TEMat

divulgación de trabajos de estudiantes de matemáticas

e-ISSN 2530-9633

VOL. 3 — MAYO 2019

Asociación Nacional de Estudiantes de Matemáticas

TEMat

divulgación de trabajos de estudiantes de matemáticas

volumen 3
mayo de 2019

https://temat.es/volumen/2019/
https://anemat.com/

Una iniciativa de la
Asociación Nacional de Estudiantes de Matemáticas

Publica

Asociación Nacional de Estudiantes de Matemáticas
Plaza de las Ciencias, 3
Despacho 525, Facultad de Ciencias Matemáticas
Universidad Complutense de Madrid
28040 – Madrid

temat@temat.es
publicaciones@anemat.com
contacto@anemat.com

Colabora

Real Sociedad Matemática Española
Plaza de las Ciencias, 3
Despacho 525, Facultad de Ciencias Matemáticas
Universidad Complutense de Madrid
28040 – Madrid

Diseño de portada: Roberto Berná Larrosa, rbernalarrosa@gmail.com

TEMat, divulgación de trabajos de estudiantes de matemáticas – volumen 3 – mayo de 2019
e-ISSN: 2530-9633
https://temat.es/

Equipo

Sobre TEMat

TEMat es una revista de divulgación de trabajos de estudiantes de matemáticas publicada sin ánimo de lucro por la Asociación Nacional de Estudiantes de Matemáticas. Se busca publicar trabajos divulgativos de matemáticas, escritos principalmente (pero no exclusivamente) por estudiantes, de todo tipo: breves reseñas, introducciones a temas de investigación complejos, o artículos explicando las bases e incluso algún pequeño resultado de trabajos desarrollados por estudiantes.

TEMat persigue el doble objetivo de dar visibilidad a la calidad y diversidad de los trabajos realizados por estudiantes de matemáticas en los centros españoles a la vez que permite a los estudiantes publicar sus primeros artículos, familiarizándose así con el proceso de redacción, revisión y corrección que va asociado a la actividad investigadora.

Se contemplan para su publicación artículos escritos en castellano de todas las áreas de las matemáticas, incluyendo álgebra, análisis, ciencias de la computación, combinatoria, educación matemática, estadística, geometría, teoría de números y cualquier otra área de las matemáticas (puras y aplicadas), así como aplicaciones científicas o tecnológicas en las que las matemáticas jueguen un papel central.

Índice general

Carta del presidente de la ANEM

Con este tercer volumen de *TEMat*, la Asociación Nacional de Estudiantes de Matemáticas sigue adelante consolidando este ambicioso proyecto, único a nivel internacional. Cada vez son más los artículos que recibe el comité editorial, muestra innegable de la madurez que ha obtenido *TEMat* en estos tres años de vida y de la ilusión que tiene el estudiantado con la revista y con fomentar la ciencia en España.

Asimismo, desde el presente volumen, la Real Sociedad Matemática Española (RSME) va a colaborar con *TEMat*, por lo que aprovechamos para expresar el agradecimiento de la ANEM. Su colaboración servirá para dar aún más difusión a *TEMat* y para ayudar en la revisión de algunos artículos, lo que aportará a la revista más visibilidad y capacidad de edición y revisión.

Siguiendo las líneas de crecimiento, pronto se publicarán los primeros volúmenes temáticos de *TEMat*, que servirán para acercar los congresos de jóvenes estudiantes a la publicación de artículos científicos.

Nada de esto sería posible sin el desinteresado trabajo de todos los revisores de los artículos. Gracias a ellos, *TEMat* puede seguir creciendo. Para finalizar, me gustaría agradecer el incansable e intachable trabajo de todo el comité editorial: nada de esto hubiera sido posible sin vosotros.

Guillem García Subies,
presidente de la ANEM.

Madrid, mayo de 2019.

TEMat

El problema de las sumas de dos cuadrados

Alberto Cobos Rábano[a]
KU Leuven
albertocobosrabano@gmail.com

Resumen: Con el pretexto de resolver el problema clásico de la representación como suma de dos cuadrados en teoría de números, introduciremos una serie de conceptos fundamentales de este campo, como son las funciones multiplicativas o la descomposición de primos en anillos de enteros. Veremos también la relación entre la teoría de números y otras ramas de las matemáticas, pues la resolución del problema se basa en el estudio de los enteros gaussianos y de la divisibilidad en este anillo. Concluiremos demostrando qué enteros son sumas de dos cuadrados de enteros, y de cuántas maneras distintas.

Abstract: With the pretext of solving the classical number theory problem of representations as a sum of two squares, we shall introduce a series of fundamental concepts of this field, such as multiplicative functions or factorizations of primes in rings of integers. We shall also see the connection between number theory and other fields of mathematics, as the solution of the problem is based on the study of Gaussian integers and their divisibility. We shall finish by showing which integers are a sum of two squares of integers, and in how many different ways.

Palabras clave: teoría de números, sumas de dos cuadrados, función multiplicativa, enteros gaussianos, primos gaussianos.

MSC2010: 11E25.

Recibido: 23 de agosto de 2018.
Aceptado: 13 de septiembre de 2018.

Agradecimientos: Quiero agradecer al profesor Luis Manuel Navas, de la Universidad de Salamanca, todo el tiempo que me ha dedicado y todo lo que he aprendido de él en el desarrollo de mi Trabajo de Fin de Grado, del cual se extrae este artículo.

Referencia: Cobos Rábano, Alberto. «El problema de las sumas de dos cuadrados». En: *TEMat*, 3 (2019), págs. 1-16. ISSN: 2530-9633. URL: https://temat.es/articulo/2019-p1.

[a] El autor estaba afiliado a la Universidad de Salamanca (USAL) durante la realización de este trabajo.

1. Introducción

El objetivo de este artículo es resolver el siguiente problema clásico de la teoría de números.

Problema 1. Determinar qué números naturales son suma de dos cuadrados de enteros. ◂

En otras palabras, queremos determinar qué $n \in \mathbb{N}$ se pueden expresar en la forma $n = a^2 + b^2$ para ciertos $a, b \in \mathbb{Z}$. Con esto ya podemos dar nuestra primera definición.

Definición 2. Diremos que una pareja $(a, b) \in \mathbb{Z}^2$ es una **representación de $n \in \mathbb{N}$ como suma de dos cuadrados** si $a^2 + b^2 = n$. Por abuso de notación, diremos también que $a^2 + b^2$ es una **representación** de n, considerando relevante tanto el orden como el signo de a y b. ◂

Relacionado con el problema de determinar los naturales que son representables, queremos resolver también el siguiente problema.

Problema 3. Dado $n \in \mathbb{N}$ representable como suma de dos cuadrados, dar una fórmula cerrada para el número de representaciones distintas de n como suma de dos cuadrados. ◂

La primera referencia histórica a problemas de sumas de cuadrados es el problema 8 de la *Aritmética* de Diofanto, que pide escribir un cuadrado como suma de dos cuadrado. Dicho problema fue retomado por Fermat, quien enunció también el problema 1. Véase el libro de Weil [9] para más información.

Comencemos con un ejemplo para comprender los problemas 1 y 3.

Ejemplo 4. Para números «pequeños» es fácil hacer las comprobaciones pertinentes a mano. Un ejemplo de número que es suma de dos cuadrados, junto con todas las representaciones posibles, es el siguiente:

$$\begin{aligned} 10 &= 1^2 + 3^2 = (-1)^2 + 3^2 = 1^2 + (-3)^2 = (-1)^2 + (-3)^2 \\ &= 3^2 + 1^2 = 3^2 + (-1)^2 = (-3)^2 + 1^2 = (-3)^2 + (-1)^2. \end{aligned}$$

Este ejemplo nos sirve para aclarar que contamos como representaciones distintas las permutaciones y los cambios de signo; es decir, estamos contando las soluciones $(x, y) \in \mathbb{Z}^2$ de la ecuación $x^2 + y^2 = n$, para cada n fijo. También 4 es suma de dos cuadrados:

$$4 = 2^2 + 0^2 = (-2)^2 + 0^2 = 0^2 + 2^2 = 0^2 + (-2)^2.$$

En este caso existen cuatro representaciones en lugar de ocho, porque cambiar de signo a 0 es no hacer nada. También es fácil comprobar que 3 no es suma de dos cuadrados, luego el problema 1 no es trivial, pues hay números que sí son representables y otros que no. También se puede comprobar que 1996 no es suma de dos cuadrados, o que

$$\begin{aligned} 2018 &= 43^2 + 13^2 = (-43)^2 + 13^2 = 43^2 + (-13)^2 = (-43)^2 + (-13)^2 \\ &= 13^2 + 43^2 = (-13)^2 + 43^2 = 13^2 + (-43)^2 = (-13)^2 + (-43)^2 \end{aligned}$$

son todas las representaciones de 2018 como sumas de dos cuadrados, para lo cual es recomendable leer primero este documento en lugar de hacer todos los cálculos «por fuerza bruta». ◂

Antes de continuar, introducimos algunos conceptos y notaciones.

Definición 5. Dados $n, k \in \mathbb{N}$, se denomina **conjunto de representaciones de n como suma de k cuadrados** y se denota $R_k(n)$ al conjunto

$$R_k(n) = \{(x_1, \dots, x_k) \in \mathbb{Z}^k \,:\, x_1^2 + \dots + x_k^2 = n\}.$$

Estamos interesados en calcular su cardinal, para lo cual consideramos la función que sobre $n \in \mathbb{N}$ toma el valor

$$r_k(n) = \#R_k(n),$$

denominada **función de suma de k cuadrados** o **función de representación** (como suma de k cuadrados). Por tanto, $r_k(n)$ es el número de representaciones de n como suma de k cuadrados de enteros, contando como distintos los cambios de signo y las permutaciones[1]. ◂

[1] También tiene sentido considerar el caso $n = 0$, siendo $R_k(0) = \{(0, \overset{k}{\dots}, 0)\}$ y, por tanto, $r_k(0) = 1$ para todo $k \in \mathbb{N}$.

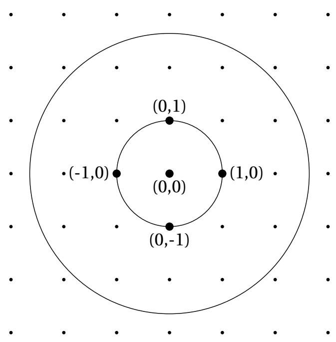

Figura 1: Representación gráfica de $R_2(1) = S^1(0,1) \cap \mathbb{Z}^2$ y $R_2(7) = S^1(0,\sqrt{7}) \cap \mathbb{Z}^2 = \emptyset$.

De la definición surge una duda curiosa: ¿por qué nos interesa contar como distintas las permutaciones y los cambios de signo? Un motivo es que la interpretación geométrica de $R_k(n)$ es que estos son los puntos que pertenecen a la esfera de centro 0 y radio $\sqrt{n}$ en $\mathbb{R}^k$ y tienen coordenadas enteras (véase la figura 1, en la que hemos representado $R_2(1)$ y $R_2(7) = \emptyset$). Por tanto, geométricamente, está claro que las permutaciones y los cambios de signo nos dan puntos distintos y deberían contarse como tales.

Ejemplo 6. Un ejemplo trivial de función de representación es el caso $k = 1$. Para $n \in \mathbb{N}$ se tiene que

$$r_1(n) = \begin{cases} 2 & \text{si } n \text{ es un cuadrado perfecto,} \\ 0 & \text{si } n \text{ no es un cuadrado perfecto,} \end{cases}$$

siendo en el primer caso $\sqrt{n}$ y $-\sqrt{n}$ las dos representaciones distintas. ◂

Sobre las funciones de representación hay muchos resultados conocidos, como por ejemplo el teorema de los cuatro cuadrados de Lagrange, que afirma que $r_4(n) > 0$ para todo $n \in \mathbb{N}$; es decir, que todo entero no negativo es suma de cuatro cuadrados de enteros. De hecho, 4 es el mínimo valor de k con esta propiedad, pues ya hemos visto que 3 no es suma de dos cuadrados y no es complicado comprobar que 7 no es suma de tres cuadrados. También se conoce una fórmula para $r_4(n)$, aunque no abordaremos aquí su demostración. No obstante, no todo está dicho sobre este tipo de problemas. Por ejemplo, el problema de Waring, «para cada $\ell \in \mathbb{N}$, ¿existe un $k \in \mathbb{N}$ tal que cada $n \in \mathbb{N}$ es suma de a lo más k potencias ℓ-ésimas?», no está completamente resuelto, pues se sabe que la respuesta es afirmativa pero se desconocen los valores óptimos de k.

Concluimos esta introducción con una pregunta: ¿qué herramientas son necesarias para resolver los problemas 1 y 3? La base fundamental que vamos a utilizar es el álgebra. Dedicaremos la sección 2 a recordar los conceptos y resultados básicos de la teoría de divisibilidad, que trata de generalizar conceptos como división, máximo común divisor o elemento primo más allá de los números enteros, y se estudia en cursos básicos de cualquier grado en Matemáticas. Aplicaremos esta teoría al estudio de los enteros gaussianos, que son los elementos de la forma $a + bi$ con $a, b \in \mathbb{Z}$ pensados como subanillo de los números complejos. En particular, queremos determinar los primos gaussianos, y relacionarlos con la solución del problema 1. Por último, introduciremos un concepto de teoría de números como son las funciones multiplicativas que nos permitirá, junto con toda la información antes recabada, solucionar el problema 3.

2. Un repaso de teoría de divisibilidad

Para facilitar la comprensión de las demás secciones, hemos decidido añadir este repaso sobre teoría de la divisibilidad. Asumimos que el lector está familiarizado con la teoría de divisibilidad en anillos como se estudia en muchos cursos introductorios de álgebra abstracta; en particular, suponemos que conoce los conceptos de elementos primos e irreducibles, unidades y asociados, y máximo común divisor. Recogeremos los principales resultados que nos harán falta posteriormente, aunque no incluiremos su demostración, al no ser este el objetivo de este documento. Puede consultarse el libro de Jacobson [6, capítulo 2].

En lo sucesivo denotaremos por A a un anillo conmutativo con unidad $1 \neq 0$[2] y sin divisores de cero (esto es, un dominio de integridad[3]) y llamaremos indistintamente **unidad** o elemento **invertible** a cualquier $a \in A$ tal que existe $b \in A$ con $ab = 1$.

Definición 7. Un elemento $d \in A$ se dice que **divide** a otro elemento $a \in A$ y se denota $d \mid a$ si $a = md$ para algún $m \in A$. En ese caso, se dice que d es un **divisor** o un **factor** de a, mientras que la igualdad $a = md$ es una **factorización** de a. En términos de ideales, $d \mid a \iff (a) \subseteq (d)$. ◂

Definición 8. Dos elementos $a, b \in A$ se dice que son **asociados** (o que a es un asociado de b, o que b es un asociado de a) si generan el mismo ideal; es decir, si $(a) = (b)$. En términos de divisibilidad, esto equivale a que $a \mid b$ y $b \mid a$. Escribiremos $a \sim b$ si a y b son asociados. En un dominio de integridad, es fácil comprobar que $a \sim b$ si y solo si $b = \epsilon a$ para alguna unidad ϵ. Claramente, ser asociados es una relación de equivalencia en A. ◂

Definición 9. Sea $c \in A$ no nulo y no invertible. Dada una factorización $c = ab$, diremos que a es un **factor (o divisor) no trivial** si a no es invertible y no es asociado de c. En otro caso, decimos que a es un **factor trivial**. En caso de que tanto a como b sean factores no triviales, decimos que la factorización $c = ab$ es una **factorización no trivial**. ◂

Definición 10. Un elemento no invertible $p \in A$, $p \neq 0$, se dice que es **irreducible** si p solamente admite factorizaciones triviales; esto es, si $p = ab$, entonces o bien a o bien b es una unidad. Por el contrario, si $\alpha \in A$ es no nulo y no invertible y admite alguna factorización no trivial, se dice que α es **compuesto**. ◂

Definición 11. Un elemento no invertible $p \in A$, $p \neq 0$, se dice que es **primo** si para cualesquiera $a, b \in A$ tales que $p \mid ab$, o bien $p \mid a$ o bien $p \mid b$. En términos de ideales, esto significa que (p) es un ideal primo (no nulo). ◂

Lema 12. *En un dominio íntegro, los elementos primos son irreducibles.*

Lema 13. *En un dominio de ideales principales*[4] *(DIP) A, un elemento no nulo y no invertible $a \in A$ es irreducible si y solo si el ideal (a) es maximal. En particular, los elementos irreducibles son primos.*

Proposición 14 (unicidad de la factorización en primos en cualquier dominio). *Sea A un dominio de integridad y sea $a \in A$ un elemento no nulo y no invertible. Si a es producto de primos*[5]*, entonces esta factorización es única salvo reordenaciones y asociados. En otras palabras, si*

$$a = p_1 p_2 \cdots p_n = q_1 q_2 \cdots q_m,$$

donde los p_i y los q_j son primos (no necesariamente distintos) y $n, m \in \mathbb{N}$, entonces $n = m$ y, tras una reordenación adecuada, $p_i \sim q_i$ para todo i.

Proposición 15 (existencia de factorización en irreducibles en dominios noetherianos). *En un dominio íntegro noetheriano*[6] *A, todo elemento no nulo y no invertible es producto finito de elementos irreducibles. En particular, esto es válido para un DIP.*

Definición 16. Un **dominio de factorización única** (DFU) es un dominio íntegro en el que todo elemento no nulo y no invertible es producto de irreducibles de manera única salvo reordenaciones y asociados. ◂

Lema 17. *En un dominio de factorización única, todo elemento irreducible es primo.*

Teorema 18. *Un dominio de ideales principales es un dominio de factorización única.*

Definición 19. Un **domino euclídeo** es un dominio íntegro A en el que existe una función $\nu : A \setminus \{0\} \to \mathbb{Z}_{\geq 0}$ tal que, si $a, b \in A$ y $b \neq 0$, existen $q, r \in A$ con $a = qb + r$ y o bien $r = 0$ o bien $\nu(r) < \nu(b)$[7]. ◂

[2] Esta condición sirve para excluir el caso del anillo $A = 0$.

[3] Recordamos que un anillo conmutativo con unidad A se dice que es **íntegro** o que es un **dominio de integridad** si para cualesquiera $a, b \in A$ se verifica la condición $ab = 0 \implies a = 0$ o $b = 0$.

[4] Recordamos que un **dominio de ideales principales** es un dominio de integridad A en el que todo ideal es de la forma (a) para algún $a \in A$.

[5] En general, un elemento de un dominio de integridad puede no tener descomposición como producto de primos. Por ello introducimos a continuación el concepto de dominio de factorización única.

[6] Recordamos que un anillo es **noetheriano** si verifica la condición de cadena ascendente; es decir, si para cada cadena de ideales $I_1 \subseteq ... \subseteq I_n \subseteq ...$ existe un n tal que $I_n = I_{n+m}$ para todo $m \in \mathbb{N}$.

[7] Sin pérdida de generalidad, podemos asumir que $\nu(a) \leq \nu(ab)$ para $a, b \in A$ con $a, b \neq 0$. A veces se considera esta desigualdad como parte de la definición de dominio euclídeo.

Lema 20. *Un domino euclídeo es un dominio de ideales principales.*

Corolario 21. *Un dominio euclídeo es un dominio de factorización única.*

Definición 22. Un máximo común divisor (mcd) de dos elementos a, b en un dominio de integridad A es un divisor común d de a y b que es maximal en la relación de divisibilidad; esto es, un $d \in A$ tal que $d \mid a$ y $d \mid b$, y si δ es cualquier otro elemento con la misma propiedad, entonces $\delta \mid d$[8]. Decimos que a, b son **primos relativos** o **coprimos** si su mcd existe y es una unidad. El máximo común divisor de dos elementos es único salvo asociados. ◂

Teorema 23 (lema de Bezout). *Si A es un DIP, un máximo común divisor de dos elementos $a, b \in A$ es cualquier elemento d tal que $(a) + (b) = (d)$.*

Concluimos con una lista de sencillas observaciones sobre el máximo común divisor y algunos ejemplos de los conceptos que hemos introducido previamente.

Corolario 24 (propiedades de divisibilidad). *Sea A un DIP y sean $a, b, c, a', b' \in A$. Entonces,*

- *Los elementos $a, b \in A$ son coprimos si y solo si existen $s, t \in A$ tales que $sa + tb = 1$.*
- *Si a, b son coprimos y $a \mid bc$, entonces $a \mid c$.*
- *Si a, b son coprimos y $a \mid c$, $b \mid c$, entonces $ab \mid c$.*
- *Los elementos a, b son ambos coprimos con c si y solo si ab es coprimo con c.*
- *Si $ab = a'b'$ y tanto a, b' como a', b son coprimos, entonces $a \sim a'$ y $b \sim b'$.*

Ejemplo 25. El anillo $\mathbb{Z}$ es un dominio euclídeo tomando como ν la función valor absoluto, por lo que también es un DIP (lema 20) y un DFU (corolario 21).

Si k es un cuerpo, es bien conocido que sus únicos ideales son (0) y $k = (1)$, luego k es un DIP y también un DFU (teorema 18); es más, se comprueba fácilmente que el anillo de polinomios en una variable con coeficientes en k, que denotamos por $k[x]$, es un dominio euclídeo tomando $\nu(p(x))$ igual al grado de $p(x)$.

En general, si un anillo A es un DFU, también $A[x]$ es un DFU; en particular, $\mathbb{Z}[x]$ es un DFU. Si $\mathbb{Z}[x]$ fuera un DIP, entonces todo elemento irreducible generaría un ideal maximal (lema 13); en particular, como x es irreducible, (x) sería maximal y, por tanto, $\mathbb{Z}[x]/(x) \simeq \mathbb{Z}$ sería un cuerpo, lo cual es claramente falso. Es decir, $\mathbb{Z}[x]$ es un DFU que no es un DIP, y, por tanto, tampoco es un dominio euclídeo (lema 20). ◂

3. Enteros gaussianos

Tras haber refrescado conceptos como la divisibilidad, los dominios de ideales principales, los dominios de factorización única, etc., es ahora el momento de ponerlos en práctica. Para ello, vamos a introducir los enteros gaussianos. Comenzaremos definiendo este anillo y exponiendo algunas de sus propiedades básicas, para enfrascarnos enseguida en la demostración de que los enteros gaussianos forman un dominio euclídeo, pudiendo, por tanto, aplicar los conocimientos adquiridos en la sección 2. Concluiremos esta sección mostrando que existe una estrecha relación entre divisibilidad en $\mathbb{Z}$ y divisibilidad en los enteros gaussianos. Todos estos detalles constituyen los primeros pasos hacia nuestro objetivo final: resolver los problemas 1 y 3. Para esta sección y la siguiente tomamos como referencia el artículo divulgativo de Conrad [2].

Definición 26. Se denomina **anillo de enteros gaussianos** al subanillo $\mathbb{Z}[\mathrm{i}]$ de $\mathbb{C}$ definido como

$$\mathbb{Z}[\mathrm{i}] = \{a + b\mathrm{i} : a, b \in \mathbb{Z}\}.$$ ◂

En los números complejos, dado $z = x + y\mathrm{i}$, se define su conjugado como $\overline{z} = x - y\mathrm{i}$. Esta operación se restringe bien a los enteros gaussianos pues, si $\alpha \in \mathbb{Z}[\mathrm{i}]$, entonces $\overline{\alpha} \in \mathbb{Z}[\mathrm{i}]$. Por otro lado, como $\mathbb{C}$ es un cuerpo, $\mathbb{Z}[\mathrm{i}]$ es un anillo íntegro. Además, podemos pensar que los enteros son enteros gaussianos mediante el morfismo inyectivo de anillos $n \mapsto n + 0\mathrm{i}$ para cada $n \in \mathbb{Z}$.

[8]Como es de esperar, si tomamos $A = \mathbb{Z}$ el concepto de máximo común divisor que acabamos de definir coincide con el habitual, y es único salvo cambio de signo.

Ejemplo 27. Dados $\alpha, \beta \in \mathbb{Z}[\mathrm{i}]$, podemos determinar si $\alpha \mid \beta$ en $\mathbb{Z}[\mathrm{i}]$ dividiendo en $\mathbb{C}$. Esto nos dará un cociente de elementos de $\mathbb{Z}[\mathrm{i}]$ (que pertenece, por tanto, al cuerpo de fracciones, $\mathbb{Q}[\mathrm{i}]$) y se trata de comprobar si este cociente vuelve a ser un elemento de $\mathbb{Z}[\mathrm{i}]$. Por ejemplo,

$$\frac{14+3\mathrm{i}}{4+5\mathrm{i}} = \frac{(14+3\mathrm{i})(4-5\mathrm{i})}{16+25} = \frac{71-58\mathrm{i}}{41} \notin \mathbb{Z}[\mathrm{i}],$$

lo cual demuestra que $(4+5\mathrm{i}) \nmid (14+3\mathrm{i})$ en $\mathbb{Z}[\mathrm{i}]$. ◂

También podemos restringir a los enteros gaussianos la norma que a $z = x + y\mathrm{i} \in \mathbb{C}$ le asigna el valor

$$N(z) = z\overline{z} = (x+y\mathrm{i})(x-y\mathrm{i}) = x^2 + y^2.$$

Está claro, además, que si $\alpha \in \mathbb{Z}[\mathrm{i}]$, entonces $N(\alpha) \in \mathbb{Z}_{\geq 0}$. La norma N nos facilita información sobre el anillo de enteros gaussianos. Además, utilizando que la conjugación compleja es multiplicativa (es decir, $\overline{z_1 z_2} = \overline{z_1} \cdot \overline{z_2}$) se comprueba fácilmente que también es multiplicativa la norma N.

Una propiedad fundamental sobre $\mathbb{Z}[\mathrm{i}]$ es que se trata de un dominio euclídeo y, por tanto, hay factorización única como producto de irreducibles (que son lo mismo que los primos) a partir de los resultados mencionados en la sección previa.

Teorema 28. *$\mathbb{Z}[\mathrm{i}]$ es un dominio euclídeo con respecto a la norma N. En otras palabras, dados $\alpha, \beta \in \mathbb{Z}[\mathrm{i}]$ con $\beta \neq 0$, existen $\gamma, \rho \in \mathbb{Z}[\mathrm{i}]$ tales que $\alpha = \beta\gamma + \rho$ y $N(\rho) < N(\beta)$. De hecho, podemos elegir ρ tal que $N(\rho) \leq \frac{1}{2}N(\beta)$.*

Demostración. Sean $\alpha, \beta \in \mathbb{Z}[\mathrm{i}]$ con $\beta \neq 0$. Dividimos en $\mathbb{C}$ para obtener que

$$\frac{\alpha}{\beta} = \frac{\alpha\overline{\beta}}{\beta\overline{\beta}} = \frac{\alpha\overline{\beta}}{N(\beta)} = \frac{m+n\mathrm{i}}{N(\beta)},$$

donde $m + n\mathrm{i} = \alpha\overline{\beta} \in \mathbb{Z}[\mathrm{i}]$. Usando que $\mathbb{Z}$ es un dominio euclídeo, podemos encontrar $q_1, q_2, r_1, r_2 \in \mathbb{Z}$ tales que

$$m = N(\beta)q_1 + r_1, \qquad n = N(\beta)q_2 + r_2$$

y $0 \leq |r_1|, |r_2| \leq \frac{1}{2}N(\beta)$. Esto es consecuencia de permitir restos negativos; es decir, en lugar de tomar restos $r_1, r_2 \in \{0, 1, \dots, N(\beta)-1\}$, permitimos elegir los restos de entre los enteros del intervalo $[-N(\beta)/2, \dots, N(\beta)/2)$. Entonces,

$$\frac{\alpha}{\beta} = \frac{N(\beta)q_1 + r_1 + (N(\beta)q_2 + r_2)\mathrm{i}}{N(\beta)} = q_1 + q_2\mathrm{i} + \frac{r_1 + r_2\mathrm{i}}{N(\beta)}.$$

Tomamos $\gamma = q_1 + q_2\mathrm{i}$, reordenamos y multiplicamos por β, para obtener que

$$\alpha - \beta\gamma = \frac{r_1 + r_2\mathrm{i}}{\overline{\beta}}.$$

Basta con comprobar que $N(\alpha - \beta\gamma) \leq \frac{1}{2}N(\beta)$, y tomar $\rho = \alpha - \beta\gamma$. Ahora bien, tomando normas a ambos lados, como $N(\overline{\beta}) = N(\beta)$, queda que

$$N(\alpha - \beta\gamma) = \frac{r_1^2 + r_2^2}{N(\beta)} \leq \frac{\frac{1}{4}N(\beta)^2 + \frac{1}{4}N(\beta)^2}{N(\beta)} = \frac{1}{2}N(\beta). \qquad \blacksquare$$

A continuación, presentamos una serie de resultados que relacionan la divisibilidad en $\mathbb{Z}$ y en $\mathbb{Z}[\mathrm{i}]$ por medio de la norma N. Algunas de estas propiedades serán claves para resolver el problema de los dos cuadrados.

Lema 29. *Sean $c \in \mathbb{Z}$ y $\alpha = a + b\mathrm{i} \in \mathbb{Z}[\mathrm{i}]$. Entonces, $c \mid \alpha$ en $\mathbb{Z}[\mathrm{i}]$ si y solo si $c \mid a$ y $c \mid b$ en $\mathbb{Z}$. En particular, tomando $b = 0$ se deduce que $c \mid a$ en $\mathbb{Z}[\mathrm{i}]$ si y solo si $c \mid a$ en $\mathbb{Z}$.*

Demostración. Si $\beta = m + n\mathrm{i} \in \mathbb{Z}[\mathrm{i}]$, $c\beta = \alpha \iff cm + cn\mathrm{i} = a + b\mathrm{i} \iff cm = a$ y $cn = b$. ■

Proposición 30 (propiedades de divisibilidad sobre la norma). *Sean $\alpha, \beta \in \mathbb{Z}[\mathrm{i}]$.*

1. *Si $\beta \mid \alpha$, entonces $N(\beta) \mid N(\alpha)$.*
2. *α es una unidad si y solo si $N(\alpha) = 1$.*
3. *Las unidades gaussianas son $\mathbb{Z}[\mathrm{i}]^* = \{\pm 1, \pm \mathrm{i}\}$.*
4. *Si $\alpha \neq 0$ y $\beta \mid \alpha$, entonces $N(\beta) = N(\alpha)$ si y solo si $\beta \sim \alpha$.*
5. *$N(\mathrm{mcd}_{\mathbb{Z}[\mathrm{i}]}(\alpha, \beta)) \mid \mathrm{mcd}_{\mathbb{Z}}(N(\alpha), N(\beta))$.*
6. *Si $\alpha, \beta \neq 0$, entonces cualquier divisor común de α, β de norma máxima es un mcd.*
7. *Si α, β tienen normas coprimas en $\mathbb{Z}$, entonces α, β son coprimos en $\mathbb{Z}[\mathrm{i}]$.*
8. *Si $N(\alpha)$ es primo en $\mathbb{Z}$, entonces α es primo en $\mathbb{Z}[\mathrm{i}]$.*

Demostración.

