UNIVERSIDAD DE SONORA

Escuela de Altos Estudios

"ALGUNOS PROBLEMAS RECIENTES EN ESTADISTICA MATEMATICA Y EN PROGRAMACION Y ANALISIS DE PATRONES Y TEORIA DEL HASHING"

TESIS

Que para obtener el Título de

LICENCIADO EN MATEMATICAS

Presenta

Fernando Avila Murillo

Hermosillo, Sonora, México 1975

A casi todo el mundo.

Al pueblo de Sonora.

A los gatos y las lechuzas.

INDICE.

INTRODUCCION

A pesar del auge que están teniendo en nuestros días las ciencias de la computación, la mayoría de las personas aún no saben lo que es una computadora. Inclusive entre los matemáticos. Esto se debe, a que se ve a las computadoras como un producto de la tecnología moderna, y por lo tanto, útil solo a los ingenieros y técnicos. Es cierto que las computadoras se usan, la mayoría de las veces, para realizar tareas engorrosas y rutinarias, en las que el hombre perdería mucho tiempo, y que ellas realizan a una velocidad asombrosa. Se olvida, sin embargo, que las computadoras tambien han motivado a plantear problemas en los que antes ni siquiera se pesnsaba.

El hombre, siempre está procesando información, muchas veces sin darse cuenta. Esta informacion es de índoles muy diversas pero la mayoría de las veces, esto no es un problema, pues hemos sido educados para reconocerla independientemente de la forma que ésta asuma (visual, sonora, etc.). Tomemos el lenguaje, por ejemplo: desde nuestra infancia, aprendimos a reconocer y diferenciar letras y sonidos. Una misma letra, se presentaba en formas muy distintas, impresa, mayúscula, etc., y no teníamos grandes problemas para distinguirla de las demás (con sus excepciones, claro); habíamos aislado los *patrones* de la escritura. Tambien aprendimos el significado de las palabras.

Nuestra evolución consiste en formar patrones mentales cada vez mas complicados; nuestro cerebro es un gran centro

de reconocimiento, análisis y adquisición de información con la cual se forman nuevos patrones.

Cuando una persona pregunta: ¿ pueden pensar las computadoras ?, en realidad lo que está preguntando es: ¿ pueden pensar como nosotros ?, y desde luego, la respuesta es no. La computadora realiza tareas, pero solo aquellas que se le han indicado en un lenguaje adecuado. Este lenguaje debe ser preciso, matemático y fácil de transmitir, y ésto ha constituido la máxima preocupacion de los programadores.

Si queremos ampliar los usos de las computadoras, hay que ampliar los canales de comunicación. Un paso en esta dirección ha sido el llamado *Análisis de Patrones* y mas concretamente, el reconocimiento de ellos. Actualmente las computadoras decifran códigos, reciben órdenes habladas y se les está enseñando a reconocer la escritura manuscrita comun y corriente.

Desde el punto de vista matemático, el planteamiento riguroso de este problema,ha sido muy fértil. Por un lado, se han construido teorias matemáticas del lenguaje, hasta llegar a un *Algebra Linguística*. Por otro lado, han surgido interesantes problemas de optimización y estimación que han enriquecido diversas ramas como la estadística, la combinatoria, etc..

El propósito de esta tesis es dar algunos ejemplos del tipo de matemáticas que se están haciendo en relación a problemas surgidos de cuestiones eminentemente prácticas, así como mostrar someramente, la manera en que los resultados

matemáticos, además de su valor intrínseco, han ido mas allá de las causas que lo produjeron.

Agradezco a todas las personas, son muchas, que me han ayudado a lo largo de mi carrera y durante la elaboración de esta tesis. Una de estas personas, es el matemático Enrique Valle Flores.

CAPITULO 0

0.1. *Funciones de distribución.* Una función de distibución F, es una función real de variable real, no decreciente, no negativa y continua por la izquierda tal que $F(-\infty)= 0$, $F(+\infty)= 1$. Si F puede escribirse en la forma $F(x)= \int_{-\infty}^{x} f(t)dt$, decimos que f es la función de densidad de F.

Sea (Ω,A,P) un espacio de probabilidad fijo y X una variable aleatoria (v.a.) en el; la función $F(x)= P(X\leq x)$ es la función de distribución de X.

Ejemplos. Generalmente es mas fácil dar la función de densidad (f.d.), pues en algunos casos, es imposible la integración exacta.

1) La densidad de Poisson.

$$p(x)= \mu^{x}e^{-\mu}/x \ , \ x=0,1,\ldots.$$

Este es un ejemplo de densidad discreta. μ es el parámetro de la densidad.

2) La densidad Rectangular. $R(\mu, \omega)$

$$f(x)= 1/\omega, \quad \mu-\omega/2<x<\mu+\omega/2$$
$$= 0 \qquad \text{otro caso.}$$

La distribución generada por esta densidad, se llama distribución uniforme.

3) La densidad exponencial negativa.

$$h(x)= \lambda e^{-\lambda x}, \ x>0$$
$$= 0 \quad \text{otro caso.}$$

0.2. *Estimación.* Un estimador de un parámetro θ, es una v.a. observable, determinada a partir de una muestra de la población de la cual θ es parámetro. Un estimador $\bar{\theta}$

es insesgado en la media, si $E(\bar{\theta})= \theta$. Un estimador insesgado de variancia finita que sea de variancia mínima en la clase de los estimadores insesgados, se llama eficiente. Si un estimador (mas correctamente, una familia de estimadores) convergen en probabilidad a θ, se dice que es un estimador consistente.

Si $(x_1,\ldots,x_n)$ es una muestra aleatoria de una población con parámetro θ, la función de verosimilitud de la muestra, es la densidad conjunta de los x_i. El estimador máximo verosimil de θ es el valor que maximiza la función de verosimilitud, viéndola como función de θ.

0,3. En toda la tesis, se usarán las notaciones mas conocidas. Por cuestiones tipográficas, $\underline{x}$ denotará al máximo entero que no excede a x. C_k^n es el número de combinaciones de n objetos tomando k de ellos.

CAPITULO I

I.1. En *Análisis de Patrones*, es útil el siguiente formalismo (Grenander): a partir de ciertas primitivas llamadas *signos* y que pueden ser figuras geómetricas, caracteres alfanuméricos, etc., formamos un conjunto C de *configuraciones legales*, constituido por vectores finitos cuyas componentes son signos y que no violan un conjunto de reglas R. Una relación de equivalencia da origen a clases de equivalencia llamadas *imágenes*. El conjunto resultante de imágenes se llama *Algebra Imagen*. Las imagenes son los objetos que pueden ser observados bajo condiciones ideales y, dependiendo de la manera en que han sido generadas, son clasificadas en *clases de patrones* que forman una familia de patrones.

Por ejemplo, las fotografías enviadas por satélites artificiales y las imágenes de televisión, están compuestas por puntos de diversos tonos. Estos puntos observados por el ojo humano, o por algún sistema especializado, son clasificados en montañas, valles, etc..

Es muy común, sin embargo, que se produzcan interferencias dando por resultado imágenes deformadas, que en muchos casos hacen imposible el reconocimiento de la imagen verdadera. Nuestro propósito es revisar algunas formas en que este problema ha sido atacado y resuelto, al menos parcialmente.

Siguiendo el formalismo descrito arriba, consideraremos un mecanismo de deformación como un mapeo del álgebra

imagen A a un conjunto A° de imágenes deformadas. El propósito del Análisis de Patrones es describir la generación de A, el mapeo a A° y diseñar algoritmos para el análisis y reconocimiento de I (imagen real) dada I° (imagen deformada).

En esta parte, seguiremos el siguiente camino: en el plano, con la topología ordinaria, consideremos un conjunto de figuras cerradas y convexas llamadas *prototipos*. Consideremos un grupo G de transformaciones del plano en si mismo. G puede ser el grupo de translaciones en R^2, el de rotaciones, el de cambios de escala, o algún subgrupo de éstos.

