_1, \ldots, \mathbf{X}_r$, y deja $\mathbf{q} = (q_1, \ldots, q_r)$ y $\mathbf{pags} = (pags_1, \ldots, pags_r)$ sean dos vectores de distribución (columna) estrictamente positivos para $\mathbf{X}$. Considere la fo $\sum_{i=1}^{metro} pags_i \ln(pags_i/q_i)$, siguiendo el programa primario de minimizando la entropía cruzada de **pags** y **q**, eso es, **q** , sujeto a restricciones de igualdad lineal:

$$\min_{\mathbf{pags}} \sum_{k=1}^{r} pags_k \ln \frac{pags_k}{q_k} \tag{1.72}$$

$$\text{sujeto a:} \quad mi_{\mathbf{pags}}[S_i(\mathbf{X})] = \sum_{k=1}^{r} S_i(\mathbf{X}_k) pags_k = y_i, \qquad i = 1, \ldots, metro \tag{1.73}$$

$$\sum_{k=1}^{r} pags_k = 1, \tag{1.74}$$

dónde $S_1, \ldots, S_{metro}$ son funciones arbitrarias.

Aquí la función objetivo es convexa, ya que es una combinación lineal de funciones de la forma $pags \ln(ordenador\ personal)$, que son convexos en R_+, para cualquier $do > 0$. Además, las funciones de restricción de igualdad son afines (de la forma $\mathbf{ap} - y$). Por lo tanto, este problema es convexo. Para obtener la solución óptima **pags*** del programa primario anterior, normalmente es más fácil resolver el problema asociado *doble* programa [9]. A continuación presentamos el procedimiento correspondiente.

1. El Lagrangiano del problema primal viene dado por

$$L(\mathbf{pags}, \boldsymbol{\lambda}, \beta) = \sum_{k=1}^{r} pags_k \ln \frac{pags_k}{q_k} + \sum_{i=1}^{metro} \lambda_i \left(\sum_{k=1}^{r} S_i(\mathbf{X}_k) pags_k - y_i \right) + \beta \left(\sum_{k=1}^{r} pags_k - 1 \right), \tag{1.75}$$

dónde $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{metro})$ es el vector multiplicador de Lagrange correspondiente a (1.73) y β es el multiplicador de Lagrange correspondiente a (1.74). Tenga en cuenta que podemos usar un signo más o menos en la segunda suma de (1.75). Elegimos este último porque luego generalizamos el mismo problema a la desigualdad. (-) restricciones en (1.73), dando lugar a un signo menos en el Lagrangiano.

2. Resolver (para fijo $\boldsymbol{\lambda}$ y β)

$$\min_{\mathbf{pags}} L(\mathbf{pags}, \boldsymbol{\lambda}, \beta) \tag{1.76}$$

resolviendo

$$\nabla_{\mathbf{pags}} L(\mathbf{pags}, \boldsymbol{\lambda}, \beta) = \mathbf{0},$$

que da el conjunto de ecuaciones

$$\nabla_{pags_k} L(\mathbf{pags}, \boldsymbol{\lambda}, \beta) = \ln \frac{pags_k}{q_k} + 1 - \sum_{i=1}^{metro} \lambda_i S_i(\mathbf{X}_k) + \beta = 0, \qquad k = 1, \ldots, r.$$

Denote la solución óptima y el valor de la función óptima obtenidos del programa (1.76) como **pags**$(\boldsymbol{\lambda}, \beta)$ y $L^*(\boldsymbol{\lambda}, \beta)$, respectivamente. Esta última es la función dual de Lagrange. Entonces escribimos

$$pags_k(\boldsymbol{\lambda}, \beta) = q_k \text{Exp}\left(-\beta - 1 + \sum_{i=1}^{metro} \lambda_i S_i(\mathbf{X}_k)\right), k = 1, \ldots, r. \tag{1.77}$$

Dado que la suma de los $\{pags_k\}$ debe ser 1, obtenemos

$$\text{mi}_{\beta} = \sum_{k=1}^{r} q_k \text{Exp}\left(-1 + \sum_{i=1}^{metro} \lambda_i S_i(\mathbf{X}_k)\right). \tag{1.78}$$

Sustituyendo **pags**$(\boldsymbol{\lambda}, \beta)$ de vuelta al Lagrangiano da

$$L^*(\boldsymbol{\lambda}, \beta) = -1 + \sum_{i=1}^{metro} \lambda_i y_i - \beta. \tag{1.79}$$