1. Es consecuencia inmediata de ser N multiplicativa.
2. Si $\alpha\beta = 1$, entonces $N(\alpha)N(\beta) = N(\alpha\beta) = N(1) = 1$ y, como $N(\alpha), N(\beta) \in \mathbb{Z}_{\geq 0}$, se debe dar que $N(\alpha) = N(\beta) = 1$. Recíprocamente, $1 = N(\alpha) = \alpha\overline{\alpha}$ indica que $\overline{\alpha}$ es el inverso de α.
3. Basta resolver $a^2 + b^2 = 1$ para $a, b \in \mathbb{Z}$.
4. Dado $\alpha \neq 0$, si $\beta \mid \alpha$ tenemos que $\alpha = \gamma\beta$. Entonces, $N(\alpha) = N(\gamma)N(\beta)$, luego $N(\alpha) = N(\beta)$ si y solo si $N(\gamma) = 1$, que por la propiedad 2 equivale a que $\gamma \in \mathbb{Z}[\mathrm{i}]^*$ y, por tanto, $\beta \sim \alpha$.
5. Por la propiedad 2, la norma es independiente de la elección de asociados. Si $\delta \mid \alpha, \beta$, por la propiedad 1 tenemos que $N(\delta) \mid N(\alpha), N(\beta)$ y, por tanto, $N(\delta) \mid \mathrm{mcd}_{\mathbb{Z}}(N(\alpha), N(\beta))$. En particular, esto es válido para $\delta = \mathrm{mcd}_{\mathbb{Z}[\mathrm{i}]}(\alpha, \beta)$.
6. Sea δ un mcd de α, β y sea d un divisor común de α, β de norma máxima. Por definición, $d \mid \delta$, luego por la propiedad 1 tenemos que $N(d) \mid N(\delta)$ y, por tanto, $N(d) \leq N(\delta)$, así que por maximalidad de $N(d)$ se tiene que $N(d) = N(\delta)$. De la propiedad 4 deducimos que $d \sim \delta$, luego d también es un mcd de α, β.
7. Por la propiedad 5, si $\mathrm{mcd}_{\mathbb{Z}}(N(\alpha), N(\beta)) = 1$, entonces $N(\mathrm{mcd}_{\mathbb{Z}[\mathrm{i}]}(\alpha, \beta)) = 1$, luego de la propiedad 2 se deduce que $\mathrm{mcd}_{\mathbb{Z}[\mathrm{i}]}(\alpha, \beta)$ es una unidad, es decir, α y β son coprimos en $\mathbb{Z}[\mathrm{i}]$.
8. Si $N(\alpha) = p$ para algún primo p y $\alpha = \beta\gamma$, entonces $p = N(\beta)N(\gamma)$. Utilizando la factorización única en $\mathbb{N}$ deducimos que $N(\beta) = 1$ o $N(\gamma) = 1$; por tanto, o bien β o bien γ es una unidad. ■

Observación 31. La propiedad 1 no es una equivalencia. En el ejemplo 27 vimos que $(4 + 5\mathrm{i}) \nmid (14 + 3\mathrm{i})$, pero $N(4 + 5\mathrm{i}) = 41 \mid 205 = N(14 + 3\mathrm{i})$.

La propiedad 4 solo es cierta bajo la hipótesis de que $\beta \mid \alpha$. Es fácil encontrar enteros gaussianos que tengan igual norma pero no sean asociados. Por ejemplo, $\alpha = 2 + \mathrm{i}$, $\beta = 2 - \mathrm{i}$ tienen norma 5; $\alpha = 3 + 4\mathrm{i}$, $\beta = 5$ tienen norma 25, y $1 + 8\mathrm{i}$, $4 + 7\mathrm{i}$ tienen norma 65.

Del mismo modo, la propiedad 8 permite comprobar la primalidad en $\mathbb{Z}[\mathrm{i}]$ a través de la norma, pero no es una equivalencia. Por ejemplo, es fácil comprobar que 3 es primo en $\mathbb{Z}[\mathrm{i}]$ pero tiene norma 9. Esto muestra, además, que no es cierto en general que $N(\mathrm{mcd}_{\mathbb{Z}[\mathrm{i}]}(\alpha, \beta)) = \mathrm{mcd}_{\mathbb{Z}}(N(\alpha), N(\beta))$. Por ejemplo, $2 - \mathrm{i}$ y $2 + \mathrm{i}$ son coprimos (pues son primos como consecuencia de la propiedad 8 y, como hemos dicho anteriormente, no son asociados), pero ambos tienen norma igual a 5. También $\alpha = 3 + 4\mathrm{i}$ y $\beta = 3 + \mathrm{i}$ son coprimos (pues $\alpha = (2 + \mathrm{i})^2$ y $\beta = (2 - \mathrm{i})(1 + \mathrm{i})$), mientras que $N(\alpha) = 25$, $N(\beta) = 10$, de modo que $\mathrm{mcd}_{\mathbb{Z}}(N(\alpha), N(\beta)) = 5$. ◂

Corolario 32. *Se verifican las siguientes afirmaciones:*

1. *Un entero gaussiano α es no nulo y no invertible si y solo si $N(\alpha) > 1$.*
2. *Supongamos que $N(\alpha) > 1$. Un factor β de α es trivial si y solo si $N(\beta) = 1$ (unidad) o $N(\beta) = N(\alpha)$ (asociado). Por tanto, existen ocho factores triviales de α, dados por las cuatro unidades $\pm 1, \pm \mathrm{i}$ y los cuatro asociados $\pm\alpha, \pm\mathrm{i}\alpha$. Además, un divisor β de α es no trivial si y solo si $1 < N(\beta) < N(\alpha)$.*

Demostración. Basta aplicar que N valora en $\mathbb{Z}_{\geq 0}$, la equivalencia entre tener norma nula y ser 0, y la propiedad 2 de la proposición 30 para demostrar la primera afirmación, mientras que la segunda afirmación se deduce de las propiedades 1, 2, 3 y 4 de la proposición 30. ■

4. Primos gaussianos y sumas de dos cuadrados

Nos proponemos ahora resolver el problema 1. Para ello, estudiaremos primero qué primos enteros son representables como sumas de dos cuadrados, lo cual está íntimamente relacionada con la divisibilidad en $\mathbb{Z}[\mathrm{i}]$ y da sentido al estudio que hemos hecho de este anillo. Esta relación motivará que dediquemos parte de nuestro tiempo a explorar los primos enteros que son primos gaussianos, la factorización en $\mathbb{Z}[\mathrm{i}]$ y a determinar todos los primos gaussianos. Concluiremos determinando la factorización de cualquier entero como producto de primos gaussianos, y utilizando este resultado para resolver el problema 1.

Comenzamos con una sencilla observación que nos ayuda a entender por qué los enteros gaussianos son esenciales para nuestro estudio.

Proposición 33. *Un entero n es representable como suma de dos cuadrados si y solo si existe un entero gaussiano α tal que $n = N(\alpha)$.*

Demostración. Esto es inmediato a partir de la expresión $N(a + b\mathrm{i}) = a^2 + b^2$. ∎

La proposición previa nos permite hacer una de las observaciones fundamentales, que era conocida por Fermat y Euler: la propiedad de ser representable como suma de dos cuadrados es una propiedad multiplicativa. Es decir, si $n_1, n_2, \dots, n_s \in \mathbb{N}$ son representables con $n_i = N(\alpha_i)$ para cada i, entonces $n_1 n_2 \cdots n_s = N(\alpha_1 \cdots \alpha_s)$ también es representable[9]. ¡Ojo!, no estamos diciendo que se trate de una equivalencia; es decir, puede suceder que $n_1 n_2 \cdots n_s$ sea representable como suma de dos cuadrados mientras que alguno de los n_i no sea representable. Siguiendo este razonamiento, estamos interesados en estudiar la representabilidad de los divisores y, por tanto, la representabilidad de los primos enteros, y en relacionar esta con la divisibilidad en $\mathbb{Z}[\mathrm{i}]$.

Teorema 34. *Un primo entero p es suma de dos cuadrados si y solo si p es compuesto en $\mathbb{Z}[\mathrm{i}]$. Es decir, p permanece primo en $\mathbb{Z}[\mathrm{i}]$ si y solo si p no es suma de dos cuadrados.*

Demostración. Si un primo entero p es de la forma $p = a^2 + b^2$, entonces $p = (a + b\mathrm{i})(a - b\mathrm{i})$ es una descomposición no trivial en $\mathbb{Z}[\mathrm{i}]$ por el corolario 32, pues $1 < N(a \pm b\mathrm{i}) = p < p^2 = N(p)$.

Recíprocamente, sea p un primo en $\mathbb{Z}$ que es compuesto en $\mathbb{Z}[\mathrm{i}]$, y sea $p = \alpha\beta$ una descomposición no trivial. Tomando normas, $p^2 = N(\alpha)N(\beta)$. Por el corolario 32, necesariamente se debe cumplir que $N(\alpha) = N(\beta) = p$. Por tanto, si $\alpha = a + b\mathrm{i}$, entonces $p = a^2 + b^2$. ∎

Corolario 35. *Sea $p \in \mathbb{N}$ un primo impar que es compuesto en $\mathbb{Z}[\mathrm{i}]$. Salvo asociados, p tiene dos factores primos distintos en $\mathbb{Z}[\mathrm{i}]$. Además, dichos factores son conjugados y tienen norma p.*

Demostración. De la demostración del teorema 34 sabemos que $N(a \pm b\mathrm{i}) = p$, luego por la propiedad 8 de la proposición 30, $a \pm b\mathrm{i}$ son primos gaussianos. Si $a + b\mathrm{i} = \epsilon(a - b\mathrm{i})$ para alguna unidad $\epsilon \in \mathbb{Z}[\mathrm{i}]$, entonces

$$\begin{cases} \epsilon = 1 & \Longrightarrow\ a + b\mathrm{i} = a - b\mathrm{i} & \Longrightarrow\ b = 0, & p = a^2, \\ \epsilon = -1 & \Longrightarrow\ a + b\mathrm{i} = -a + b\mathrm{i} & \Longrightarrow\ a = 0, & p = b^2, \\ \epsilon = \mathrm{i} & \Longrightarrow\ a + b\mathrm{i} = b + a\mathrm{i} & \Longrightarrow\ b = a, & p = 2a^2, \\ \epsilon = -\mathrm{i} & \Longrightarrow\ a + b\mathrm{i} = -b - a\mathrm{i} & \Longrightarrow\ b = -a, & p = 2a^2. \end{cases}$$

En cualquier caso, esto es imposible para un primo impar p, por lo que $a \pm b\mathrm{i}$ no son asociados. ∎

Comenzaremos estudiando los primos enteros que son sumas de dos cuadrados. Para ello, el teorema 34 nos indica que es conveniente estudiar la factorización en $\mathbb{Z}[\mathrm{i}]$ de los primos enteros. El siguiente lema nos indica que de este modo se recuperan todos los primos gaussianos.

Lema 36. *Todo primo gaussiano π divide a algún primo entero p en $\mathbb{Z}[\mathrm{i}]$.*

Demostración. Basta observar que $N(\pi) = \pi\overline{\pi}$ en $\mathbb{Z}[\mathrm{i}]$. Como $N(\pi) > 1$ y $N(\pi) \in \mathbb{N}$, existe la descomposición en primos enteros $\pi\overline{\pi} = N(\pi) = p_1 \cdots p_r$, y como π es primo, debe dividir a alguno de los p_i. ∎

[9] Esta sencilla idea, basada en la multiplicatividad de N, resulta clave para demostrar el teorema de los cuatro cuadrados de Lagrange que hemos enunciado más adelante como el teorema 58, pues en la práctica se demuestra que todo primo entero es representable.

Observación 37. El primo $p = 2$ factoriza como $2 = (1 + \mathrm{i})(1 - \mathrm{i})$, donde $1 \pm \mathrm{i}$ son primos gaussianos de norma 2, pero son asociados pues $1 - \mathrm{i} = -\mathrm{i}(1 + \mathrm{i})$, luego $2 = -\mathrm{i}(1 + \mathrm{i})^2$, apareciendo dos veces el primo gaussiano $1 + \mathrm{i}$. El hecho de que en una factorización aparezca un primo con multiplicidad mayor que 1 se conoce como **ramificación**. En general, se utiliza la siguiente terminología para clasificar la factorización de un primo entero en $\mathbb{Z}[\mathrm{i}]$. ◂

Definición 38. Sea $p \in \mathbb{N}$ un primo entero.

- Si p permanece primo en $\mathbb{Z}[\mathrm{i}]$, se dice que p es un primo **inerte**.
- Si p es compuesto en $\mathbb{Z}[\mathrm{i}]$, decimos que p es un primo que **descompone**.
- Si $p = 2$, se dice que 2 es el primo **ramificado**. ◂

Utilizando esta terminología, hemos visto hasta el momento que p es inerte si y solo si p no es suma de dos cuadrados; que los compuestos impares p descomponen con dos factores primos no asociados, y que $p = 2$ es ramificado y es asociado del cuadrado del primo $1 + \mathrm{i}$. Por otro lado, el teorema 34 ha resultado ser una herramienta útil en nuestro estudio, pero no resuelve el problema por sí solo. Necesitamos la caracterización de primos que son sumas de dos cuadrados, debida a Fermat.

Teorema 39 (clasificación de los primos que son sumas de dos cuadrados). *Sea p un primo entero. Las siguientes afirmaciones son equivalentes:*

1. $p = 2$ *o* $p \equiv 1 \text{ mód } 4$.
2. *La ecuación* $x^2 \equiv -1 \text{ mód } p$ *tiene solución; es decir,* -1 *es un cuadrado módulo* p.
3. $p = a^2 + b^2$ *para ciertos* $a, b \in \mathbb{Z}$.

En particular, los primos enteros p congruentes con $3 \text{ mód } 4$ *son precisamente los primos enteros que no son suma de dos cuadrados.*

Demostración. **3 $\Longrightarrow$ 1:** Basta utilizar que los cuadrados módulo 4 son 0 y 1, por lo que cualquier entero $n \equiv 3 \text{ mód } 4$ no puede ser suma de dos cuadrados en $\mathbb{Z}/(4)$, y menos aún en $\mathbb{Z}$.

1 $\Longrightarrow$ 2: Para $p = 2$, $x^2 \equiv -1 \text{ mód } p$ tiene solución $x = 1$. Para un primo impar p, podemos considerar la siguiente factorización en $\mathbb{F}_p[x]$:

$$x^{p-1} - 1 = (x^{\frac{(p-1)}{2}} - 1)(x^{\frac{(p-1)}{2}} + 1). \tag{1}$$

Aplicando el teorema de Fermat[10], todo elemento a no nulo de $\mathbb{F}_p$ verifica que $a^{p-1} - 1 \equiv 0 \text{ mód } p$ o, lo que es lo mismo, existen al menos $p - 1$ raíces distintas en $\mathbb{F}_p$ del polinomio $x^{p-1} - 1$. En el lado derecho de la ecuación (1), al ser $\mathbb{F}_p[x]$ un DFU (ejemplo 25), el primer polinomio del lado derecho no puede tener más raíces que su grado, que es $(p - 1)/2$, luego el segundo polinomio del lado derecho debe tener alguna raíz sobre $\mathbb{F}_p$; en otras palabras, existe $a \in \mathbb{Z}$ tal que $a^{(p-1)/2} \equiv -1 \text{ mód } p$. Si, además, $p \equiv 1 \text{ mód } 4$, entonces $(p - 1)/4 \in \mathbb{N}$ y $x = a^{(p-1)/4}$ satisface que $x^2 \equiv a^{(p-1)/2} \equiv -1 \text{ mód } p$.

2 $\Longrightarrow$ 3: Por el teorema 34, basta demostrar que p es compuesto en $\mathbb{Z}[\mathrm{i}]$. Si $x \in \mathbb{Z}$ es una solución de $x^2 \equiv -1 \text{ mód } p$, entonces $p \mid (x^2 + 1)$ en $\mathbb{Z}$, luego $p \mid (x^2 + 1) = (x + \mathrm{i})(x - \mathrm{i})$ en $\mathbb{Z}[\mathrm{i}]$. Si p fuera primo en $\mathbb{Z}[\mathrm{i}]$, tendríamos que $p \mid (x + \mathrm{i})$ o $p \mid (x - \mathrm{i})$, y del lema 29 deduciríamos que $p \mid 1$, lo cual es imposible. Por tanto, p es compuesto en $\mathbb{Z}[\mathrm{i}]$. ■

Resumimos los resultados previos en los siguientes teoremas.

Teorema 40 (factorización de primos enteros como producto de primos gaussianos). *Sea $p \in \mathbb{N}$ un primo. La descomposición de p en $\mathbb{Z}[\mathrm{i}]$ queda determinada por la congruencia de p módulo* 4 *como sigue:*

1. *Si* $p \equiv 3 \text{ mód } 4$*, entonces p permanece primo en* $\mathbb{Z}[\mathrm{i}]$ *(inerte).*
2. *Si* $p \equiv 1 \text{ mód } 4$*, entonces* $p = \pi\overline{\pi}$*, donde* $\pi, \overline{\pi}$ *son primos conjugados y no asociados (descompone).*
3. $2 = (1 + \mathrm{i})(1 - \mathrm{i}) = -\mathrm{i}\,(1 + \mathrm{i})^2$ *(ramificado).*

Demostración. Basta combinar la observación 37, los teoremas 39 y 34 y el corolario 35. ■

[10] Si $a \not\equiv 0 \text{ mód } p$, entonces $a^{p-1} \equiv 1 \text{ mód } p$. Aunque se pueden dar demostraciones sencillas y directas, también es un corolario inmediato del teorema de Lagrange en teoría de grupos, pues el grupo de unidades $\mathbb{F}_p^*$ tiene orden $p - 1$.

Teorema 41. *Sea $\alpha \in \mathbb{Z}[\mathrm{i}]$ un primo gaussiano. Entonces, salvo asociados, α debe ser de uno de los siguientes tipos:*

1. $p \in \mathbb{N}$ *es un primo entero con* $p \equiv 3 \text{ mód } 4$*,*
2. π *o* $\overline{\pi}$ *con* $N(\pi) = p \in \mathbb{N}$ *primo y* $p \equiv 1 \text{ mód } 4$*, o*
3. $1 + \mathrm{i}$.

Además, todos los tipos de enteros gaussianos arriba descritos, al igual que sus asociados, son primos gaussianos, por lo que hemos descrito todos los primos gaussianos.

Demostración. Por el lema 36, cada primo gaussiano divide a un primo entero. Como $\mathbb{Z}[\mathrm{i}]$ tiene factorización única, los primos de $\mathbb{Z}[\mathrm{i}]$ deben ser los primos gaussianos que aparecen en el teorema 40. Por último, está claro que los enteros gaussianos que aparecen en el enunciado son primos, pues en los dos últimos tipos $N(\alpha)$ es primo y la propiedad 8 de la proposición 30 lo concluye, y en el primer tipo es consecuencia del teorema 40. ■

En particular, del teorema 41 se deduce que la norma de cualquier primo gaussiano es p o p^2, siendo p un primo entero. Por último, del teorema 40 se deduce la factorización de cualquier entero dentro de $\mathbb{Z}[\mathrm{i}]$.

Corolario 42. *Sea $n \geq 2$ un entero, y sean $p_1, \dots, p_r$ los factores primos de n congruentes con $1 \text{ mód } 4$ y $q_1, \dots, q_s$ los factores primos de n congruentes con $3 \text{ mód } 4$, de modo que la descomposición de n como producto de primos en $\mathbb{Z}$ es la siguiente:*

$$n = 2^c p_1^{n_1} \cdots p_r^{n_r} q_1^{m_1} \cdots q_s^{m_s}$$

para ciertos $n_\ell, m_j \geq 1$ y $c \in \mathbb{Z}_{\geq 0}$. Para cada $1 \leq \ell \leq r$, sea $p_\ell = \pi_\ell \overline{\pi}_\ell$ la descomposición de p_ℓ como producto de dos primos conjugados no asociados en $\mathbb{Z}[\mathrm{i}]$. Entonces, n factoriza en $\mathbb{Z}[\mathrm{i}]$ del siguiente modo:

$$n = (1 + \mathrm{i})^{2c} \pi_1^{n_1} \overline{\pi}_1^{n_1} \cdots \pi_r^{n_r} \overline{\pi}_r^{n_r} q_1^{m_1} \cdots q_s^{m_s}.$$

Además, dicha factorización es única salvo el orden de los factores y asociados.

Demostración. Es consecuencia de ser $\mathbb{Z}[\mathrm{i}]$ un DFU (por el teorema 28 y el corolario 21) y del teorema 41. ■

Terminamos esta sección con el resultado principal, la solución completa al problema 1, que viene dada por medio de los factores primos del entero en cuestión. En el teorema está excluido el caso $n = 1$, pero está claro que $1 = 1^2 + 0^2 = (-1)^2 + 0^2 = 0^2 + 1^2 = 0^2 + (-1)^2$ son todas sus representaciones, y, en particular, es representable.

Teorema 43. *Un entero $n > 1$ es suma de dos cuadrados si y solo si cada factor primo p de n verificando $p \equiv 3 \text{ mód } 4$ tiene multiplicidad par.*

Demostración. Ya hemos comentado que ser suma de dos cuadrados es una propiedad multiplicativa tras la proposición 33. El primo $p = 2$ y cualquier primo $p \equiv 1 \text{ mód } 4$ son sumas de dos cuadrados por el teorema 39; por tanto, también lo son sus potencias. Por otro lado, un primo $p \equiv 3 \text{ mód } 4$ no es suma de dos cuadrados, pero p^2 sí lo es y, por tanto, cualquier potencia par de p es también suma de dos cuadrados. En definitiva, cualquier n en el que los primos $p \equiv 3 \text{ mód } 4$ aparecen con multiplicidad par es una suma de dos cuadrados.

Sea ahora $n > 1$ suma de dos cuadrados que tiene un factor primo $p \equiv 3 \text{ mód } 4$ (pues en otro caso no hay nada que demostrar). Si $n = a^2 + b^2$, entonces $p \mid n = (a + b\mathrm{i})(a - b\mathrm{i})$ en $\mathbb{Z}[\mathrm{i}]$, y como p es inerte, o bien $p \mid (a + b\mathrm{i})$ o bien $p \mid (a - b\mathrm{i})$. En cualquier caso, la conclusión es que $p \mid a$ y $p \mid b$ en $\mathbb{Z}$ por el lema 29. De $a = pA$ y $b = pB$ se deduce que $n = a^2 + b^2 = p^2(A^2 + B^2)$, luego $p^2 \mid n$ y $n/p^2 = A^2 + B^2$ es suma de dos cuadrados. Si m es la multiplicidad de p como divisor de n, escribiendo $m = 2k + r$ para $r = 0$ o 1 y aplicando este resultado recursivamente k veces, concluimos que $n' = n/p^{2k}$ es suma de dos cuadrados. El caso $r = 1$ implicaría que $p \mid n'$ pero $p^2 \nmid n'$, pero esto no es posible por el razonamiento previo (sustituyendo n por n'). Por tanto, se tiene que $r = 0$ o, lo que es lo mismo, la multiplicidad de p es par. ■

5. Número de representaciones como suma de dos cuadrados

Para concluir, queremos utilizar los resultados anteriores para resolver el problema 3. La fórmula buscada dependerá, como uno podría imaginar, de los divisores de n que son congruentes con 1 y con 3 mód 4; más concretamente, de su cardinal. La fórmula que queremos demostrar es la siguiente:

$$r_2(n) = 4(d_1(n) - d_3(n)), \tag{2}$$

donde $d_j(n)$ denota el número de divisores de n congruentes con j mód 4 para $j \in \{1, 3\}$. Para la demostración necesitamos realizar varias comprobaciones, por lo que hemos decidido dividir la prueba en una serie de pasos.

- Partiendo de la factorización de n, observar que podemos eliminar las potencias pares de primos $p \equiv 3$ mód 4, al igual que el factor 2 (con su potencia).
- Interpretar el factor 4 de la ecuación (2), así como las funciones $r_2(n)$ y $r_2(n)/4$, en términos de ideales del anillo $\mathbb{Z}[\mathrm{i}]$.
- Demostrar que, tras dividir la ecuación (2) por 4, ambos términos son multiplicativos en n; es decir, que si demostramos la fórmula para n, m coprimos, entonces es válida para nm.
- Demostrar que la fórmula es cierta para las potencias de primos enteros $n = p^e$.

La mayoría de estas ideas se encuentran en el libro de Hardy y Wright [5, sección 20] con un enfoque ligeramente distinto.

Proposición 44. *Dado $n \in \mathbb{N}$, $r_2(n) = r_2(2n)$.*

Demostración. Sea $n = a^2+b^2$ una representación de n. Entonces, se puede comprobar que $(a-b)^2+(a+b)^2$ es una representación de $2n$. Recíprocamente, sea $2n = c^2+d^2$. Entonces, $c \equiv d$ mód 2 y $n = \left(\frac{c+d}{2}\right)^2 + \left(\frac{c-d}{2}\right)^2$. Por tanto, $(a, b) \mapsto (a-b, a+b)$ es una biyección entre $R_2(n)$ y $R_2(2n)$. ∎

Observación 45. El hecho de que si $n = a^2 + b^2$, entonces $2n = (a-b)^2 + (a+b)^2$, es también consecuencia de la igualdad $(a^2+b^2)(c^2+d^2) = (ac-bd)^2+(ad+bc)^2$ para cualesquiera $a, b, c, d \in \mathbb{R}$, lo cual se deduce de que N sea multiplicativa y, por tanto, $N(a+b\mathrm{i})N(c+d\mathrm{i}) = N((a+b\mathrm{i})(c+d\mathrm{i}))$. Dicha fórmula nos indica que el producto de enteros representables es representable, y nos da un modo de obtener una representación de nm conocidas representaciones de n y de m, que es lo que aplicamos para $m = 2 = 1^2 + 1^2$. Solamente existen fórmulas similares que expresan el producto de dos sumas de k cuadrados como una suma de k cuadrados para $k \in \{1, 2, 4, 8\}$, como afirma el problema de Hurwitz; para más información, véase el libro de Jacobson [6, sección 7.6]. ◂

La proposición 44 nos indica que, a la hora de calcular $r_2(n)$, podemos primero simplificar n dividiendo por 2 tantas veces como sea posible. El análogo para los primos $p \equiv 3$ mód 4 que presentamos a continuación nos permite simplificar las potencias de p por pares, lo cual tiene sentido porque recordamos que, si p es divisor de n con multiplicidad impar, entonces n no es representable.

Proposición 46. *Sean $n, p \in \mathbb{N}$ con p primo y $p \equiv 3$ mód 4. Entonces, $r_2(n) = r_2(p^2 n)$.*

Demostración. Si $n = a^2 + b^2$, entonces $p^2n = (pa)^2 + (pb)^2$. Recíprocamente, si $p^2n = c^2 + d^2$, entonces, repitiendo el argumento del teorema 43 (es decir, que como p es inerte, $p \mid c^2 + d^2 = (c+d\mathrm{i})(c-d\mathrm{i})$ en $\mathbb{Z}[\mathrm{i}]$ y, por tanto, $p \mid c, d$ en $\mathbb{Z}$), vemos que $n = (c/p)^2 + (d/p)^2$ es una representación de n. Por tanto, $(a, b) \mapsto (pa, pb)$ es una biyección entre $R_2(n)$ y $R_2(p^2n)$. ∎

Hasta el momento no hemos tenido que lidiar con el tema de las permutaciones y cambios de signo entre representantes. Con el fin de evitar este detalle, consideramos una variación sobre el conjunto de representaciones.

Definición 47. Una representación $n = a^2 + b^2$ se dice **positiva** si $a > 0$ y $b \geq 0$. Se define el **conjunto de representaciones positivas de** n como sigue:

$$R_2^+(n) = \{(x, y) \in \mathbb{Z}^2 \,:\, x^2 + y^2 = n, x > 0, y \geq 0\}.$$

Definimos también la función $r_2^+(n) = \#R_2^+(n)$ que cuenta las representaciones positivas de n. ◂

Vía la biyección $\mathbb{Z}^2 \to \mathbb{Z}[\mathrm{i}] : (a, b) \mapsto \gamma = a + b\mathrm{i}$, podemos pensar que $R_2(n)$ y $R_2^+(n)$ son subconjuntos de $\mathbb{Z}[\mathrm{i}]$, y diremos que $a + b\mathrm{i}$ es un entero gaussiano **positivo** si $(a, b) \in R_2^+(n)$. Bajo esta correspondencia, $r_2(n)$ es claramente el número de enteros gaussianos γ tales que $N(\gamma) = n$. El siguiente resultado nos dice que al considerar solamente generadores positivos estamos eliminando la multiplicación por 4 que aparece al considerar asociados (pues hay precisamente cuatro unidades), de modo que $r_2(n)$ cuenta los cuatro asociados de un $\alpha \neq 0$ como representaciones distintas, mientras que $r_2^+(n)$ cuenta solamente una de ellas.

Lema 48. *Sea $n \in \mathbb{N}$. Si pensamos $(x, y) \in R_2(n)$ como el entero gaussiano $\alpha = x + y\mathrm{i} \neq 0$, entonces α tiene un único asociado β que pertenece a $R_2^+(n)$. Por tanto, $r_2(n) = 4r_2^+(n)$.*

Demostración. Que los cuatro asociados distintos de α,

$$\{\alpha = x + y\mathrm{i}, \quad \mathrm{i}\alpha = -y + x\mathrm{i}, \quad -\alpha = -x - y\mathrm{i}, \quad -\mathrm{i}\alpha = y - x\mathrm{i}\},$$

pertenecen a $R_2(n)$, pero solo uno de ellos pertenece a $R_2^+(n)$, es una mera comprobación. La fórmula es una consecuencia inmediata. ■

Observación 49. Recordemos que $\alpha \sim \beta \iff (\alpha) = (\beta)$. Por tanto, estamos seleccionando el único generador positivo de cada ideal (no nulo). Es más, podemos definir una norma N' sobre el conjunto de ideales del siguiente modo: para $I = (\alpha)$, se define $N'(I) = N(\alpha)$. N' está bien definida porque $(\alpha) = (\beta) \iff \alpha = \epsilon\beta$ para alguna unidad ϵ, y $N(\alpha) = N(\epsilon)N(\beta) = N(\beta)$ por ser N multiplicativa y por la propiedad 2 de la proposición 30. Además, N' es multiplicativa por serlo N. Por abuso de notación, denotaremos a ambas normas como N. Teniendo esto en cuenta, $r_2^+(n)$ contabiliza el número de ideales de $\mathbb{Z}[\mathrm{i}]$ de norma n (vía la correspondencia que asocia a cada $(a, b) \in R_2^+(n)$ el ideal $(a + b\mathrm{i})$), a diferencia de $r_2(n)$, que contabiliza el número de elementos de norma n. ◂

Intentemos ahora entender el factor 4 de la ecuación (2), así como el hecho de que en el ejemplo 4 el número 4 tuviera cuatro representaciones mientras que el número 10 tenía ocho representaciones, cuando en ambos casos hay una sola representación salvo permutaciones y asociados. Para ello consideramos el conjunto de elementos no nulos de $\mathbb{Z}[\mathrm{i}]$, y la acción sobre él (por multiplicación) del grupo de unidades $\mathbb{Z}[\mathrm{i}]^*$, que identificamos con el grupo cíclico C_4 generado por la multiplicación por i. La acción es libre[11] y, por tanto, cada órbita[12] tiene longitud 4; de hecho, la órbita de $\alpha \neq 0$ es $\{\mathrm{i}\alpha, -\alpha, -\mathrm{i}\alpha, \alpha\}$. Podemos considerar también la acción (sobre el conjunto de elementos no nulos de $\mathbb{Z}[\mathrm{i}]$) del grupo diédrico D_4, tomando como generadores la multiplicación por i y la conjugación. En ese caso, la órbita de $\alpha = x + y\mathrm{i} \neq 0$ es

$$\begin{aligned}
\alpha &= x + y\mathrm{i} \leftrightarrow (x, y), & \overline{\alpha} &= x - y\mathrm{i} \leftrightarrow (x, -y), \\
\mathrm{i}\alpha &= -y + x\mathrm{i} \leftrightarrow (-y, x), & \mathrm{i}\overline{\alpha} &= y + x\mathrm{i} \leftrightarrow (y, x), \\
-\alpha &= -x - y\mathrm{i} \leftrightarrow (-x, -y), & -\overline{\alpha} &= -x + y\mathrm{i} \leftrightarrow (-x, y), \\
-\mathrm{i}\alpha &= y - x\mathrm{i} \leftrightarrow (y, -x), & -\mathrm{i}\overline{\alpha} &= -y - x\mathrm{i} \leftrightarrow (-y, -x).
\end{aligned}$$

En términos de ideales, C_4 permuta los generadores de (α), mientras que D_4 permuta los generadores de (α) y los de su ideal conjugado $(\overline{\alpha})$. Como siempre hay cuatro asociados distintos, la órbita puede tener longitud 4 u 8, y tiene longitud 4 si y solo si $(\alpha) = (\overline{\alpha})$. Esto último ocurre si y solo si

- o bien $x = 0$ o $y = 0$, caso en que (α) está generado por un número natural,
- o bien $x = y$, caso en que (α) está generado por $m(1 + \mathrm{i})$ para algún $m \in \mathbb{N}$.