El conjunto de signos será $S=G(\text{ Proto })$. Si $s_1, s_2, \ldots, s_\nu$ son signos y $B(x_1,\ldots,x_\nu)$ es una función Booleana de ν variables, hablaremos de la imagen $I= B(s_1,\ldots,s_\nu)$, es decir, las imágenes se forman mediante combinaciones conjuntistas de los signos. Diremos que dos configuraciones son equivalentes si originan la misma imagen; por ejemplo, si $E \subset F$ las imágenes resultantes E y EF son iguales, aunque provienen de configuraciones distintas. Es pertinente observar que si no queremos un álgebra imagen demasiado grande, el número de prototipos y el número de variables de B debe ser pequeño.

Ejemplo.- El prototipo es el círculo unitario, G= Translaciones×Homotecias. Dependiendo de B se pueden obtener figuras como las siguientes:

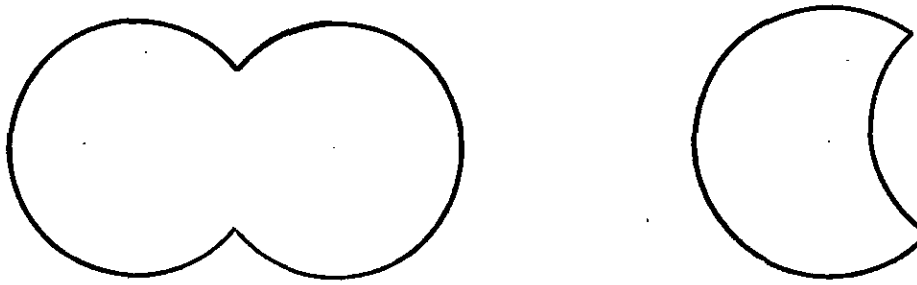

Figura 1

En ausencia de mayor información sobre el mecanismo de deformación, es razonable considerar a éste, como un proceso estocástico cuya acción sobre la imagen da por resultado:

a) Que sólo conozcamos de la figura una muestra aleatoria de ν puntos, distribuida uniformemente sobre ella.

b) Que obtengamos la realización de un proceso de Poisson con parametro μ dentro de la imagen.

c) Alguna generalización de los anteriores.

Sólo estos casos serán tratados, aunque las técnicas usadas permiten, en algunas situaciones, ser extendidas a casos no considerados.

El método para reconstruir la imagen verdadera I consistirá en calcular un cierto conjunto I* a partir de la imagen deformada. Se tratará de que I* esté cercano, en algún sentido, a I; por ejemplo, minimizando la expresión $m(I \Delta I^*)$, donde m es la medida de Lebesgue en R^2.

Si la imagen deformada es el conjunto $I^\circ = \{ x_1, \ldots, x_\nu \}$, formado por puntos del plano, definimos para el caso a), la función de verosimiltud en la variable (desconocida) I :

$L_1(I)= 1 / m(I)^{\nu}$ cuando $I^{\circ} \subset I$

$= 0$, en otro caso.

El estimador máximo-verosimil I* es el que resuelve el problema

(1) $m(I)= \underset{I^{\circ} \subset I}{\text{mínimo}}$

Para el caso b) ponemos $L_2(I)= \mu^{\nu} \exp\{ m(I)\mu \}$ cuando los puntos de la muestra estan desordenados, si no dividimos entre $\nu!$. El estimador MV sigue siendo el que resuelve (1).

Nótese la restricción $I^* \subset I$.

Con este planteamiento preliminar, podemos definir entonces la *Geometría Estadística* como el estudio de cómo restaurar figuras geométricas cuando sólo tenemos acceso a una versión deformada de ellas (Grenander).

I.2. *El círculo.* Empezaremos por el caso en que la imagen es la de un círculo descrito por tres parámetros: el radio y las coordenadas de su centro. Este ha sido generado por el prototipo círculo unitario y el grupo de translaciones y cambios de escala. En este caso, el álgebra imagen tiene 3 dimensiones.

Primero supongamos que el centro del círculo está situado en el origen de coordenadas. El problema es estimar el radio R.

De todas las formas posibles de resolver este problema, se usará una que sea aplicable posteriormente a figuras más complicadas. Se usarán las llamadas *características de frontera.* Este concepto quedará claro más

abajo.

Deaseamos un algoritmo eficiente en el sentido estadístico, y que nos diga cuándo la figura resultante no es un círculo. El algoritmo estará basado en los diámetros empíricos obtenidos a partir de la imagen deformada. Un diámetro en la dirección ϕ, es la distancia entre dos líneas soporte en una dirección ortogonal a ϕ.

Un algoritmo basado en diámetros empíricos es invariante bajo el grupo de transformaciones.

Estudiemos la siguiente figura, que representa al círculo unitario y una muestra aleatoria de puntos de el:

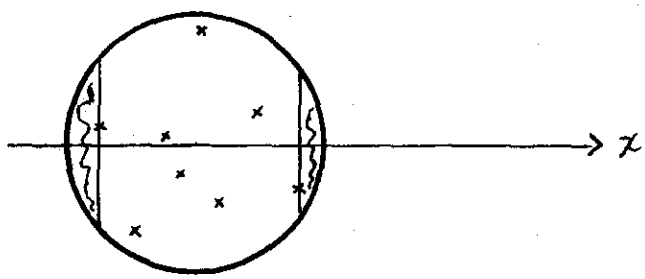

Figura 2.

Denotamos los puntos de la muestra por (x_κ, y_κ), $\kappa = 1,\ldots,\nu$. Ponemos $X_{max} = \max_\kappa(x_\kappa)$, $X_{min} = \min_\kappa(x_\kappa)$.

Para encontrar la distribución conjunta de éstos estadísticos de orden, notamos que:

$$\lim_{\nu\to\infty} X_{max} = 1 \qquad \lim_{\nu\to\infty} X_{min} = -1$$

Escribimos entonces:

$$X_{max} = 1 - u/n^\alpha \qquad X_{min} = -1 + v/n^\alpha$$

u y v son variables aleatorias positivas y α es un parámetro que se determinará posterirmente.

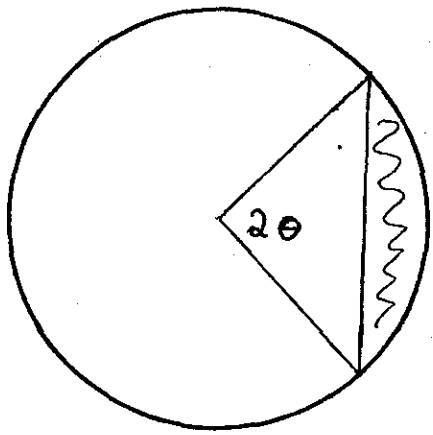

Figura 3

Sea $G(a,b)= P\{ u\geq a , v\geq b \}= P\{ -1 + b / \nu^{\alpha} \leq x_{\kappa} \leq 1 - a / \nu^{\alpha} ,\kappa=1,...,\nu \}.$ (1)

En la figura, el area de la parte rayada es

$A(\theta)= \theta - \cos\theta\operatorname{sen}\theta$ (2)

Haciendo $\cos\theta= 1 - x$, y usando el desarrollo de Taylor, válido para x pequeñas, obtenemos:

$\theta\simeq \sqrt{2x}$ (3)

Sustituyendo (3) en (2)

$A(\theta)= A(x)\simeq 4\sqrt{2}x^{1.5} / 3$ (4)

Como los x_{κ} son independientes:

$G(a,b)= (\pi - A(a/\nu^{\alpha}) - A(b/\nu^{\alpha}))^{\nu} / \pi^{\nu}$ (5)

Sustituyendo (4) en (5):

$G(a,b)= (1 - 4\sqrt{2}(a^{1.5} + b^{1.5}))^{\nu} /(3\pi\nu^{1.5\alpha})^{\nu}$

Tomamos $\alpha=2/3$ y vemos que:

$\lim_{\nu\to\infty} G(a,b)= \exp\{ -4\sqrt{2}(a^{1.5} + b^{1.5})/(3\pi) \}.$

El resultado obtenido, nos dice varias cosas: u y v, son asintoticamente independientes y tienen una función de distribución del tipo Weibull:

$H(x) = 1 - \exp\{ -\lambda x^{1.5} \}$ para $x>0$

$= 0$ otro valor.

