Capítulo 3

Distancia entre lenguajes estocásticos

Aunque no hubiera azar en este mundo, nuestra ignorancia de la causa real de un suceso tendría la misma influencia sobre el entendimiento.
David Hume, Enquiry concerning the human Understanding.

En este capítulo se introducen algunas herramientas que utilizaremos en el estudio de los métodos de identificación de lenguajes estocásticos. En particular, para evaluar la calidad de los modelos obtenidos encontraremos métodos eficientes para calcular la distancia de Kullback–Leibler entre lenguajes racionales, una medida de la similitud entre distribuciones probabilísticas que se basa en la teoría de la información. El cálculo de la distancia entre modelos de Markov ha sido estudiado anteriormente (Ziv & Merhav 1993, Kesidis & Walrand 1993). En lo que sigue generalizaremos estos procedimientos al caso de los autómatas finitos estocásticos, tanto de cadenas (en la sección 3.2) como de árboles (en la sección 3.4). Además, presentaremos distin-

tos criterios de contraste estadístico para variables aleatorias (3.5) y discutiremos las ventajas e inconvenientes de identificar o estimar una probabilidad (3.6).

3.1 Entropía de un lenguaje estocástico

Dado un lenguaje estocástico L con distribución de probabilidad $p(x|L)$ sobre $\mathcal{A}^*$, se define la entropía del lenguaje como el valor esperado (cambiado de signo) del logaritmo de la función de probabilidad

$$H(L) = -\sum_{x \in \mathcal{A}^*} p(x|L) \log p(x|L) \tag{3.1}$$

donde se toma por convención $0 \log 0 = 0$. Si el logaritmo se toma en base dos, el resultado se expresa en bits. La entropía es siempre un número positivo y constituye una medida del desorden del lenguaje. Por ejemplo, para un lenguaje finito formado por N cadenas equiprobables, la entropía es $\log N$ bits. Este número está relacionado con la longitud media de las cadenas si se utiliza una codificación óptima (recuérdese que el número de mensajes codificables con cadenas binarias de longitud l es 2 elevado a l). También puede interpretarse como el número esperado de preguntas (con respuesta si/no) necesarias para identificar el resultado de un experimento aleatorio realizado a partir de la distribución $p(x|L)$, suponiendo, de nuevo, que la estrategia seguida es óptima.

Es sencillo probar que si un lenguaje consta de N cadenas no equiprobables $\{x_1, x_2, ..., x_N\}$ la entropía es siempre menor que $\log N$. En el caso extremo de que todas las probabilidades son nulas excepto una que es uno, la entropía es trivialmente nula. Teniendo en cuenta que la suma de las probabilidades de todas las cadenas es uno, se trata pues de busca el máximo de la función H sobre la variedad lineal $\sum_{k=1}^{N} p_k = 1$. Introduciendo el correspondiente multiplicador de Lagrange

$$\lambda \sum_{k=1}^{N} \Delta p_k = 0, \tag{3.2}$$

donde λ es un número real arbitrario y Δ representa un diferencial, y derivando H se obtiene que la condición de máximo es

$$\Delta H = \sum_{k=1}^{N} (\Delta p_k \ \log p_k + \Delta p_k - \lambda \Delta p_k) = 0, \tag{3.3}$$

por lo que

$$\sum_{k=1}^{N} \Delta p_k \ (\log p_k + 1 - \lambda) = 0 \tag{3.4}$$

La introducción del multiplicador garantiza que todos los diferenciales Δp_k son independientes y, por tanto, $p_k = \lambda - 1$ para todas las cadenas x_k. Por tanto, el máximo se alcanza en el caso de que todas las cadenas son equiprobables. Esta disminución de la entropía puede ser interpretada en términos de longitud media de la codificación como sigue: la codificación puede ser optimizada si a las cadenas más probables se les asignan los códigos más cortos y a las menos probables los más largos. También el número de preguntas en el problema de la predicción de un resultado puede disminuirse en la misma proporción siguiendo una estrategia de agrupamiento, que divida en cada paso las cadenas en dos clases aproximadamente equiprobables.

Dos lenguajes con la misma entropía no son, en general, idénticos. Sin embargo la magnitud:

$$H(L_1, L_2) = \sum_{x \in \mathcal{A}^*} p(x|L_1) \log \frac{p(x|L_1)}{p(x|L_2)} \tag{3.5}$$

presenta la propiedad de que $H(L_1, L_2) = 0$ si y sólo si $p(x|L_1) = p(x|L_2)$ para todas las cadenas $x \in \mathcal{A}^*$. Esta magnitud es conocida con el nombre de entropía relativa o distancia de Kullback–Leibler, si bien no se trata de una auténtica distancia pues ni es simétrica ni satisface la desigualdad triangular. La entropía relativa indica la penalización (en bits) que se sufre por utilizar una distribución errónea en lugar de la correcta para diseñar la estrategia óptima en los problemas de codificación o de predicción de un resultado.