3. Resuelve el *doble* programa

$$\underset{\boldsymbol{\lambda}, \beta}{\text{máximo}} L^*(\boldsymbol{\lambda}, \beta). \tag{1.80}$$

Ya que β y $\boldsymbol{\lambda}$ están relacionados a través de (1.78), podemos resolver (1.80) sustituyendo el correspondiente $\beta(\boldsymbol{\lambda})$ en (1.79) y optimizando la función resultante:

$$D(\boldsymbol{\lambda}) = -1 + \sum_{i=1}^{metro} \lambda_i y_i - \text{en}\left\{\sum_{k=1}^{r} q_k \text{Exp}\left\{-1 + \sum_{i=1}^{metro} \lambda_i S_i(\mathbf{X}_k)\right\}\right\}. \tag{1.81}$$

Ya que $D(\boldsymbol{\lambda})$ es continuamente diferenciable y cóncava con respecto a $\boldsymbol{\lambda}$, podemos derivar la solución óptima, $\boldsymbol{\lambda}^*$, al resolver

$$\nabla_{\lambda} D(\boldsymbol{\lambda}) = \mathbf{0}, \tag{1.82}$$

que se puede escribir por componentes en la siguiente forma explícita:

$$\begin{aligned} \nabla_{\lambda_j} D(\boldsymbol{\lambda}) &= y_i - \frac{\sum_{k=1}^{r} S_i(\mathbf{X}_k) q_k \text{Exp}\left\{-1 + \sum_{j=1}^{metro} \lambda_j S_j(\mathbf{X}_k)\right\}}{\sum_{k=1}^{r} q_k \text{Exp}\left\{-1 + \sum_{j=1}^{metro} \lambda_j S_j(\mathbf{X}_k)\right\}} \\ &= y_i - \frac{\text{mi}_{\mathbf{q}} S_i(\mathbf{X}) \text{Exp}\left[\left\{-1 + \sum_{j=1}^{metro} \lambda_j S_j(\mathbf{X})\right\}\right]}{\text{mi}_{\mathbf{q}} \text{Exp}\left[\left\{-1 + \sum_{j=1}^{metro} \lambda_j S_j(\mathbf{X})\right\}\right]} = 0 \end{aligned} \tag{1.83}$$

por $j = 1, \ldots, metro$. El vector óptimo $\boldsymbol{\lambda}^* = (\lambda^*_1, \ldots, \lambda^*_{metro})$ puede ser encontrado por resolviendo (1.83) numéricamente. Tenga en cuenta que si el programa primal tiene una solución óptima interior no vacía, entonces el programa dual tiene una solución óptima $\boldsymbol{\lambda}^*$.

4. Finalmente, sustituya $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$ y $\beta = \beta(\boldsymbol{\lambda}^*)$ de nuevo en (1.77) para obtener la solución del programa MinxEnt original.

Es importante señalar que no es necesario imponer explícitamente las condiciones $pags_i > 0$, $i = 1, \ldots, norte$, porque las cantidades $\{pags_i\}$ en (1.77) son automáticamente estrictamente positivos. Esta es una propiedad crucial de la distancia CE; ver también

[1]. Es instructivo (vea el problema 1.37) verificar cómo la adición de las restricciones de no negatividad afecta el procedimiento anterior.

Cuando las restricciones de desigualdadmi$_{\mathbf{pags}}$[S(**X**)] -$y_i$se usan en (1.73) en lugar de restricciones de igualdad, el procedimiento de solución sigue siendo casi el mismo. La única diferencia es que el vector multiplicador de Lagrange$\boldsymbol{\lambda}$ahora debe ser no negativo. De ello se deduce que el programa dual se convierte en

$$\max_{\boldsymbol{\lambda}} \quad D(\boldsymbol{\lambda})$$
$$\text{sujeto a:} \quad \boldsymbol{\lambda} \geq \mathbf{0},$$

con$D$($\boldsymbol{\lambda}$) dado en (1.81).

Una generalización adicional es reemplazar el problema de optimización discreta anterior con un*funcional*problema de optimizacion. Este tema se discutirá en el Capítulo 8. En particular, la Sección 8.9 trata sobre el método MinxEnt, que involucra un problema funcional de MinxEnt.