Por último, si incluimos al elemento $0 = 0 + 0\mathrm{i} \leftrightarrow (0, 0)$, se trataría del único punto fijo de las acciones antes mencionadas.

Continuamos simplificando el problema 3: veremos que $r_2^+(n)$ es una función multiplicativa en n; es decir, si $n, m \in \mathbb{N}$ y $\operatorname{mcd}(m, n) = 1$, entonces $r_2^+(mn) = r_2^+(m)r_2^+(n)$. La multiplicatividad es una propiedad importante de funciones sobre $\mathbb{N}$ que reduce considerablemente la dificultad de determinar su valor, y aparece frecuentemente al tomar enfoques aritméticos, como es nuestro caso. Por este motivo es conveniente que tratemos algunos resultados básicos sobre estas funciones.

[11] Recordamos que una acción de un grupo G en un conjunto X se dice **libre** si de la igualdad $g \cdot x = h \cdot x$, siendo $g, h \in G, x \in X$, se deduce que $g = h$. En nuestro caso esto está claro pues $\epsilon\alpha = \epsilon'\alpha \implies (\epsilon - \epsilon')\alpha = 0$ y, por integridad, se tiene que $\epsilon = \epsilon'$.

[12] La **órbita** de un elemento x (siendo G una acción en X) es el conjunto $\{g \cdot x\}_{g \in G}$ de los trasladados de x por G.

5.1. Apunte sobre funciones multiplicativas

Definición 50. Sea M un monoide[13] con elemento unidad 1. Una **función aritmética M-valorada** (o simplemente una **función aritmética** cuando no haya ambigüedad sobre M) es una función $f : \mathbb{N} \to M$. ◂

Definición 51. Una función aritmética f se dice que es **multiplicativa** si $f(1) = 1$ y si para cualesquiera $n, m \in \mathbb{N}$ tales que $\mathrm{mcd}(n, m) = 1$ se verifica que $f(nm) = f(n)f(m)$. Si, además, f verifica que $f(nm) = f(n)(m)$ para cualesquiera $n, m \in \mathbb{N}$ (independientemente de su mcd), entonces se dice que f es **completamente multiplicativa**. ◂

Un ejemplo trivial de función completamente multiplicativa es la función constante 1. Para $M = \mathbb{C}$ también lo es la exponenciación compleja $f(n) = n^s$. Un ejemplo menos obvio y que está relacionado con las representaciones por dos cuadrados es el siguiente.

Ejemplo 52. Es fácil comprobar que la función[14]

$$\chi(n) = \left.\begin{cases} (-1)^{\frac{n-1}{2}} & \text{si } n \equiv 1 \text{ mód } 2, \\ 0 & \text{si } n \equiv 0 \text{ mód } 2, \end{cases}\right\} = \begin{cases} 1 & \text{si } n \equiv 1 \text{ mód } 4, \\ -1 & \text{si } n \equiv 3 \text{ mód } 4, \\ 0 & \text{si } n \equiv 0 \text{ mód } 2 \end{cases} \tag{3}$$

es una función completamente multiplicativa con valores en el submonoide $\{-1, 0, 1\}$ de $(\mathbb{N}, \cdot, 1)$. ◂

También es fácil comprobar la siguiente proposición, por lo que dejamos la demostración para el lector.

Proposición 53.

- *Una función $f : \mathbb{N} \to M$ es multiplicativa si y solo si $f(1) = 1$ y para cualesquiera $p, n \in \mathbb{N}$ y $e \in \mathbb{N}$, siendo p primo y $\mathrm{mcd}(p, n) = 1$, se verifica que $f(p^e n) = f(p^e)f(n)$.*
- *Dos funciones multiplicativas $f, g : \mathbb{N} \to M$ son iguales si y solo si $f(p^e) = g(p^e)$ para cada primo $p \in \mathbb{N}$ y cada $e \in \mathbb{N}$.*

Pasamos ahora a introducir la convolución de Dirichlet, que es una operación de funciones aritméticas y que preserva la multiplicatividad. Este hecho nos será muy útil para determinar la expresión de $r_2^+(n)$. De ahora en adelante consideramos solamente funciones aritméticas $\mathbb{C}$-valoradas.

Definición 54. Sean f, g funciones aritméticas. Definimos la **convolución de Dirichlet** de f y g como

$$(f * g)(n) = \sum_{d|n} f(d)g\left(\frac{n}{d}\right) = \sum_{ab=n} f(a)g(b),$$

donde la suma se toma sobre todos los divisores positivos de n. ◂

Proposición 55. *Si f, g son funciones multiplicativas, entonces $f * g$ también es multiplicativa.*

Demostración. Está claro que $(f * g)(1) = 1$. Basta demostrar que $(f * g)(p^e n) = (f * g)(p^e)(f * g)(n)$ si p es primo, $e \in \mathbb{N}$ y $\mathrm{mcd}(p, n) = 1$. Observamos que todo divisor d de $p^e n$ es de la forma $d = p^c d'$ para $0 \le c \le e$ y $d' \mid n$ únicos. Por tanto, se tiene que

$$\begin{aligned}(f * g)(p^e n) &= \sum_{d|p^e n} f(d)g\left(\frac{p^e n}{d}\right) = \sum_{\substack{0\le c\le e \\ d'|n}} f(p^c d')g\left(\frac{p^e n}{p^c d'}\right) = \sum_{\substack{0\le c\le e \\ d'|n}} f(p^c)f(d')g(p^{e-c})g\left(\frac{n}{d'}\right) \\ &= \left(\sum_{0\le c\le e} f(p^c)g\left(\frac{p^e}{p^c}\right)\right)\left(\sum_{d'|n} f(d')g\left(\frac{n}{d'}\right)\right) = (f * g)(p^e)\,(f * g)(n).\end{aligned}$$ ∎

[13] Un **monoide** es un conjunto M con una operación interna asociativa y con elemento neutro. El lector puede pensar que se trata de un grupo pues tomaremos $M = \mathbb{C}$, pero no es necesaria la existencia de elemento opuesto para la definición.

[14] En teoría de números, χ se conoce como **símbolo de reciprocidad cuadrática de** -1 o como el **carácter de Dirichlet no trivial módulo** 4.

5.2. Expresión de la función de representación

Veamos ahora la principal razón por la que hemos introducido las funciones multiplicativas.

Teorema 56. *La función $r_2^+(n)\colon \mathbb{N} \to \mathbb{Z}_{\geq 0}$ es multiplicativa.*

Demostración. Está claro que $r_2^+(1) = 1$. Queremos demostrar que $r_2^+(p^e n) = r_2^+(p^e) r_2^+(n)$ para cualquier primo p y cualesquiera $n, e \in \mathbb{N}$ con $\operatorname{mcd}(p, n) = 1$. Pensando los elementos de $R_2^+(m)$ como ideales de norma m, la multiplicatividad de la norma demuestra que la aplicación de multiplicación de ideales $R_2^+(p^e) \times R_2^+(n) \to R_2^+(p^e n)\colon ((\beta), (\gamma)) \mapsto (\beta\gamma)$ está bien definida. Basta demostrar que se trata de una biyección.

La inyectividad se sigue de que β, γ deben ser coprimos en $\mathbb{Z}[\mathrm{i}]$ porque sus normas son coprimas en $\mathbb{Z}$ (propiedad 7 de la proposición 30) y del corolario 24. En efecto, si $\beta\gamma = \beta'\gamma'$ con $N(\beta) = N(\beta') = p^e$ y $N(\gamma) = N(\gamma') = n$, entonces $\operatorname{mcd}(N(\beta), N(\gamma')) = 1$ implica que $\operatorname{mcd}(\beta, \gamma') = 1$, luego $\beta \mid \beta'$. Del mismo modo se deduce que $\beta' \mid \beta$, luego β, β' son asociados, y lo análogo es cierto para γ, γ'. En términos de ideales, esto significa que $(\beta) = (\beta')$ y $(\gamma) = (\gamma')$.

La epiyectividad equivale a demostrar que si $\alpha \in \mathbb{Z}[\mathrm{i}]$ tiene norma $p^e n$ entonces se puede factorizar $\alpha = \beta\gamma$ con $N(\beta) = p^e$ y $N(\gamma) = n$. Consideramos la factorización de α en $\mathbb{Z}[\mathrm{i}]$, $\alpha = \pi_1 \cdots \pi_r$, siendo π_i primos gaussianos contados con multiplicidad y únicos salvo asociados. Entonces, tenemos que $p^e n = N(\alpha) = N(\pi_1) \cdots N(\pi_r)$. Algunas de las normas $N(\pi_j)$ deben ser divisibles por p. Reordenando, podemos suponer que son $\pi_1, \dots, \pi_s$, con $1 \leq s \leq r$, y entonces $p \nmid N(\pi_{s+1}), \dots, N(\pi_r)$. Sabemos que $N(\pi_j)$ debe ser de la forma q o q^2 para algún primo entero q, luego $N(\pi_j) = p$ o p^2 para $1 \leq j \leq s$, mientras que $\operatorname{mcd}(p, N(\pi_{s+1} \cdots \pi_r)) = 1$. Necesariamente $N(\pi_1 \cdots \pi_s) = p^e$ y $N(\pi_{s+1} \cdots \pi_r) = n$. Por tanto, debe darse que $\beta = \pi_1 \cdots \pi_s$ y $\gamma = \pi_{s+1} \cdots \pi_r$ salvo asociados. ■

Finalmente, podemos resolver el problema 3 utilizando los resultados previos.

Teorema 57. *Para cada $n \in \mathbb{N}$, $r_2^+(n) = d_1(n) - d_3(n)$ y, por tanto, $r_2(n) = 4(d_1(n) - d_3(n))$.*

Demostración. Basta probar la fórmula para $r_2^+(n)$ y aplicar el lema 48. Para ello, consideramos de nuevo la función χ del ejemplo 52. Por definición se tiene que

$$(\chi * 1)(n) = \sum_{d \mid n} \chi(d) 1(n/d) = \sum_{d \mid n} \chi(d) = \sum_{\substack{d \mid n \\ d \equiv 1 \text{ mód } 4}} 1 - \sum_{\substack{d \mid n \\ d \equiv 3 \text{ mód } 4}} 1 = d_1(n) - d_3(n).$$

Por tanto, tenemos que demostrar que $r_2^+(n) = (\chi * 1)(n)$. Ambas funciones son multiplicativas: para $r_2^+(n)$ es consecuencia del teorema 56, mientras que $\chi * 1$ es multiplicativa por serlo las funciones χ y 1, como consecuencia de la proposición 55. De la proposición 53 se deduce que basta con demostrar que ambas funciones coinciden sobre las potencias de primos, para lo cual utilizaremos la siguiente propiedad: si p es primo y $e \in \mathbb{N}$, entonces los divisores de p^e son precisamente p^c variando c en el conjunto $\{0, 1, \dots, e\}$. Separamos a continuación la demostración según las congruencias módulo 4.

- Para $p = 2$, tenemos que $r_2^+(2^e) = r_2^+(1) = 1$ por la proposición 44. Por otro lado, todos los divisores de 2^e distintos de $d = 1$ son pares, luego $d_1(2^e) - d_3(2^e) = 1$ y se da la igualdad.
- Para un primo $p \equiv 3$ mód 4, aplicando la proposición 46 y el teorema 43 tenemos que

  $$r_2^+(p^e) = \begin{cases} r_2^+(1) = 1 & \text{si } e \equiv 0 \text{ mód } 2, \\ 0 & \text{si } e \equiv 1 \text{ mód } 2. \end{cases}$$

  Observamos que $p^c \equiv 1$ mód 4 si y solo si c es par y que $p^c \equiv 3$ mód 4 si y solo si c es impar, por lo que $d_1(p^e) - d_3(p^e)$ es la diferencia entre la cantidad de números pares e impares en el conjunto $\{0, 1, \dots, e\}$, que claramente coincide con el valor arriba indicado en la expresión de $r_2^+(p^e)$.
- Si $p \equiv 1$ mód 4, entonces todos los divisores de p^e son congruentes con 1 mód 4. Así, $d_1(p^e) - d_3(p^e)$ es precisamente el número de divisores de p^e, que es $e+1$. Por otro lado, $r_2^+(p^e)$ es el número de ideales (α) de norma p^e. Como p es un primo que descompone, esto es, $p = \pi\overline{\pi}$ con $N(\pi) = N(\overline{\pi}) = p$, tenemos que $\alpha\overline{\alpha} = \pi^e\overline{\pi}^e$ y, por la factorización única, los ideales mencionados son de la forma $(\alpha) = (\pi)^r(\overline{\pi})^s$ con $r + s = e$ y $0 \leq r, s$, luego hay precisamente $e + 1$ de ellos. ■

6. Conclusiones

En este artículo hemos visto cómo la teoría de números puede nutrirse del álgebra y hemos dado unas pinceladas de algunos conceptos de la teoría algebraica de números, como son los primos que descomponen, que ramifican o que son inertes, y también de algunos conceptos de teoría analítica de números, como las funciones multiplicativas y la convolución de Dirichlet. Todos estos detalles nos han servido para resolver un problema clásico que muestra muy bien cómo es la teoría de números, en la que para la demostración de un enunciado aparentemente inocente uno debe valerse de muy diversas herramientas.

No queremos terminar sin comentar algunos problemas relacionados con la suma de dos cuadrados que no hemos podido tratar. Por ejemplo, hemos estudiado el número representaciones que existen de un entero n, pero no hemos hablado de cómo obtenerlas. Puede encontrarse un ejemplo de este cálculo en el libro de Niven, Zuckerman y Montgomery [7, capítulo 3, ejemplo 3]. En general, el problema se reduce a calcular las representaciones para los primos que dividan a n y utilizar la observación 45. Claramente es necesario el uso de ordenadores para valores grandes de n, siendo interesante estudiar los algoritmos de resolución de este problema.

Similar al estudio que hemos hecho para dos cuadrados, uno puede preguntarse por el valor de la función $r_k(n)$ para distintos valores de k. Está claro que en nuestro estudio ha resultado fundamental utilizar los enteros gaussianos $\mathbb{Z}[\mathrm{i}]$. El caso $k = 4$ se puede estudiar de manera similar, sustituyendo $\mathbb{C}$ por los cuaterniones[15] y $\mathbb{Z}[\mathrm{i}]$ por los cuaterniones con coeficientes enteros[16]. Uno puede de este modo demostrar el siguiente teorema.

Teorema 58 (teorema de los cuatro cuadrados de Lagrange). *Todo entero no negativo es suma de cuatro cuadrados de enteros.*

El teorema es equivalente a afirmar que $R_4(n) \neq \emptyset$ para todo $n \in \mathbb{N}$, y también a que $r_4(n) \geq 1$ para todo $n \in \mathbb{N}$. Se conoce, además, la fórmula explícita de $r_4(n)$, dada por Jacobi.

Teorema 59. *Para* $n \in \mathbb{N}$, $r_4(n) = 8 \sum_{4 \nmid d \mid n} d$.

La suma se toma sobre los divisores de n que no son divisibles por 4. Existen varias maneras de demostrar esta fórmula: una es observando que una suma de cuatro cuadrados no es más que dos sumas de dos cuadrados, luego la fórmula de $r_4(n)$ puede deducirse si conocemos la expresión de $r_2(n)$ sin excesivo trabajo (véase el libro de Davidoff, Sarnak y Valette [4, sección 2.4], donde se trata el caso n impar, pero el caso n par es consecuencia de los resultados que allí aparecen); otra manera consiste en estudiar funciones modulares, que son interesantes por sí mismas y están muy presentes en la actualidad en la teoría de números; también se puede intentar seguir los pasos que hemos dado en este artículo y utilizar la factorización sobre los cuaterniones con coeficientes enteros, aunque esta opción es bastante complicada porque se trata de un anillo no conmutativo (para estudiar dicha factorización puede consultarse el libro de Conway y Smith [3, capítulo 5]), y también existen demostraciones elementales, como, por ejemplo, el artículo de Spearman y Williams [8] o el libro de Williams [10, capítulo 9], donde se hace una demostración basada en unas identidades de Liouville nada evidentes, aunque elementales.

Referencias

[1] Cobos Rábano, Alberto. *On certain algebraic, arithmetic and topological properties of quaternions.* Trabajo de Fin de Grado. Universidad de Salamanca, 2018.

[2] Conrad, Keith. *The gaussian integers.* Expository papers. 2016. URL: `http://www.math.uconn.edu/~kconrad/blurbs/`.

[15] Por cuaterniones nos referimos a los **cuaterniones de Hamilton**, que son, salvo isomorfismo algebraico, el único $\mathbb{R}$-álgebra de dimensión finita no conmutativa. Más explícitamente, se trata de elementos de la forma $a + b\mathrm{i} + c\mathrm{j} + d\mathrm{k}$ tales que $a, b, c, d \in \mathbb{R}$ y donde $\mathrm{i}^2 = \mathrm{j}^2 = \mathrm{k}^2 = \mathrm{ijk} = -1$.

[16] Para ser más precisos, es conveniente considerar un subanillo de los cuaterniones con coeficientes enteros, denominado **anillo de cuaterniones de Hurwitz**, por ser este un dominio euclídeo. Se muestra así que la teoría de divisibilidad es relevante en el problema de determinar $r_k(n)$ en general, y no solo en el caso $k = 2$.

[3] Conway, John H. y Smith, Derek A. *On quaternions and octonions: their geometry, arithmetic, and symmetry.* A K Peters, Ltd., 2003, págs. xii+159. ISBN: 978-1-56881-134-5.

[4] Davidoff, Giuliana; Sarnak, Peter, y Valette, Alain. *Elementary number theory, group theory, and Ramanujan graphs.* Vol. 55. London Mathematical Society Student Texts. Cambridge University Press, Cambridge, 2003, págs. x+144. https://doi.org/10.1017/CBO9780511615825.

[5] Hardy, Godfrey H. y Wright, Edward M. *An introduction to the theory of numbers.* 3rd ed. Oxford, at the Clarendon Press, 1954, págs. xvi+419.

[6] Jacobson, Nathan. *Basic algebra. I.* Second edition. W. H. Freeman y Company, New York, 1986, págs. xviii+499. ISBN: 978-0-7167-1480-4.

[7] Niven, Ivan; Zuckerman, Herbert S., y Montgomery, Hugh L. *An introduction to the theory of numbers.* Fifth. John Wiley & Sons, Inc., New York, 1991, págs. xiv+529. ISBN: 978-0-471-62546-9.

[8] Spearman, Blair K. y Williams, Kenneth S. «The simplest arithmetic proof of Jacobi's four squares theorem». En: *Far East Journal of Mathematical Sciences (FJMS)* 2.3 (2000), págs. 433-439. ISSN: 0972-0871.

[9] Weil, André. *Number theory. An approach through history from Hammurapi to Legendre.* Reprint of the 1984 edition. Modern Birkhäuser Classics. Birkhäuser Boston, Inc., Boston, MA, 2007, págs. xxii+377. ISBN: 978-0-8176-4565-6.

[10] Williams, Kenneth S. *Number theory in the spirit of Liouville.* Vol. 76. London Mathematical Society Student Texts. Cambridge University Press, Cambridge, 2011, págs. xviii+287. ISBN: 978-0-521-17562-3.

TEMat

Geometría diferencial en el estudio de imágenes médicas

Clara Rodríguez Pérez
Universidad de La Laguna
clara8794@hotmail.com

Resumen: La teoría de variedades de Fréchet riemannianas nos proporciona un marco teórico para abordar el estudio del espacio $\mathbb{M}$ de curvas planas cerradas regulares, de modo que se pueda introducir la distancia entre dos de estas curvas. En este artículo presentamos una noción de curva media extrínseca de una muestra finita de curvas planas cerradas regulares. El término «extrínseco» se refiere al hecho de que embebemos $\mathbb{M}$ en un espacio euclídeo para poder calcular la media de forma natural y, desde esta, dar la noción de curva media extrínseca en $\mathbb{M}$. Además, presentamos un algoritmo para el cálculo de la curva media extrínseca y lo aplicamos al estudio de imágenes médicas. Para ello, previamente lo discretizamos y lo implementamos en MATLAB. Este trabajo está basado en resultados de Gual-Arnau, Ibáñez Gual y Simó Vidal [9].

Abstract: The Riemmanian Geometry on Fréchet manifolds gives us a theorical framework in order to study the space $\mathbb{M}$ of regular closed plane curves, so that we can introduce the distance between two such curves. In this article, we present a notion of the extrinsic mean curve of a finite sample of regular closed plane curves. The word «extrinsic» refers to the fact that we have embedded $\mathbb{M}$ into an Euclidean space in order to compute the mean in a natural way and, from it, give the notion of an extrinsic mean curve. Furthemore, we develop an algorithm to calculate the extrinsic mean curve and we use it to study medical images. To do that, we previously discretize and implement the algorithm with MATLAB. This work is based on results of Gual-Arnau, Ibáñez Gual, and Simó Vidal [9].

Palabras clave: variedades de Fréchet, curvas planas cerradas regulares, métricas riemannianas, distancia entre curvas, curva media extrínseca, aplicaciones a imágenes médicas.

MSC2010: 58B10, 58B20, 58D15.

Recibido: 24 de enero de 2018.
Aceptado: 30 de enero de 2019.

Agradecimientos: En primer lugar, me gustaría dar las gracias a Ximo Gual-Arnau por plantear la temática de este artículo y a mis profesores Juan Carlos Marrero González y Edith Padrón Fernández por su inestimable ayuda para publicarlo. En segundo lugar, quiero expresar un profundo agradecimiento a mi tía Mariluz Rodríguez Palmero por su optimismo e implicación y a mi familia por toda la confianza, cariño y apoyo que me ha dado desde que inicié el viaje por las matemáticas. Dedicado a mi padre Miguel José Rodríguez Palmero.

Referencia: Rodríguez Pérez, Clara. «Geometría diferencial en el estudio de imágenes médicas». En: *TEMat*, 3 (2019), págs. 17-31. ISSN: 2530-9633. URL: https://temat.es/articulo/2019-p17.

1. Introducción

Cuando un radiólogo se enfrenta con la decisión de radiar un tumor, su experiencia e intuición son dos herramientas fundamentales. Sobre la imagen computarizada en 2D, el médico señala una curva cerrada cuyo interior debe ser radiado. Sin embargo, puede ocurrir que los contornos señalados para un mismo tumor por profesionales diferentes de la medicina no sean iguales. Delimitar con la mayor exactitud posible estos contornos es significativamente importante. Es necesario radiar la mayor masa tumoral posible, mientras se intenta dañar la menor cantidad de tejido no afectado por el tumor. Durante los últimos años, distintos grupos de investigación han intentado crear herramientas más eficaces para enfrentarse a este problema (véanse, por ejemplo, el libro de Bookstein [2] y los artículos de Goodall [8] y Kendall [11]).

En este trabajo nos planteamos la obtención de la curva media extrínseca de una muestra finita de curvas delimitadas sobre una imagen plana por varios profesionales de la medicina.

Así, las matemáticas, de una manera transversal, juegan un papel fundamental en el estudio de esta problemática. El término transversal se justifica porque en la resolución del problema intervienen diferentes áreas:

- La *geometría diferencial* de variedades riemannianas de Fréchet infinito dimensionales, ya que el espacio de estas curvas determina una variedad de este tipo.
- El *análisis funcional* de los espacios de Fréchet sobre los que se modelan estas variedades.
- La *estadística* para hallar una media de las curvas cerradas consideradas.
- El *análisis numérico* para implementar computacionalmente el proceso.

Para estudiar este problema, el primer elemento que debemos considerar es el espacio de las *formas planas*. Pero, ¿qué es matemáticamente una forma plana? En algunos trabajos (véanse, por ejemplo, los textos de Dryden y Mardia [4] y de Kendall *et al.* [12]) se considera una forma plana como un número finito de puntos del plano. La teoría matemática desarrollada en torno a esta descripción depende de la elección de puntos que se haya hecho, por lo que en muchas ocasiones es poco eficiente.

La otra opción (propuesta por Azencott, Coldefy y Younes [1] y detallada por Younes [16]) es considerar una forma como una curva diferenciable cerrada, de modo que podamos referirnos al objeto a través de su contorno. Para desarrollar esta perspectiva debemos trabajar con un concepto matemático complejo: las variedades de Fréchet riemannianas infinito dimensionales (véanse, por ejemplo, el artículo de Hamilton [10] y las referencias contenidas en él).

Una de las desventajas de considerar variedades infinito dimensionales es la cantidad de problemas que plantea relacionados con el cálculo de herramientas estadísticas como la media. Si tomamos una muestra de puntos de un espacio euclídeo, es sencillo calcular la media. En cambio, en el caso del espacio de curvas planas cerradas regulares, al que denotamos por $\mathbb{M}$, la tarea no es tan simple.

El primer problema en estos espacios es cómo definir la distancia entre curvas. Un primer análisis muestra que $\mathbb{M}$ es una variedad de Fréchet sobre la que se puede definir una métrica riemanniana (véase, por ejemplo, el artículo de Sundaramoorthi *et al.* [15]). Por lo tanto, una manera de establecer la distancia entre curvas es considerar la correspondiente distancia riemanniana sobre $\mathbb{M}$. En este caso, obtenemos una media intrínseca, como se muestra en el trabajo de fin de máster de Flores Compañ [5] y el artículo de Flores Compañ *et al.* [6].

Otra opción es embeber la variedad riemanniana $\mathbb{M}$ en un espacio euclídeo y, desde allí, calcular la media de la muestra. En general, esta media no está en la variedad, así que se debe establecer un procedimiento para encontrar la curva de $\mathbb{M}$ que minimice la distancia con la curva media. La nueva curva se denomina curva media extrínseca. Este tipo de medida es interesante cuando las curvas representan fronteras de superficies anatómicas donde las rotaciones, traslaciones o dilataciones son elementos que se deben tener en cuenta en las conclusiones médicas.

El objetivo de este trabajo es indagar acerca de la segunda posibilidad. Así, en este artículo presentamos la construcción propuesta en el artículo de Gual-Arnau, Ibáñez Gual y Simó Vidal [9] de una media extrínseca, clarificando algunos conceptos y explicando de forma genérica el proceso. Además, mostramos un algoritmo que se deriva de la fundamentación teórica y se aplica a un caso médico, con el fin de probar su posible aplicabilidad al mundo real. El trabajo de fin de grado en el que se basa este artículo es el de Rodríguez Pérez [14].

2. El espacio de las curvas planas diferenciables como un espacio de Fréchet

Antes de enfrentarnos al problema de hallar la curva media extrínseca debemos estudiar el espacio donde vamos a trabajar. Este es el espacio de las curvas planas cerradas diferenciables, que denotamos por $C^\infty(\mathbb{S}^1, \mathbb{R}^2)$, compuesto por las funciones C^∞-diferenciables de $\mathbb{S}^1$ en $\mathbb{R}^2$, es decir,

$$C^\infty(\mathbb{S}^1, \mathbb{R}^2) = \{f : \mathbb{S}^1 \to \mathbb{R}^2 \mid f \text{ es } C^\infty\text{-diferenciable}\}.$$

Nótese que se trata de un espacio infinito dimensional, por lo que hay ciertas propiedades y definiciones geométricas y topológicas de los espacios finito dimensionales que no se cumplen en general. Por ejemplo, el teorema de la función inversa.

A continuación, recordaremos brevemente algunas nociones matemáticas que serán útiles para asignar una estructura geométrico-algebraica a $C^\infty(\mathbb{S}^1, \mathbb{R}^2)$.

Definición 1. Una **seminorma** es una función $\mathfrak{p} : V \to \mathbb{R}$ sobre un espacio vectorial V que satisface las siguientes propiedades:

I) para cualesquiera $u, v \in V$, $\mathfrak{p}(u + v) \leq \mathfrak{p}(u) + \mathfrak{p}(v)$,

II) para todo $u \in V$ y para todo $\lambda \in \mathbb{R}$, $\mathfrak{p}(\lambda u) = |\lambda|\mathfrak{p}(u)$. ◂

Usando las propiedades anteriores, obtenemos de II) que $\mathfrak{p}(0) = 0$ y de I) y II), que $\mathfrak{p}(u) \geq 0$ para todo $u \in V$; sin embargo, $\mathfrak{p}(u) = 0$ no implica que $u = 0$. Por tanto, toda norma es una seminorma, pero el recíproco no es cierto en general. Por ejemplo, en $\mathbb{R}^2$ podemos definir la seminorma $\mathfrak{p}((x, y)) = |y|$ para $(x, y) \in \mathbb{R}^2$. Claramente, para todo $(x, 0)$ con $x \in \mathbb{R},\ x \neq 0$, se tiene que $\mathfrak{p}((x, 0)) = 0$, por lo que $\mathfrak{p}$ no puede ser una norma.

Así, dado un conjunto de seminormas $\{\mathfrak{p}_i\}_{i\in\mathbb{I}}$ sobre un espacio vectorial V, donde $\mathbb{I}$ es un conjunto cualquiera de índices, y un punto u_0 de V, podemos generar una base de entornos de u_0 del siguiente modo:

$$U_{\mathbb{B},\varepsilon}(u_0) = \{u \in V \mid \mathfrak{p}_i(u - u_0) < \varepsilon,\ \forall\, i \in \mathbb{B}\},$$

donde $\mathbb{B}$ es un subconjunto finito de $\mathbb{I}$ no vacío y $\varepsilon > 0$.

De esta manera, sobre V podemos inducir una topología asociada a esta base de entornos, respecto de la cual las aplicaciones $\mathfrak{p}_i$ son aplicaciones continuas. En este caso se puede probar que, respecto de esta topología, las operaciones asociadas a la estructura de espacio vectorial son aplicaciones continuas. Por consiguiente, todo espacio vectorial con una familia de seminormas define un espacio vectorial topológico.