El exponente $\alpha=2/3$, dá la rapidez de convergencia de X_{max} y X_{min}.

El diámetro empírico en la dirección x, es $D_1 = X_{max} - X_{min}$. Introducimos la v.a. $\zeta = u + v$, cuya distribución asintótica es $K(x) = H*H$, que puede ser escrita así:

$$K(x) = \int_0^X H(x-y)h(y)dy \quad , \; h(y) \text{ la f.d. de H.}$$

En efecto:

$$K(x) = \int_0^X (h*h)(u)du = \int_0^X \left(\int_0^\infty h(y)h(u-y)dy\right)du =$$

$$= \int_0^\infty h(y) \left(\int_0^X h(u-y)du \right)dy = \int_0^\infty h(y) \left(\int_0^{x-y} h(u)\right)dy$$

$$= \int_0^\infty h(y)H(x-y)dy = \int_0^X h(y)H(x-y)dy \quad .$$

Escogemos m (m natural fijo) direcciones $\phi_1,\ldots,\phi_m$ y efectuamos el proceso anterior, obteniendo m diámetros empíricos D_i asintóticamente independientes, siempre y cuando los segmentos correspondientes no se corten, pues el procedimiento usado para obtener G, ya no valdría. Esto vale para m pequeñas y muestras grandes.

Se tiene: $\zeta = u + v = (2 - D_1)n^{2/3}$, y para cada diámetro D_k se puede definir una v.a. ζ_k de manera que:

$$\lim_{\nu\to\infty} P\left(\max_k D_k \leq 2 - \sigma n^{-2/3} \right) = \lim_{\nu\to\infty} P\left(\min_k \zeta_k \geq \sigma \right) =$$

$$= 1 - K^m(\sigma) \ .$$

Como para un círculo de radio R, las distribuciones probabilistas no cambian, podemos usar el estimador

$$R^* = \max D_k \ / \ (2 - \sigma)n^{-2/3} \ .$$

En particular, si escogemos σ tal que satisfaga la ecuación $K^m(\sigma) = 1/2$, obtenemos un estimador asintóticamente insesgado en la mediana.

Supongamos ahora, que el círculo ha sido transladado, de tal manera, que hay que determinar las coordenadas (x_0, y_0) del centro. Usando los diámetros empíricos obtenidos anteriormente, y un conjunto fijo de direcciones Φ, escribimos:

$$X^{\phi}_{max} = x_0\cos\phi + y_0\operatorname{sen}\phi + R(1 - u/n^{\alpha})$$

$$X^{\phi}_{min} = x_0\cos\phi + y_0\operatorname{sen}\phi + R(-1 + v/n^{\alpha}).$$

para las mediciones máxima y mínima en la dirección ϕ respectivamente. El estimador obvio es el promedio, o punto medio

$$X^{\phi} = x_0\cos\phi + y_0\operatorname{sen}\phi + R\omega/n^{\alpha} \ , \ \omega = u - v.$$

ω es una v.a. con media 0 y variancia $2\sigma_H^2$. Escogemos una combinación lineal de X^{ϕ}, que nos dé un estimador lineal insesgado:

$$X_0^* = \sum_{\phi\in\Phi} c_{\phi} X^{\phi} \quad \text{con} \quad \sum c_{\phi}\cos\phi = 1 \ , \ \sum c_{\phi}\operatorname{sen}\phi = 0.$$

La variancia del estimador es:

$$V(X_0^*) = (2R^2\sigma^2 \sum c_\phi^2)/n^{2\alpha}$$

Como $\alpha=2/3$, vemos que $V(X^*) = O(n^{-4/3})$, que en particular, dá una convergencia más rápida que $\bar{x}$. Para calcular Y_0^*, se usa el mismo procedimiento.

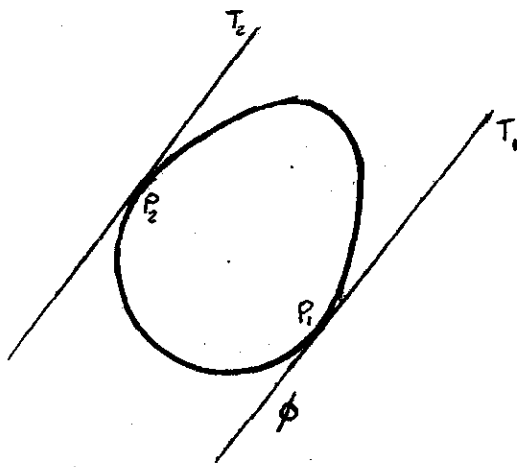

Figura 4

El análisis es parecido cuando las imágenes son de la forma I= gV, donde V es una figura cerrada convexa y con curvatura continua. Si G= cambios de escala, y hay que calcular el factór R, se procede así:

En la figura, T_1 y T_2 son 2 líneas soporte en la dirección ϕ, sus puntos de contacto son P_1 y P_2 y los radios de curvatura son $R_1(\phi)$ y $R_2(\phi)$.

Con un eje de coordenadas ortogonal a T_1 y T_2, introducimos X_{max} y X_{min} por medio de las fórmulas:

$$X_{max}^{\phi} = R(X_2^{\phi} - k_1(\phi)u/n^{\alpha})$$

$$X_{min}^{\phi} = R(X_1^{\phi} + k_2(\phi)v/n^{\alpha})$$

Los factores k_1 y k_2 se introducen para medir la desviación a un círculo (el de curvatura) en los puntos de contacto,

$$k_i(\phi) = (\pi R_i^2(\phi))/m(V) \qquad i = 1,2.$$

Tomamos $R^{\phi} = (X^{\phi}_{max} - X^{\phi}_{min})/(X^{\phi}_2 - X^{\phi}_1)$

Se puede escribir:

$$R^{\phi} = R - R(k_1 u + k_2 v)/(X^{\phi}_2 - X^{\phi}_1)n^{\alpha}$$

El estimador $\sum c_{\phi} R^{\phi} = R^*$ es insesgado si:

$$\sum c_{\phi}(1 - \mu(k_1 + k_2)/(X^{\phi}_2 - X^{\phi}_1)n^{\alpha} = 1$$

Es de variancia mínima si:

$$\sum c^2_{\phi}(k_1 + k_2)/(X^{\phi}_2 - X^{\phi}_1)^2 = \min.$$

En todas las fórmulas, μ es la media de la distribución H. De nuevo, $V(R^*) = O(n^{-4/3})$.

Ahora, pódemos permitir operaciones conjuntistas entre imágenes. Los métodos seguidos, son extendibles a casos como los de la figura 5; sin embargo, en el caso de las figuras con picos, la distribución probabilista ya no es H, sino $D(x) = 1 - \exp\{ -x^2 \cos\phi \cos(v/2)/A^2 \}$, $x>0$, v el ángulo de la esquina y ϕ la dirección en que nos acercamso. Por esto, es deseable usar otro tipo de métodos menos paramétricos y aplicables a casos mas generales.