PROBLEMAS

Teoría de probabilidad

1.1 Demuestre los siguientes resultados, usando las propiedades de la medida de probabilidad en la Definición 1.2.1 (aquíAyBson eventos):

a)PAGS(A^c) = 1 –PAGS(A).

b)PAGS($A \cup B$) =PAGS(A) +PAGS(B)–PAGS($A \cap B$).

1.2 Demuestre la regla del producto (1.4) para el caso de tres eventos.

1.3 Sacamos tres bolas consecutivamente de un bol que contiene exactamente cinco bolas blancas y cinco negras, sin volver a colocarlas. ¿Cuál es la probabilidad de que todas las bolas extraídas sean negras?

1.4 Considere el experimento aleatorio en el que lanzamos una moneda sesgada hasta que sale cara. Suponga que la probabilidad de cara en cualquier lanzamiento es$pags$. Dejar$X$Sea el número de lanzamientos necesarios. Muestra esa$X$~GRAMO($pags$).

1.5 En una sala con mucha gente, le preguntamos a cada persona su cumpleaños (día y mes). Dejar$norte$ sea el número de personas consultadas hasta que obtengamos un cumpleaños "duplicado".

a)CalcularPAGS($norte > norte$),$norte$=0,1,2, ...

b)Para cual$norte$tenemosPAGS($norte$-$norte$) - 1/2?

C)Usa una computadora para calcularMI[$norte$].

1.6 Dejar$X$yYsean variables aleatorias normales estándar independientes, y seantuyV ser variables aleatorias que se derivan de$X$yYmediante la transformada lineal mación

$$\begin{pmatrix} tu \\ V \end{pmatrix} = \begin{pmatrix} \text{pecado}\alpha & -\text{porque}\alpha \\ \text{porque}\alpha & \text{pecado}\alpha \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}.$$

a)Derive el pdf conjunto detuyV.

b)Muestra esatuyVson independientes y normalmente distribuidos estándar.

1.7 Dejar$X$~Exp(λ). Muestre que el*propiedad sin memoria*tiene: para todos$calle$-0,

$$\text{PAGS}(X > t+s \mid X > t) = \text{PAGS}(X > s).$$

1.8 Dejar X_1, X_2, X_3 sean variables aleatorias de Bernoulli independientes con probabilidades de éxito 1/2, 1/3 y 1/4, respectivamente. Dé su pdf conjunta condicional, dado que $X_1+X_2+X_3=2$.

1.9 Verifique las expectativas y las variaciones en la Tabla 1.3.

1.10 Dejar X y Y tener densidad articular F dada por

$$F(x, y) = cxy, 0 \le y \le x \le 1.$$

a) Determinar la constante de normalización C.
b) Determinar PAGS($X+2Y \ge 1$).

1.11 Dejar $X \sim \text{Exp}(\lambda)$ y $Y \sim \text{Exp}(m)$ Se independiente. Muestra esa

a) $\min(X, Y) \sim \text{Exp}(\lambda+m)$,

b) $\text{PAGS}(X < Y \mid \min(X, Y)) = \dfrac{\lambda}{\lambda+m}$.

1.12 Verifique las propiedades de varianza y covarianza en la tabla 1.4.

1.13 Demuestre que el coeficiente de correlación siempre se encuentra entre −1 y 1. [Sugerencia: utilice el hecho de que la varianza de *hacha*+ Y siempre es no negativo, para cualquier a.]

1.14 Considere los ejemplos 1.1 y 1.2. Definir X como la función que asigna el número $X_1+\cdots+X_{norte}$ a cada resultado $\omega = (X_1, \ldots, X_{norte})$. El evento de que haya exactamente k cabezas adentro $norte$ los lanzamientos se pueden escribir como

$$\{\omega \in \Omega : X(\omega) = k\}.$$

Si lo abreviamos a $\{X=k\}$, y abreviar aún más PAGS($\{X=k\}$) a PAGS($X=k$), entonces obtenemos exactamente (1.7). Verifique que siempre se pueden ver las variables aleatorias de esta manera, es decir, como funciones de valor real en Ω, y que las probabilidades como PAGS($X \le x$) debe interpretarse como PAGS($\{\omega \in \Omega : X(\omega) \le x\}$).