En lo que sigue, denotaremos mediante $p_L(a|x)$ la probabilidad de observación del símbolo a del alfabeto después del prefijo x, es decir, la probabilidad de a condicionada a la aparición previa de x:

$$p_L(a|x) = \frac{p(xa\mathcal{A}^*|L)}{p(x\mathcal{A}^*|L)}. \tag{3.6}$$

De forma análoga, y consistentemente con la ecuación (2.6), $p_L(\lambda|x)$ representará la probabilidad de que se observe un final de cadena después del prefijo x:

$$p_L(\lambda|x) = \frac{p(x|L)}{p(x\mathcal{A}^*|L)} \tag{3.7}$$

Con estos convenios, la probabilidad, por ejemplo, $p(ab|L)$ para la cadena ab en el lenguaje L satisface:

$$p(ab|L) = p_L(a|\lambda)\; p_L(b|a)\; p_L(\lambda|ab) \tag{3.8}$$

y su logaritmo correspondiente es, por tanto,

$$\log p(ab|L) = \log p_L(a|\lambda) + \log p_L(b|a) + \log p_L(\lambda|ab) \tag{3.9}$$

Es decir, en el cálculo de la entropía (3.1) encontraremos el término $\log p_L(b|a)$ en todos los sumandos asociados a cadenas que empiezan por ab. En general, el factor $\log p_L(a|x)$, con $a \in \mathcal{A}$, aparece para todas las cadenas que empiezan por xa, que denotaremos como $xa\mathcal{A}^*$, mientras que el factor $\log p_L(\lambda|x)$ sólo multiplica a $p(x|L)$ y podemos escribir:

$$H(L) = -\sum_{x\in\mathcal{A}^*}\sum_{a\in\mathcal{A}} p(xa\mathcal{A}^*|L)\log p_L(a|x) - \sum_{x\in\mathcal{A}^*} p(x|L)\log p_L(\lambda|x). \tag{3.10}$$

Usando (3.6) y (3.7), podemos reescribir la ecuación anterior en una forma más sencilla:

$$H(L) = -\sum_{x\in\mathcal{A}^*}\sum_{a\in\mathcal{A}^\dagger} p(x\mathcal{A}^*|L)p_L(a|x)\log p_L(a|x) \tag{3.11}$$

donde $\mathcal{A}^\dagger = \mathcal{A}\cup\{\lambda\}$.

Si L es generado por una gramática regular estocástica, existe un autómata asociado $A = (Q, \mathcal{A}, \delta, q_I, p)$ que genera L y $p_L(a|x)$ puede tomar sólo un número finito de valores distintos. De hecho, para todas las cadenas x que satisfacen que $\delta(q_I, x) = q_i$ y para todos los símbolos $a \in \mathcal{A}^\dagger$ se obtiene $p_L(a|x) = p(q_i, a)$. Los diferentes subconjuntos $L_i = \{x \in \mathcal{A}^* : \delta(q_I, x) = q_i\}$ definen una partición en L y, si definimos

$$c_i = \sum_{x \in L_i} p(x\mathcal{A}^*|L), \tag{3.12}$$

obtenemos que la entropía es

$$H(L) = -\sum_{q_i \in Q} \sum_{a \in \mathcal{A}^\dagger} c_i \; p(q_i, a) \log p(q_i, a) \; , \tag{3.13}$$

expresión que puede calcularse fácilmente si se conocen los coeficientes c_i.

Nótese que $\lambda \in L_I$ y además $p(\mathcal{A}^*|L) = 1$. Esto nos permite tratar separadamente el caso especial $x = \lambda$ y escribir

$$c_i = \delta_{iI} + \sum_{x \in \mathcal{A}^*} \sum_{\substack{a \in \mathcal{A}: \\ xa \in L_i}} p(xa\mathcal{A}^*|L) \tag{3.14}$$

donde I es el índice del estado inicial q_I y δ_{ij} es la delta de Kronecker:

$$\delta_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{en caso contrario} \end{cases} \tag{3.15}$$

Como $L = \bigcup_j L_j$ y para cada $x \in L_j$,

$$p(xa\mathcal{A}^*|L) = p(x\mathcal{A}^*|L) p(q_j, a) \; , \tag{3.16}$$

se obtiene

$$c_i = \delta_{iI} + \sum_{j=1}^{|Q|} \sum_{x \in L_j} \sum_{\substack{a \in \mathcal{A}: \\ \delta(q_j, a) = q_i}} p(x\mathcal{A}^*|L) \; p(q_j, a) \tag{3.17}$$

Por tanto, los coeficientes c_i pueden ser calculados resolviendo el sistema de ecuaciones

$$c_i = \sum_{j=1}^{|Q|} \Lambda_{ij}\ c_j + \delta_{iI} \tag{3.18}$$

donde

$$\Lambda_{ij} = \sum_{\substack{a \in \mathcal{A}: \\ \delta(q_j, a) = q_i}} p(q_j, a). \tag{3.19}$$

y $c_i^{[0]} = 0$. La inversión de la matriz Λ_{ij} suele ser costosa, por lo que resulta más eficiente un cálculo iterativo de los coeficientes:

$$c_i^{[t+1]} = \sum_{j=1}^{|Q|} \Lambda_{ij}\ c_j^{[t]} + \delta_{iI} \tag{3.20}$$

Es sencillo comprobar por inducción en t que $c_i^{[t+1]} \geq c_i^{[t]}$ y que al mismo tiempo $c_i^{[t]} \leq c_i$, por lo que el cálculo iterativo converge rápidamente al valor correcto.