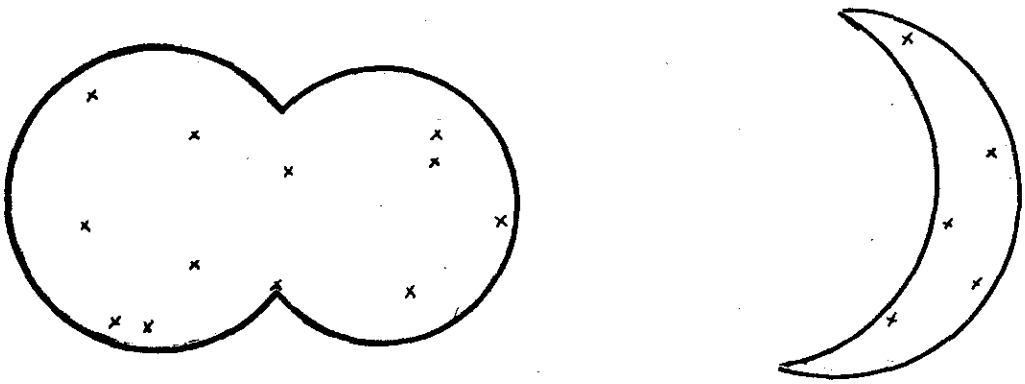

Figura 5.

Antes de pasar a métodos generales, examinaremos un caso muy interesante:

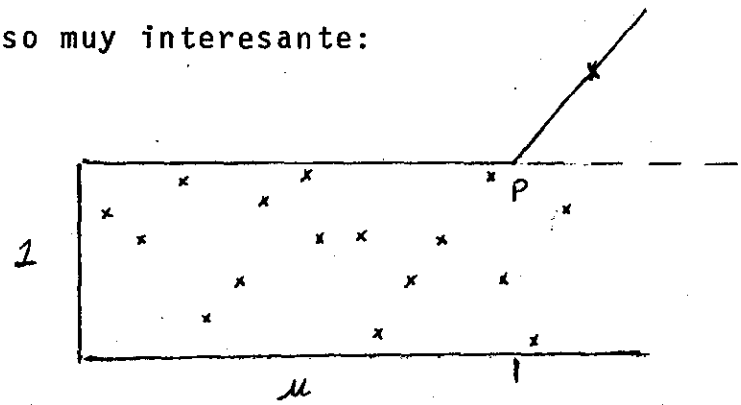

Figura 6.

Se tiene por imagen una banda de ancho unitario y una transversal tocándola en un punto P a u unidades del origen. El problema es estimar u, primero bajo la hipótesis de que tenemos acceso a una cantidad ilimitada de información.

Sea $\{x_i\}$ una sucesión de v.a. independientes, $x_i \varepsilon R(0,a_i)$, $a_i = 1 + \phi(i/n)/n$, donde ϕ es no decreciente y continua escalonadamente.

De la figura se vé, que $a_i = 1$ para $i = 1,\ldots,i_0$, y queremos ver cual es el subíndice $i_0 = \alpha n$, observando los x_i.

Hacemos $\lambda = 1 + \delta/n$, $\delta > 0$ y escogemos el subíndice para el cual

$x_1 \leqslant \lambda,\ x_2 \leqslant \lambda, \ldots\ x_{k-1} \leqslant \lambda,\ x_k > \lambda.$

Pero esto ocurre con probabilidad

$$p_k = \prod_{1}^{k-1} P(\ x_i \leqslant \lambda\ ) \cdot P(\ x_k > \lambda\ )$$

Hacemos:

$$Q_k = \prod_{1}^{k} P(x_i \leqslant \lambda\ )$$

La probabilidad de que k no exceda un cierto tope nx, nos la da la distribución:

$$F_n(x) = P(\ k \leqslant nx\ ) = p_1 + p_2 + \ldots + p_{\underline{nx}} = 1 - Q_{\underline{nx}}.$$

De la forma de las a_i, vemos que si $\phi(i/n) < \delta$, los factores correspondientes de Q son 1, y tomando la raíz más pequeña x' de la ecuación $\phi(x)=\delta$, resulta:

$$Q_{\underline{nx}} = \prod_{x'<i/n<x} P(\ x_i < \lambda\ ) = \prod_{x'<i/n<x} (\ 1 + \delta/n\ )/(\ 1 + \phi(i/n)/n\ )$$

Si n es grande:

$$\ln Q_{\underline{nx}} = \sum_{x'<i/n<x} (\delta - \phi(i/n))/n + o(1/n)$$

Teorema. La distribución límite F, del tiempo k, cuando las observaciones pasan el nivel λ, es

$$F(x) = \lim_{n\to\infty} P(\ k \leqslant nx\ ) = 1 - \exp\{\ \int_{x'}^{x} (\delta - \phi(x)\ )dx\ \},\ x > x'$$

En la figura anterior, $\phi(x) = \beta(x - \alpha\ )$ cuando $x > \alpha$ y $\phi(0) = 0$ en otro caso. β es la pendiente de la transversal.

En este caso, $x' = \alpha + \delta/\beta$ y

$$F(x) = 1 - \exp\{\ -\beta(x - \alpha\ )^2/2 + (\ x - \alpha - \delta/\beta\ ) + \delta^2/2\beta\}$$

Además, Mediana(x)= $\alpha+ \delta/\beta +\sqrt{2\ln 2/\beta}= x' +(2\ln 2/\beta)^{1/2}$

Un cambio más brusco de la trayectoria, queda descrito por:

$\phi(x)= 1$ si $0\leqslant x\leqslant\alpha$

$= 1+d$ otro caso

con $d>\delta$, y para esta situación:

Corolario. Para un cambio súbito de las a_i, la distribución límite está dada por:

$F(x)= 1 - \exp\{ -(d - \delta)(x - \alpha) \}$, $x>\alpha$.

Bajo las mismas hipotésis, supongamos ahora, que solo se tiene acceso a N= $\underline{cn}$ observaciones, $c>\alpha$. Construimos la función de verosimilitud, para $\alpha'>\alpha$:

$$L(\alpha')= \prod_{i=\underline{\alpha' n}}^{\underline{cn}} 1/(1+(\beta/n)(i/n-\alpha')) \quad \text{si } x_i\leqslant 1 +(\beta/n)(i/n-\alpha) \text{ para } \underline{\alpha' n}\leqslant i\leqslant \underline{cn} .$$

$L(\alpha')= 0$ en otro caso.

El estimador máximo-verosimil α^*, es el máximo valor de α' para el cual se sigue teniendo

$x_i\leqslant 1+(\beta/n)(i/n-\alpha')$, con i entre $\underline{\alpha' n}$ y $\underline{cn}$.

Calculamos:

$$G_n(t)= P(\alpha^*\leqslant t)= 1 - \prod_{\alpha\leqslant i/n\leqslant t}(1/\omega_i) \prod_{t\leqslant i/n\leqslant c}(1+(\beta/n)(i/n-t))/\omega_i$$

donde $\omega_i= 1 + (\beta/n)(i/n- \alpha)$.

Obtenemos:

$$G(t)= \lim_{n\to\infty} G_n(t)= 1 - \exp\{ -\beta((t-\alpha)^2/2 + (c-t)(t-\alpha)\}$$

y se puede enunciar:

Teorema. El e.m.v. cuando sólo se tiene acceso a $\underline{cn}$

observaciones, tiene la distribución límite:

$$G(t) = 1 - \exp\{-\beta((t-\alpha)^2/2 + (t-\alpha)(c-t)\}, \quad \alpha<t<c$$
$$= 1 \qquad t>c$$

I.3. *El mecanismo de deformación es un proceso de Poisson.* Consideremos dos funciones continuas ϕ_1 y ϕ_2 con $\phi_1(x)<\phi_2(x)$ para toda $x\varepsilon(0,1)$, y los signos $s_1=\{(x,y)|\ x\varepsilon(0,1),\ y\leqslant\phi_2(x)\}$ y $s_2=\{(x,y)|\ x\varepsilon(0,1),\ y\geqslant\phi_1(x)\}$.