1.15 Muestra esa

$$\text{Var}\left(\sum_{i=1}^{norte} X_i\right) = \sum_{i=1}^{norte} \text{Var}(X_i) + 2 \sum_{yo<j} \text{cov}(X_i, X_j).$$

1.16 Sea Σ la matriz de covarianza de un vector columna aleatorio **X**. Escribe **Y** = **X** − ***m***, dónde ***m*** es el vector de esperanza de **X**. Por lo tanto Σ = MI[**AA**]. Demuestre que Σ es semidefinido positivo. Es decir, para cualquier vector **tu**, tenemos **tu**Σ**tu** ≥ 0.

1.17 Suponer $Y \sim \text{Gama}(norte, 1)$. Demuestra que para todos $x \ge 0$

$$\text{PAGS}(Y \le x) = 1 - \sum_{k=0}^{norte-1} mi^{-\lambda x} \frac{(\lambda x)^k}{k!}. \tag{1.84}$$

1.18 Considere el experimento aleatorio donde dibujamos de manera uniforme e independiente $norte$ números, $X_1, \ldots, X_{norte}$, del intervalo [0,1].

a) Dejar $METRO$ ser el más pequeño de los $norte$ números. Expresar $METRO$ en términos de $X_1, \ldots, X_{norte}$.

b) Determinar el pdf de $METRO$.

1.19 Dejar $Y = mi^{X}$, dónde $X \sim$ NORTE(0,1).

a) Determinar el pdf de Y.

b) Determine el valor esperado de Y.

1.20 Seleccionamos un punto (X, Y) del triangulo (0,0)–(1,0)–(1,1) de tal manera que X tiene una distribución uniforme en (0,1) y la distribución condicional de Y dado $X = X$ es uniforme en $(0, X)$.

a) Determine la función de densidad de probabilidad conjunta de X y Y.

b) Determinar el pdf de Y.

C) Determine la función de densidad de probabilidad condicional de X dado $Y = y$ para todos $y \in (0,1)$.

d) Calcular MI[X / $Y = y$] para todos $y \in (0,1)$.

mi) Determinar las expectativas de X y Y.

Procesos de veneno

1.21 Dejar $\{norte_t, t \geq 0\}$ sea un proceso de Poisson con tasa $\lambda = 2$. Encuentra

a) PAGS($norte_2 = 1$, $norte_3 = 4$, $norte_5 = 5$),

b) PAGS($norte_4 = 3$ / $N_2 = 1$, $norte_3 = 2$),

C) MI[$norte_4$ / $norte_2 = 2$],

d) PAGS($norte[2,7] = 4$, $norte[3,8] = 6$),

mi) MI[$norte[4,6]$ / $norte[1,5] = 3$].

1.22 Demostrar que para cualquier fijo $k \in$ NORTE, $t > 0$ y $\lambda > 0$,

$$\lim_{norte \to \infty} \binom{norte}{k} \left(\frac{\lambda t}{norte}\right)^{k} \left(1 - \frac{\lambda t}{norte}\right)^{n-k} = \frac{(\lambda t)^{k}}{k!} mi^{-\lambda t}.$$

[Sugerencia: escriba el coeficiente binomial y use el hecho de que $\lim_{norte \to \infty} \left(1 - \frac{\lambda Terntesse}{norte}\right) = mi^{-\lambda t}$.]

1.23 Considere la aproximación de Bernoulli en la Sección 1.12. Dejar $tu_1, tu_2, \ldots$ denotan los tiempos de éxito para el proceso de Bernoulli X.

a) Verifique que los tiempos de "interéxito" $tu_1, tu_2 - tu_1, \ldots$ son independientes y tienen una distribución geométrica con parámetro $pags = \lambda h$.

b) Para pequeños h y $norte = t/h$, demostrar que la relación PAGS($A_1 > t$) ≈ PAGS($tu_1 > norte$) lleva en el límite, como $norte \to \infty$, a

$$\text{PAGS}(A_1 > t) = mi^{-\lambda t}.$$

1.24 Si $\{norte_t, t \geq 0\}$ es un proceso de Poisson con tasa λ, demuestre que para $0 < tu < t$ y $j = 0,1,2, \ldots, norte$,

$$\text{PAGS}(norte_{tu} = j \mid N_t = norte) = \binom{norte}{jt} \left(\frac{tu}{t}\right)^{n-j},$$

es decir, la distribución condicional de $norte_{tu}$ dado $norte_t = norte$ es binomial con parámetros $norte$ y $Utah$.

Procesos de Markov

1.25 Determine el pdf (discreto) de cada X_{norte}, $norte$=0,1,2, . . ., para el paseo aleatorio del ejemplo 1.10. Además, calcule MI[X_{norte}] y la varianza de X_{norte} para cada $norte$.