3.2 Entropía relativa entre lenguajes

Es posible aplicar un procedimiento semejante al de la sección anterior para calcular la entropía relativa entre dos lenguajes estocásticos L y L', generados por A y A' respectivamente. En este caso,

$$H(L, L') = \sum_{q_i \in Q} \sum_{q'_j \in Q'} \sum_{a \in \mathcal{A}^\dagger} c_{ij}\ p(q_i, a) \log \frac{p(q_i, a)}{p'(q'_j, a)} \tag{3.21}$$

con los coeficientes

$$c_{ij} = \sum_{x \in L_{ij}} p(x\mathcal{A}^* | L), \tag{3.22}$$

donde

$$L_{ij} = \{x \in \mathcal{A}^* : \delta(q_I, x) = q_i \wedge \delta'(q'_{I'}, x) = q'_j\}. \tag{3.23}$$

Los coeficientes c_{ij} pueden ser calculados mediante la relación:

$$c_{ij}^{[t+1]} = \delta_{iI}\delta_{I'j} + \sum_{k=1}^{|Q|} \sum_{l=1}^{|Q'|} \Lambda_{ijkl}\ c_{kl}^{[t]} \tag{3.24}$$

donde

$$\Lambda_{ijkl} = \sum_{\substack{a \in \mathcal{A}: \\ \delta(q_k, a) = q_i \\ \delta'(q'_l, a) = q'_j}} p(q_k, a). \tag{3.25}$$

La expresión anterior para los coeficientes $c_{ij}^{[t+1]}$ puede probarse de forma sencilla siguiendo los mismos pasos de la sección anterior si se observa que $\lambda \in L_{II'}$ y que los antiguos coeficientes c_i satisfacen

$$c_i = \sum_{j=1}^{|Q'|} c_{ij}. \tag{3.26}$$

La figura 3.1 representa la entropía relativa entre dos lenguajes generados por dos gramáticas elegidas al azar, cada una con 10 estados y 30 reglas, ambas con el mismo alfabeto de trabajo $\mathcal{A} = \{0, 1\}$. La línea sólida representa el resultado del cálculo algorítmico mientras que los puntos son los resultados y desviaciones obtenidos cuando la entropía relativa se calcula mediante conjuntos de prueba de distintos tamaños. Se observa que, incluso para gramáticas sencillas como éstas, la convergencia al valor correcto es bastante lenta y que se requieren muestras enormes para que la estimación de la distancia entre los lenguajes sea fiable. Por ello, el procedimiento descrito en esta sección puede ser utilizado para conseguir una comprobación más precisa de los modelos obtenidos mediante inferencia gramatical.

3.3 Entropía de un lenguaje de árboles

En la sección 3.1 vimos cómo la entropía de un lenguaje regular estocástico L puede ser calculada de forma eficiente si se conoce la colección de coeficientes $c_i = \sum_{x \in L_i} p(x\mathcal{A}^*|L)$. Dichos coeficientes pueden interpretarse como el valor esperado del número de nodos de tipo q_i que se utilizan en el análisis de una cadena w elegida al azar según la distribución $p(w|L)$. Es posible realizar un razonamiento análogo para el caso de un lenguaje estocástico racional de árboles T generado

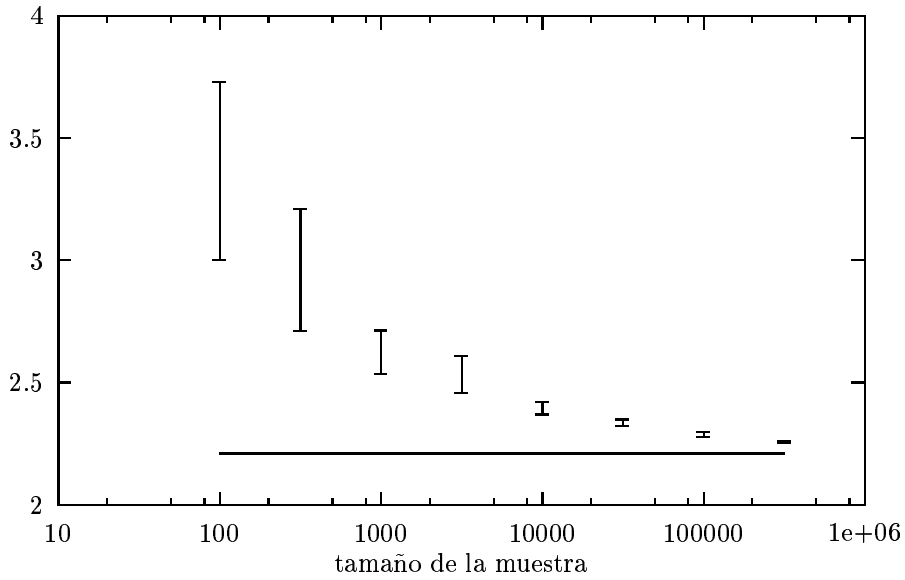

Figura 3.1: Entropía relativa en bits entre dos gramáticas de tamaño 10 generadas al azar. Línea continua: cálculo exacto. Puntos: estimación mediante muestras.

mediante el autómata $A = (Q, V, \delta, p, r)$. En este caso, la probabilidad de generación de un árbol t se compone de dos factores multiplicativos:

- Por un lado, la probabilidad $r(q)$ de que la raíz del árbol sea del tipo $q = \delta(t)$.
- Por otro, la probabilidad $p(t|q)$ de que se genere el árbol t a partir de un nodo del tipo $q = \delta(t)$; esta probabilidad ha de evaluarse recursivamente.

Recordemos que, por ejemplo, si $t = f(t_1, t_2)$, entonces

$$p(t|A) = r(\delta(t))\ p_2(f, \delta(t_1), \delta(t_2))\ p(t_1|\delta(t_1))\ p(t_2|\delta(t_2))\ . \qquad (3.27)$$

La condición de normalización (2.9) garantiza que $p(t|q) = 0$ para todo $q \neq \delta(t)$.