La imagen es $I=(g_1s_1)(g_2s_2)$, donde g_1 y g_2 son traslaciones en la dirección del eje y, de a_1 y a_2 unidades respectivamente (a_1 y a_2 no negativos).

El proceso de Poisson tiene parámetro y dá una imagen deformada $I^\circ=\{(x_1,y_1),\ldots,(x_n,y_n)\}$. La situación se muestra en la figura:

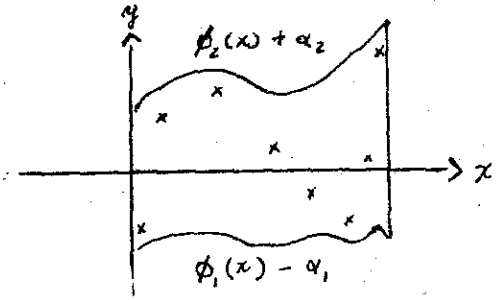

Figura 7.

Para encontrar el e.m.v. de a_1 y a_2, basta minimizar el área $A(a_1,a_2)= A(0,0) + a_1 + a_2$, así que basta minimizar a_1 y a_2 separadamente. Los estimadores son:

$$a_1^*= \max_i\{\phi_1(x_i) - y_i\} \qquad a_2^*= \max_i\{y_i - \phi_2(x_i)\} \ .$$

Para encontrar su distribución conjunta, es suficiente observar que si $s\leqslant a_1$ y $t\leqslant a_2$, entonces $F(s,t)= P(a_1^*\leqslant s, a_2^*\leqslant t)= P(\phi_1(x_i)-s\leqslant y_i\leqslant\phi_2(x_i)-t,\ i=1,\ldots,n)= \exp\{-\mu(A(a_1,a_2) - A(s,t)\}= \exp\{-\mu(a_1-s) -\mu(a_2-t)\}$. Esto nos di-

ce que los estimadores a_1^* y a_2^* son estocásticamente independientes, y el error $a_i - a_i^*$ tiene una distribución exponencial par i=1,2.

I.4. *Comportamiento del conjunto margen.* Consideremos una imagen de la forma $I=gI_0$, generada a parti de un prototipo I_0 fijo en R^2, convexo, cuya frontera tiene curvatura continua y positiva. La imagen deformada es $I^\circ = \{ (x_1,y_1),\ldots,(x_n,y_n) \} \subset I$, y la única hipótesis es que el elemento del grupo de transformaciones es único para cada imagen: $g_1 I_0 = g_2 I_0 \rightarrow g_1 = g_2$.

El problema de determinar qué elemento del grupo de transformaciones produce a I, se reduce a una búsqueda en el conjunto $\gamma_n = \{ g \in G \mid g^{-1}(I^\circ) \subset I \}$, y los algoritmos que den resultados en γ_n, se llamarán *admisibles*. En cada caso, el algoritmo admisible usado, se determina por criterios adicionales, como error cuadrático medio, verosimilitud, etc..

Los γ_n son conjuntos estocásticos que forman una sucesión decreciente. Si formamos la cubierta convexa de I° y la denotamos por $\hat{I}^\circ$, es fácil ver que $P(\hat{I}^\circ \mid I) = 1$, pues siempre $\hat{I}^\circ \quad I$; pero entonces, si $I = g_v I_0$, vemos que

$$\gamma_n = \{ g \mid g^{-1}(\hat{I}^\circ) \subset I_0 \},$$

$$\lim_{n\to\infty} \gamma_n = \{ g \mid g^{-1}(g_v I_0) = I_0 \}$$

de donde, por nuestra hipótesis, $g^{-1} g_v = e$, lo cual implica:

Teorema. Cualquier algoritmo admisible, es consistente en el sentido estadístico.

Demostración. Se ha visto que $\lim_{n\to\infty} \gamma_n = \{g_v\}$.

Ejemplos. G es el grupo de translaciones en R^2. Escribimos sus elementos en la forma g= (a,b), y ponemos g_v= (α,β). En este caso

$$\gamma_n = \{ (a,b) \mid (x_i-a, y_i-b) \in I_0, i= 1,\ldots,n \}$$

Definiendo

$$E_i = \{ (a,b) \mid \exists (x,y) \in I_0 \text{ tal que } (a,b) = (x_i-x, y_i-y) \},$$

los E_i son conjuntos estocásticos independientes y convexos que satisfacen:

$$\gamma_n = \bigcap_{i=1}^{n} E_i$$

γ_n es convexo.

Sin examinarlo en detalle, cuando G es el subgrupo del grupo lineal de R^2 formado por los elementos de determinante positivo, definimos

$$E_i = \{ g \mid (g_{11}x_i+g_{12}y_i, g_{21}x_i+g_{22}y_i) \in I_0 \},$$

cuando

$$g^{-1} = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}$$

y se sigue verificando que $\bigcap_{i=1}^{n} E_i = \gamma_n$, γ_n es convexo.

Para este ejemplo, se tiene el siguiente lema:

Lema. En el caso descrito arriba, estimación m.v. de g_v, es equivalente a un problema de programación cuadrática.

Demostración. Poniendo la función de verosimilitud

$$L(I) = m(I)^{-n},$$

debemos minimizar m(I)= det(g)m(I_0), y por lo tanto, basta minimizar la función cuadrática $\det(g) = g_{11}g_{22}-g_{12}g_{21}$.

Por último, veamos un teorema sobre el "tamaño promedio" de los conjuntos γ_n.

Teorema. Si G es el grupo aditivo de las translaciones, el conjunto de margen γ_n verifica:

$$\lim_{n\to\infty} n^2 E(m(\gamma_n)) = K\int_0^{2\pi} d\phi / D_\phi^2$$

donde K es una constante positiva, D_ϕ es el diámetro asociado a dos lineas de soporte en la dirección ϕ

Demostración. Denotemos por A_0 el área de I_0; por g al elemento (x,y); por $I_g = I_0 + (x,y)$ y por f al indicador de I_0. Usando el ejemplo anterior, vemos que el indicador de γ_n puede ser escrito como

$$q(x,y) = \prod_{i=1}^{n} f(x_i - x, y_i - y), \text{ así que}$$

$$m(\gamma_n) = \int_{R^2} \prod f(x_i - x, y_i - y).$$

Como los puntos (x_i, y_i) son estocásticamente independientes, con función de densidad $A_0^{-1} f(x,y)$,

$$Em(\gamma_n) = \int_{R^2} (A_0^{-1} \int_{R^2} f(u-x, v-y) f(u,v) du dv)^n dx dy$$

Definimos $A_0 K(x,y) = \int_{R^2} f(u-x, v-y) f(u,v)$. Se tiene:

$$K(x,y) = A_0^{-1} \int_{I_0 I_g} f(u-x, v-y) f(u,v).$$

$$Em(\gamma_n) = \int_{R^2} K^n(x,y) dx dy$$

Obviamente $K(x,y) \leq 1$, y si el vector (x,y) es pequeño y apunta en la dirección ϕ, ponemos D_ϕ para el diámetro asociado en la dirección ϕ. Pero la integral que representa

a K, es el área de la región $I_0 I_g$, y por lo tanto, puede ser aproximada por $A_0 - ||g||D_\phi(1 + o(1))$, de donde, asintóticamente,

$$K(x,y) = 1 - ||g||D_\phi/A_0 \simeq \exp\{ -||g||D_\phi/A_0 \} .$$

De aquí que;

$$Em(\gamma_n) = \int_{R^2} \exp\{ -n||g||D_\phi/A_0 \} dxdy$$

Pasando a coordenadas polares:

$$Em(\gamma_n) = \int_0^{2\pi}\int_0^{\infty} \exp\{ -nD_\phi \rho/A_0 \} \rho d\rho d\phi$$

La integral $\int_0^{\infty} \exp\{ bt \} t^n$ tiene el valor $\Gamma(n+1)/b^{n+1}$; simplificando la expresión de arriba, obtenemos:

$$Em(\gamma_n) = A_0 n^{-2} \int_0^{2\pi} d\phi / D_\phi^2 \quad .$$

Nota. Se puede enunciar unteorema análogo para el caso de rotaciones, y otro para el caso de cambios de escala, pero no se obtiene nada nuevo.