Definición 2. Se dice que un espacio vectorial topológico V es un **espacio de Fréchet** si y solo si satisface las siguientes propiedades:

1. V es un espacio Hausdorff, es decir, para cualesquiera dos puntos de V existen dos entornos cuya intersección es vacía.
2. La topología sobre V está inducida por una familia numerable de seminormas $\{\mathfrak{p}_n\}_{n\in\mathbb{N}}$. Esto significa que un subconjunto U de V es abierto si y solo si para todo $u \in U$ existen $k \geq 0$ y $\varepsilon > 0$ tales que el conjunto
$$\{v \in V \mid \mathfrak{p}_n(v - u) < \varepsilon, \forall\, n \leq k\}$$
está contenido en U.
3. V es un espacio vectorial completo respecto de la familia de seminormas, esto es, toda suceción de Cauchy es convergente respecto de $\mathfrak{p}_n$ para todo $n \in \mathbb{N}$. ◂

De hecho, las condiciones segunda y la tercera de la definición 2 son equivalentes a que se pueda definir una distancia d invariante por traslación sobre el espacio vectorial V y que, respecto de esa métrica, el espacio sea completo. La relación entre la métrica d y las seminormas es

$$d(u, v) = \sum_{k=0}^{\infty} \frac{1}{2^k} \frac{\mathfrak{p}_k(u - v)}{1 + \mathfrak{p}_k(u - v)}.$$

A continuación, veremos una manera de construir ejemplos de espacios de Fréchet.

Sea V un espacio vectorial con un conjunto contable de seminormas $\{\mathfrak{p}_n\}_{n\in\mathbb{N}}$. Si estas cumplen las siguientes propiedades:

F_1 Si $v \in V$ y $\mathfrak{p}_n(v) = 0$ para todo $n \in \mathbb{N}$, entonces $v = 0$,

F_2 Si $(v_i)_{i\in\mathbb{N}}$ es una sucesión de Cauchy en V respecto de cada seminorma $\mathfrak{p}_n$, entonces existe $v \in V$ tal que la sucesión v_i converge a v respecto de cada seminorma $\mathfrak{p}_n$,

entonces V es un espacio de Fréchet. Esto es así ya que la condición F_2 garantiza obviamente que V es completo respecto de cada seminorma y F_1 garantiza que V es Hausdorff.

Usando este método y el conjunto de seminormas $\{\mathfrak{p}_n\}_{n\in\mathbb{N}}$ definido por

$$\mathfrak{p}_n(f) = \max_{\theta\in\mathbb{S}^1}\{\|f^{(k)}(\theta)\| \mid k \leq n\},$$

se puede probar que el espacio de las curvas cerradas en el plano $C^\infty(\mathbb{S}^1, \mathbb{R}^2)$ es un espacio de Fréchet infinito dimensional (para más detalles, se puede consultar el trabajo de Rodríguez Pérez [14]).

3. El espacio de curvas planas cerradas regulares como variedad de Fréchet

En la sección anterior conocimos el espacio donde debemos trabajar, $C^\infty(\mathbb{S}^1, \mathbb{R}^2)$. Sin embargo, en nuestro estudio nos interesa considerar el espacio de curvas cerradas en el plano que cumplan una cierta propiedad de regularidad: el vector tangente en cada punto de la curva debe ser distinto de cero. A este conjunto lo denotamos por

$$\mathbb{M} = \{f \in C^\infty(\mathbb{S}^1, \mathbb{R}^2) \mid \|f'(\theta)\| \neq 0,\ \forall\, \theta \in \mathbb{S}^1\},$$

siendo $\|\cdot\|$ la norma en $\mathbb{R}^2$, y se puede demostrar que es un abierto de $C^\infty(\mathbb{S}^1, \mathbb{R}^2)$ [14].

Además, no solo queremos que las curvas cerradas en el plano sean regulares, sino también que estén normalizadas, es decir, que tengan longitud 1 y centroide el origen. El centroide de un objeto está relacionado con la noción física del centro de masas. En el caso de una curva cerrada $f\colon \mathbb{S}^1 \to \mathbb{R}^2$ se define como

$$\overline{c}(f) = \frac{1}{L(f)} \int_0^{2\pi} \|f'(\theta)\| f(\theta)\, \mathrm{d}\theta \in \mathbb{R}^2,$$

donde

$$L(f) = \int_0^{2\pi} \|f'(\theta)\|\, \mathrm{d}\theta. \tag{1}$$

Nota 3. Se debe entender la función $\overline{c}(f)$ como la integración en cada una de las componentes de la función $\|f'(\theta)\| f(\theta)$. ◂

A este nuevo conjunto lo denotamos por $\mathbb{M}_d = \{f \in \mathbb{M} \mid L(f) = 1, \overline{c}(f) = (0,0)\}$.

Sin embargo, este subconjunto de $\mathbb{M}$ no es un abierto de $C^\infty(\mathbb{S}^1, \mathbb{R}^2)$. Este hecho nos obliga a trabajar con una nueva categoría de espacio: las variedades de Fréchet.

Definición 4. Sea M un espacio topológico Hausdorff. Una **carta** sobre M modelada sobre un espacio de Fréchet E es una aplicación $\phi\colon U \subset M \to E$ continua tal que U es un abierto de M, $\phi(U)$ es un abierto de E y $\phi\colon U \subset M \to \phi(U)$ es un homeomorfismo. Un **atlas** sobre M modelado sobre E es un conjunto de cartas $\{(U_\alpha, \phi_\alpha)\}_{\alpha\in\mathbb{A}}$, siendo $\mathbb{A}$ un conjunto de índices, modeladas sobre E, tales que $M = \bigcup_{\alpha\in\mathbb{A}} U_\alpha$ y, para cualesquiera α y β que cumplen que $U_\alpha \cap U_\beta \neq \emptyset$, se tiene que

$$\phi_\beta \circ \phi_\alpha^{-1}\colon \phi_\alpha(U_\alpha \cap U_\beta) \subseteq E \to \phi_\beta(U_\alpha \cap U_\beta) \subseteq E$$

es diferenciable. Una **estructura diferenciable** sobre M es un atlas maximal con respecto a esta última condición. En tal caso, se dice que M es una **variedad de Fréchet**.

Una **subvariedad** de una variedad de Fréchet M modelada sobre E es un subconjunto N de M tal que para todo $x \in N$ existe una carta (U, ϕ) de M tal que $\phi(U \cap N) = \phi(U) \cap E_1$, siendo E_1 un subespacio lineal cerrado de E. ◂

Téngase en cuenta que la noción de variedad de Fréchet es una generalización del concepto de variedad diferenciable n-dimensional, donde hemos reemplazado $\mathbb{R}^n$ por el espacio de Fréchet E.

Ejemplo 5. Un espacio de Fréchet E es naturalmente una variedad de Fréchet. Basta tomar el atlas formado por la carta (E, Id), donde $Id: E \to E$ es la aplicación identidad en E. ◂

Ejemplo 6. Un abierto U en M es una subvariedad de M. ◂

Por lo tanto, $\mathbb{M}$ es una subvariedad de Fréchet del espacio de Fréchet $C^\infty(\mathbb{S}^1, \mathbb{R}^2)$, pero ¿qué ocurre con $\mathbb{M}_d$?

Tras la definición 4, extender la noción de curva diferenciable sobre una variedad de Fréchet resulta natural.

Definición 7. En una variedad de Fréchet M modelada sobre un espacio de Fréchet E, una curva (aplicación continua) $\gamma: (-\varepsilon, \varepsilon) \to M$ es **diferenciable** si para todo $t_0 \in (-\varepsilon, \varepsilon)$ existen un $\delta > 0$ y una carta (U, ϕ) de M tales que $(t_0 - \delta, t_0 + \delta) \subset (-\varepsilon, \varepsilon)$, $\gamma(t_0 - \delta, t_0 + \delta) \subset U$ y

$$\phi \circ \gamma: (t_0 - \delta, t_0 + \delta) \to \phi(U) \subset E$$

es una aplicación diferenciable [13]. ◂

Supongamos ahora que $F: M_1 \to M_2$ es una aplicación continua entre variedades de Fréchet modeladas sobre E_1 y E_2, respectivamente. La aplicación F es **diferenciable** en $p \in M_1$ si existe una carta (U_1, ϕ_1) de M_1 donde U_1 contiene a p y otra carta (U_2, ϕ_2) de M_2 donde U_2 contiene a $F(p)$, con $F(U_1) \subseteq U_2$, tales que la aplicación

$$\phi_2 \circ F \circ \phi_1^{-1}: \phi_1(U_1) \subseteq E_1 \to \phi_2(U_2) \subseteq E_2$$

es diferenciable. La aplicación F es diferenciable cuando lo es en cada uno de sus puntos.

En variedades finito dimensionales, un vector tangente en un punto de la variedad puede definirse de dos formas equivalentes: como una derivación en el espacio de funciones reales diferenciables en un entorno del punto, o como un vector tangente de una curva en la variedad que pasa por el punto. Esta equivalencia no ocurre en el caso infinito dimensional: todo vector velocidad de una curva es una derivación, pero no todas las derivaciones son inducidas desde una curva sobre la variedad. Nosotros utilizaremos los vectores velocidad de las curvas sobre la variedad, esto es, trabajaremos con el espacio tangente cinemático.

Si M es una variedad de Fréchet y $p \in M$, entonces el **espacio tangente cinemático** T_pM en p es el espacio cociente de curvas diferenciables $\gamma: (-\varepsilon, \varepsilon) \to M$ tales que $\gamma(0) = p$, con la relación de equivalencia

$$\gamma \sim \overline{\gamma} \iff \gamma'(0) = \overline{\gamma}'(0).$$

Se puede probar que T_pM es un espacio vectorial. Si U es un abierto de M, resulta que T_pU es isomorfo a T_pM y, en el caso de un espacio de Fréchet E, se tiene que T_pE es difeomorfo a E.

Seguidamente, supongamos que $F: M_1 \to M_2$ es una aplicación diferenciable en p.

Definición 8. La **aplicación inducida** por F entre los espacios tangentes T_pM_1 y $T_{F(p)}M_2$, o aplicación tangente en p, es la aplicación lineal $T_pF: T_pM_1 \to T_{F(p)}M_2$ tal que para todo $v \in T_pM_1$ se tiene que $T_pF(v) = (F \circ \gamma)'(0)$, donde $\gamma: (-\varepsilon, \varepsilon) \to M_1$ es una curva en M_1 tal que $\gamma(0) = p$ y $\gamma'(0) = v$. ◂

En general, cuando estamos trabajando con variedades de Fréchet, el teorema de la función inversa no es cierto. Es necesario que las variedades de Fréchet sean de un tipo especial, muy cercano a las variedades de Banach: las variedades Tame-Fréchet. Estas variedades no son objeto de estudio de este artículo. Solo necesitamos el siguiente resultado, que se demuestra a partir del corolario 2.3.2 (pág. 146) y del teorema 2.3.1 (pág. 196) del artículo de Hamilton [10].

Teorema 9. *Sean M_1 una variedad compacta, M_2 una variedad de dimensión finita, V un espacio vectorial finito dimensional, $F\colon\ U \subseteq C^\infty(M_1, M_2) \to V$ una aplicación diferenciable en un abierto U de $C^\infty(M_1, M_2)$ y $\gamma_0 \in U$. Si las aplicaciones lineales*

$$T_\gamma F\colon\ T_\gamma C^\infty(M_1, M_2) \to T_{F(\gamma)}V, \quad \forall\, \gamma \in F^{-1}(F(\gamma_0)),$$

son suprayectivas, entonces $F^{-1}(F(\gamma_0))$ es una subvariedad de U y

$$T_\gamma\left(F^{-1}(F(\gamma_0))\right) = \operatorname{Ker} T_\gamma F, \quad \forall\, \gamma \in F^{-1}(F(\gamma_0)).$$

De esta manera, volviendo a nuestro problema real, podemos tomar la aplicación diferenciable

$$(L, \overline{c})\colon\ \mathbb{M} \subset C^\infty(\mathbb{S}^1, \mathbb{R}^2) \to \mathbb{R} \times \mathbb{R}^2$$

definida por $(L, \overline{c})(f) = (L(f), \overline{c}(f))$. Está claro que $\mathbb{M}_d = (L, \overline{c})^{-1}(1, (0, 0))$. Por lo tanto, usando el teorema anterior podemos demostrar que $\mathbb{M}_d$ es una subvariedad de $\mathbb{M}$ y que $T_f \mathbb{M}_d = \operatorname{Ker} T_f(L, \overline{c})$.

4. Una métrica riemanniana sobre el espacio de las curvas cerradas regulares en el plano

A continuación, trataremos de definir una métrica en $\mathbb{M}$. Para ello, recordamos brevemente la noción de métrica de Riemann sobre una variedad de Fréchet.

Sean M una variedad de Fréchet modelada sobre E y $TM = \bigcup_{p\in M} T_pM$ el fibrado tangente de M. Denotamos por $\pi\colon\ TM \to M$ a la proyección $\pi(p, [\gamma]) = p$. Entonces, TM es una variedad de Fréchet modelada sobre $E \times E$ que hace a π diferenciable. En efecto, basta tomar un atlas de TM de la siguiente manera: $\{(\pi^{-1}(U_\alpha), \phi_\alpha)\}_{\alpha\in\mathbb{A}}$ tal que $\phi_\alpha\colon\ \pi^{-1}(U_\alpha) \to E \times E$ está definida como

$$\phi_\alpha(p, [\gamma]) = (\varphi_\alpha(p), (\varphi_\alpha \circ \gamma)'(0)),$$

siendo $\{(U_\alpha, \varphi_\alpha)\}_{\alpha\in\mathbb{A}}$ un atlas diferenciable de M.

Asimismo, el producto fibrado $TM \times_M TM \coloneqq \{(p, [\gamma_1], [\gamma_2]) \mid p \in M \ \text{ y } \ [\gamma_1], [\gamma_2] \in T_pM\}$ también es una variedad de Fréchet modelada sobre $E \times E \times E$. Basta considerar la proyección $\widetilde{\pi}\colon\ TM \times_M TM \to M$ y el atlas $\{(\widetilde{\pi}^{-1}(U_\alpha), \widetilde{\phi}_\alpha\}_{\alpha\in\mathbb{A}}$, tal que

$$\begin{array}{rrcl} \widetilde{\phi}_\alpha\colon & \widetilde{\pi}^{-1}(U_\alpha) & \to & E \times E \times E \\ & (p, [\gamma_1], [\gamma_2]) & \mapsto & (\varphi_\alpha(p), (\varphi_\alpha \circ \gamma_1)'(0), (\varphi_\alpha \circ \gamma_2)'(0))\,, \end{array}$$

donde $\{(U_\alpha, \varphi_\alpha)\}_{\alpha\in\mathbb{A}}$ es un atlas de M.

Ahora podemos definir la noción de métrica riemanniana en una variedad de Fréchet.

Definición 10. Sea M una variedad. Una **métrica riemanniana** es una aplicación diferenciable

$$\langle\cdot,\cdot\rangle_M\colon\ TM \times_M TM \to \mathbb{R}$$

tal que para todo $p \in M$ la aplicación $\langle\cdot,\cdot\rangle_M\colon\ T_pM \times T_pM \to \mathbb{R}$ es un producto escalar sobre T_pM. ◂

Si regresamos al problema de definir una métrica sobre $\mathbb{M}$, podemos definir la siguiente descomposición de los elementos h de $T_f\mathbb{M}$ [15]:

$$h(\theta) = h^t + h^l(f(\theta) - \overline{c}(f)) + L(f)h^d(\theta),$$

donde

1. $h^t = T_f\overline{c}(h) \in \mathbb{R}^2$,
2. $h^l = T_f(\ln L)(h) \in \mathbb{R}$,

3. $h^d = \frac{1}{L(f)}\left(h - h^t - h^l(f - \overline{c}(f))\right) \in \operatorname{Ker} T_f(L, \overline{c})$,

siendo L la aplicación longitud definida en la ecuación (1).

De este modo, se demuestra que la aplicación

$$\begin{array}{rccl} \phi : & T_f\mathbb{M} & \to & \mathbb{R}^2 \times \mathbb{R} \times \operatorname{Ker} T_f(L, \overline{c}) \\ & h & \mapsto & (h^t, h^l, h^d) \end{array}$$

es un isomorfismo de espacios vectoriales, lo que nos sugiere introducir la siguiente métrica riemanniana sobre $\mathbb{M}$ [15].

Proposición 11. *Sean h y k dos elementos de $T_f\mathbb{M}$. Entonces, la operación $\langle\cdot,\cdot\rangle_{\mathbb{M}} : T_f\mathbb{M} \times T_f\mathbb{M} \to \mathbb{R}$ dada por*

$$\langle h, k\rangle_{\mathbb{M}} = \langle h^t, k^t\rangle + h^l k^l + L(f)\int_0^{2\pi} \frac{\langle (h^d)'(\theta), (k^d)'(\theta)\rangle}{\|f'(\theta)\|}\,\mathrm{d}\theta$$

define una métrica riemanniana sobre $\mathbb{M}$, donde $\langle\cdot,\cdot\rangle$ es el producto escalar estándar en $\mathbb{R}^2$.

Nota 12. Sea $F : (M_1, \langle\cdot,\cdot\rangle_{M_1}) \to (M_2, \langle\cdot,\cdot\rangle_{M_2})$ un difeomorfismo entre variedades riemannianas de Fréchet. Diremos que F es una **isometría** si para todo $p \in M_1$ se tiene que

$$\langle T_pF([f_1]), T_pF([f_2])\rangle_{M_2} = \langle [f_1], [f_2]\rangle_{M_1}, \quad \forall [f_1], [f_2] \in T_pM_1.$$ ◂

Es más, $\mathbb{M}$ con esta métrica es isométrico a $\mathbb{R}^2 \times \mathbb{R} \times \mathbb{M}_d$ con la métrica derivada de los productos escalares correspondientes a $\mathbb{R}^2$ y $\mathbb{R}$, y el producto sobre $T_{f_0}\mathbb{M}_d$ siguiente :

$$\langle h_0, k_0\rangle_{\mathbb{M}_d} = \int_0^{2\pi} \frac{\langle h_0'(\theta), k_0'(\theta)\rangle}{\|f_0'(\theta)\|}\,\mathrm{d}\theta, \tag{2}$$

donde $f_0 \in \mathbb{M}_d$ y $h_0, k_0 \in T_{f_0}\mathbb{M}_d$.

Proposición 13. *La aplicación $\phi_1 : \mathbb{M} \to \mathbb{R}^2 \times \mathbb{R} \times \mathbb{M}_d$ definida por*

$$\phi_1(f) = \left(\overline{c}(f), \ln L(f), \frac{f - \overline{c}(f)}{L(f)}\right)$$

es una isometría respecto de $\langle\cdot,\cdot\rangle_{\mathbb{M}}$ y la métrica determinada por la estándar sobre $\mathbb{R}^2$, $\mathbb{R}$ y $\langle\cdot,\cdot\rangle_{\mathbb{M}_d}$.

Nota 14. Si restringimos $\langle\cdot,\cdot\rangle_{\mathbb{M}}$ a $T_{f_0}\mathbb{M}_d$, obtenemos $\langle\cdot,\cdot\rangle_{\mathbb{M}_d}$. ◂

Antes de finalizar esta sección, presentaremos otra descripción de la variedad riemanniana $(\mathbb{M}_d, \langle\cdot,\cdot\rangle_{\mathbb{M}_d})$ que nos será útil en lo sucesivo.

Proposición 15. *Si $\mathbb{M}_d' = \{f \in \mathbb{M} \mid L(f) = 1, f(0) = (0,0)\}$, entonces*

a) *$\mathbb{M}_d'$ es una subvariedad de $\mathbb{M}$ y su espacio tangente en $f \in \mathbb{M}_d'$ es*

$$T_f\mathbb{M}_d' = \{h \in T_f\mathbb{M} \mid T_fL(h) = 0, h(0) = (0,0)\}.$$

b) *La aplicación $\rho : \mathbb{M}_d \to \mathbb{M}_d'$ dada por $\rho(f) = f - f(0)$ es un difeomorfismo.*

c) *Si sobre $\mathbb{M}_d'$ consideramos la métrica riemanniana*

$$\langle h, k\rangle_{\mathbb{M}_d'} = \int_0^{2\pi} \frac{\langle h'(\theta), k'(\theta)\rangle}{\|f'(\theta)\|}\,\mathrm{d}\theta$$

con $h, k \in T_f\mathbb{M}_d'$ y $f \in \mathbb{M}_d'$, entonces $\rho : (\mathbb{M}_d, \langle\cdot,\cdot\rangle_{\mathbb{M}_d}) \to (\mathbb{M}_d', \langle\cdot,\cdot\rangle_{\mathbb{M}_d'})$ es una isometría.

5. El embebimiento

Queremos hallar la curva media extrínseca de una familia finita de curvas de $\mathbb{M}$ y, para ello, la estrategia será embeber esta variedad en un espacio vectorial euclídeo. Recordamos que $\mathbb{M}$ es difeomorfo al espacio producto $\mathbb{R}^2 \times \mathbb{R} \times \mathbb{M}_d$, por lo que podemos asumir que $\mathbb{M}_d$ codifica la parte no euclídea de $\mathbb{M}$. Por lo tanto, debemos preguntarnos en qué espacio vectorial euclídeo podemos embeber la subvariedad $\mathbb{M}_d$. La respuesta a esta pregunta la da el espacio euclídeo infinito dimensional $V = C^\infty(\mathbb{S}^1, \mathbb{R})$, cuyo producto escalar es

$$\langle a, b\rangle = \int_0^{2\pi} a(\theta) b(\theta)\, \mathrm{d}\theta.$$

Así, el embebimiento que vamos a considerar es

$$\mathbb{M} \cong \mathbb{R}^2 \times \mathbb{R} \times \mathbb{M}_d \hookrightarrow \mathbb{R}^2 \times \mathbb{R} \times V \times V.$$

A continuación, lo describimos con detalle.

Tomamos en $V \times V$ el conjunto de pares ortonormales bajo una cierta condición abierta

$$St^0(2, V) = \{(a, b) \in St(2, V) \mid (a(\theta), b(\theta)) \neq (0, 0)\ \forall\, \theta \in \mathbb{S}^1\},$$

donde $St(2, V) = \{(a, b) \in V \times V \mid \|a\|_V = \|b\|_V = 1, \langle a, b\rangle_V = 0\}$.

Se puede demostrar que $St(2, V)$ es una subvariedad de $V \times V$ y que $St^0(2, V)$ es un abierto de $St(2, V)$ y, por lo tanto, una subvariedad.

Asimismo, tomamos la aplicación

$$\begin{array}{rccl} \Psi: & St^0(2, V) & \to & \mathbb{M}_d' \\ & (a, b) & \mapsto & \Psi(a, b)(\theta) = \dfrac{1}{2}\displaystyle\int_0^\theta (a(s) + ib(s))^2\, \mathrm{d}s, \end{array}$$

que no es inyectiva, pues $\Psi(a, b) = \Psi(-a, -b)$.

Definición 16. Sea $p : \tilde{X} \to X$ una aplicación continua. Diremos que el subconjunto abierto $U \subset X$ está propiamente recubierto por p si $p^{-1}(U)$ es la unión disjunta de subconjuntos abiertos de $\tilde{X}$, cada uno de los cuales se aplica por p homeomórficamente sobre U. Se dice que la aplicación continua $p : \tilde{X} \to X$ es una **aplicación recubridora** si todo punto $x \in X$ tiene un entorno abierto propiamente recubierto por p. Entonces diremos que $p : \tilde{X} \to X$ es un recubrimiento, $\tilde{X}$ es el espacio recubridor de X y X es el espacio base de la aplicación recubridora p. ◂

De este modo, teniendo en cuenta la definición anterior, se puede demostrar que Ψ es una aplicación recubridora de dos hojas, las cuales denotamos por $(St^0(2, V))^+$, de manera que $St^0(2, V)$ es un recubridor de dos hojas de $\mathbb{M}_d'$.

Por consiguiente, $\mathbb{M}_d'$ es difeomorfa a cada una de las hojas del recubrimiento. Como $\mathbb{M}_d$ y $\mathbb{M}_d'$ son difeomorfos, podemos concluir que $\mathbb{M}_d$ también es difeomorfa con cada una de las hojas del recubrimiento y, por lo tanto,

$$\mathbb{M} \cong \mathbb{R}^2 \times \mathbb{R} \times \mathbb{M}_d \cong \mathbb{R}^2 \times \mathbb{R} \times (St^0(2, V))^+,$$

donde el difeomorfismo se define a través de la aplicación

$$\begin{array}{rccl} \phi: & \mathbb{R}^2 \times \mathbb{R} \times (St^0(2, V))^+ & \to & \mathbb{M} \\ & (v, l, (a, b)) & \mapsto & v + e^l \left(\Psi(a, b) - \bar{c}\,(\Psi(a, b))\right). \end{array} \tag{3}$$

Así, el embebimiento que estábamos buscando es

$$\begin{array}{ccl} \mathbb{M} & \hookrightarrow & \mathbb{R}^2 \times \mathbb{R} \times (St^0(2, V))^+ \subset \mathbb{R}^2 \times \mathbb{R} \times V \times V \\ f & \hookrightarrow & \left(\bar{c}(f), \ln L(f), (a, b)\right) \end{array},$$

donde

$$a(\theta) = \sqrt{2\|f'(\theta)\|}\cos\left(\frac{\alpha(\theta)}{2}\right) \quad \text{y} \quad b(\theta) = \sqrt{2\|f'(\theta)\|}\,\mathrm{sen}\left(\frac{\alpha(\theta)}{2}\right),$$

siendo $\alpha(\theta)$ el ángulo que forma $f'(\theta)$ con el eje OX medido en el sentido contrario a las agujas del reloj.

6. La curva media extrínseca de una familia finita de curvas planas cerradas regulares

Llegados a este punto, solo nos queda definir la curva media extrínseca de un conjunto finito de curvas $\{f_1, f_2, \cdots, f_n\}$ de $\mathbb{M}$. Estas curvas se pueden identificar con los elementos de $\mathbb{R}^2 \times \mathbb{R} \times V \times V$

$$\left(\overline{c}(f_i), \ln L(f_i), (a_i, b_i)\right),$$

de manera que podemos definir la media en $\mathbb{R}^2 \times \mathbb{R} \times V \times V$ como $(\overline{c}_m, \ln L_m, (a_m, b_m))$, donde

$$\overline{c}_m = \frac{1}{n}\sum_{i=1}^{n} \overline{c}(f_i), \qquad L_m = \sqrt[n]{\prod_{i=1}^{n} L(f_i)},$$

$$a_m(\theta) = \frac{1}{n}\sum_{i=1}^{n} a_i(\theta), \qquad b_m(\theta) = \frac{1}{n}\sum_{i=1}^{n} b_i(\theta).$$

Nota 17. Adviértase que L_m es la media geométrica de $L(f_i)$ para todo $i \in \{1, \ldots, n\}$ y que

$$\ln L_m = \frac{1}{n}\sum_{i=1}^{n} \ln L(f_i).$$ ◂

Lo que nos interesa es que $(\overline{c}_m, \ln L_m, (a_m, b_m))$ pertenezca a $\mathbb{R}^2 \times \mathbb{R} \times St^0(2, V)$, pero en general (a_m, b_m) no pertenece a $St^0(2, V)$. Por lo tanto, no es posible obtener el correspondiente elemento de $\mathbb{M}$ asociado por la aplicación ϕ dada en la ecuación (3). La estrategia para subsanar este inconveniente consiste en aplicar el proceso de normalización de Gram-Schmidt a (a_m, b_m) para conseguir un punto de $St^0(2, V)$. Los puntos restantes de este espacio son rotaciones de dicho punto. Así, podemos calcular la rotación correspondiente a la mínima distancia con (a_m, b_m). Para ello, debemos garantizar que ese elemento existe y es único, por lo que suponemos que (a_m, b_m) es un **punto no focal** de $St^0(2, V)$, esto es, existe un único $(a_m^0, b_m^0) \in St^0(2, V)$ tal que

$$d_0\left((a_m, b_m), St^0(2, V)\right) = d_0\left((a_m, b_m), (a_m^0, b_m^0)\right). \tag{4}$$

Para ver con más detalle este proceso se puede consultar el trabajo de Rodríguez Pérez [14].

Nota 18. Esta condición de focalidad se puede garantizar en gran parte de los casos prácticos. Además, en este caso, como $St(2, V)$ es un cerrado de $V \times V$, siempre existe $(\overline{a}_m, \overline{b}_m) \in St(2, V)$ tal que

$$d_0\left((a_m, b_m), St(2, V)\right) = d_0\left((a_m, b_m), (\overline{a}_m, \overline{b}_m)\right).$$ ◂

Para finalizar, todo el desarrollo teórico que hemos realizado nos permite cumplir nuestro objetivo: definir la curva media extrínseca.

Definición 19. La **curva media extrínseca** de una muestra $\{f_1, f_2, \ldots, f_n\}$ de curvas planas cerradas regulares es la curva obtenida de la siguiente fórmula:

$$f_m = \overline{c}_m + \sqrt[n]{\prod_{i=1}^{n} L(f_i)}\left(\frac{1}{2}\int_0^{\bullet} \left(a_m^0(s) + i b_m^0(s)\right)^2 \mathrm{d}s - \overline{c}\left(\frac{1}{2}\int_0^{\bullet} \left(a_m^0(s) + i b_m^0(s)\right)^2 \mathrm{d}s\right)\right).$$ ◂

Nota 20. Obsérvese que $f_m = \phi\left(\overline{c}_m, \ln L_m, (a_m^0, b_m^0)\right)$. ◂

7. El algoritmo

La teoría desarrollada en las secciones anteriores nos permite construir un algoritmo para resolver el problema de encontrar la curva media extrínseca. A continuación lo describiremos paso a paso y, posteriormente, estudiaremos un ejemplo sencillo. Así, el algoritmo se desarrolla en los siguientes pasos:

1. Establecer las curvas $\{f_1, f_2, \dots, f_n\}$ planas cerradas y regulares que forman la muestra.
2. Calcular los correspondientes elementos de cada curva en $\mathbb{R}^2 \times \mathbb{R} \times St^0(2, V)$, es decir,
$$\left(\bar{c}(f_i), \ln L(f_i), (a_i, b_i)\right),$$
donde
$$L(f_i) = \int_0^{2\pi} \|f_i'(\theta)\| \, \mathrm{d}\theta, \qquad \bar{c}(f_i) = \frac{1}{L(f_i)} \int_0^{2\pi} \|f_i'(\theta)\| f_i(\theta) \, \mathrm{d}\theta,$$
$$a_i(\theta) = \sqrt{\frac{2\|f_i'(\theta)\|}{L(f_i)}} \cos\left(\frac{\alpha_i(\theta)}{2}\right), \qquad b_i(\theta) = \sqrt{\frac{2\|f_i'(\theta)\|}{L(f_i)}} \operatorname{sen}\left(\frac{\alpha_i(\theta)}{2}\right),$$
siendo $\alpha_i(\theta)$ el ángulo que forma $f_i'(\theta) = (x_i'(\theta), y_i'(\theta))$ con el eje OX.
3. Calcular el valor medio $\left(\bar{c}_m, \ln L_m, (a_m, b_m)\right)$ en $\mathbb{R}^2 \times \mathbb{R} \times V \times V$ con las siguientes fórmulas:
$$\bar{c}_m = \frac{1}{n} \sum_{i=1}^{n} \bar{c}(f_i), \quad \ln L_m = \frac{1}{n} \sum_{i=1}^{n} \ln L(f_i), \quad a_m = \frac{1}{n} \sum_{i=1}^{n} a_i \quad \text{y} \quad b_m = \frac{1}{n} \sum_{i=1}^{n} b_i.$$
4. Construir el par $(a_m^0, b_m^0) \in St^0(2, V)$ que minimiza la distancia con $(a_m, b_m) \in V \times V$. Consideramos los elementos de $V \times V$
$$u_1 = a_m \qquad \text{y} \qquad u_2 = b_m - \frac{\langle a_m, b_m \rangle_V}{\langle a_m, a_m \rangle_V} a_m$$
y los normalizamos, obteniendo el par $\left(a_g = \frac{u_1}{\|u_1\|_V}, b_g = \frac{u_2}{\|u_2\|_V}\right)$. Por último, calculamos el elemento buscado con la fórmula
$$(a_m^0, b_m^0) = (a_g \cos\alpha - b_g \operatorname{sen}\alpha, a_g \operatorname{sen}\alpha + b_g \cos\alpha),$$
donde α es el ángulo
$$\alpha = \arctan \frac{\langle b_m, a_g \rangle_V}{\langle a_m, a_g \rangle_V + \langle b_m, b_g \rangle_V} \in \left[0, \frac{\pi}{2}\right[\cup \left]\frac{3\pi}{2}, 2\pi\right[.$$
5. Calcular la curva $\Psi(a_m^0, b_m^0) \in \mathbb{M}_d'$ mediante la expresión
$$\Psi(a_m^0, b_m^0)(\theta) = \frac{1}{2} \int_0^{\theta} \left(a_m^0(s) + i b_m^0(s)\right)^2 \mathrm{d}s.$$
6. Calcular el centroide de la curva $\Psi(a_m^0, b_m^0) \in \mathbb{M}_d'$, esto es,
$$\bar{c}\left(\Psi(a_m^0, b_m^0)\right) = \int_0^{2\pi} \left\|\left(\Psi(a_m^0, b_m^0)\right)'(\theta)\right\| \Psi(a_m^0, b_m^0)(\theta) \, \mathrm{d}\theta.$$
7. Construir la curva media extrínseca en $\mathbb{M}$, es decir,
$$f_m = \bar{c}_m + \sqrt[n]{\prod_{i=1}^{n} L(f_i)} \left(\Psi(a_m^0, b_m^0) - \bar{c}\left(\Psi(a_m^0, b_m^0)\right)\right).$$

Ejemplo 21. Para entender el proceso, presentamos un ejemplo matemático con tres circunferencias con centro el origen y radios 1, 2 y 6, respectivamente.