Para simplificar la notación, en lo que sigue denotaremos los estados del autómata indicando únicamente el ordinal correspondiente: $Q = \{1, 2, ..., N\}$, de forma que $q \in Q$ es un natural entre 1 y N. Por otra parte, n denota el máximo número de descendientes que puede tener un nodo en el lenguaje, valor que viene determinado por el conjunto δ de funciones de transición.

La entropía de un lenguaje de árboles T generado por el autómata A es

$$H(T) = -\sum_{t \in V^T} p(t|A) \log p(t|A). \qquad (3.28)$$

En el cálculo de la expresión anterior, el factor $\log r(q)$ multiplica a la probabilidad de todos los árboles t tales que $\delta(t) = q$, por lo que $\log r(q)$ aparecerá multiplicando a

$$\sum_{\substack{t \in V^T \\ \delta(t) = q}} p(t|A) = r(q), \qquad (3.29)$$

es decir, los términos asociados a las probabilidades $r(q)$ contribuyen a la entropía con un sumando

$$H_r(T) = -\sum_{q \in Q} r(q) \log r(q) \qquad (3.30)$$

Por otro lado, si tomamos por ejemplo $p_2(f, q_1, q_2)$, siendo $f \in V$ y $q_1, q_2 \in Q$, el factor $\log p_2(f, q_1, q_2)$ aparecerá multiplicando a la probabilidad de aquellos árboles que contienen un nodo del tipo $q = \delta_2(f, q_1, q_2)$ etiquetado f y que genera dos descendientes del tipo q_1 y q_2 respectivamente. Además, si hubiese m nodos con tales características en el mismo árbol t, $p(t|A)$ aparecería m veces. Es decir, $\log p_2(f, q_1, q_2)$ aparecerá multiplicado por $c_q p_2(f, q_1, q_2)$, siendo c_q el número esperado de nodos de tipo q en un árbol elegido al azar según $p(t|A)$.

Resumiendo, podemos descomponer la entropía del lenguaje T en dos términos

$$H(T) = H_r(T) + H_p(T) \tag{3.31}$$

donde $H_r(T)$ viene dado por la expresión (3.30) y $H_p(T)$ es

$$\begin{aligned} H_p(T) \quad = \quad & \sum_{k=0}^{n} \sum_{f \in V} \sum_{q_1,\, q_2,\, ...,\, q_k \,\in\, Q} c_{\delta_k(f, q_1, q_2, \ldots q_k)} \\ & p_k(f, q_1, q_2, ..., q_k) \log p_k(f, q_1, q_2, ..., q_k) \; . \end{aligned} \tag{3.32}$$

El cálculo de (3.30) y (3.32) es inmediato si se conocen los coeficientes c_q, que pueden ser evaluados de forma iterativa mediante la relación siguiente:

$$c_i^{[t+1]} = r(i) + \sum_{j \in Q} \Lambda_{ij} c_j^{[t]} \tag{3.33}$$

donde

$$\Lambda_{ij} = \sum_{k=0}^{n} \sum_{f \in V} \sum_{\substack{q_1,\, q_2,\, ...,\, q_k \,\in\, Q\,: \\ \delta(f, q_1, ..., q_k) = j}} p_k(f, q_1, q_2, ..., q_k)(\delta_{iq_1} + \delta_{iq_2} + \cdots + \delta_{iq_k}) \tag{3.34}$$

y para $t = 0$, se elige $c_i^{[0]} = 0$.

3.4 Entropía relativa entre lenguajes de árboles

La entropía relativa entre dos lenguajes de árboles T y T', generados por los autómatas $A = (Q, V, \delta, p, r)$ y $A' = (Q', V, \delta', p', r')$ respectivamente, viene dada por

$$H(T, T') = \sum_{t \in V^T} p(t|A) \log \frac{p(t|A)}{p(t|A')} \, . \tag{3.35}$$

Para el cálculo de esta entropía, definimos la probabilidad $\eta_{qq'}$ de que el nodo $q \in Q$ genere un subárbol t tal que $\delta'(t) = q'$.

Usando los coeficientes η_{ij}, podemos escribir la contribución a la entropía relativa de los términos del tipo $\log r'(q')$ de la siguiente forma:

$$-\sum_{t \in V^T} p(t|A) \ \log r'(\delta'(t)) = \sum_{i \in Q} \sum_{j \in Q'} r(i) \ \eta_{ij} \ \log r'(j) \tag{3.36}$$

De esta forma, obtenemos que la entropía relativa es

$$H(T, T') = H(T) + H_r(T, T') + H_p(T, T') \tag{3.37}$$

donde el término $H_r(T, T')$ es

$$H_r(T, T') = -\sum_{i \in Q} \sum_{j \in Q'} \eta_{ij} \ r(i) \log r'(j) \tag{3.38}$$

y el término $H_p(T, T')$ es

$$\begin{aligned} H_p(T, T') \ = \ & -\sum_{k=0}^{n} \sum_{f \in V} \sum_{i_1, i_2, \ldots, i_k \in Q} \sum_{j_1, j_2, \ldots, j_k \in Q'} c_{\delta_k(f, i_1, i_2, \ldots, i_k)} \\ & p_k(f, i_1, i_2, .., i_k) \log p'_k(f, j_1, j_2, ..., j_k) \ \eta_{i_1 j_1} \eta_{i_2 j_j} \cdots \eta_{i_k j_k} \end{aligned} \tag{3.39}$$

En esta última expresión hemos usado los mismos coeficientes c_i definidos en la sección anterior.