I,5. *Algoritmos generales.* Los métodos anteriores se caracterizan por usar propiedades de la frontera de I. Sacrificando esto, se pueden construir algoritmos más generales que resultan consistentes en el sentido siguiente:

Definimos el error esperado de I y I* como

$$e(I,I^*) = Em(I \Delta I^*)$$

Los algoritmos consistentes son los que hacen tender a 0 éste error cuando los parámetros de los mecanismos de deformación crecen.

La imagen deformada I° no sirve como estimador de I pues $e(I^\circ,I)= Em(I)$. Para remediar esto, cuando $I^\circ= \{x_1, \ldots,x_n\}$, cubrimos cada x_i con discos de radio ρ y centrados en el punto; la unión de estos discos nos dá un estimador consistente si escogemos el radio adecuado. El teorema siguiente nos dice como hacerlo.

Teorema. Sea $I^\circ= \{x_1,\ldots,x_n\}$ una imagen deformada de I. $I^*= \bigcup_{i=1}^{n} C_\rho(x_i)$, donde los $C_\rho(x_i)$ son discos de radio ρ y centro x_i, define un algoritmo de reconstrucción consistente para cualquier imagen I cuando el mecanismo de deformación es del tipo a) y ρ se escoge de manera que $\rho\to 0$, $n\rho^2\to\infty$.

Demostración. Denotamos por I(x), I*(x) los indicadores de I, I* respectivamente. Se tiene:

$$Ee(I,I^*)= Em(I\Delta I^*)= \int_I (1-EI^*(x))dm(x) + \int_{R^2-I} EI^*(x)dm(x)$$

Construimos

$D'_\rho= \{x\in I \mid d(x,\partial I)\geq\rho\}$

$D''_\rho= \{x\in R^2-I \mid d(x,\partial I)\geq\rho\}$

Si $x\in D'_\rho$, calculamos

$$EI^*(x)= P(x\in I^*)= 1 - P(x\notin I^*)= 1 - \prod_{i=1}^{n} P(x\in C_\rho^c(x_i))=$$

$$= 1 - (1 - \pi\rho^2/m(I))^n \simeq 1 - \exp\{ n\pi\rho^2/m(I)\}.$$

Analogamente, $EI^*(x)= 0$ para $x\in D''_\rho$. Pero cuando $\rho\to 0$ tenemos que:

Rts-387

$$m(D_\rho^c \cap D_\rho^c) \to 0$$

$$E(I^*(x)) \to 1$$

y con esto, obtenemos el resultado.

Este no es el único estimador que se puede construir; por ejemplo, en la figura

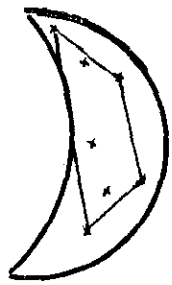

Figura 8.

una vez formada la cubierta convexa de I°, denotamos por P_1 y P_2 los puntos vértices del lado de mayor longitud. Formamos círculos de radio ρ y que pasan por P_1 y P_2. El estimador de I será $I^* = C(\rho_1) \cap C^c(\rho_2)$, donde ρ_1 y ρ_2 hacen mínima la expresión $m(I^*)$ con la restricción $I^* \subset I$.

Esto sugiere que para imágenes de la forma $I = B(s_1, \ldots, s_n)$, donde B es una función booleana, tomar como estimador $I^* = \bigcap B(s_{\alpha 1}, \ldots, s_{\alpha n})$, donde se intersecta sobre todos los $(s_{\alpha 1}, \ldots, s_{\alpha n})$ que hacen $I^\circ \subset B(s_{\alpha 1}, \ldots, s_{\alpha n})$. Los $s_{\alpha i}$ se generan a partir de los prototipos por medio detransformaciones de similaridad. Bajo hipótesis débiles, se demuestra que éste es un estimador consistente.

CAPITULO II

II.1. *Planteamiento del problema.* El problema general que atacaremos en este capítulo, es el de como guardar y recobrar información en una computadora. Esta información puede ser numérica, por ejemplo, en una etapa de un algoritmo para resolver numericamente ecuaciones diferenciales, se ha guardado en la memoria de la computadora una tabla de valores de una función f(x); en una etapa posterior es necesario buscar el valor correspondiente a un cierto x. Algunas veces la información es menos matemática, por ejemplo, buscamos la traducción de una cierta palabra en ruso.

La manera en que este problema será resuelto, será un tanto probabilista. Empezaremos por fijar la nomenclatura.

Vamos a suponer que un conjunto de N *datos* (records), han sido guardados y el problema es encontrar el que necesitamos. Cada dato tiene asociado una *clave* (key) que lo identifica univocamente. El conjunto de datos, será llamado la *tabla* (file).

El método del *Hashing* recupera información así: un número m (m>N) de posiciones de memoria dentro de la computadora son usadas para guardar los datos; suponemos que un dato cabe en una localidad de memoria. Estas posiciones serán $E_1,\ldots,E_m$. Tambien son reservadas m localidades $T_1,\ldots,T_m$ para las claves. Si T_i no está vacía,

E_i contiene el dato correspondiente a la clave guardada en T_i. Para guardar los datos, tomamos una clave x y calculamos una cierta función h(x) que toma valores enteros entre 1 y m; guardamos a x en $T_{h(x)}$ y el dato correspondiente en $E_{h(x)}$. En principio, buscar la localidad que guarda a x, es buscar el dato cuya clave es x.

Es necesario que h sea facil de calcular, y que $h(x) \neq h(y)$ para $x \neq y$, pero con N claves y m localidades, la probabilidad de encontrar una de esas funciones *hash* es $P = (m)_N / m^N$, un número muy pequeño si N es grande. Ademas, generalmente no sabemos por adelantado que claves van a ser usadas, así que h debe funcionar para cualquier conjunto potencial de claves. Es inevitable pues, que debemos permitir que sucedan *colisiones*, esto es, que para algunas parejas $x \neq y$ ocurra que $h(x) = h(y)$

En la práctica las funciones utilizadas, son aquellas que, con dominio el conjunto de todas las posibles claves, y contradominio { 1,...,m }, den a una colisión una probabilidad de ocurrencia de 1 / m. Un ejemplo bastante bueno de estas funciones, es $h(x) \equiv x \pmod{m} + 1$ donde escogemos a m primo. Este tipo de funciones, es muy frecuentemente favorecido en los *Métodos Monte Carlo* (véase Knuth). Una manera de resolver colisiones, es buscar en las localidades h(x) + 1, h(x) + 2,..., hasta encontrar una vacía.

El algoritmo siguiente, resume la discusion anterior

Paso 1. A h(x) lo llamamos i

Paso 2. Si T_i contiene a x, sacamos la información

de E_i. El algoritmo termina.

Paso 3. Si T_i está vacía, es por que la clave x no está guardada. Se guarda en T_i.

Paso 4. Si i = m, ponemos i = 1; si no, aumentamos i en una unidad. Regresamos al paso 2.

Este será el *Algoritmo L*.

El algoritmo sirve tanto para guardar la información como para recobrarla, pues al empezar, todas las T_i están vacías; para guardar una clave (y el dato correspondiente), calculamos h(x) y buscamos a x por las T_i; no lo encontraremos y el algoritmo parará en el paso 3: guardamos a x en T_i. De ahora en adelante, cuando se nos dé x, encotraremos la posición T_i por medio del algoritmo y así recuperaremos la información.