Consideramos las curvas $f_1(\theta) = (\cos\theta, \operatorname{sen}\theta)$, $f_2(\theta) = (2\cos\theta, 2\operatorname{sen}\theta)$ y $f_3(\theta) = (6\cos\theta, 6\operatorname{sen}\theta)$. Usando la fórmula de la longitud tenemos que $L(f_1) = 2\pi$, $L(f_2) = 4\pi$ y $L(f_3) = 12\pi$. Por otro lado, el centroide de todas ellas es el punto $(0, 0)$. Como
$$f_1'(\theta) = (-\operatorname{sen}\theta, \cos\theta) = \left(\cos\left(\theta + \frac{\pi}{2}\right), \operatorname{sen}\left(\theta + \frac{\pi}{2}\right)\right),$$

tenemos que el ángulo que forma $f_1'(\theta)$ con el eje OX es $\alpha_1(\theta) = \theta + \pi/2$ y, de la misma manera, deducimos que $\alpha_1(\theta) = \alpha_2(\theta) = \alpha_3(\theta)$.

Con estos datos, estamos en condiciones de construir los pares (a_i, b_i) correspondientes a cada curva, esto es,

$$(a_1(\theta), b_1(\theta)) = (a_2(\theta), b_2(\theta)) = (a_3(\theta), b_3(\theta)) = \left(\sqrt{\frac{1}{\pi}}\cos\left(\frac{\theta}{2} + \frac{\pi}{4}\right), \sqrt{\frac{1}{\pi}}\operatorname{sen}\left(\frac{\theta}{2} + \frac{\pi}{4}\right)\right).$$

De este modo, el par (a_m, b_m) es

$$(a_m(\theta), b_m(\theta)) = \left(\sqrt{\frac{1}{\pi}}\cos\left(\frac{\theta}{2} + \frac{\pi}{4}\right), \sqrt{\frac{1}{\pi}}\operatorname{sen}\left(\frac{\theta}{2} + \frac{\pi}{4}\right)\right).$$

Nótese que $(a_m, b_m) \in St^0(2, V)$, por lo que $(a_m^0, b_m^0) = (a_m, b_m)$.

Finalmente, calculamos la curva imagen por la aplicación Ψ, es decir,

$$\Psi(a_m^0, b_m^0)(\theta) = \frac{1}{2}\int_0^\theta \left(a_m^0(s) + ib_m^0(s)\right)^2 \mathrm{d}s = \frac{1}{2\pi}(\cos(\theta) - 1, \operatorname{sen}(\theta)),$$

y su centroide, o sea,

$$\bar{c}(\Psi(a_m^0, b_m^0)) = \left(\frac{-1}{2\pi}, 0\right).$$

Así, concluimos que la curva media en este ejemplo es la circunferencia de radio $\sqrt[3]{12}$ y centro $(0, 0)$

$$f_m(\theta) = \sqrt[3]{12}(\cos\theta, \operatorname{sen}\theta).$$ ◂

Nota 22. Téngase en cuenta que la curva media extrínseca es una circunferencia cuyo radio es la media geométrica de los radios de la muestra. Este tipo de media es menos sensible a valores extremos que la media aritmética y aparece como consecuencia de la introducción del logaritmo neperiano en la segunda componente de $\mathbb{M} \cong \mathbb{R}^2 \times \mathbb{R} \times (St^0(2, V))^+$. Es necesario aplicar el logaritmo neperiano porque tomar directamente las longitudes implicaría trabajar en $\mathbb{R}^2 \times \mathbb{R}^+ \times V \times V$ y este espacio no es un espacio vectorial. ◂

8. Una discretización del algoritmo

En el ejemplo 21 de la sección anterior tenemos las parametrizaciones de las curvas, pero esto no suele ser así. Lo normal es que conozcamos una serie de puntos de cada curva, por lo que debemos discretizar el proceso usando métodos numéricos de aproximación. En concreto, debemos discretizar dos operaciones: la derivación y la integración.

Para la elaboración del código en MATLAB[1], hemos optado por utilizar los métodos numéricos que a continuación describimos:

1. Para el cálculo de la derivada utilizamos el método del punto medio con cinco nodos, es decir,

   $$f'(\theta) \approx \frac{1}{12h}\Big[-f(\theta + 2h) + 8f(\theta + h) - 8f(\theta - h) + f(\theta - 2h)\Big],$$

   donde $\theta \in \{0, h, 2h, \dots, 2\pi\}$ y h es el tamaño de paso. El error cometido al usar este método es

   $$\frac{h^4}{30}f^{(5)}(\xi), \quad \xi \in [\theta - 2h, \theta + 2h],$$

   por lo que es un método de orden 5 [3, pág. 178].

[1] Archivo `curva2.m`, disponible en https://temat.es/articulo/2019-p17/a-curva2-m.

2. Para el cálculo de las integrales usamos el método trapezoidal compuesto, esto es,

$$\int_0^{2\pi} f(\theta)\,\mathrm{d}\theta \approx \frac{h}{2}\Big[f(0) + f(2\pi) + 2\sum_{i=2}^{N-1} f(\theta_i)\Big],$$

donde $\theta_i \in \{0, h, 2h, \dots, 2\pi\}$. En este caso, el error cometido por el método es

$$\frac{2\pi h^2}{12} f''(\xi), \quad \xi \in [0, 2\pi]$$

y, por tanto, tiene orden 2 [3, pág. 194].

En lo que se refiere a los datos del algoritmo, debemos aportar como datos de entrada el número de curvas de la muestra (n), el número de puntos que se toman de cada curva (N), el tamaño de paso h y el conjunto de puntos escogidos. Debido al uso del método trapezoidal compuesto, es preciso poner una restricción sobre los puntos, pues tienen que ser tomados equiespaciadamente con tamaño de paso $h = \dfrac{2\pi}{N-1}$. De esta manera, se obtiene como dato de salida una representación de las curvas de la muestra junto con la curva media extrínseca.

Nota 23. Se podría plantear el uso de un método de mayor orden para la aproximación integral o, incluso, métodos que no exijan que los puntos sean equiespaciados. No obstante, se debe tener en cuenta que la integral del paso 5 es una integral acumulativa, por lo que no debe haber condiciones sobre el número de puntos a escoger. ◂

Así, si aplicamos el algoritmo al ejemplo 21 de las circunferencias descritas tomando 18 puntos, obtenemos el siguiente resultado, donde la curva media extrínseca es la curva dibujada en negro.

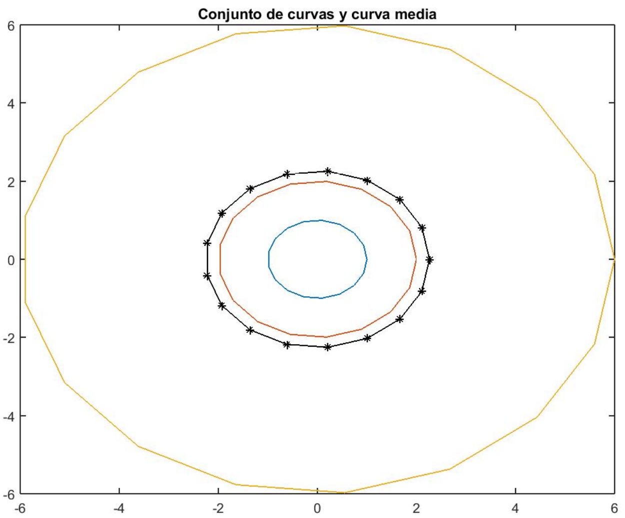

Figura 1: Curvas de radios 1, 2 y 6 y curva media extrínseca.

Como se observa, el algoritmo genera una curva que resulta ser una circunferencia centrada en el origen y de radio 2,26, por lo que se comete un error de 0,03.

9. Aplicación a una imagen médica

Para finalizar este artículo, aplicamos el algoritmo al problema de delimitar un tumor cerebral en una radiografía (figura 2) tomada por Gaillard [7]. Con la ayuda de tres expertos, hemos delimitado el tumor usando el programa GeoGebra. Las tres curvas que formaron la muestra son las que aparecen en la imagen de la derecha de la figura 2.

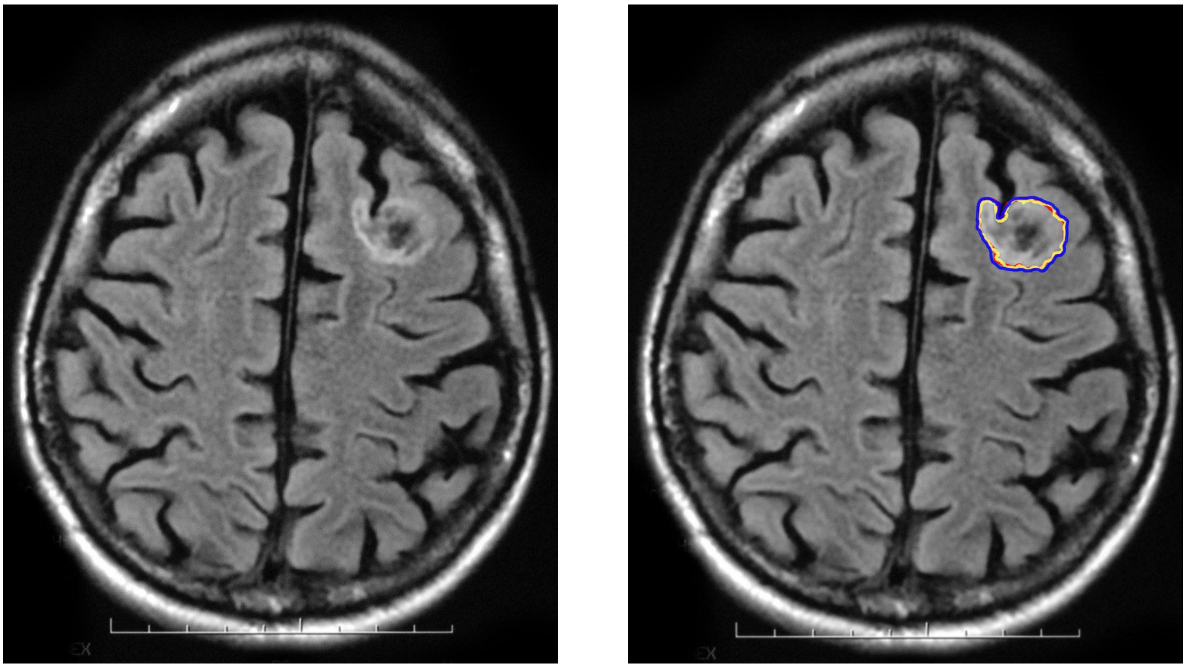

Figura 2: Tumor neuroepitelial disembrioplástico y curvas de la muestra. Imagen original (izquierda) tomada por Gaillard [7] con licencia CC BY 3.0 Internacional.

Nota 24. Existen programas más sofisticados para realizar estas delimitaciones, pero en este artículo solo queremos mostrar el proceso. ◂

Debido a que el tumor no es convexo, no basta utilizar 18 puntos como en el ejemplo anterior. En este caso, para poder tomar el mismo número de puntos en las tres curvas, se ha elegido un punto inicial con un ángulo θ_0 y el resto de puntos se han considerado con un tamaño de paso de 0,2244, lo que genera 33 puntos. De este modo, se han obtenido los siguientes resultados:

Cuadro 1: Longitudes y centroides de las curvas.

	Longitud	Centroide
Curva amarilla (c1)	4,676 413	$(-0{,}083\,440, -0{,}021\,952)$
Curva roja (c2)	4,677 151	$(-0{,}092\,342, -0{,}021\,112)$
Curva azul (c3)	4,946 474	$(-0{,}086\,342, -0{,}032\,029)$
f_m	4,602 473	$(-0{,}087\,374, -0{,}025\,031)$

En la figura 3, se representa el resultado del algoritmo en la imagen de la izquierda y, en la de la derecha, el conjunto de curvas sobre la radiografía, con la curva media extrínseca señalada en verde.

10. Conclusiones

El análisis de formas resulta ser un área de investigación de las matemáticas que permite dar respuestas a problemas reales de diferentes ámbitos. Por ejemplo, en el área de la oncología, podemos plantear problemas como la evolución de un tumor o la decisión de qué zonas se deben radiar y qué zonas no. Pero también en otros contextos, como en el ámbito de la industria textil, donde podemos plantear la búsqueda de una talla estándar; en la agricultura, donde podemos encontrar el área óptima para aplicar fertilizante a los cultivos; en el área de la geología, detectando la ubicación aproximada de yacimientos de petróleo y de minerales; en el ámbito de la vulcanología, donde podemos analizar la evolución de la

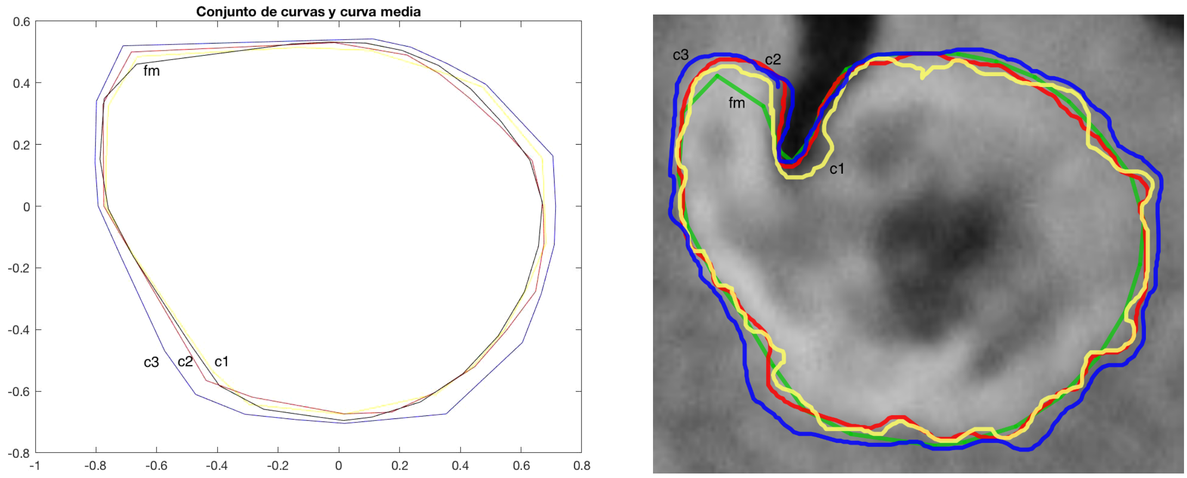

Figura 3: Resultado del algoritmo y representación del conjunto de curvas sobre la radiografía de la figura 2.

actividad volcánica, o en el ámbito de la aeronomía, en el que podemos dar una forma plana aproximada de posibles nubes de gases que se encuentren en la atmósfera.

En este artículo se ha introducido al lector en el fascinante mundo de las variedades de Fréchet y se ha explicado el proceso contenido en el artículo de Gual-Arnau, Ibáñez Gual y Simó Vidal [9]. A su vez, el algoritmo desarrollado ha resultado ser eficiente con el ejemplo matemático de las tres circunferencias concéntricas y con el ejemplo médico del tumor neuroepitelial disembrioplástico. Aunque para establecer la eficacia del mismo es necesario estudiar más casos y hacer un análisis de errores, podemos concluir que en este artículo se ha cumplido el objetivo planteado al inicio del mismo.

En particular, hemos tratado el problema del cálculo de la forma media a partir de una muestra de formas. Este problema se puede contemplar desde dos perspectivas diferentes: calcular una media intrínseca o una media extrínseca. La primera perspectiva implica trabajar con la estructura geométrica de la variedad riemanniana de Fréchet del espacio de curvas planas cerradas regulares. La otra opción consiste en embeber esta variedad en un espacio euclídeo en el que podamos calcular la media y posteriormente, encontrar la curva (media extrínseca) en $\mathbb{M}$ que minimiza la distancia con la media en el espacio embebido. La media extrínseca [9] se utiliza para evaluar la variabilidad de las observaciones que se hacen de una misma imagen. Sin embargo, sería interesante comparar ambas perspectivas (la intrínseca y la extrínseca) para el cálculo de la media.

Asimismo, recientes investigaciones plantean nuevos retos con curvas o superficies cerradas en el espacio o con curvas planas no cerradas, las cuales son útiles para modelar los surcos del cerebro. Por consiguiente, podemos afirmar que la aplicación de la geometría en el espacio de las formas es un campo abierto a múltiples investigaciones.

Referencias

[1] Azencott, Robert; Coldefy, François, y Younes, Laurent. «A distance for elastic matching in object recognition.» En: *Proceedings of 13th International Conference on Pattern Recognition* (Vienna). IEEE. 1996, págs. 687-691. https://doi.org/10.1109/ICPR.1996.546112.

[2] Bookstein, Fred L. *The Measurement of Biological Shape and Shape Change*. Vol. 24. Lecture Notes in Biomathematics. Springer Science y Business Media, New York, 1978. https://doi.org/10.1007/978-3-642-93093-5.

[3] Burden, Richard L. y Faires, J. Douglas. *Numerical Analysis*. 9.ª ed. Brooks/Cole, Cengage Learning, 2011. isbn: 978-0-538-73351-9.

[4] Dryden, Ian L. y Mardia, Kanti V. *Statistical shape analysis.* Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 1998, págs. xx+347. isbn: 978-0-471-95816-1.

[5] Flores Compañ, María Jesús. *Media muestral intrínseca en el espacio de curvas planas.* Trabajo de Fin de Máster. Universitat Jaume I, 2014. url: http://hdl.handle.net/10234/108703.

[6] Flores Compañ, María Jesús; Gual-Arnau, Ximo; Ibañez Gual, M. Victoria, y Simó Vidal, Amelia. «Intrinsic sample mean in the space of planar shapes». En: *Pattern Recognition* 60 (2016), págs. 164-176. issn: 0031-3203. https://doi.org/10.1016/j.patcog.2016.04.025.

[7] Gaillard, Frank. *Dysembryoplastic neuroepithelial tumour, MRI FLAIR.* Imagen. 2008. url: https://commons.wikimedia.org/wiki/File:DNET02.jpg.

[8] Goodall, Colin. «Procrustes Methods in the Statistical Analysis of Shape». En: *Journal of the Royal Statistical Society: Series B (Methodological)* 53.2 (1991), págs. 285-321. issn: 0035-9246. https://doi.org/10.1111/j.2517-6161.1991.tb01825.x.

[9] Gual-Arnau, Ximo; Ibáñez Gual, M. Victoria, y Simó Vidal, Amelia. «A new extrinsic sample mean in the shape space with applications to the boundaries of anatomical structures». En: *Biometrical Journal* 57.3 (2015), págs. 502-516. issn: 0323-3847. https://doi.org/10.1002/bimj.201400097.

[10] Hamilton, Richard S. «The inverse function theorem of Nash and Moser». En: *American Mathematical Society. Bulletin. New Series* 7.1 (1982), págs. 65-222. issn: 0273-0979. https://doi.org/10.1090/S0273-0979-1982-15004-2.

[11] Kendall, David G. «Shape manifolds, Procrustean metrics, and complex projective spaces». En: *The Bulletin of the London Mathematical Society* 16.2 (1984), págs. 81-121. issn: 0024-6093. https://doi.org/10.1112/blms/16.2.81.

[12] Kendall, David G.; Barden, Dennis.; Carne, Thomas K., y Le, Huiling. *Shape and shape theory.* Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 1999, págs. xii+306. https://doi.org/10.1002/9780470317006.

[13] Kriegl, Andreas y Michor, Peter W. *The convenient setting of global analysis.* Vol. 53. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 1997, págs. x+618. https://doi.org/10.1090/surv/053.

[14] Rodríguez Pérez, Clara. *Geometría Diferencial en el estudio de imágenes médicas.* Trabajo de Fin de Grado. Universidad de La Laguna, 2017. url: https://riull.ull.es/xmlui/handle/915/6757.

[15] Sundaramoorthi, Ganesh; Mennucci, Andrea; Soatto, Stefano, y Yezzi, Anthony. «A new geometric metric in the space of curves, and applications to tracking deforming objects by prediction and filtering». En: *SIAM Journal on Imaging Sciences* 4.1 (2011), págs. 109-145. issn: 1936-4954. https://doi.org/10.1137/090781139.

[16] Younes, Laurent. «Optimal matching between shapes via elastic deformations». En: *Image and Vision Computing* 17.5 (1999), págs. 381-389. issn: 0262-8856. https://doi.org/10.1016/S0262-8856(98)00125-5.

TEMat

El teorema de Karush-Kuhn-Tucker, una generalización del teorema de los multiplicadores de Lagrange, y programación convexa

Fco. Javier Martínez Sánchez
Facultad de Ciencias,
Universidad de Granada
javims282@gmail.com

Resumen: El presente artículo pretende mostrar una generalización del teorema de los multiplicadores de Lagrange, que resuelve problemas de optimización condicionados solo a restricciones de igualdad. El teorema de Karush-Kuhn-Tucker es una extensión suya que resuelve problemas de optimización condicionados tanto a restricciones de igualdad como de desigualdad. En la primera sección del presente texto, se enuncia y comenta el teorema de Lagrange y se incluye un ejemplo de aplicación. En la segunda sección, se enuncia y se demuestra el teorema que extiende al teorema de Lagrange, incluyendo un ejemplo ilustrativo. En la tercera y última sección, se hace una breve introducción a la programación convexa y cóncava y se prueba la condición suficiente en programación convexa y cóncava.

Abstract: This paper expects to show a generalization of the Lagrange multiplier rule, which solves optimization problems with only equality constraints. The Karush-Kuhn-Tucker theorem is an extention of this result in which inequality constraints are also considered. In the first section of this text, we discuss the Lagrange multiplier rule, including one example. In the second one, we prove the Karush-Kuhn-Tucker theorem, including another example. In the third and last one, we make a brief introduction to convex and concave programming and we prove a sufficient condition in convex and concave programming.

Palabras clave: optimización condicionada, programación no lineal, Lagrange, Karush-Kuhn-Tucker, programación convexa, programación cóncava.

MSC2010: 90C30.

Recibido: 28 de octubre de 2018.
Aceptado: 4 de marzo de 2019.

Agradecimientos: A ti, lector.

Referencia: Martínez Sánchez, Fco. Javier. «El teorema de Karush-Kuhn-Tucker, una generalización del teorema de los multiplicadores de Lagrange, y programación convexa». En: *TEMat*, 3 (2019), págs. 33-44. issn: 2530-9633. url: https://temat.es/articulo/2019-p33.

1. Introducción

A lo largo de la historia, la *optimización* ha resultado fructífera a la hora de resolver numerosos problemas de naturaleza variada tanto en matemáticas como en física o economía, entre otras. La optimización es el campo de las matemáticas dedicado a minimizar o maximizar una determinada función, distinguiéndose dos grandes ramas dentro de esta, a saber: *optimización libre* y *optimización condicionada*.

Hoy en día es habitual, en el curso de Cálculo, estudiar que los valores extremos de una función real y derivable f definida en un intervalo abierto $I \subset \mathbb{R}$ se encuentran entre los puntos $x \in I$ tales que $f'(x) = 0$. Y, dado $n \in \mathbb{N}$, los valores extremos de una función real de varias variables y diferenciable definida en un subconjunto abierto $\Omega \subset \mathbb{R}^n$ se encuentran entre los puntos $\boldsymbol{x} \in \Omega$ tales que $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$.

Dados $n \in \mathbb{N}$ un número natural, $\Omega \subset \mathbb{R}^n$ un subconjunto no vacío de $\mathbb{R}^n$ y $f : \Omega \to \mathbb{R}$ una función real definida en Ω, la optimización libre trata de resolver el siguiente problema:

$$\begin{cases} \text{minimizar/maximizar } f(\boldsymbol{x}) \\ \quad \boldsymbol{x} \in \Omega. \end{cases}$$

Pero ¿qué pasa cuando se buscan el mínimo y máximo de una función condicionados a ciertas restricciones o ligaduras? La optimización condicionada se encarga de responder a esta pregunta. En los siglos XVII y XVIII, grandes matemáticos (en especial, Lagrange) se ocuparon de obtener máximos y mínimos condicionados de determinadas funciones. A mediados del siglo XVIII, Lagrange publicó un método para resolver tales problemas de optimización condicionada solo a restricciones de igualdad: el *método de los multiplicadores de Lagrange*.

El objetivo del presente texto es mostrar una extensión del teorema de Lagrange que sirva para resolver un programa mixto, que es un problema de optimización condicionada donde se minimiza o maximiza una función sujeta a ambos tipos de restricciones: de igualdad y de desigualdad.

La *programación lineal* versa sobre la resolución del programa mixto lineal, esto es, el problema de optimización condicionada a restricciones mixtas (de igualdad y desigualdad)

$$(\star) \qquad \begin{cases} \text{minimizar/maximizar } f(\boldsymbol{x}) \text{ sujeto a} \\ \quad f_1(\boldsymbol{x}) = 0, \dots, f_k(\boldsymbol{x}) = 0, f_{k+1}(\boldsymbol{x}) \leq 0, \dots, f_m(\boldsymbol{x}) \leq 0, \\ \quad \boldsymbol{x} \in \Omega, \end{cases}$$

donde $\Omega \subset \mathbb{R}^n$, $n, k, m \in \mathbb{N}$ y todas las funciones involucradas $f, f_1, \dots, f_m$ son funciones lineales.

Exceptuando al matemático francés G. Monge (1746-1818), quien en 1776 se interesó por problemas de este tipo, debemos remontarnos al año 1939 para encontrar nuevos estudios relacionados con los métodos de la actual programación lineal, entre los que destacan los siguientes matemáticos:

Leonid V. Kantoróvich (1912-1986) publicó una extensa monografía titulada «Mathematical methods of organizing and planning production» [4] en 1939, en la que se hace corresponder una extensa gama de problemas con una teoría matemática concisa.

Tjalling C. Koopmans (1910-1985) formuló el *problema de transporte* para conseguir determinar los planes de embarque al mínimo coste total, conociendo de antemano la disponibilidad y demanda de cada puerto, con la ayuda de Kantarovitch. Ambos fueron galardonados con el Premio Nobel de Economía en 1975.

George B. Dantzig (1914-2015) desarrolló un método iterativo y eficaz de resolución del programa lineal, llamado *método símplex*, utilizado para resolver el problema del puente aéreo de Berlín. Dantzig recibió el Premio de Teoría John von Neumann de la Sociedad Americana de Investigación Operativa del año 1975.

John von Neumann (1903-1957) estableció los fundamentos matemáticos de la programación lineal en 1947, al relacionar esta con la teoría de juegos, que había publicado tres años antes, junto con Oskar Morgenstern, en el libro *Theory of Games and Economic Behavior* [11].

Martínez Sánchez [9] y Dantzig [3] ofrecen una información más ampliada de lo expuesto arriba. Este trabajo trata sobre el teorema fundamental dedicado a resolver el programa mixto general (como el programa ($\star$) pero donde las funciones involucradas no tienen por qué ser lineales).

2. El teorema de Langrange

En esta sección, se define la noción de programa con restricciones de igualdad, se enuncia el teorema de Lagrange (utilizado para resolver tales programas) y se muestra un ejemplo ilustrativo.

Definición 1 (programa con restricciones de igualdad). Dados $n, m \in \mathbb{N}$, $\Omega \subset \mathbb{R}^n$ un subconjunto abierto y no vacío de $\mathbb{R}^n$ y $m+1$ funciones reales $f, g_1, \dots, g_m$ de clase $\mathcal{C}^1$ definidas en Ω, un **programa con restricciones de igualdad** es un problema de optimización condicionada de la forma

$$\text{(PI)} \qquad \begin{cases} \min/\max f(\boldsymbol{x}) \text{ sujeto a} \\ \quad g_1(\boldsymbol{x}) = 0, \dots, g_m(\boldsymbol{x}) = 0, \\ \quad \boldsymbol{x} \in \Omega. \end{cases}$$

Lagrange publicó de manera oficial su resultado en su obra *Mécanique analytique* [8] de 1788 (aunque obtuvo el resultado en agosto 1755 cuando se lo envió por carta a su amigo L. Euler). A continuación se enuncia la versión del teorema de Lagrange del libro de Apostol [1].

Teorema 2 (Lagrange, 1788). *En la situación de la definición 1, si $m < n$, f tiene un extremo condicionado por $g_1(\boldsymbol{x}) = 0, \dots, g_m(\boldsymbol{x}) = 0$ en $\boldsymbol{x}^* \in \Omega$ y la matriz jacobiana de $\mathbf{g} = (g_1, \dots, g_m)$ en $\boldsymbol{x}^*$ tiene rango máximo m, entonces existen m números reales $\lambda_1, \dots, \lambda_m \in \mathbb{R}$ tales que*

$$\nabla f(\boldsymbol{x}^*) + \sum_{k=1}^{m} \lambda_k \nabla g_k(\boldsymbol{x}^*) = \mathbf{0}. \qquad (1)$$

La demostración mas común de este teorema hace uso del teorema de la función implícita, hecho por el cual se exige como hipótesis que $m < n$, y puede verse en el libro de Apostol [1]. Sin embargo, esta hipótesis quedará eliminada en la versión general del teorema (teorema de Karush-Kuhn-Tucker). En honor a Lagrange, la ecuación (1) recibe el nombre de **condición de Lagrange** y los escalares $\lambda_1, \dots, \lambda_m$ se denominan **multiplicadores de Lagrange**. Es conveniente advertir que el recíproco del teorema de Lagrange no es cierto: es posible que la condición de Lagrange tenga solución $\boldsymbol{x}^* \in \Omega$ pero que $\boldsymbol{x}^*$ no sea ni mínimo ni máximo de f condicionado a las restricciones de igualdad $g_1(\boldsymbol{x}) = 0, \dots, g_m(\boldsymbol{x}) = 0$. Es conocido que los multiplicadores de Lagrange tienen una interpretación económica. No es el objetivo de este trabajo exponer esta interpretación, pero si el lector lo desea podrá encontrar información al respecto en el libro de Sydsaeter y Hammond [14].