Los coeficientes η_{ij} pueden ser calculados de forma sencilla mediante un procedimiento iterativo:

$$\eta_{ij}^{[t+1]} = \sum_{k=0}^{n} \sum_{f \in V} \sum_{\substack{i_1, i_2, \ldots i_k \in Q : \\ \delta(f, i_1, \ldots, i_k) = i}} \sum_{\substack{j_1, j_2, \ldots, j_k \in Q' : \\ \delta'(f, j_1, \ldots, j_k) = j}} p_k(f, i_1, i_2, \ldots, i_k)\, \eta_{i_1 j_1}^{[t]} \eta_{i_2 j_2}^{[t]} \cdots \eta_{i_k j_k}^{[t]} \tag{3.40}$$

donde para $t = 0$ se toma $\eta_{ij}^{[0]} = 0$.

En la figura 3.2 se observa cómo la entropía relativa entre lenguajes de árboles converge muy lentamente al valor correcto cuando éste se estima a partir de muestras. Dicha gráfica representa la entropía relativa de dos gramáticas sencillas (con seis reglas cada una) que generan expresiones aritméticas:

expresión $\rightarrow$ *expresión* $+$ *término*

expresión $\rightarrow$ *término*

término $\rightarrow$ *término factor*

término $\rightarrow$ *factor*

factor $\rightarrow$ **número**

factor $\rightarrow$ (*expresión*)

donde las variables aparecen en cursiva y los terminales en negrita. Los resultados muestran que incluso con muestras enormes, el valor obtenido para la entropía está lejos del correcto. La convergencia es mucho peor que en el caso de los lenguajes de cadenas. Esto puede explicarse si se observa el hecho de que el número de elementos en un lenguaje de árboles es enorme comparado con el caso de las cadenas. Por ejemplo, mientras el número de cadenas binarias de longitud máxima L es $2^{L+1} - 1$, el de árboles binarios (es decir, con dos o cero descendientes por nodo) etiquetados con $V = \{0, 1\}$ y de profundidad

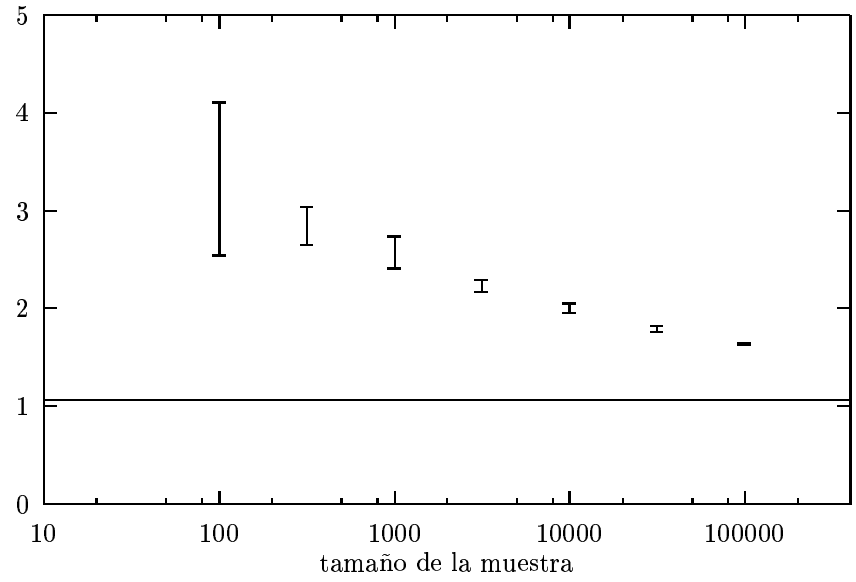

Figura 3.2: Entropía relativa en bits entre dos gramáticas de árboles con seis reglas cada una. Línea continua: cálculo exacto. Puntos: estimación mediante muestras.

máxima d crece más rápido que 2 elevado a 2^d, es decir, mientras el primero crece exponencialmente el segundo lo hace de forma exponencialmente exponencial.

3.5 Convergencia de una variable aleatoria

Los procesos aleatorios presentan una marcada tendencia a la regularidad (la llamada regularidad estadística), sobre todo cuando el número de experimentos es grande. De hecho, es posible establecer distintas cotas superiores para la diferencia entre los valores de una variable aleatoria x obtenidos a partir de una muestra y su valor esperado $\mu = \mathrm{E}(x)$. Una de estas cotas es la expresada por la desigualdad de Chebychev (Feller 1950). Recordemos que, dada un distribución de

probabilidad $f(x)$, el valor esperado de la variable aleatoria x es:

$$\mathrm{E}(x) = \int dx\ f(x)\ x \tag{3.41}$$

En particular, $\mathrm{E}(x-\mu) = 0$, por lo que normalmente se utiliza $(x-\mu)^2$ para medir la dispersión de la variable x. Al valor esperado de $(x-\mu)^2$ se le denomina *varianza* de x:

$$\mathrm{Var}(x) = \int dx\ f(x)(x-\mu)^2 \tag{3.42}$$

Dado que el integrando en(3.42) es siempre positivo, si excluimos de la integral un entorno del valor esperado μ se obtiene la siguiente desigualdad:

$$\begin{aligned} \mathrm{Var}(x) &\geq \int_{|x-\mu|\geq\epsilon} dx\ f(x)(x-\mu)^2 \geq \\ &\geq \int_{|x-\mu|\geq\epsilon} dx\ f(x)\ \epsilon^2 = \epsilon^2\, p\,(|x-\mu| \geq \epsilon) \end{aligned} \tag{3.43}$$