II.2. *Aspecto Matemático*. El problema básico planteado en relación al algoritmo anterior, es determinar el promedio de veces que se repite el paso 2 hasta encontrar x. Veamos un ejemplo con 10 claves A, B,..., J y 10 localidades $T_1,\ldots, T_{10}$. La función hash será h(x) = i-ésimo dígito de π. Al guardar las claves, obtenemos la siguiente configuración simbólica:

1	2	3	4	5	6	7	8	9	10
B	D	A	C	E	G	H	I	F	J

Al cabo de cierto tiempo, algunos pedazos de la tabla empiezan a congestionarse y las configuraciones resultantes no pueden considerarse aleatorias. Por ejemplo despues de colocar las primeras 7 claves, obtenemos las

sigiuentes posiciones ocupadas:

1	2	3	4	5	6	7	8	9	10
B	D	A	C	E	G			F	

Para la clave H, si suponemos que empieza en una posición aleatoria, la probabilidad de que quede en T_7 es 0.6, pero la probabilidad de que termine en T_8 es 0.1.

Las distancias recorridas por las 10 claves son: 0, 0, 0, 1, 0, 0, 4, 1, 3, 7.

Para el análisis matemático, empecemos por suponer que los valores de la función hash son $a_1,\ldots,a_m$. Esto será una sucesió hash. Calculamos $u_K(m,n)$, el número de sucesiones hash parciales $a_1,\ldots,a_n$, tales que, despues de haber colocado las primeras n claves, la posición K queda vacía. Por consideraciones de simetría, se debe tener $u_1(m,n)= u_2(m,n)= \ldots = u_K(m,n)$. Llamemos u(m,n) a ese número. Entonces:

$$\sum_{K=1}^{m} u_K(m,n) = mu(m,n)$$

Cada sucesión $a_1,\ldots,a_n$ deja vacías (m-n) posiciones, es decir, es contabilizada por (m-n) de los $u_K(m,n)$. Como hay m^n sucesiones de n términos, obtenemos la ecuacion:

$$mu(m,n) = (m-n)m^n \ .$$

Denotemos ahora por v(m,n,k) el número de sucesiones hash parciales, tales que, despues de guardar n claves, las posiciones 1, 2,..., k, quedan ocupadas y las k + 1 y m vacías. Para que las posiciones 1,..., k queden ocupadas

es necesario que en la sucesón $a_1,\ldots,a_n$, haya k números menores que k + 1, pues como la localidad m deb quedar vacía, esas posiciones no se llenaron a partir de números más grandes; pero estos números, forman una de las sucesiones contadas por u(k+1,k). Los números restantes, forman una sucesión enumerada por u(m-1-k,n-k).

Todas las sucesiones contadas por v(m,n,k), pueden ser construidas intercalando una de las sucesiones contadas por u(k+1,k) con una de las contadas por u(m-1-k,n-k). Obtenemos:

$$v(m,n,k)= C_n^m u(k+1,k)u(m-1-k,n-k)$$

$$= C_n^m\ (k+1)^{k-1}(m-k-1)^{n-k-1}(m-n-1).$$

La distancia recorrida por la n-ésima clave es k, si, y solo si, la sucesión parcial $a_1,\ldots,a_{n-1}$ ha ocupado las localidades a_n, $a_n+1,\ldots,a_n+k-1$ y dejado vacía la a_n+k.

La probabilidad $p_k(m,n)$ de que la n-ésima clave deba recorrer k localidades hasta encontrar una vacía, es:

$$p_k(m,n)= \sum_{r\geqslant 0} v(m,n-1,k+r)\ /\ m^{n-1}$$

En efecto, hay m^{n-1} sucesiones de n-1 términos, de éstas nos interesan las que dejan ocupadas las localidades $a_n,\ldots,a_n+k-1$ y dejan vacía la a_n+k. Por simetía, estas son las enumeradas por v(m,n-1,k+r), con $r\geqslant 0$.

Teorema. La distancia promedio d(m,n) que la n-ésima clave debe recorrer hasta llegar a una localidad vacía, es:

$$2d(m,n)= (\ 2(n-1)/m + 3(n-1)(n-2)/m^2 + \ldots\ ).$$

Demostración. Por probabilidad elemental, sabemos que:

$$d(m,n)= \sum_{k\geqslant 0} k p_k(m,n)$$

Sustituyendo el valor de p_k, obtenemos:

$$d(m,n)= \sum_{k\geqslant 0} k \left(\sum_{r\geqslant k} v(m,n-1,r)\right)/m^{n-1} =$$

$$= \sum_{r\geqslant k\geqslant 0} (m-n)m^{1-n} k C_r^{n-1} (r+1)^{r-1} (m-r-1)^{n-r-2}$$

$$= (1/2)(m-n)m^{1-n} \sum_{r\geqslant 0} r C_r^{n-1} (r+1)^r \quad (m-r-1)^{n-r-2}$$

En el último paso se usó la identidad $\sum_{k=1}^{r} k = r(r+1)/2$.

Hemos obtenido la ecuación:

$$2m^{n-1} d(m,n)= \sum_{r\geqslant 0} r C_r^{n-1} (r+1)^r (m-r-1)^{n-r-2} (m-n)$$

$$= (n-1) \sum_{k\geqslant 0} C_k^{n-2} (k+2)^{k+1} (m-2-k)^{n-k-3} (m-n)$$

En la última igualdad, se hizo el cambio $r-1= k$.

Estudiemos la función:

$$S(n,x,y)= \sum_{k\geqslant 0} C_k^{n-2} (x+k)^{k+1} (y-k)^{n-k-3} (y-n+2)$$

Desarrollamos:

$$S(n,x,y)= \sum_{k\geq 0} C_k^{n-2} x(x+k)^k (y-k)^{n-k-3}(y-n+2) \quad +$$

$$\sum_{k\geq 0} kC_k^{n-2} (x+k)^k (y-k)^{n-k-3}(y-n+2)$$

En la primer suma, hacemos u= n-2-k; en la segunda, v= k-1, para k>0. Obtenemos:

$$S(n,x,y)= x\sum_{u>0} C_u^{n-2}(x+n-2-u)^{n-2-u}(y-n+2)^{u-1}(y-n+2) \quad +$$

$$(n-2)\sum_{v>0} C_v^{n-3}(x+1+v)^{v+1}(y-1-v)^{n-v-4}(y-n+2)$$

Recordando la fórmula de Abel, válida para reales arbitrarios a y b ($a\neq 0$) y t, entero positivo:

$$(a+b)^t= \sum_{r\geq 0} C_r^t a(a-rc)^{r-1}(b+rc)^{t-r} \qquad c \text{ real arbitrario.}$$

S puede escribirse:

$$S(n,x,y)= x(x+y)^{n-2} + S(n-1,x+1,y-1) \quad . \qquad (^{\alpha})$$

Regresando a la ecuación para $2m^{n-1}d(m,n)$, vemos que la suma de la derecha es S(n,2,m-2); para obtener el resultado, basta desarrollar la ecuación ($^{\alpha}$).

Del teorema, obtendremos unos corolarios muy importantes:

Corolario. La distancia promedio $\delta(m,n)$, recorrida por las primeras n claves, satisface:

$$\delta(m,n)= (n-1)/(2m) + (n-1)\delta(m,n-1)/m$$

En efecto, $\delta(m,n) = \left(\sum_{k=1}^{n} d(m,k) \right)/n = \left((n-1)/m + \ldots \right)/2$

Una simple factorización, dá el resultado.

Corolario. Cuando n=m, la distancia promedio recorrida por cada clave es:

$\delta(m,m) = \left((m-1)/m + (m-1)(m-2)/m^2 + \ldots \right)/2$

En éste caso, todas las localidades reservadas quedan ocupadas.