1.26 Dejar $\{X_{norte},\ norte \in \text{norte}\}$ ser una cadena de Markov con espacio de estado $\{0,1,2\}$, matriz de transición

$$PAGS = \begin{pmatrix} 0.3 & 0.1 & 0.6 \\ 0.4 & 0.4 & 0.2 \\ 0.1 & 0.7 & 0.2 \end{pmatrix},$$

y distribución inicial π = (0.2,0.5,0.3). Determinar

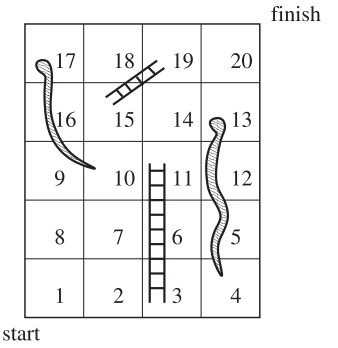
a) PAGS(X_1=2),
b) PAGS(X_2=2),
C) PAGS(X_3=2 / X_0=0),
d) PAGS(X_0=1 / X_1=2),
mi) PAGS(X_1=1, X_3=1).

1.27 Dos perros albergan un número total de $metro$ Floridafácil Spot inicialmente tiene b Floridaeas y Lassie tiene el resto $m-b$. Las pulgas han acordado la siguiente política de inmigración: en todo momento $norte$=1,2. . ., se selecciona al azar una pulga de la población total y esa pulga saltará de un perro a otro. Describe la población de pulgas en Spot como una cadena de Markov y encuentra su distribución estacionaria.

1.28 Clasifique los estados de la cadena de Markov con la siguiente matriz de transición:

$$PAGS = \begin{pmatrix} 0.0 & 0.3 & 0.6 & 0.0 & 0.1 \\ 0.0 & 0.3 & 0.0 & 0.7 & 0.0 \\ 0.3 & 0.1 & 0.6 & 0.0 & 0.0 \\ 0.0 & 0.1 & 0.0 & 0.9 & 0.0 \\ 0.1 & 0.1 & 0.2 & 0.0 & 0.6 \end{pmatrix}.$$

1.29 Considere el siguiente juego de serpientes y escaleras. Dejar $norte$ sea el número de lanzamientos requeridos para llegar al final usando un dado justo. Calcular la esperanza de $norte$ usando una computadora.



1.30 La Sra. Ella Brum camina de un lado a otro entre su casa y su oficina todos los días. Tiene tres paraguas, que se distribuyen en dos paragüeros (uno en casa y otro en el trabajo). Cuando no llueve, la Sra. Brum camina sin paraguas. Cuando llueve, toma un paraguas del puesto en el lugar de

su partida, siempre que haya uno disponible. Suponga que la probabilidad de que esté lloviendo en el momento de cualquier salida es p. Dejar X_n indique el número de sombrillas disponibles en el lugar donde Ella llega después del número de caminata n; $n = 1, 2, \ldots$, incluido el que posiblemente traiga consigo. Calcula la probabilidad límite de que llueva y no haya paraguas disponible.

1.31 Se suelta un ratón en el laberinto de la figura 1.10. De cada compartimento, el ratón elige uno de los compartimentos adyacentes con igual probabilidad, independientemente del pasado. El ratón pasa una cantidad de tiempo distribuida exponencialmente en cada compartimento. El tiempo medio de permanencia en cada uno de los compartimentos 1, 3 y 4 es de dos segundos; el tiempo medio de permanencia en los compartimentos 2, 5 y 6 es de cuatro segundos. Dejar $\{X_t, t \geq 0\}$ sea el proceso de salto de Markov que describe la posición del ratón por veces $t \geq 0$. Suponga que el mouse comienza en el compartimento 1 en el momento $t = 0$.

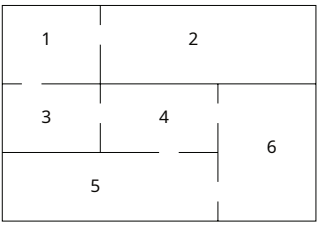


Figura 1.10: Un laberinto.

¿Cuáles son las probabilidades de que el ratón se encuentre en cada uno de los compartimentos $1, 2, \ldots, 6$ en algún momento t muy lejos en el futuro?