De donde se deduce que la probabilidad α de que la variable difiera de la media en un valor mayor que ϵ esta acotada:

$$\alpha = p\,(|x-\mu| \geq \epsilon)\ \leq\ \frac{\mathrm{Var}(x)}{\epsilon^2} \tag{3.44}$$

Por tanto, $\mathrm{Var}(x)$ es una medida de la variabilidad de x. A la probabilidad α se le llama *nivel de significación* y al valor $1-\alpha$ se le denomina *nivel de confianza*. En general interesa que α sea pequeño (por ejemplo, menor que 0.1). Esto es posible si la varianza es pequeña, o bien si se toma un intervalo suficientemente amplio (es decir, ϵ suficientemente grande).

Con frecuencia, se fija el valor de α y se busca el intervalo más pequeño que satisface (3.44). En ese caso, una forma más conveniente de escribir (3.44) es:

$$p\left(|x-\mu| < \sqrt{\frac{\mathrm{Var}(x)}{\alpha}}\right) > 1-\alpha \tag{3.45}$$

Es decir, el resultado de un experimento de la variable x se encuentra con probabilidad mayor que $1-\alpha$ a una distancia menor que $\epsilon = \sqrt{\frac{1}{\alpha}\mathrm{Var}(x)}$ del valor esperado μ.

Es sencillo comprobar que $\overline{x}$, la media aritmética obtenida después de n experimentos aleatorios de la variable x, es una nueva variable aleatoria con valor esperado $\mathrm{E}(\overline{x}) = \mathrm{E}(x)$ y varianza $\mathrm{Var}(\overline{x}) = \frac{1}{n}\mathrm{Var}(x)$. Por tanto, reescribiendo la ecuación anterior para el caso particular de $\overline{x}$ obtenemos:

$$p\left(|\overline{x}-\mu| < \sqrt{\frac{\mathrm{Var}(x)}{n\alpha}}\right) > 1-\alpha \tag{3.46}$$

Otra cota para esta diferencia viene dada por el límite de Hoeffding (Hoeffding 1963), que es válido para variables de Bernoulli, es decir, aquellas que sólo pueden presentar dos resultados: éxito o fracaso. Sea x una variable de Bernoulli con probabilidad de éxito p y sea

$$\epsilon_\alpha(n) = \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}\,. \tag{3.47}$$

Entonces,

$$p(|\bar{x}-p| < \epsilon_\alpha(n)) > 1-\alpha \tag{3.48}$$

donde $\bar{x} = f_n(x)/n$ y $f_n(x)$ representa el número de éxitos después de n experimentos.

La distribución asociada a una variable de Bernoulli es la distribución binomial que, para grandes valores de n, es aproximadamente una distribución gaussiana o normal. La *distribución normal* para una variable de media cero y varianza uno es

$$p(x>a) = \frac{1}{\sqrt{2\pi}}\int_a^\infty dx\ \exp(-\frac{x^2}{2}) \tag{3.49}$$

y por tanto

$$\begin{aligned} p(x>a) &= \frac{1}{\sqrt{2\pi}}\int_0^\infty dt\ \exp(-\frac{(t+a)^2}{2}) = \\ &= \frac{1}{\sqrt{2\pi}}\int_0^\infty dt\ \exp(-\frac{t^2}{2})\exp(-at)\exp(-\frac{a^2}{2}) \leq \end{aligned}$$

$$\leq \frac{1}{\sqrt{2\pi}} \int_0^\infty dt \, \exp(-\frac{t^2}{2}) \exp(-\frac{a^2}{2}) =$$
$$= \exp(-\frac{a^2}{2}) \tag{3.50}$$

Si definimos $\alpha = p(|x| > \epsilon)$, entonces

$$\epsilon = \sqrt{2 \log \frac{2}{\alpha}} \tag{3.51}$$

es la anchura del intervalo que garantiza que $x \in [-\epsilon, \epsilon]$ con probabilidad mayor que $1 - \alpha$.

Por otro lado, para una variable de Bernoulli x con probabilidad p, la varianza de $\bar{x}$ es $p(1-p)/n$ que es siempre menor o igual que $1/4n$. Por tanto, es sencillo comprobar que el rango asociado a $\bar{x}$ es:

$$\epsilon = \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}} \tag{3.52}$$

que coincide con el valor de la cota (3.47). Importa destacar que la cota de Hoeffding no es únicamente válida cuando n es grande, sino que es válida para cualquier valor de n (Hoeffding 1963).

Otro límite de interés es la regla del logaritmo iterado (Feller 1950), que permite escribir que con probabilidad 1

$$|\bar{x} - p| < \sqrt{\frac{\log \log n}{n}} \tag{3.53}$$

excepto para un número finito de valores de n.