A continuación, se expone un ejemplo sencillo de aplicación del teorema de Lagrange.

Ejemplo 3. Optimización de la función $f : \mathbb{R}^2 \to \mathbb{R}$ dada por $f(x, y) = 2xy$ para todo $(x, y) \in \mathbb{R}^2$ sujeta a la restricción $g(x, y) = x^2 + y^2 \leq 1$.

La función f es continua y el conjunto $K = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$ es compacto, luego el teorema de Weierstrass garantiza la existencia de mínimo y máximo de f en K. A la hora de resolver el problema, se distinguen dos casos según si el punto donde f tiene un extremo global pertenece al interior o a la frontera de K.

Por un lado, si el punto donde f tiene un extremo global pertenece al interior de K, entonces se aplica la condición necesaria de existencia de extremo en puntos interiores: las derivadas parciales de f en dicho punto deben anularse, dando lugar a un único punto candidato a extremo:

$$\nabla f(x, y) = 0 \iff \begin{cases} 2y = 0, \\ 2x = 0 \end{cases} \iff (x, y) = (0, 0).$$

Por otro lado, si el punto donde f tiene un extremo global pertenece a la frontera de K, $\mathbb{S} = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$, entonces se aplica el método de los multiplicadores de Lagrange:

$$\text{Condición de Lagrange: } \begin{cases} 2y + 2\lambda x = 0, \\ 2x + 2\lambda y = 0, \\ x^2 + y^2 = 1 \end{cases} \implies \lambda = \pm 1. \qquad \text{Solución: } \begin{cases} \lambda = 1 \implies \left(\pm\frac{\sqrt{2}}{2}, \mp\frac{\sqrt{2}}{2}\right), \\ \lambda = -1 \implies \left(\pm\frac{\sqrt{2}}{2}, \pm\frac{\sqrt{2}}{2}\right). \end{cases}$$

Basta comprobar los valores que toma f en cada uno de los puntos obtenidos,

$$f(0,0)=0, \qquad f\left(\pm\frac{\sqrt{2}}{2},\mp\frac{\sqrt{2}}{2}\right)=-1, \qquad f\left(\pm\frac{\sqrt{2}}{2},\pm\frac{\sqrt{2}}{2}\right)=1,$$

y se concluye entonces que el mínimo de f condicionado a $x^2+y^2 \leq 1$ vale -1 y se alcanza en los puntos $\left(\pm\frac{\sqrt{2}}{2},\mp\frac{\sqrt{2}}{2}\right)$ y que el máximo de f condicionado a $x^2+y^2 \leq 1$ vale 1 y se alcanza en los puntos $\left(\pm\frac{\sqrt{2}}{2},\pm\frac{\sqrt{2}}{2}\right)$. ◂

Las aplicaciones del teorema de Lagrange son muy variadas y numerosas. En particular, resulta verdaderamente útil para demostrar resultados relevantes de análisis matemático como, por ejemplo, las constantes de equivalencia óptimas entre las normas $\|\cdot\|_1$ y $\|\cdot\|_\infty$ de $\mathbb{R}^n$, la desigualdad de Cauchy-Schwarz, el hecho de que toda matriz simétrica real es diagonalizable en $\mathbb{R}$ o la desigualdad de Hadamard, así como para resolver problemas de carácter geométrico, entre otros. En el trabajo de fin de grado de Martínez Sánchez [9], el lector encontrará todo lo anterior. Además, Wu y Wu [15] realizan varias demostraciones de la desigualdad de Cauchy-Schwarz bastante sorprendentes por su sencillez.

Al intentar resolver un problema de extremos condicionados por el método de los multiplicadores de Lagrange, teóricamente es sencillo determinar el sistema de Lagrange asociado, pero en la práctica no siempre existe un procedimiento simple y rápido para resolverlo de manera exacta. En esa situación, una posibilidad es aplicar métodos numéricos con los que se obtengan buenas aproximaciones de la solución del sistema de Lagrange. El lector puede consultar algunos de estos métodos numéricos en el libro de Peressini, Sullivan y Uhl [12].

Ahora bien, ¿qué ocurre cuando se incluyen restricciones de desigualdad en el programa (PI)? El teorema de Karush-Kuhn-Tucker tiene la respuesta, como se verá en la siguiente sección.

3. El teorema de Karush-Kuhn-Tucker

El *teorema de Karush-Kuhn-Tucker* es el primer y principal resultado de toda una teoría que se desarrolló a su alrededor y que dio lugar posteriormente a la *programación no lineal*. En esta sección se enuncia y demuestra dicho teorema, acompañado de un ejemplo ilustrativo.

Antes de enunciar y demostrar el mencionado teorema, veamos sus distintos orígenes históricos, así como su relación con la Segunda Guerra Mundial. En lo que sigue, se hará un breve resumen del artículo de Kjeldsen [6]. Si el lector está interesado puede consultar dicho artículo para profundizar aún más en este tema. Básicamente, el teorema de Karush-Kuhn-Tucker tuvo dos orígenes muy distintos.

En primer lugar, hay que hablar del matemático estadounidense William Karush (1917-1997), quien cursó los estudios de matemáticas en la Universidad de Chicago y cuyo trabajo de fin de máster tenía como título «Minima of functions of several variables with inequalities as side conditions» (1939) [5]. La motivación de Karush era extender un artículo publicado el año anterior por el que, en aquel momento, era el jefe del Departamento de Matemáticas de la Universidad de Chicago, A. Bliss, y que tenía por título «Normality and abnormality in the calculus of variations» [2]. El resultado que demostró Karush en su trabajo en 1939 pertenece indudablemente al campo de la programación no lineal, pero esta área no existía en aquel momento.

El Departamento de Matemáticas de la Universidad de Chicago, fundado con la apertura de la misma en 1892, estaba dirigido por E. Moore junto con G. Bolza y H. Maschke, quienes condujeron al departamento a ser uno de los mas influyentes en matemáticas en Estados Unidos, especialmente en cálculo de variaciones. Bolza estaba profundamente interesado en esta rama de las matemáticas[1] y creó un amplio y fuerte grupo de investigación dedicado única y exclusivamente al cálculo de variaciones. Este grupo fue conocido posteriormente como la *Escuela de Chicago* en cálculo de variaciones y estaba formada tanto por profesores como por alumnos interesados.

[1] Hay que mencionar que Bolza se interesó en el cálculo de variaciones tras asistir a una conferencia de K. Weierstrass en 1879.

En 1908, Maschke falleció y, dos años después, Bolza regresó a Alemania, su país natal. Chicago perdió así a dos líderes matemáticos, lo que se tradujo en un declive que acabó con la llegada al departamento de un nuevo equipo liderado por A. Bliss. Entre 1927 y 1941, el nuevo departamento y, sobre todo, Bliss, que fue alumno de Bolza, continuaron con la tradición de los anteriores líderes y, de nuevo, el departamento se caracterizó por un intenso estudio en cálculo de variaciones que ocupó la mayor parte de la investigación matemática en Chicago. De hecho, entre 1927 y 1937, Bliss dirigió treinta y cinco tesis doctorales, de las cuales treinta y cuatro pertenecían al cálculo de variaciones.

Como estudiante en Chicago, Karush fue producto de esta tradición. En su trabajo, Karush demostró una condición necesaria para la existencia de mínimo local de una función de varias variables $f = f(x_1, \dots, x_n)$ sujeta a desigualdades de la forma $g_1(x) \geq 0, \dots, g_m(x) \geq 0$ con $n, m \in \mathbb{N}$. Karush llevó a cabo su trabajo en 1939 mientras que la Escuela de Chicago se centraba en problemas de cálculo de variaciones con restricciones de desigualdad, donde se minimizaban funcionales de la forma

$$\mathcal{F}[\varphi] = \int_a^b F(x, \varphi(x), \varphi'(x))\, \mathrm{d}x$$

en el conjunto $D = \{\varphi \in \mathcal{C}^1(a,b) : \varphi(a) = A, \varphi(b) = B\}$ para $A, B \in \mathbb{R}$ fijos, $a < b$ y F una función conocida.

Así pues, el trabajo de Karush fue concebido como una versión finitodimensional de los problemas que se atacaban en el cálculo de variaciones y, por lo tanto, en el ambiente de la Escuela de Chicago, carecía de interés y pasó desapercibido. Nadie lo animó a publicarlo y quedó en el olvido durante muchos años.

En segundo lugar, encontramos a los matemáticos Albert W. Tucker (Canadá, 1905-1995) y Harold W. Kuhn (California, 1925-2014), que eran, respectivamente, profesor y alumno en la Universidad de Princeton. Kuhn y Tucker dieron una conferencia en verano de 1950 en Berkeley (Simposio de Berkeley), California, donde enunciaron y demostraron su descubrimiento (lo que hoy en día conocemos por teorema de Karush-Kuhn-Tucker).

En esta conferencia aparece por primera vez en la historia el nombre *programación no lineal*. El objetivo de Kuhn y Tucker era generalizar la programación lineal, que ya había surgido años antes de la mano de Dantzig. A diferencia de Karush, Kuhn y Tucker no tuvieron ningún inconveniente, adquirieron fama casi instantánea en el mundo de las matemáticas y la gente empezó a referirse al resultado como el teorema de Kuhn-Tucker a secas.

Lo que Kuhn y Tucker no sabían es que su resultado no era para nada novedoso. Karush, once años antes, obtuvo prácticamente lo mismo, solo que él utilizaba otras herramientas y notación en la demostración. En 1975, cuando Kuhn y Tucker se enteraron de que el teorema ya había sido probado por Karush en 1939, le escribieron inmediatamente por carta para reconocer su trabajo y prioridad en este asunto.

Se trata, pues, de un descubrimiento múltiple y hoy en día la comunidad matemática se refiere al resultado como **teorema de Karush-Kuhn-Tucker**, apareciendo el apellido Karush en primer lugar ya que este lo demostró once años antes que Kuhn y Tucker; pero, ¿por qué el resultado de Karush pasó desapercibido y tan solo once años después el mismo resultado tuvo tanta fama y reconocimiento? Pues bien, la respuesta a esta pregunta, según la profesora e historiadora de matemáticas T. H. Kjeldsen, está en el contexto histórico y social en el que nació el teorema. El fin de la Segunda Guerra Mundial supuso también la igualdad entre matemática pura y matemática aplicada. Antes de la guerra, la matemática pura era la que gobernaba y dominaba el mundo de las matemáticas pero, durante la guerra, muchos matemáticos dedicados hasta entonces a la investigación en matemática aplicada fueron contratados por diversas organizaciones involucradas en la guerra, como, por ejemplo, el Ejército, para que diseñaran métodos de planificación de programas, una herramienta de las Fuerzas Armadas para llevar a cabo enormes planteamientos logísticos.

Es más, el propio G. Dantzig, contratado por las Fuerzas Armadas en 1941, fue el principal responsable de lo que hoy en día se conoce como programación lineal (que nació durante la guerra bajo el nombre de *programación en estructura lineal*) y del famoso y sencillo método que resuelve un programa lineal, a saber, el *método símplex* (ideado por el propio Dantzig). Más información sobre este método puede verse en el artículo de Peressini, Sullivan y Uhl [12] si el lector está interesado.

Tras la guerra, Dantzig utilizó sus estudios acerca de la programación lineal para resolver el problema del puente aéreo de Berlín: a mediados de 1948, en plena Guerra Fría, la URSS bloqueó las comunicaciones terrestres entre las zonas occidentales alemanas ocupadas por los Aliados y la ciudad de Berlín, y Dantzig, utilizando la programación lineal, diseñó un plan de abastecimiento aéreo minimizando los costes que, en pocos meses, consiguió igualar a los suministros realizados por carretera y ferrocarril antes del bloqueo.

Los países se percataron de que para conseguir ser una potencia mundial debían ser fuertes no solo en el ejército, sino también en ciencia e investigación, y fue ahí donde aumentó la importancia de la matemática aplicada y, en especial, de la programación lineal. Tiene sentido, entonces, que, en ese ambiente, la conferencia sobre programación no lineal de Kuhn y Tucker en 1950 fuera acogida de manera excelente y que, sin embargo, al trabajo de Karush (anterior a la Segunda Guerra Mundial) no se le diera la importancia correspondiente. Kjeldsen [6] y Kuhn [7] ofrecen información sobre todo esto y más. Además, Prékopa [13] permite al lector encontrar la relación de estos problemas de optimización con principios de la física, así como la demostración del *lema de Farkas*, una de las principales herramientas que permitió a Karush demostrar su resultado.

Volviendo ya al contenido matemático, se definen a continuación lo que se entiende por un programa mixto, un punto factible y un punto regular, con lo que estaremos en disposición de enunciar y demostrar el teorema de Karush-Kuhn-Tucker.

Definición 4 (programa mixto). Dados $n, p, q \in \mathbb{N}$, Ω un subconjunto abierto y no vacío de $\mathbb{R}^n$ y $1 + p + q$ funciones reales $f, g_1, \dots, g_p, h_1, \dots, h_q$ de clase $\mathcal{C}^1$ definidas en Ω, un **programa mixto** es un problema de optimización condicionada de la siguiente forma:

$$\text{(PM}^-\text{)} \quad \begin{cases} \text{minimizar } f(\boldsymbol{x}) \text{ sujeto a} \\ \quad g_1(\boldsymbol{x}) = 0, \dots, g_p(\boldsymbol{x}) = 0, \\ \quad h_1(\boldsymbol{x}) \le 0, \dots, h_q(\boldsymbol{x}) \le 0, \\ \quad \boldsymbol{x} \in \Omega; \end{cases} \qquad \text{(PM}^+\text{)} \quad \begin{cases} \text{maximizar } f(\boldsymbol{x}) \text{ sujeto a} \\ \quad g_1(\boldsymbol{x}) = 0, \dots, g_p(\boldsymbol{x}) = 0, \\ \quad h_1(\boldsymbol{x}) \le 0, \dots, h_q(\boldsymbol{x}) \le 0, \\ \quad \boldsymbol{x} \in \Omega. \end{cases}$$

En lo que sigue, escribiremos (PM) para referirnos indistintamente a (PM^-) o (PM^+). ◂

La función f se denomina *función objetivo* del problema (PM) y las igualdades $\{g_i(\boldsymbol{x}) = 0 : i = 1, \dots, p\}$ y desigualdades $\{h_j(\boldsymbol{x}) \le 0 : j = 1, \dots, q\}$ se denominan *restricciones* de igualdad y desigualdad de (PM), respectivamente.

Definición 5 (punto factible). Un **punto factible** para el problema (PM) es un punto $\boldsymbol{x}^* \in \Omega$ tal que $g_i(\boldsymbol{x}^*) = 0$ y $h_j(\boldsymbol{x}^*) \le 0$ para todo $i = 1, \dots, p$ y $j = 1, \dots, q$. ◂

Definición 6 (punto regular). Un **punto regular** para el problema (PM) es un punto factible $\boldsymbol{x}^* \in \Omega$ de (PM) tal que el conjunto de vectores siguiente es linealmente independiente:

$$\{\nabla g_i(\boldsymbol{x}^*), \nabla h_j(\boldsymbol{x}^*) \in \mathbb{R}^n : i \in \{1, \dots, p\}, j \in J(\boldsymbol{x}^*)\},$$

donde $J(\boldsymbol{x}^*) = \{k \in \{1, \dots, q\} : h_k(\boldsymbol{x}^*) = 0\}$. ◂

***Nota* 7.** Nótese que, bajo las hipótesis del teorema de Lagrange, el punto $\boldsymbol{x}^*$ es regular. ◂

La siguiente versión del teorema de Karush-Kuhn-Tucker y la demostración aquí expuestas no son más que una traducción del artículo de McShane [10].

Teorema 8 (Karush-Kuhn-Tucker, 1939 y 1950). *En la situación de la definición 4, si f tiene un mínimo (resp. máximo) condicionado por $g_i(\boldsymbol{x}) = 0$ para $i = 1, \dots, p$ y por $h_j(\boldsymbol{x}) \le 0$ para $j = 1, \dots, q$, entonces existen números reales $\lambda_0, \lambda_1, \dots, \lambda_p, \mu_1, \dots, \mu_q \in \mathbb{R}$, no todos nulos, tales que*

$$\lambda_0 \nabla f(\boldsymbol{x}) + \sum_{i=1}^{p} \lambda_i \nabla g_i(\boldsymbol{x}) + \sum_{j=1}^{q} \mu_j \nabla h_j(\boldsymbol{x}) = \mathbf{0}. \tag{2}$$

Además,

- **(I)** $\lambda_0 \ge 0$ *y* $\mu_j \ge 0$ *(resp.* $\mu_j \le 0$*) para todo* $j = 1, \dots, q$.
- **(II)** $\mu_j h_j(\boldsymbol{x}^*) = 0$ *para* $j = 1, \dots, q$.
- **(III)** *Si* $\boldsymbol{x}^*$ *es un punto regular, entonces se puede tomar* $\lambda_0 = 1$.

Demostración. A continuación, se demuestra el teorema para el caso de mínimo condicionado. El caso de máximo condicionado se deduce de inmediato de este ya que basta aplicarlo a la función $-f$.

Sin pérdida de generalidad asumimos que $\boldsymbol{x}^* = \mathbf{0}$, $f(\boldsymbol{x}^*) = 0$ (esto siempre es posible tras considerar una traslación adecuada) y que, para algún $z \in \mathbb{N}$ con $0 \leq z \leq q$, se tiene $h_1(\boldsymbol{x}^*) = 0, \dots, h_z(\boldsymbol{x}^*) = 0$ y $h_{z+1}(\boldsymbol{x}^*) < 0, \dots, h_q(\boldsymbol{x}^*) < 0$ (basta ordenar las funciones $h_1, \dots, h_q$ de forma que las z primeras se anulen en $\boldsymbol{x}^*$ y las restantes sean menores que cero en $\boldsymbol{x}^*$). Nótese que el caso $z = 0$ corresponde a que en todas las desigualdades se dé la desigualdad estricta, $h_j(\boldsymbol{x}^*) < 0$ para todo $j = 1, \dots, q$, mientras que el caso $z = q$ corresponde a que en todas se dé la igualdad, $h_j(\boldsymbol{x}^*) = 0$ para todo $j = 1, \dots, q$.

Por ser $\boldsymbol{x}^* = \mathbf{0}$ un punto interior de Ω, existe $\varepsilon > 0$ tal que la bola abierta $B(\mathbf{0}, \varepsilon)$ de centro el origen y radio ε está contenida en Ω, y es inmediato que la bola cerrada $\overline{B}(\mathbf{0}, \varepsilon_1)$ de centro el origen y radio $\varepsilon_1 = \varepsilon/2$ está contenida en Ω. Por el teorema de conservación del signo, existe $\varepsilon_2 > 0$ tal que las restricciones h_j con $j = z+1, \dots, q$ son negativas en $\overline{B}(\mathbf{0}, \varepsilon_2)$. Sea $\varepsilon_0 = \min\{\varepsilon_1, \varepsilon_2\}$; entonces, la bola cerrada $\overline{B}(\mathbf{0}, \varepsilon_0)$ está contenida en Ω y las restricciones h_j para $j = z+1, \dots, q$ son negativas en $\overline{B}(\mathbf{0}, \varepsilon_0)$.

Lema 9. *Para cada $\varepsilon > 0$ con $\varepsilon \leq \varepsilon_0$ existe $N \in \mathbb{N}$ tal que*

$$f(\boldsymbol{x}) + \|\boldsymbol{x}\|^2 + N\left(\sum_{i=1}^{p} g_i(\boldsymbol{x})^2 + \sum_{j=1}^{z} h_j^+(\boldsymbol{x})^2\right) > 0 \qquad \forall\, \boldsymbol{x} \in \mathbb{S}(\varepsilon), \tag{3}$$

donde $h_j^+(\boldsymbol{x}) = \max\{h_j(\boldsymbol{x}), 0\}$ para cada $j = 1, \dots, z$ y $\mathbb{S}(\varepsilon) = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| = \varepsilon\}$.

Demostración. Razonemos por reducción al absurdo y supongamos que el enunciado es falso:

> Existe $\tilde{\varepsilon} \in (0, \varepsilon_0]$ de forma que a cada $N \in \mathbb{N}$ le corresponde un punto $\boldsymbol{x}_N \in \mathbb{S}(\tilde{\varepsilon})$ con $f(\boldsymbol{x}_N) + \|\boldsymbol{x}_N\|^2 + N\left(\sum_{i=1}^{p} g_i(\boldsymbol{x}_N)^2 + \sum_{j=1}^{z} h_j^+(\boldsymbol{x}_N)^2\right) \leq 0$.

Tomamos una sucesión de números naturales $\{N_m\}_{m\in\mathbb{N}}$ creciente y tendiendo a infinito y la sucesión de los correspondientes puntos $\{\boldsymbol{x}_m\}_{m\in\mathbb{N}}$ en $\mathbb{S}(\tilde{\varepsilon})$ de manera que

$$f(\boldsymbol{x}_m) + \|\boldsymbol{x}_m\|^2 + N_m\left(\sum_{i=1}^{p} g_i(\boldsymbol{x}_m)^2 + \sum_{j=1}^{z} h_j^+(\boldsymbol{x}_m)^2\right) \leq 0 \qquad \forall\, m \in \mathbb{N}. \tag{4}$$

Como $\{\boldsymbol{x}_m\}_{m\in\mathbb{N}}$ es una sucesión acotada de vectores de $\mathbb{R}^n$, el teorema de Bolzano-Weierstrass garantiza la existencia de una sucesión parcial de $\{\boldsymbol{x}_m\}_{m\in\mathbb{N}}$ convergente a un punto $\boldsymbol{x}^0 \in \mathbb{R}^n$. Supongamos, sin pérdida de generalidad, que dicha sucesión parcial es desde un principio la sucesión original $\{\boldsymbol{x}_m\}_{m\in\mathbb{N}}$ y, en virtud de la continuidad de la función objetivo f y de la función norma en $\mathbb{R}^n$, se tiene que

$$\lim f(\boldsymbol{x}_m) = f(\boldsymbol{x}^0) \text{ y}$$

$$\|\boldsymbol{x}^0\| = \|\lim \boldsymbol{x}_m\| = \lim \|\boldsymbol{x}_m\| = \lim \tilde{\varepsilon} = \tilde{\varepsilon}.$$

Dividiendo ahora ambos miembros de la desigualdad (4) por N_m obtenemos que

$$\frac{f(\boldsymbol{x}_m)}{N_m} + \frac{\|\boldsymbol{x}_m\|^2}{N_m} + \left(\sum_{i=1}^{p} g_i(\boldsymbol{x}_m)^2 + \sum_{j=1}^{z} h_j^+(\boldsymbol{x}_m)^2\right) \leq 0 \qquad \forall\, m \in \mathbb{N}.$$

Tomando límite cuando m tiende a infinito en la expresión anterior y usando que $\lim N_m = \infty$, $\lim f(\boldsymbol{x}_m) = f(\boldsymbol{x}^0)$ y $\lim \|\boldsymbol{x}_m\|^2 = \tilde{\varepsilon}^2$, lo que implica que $\lim \frac{f(\boldsymbol{x}_m)}{N_m} = \lim \frac{\|\boldsymbol{x}_m\|^2}{N_m} = 0$, se sigue que

$$\sum_{i=1}^{p} g_i(\boldsymbol{x}^0)^2 + \sum_{j=1}^{z} h_j^+(\boldsymbol{x}^0)^2 \leq 0,$$

de donde se deduce que $\boldsymbol{x}^0$ satisface que $g_i(\boldsymbol{x}^0) = 0$ para $i = 1, \dots, p$ y $h_j^+(\boldsymbol{x}^0) = 0$ para $j = 1, \dots, z$. Así que $\lim f(\boldsymbol{x}_m) = f(\boldsymbol{x}^0) \geq f(\boldsymbol{x}^* = \mathbf{0}) = 0$, pero por (4) se desprende que $f(\boldsymbol{x}_m) \leq -\tilde{\varepsilon}^2 < 0$ para todo $m \in \mathbb{N}$, luego $\lim f(\boldsymbol{x}_m) = f(\boldsymbol{x}^0) < 0$ y se llega a una contradicción. ∎

Lema 10. *Para cada $\varepsilon > 0$ con $\varepsilon \leq \varepsilon_0$ existen un punto $\overline{\boldsymbol{x}} = (\overline{x}_1, \dots, \overline{x}_n) \in \mathbb{R}^n$ y un vector unitario $(\lambda_0, \lambda_1, \dots, \lambda_p, \mu_1, \dots, \mu_z)$ con componentes $\lambda_0, \mu_1, \dots, \mu_z$ no negativas tales que $\|\overline{\boldsymbol{x}}\| < \varepsilon$ y*

$$\lambda_0 \left[D_k f(\overline{\boldsymbol{x}}) + 2\overline{x}_k \right] + \sum_{i=1}^{p} \lambda_i D_k g_i(\overline{\boldsymbol{x}}) + \sum_{j=1}^{z} \mu_j D_k h_j(\overline{\boldsymbol{x}}) = 0 \qquad \forall\, k = 1, \dots, n. \tag{5}$$

Demostración. Sea $\tilde{\varepsilon} \in (0, \varepsilon_0]$ fijo pero arbitrario y $N \in \mathbb{N}$ el número natural dado por el lema 9. Considérese la función $F\colon \overline{B}(\mathbf{0}, \tilde{\varepsilon}) \subset \Omega \to \mathbb{R}$ definida por

$$F(\boldsymbol{x}) = f(\boldsymbol{x}) + \|\boldsymbol{x}\|^2 + N\left(\sum_{i=1}^{p} g_i(\boldsymbol{x})^2 + \sum_{j=1}^{z} h_j^+(\boldsymbol{x})^2 \right) \qquad \forall\, \boldsymbol{x} \in \overline{B}(\mathbf{0}, \tilde{\varepsilon}).$$

Como F es continua y $\overline{B}(\mathbf{0}, \tilde{\varepsilon})$, compacto, el teorema de Weierstrass asegura la existencia de un punto $\overline{\boldsymbol{x}} \in \overline{B}(\mathbf{0}, \tilde{\varepsilon})$ donde F alcanza su mínimo global; en particular, $F(\overline{\boldsymbol{x}}) \leq F(\boldsymbol{x}^* = \mathbf{0}) = 0$, y el lema 9 afirma que $\|\overline{\boldsymbol{x}}\| < \tilde{\varepsilon}$ ($\overline{\boldsymbol{x}}$ es un punto interior de $B(\mathbf{0}, \tilde{\varepsilon})$). Así, todas las derivadas parciales de primer orden de F deben anularse en $\overline{\boldsymbol{x}}$:

$$D_k f(\overline{\boldsymbol{x}}) + 2\overline{x}_k + 2N\left(\sum_{i=1}^{p} g_i(\overline{\boldsymbol{x}}) D_k g_i(\overline{\boldsymbol{x}}) + \sum_{j=1}^{z} h_j^+(\overline{\boldsymbol{x}}) D_k h_j(\overline{\boldsymbol{x}}) \right) = 0 \qquad \forall\, k = 1, \dots, n, \tag{6}$$

donde se ha usado que la función $(h_j^+)^2$ $(j = 1, \dots, z)$ es diferenciable en $B(\mathbf{0}, \tilde{\varepsilon})$ con

$$D_k (h_j^+)^2(\boldsymbol{x}) = 2h_j^+(\boldsymbol{x}) D_k h_j(\boldsymbol{x}) \qquad \forall\, \boldsymbol{x} \in B(\mathbf{0}, \tilde{\varepsilon})$$

para cualquier $k = 1, \dots, n$, lo cual se demuestra en el libro de Peressini, Sullivan y Uhl [12, capítulo 6] teniendo en cuenta que

$$h_j^+(\boldsymbol{x}) = \frac{h_j(\boldsymbol{x}) + |h_j(\boldsymbol{x})|}{2} \qquad \forall\, \boldsymbol{x} \in B(\mathbf{0}, \tilde{\varepsilon}).$$

Tomando $\tau = \left[1 + 4N^2 \left(\sum_{i=1}^{p} g_i(\overline{\boldsymbol{x}})^2 + \sum_{j=1}^{z} h_j^+(\overline{\boldsymbol{x}})^2 \right) \right]^{1/2} > 0$ y definiendo

$$\lambda_0 = \frac{1}{\tau}, \qquad \lambda_i = 2N\frac{g_i(\overline{\boldsymbol{x}})}{\tau} \quad (i = 1, \dots, p), \qquad \mu_j = \begin{cases} 2N\dfrac{h_j^+(\overline{\boldsymbol{x}})}{\tau} & j = 1, \dots, z; \\ 0 & j = z+1, \dots, q, \end{cases}$$

es fácil comprobar que el vector $(\lambda_0, \lambda_1, \dots, \lambda_p, \mu_1, \dots, \mu_z)$ es unitario, λ_0 y μ_j $(j = 1, \dots, z)$ son no negativos y dividiendo ambos miembros de la condición (6) por τ se consigue la igualdad (5). ∎

Finalmente, tomamos una sucesión decreciente de números reales y positivos $\{\delta_m\}_{m \in \mathbb{N}}$ tendiendo a cero con $\delta_1 < \varepsilon_0$. Para cada $m \in \mathbb{N}$, elegimos un punto $\overline{\boldsymbol{x}}_m \in \mathbb{R}^n$ con $\|\overline{\boldsymbol{x}}_m\| < \delta_m$ y un vector unitario $(\lambda_{0,m}, \lambda_{1,m}, \dots, \lambda_{p,m}, \mu_{1,m}, \dots, \mu_{z,m}, 0, \dots, 0)$ con componentes $\lambda_{0,m}$ y $\mu_{j,m}$ $(j = 1, \dots, z)$ no negativas tales que se cumple la igualdad (5) (esto es posible por el lema 10). De nuevo, por el teorema de Bolzano-Weierstrass, existe una sucesión parcial para la cual los vectores unitarios convergen a un límite $(\lambda_0, \lambda_1, \dots, \lambda_p, \mu_1, \dots, \mu_q)$.

Como $\{\overline{\boldsymbol{x}}_m\}_{m \in \mathbb{N}} \to \boldsymbol{x}^* = \mathbf{0}$, la ecuación (5) se cumple para este vector límite y esto demuestra el teorema salvo el apartado (iii), que se comenta a continuación: si $\boldsymbol{x}^*$ es un punto regular, no puede ser que $\lambda_0 = 0$, pues entonces la condición (2) contradice la independencia lineal de los vectores $\nabla g_1(\boldsymbol{x}^0), \dots, \nabla g_p(\boldsymbol{x}^*)$, $\nabla h_1(\boldsymbol{x}^*), \dots, \nabla h_z(\boldsymbol{x}^*)$, y esto concluye la demostración del teorema. ∎

Esta demostración se le debe a Edward J. McShane (1904-1989) y fue publicada en un artículo para *The American Mathematical Monthly* en el año 1973 [10], y es realmente asombrosa: únicamente aplica teoremas elementales del análisis matemático como son el teorema de Weierstrass, el teorema de Bolzano-Weierstrass y la condición necesaria de existencia de extremo en un punto interior.