1.32 en un $M/M/\infty$-sistema de colas, los clientes llegan de acuerdo con un proceso de Poisson con tasa a. Cada cliente que ingresa es inmediatamente atendido por uno de una infinidad de servidores; por lo tanto, no hay cola. Los tiempos de servicio están distribuidos exponencialmente, con media $1/b$. Todos los tiempos de servicio y entre llegadas son independientes. Dejar X_t sea el número de clientes en el sistema en ese momento t. Demuestre que la distribución límite de X_t, como $t \to \infty$, es Poisson con parámetro a/b.

Mejoramiento

1.33 Dejar **a** y deja **X** ser n-vectores columna dimensionales. Muestra esa $\nabla_{\mathbf{x}}\mathbf{hacha} = \mathbf{a}$.

1.34 Dejar A ser un simétrico $n \times n$ matriz y **X** frijol n-vector columna dimensional. Muestra eso $\frac{1}{2}\nabla_{\mathbf{X}}\mathbf{X}'A\mathbf{X} = A\mathbf{X}$. ¿Cuál es el gradiente si A no es simetrico?

1.35 Demuestre que la distribución óptima **pags** * en el ejemplo 1.16 está dada por la distribución uniforme.

1.36 Derive el programa (1.71).

1.37 Considere el programa MinxEnt

$$\min_{\mathbf{pags}} \sum_{i=1}^{norte} pags_i \ln \frac{pags_i}{q_i}$$

$$\text{sujeto a: } \mathbf{pags} \geq \mathbf{0}, \quad A\mathbf{pags} = \mathbf{b}, \quad \sum_{i=1}^{norte} pags_i = 1,$$

dónde **pags** y **q** son vectores de distribución de probabilidad y A es un *metro* × *norte* matriz.

a) Demuestre que el Lagrangiano para este problema es de la forma

$$L(\mathbf{pags}, \boldsymbol{\lambda}, \beta, \boldsymbol{\mu}) = \mathbf{pags}^{\top}\ln(\mathbf{pags}) - \boldsymbol{\lambda}^{\top}(A\mathbf{pags} - \mathbf{b}) - \boldsymbol{\mu}\mathbf{pags} + \beta(\mathbf{1}^{\top}\mathbf{pag} - 1).$$

b) Muestra esa $pags_i = q_i \exp(-\beta - 1 + \mu_i + \sum_{j=1}^{metro} \lambda_j a_{ji})$, por $i = 1, \ldots, norte$.

C) Explique por qué, como resultado de las condiciones KKT, el óptimo $\boldsymbol{\mu}^*$ debe ser igual al vector cero.

d) Demuestre que la solución para este programa MinxEnt es exactamente la misma que para el programa donde se omiten las restricciones de no negatividad.

Otras lecturas

Una introducción fácil a la teoría de la probabilidad con muchos ejemplos es [13], y un libro de texto más detallado es [8]. Una referencia clásica es [6]. En [3] se da un tratamiento preciso y accesible de varios procesos estocásticos. Para la optimización convexa nos referimos a [2] y [7].

REFERENCIAS

1. ZI Botev, DP Kroese y T. Taimre. Métodos generalizados de entropía cruzada para la simulación y optimización de eventos raros. *Simulación: transacciones de la Society for Modeling and Simulation International*, 83(11):785–806, 2007.
2. S. Boyd y L. Vandenberghe. *Optimizacion convexa*. Prensa de la Universidad de Cambridge, Cambridge, Reino Unido, 2004.
3. E. Çinlar. *Introducción a los Procesos Estocásticos*. Prentice Hall, Englewood Cliffs, Nueva Jersey, 1975.
4. TM Cover y JA Thomas. *Elementos de la teoría de la información*. John Wiley & Sons, Nueva York, 1991.
5. CW Curtis. *Álgebra lineal: un enfoque introductorio*. Springer-Verlag, Nueva York, 1984.
6. W. Feller. *Una introducción a la teoría de la probabilidad y sus aplicaciones*, volumen 1. John Wiley & Sons, Nueva York, 2ª edición, 1970.
7. R. Fletcher. *Métodos prácticos de optimización*. John Wiley & Sons, Nueva York, 1987.
8. GR Grimmett y DR Stirzaker. *Procesos de probabilidad y aleatorios*. Oxford University Press, Oxford, 3.ª edición, 2001.
9. JN Kapur y HK Kesavan. *Principios de optimización de entropía con aplicaciones*. Prensa académica, Nueva York, 1992.