3.6 Aproximación contra identificación

Es llamativo el hecho de que puede definirse un procedimiento para determinar (en el límite) si un número es racional o irracional (Cover 1973). Sin embargo, en un ordenador los números reales son representados mediante conjuntos finitos de bits. Esto hace que, a efectos prácticos, podamos tratar las probabilidades como números racionales. Por otro lado, calcular una probabilidad con un método

que garantice su identificación en el límite no siempre es preferible a simplemente estimar dicha probabilidad. En efecto, si aproximamos la probabilidad $p(x)$ mediante el cociente $f_n(x)/n$, donde $f_n(x)$ es el número de veces que se observa el suceso $x \in \Omega$ en una serie de n experimentos, obtenemos con frecuencia un valor que se acerca al valor real $p(x)$ mucho más rápidamente que el proporcionado por el método que identifica el valor correcto en el límite. Si bien, para un número suficiente de experimentos, la identificación en el límite proporciona el resultado exacto (siempre y cuando la probabilidad sea un número racional), el número de experimentos necesarios para que esto ocurra puede ser muy grande. No debemos olvidar que, en la práctica, sólo disponemos de muestras de tamaño finito y a menudo no demasiado grandes.

Por ejemplo, aunque el algoritmo de la figura 3.3 garantiza la identificación en el límite, de la probabilidad buscada, ésta se produce en general demasiado tarde como para que resulte interesante su utilización en experimentos reales. La comparación de las gráficas 3.6 y 3.7 muestra que, salvo para valores racionales muy sencillos para $p(x)$, la identificación necesita un número de experimentos demasiado grande como para que resulte práctica. La gráfica 3.8, sugiere que en el caso de que la probabilidad sea irracional, puede ser siempre preferible la estimación del valor de $p(x)$.

Esta conclusión se puede generalizar inmediatamente al caso en el que el número de probabilidades a considerar es finito. Sin embargo, en los problemas que estudiaremos más adelante encontraremos que el número de probabilidades a determinar es a menudo infinito. En consecuencia, investigaremos la forma de reducir este número de forma que sea finito.

Para terminar esta sección, justificaremos que el algoritmo 3.3 garantiza la identificación en el límite de la probabilidad $p(x)$ si su valor es un número racional.

En el algoritmo 3.3, la función `next`(q) proporciona el siguiente número racional a uno dado q en el intervalo [0,1], siguiendo una or-

denación inspirada en el método de Cantor que recorre todo $\mathbb{Q} \cup [0,1]$. La función $\texttt{diff}(f, n, q)$ utiliza la prueba del logaritmo iterado (3.53) que garantiza que si $q = p(x)$ es el racional correcto el resultado es incorrecto (devuelve FALSO) para un número finito de valores de n.

Por otro lado, si $q = p(x)$, $q' \neq p(x)$ y

$$|q - q'| > 2\sqrt{\frac{\log\log n}{n}} \tag{3.54}$$

entonces, aplicando la desigualdad triangular a la regla del logaritmo iterado (3.53) se obtiene que

$$\left|\frac{f}{n} - q'\right| > \sqrt{\frac{\log\log n}{n}} \tag{3.55}$$

excepto para un número finito de valores de n. Como el cociente $\log\log n/n$ tiende a cero, la condición (3.54) se satisface siempre que n sea lo suficientemente grande. Por consiguiente, cualquier valor $q' \neq p(x)$ es rechazado para un n suficientemente grande.

Dada una ordenación $\{q_1, q_2, ...\}$ de los racionales en [0,1], sólo existe un número finito de racionales anteriores al valor buscado $q_k = q$, y por ello, sólo hay que seleccionar una cota inferior n_0 para n entre un conjunto finito de valores. Para cualquier $n > n_0$, todos los valores $q' \neq p(x)$ son rechazados y $q = p(x)$ es aceptado, es decir, se alcanza la identificación en el límite de $p(x)$.

```
algorithm identifica_p
input: f (número de observaciones de x)
       n (número de experimentos)
output: q ∈ ℕ × ℕ ≃ ℚ

begin algorithm
  q = (0, 1)
  do ( mientras diff(q, f, n) )
     q = next(q)
  end do
  return q
end algorithm
```

Figura 3.3: Algoritmo que identifica $p(x) \in \mathbb{Q}$.

```
algorithm diff
input: f ∈ ℕ
       n ∈ ℕ
       q = (x, y) ∈ ℕ × ℕ
output: boolean

begin algorithm
  return |f/n − x/y| > √((log log n)/n)
end algorithm
```

Figura 3.4: Algoritmo que compara q y f/n.

algorithm next
input: $q = (x, y) \in \mathbb{N} \times \mathbb{N}$
output: $q' \in \mathbb{N} \times \mathbb{N}$

begin algorithm
$$q' = \begin{cases} (x+1, y-1) & \text{si } x+1 \leq y-1 \\ (1, x+y) & \text{en caso contrario} \end{cases}$$
return q'
end algorithm

Figura 3.5: Siguiente racional a q en $[0, 1]$.