Nótese que si $q = 0$ y $p > 0$, entonces el problema (PM) incluye solo restricciones de igualdad y el clásico método de los multiplicadores de Lagrange aporta una condición necesaria de existencia de solución del programa, y el caso extremo $p = q = 0$ da lugar a un problema de optimización libre.

En honor a Karush, Kuhn y Tucker, la ecuación (2) junto con los apartados (I) y (II) del teorema 8 reciben el nombre de **condiciones de Karush-Kuhn-Tucker**, los escalares $\lambda_1, \dots, \lambda_p$ (asociados a las restricciones de igualdad) se denominan **multiplicadores de Lagrange** y los escalares $\mu_1, \dots, \mu_q$ (asociados a las restricciones de desigualdad) se denominan **multiplicadores de Karush-Kuhn-Tucker**. Al igual que el teorema de Lagrange, es conveniente advertir que el teorema de Karush-Kuhn-Tucker solo aporta una condición necesaria, y no suficiente, de existencia de solución del programa (PM).

Obsérvese que no se impone ninguna hipótesis sobre el número de variables n y el número de restricciones p y q, a diferencia del teorema de Lagrange, donde se exigía que el número de restricciones de igualdad m fuese menor que el número de variables n. En el caso $q = 0$ (solamente restricciones de igualdad están presentes), las condiciones (I) y (II) del ecuación (2) carecen de sentido.

En la práctica, si se sabe de antemano que existe un punto $\boldsymbol{x}^* \in \Omega$ que es solución de (PM), entonces $\boldsymbol{x}^*$ cumple las siguientes condiciones:

Condición estacionaria: $\lambda_0 \nabla f(\boldsymbol{x}) + \sum_{i=1}^{p} \lambda_i \nabla g_i(\boldsymbol{x}) + \sum_{j=1}^{q} \mu_j \nabla h_j(\boldsymbol{x}) = \boldsymbol{0}$.

Condición de factibilidad: $g_i(\boldsymbol{x}^*) = 0$ para $i = 1, \dots, p$ y $h_j(\boldsymbol{x}^*) \le 0$ para $j = 1, \dots, q$.

Condición de holgura: $\mu_j h_j(\boldsymbol{x}^*) = 0$ para $j = 1, \dots, q$.

Condición de signo: $\lambda_0 \ge 0$ y $\mu_j \ge 0$ para todo $j = 1, \dots, q$ (para mínimo condicionado) o $\mu_j \le 0$ para todo $j = 1, \dots, q$ (para máximo condicionado).

A continuación, se expone un primer ejemplo sencillo de aplicación del teorema de Karush-Kuhn-Tucker.

Ejemplo 11. Resolver el siguiente problema de optimización:

$$\begin{cases} \text{optimizar } f(x,y,z) = x + y + z \text{ sujeto a} \\ \quad h_1(x,y,z) = (y-1)^2 + z^2 \le 1, \\ \quad h_2(x,y,z) = x^2 + (y-1)^2 + z^2 \le 3. \end{cases}$$

Por el teorema de Weierstrass, existe solución (máximo y mínimo) del programa considerado y las condiciones de Karush-Kuhn-Tucker son las siguientes:

Condición estacionaria:

$$\begin{cases} 1 + 2\mu_2 x = 0, \\ 1 + 2\mu_1(y-1) + 2\mu_2(y-1) = 0, \\ 1 + 2\mu_1 z + 2\mu_2 z = 0. \end{cases}$$

Condición de factibilidad:

$$\begin{cases} (y-1)^2 + z^2 \le 1, \\ x^2 + (y-1)^2 + z^2 \le 3. \end{cases}$$

Condición de holgura:

$$\begin{cases} \mu_1 \left[(y-1)^2 + z^2 - 1\right] = 0, \\ \mu_2 \left[x^2 + (y-1)^2 + z^2 - 3\right] = 0. \end{cases}$$

Condición de signo:

$$\begin{cases} \mu_1, \mu_2 \ge 0 \longrightarrow \text{mínimo local}, \\ \mu_1, \mu_2 \le 0 \longrightarrow \text{máximo local}. \end{cases}$$

A partir de la condición de holgura se distinguen cuatro casos:

$$\begin{cases} \mu_1 = 0 & \Longrightarrow \begin{cases} \mu_2 = 0, & \text{(caso I)} \\ x^2 + (y-1)^2 + z^2 - 3 = 0; & \text{(caso II)} \end{cases} \\ (y-1)^2 + z^2 - 1 = 0 & \Longrightarrow \begin{cases} \mu_2 = 0, & \text{(caso III)} \\ x^2 + (y-1)^2 + z^2 - 3 = 0, & \text{(caso IV)} \end{cases} \end{cases}$$

pero la primera ecuación de la condición estacionaria obliga a que $\mu_2 \neq 0$, así que los solo se deben comprobar los casos II y IV.

Caso II: para el caso II, un sencillo cálculo más o menos breve proporciona los siguientes dos puntos de Karush-Kuhn-Tucker con sus correspondientes multiplicadores:

$$P_1 = (1, 2, 1), \qquad \boldsymbol{\mu} = \left(0, -\frac{1}{2}\right),$$

$$P_2 = (-1, 0, -1), \qquad \boldsymbol{\mu} = \left(0, \ \frac{1}{2}\right).$$

Caso IV: para el caso IV, un sencillo cálculo más o menos breve proporciona los siguientes cuatro puntos de Karush-Kuhn-Tucker con sus correspondientes multiplicadores:

$$
\begin{aligned}
P_3 &= \left(\sqrt{2}, 1 + \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right), & \mu &= \left(-\frac{1}{2\sqrt{2}}, -\frac{1}{2\sqrt{2}}\right), \\
P_4 &= \left(\sqrt{2}, 1 - \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right), & \mu &= \left(\frac{3}{2\sqrt{2}}, -\frac{1}{2\sqrt{2}}\right), \\
P_5 &= \left(-\sqrt{2}, 1 + \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right), & \mu &= \left(-\frac{3}{2\sqrt{2}}, \frac{1}{2\sqrt{2}}\right), \\
P_6 &= \left(-\sqrt{2}, 1 - \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right), & \mu &= \left(\frac{1}{2\sqrt{2}}, \frac{1}{2\sqrt{2}}\right).
\end{aligned}
$$

Finalmente, se aplican las dos condiciones aún no usadas (factibilidad y signo) y se resumen los resultados obtenidos en el cuadro 1. ◂

Cuadro 1: Resumen de los resultados del ejemplo 11.

P	μ	Factibilidad	Signo	Conclusión
P_1	$\left(0, -\frac{1}{2}\right)$	NO	-	-
P_2	$\left(0, \frac{1}{2}\right)$	NO	-	-
P_3	$\left(-\frac{1}{2\sqrt{2}}, -\frac{1}{2\sqrt{2}}\right)$	SÍ	Negativo	Máximo condicionado
P_4	$\left(\frac{3}{2\sqrt{2}}, -\frac{1}{2\sqrt{2}}\right)$	SÍ	NO	-
P_5	$\left(-\frac{3}{2\sqrt{2}}, \frac{1}{2\sqrt{2}}\right)$	SÍ	NO	-
P_6	$\left(\frac{1}{2\sqrt{2}}, \frac{1}{2\sqrt{2}}\right)$	SÍ	Positivo	Mínimo condicionado

4. Programación convexa y cóncava

En esta última sección se expone lo que se entiende por programa convexo y cóncavo y se prueba que la condición necesaria de existencia de solución para programas mixtos que aporta el teorema de Karush-Kuhn-Tucker es también condición suficiente bajo hipótesis de convexidad o concavidad. Antes de ello, recordemos el siguiente resultado que caracteriza a las funciones diferenciables que son convexas o cóncavas.

Proposición 12 (caracterización de funciones convexas y cóncavas). *Sean $n \in \mathbb{N}$, $\Omega \subset \mathbb{R}^n$ un subconjunto abierto, convexo y no vacío de $\mathbb{R}^n$ y $f\colon \Omega \to \mathbb{R}$ una función real y diferenciable en Ω. Entonces,*

- *f es convexa si y solo si $f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle \leq f(\boldsymbol{y})$ para todo $\boldsymbol{x}, \boldsymbol{y} \in \Omega$.*
- *f es estrictamente convexa si y solo si $f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle < f(\boldsymbol{y})$ para todo $\boldsymbol{x}, \boldsymbol{y} \in \Omega$ con $\boldsymbol{x} \neq \boldsymbol{y}$.*
- *f es cóncava si y solo si $f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle \geq f(\boldsymbol{y})$ para todo $\boldsymbol{x}, \boldsymbol{y} \in \Omega$.*
- *f es estrictamente cóncava si y solo si $f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle > f(\boldsymbol{y})$ para todo $\boldsymbol{x}, \boldsymbol{y} \in \Omega$ con $\boldsymbol{x} \neq \boldsymbol{y}$.*

Definición 13 (programa convexo). Dados $n, m \in \mathbb{N}$, $\Omega \subset \mathbb{R}^n$ un subconjunto abierto, convexo y no vacío de $\mathbb{R}^n$ y $f, h_1, \dots, h_m$ funciones reales, diferenciables y convexas en Ω, un **programa convexo** es un problema de minimización condicionada de la siguiente forma:

$$(\mathrm{PC}^-) \qquad \begin{cases} \text{minimizar } f(\boldsymbol{x}) \text{ sujeto a} \\ \quad h_1(\boldsymbol{x}) \leq 0, \dots, h_m(\boldsymbol{x}) \leq 0, \\ \quad \boldsymbol{x} \in \Omega. \end{cases}$$ ◂

Definición 14 (programa cóncavo). Dados $n, m \in \mathbb{N}$, $\Omega \subset \mathbb{R}^n$ un subconjunto abierto, convexo y no vacío de $\mathbb{R}^n$ y $f, h_1, \dots, h_m$ funciones reales, diferenciables y cóncavas en Ω, un **programa cóncavo** es un problema de maximización condicionada de la siguiente forma:

$$(\mathrm{PC}^+) \qquad \begin{cases} \text{maximizar } f(\boldsymbol{x}) \text{ sujeto a} \\ \quad h_1(\boldsymbol{x}) \leq 0, \dots, h_m(\boldsymbol{x}) \leq 0, \\ \quad \boldsymbol{x} \in \Omega. \end{cases}$$ ◂

Teorema 15 (condición suficiente para programas convexos). *Si $\boldsymbol{x}^* \in \Omega$ es un punto factible y regular para el programa convexo y cumple las condiciones de Karush-Kuhn-Tucker, entonces $\boldsymbol{x}^*$ es solución del programa convexo, esto es, $\boldsymbol{x}^*$ es un mínimo global de f condicionado a $h_k(\boldsymbol{x}) \leq 0$ para $k = 1, \dots, m$.*

Demostración. Como $\boldsymbol{x}^*$ es un punto regular por hipótesis, se puede tomar $\lambda_0 = 1$. Por hipótesis, existen escalares $\mu_1, \dots, \mu_m \in \mathbb{R}_0^+$ no todos nulos tales que $\nabla f(\boldsymbol{x}^*) + \sum_{k=1}^m \mu_k \nabla h_k(\boldsymbol{x}^*) = 0$. Como $\mu_k \geq 0$ para cada $k = 1, \dots, m$ y $h_k(\boldsymbol{x}) \leq 0$ para cada $\boldsymbol{x} \in \Omega$ y $k = 1, \dots, m$, se tiene que $\mu_k h_k(\boldsymbol{x}) \leq 0$ para cada $\boldsymbol{x} \in \Omega$ y $k = 1, \dots, m$, y esto, junto con la proposición 12, permite escribir lo siguiente para cualquier punto factible $\boldsymbol{x} \in \Omega$:

$$\begin{aligned} f(\boldsymbol{x}) &\geq f(\boldsymbol{x}) + \sum_{k=1}^m \mu_k h_k(\boldsymbol{x}) \\ &\geq (f(\boldsymbol{x}^*) + \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^* \rangle) + \sum_{k=1}^m \mu_k \left(h_k(\boldsymbol{x}^*) + \langle \nabla h_k(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^* \rangle \right) \\ &= f(\boldsymbol{x}^*) + \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^* \rangle + \sum_{k=1}^m \underbrace{\mu_k h_k(\boldsymbol{x}^*)}_{0} + \langle \sum_{k=1}^m \mu_k \nabla h_k(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^* \rangle \\ &= f(\boldsymbol{x}^*) + \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^* \rangle + \langle \sum_{k=1}^m \mu_k \nabla h_k(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^* \rangle \\ &= f(\boldsymbol{x}^*) + \langle \underbrace{\nabla f(\boldsymbol{x}^*) + \sum_{k=1}^m \mu_k \nabla h_k(\boldsymbol{x}^*)}_{\boldsymbol{0}}, \boldsymbol{x} - \boldsymbol{x}^* \rangle = f(\boldsymbol{x}^*), \end{aligned}$$

lo que significa que $\boldsymbol{x}^*$ es un mínimo global de f condicionado a $h_k(\boldsymbol{x}) \leq 0$ para cada $k = 1, \dots, m$, como se quería. ∎

Una demostración análoga a la recién expuesta probaría la condición suficiente para programas cóncavos. Animamos al lector a intentarlo.

Nota 16. En la situación del teorema 15, de la proposición 12 se deduce que, si la función objetivo f de (PC^-) (resp. (PC^+)) es estrictamente convexa (resp. estrictamente cóncava), entonces el mínimo global (resp. máximo global) de (PC^-) (resp. (PC^+)) es único. ◂

Como comentario final, hay que destacar el alcance que tuvo y tienen los resultados aquí expuestos, con mención especial a la programación convexa. La convexidad es una herramienta eficaz que aporta una condición suficiente a la hora de resolver ciertos problemas de optimización y, además, grandes ramas dentro de la programación, como la programación lineal (donde la función a optimizar es una función lineal), geométrica (donde la función a optimizar es un posinomio[2]) o cuadrática (donde la función a

[2] Un posinomio tiene la misma expresión que un polinomio en varias variables $x_1, \dots, x_n$, pero aquí las variables son solo positivas, los coeficientes son solo positivos y los exponentes son reales (positivos, negativos o cero).

optimizar es una función cuadrática), pueden atacarse aplicando el teorema 15. Todo esto puede verse detalladamente en el trabajo final de grado de Martínez Sánchez [9], así como la *programación convexa dual*, que a grandes rasgos, trata de asociar a cada programa convexo de minimización (PC^-) un programa de maximización libre (sin restricciones) llamado *programa convexo dual* y que, a menudo, es mas fácil de resolver que el propio (PC^-) y cuyas soluciones pueden usarse para generar soluciones de (PC^-). Véase el libro de Peressini, Sullivan y Uhl [12] para más información al respecto.

Referencias

[1] Apostol, Tom M. *Análisis matemático*. Trad. por Vélez Cantarell, Francisco. Barcelona: Reverté, 1960.

[2] Bliss, G. A. «Normality and abnormality in the calculus of variations». En: *Transactions of the American Mathematical Society* 43.3 (1938), págs. 365-376. issn: 0002-9947. https://doi.org/10.2307/1990066.

[3] Dantzig, George B. *Linear programming and extensions*. RAND Corporation, 1963, págs. xvi+625. https://doi.org/10.7249/R366.

[4] Kantorovich, L. V. «Mathematical methods of organizing and planning production». En: *Management Science. Journal of the Institute of Management Science. Application and Theory Series* 6 (1959/1960), págs. 366-422. issn: 0025-1909. https://doi.org/10.1287/mnsc.6.4.366.

[5] Karush, William. *Minima of functions of several variables with inequalities as side conditions*. Thesis (SM). The University of Chicago, 1939, pág. 25. url: http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqdiss&rft_dat=xri:pqdiss:TM15121.

[6] Kjeldsen, Tinne Hoff. «A contextualized historical analysis of the Kuhn-Tucker theorem in nonlinear programming: the impact of World War II». En: *Historia Mathematica* 27.4 (2000), págs. 331-361. issn: 0315-0860. https://doi.org/10.1006/hmat.2000.2289.

[7] Kuhn, Harold W. «Nonlinear programming: a historical view». En: *Nonlinear programming Proceedings of a Symposium in Applied Mathematics Held in New York City*. Vol. 9. SIAM-AMS Proceedings. American Mathematical Society, 1976, págs. 1-26.

[8] Lagrange, Joseph-Louis. *Mécanique analytique*. Edición revisada. Librairie Scientifique et Technique Albert Blanchard, 1965.

[9] Martínez Sánchez, Francisco Javier. *Una generalización del teorema de los multiplicadores de Lagrange: condiciones de Karush-Kuhn-Tucker en programación no lineal*. Trabajo de Fin de Grado. Universidad de Granada, 2018. url: https://www.ugr.es/~acanada/docencia/matematicas/TFG-definitivo-2julio2018.pdf.

[10] McShane, Edward J. «The Lagrange Multiplier Rule». En: *The American Mathematical Monthly* 80.8 (1973), págs. 922-925. https://doi.org/10.1080/00029890.1973.11993409.

[11] Neumann, John von y Morgenstern, Oskar. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, New Jersey, 1944, págs. xviii+625.

[12] Peressini, Anthony L.; Sullivan, Francis E., y Uhl, J. J., Jr. *The mathematics of nonlinear programming*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 1988, págs. x+273. isbn: 978-0-387-96614-4.

[13] Prékopa, András. «On the development of optimization theory». En: *The American Mathematical Monthly* 87.7 (1980), págs. 527-542. issn: 0002-9890. https://doi.org/10.2307/2321417.

[14] Sydsaeter, Knut y Hammond, Peter. *Matemáticas para el análisis económico*. Prentice Hall, 1996. isbn: 978-0-13-240615-4.

[15] Wu, Hui-Hua y Wu, Shanhe. «Various proofs of the Cauchy-Schwarz inequality». En: *Octogon Mathematical Magazine* 17.1 (2009), págs. 221-229. url: http://www.uni-miskolc.hu/~matsefi/Octogon/volumes/volume1/article1_19.pdf.

TEMat

Códigos de Reed-Muller: las matemáticas detrás de las primeras fotografías del planeta rojo

Andoni De Arriba De La Hera[a]
Instituto de Ciencias Matemáticas (ICMAT)
andoni.dearriba@icmat.es

Resumen: Este artículo tiene como objetivo presentar y estudiar los llamados códigos de Reed-Muller binarios. Estos son un tipo muy especial de códigos que, además, han jugado un papel fundamental en nuestra historia, puesto que fueron los responsables de que se obtuvieran las primeras fotografías en blanco y negro de la superficie marciana.

El artículo comienza con una breve introducción que tiene como objetivo situar este en contexto, así como fijar algunas de las notaciones básicas. Después, se hace un rápido repaso al mundo de los códigos desde un punto de vista matemático, estudiando todas las nociones básicas necesarias para la correcta comprensión del artículo. Con esto se pretende que cualquier lector mínimamente familiarizado con las matemáticas pueda disfrutar de la lectura. Para terminar, a modo de aplicación práctica, aparecen enlaces a programas diseñados en Mathematica que permiten interactuar con la familia de códigos estudiada.

Abstract: This paper aims to present and study the so-called binary Reed-Muller codes. These are a very special type of codes that have played a fundamental role in our history, since they were responsible for obtaining the first black and white photographies of the Martian surface.

The paper begins with a brief introduction that aims to place the work in context, as well as to fix some of the basic notations. Later on, we quickly review the world of codes from a mathematical point of view, studying all the basic notions that will be necessary for the correct understanding of the paper. With this, we hope that any reader minimally familiarized with mathematics will be able to enjoy reading the paper. To finish, as a practical application, we include links to algorithms designed in Mathematica that allow us to work with the studied code family.

Palabras clave: transmisión de información, códigos detectores y correctores de errores, códigos lineales, alfabeto, letras, palabras, codificar, decodificar.

MSC2010: 94B05.

Recibido: 10 de septiembre de 2018.
Aceptado: 24 de febrero de 2019.

Agradecimientos: Quiero agradecer a la ANEM la oportunidad que ofrece a los jóvenes investigadores con la creación de esta revista, y, en especial, a los editores por su dedicación a la misma y, más concretamente, por su insistencia en que escribiera este artículo. Quisiera también dar las gracias a los revisores encargados para este por el arduo trabajo llevado a cabo en la revisión. Finalmente, no puedo dejar sin mencionar a quien fue mi directora de TFG, M.ª Asunción García Sánchez, por toda la ayuda que me brindó para la correcta realización del mismo, ya que este artículo está basado en dicho trabajo.

Referencia: De Arriba De La Hera, Andoni. «Códigos de Reed-Muller: las matemáticas detrás de las primeras fotografías del planeta rojo». En: *TEMat*, 3 (2019), págs. 45-61. ISSN: 2530-9633. URL: https://temat.es/articulo/2019-p45.

[a] El autor estaba afiliado a la Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU) durante el desarrollo del trabajo del que parte este artículo.

1. Introducción

Este artículo tiene como objetivo estudiar los *códigos de Reed-Muller*. Esta es una de las familias más antiguas y mejor conocidas entre los *códigos lineales*. En concreto, son un tipo muy especial de códigos *detectores y correctores de errores*, con ricas propiedades algebraicas, que se utilizan habitualmente en la *transmisión de información*. Los estudios que se van a tratar en el artículo se situan en una de las aplicaciones más actuales del álgebra: la *teoría de la información*, cuyas bases fueron establecidas por Claude Elwood Shannon[1] [5], quien, a día de hoy, es considerado el padre de toda esta teoría. Hoy día, la teoría de la información es la rama de las matemáticas y la computación que se ocupa del estudio de la información y de todo lo relacionado con ella.

Supongamos que un emisor desea enviar un mensaje $\boldsymbol{x}$ a través de un canal para que lo reciba un receptor (proceso que se conoce por *transmisión de información*). A lo largo de este proceso, el mensaje $\boldsymbol{x}$ suele verse alterado debido al «ruido» del canal, de manera que el mensaje recibido por el receptor pasa a ser $\boldsymbol{x}'$, donde, en general, se tiene que $\boldsymbol{x}' \neq \boldsymbol{x}$. Aquí es donde entran los llamados códigos detectores y correctores de errores. La idea consiste en que, antes de enviar el mensaje $\boldsymbol{x}$, el emisor lo *codifica* como $\boldsymbol{c}$, añadiéndole información redundante. De esta manera, si en el canal se produce un *error* $\boldsymbol{e}$ debido al cual se recibe el mensaje alterado $\boldsymbol{c}' = \boldsymbol{c} + \boldsymbol{e}$, tras *decodificar* este, el receptor debería ser capaz de recuperar $\boldsymbol{c}$ y, de ahí, deducir $\boldsymbol{x}$. El objetivo que se busca con todo esto es el de lograr que este proceso tenga éxito de la manera lo más eficiente posible (tanto en tiempo como en memoria).

Salvo que se diga lo contrario, nuestros mensajes serán vectores (que llamaremos **palabras**) del $\mathbb{F}_q$-espacio vectorial finito $\mathbb{F}_q^n$, siendo $q = p^t$ con p primo ($\mathbb{F}_q$ es lo que llamaremos **alfabeto**, mientras que a sus elementos los denominaremos **letras**). Por simplicidad, se denota a las palabras de nuestro alfabeto por

$$\boldsymbol{x} = (x_1, x_2, \dots, x_n) \underset{\text{not.}}{\equiv} x_1 x_2 \dots x_n.$$

De esta manera, podemos definir matemáticamente un **código** como un subconjunto no vacío de palabras (a las cuales nos referiremos de manera natural por **palabras código**) de un mismo alfabeto.

La idea es hacer un estudio completo de los *códigos de Reed-Muller* en el *caso binario*. Concretamente, hablar de las tres construcciones conocidas para los mismos, así como del método de decodificación propio por el cual son tan interesantes. En primer lugar, se incluye un rápido repaso en teoría de códigos lineales, haciendo especial hincapié en aquello que hará falta para la correcta comprensión del artículo. Hecho esto, entramos en materia con el objeto de estudio y, una vez terminado, aparecen algunos de los programas diseñados en Mathematica para el Trabajo de Fin de Grado de De Arriba De La Hera [1]. Salvo que se diga lo contrario, todas las demostraciones de los resultados que aquí se utilizan pueden encontrarse en el mismo. Además, se da por hecho que el lector está familiarizado con los conceptos y resultados básicos del álgebra y la geometría, sobre todo en los casos finitos. También conviene tener un mínimo de conocimiento de combinatoria, dado que a lo largo del artículo aparecen en repetidas ocasiones resultados en los que es necesario hacer uso de coeficientes binomiales, así como algunas de las relaciones más importantes que se conocen entre los mismos.

2. Repaso a la teoría de códigos lineales

A primera vista, parece muy complicado trabajar con la definición de código dada desde un punto de vista matemático. Es por esta razón que resulta natural restringirse a familias de códigos más manejables y fáciles de implementar. En concreto, nos centramos en los llamados códigos lineales.

2.1. Nociones básicas

Definición 1. Dados $s \leq n$ números naturales, llamamos **código lineal** de longitud n y dimensión s sobre $\mathbb{F}_q$ a un $\mathbb{F}_q$-subespacio vectorial de $\mathbb{F}_q^n$ de dimensión s. Nos referimos a estos por códigos lineales q-arios de longitud n y dimensión s. ◂

[1] Matemático, ingeniero eléctrico y criptógrafo americano; 30 abril, 1916 - 24 febrero, 2001.

Sea $\mathcal{C}$ un código lineal q-ario de longitud n y dimensión s. La principal ventaja que tiene el uso de códigos lineales es que, por tratarse de subespacios vectoriales, admiten bases de la forma $\mathcal{B} = \{\boldsymbol{c}_1, \dots, \boldsymbol{c}_s\}$, de modo que toda palabra código de $\mathcal{C}$ puede expresarse de manera única como combinación lineal de estas palabras básicas. Así, escribiendo $\boldsymbol{c}_i = c_{i1} \dots c_{in}$ para todo $i \in \{1, \dots, s\}$, podemos construir la conocida como **matriz generadora** del código lineal, la cual se corresponde con

$$G = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & & \vdots \\ c_{s1} & c_{s2} & \cdots & c_{sn} \end{pmatrix} \in \mathrm{Mat}_{s \times n}(\mathbb{F}_q).$$

Es obvio que para toda palabra código $\boldsymbol{c} \in \mathcal{C}$ que tomemos existen únicos escalares $\alpha_1, \dots, \alpha_s \in \mathbb{F}_q$ tales que $\boldsymbol{c} = (\alpha_1 \ \cdots \ \alpha_s)G$. Por tanto, para dar un código lineal, basta dar una matriz generadora del mismo.

Se define por **distancia de Hamming** entre dos palabras $\boldsymbol{x}$ e $\boldsymbol{y}$ de igual longitud al entero no negativo

(1) $$d(\boldsymbol{x}, \boldsymbol{y}) \coloneqq |\{i \in \{1, \dots, n\} \mid x_i \neq y_i\}| .$$

Por otra parte, se define como **peso** de una palabra $\boldsymbol{x}$ al entero no negativo

(2) $$\omega(\boldsymbol{x}) \coloneqq |\{i \in \{1, \dots, n\} \mid x_i \neq 0\}| .$$

Dado un código $\mathcal{C}$ arbitrario (no necesariamente lineal), a partir de (1) y (2) podemos introducir

$$d \underset{\text{not.}}{\equiv} d(\mathcal{C}) \coloneqq \text{mín}\{d(\boldsymbol{c}, \boldsymbol{c}') \mid \boldsymbol{c}, \boldsymbol{c}' \in \mathcal{C}, \boldsymbol{c} \neq \boldsymbol{c}'\}$$

y

$$\omega \underset{\text{not.}}{\equiv} \omega(\mathcal{C}) \coloneqq \text{mín}\{\omega(\boldsymbol{c}) \mid \boldsymbol{c} \in \mathcal{C}, \boldsymbol{c} \neq \boldsymbol{0}\},$$

donde d se conoce como **distancia mínima** del código, mientras que ω es el llamado **peso mínimo** del mismo. Si $\mathcal{C}$ es un código lineal, se comprueba fácilmente que $d = \omega$.

Un código lineal $\mathcal{C}$ *detecta hasta t errores* si, recibida $\boldsymbol{y} = \boldsymbol{c} + \boldsymbol{e}$ (siendo $\boldsymbol{c}$ la palabra código enviada y $\boldsymbol{e}$ el error dado en la transmisión, el cual se representa también como una palabra), con $0 < \omega(\boldsymbol{e}) \leq t$, entonces podemos asegurar que $\boldsymbol{y} \notin \mathcal{C}$. A su vez, un código lineal $\mathcal{C}$ *corrige hasta t errores* si, recibida $\boldsymbol{y}$, existe a lo más una palabra código $\boldsymbol{c} \in \mathcal{C}$ satisfaciendo que $d(\boldsymbol{y}, \boldsymbol{c}) \leq t$. No resulta difícil percatarse de que cualquier código $\mathcal{C}$ detecta hasta $d-1$ errores, mientras que corrige hasta $\left\lfloor \frac{d-1}{2} \right\rfloor$.

Finalmente, se define como **código dual** de un código lineal $\mathcal{C} \subseteq \mathbb{F}_q^n$ de dimensión s al conjunto

$$\mathcal{C}^\perp \coloneqq \left\{\boldsymbol{x} \in \mathbb{F}_q^n \mid \langle \boldsymbol{x}, \boldsymbol{c} \rangle = 0 \ \forall\, \boldsymbol{c} \in \mathcal{C}\right\},$$

donde $\langle \cdot, \cdot \rangle$ denota el producto escalar estándar en $\mathbb{F}_q^n$. Este vuelve a ser un código lineal q-ario de longitud n, pero de dimensión $n-s$ en este caso. Por tanto, se cumple que $\dim(\mathcal{C}) + \dim(\mathcal{C}^\perp) = n$. Más aún, este hecho nos permite afirmar que $\mathcal{C}^\perp$ admite una matriz generadora $H \in \mathrm{Mat}_{(n-s)\times n}(\mathbb{F}_q)$. A esta se la conoce como **matriz de control** del código lineal $\mathcal{C}$ inicial. Entre las propiedades más importantes de H destaca que podemos definir $\mathcal{C}$ a partir de esta. En efecto, no resulta complicado comprobar que $\mathcal{C} = \{\boldsymbol{x} \in \mathbb{F}_q^n \mid \boldsymbol{x}H^\top = \boldsymbol{0}\}$. Otra propiedad a tener en cuenta es que $(\mathcal{C}^\perp)^\perp = \mathcal{C}$. Esto se debe a que toda matriz generadora G y de control H para el código lineal $\mathcal{C}$ están relacionadas mediante la igualdad $GH^\top = 0$ (o, equivalentemente, $HG^\top = 0$).