Es interesante ver algunos casos asintóticos, por ejemplo, si $\alpha=m/n$ es el cociente de localidades ocupadas entre el total de localidades, se puede demostrar que si hacemos tender m a infinito, manteniendo a α fija, $\delta(m,\alpha m)$ tiende a $\alpha/2(1-\alpha)$. Ademas, para el caso n=m, si m es bastante grande, se tiene:

$\delta(m,m) \simeq (\pi m/8)^{1/2} - 2/3 = O(m^{1/2})$.

II,3. *Variantes del problema*. La técnica usada para resolver colisiones, se llama *búsqueda lineal*; como el principal problema que ocasiona es el del congestionamiento, se han creado otras formas de búsqueda. Al respecto, hay el siguiente teorema:

Teorema. El promedio de veces que se necesita buscar una clave por medio de búsqueda lineal, es independiente del órden en que hayan sido guardadas las claves (Peterson 1957).

Esto quiere decir, bajo la hipótesis de que todas las localidades son igualmente elegibles, que la sucesión 314 1592653, dá la misma configuración (localidades ocupadas) que la 3562951413 o que cualquier otra permutación de ella.

En particular, guardarlas simultaneamente no alivia nada.

Demostración. Para demostrar el teorema, supongamos que la sucesión $a_1,\ldots,a_n$, dá lugar a la configuración $b_1,\ldots,b_n$. Basta demostrar el teorema para la sucesión $a_1,\ldots,a_{i-1},a_{1+i},a_i,\ldots,a_n$, donde $1<i<n$. Hasta b_{i-1}, estas posiciones son las mismas para la segunda sucesión. La i-ésima clave, en la segunda sucesión es mandada a a_{i+1} y debe quedar guardada en b_{i+1}, a menos que se encuentre a b_i (que para la segunda sucesión esta vacía); si no la encuentra, la clave (i+1) mandada a a_i, termina en b_i; si la encuentra, la clave (i+1) debe terminar en b_{i+1} (el resto de las localidades no se han modificado). En el peor de los casos, las claves i y (i+1), han intercambiado localidades; pero entonces, el promedio es el mismo. El número de pruebas para la clave (i+1), ha disminuido en la misma cantidad que ha aumentado para la i-ésima.

En el algoritmo L, podemos cambiar el paso 3 $i\leftarrow i+1$ por $i\leftarrow i-c$ (o bien, $i+c$) donde $c>0$ y $(c,m)= 1$. Aunque una c fija no reduce el congestionamiento, podemos mejorar el algoritmo, haciendo que c dependa de las claves (Balbine 1968). Esto nos lleva al algoritmo D, conocido como *Hashing doble*:

Paso 1. $i\leftarrow h_1(x)$.

Paso 2. Si T_i está vacía, ir al paso 6; si no, y T_i contiene a x, sacamos la información de E_i. El algoritmo termina satisfactoriamente.

Paso 3. Si $x \notin T_i$, $c \leftarrow h_2(x)$

Paso 4. $i \leftarrow i-c$; si $i<0$, $i \leftarrow i+m$.

Paso 5. Si T_i está vacía, ir al paso 6; si $x \in T_i$, hacer como en el paso 2; si $x \notin T_i$, ir al paso 4.

Paso 6. Guardamos a x en T_i, y la información correspondiente en E_i.

Una buena elección de funciones hash h_1 y h_2, es tomar m de tal manera que (m-2) y m sean primos; si $h_1(x) \equiv x(\text{ mod } m) + 1$, podemos escoger $h_2(x) \equiv x(\text{ mod } m-2) + 2$. La razón, es que con este tipo de funciones, tenemos una cierta independencia probabilista: la probabilidad de que a una misma clave la manden al mismo valor, es del orden m^{-2}.

Otra técnica que evita congestionamientos, es la llamada *Hashing simple:* Tomamos una matriz S de m×m enteros, tal que cada renglón es una permutación de { 1,...,m} y la primer columna, contiene a los enteros 1,...,m ordenadamente. Por ejemplo, para m=4, una matriz de hashing simple es

$$S= \begin{pmatrix} 1 & 3 & 2 & 4 \\ 2 & 1 & 4 & 3 \\ 3 & 2 & 4 & 1 \\ 4 & 3 & 2 & 1 \end{pmatrix}$$

La función hash h(x), selecciona un renglon de S, y para buscar un dato, seguimos el orden dictado por ese renglón. Tenemos el mismo algoritmo L, con la modificación:

Paso 4' $i \leftarrow$ siguiente valor en el renglón h(x) de la matriz S; volver al paso 2

La búsqueda lineal es un caso particular del hashing simple. Para el caso m=4, la matriz es:

$$S = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{pmatrix}$$

En lo que sigue, llamaremos *hashing cíclico* al que utiliza búsqueda lineal.

Hay $((m-1)!)^m$ matrices que cumplen los requisitos pedidos para una matriz de hashing simple; para cada una de ellas es posible definir las medidas $d_1(m,n)$ y $\delta_1(m,n)$ análogas a las del hashing cíclico. El problema es determinar en promedio $d_1(m,n)$, para una matriz escogida aleatoriamente. Se conocen las fórmulas, pero no la manera de simplificarlas; sin embargo, se sabe que para m grandes, $\bar{\delta}_1(m,m) \simeq \ln m + \gamma - 1.5$, donde γ es la constante de Euler. Es claro que este promedio es mejor que el obtenido anteriormente, pero para una matriz determinada, su promedio puede ser mucho mas pequeño que $\bar{\delta}_1$, así que vale la pena investigar que tan chico puede ser δ_1, y para que matrices se obtiene este valor. Otro problema asociado a éste, es el de construir matrices que tengan promedio cercano al óptimo, y para las cuales el paso 4' sea fácil de realizar.

Por ejemplo, se sabe que para m=4, la matriz óptima es la de hashing cíclico. El problema no ha sido resuelto en general, pero existe la conjetura de que el mejor esquema de hashing simple (el que minimiza δ_1), es el obtenido con matrices construidas de la siguiente manera: cada renglon es obtenido a partir del anterior, sumando 1 módulo m. Estas matrices son muy prácticas, pues uno solo ne

cesita conocer un renglón, para conocerla completamente. Un ejemplo de estas matrices es:

$$S=\begin{pmatrix} 1 & 2 & 4 & 3 \\ 2 & 3 & 1 & 4 \\ 3 & 4 & 2 & 1 \\ 4 & 1 & 3 & 2 \end{pmatrix}$$

Este esquema, se conoce como *hashing cíclico generalizado*.

En éste campo, hay varias conjeturas, pero es difícil fundamentarlas empíricamente y las demostraciones formales parecen necesitar un método aún no descubierto.

REFERENCIAS

1. Coxeter H.S.M. *Introduction to Geometry. J Wiley (1961)*

2. Floyd R.W. *Nondeterministic Algorithms. JACM Vol 14 No. 4 Octubre 1967.*

3. Grenander Ulf. *Statistical Geometry: a tool for Pattern Analysis. Bull. of the Am. Math. So. Vol 79 No. 5 Septiembre 1973.*

4. Kendall M.G. & Moran P.A.P. *Geometrical Probability Griffin's monographs 1963.*

5. Knuth D.E. *The art of computer programming, Vol. I Fundamental algorithms, 1973; Vol II Seminumerical algorithms, 1969; Vol III Sorting and Searching 1973 Addison-Wesley Publishing Co.*

6. Knuth D.E. *Computer science and its relations to mathematics. Am. Math. Monthly Abril 1974.*

7. Mendel & Fu *Adaptive, Learning and Pattern Recognition Systems. Theory and Practice. Academic Press 1970.*

8. Schwartz J.T. *Semantic and Syntactic Issues in Programming. Bull. Am. Math. So. Vol 80 No2 Marzo 1974.*

9. Wilks S.S. *Mathematical Statistics. J. Wiley 1962.*