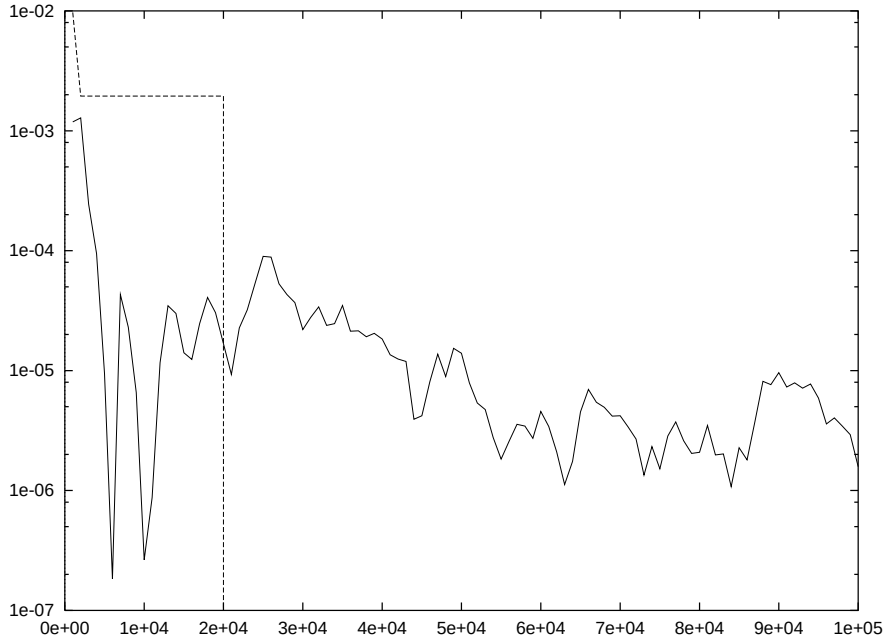

Figura 3.6: Entropía relativa entre $p(x) = 0.875 = \frac{7}{8}$ y la probabilidad experimental en función del número de experimentos. Línea continua: resultado de la estimación numérica. Línea de puntos: algoritmo de identificación. Esta última baja hasta cero a partir de 20000 experimentos.

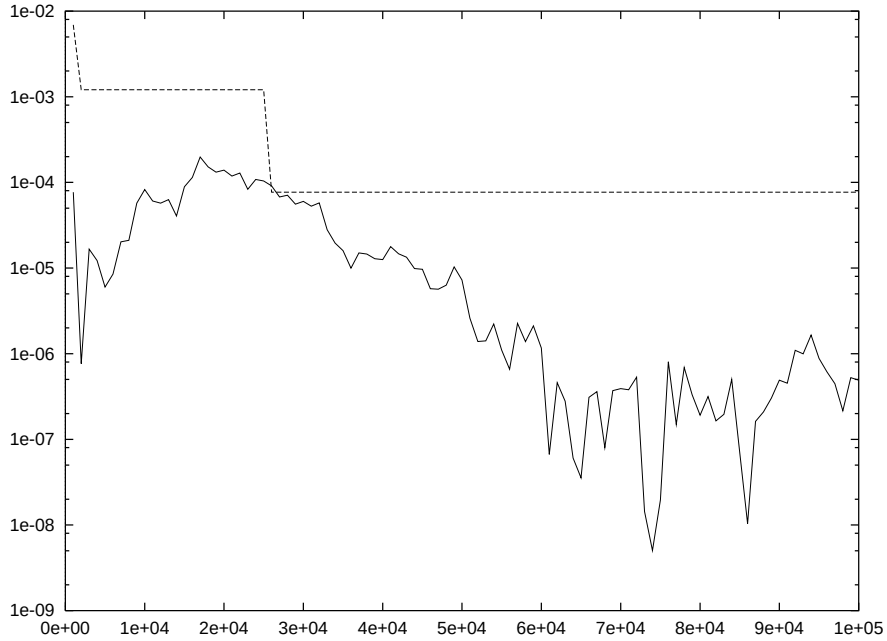

Figura 3.7: Entropía relativa entre $p(x) = 0.62$ y la probabilidad experimental en función del número de experimentos. Línea continua: estimación numérica. Línea de puntos: algoritmo de identificación.

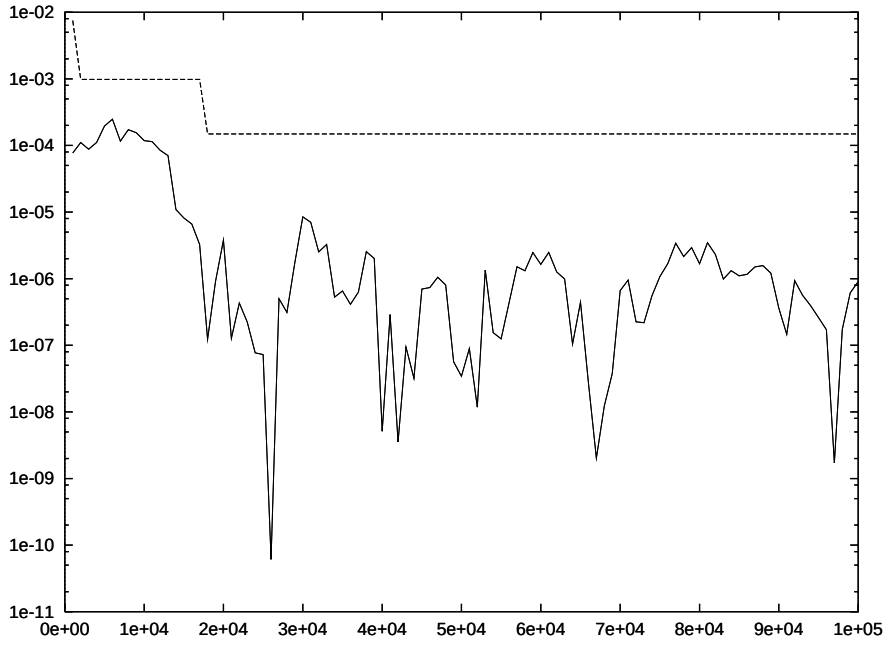

Figura 3.8: Entropía relativa entre $p(x) = \Phi = (-1 + 5^{\frac{1}{2}})/2$ y la probabilidad experimental en función del número de experimentos. Línea continua: estimación numérica. Línea de puntos: algoritmo de identificación.