Ejemplo 2. El ejemplo típico son los **códigos de Hamming**. Pese a que estos pueden construirse sobre cualquier alfabeto, su construcción binaria es muy sencilla: dados s un natural arbitrario y $n = 2^s - 1$, el *código de Hamming binario de orden s* (longitud n y dimensión s) tiene por matriz de control aquella cuyas columnas son las n palabras no nulas de $\mathbb{F}_2^s$ escritas en forma ascendente (es decir, la representación binaria ordenada de los números del 1 al n). En todos los casos la distancia mínima es 3. ◂

Ejemplo 3. Otro ejemplo interesante son los **códigos de Hadamard**. Dado s un natural, este es el código lineal binario de longitud $n = 2^s$ y dimensión s cuya matriz generadora se construye por columnas como sigue: para cada $i \leq n$ natural, la i-ésima columna se corresponde con los bits de la representación binaria del número entero i. Este es, como ya veremos, un caso especial de código de Reed-Muller binario. ◂

2.2. Procesos de codificación y decodificación

Ya hemos comentado al comienzo que un *proceso de codificación* no es más que aquel a través del cual añadimos información redundante a nuestras palabras con el fin de que, al emplear el correspondiente *proceso de decodificación*, podamos recuperar estas si se produce algún error durante la transmisión de información. Matemáticamente, esto significa que transformamos cada palabra que queremos transmitir en palabras código. Esto no resulta una tarea sencilla en general (pues no parece existir un procedimiento estándar a través del cual se asocia a cada una de estas palabras una palabra código). Sin embargo, cuando tenemos códigos lineales entre manos, sí que se tiene un método bastante general gracias a que en estos casos se tiene una matriz generadora del código. En efecto, basta multiplicar esta con cada palabra a transmitir para obtener palabras código con las que trabajar en cada caso.

2.2.1. Codificación por matrices generadoras dadas en forma estándar

Ya hemos dicho que para dar un código lineal $\mathcal{C}$ es suficiente dar una matriz generadora. Nos preguntamos ahora: ¿se puede obtener una matriz generadora de expresión lo más sencilla posible? La respuesta a esta pregunta nos la da el resultado del álgebra lineal que nos dice que toda matriz $G \in \mathrm{Mat}_{s\times n}(\mathbb{F}_q)$ (con $s \leq n$) de rango máximo s puede llevarse, realizando operaciones elementales en filas y columnas, a una matriz de la forma $(I_s \mid B)$ con $B \in \mathrm{Mat}_{s\times(n-s)}(\mathbb{F}_q)$. A esta se la conoce por *forma estándar* de G. Sin embargo, en general, para que el código lineal que tenga a esta por matriz generadora coincida con $\mathcal{C}$, deben realizarse estas transformaciones elementales solo por filas. Luego no todo código lineal admite una matriz generadora de este tipo. La importancia de las matrices generadoras dadas en forma estándar radica en lo fácil que resulta codificar con ellas ya que, dada una palabra $\boldsymbol{x} \equiv x_1 \dots x_s \in \mathbb{F}_q^s$ arbitraria, esta se codifica como

$$\boldsymbol{x}\,(I_s \mid B) = (x_1 \dots x_s\ \underbrace{c_{s+1} \dots c_n}_{\text{redundancias}}\,) \in \mathcal{C},$$

donde es evidente que esta se corresponde con una palabra código en la que las s primeras letras son, precisamente, las de la palabra original $\boldsymbol{x}$. Si $G = (I_s \mid B) \in \mathrm{Mat}_{s\times n}(\mathbb{F}_q)$ es una matriz generadora para un cierto código lineal, entonces $H = (-B^\top \mid I_{n-s}) \in \mathrm{Mat}_{(n-s)\times n}(\mathbb{F}_q)$ es una matriz de control para el mismo.

2.2.2. Métodos generales de decodificación

Recordemos que el principal objetivo de la decodificación es recuperar las palabras enviadas durante la transmisión de información que pueden haberse visto alteradas a lo largo de este proceso. Para ello, el *método de decodificación* debe ser capaz de detectar y, si es posible, corregir los errores que puedan haberse dado. Matemáticamente, recibida una palabra donde se ha dado un error de peso no nulo, hemos de ser capaces de detectar que esta no es una palabra código y hallar «aquellas más próximas» para corregir este. Cuando solo existe una palabra código en dichas condiciones, la *decodificación* es *única*. Existen dos métodos de decodificación generales válidos para todo código lineal $\mathcal{C} \subseteq \mathbb{F}_q^n$ con distancia mínima d.

Primero se tiene el llamado *método de decodificación basado en líderes*. Este se basa en la relación de equivalencia sobre $\mathbb{F}_q^n$ siguiente:

$$\forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{F}_q^n,\ \boldsymbol{x} \mathfrak{R} \boldsymbol{y} \iff \boldsymbol{x} - \boldsymbol{y} \in \mathcal{C}.$$

Supongamos que se desea decodificar $\boldsymbol{z} \in \mathbb{F}_q^n$. Se buscan en $[\boldsymbol{z}]$ (clase representada por $\boldsymbol{z}$ en la relación de equivalencia anterior) las palabras de menor peso posible (los *líderes* de $[\boldsymbol{z}]$). Sea una de estas $\boldsymbol{e}_{\boldsymbol{z}}$. Entonces, se decodifica $\boldsymbol{z}$ por $\boldsymbol{z} - \boldsymbol{e}_{\boldsymbol{z}}$, que es una palabra código (se toma como error al líder elegido). Esta palabra $\boldsymbol{z}$ admite decodificación única en caso de que $\omega(\boldsymbol{e}_{\boldsymbol{z}}) \leq \left\lfloor \frac{d-1}{2} \right\rfloor$. Este método es útil cuando resulta sencillo enumerar explícitamente las palabras del código.

Otro método de decodificación alternativo válido para cuando conocemos la matriz de control H para $\mathcal{C}$ es el llamado *método de decodificación basado en síndromes*. Este, definiendo $S(\boldsymbol{x}) \coloneqq \boldsymbol{x}H^\top$ como el *síndrome* de cada palabra $\boldsymbol{x} \in \mathbb{F}_q^n$ respecto de H, se basa en la relación de equivalencia sobre $\mathbb{F}_q^n$ siguiente:

$$\forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{F}_q^n,\ \boldsymbol{x} \sim \boldsymbol{y} \iff S(\boldsymbol{x}) = S(\boldsymbol{y}).$$

Supongamos que se desea decodificar $\boldsymbol{z} \in \mathbb{F}_q^n$. Si $S(\boldsymbol{z}) = \mathbf{0}$, se tiene que $\boldsymbol{z}$ es una palabra código y la decodificamos tal cual. En caso contrario, hay que buscar en $\bar{\boldsymbol{z}}$ (clase representada por $\boldsymbol{z}$ en la relación de equivalencia anterior) una palabra $\boldsymbol{e_z}$ de peso lo mínimo posible, para decodificar $\boldsymbol{z}$ como $\boldsymbol{z} - \boldsymbol{e_z}$, que es claramente una palabra código. Para obtener $\boldsymbol{e_z}$ en la práctica, construimos una tabla con los síndromes de las palabras del espacio vectorial total, ordenadas por pesos de menor a mayor. Habitualmente, se construye una tabla con los síndromes de las palabras con peso hasta $\left\lfloor \frac{d-1}{2} \right\rfloor$. A esta se la conoce por *tabla de síndromes*. Si el síndrome de nuestra palabra coincide con alguno de la tabla, podemos asegurar que la decodificación será única. Sin embargo, esto no es así cuando el peso de la palabra líder es mayor que $\left\lfloor \frac{d-1}{2} \right\rfloor$. Este método es útil cuando es fácil calcular una matriz de control del código.

Existen otros métodos de decodificación propios para ciertos tipos de códigos lineales más eficaces que estos dos, como puede ser el *método de decodificación cíclica*, válido para los llamados códigos cíclicos. Más adelante aparecerá otro método de decodificación, que tiene especial importancia cuando se trabaja con códigos de Reed-Muller binarios. Este es el llamado *método de decodificación por mayoría*.

Ejemplo 4. Consideremos el código de Hamming binario de orden 3. Una matriz de control para este código es

$$H = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

Para encontrar las palabras del código, basta con resolver las ecuaciones determinadas por H siguientes:

$$\begin{cases} & & & + x_4 & + x_5 & + x_6 & + x_7 = 0; \\ & + x_2 & + x_3 & & & + x_6 & + x_7 = 0; \\ + x_1 & & + x_3 & & + x_5 & & + x_7 = 0. \end{cases}$$

Podemos obtener una base tomando x_3, x_5, x_6 y x_7 como variables libres. Dando a tres de estas el valor 0 y 1 a la restante, obtenemos $\mathcal{B} = \{1110000, 1001100, 0101010, 1101001\}$. Denotando al código por $\mathcal{H}(3)$, tenemos que

$$\mathcal{H}(3) = \left\{ \begin{matrix} 0000000 & 1110000 & 1001100 & 0101010 & 1101001 & 0111100 & 1011010 & 0011001 \\ 1100110 & 0100101 & 1000011 & 0010110 & 1010101 & 0110011 & 0001111 & 1111111 \end{matrix} \right\}.$$

Considerando la palabra 1010101, que pertenece al código, vamos a reemplazar el último 1 por un 0. Ahora, empleando los dos métodos que se acaban de explicar, vamos a recuperar la palabra original.

- Aplicando, por un lado, el método de decodificación basado en líderes, tenemos que calcular la clase de equivalencia $[1010100] \equiv \{\boldsymbol{x} \mid \boldsymbol{x} - 1010100 \in \mathcal{H}(3)\}$. Esta es

$$\left\{ \begin{matrix} 1010100 & 0100100 & 0011000 & 1111110 & 0111101 & 1101000 & 0001110 & 1001101 \\ 0110010 & 1110001 & 0010111 & 1000010 & \mathbf{0000001} & 1100111 & 1011011 & 0101011 \end{matrix} \right\}.$$

  Como hay una única palabra de peso 1, la decodificación es única, y es 1010101.

- Si, por el contrario, aplicamos el método de decodificación basado en síndromes, calculamos tanto $S(1010100) = (1010100)H^{\top} = (111)$ como los síndromes de las palabras con peso hasta 1, para ver cuáles coinciden con este. Omitiendo la palabra nula, tenemos la siguiente tabla de síndromes:

palabras	1000000	0100000	0010000	0001000	0000100	0000010	0000001
síndromes	001	010	011	100	101	110	**111**

  Como hay una única palabra de mismo síndrome, la decodificación es única, y es 1010101.

En resumen, la decodificación ha sido la que cabía esperar empleando ambos métodos. ◂

2.3. Algunas construcciones de códigos lineales

Se incluyen a continuación tres construcciones interesantes de códigos lineales. Resulta un buen ejercicio de repaso comprobar que, efectivamente, lo son, así como que se cumplen todas las propiedades que se van a enunciar para las mismas. Todas estas comprobaciones pueden encontrarse en el trabajo de De Arriba De La Hera [1, capítulo 1, Problemas Resueltos].

Ejemplo 5 (código suma). Sean $\mathcal{C}_1, \mathcal{C}_2 \subseteq \mathbb{F}_q^n$ dos códigos lineales con distancia mínima d_i, dimensión k_i y matriz generadora $G_i \in \mathrm{Mat}_{k_i \times n}(\mathbb{F}_q)$, respectivamente, siendo $i \in \{1, 2\}$. Se demuestra que la **suma** de $\mathcal{C}_1$ y $\mathcal{C}_2$, dada por

$$\mathcal{C}_1 + \mathcal{C}_2 \coloneqq \{\boldsymbol{c}_1 + \boldsymbol{c}_2 \mid \boldsymbol{c}_i \in \mathcal{C}_i \text{ con } i \in \{1, 2\}\},$$

es un código lineal q-ario con distancia mínima $d \leq \min\{d_1, d_2\}$. Se puede probar, además, que, si $\mathcal{C}_1 \cap \mathcal{C}_2 = \{\boldsymbol{0}\}$, entonces $\dim(\mathcal{C}_1 + \mathcal{C}_2) = k_1 + k_2$ y una matriz generadora de $\mathcal{C}_1 + \mathcal{C}_2$ viene dada por

$$G = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix}.$$

◂

Ejemplo 6 (código concatenación). Sean $\mathcal{C}_i \subseteq \mathbb{F}_q^{n_i}$ códigos lineales de longitud n_i, distancia mínima d_i, dimensión k_i, matriz generadora $G_i \in \mathrm{Mat}_{k_i \times n_i}(\mathbb{F}_q)$ y matriz de control $H_i \in \mathrm{Mat}_{(n-k_i) \times n}(\mathbb{F}_q)$, respectivamente, siendo $i \in \{1, 2\}$. Se demuestra que la **concatenación** de $\mathcal{C}_1$ con $\mathcal{C}_2$, dada por

$$\mathcal{C}_1 * \mathcal{C}_2 \coloneqq \{\boldsymbol{c}_1 * \boldsymbol{c}_2 = c_{11} \dots c_{1n_1} c_{21} \dots c_{2n_2} \mid \boldsymbol{c}_i \in C_i \text{ con } i \in \{1, 2\}\},$$

es un código lineal q-ario de longitud $n_1 + n_2$, dimensión $k_1 + k_2$, distancia mínima $d = \min\{d_1, d_2\}$ y matrices generadora y de control

$$G = \begin{pmatrix} G_1 & 0 \\ 0 & G_2 \end{pmatrix} \qquad \text{y} \qquad H = \begin{pmatrix} H_1 & 0 \\ 0 & H_2 \end{pmatrix}.$$

◂

Ejemplo 7 (construcción de Plotkin). Sean $\mathcal{C}_1, \mathcal{C}_2 \subseteq \mathbb{F}_q^n$ dos códigos lineales de distancia mínima d_i, dimensión k_i, matriz generadora $G_i \in \mathrm{Mat}_{k_i \times n}(\mathbb{F}_q)$ y matriz de control $H_i \in \mathrm{Mat}_{(n-k_i) \times n}(\mathbb{F}_q)$, respectivamente, siendo $i \in \{1, 2\}$. Se demuestra que el conjunto definido por

$$\mathcal{C}_1 \circledast \mathcal{C}_2 \coloneqq \{(\boldsymbol{c}_1 \mid \boldsymbol{c}_1 + \boldsymbol{c}_2) \mid \boldsymbol{c}_i \in \mathcal{C}_i \text{ con } i \in \{1, 2\}\},$$

donde $(\boldsymbol{c}_1 \mid \boldsymbol{c}_1 + \boldsymbol{c}_2) \coloneqq \boldsymbol{c}_1 * (\boldsymbol{c}_1 + \boldsymbol{c}_2) = \boldsymbol{c}_1 * \boldsymbol{c}_1 + \boldsymbol{0} * \boldsymbol{c}_2$ para todo $\boldsymbol{c}_1 \in \mathcal{C}_1, \boldsymbol{c}_2 \in \mathcal{C}_2$, es un código lineal q-ario de longitud $2n$, dimensión $k_1 + k_2$, distancia mínima $d = \min\{2d_1, d_2\}$ y matrices generadora y de control

$$G = \begin{pmatrix} G_1 & G_1 \\ 0 & G_2 \end{pmatrix} \qquad \text{y} \qquad H = \begin{pmatrix} H_1 & 0 \\ -H_2 & H_2 \end{pmatrix}.$$

◂

Nota 8. El código lineal presentado en el ejemplo 7 jugará un papel muy importante cuando tratemos la segunda de las construcciones para los códigos de Reed-Muller binarios. ◂

3. Códigos de Reed-Muller binarios

3.1. Aspectos históricos

Los códigos de Reed-Muller son una familia infinita de códigos lineales, que toman su nombre de los dos matemáticos que los propusieron en el año 1954, prácticamente al mismo tiempo, en dos trabajos independientes: Irving Stoy Reed[2] y David Eugene Muller[3]. Ambos se ocuparon de introducir estos en el caso binario. Hoy se sabe que el primero en realizar la primera de las construcciones para estos códigos en su forma binaria fue Muller [3], mientras que su estudio en detalle y la sencilla decodificación por la que son tan conocidos e importantes en este caso binario es obra de Reed [4].

Los códigos de Reed-Muller tienen una gran importancia en la historia. Su estudio en la década de los años 50 fue fundamental para que en los años posteriores se hiciesen grandes avances en la exploración espacial. Así, desde 1969 hasta 1977, todas las naves espaciales de la NASA iban equipadas con un código de Reed-Muller binario de longitud 32, dimensión 6 y distancia mínima 16. Se escogió dicho código dado que el cociente entre la dimensión del mismo y su longitud es (relativamente) pequeño para la amplia distancia mínima que posee. Por esta razón podemos decir que nos encontramos ante un código de *bajo coste* y buenas capacidades para *corregir errores*.

[2] Matemático e ingeniero estadounidense; 12 noviembre, 1923 - 11 septiembre, 2012.

[3] Matemático e informático teórico estadounidense; 2 noviembre, 1924 - 27 abril, 2008.

Una de las misiones más destacadas que se llevó a cabo con el uso de estos códigos, y que pasó a la historia por su gran impacto, fue la realizada en los años 70 por la sonda Mariner 9, ya que esta fue la primera que permitió la observación fotográfica de la superficie marciana. Esta sonda fue lanzada el 30 de mayo del 1971, llegando a su destino el 13 de noviembre del mismo año, convirtiéndose así en la primera nave espacial en orbitar un planeta distinto al nuestro. Científicamente, esta misión, que constituyó una continuación de las observaciones adquiridas por las sondas Mariner 6 y 7, tenía como objetivo mostrar las primeras fotografías de Marte. En un principio la misión se complicó debido a las grandes tormentas de arena que se dieron sobre todo el conjunto de la superficie del planeta. Sin embargo, en 1972, cuando por fin amainaron dichas tormentas, se obtuvieron estas primeras fotografías en blanco y negro, las cuales cambiaron completamente la visión que se tenía hasta entonces del planeta rojo (figura 1). La sonda tomó fotografías en blanco y negro de $600 \times 600 = 360\,000$ píxeles, donde a cada píxel se le asignó una 6-tupla para representar el brillo. Cada píxel era codificado como una palabra de longitud 32 (se emplearon 26 bits de redundancia). Fue necesario usar un código con palabras de gran longitud, pues los errores de transmisión debían minimizarse al máximo dada la enorme distancia que los mensajes recorrían desde Marte a la Tierra. Era, además, imprescindible, dado el tiempo requerido por cada transmisión, que la decodificación fuese posible en la mayor parte de los casos y con garantías de unicidad.

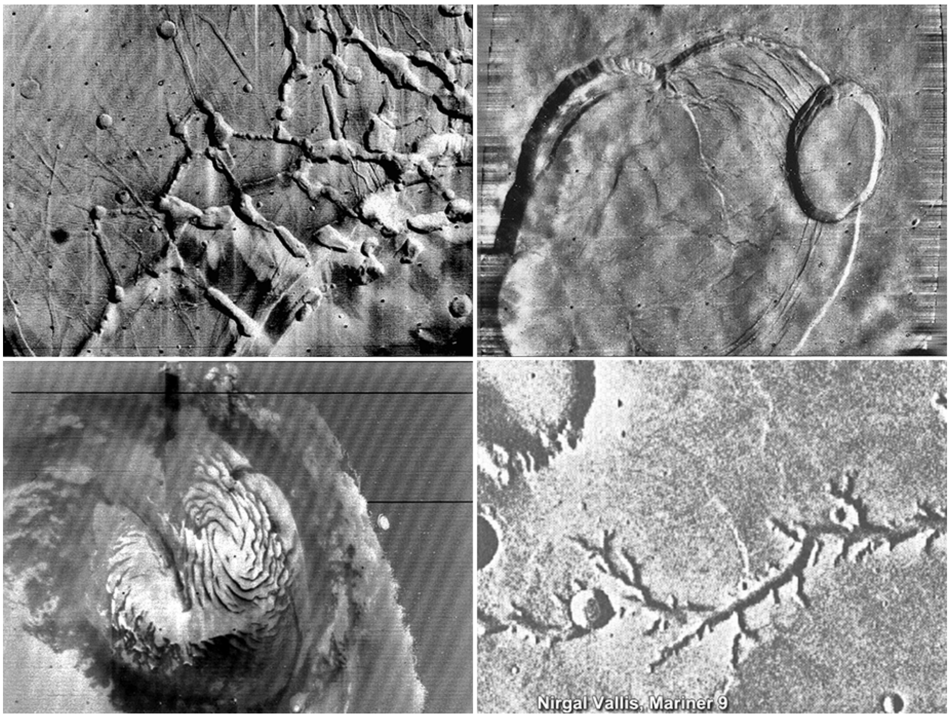

Figura 1: Imágenes, facilitadas por la NASA bajo dominio público, captadas por la sonda Mariner 9.

3.2. Construcciones y propiedades principales

Como ya se ha adelantado, existen tres formas de introducir los códigos de Reed-Muller binarios. Cada una de ellas resulta importante por motivos distintos, como ya iremos comprobando a lo largo del artículo.

3.2.1. Construcción original de Muller

Esta es la primera construcción conocida, debida a D. E. Muller. Además de la referencia principal [1], ha sido necesario también consultar las notas de Iranzo Aznar y Pérez Monasor [2, Lección 7] para la redacción de esta parte por motivos que se entenderán más adelante (véase la nota 19). Para introducir esta construcción en su versión original, es necesaria la conocida como *teoría de Boole*, la cual se pretende estudiar brevemente a continuación.

Definición 9. Sea m un número natural. Una aplicación $f\colon \mathbb{F}_2^m \to \mathbb{F}_2$ es una **función booleana** de m variables. Se denota al conjunto de todas las funciones booleanas de m variables por $\mathfrak{B}_m$. ◂

Observación 10. Podemos dotar a $\mathfrak{B}_m$ con una estructura de anillo conmutativo y unitario. Más aún, se tiene que este conjunto también posee estructura de $\mathbb{F}_2$-espacio vectorial. En definitiva, nos encontramos ante una $\mathbb{F}_2$-álgebra conmutativa y unitaria, la cual se conoce usualmente por **álgebra de Boole**. ◂

A partir de este momento, y salvo que se diga lo contrario, vamos a trabajar con el álgebra $\mathbb{F}_2^m$ dado m un entero no negativo, y tendremos siempre que $n = 2^m$.

Observación 11. Si se fija un orden en los n elementos de $\mathbb{F}_2^m$, es posible describir toda función booleana $f\colon \mathbb{F}_2^m \to \mathbb{F}_2$ de manera unívoca a través de una tabla con todos los elementos de $\mathbb{F}_2^m$ y los respectivos valores que toma f en cada uno de estos. A esta se la conoce por *tabla de verdad* asociada a f. Además, fijado un orden $\mathbb{F}_2^m = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ y dada $f \in \mathfrak{B}_m$ arbitraria, inducimos la siguiente notación:

$$f_i \underset{\text{not.}}{\equiv} f(\boldsymbol{v}_i), \quad \forall\, i \in \{1, \dots, n\}.$$

Llamaremos a este valor *coordenada i-ésima* de f bajo el orden establecido en $\mathbb{F}_2^m$. En estas condiciones, toda aplicación $f \in \mathfrak{B}_m$ puede identificarse de manera unívoca empleando tablas de verdad con la correspondiente palabra $\boldsymbol{f} \equiv f_1 \dots f_n \in \mathbb{F}_2^n$. Así, lo natural es trabajar con el conjunto $\mathbb{F}_2^m = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ ordenado cuando se tengan funciones booleanas. A $\boldsymbol{f} \in \mathbb{F}_2^n$ se la conoce por *palabra característica* de f bajo el orden establecido. En particular, se deduce de este hecho que $|\mathfrak{B}_m| = 2^n < \infty$. ◂

Definición 12. Definimos el *anillo de los polinomios booleanos* de m indeterminadas como el cociente

$$\mathcal{P}_m \underset{\text{not.}}{\equiv} \frac{\mathbb{F}_2[x_1, \dots, x_m]}{(x_1^2 - x_1, \dots, x_m^2 - x_m)}.$$

Cada clase de equivalencia $\overline{F} = F + (x_1^2 - x_1, \dots, x_m^2 - x_m) \in \mathcal{P}_m$ posee un único representante

$$F^* \underset{\text{not.}}{\equiv} \sum a_{i_1 \dots i_m} x_1^{i_1} \cdots x_m^{i_m} \in \mathbb{F}_2[x_1, \dots, x_m] \quad \left(a_{i_1 \dots i_m} \in \mathbb{F}_2\right),$$

verificando que $i_1, \dots, i_m \in \{0, 1\}$. Esta se conoce por *forma reducida* del polinomio F. Además, es evidente que esta está unívocamente determinada aplicando a cada monomio de F las reglas

$$\underbrace{x_i x_j = x_j x_i}_{\text{conmutatividad en } \mathbb{F}_2[x_1, \dots, x_m]} \qquad \text{y} \qquad \underbrace{x_i^2 = x_i,}_{\text{pequeño teorema de Fermat en } \mathbb{F}_2}$$

para cualesquiera $i, j \in \{1, \dots, m\}$ diferentes, hasta que los factores del polinomio resultante sean todos distintos. Resulta natural definir en estas condiciones el *grado* de un polinomio booleano como el grado, entendido en el sentido usual, del correspondiente polinomio en su forma reducida. ◂

Observación 13. No es complicado percatarse de que $\mathcal{P}_m$ tiene también estructura de espacio vectorial sobre $\mathbb{F}_2$. De hecho, el conjunto $\mathcal{B} = \{x_1^{r_1} \cdots x_m^{r_m} \mid r_i \in \{0, 1\}\ \forall\, i \in \{1, \dots, m\}\}$ es una base de este (en particular, $\mathcal{P}_m$ es finito como espacio vectorial). Así, dado que el número de monomios booleanos de m indeterminadas y grado k es $\binom{m}{k}$ trivialmente, no resulta complicado comprobar que $|\mathcal{P}_m| = 2^m$. ◂

De ahora en adelante, y salvo que se diga lo contrario, nos restringimos al uso de polinomios en forma reducida cuando tratemos con los elementos de $\mathcal{P}_m$ y vamos a suponer que tenemos fijado un orden $\mathbb{F}_2^m = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$. En particular, utilizar palabras en $\mathbb{F}_2^n$ y funciones booleanas de m variables resulta equivalente. Razonando por inducción sobre m, el número de indeterminadas, obtenemos el siguiente resultado, necesario para probar el que sigue y que resulta clave para llevar a cabo esta construcción.

Lema 14. *Sea $F \in \mathbb{F}_2[x_1, \dots, x_m]$ un polinomio arbitrario no necesariamente dado en forma reducida. Si se cumple que $F(u_1, \dots, u_m) = 0$ para todo $u_1, \dots, u_m \in \mathbb{F}_2$, entonces $F^* = 0$.*

Teorema 15. *Los conjuntos $\mathfrak{B}_m$ y $\mathcal{P}_m$ son isomorfos como $\mathbb{F}_2$-álgebras conmutativas y unitarias, a través de la aplicación que asocia a cada polinomio booleano F la única función booleana f tal que*

$$F(u_1, u_2, \dots, u_m) = f(\boldsymbol{u}), \quad \forall\, \boldsymbol{u} = (u_1, u_2, \dots, u_m) \in \mathbb{F}_2^m.$$

En consecuencia, cada polinomio en forma reducida tiene asociada una, y solo una, palabra de $\mathbb{F}_2^n$.

Demostración. Veamos en primer lugar que la aplicación

$$\begin{array}{rccl} \psi_m : & \mathbb{F}_2[x_1, \dots, x_m] & \longrightarrow & \mathbb{F}_2^n \\ & F & \longmapsto & \psi_m(F) := \boldsymbol{f} \equiv f_1 \dots f_n \end{array} \tag{3}$$

es un epimorfismo de álgebras sobre $\mathbb{F}_2$. El único paso no trivial es comprobar la sobreyectividad. Para ello, dado $\boldsymbol{x} \in \mathbb{F}_2^n$ arbitrario, suponiendo que $\boldsymbol{v}_i = (\alpha_1^i, \dots, \alpha_m^i) \in \mathbb{F}_2^m$ para todo $i \in \{1, \dots, n\}$, definimos

$$F_{\boldsymbol{v}_i}(x_1, \dots, x_m) := \prod_{j=1}^{m} \left(1 - \left(x_j - \alpha_j^i\right)\right) \in \mathbb{F}_2[x_1, \dots, x_m], \quad \forall\, i \in \{1, \dots, n\}.$$

Por el pequeño teorema de Fermat, se tiene que, para todo $\boldsymbol{w} = (w_1, \dots, w_m) \in \mathbb{F}_2^m$ y todo $i \in \{1, \dots, n\}$,

$$F_{\boldsymbol{v}_i}(w_1, \dots, w_m) = \begin{cases} 1, & \text{si } \boldsymbol{w} = \boldsymbol{v}_i; \\ 0, & \text{si } \boldsymbol{w} \neq \boldsymbol{v}_i. \end{cases}$$

De esta manera, resulta inmediato comprobar, por cómo se define la aplicación (3), si escribimos las imágenes explícitamente, que el conjunto $\mathcal{B} = \{\psi_m(F_{\boldsymbol{v}_i}) \mid i \in \{1, \dots, n\}\}$ se corresponde con la base canónica de $\mathbb{F}_2^n$. En consecuencia, existen escalares $k_i \in \mathbb{F}_2$ para cada $i \in \{1, \dots, n\}$ tales que

$$\boldsymbol{x} = \sum_{i=1}^{n} k_i \psi_m(F_{\boldsymbol{v}_i}) = \psi_m\left(\sum_{i=1}^{n} k_i F_{\boldsymbol{v}_i}\right).$$

Así, tomando $F = \sum_{i=1}^{n} k_i F_{\boldsymbol{v}_i}$, es evidente que $\psi_m(F) = \boldsymbol{x}$. Queda así probada la sobreyectividad de (3). Hecho esto, aplicando el primer teorema de isomorfía para álgebras, se sigue que

$$\frac{\mathbb{F}_2[x_1, \dots, x_m]}{\operatorname{Ker}(\psi_m)} \cong \operatorname{Im}(\psi_m) \equiv \mathbb{F}_2^n.$$

Basta probar, por tanto, que $\operatorname{Ker}(\psi_m) = (x_1^2 - x_1, \dots, x_m^2 - x_m)$ y habremos terminado. El contenido no trivial es $\operatorname{Ker}(\psi_m) \subseteq (x_1^2 - x_1, \dots, x_m^2 - x_m)$ y se argumenta por reducción al absurdo: dado $F \in \operatorname{Ker}(\psi_m)$ arbitrario, vamos a suponer que $F \notin (x_1^2 - x_1, \dots, x_m^2 - x_m)$ para llegar a una contradicción. En estas condiciones, suponiendo que R es la forma reducida de F, se tiene que la clase de equivalencia para F puede escribirse como $\overline{F} = R + (x_1^2 - x_1, \dots, x_m^2 - x_m)$, donde $R \neq 0$ necesariamente. Pero, como $\psi_m(F) = \mathbf{0}$ por hipótesis, estamos ante las condiciones del lema 14, luego $F^* = R = 0$, en contra de la suposición hecha. ∎

Sea r un número natural tal que $r \leq m$. Denotaremos por $\mathcal{P}_m(r)$ al conjunto de los polinomios booleanos de m indeterminadas en $\mathcal{P}_m$ con grado menor o igual que r. Este es trivialmente un $\mathbb{F}_2$-subespacio vectorial de $\mathcal{P}_m$ de dimensión finita. Empleando entonces el teorema 15 que acabamos de probar, estamos en condiciones de dar la siguiente definición importante.

Definición 16. Se define por **código de Reed-Muller** binario $\mathcal{RM}(r, m)$ de orden r y longitud n a la imagen directa de $\mathcal{P}_m(r)$ a través de la aplicación dada en (3). Dicho de otra manera, este es el conjunto de todas las palabras binarias de longitud n asociadas a todos los polinomios booleanos de m indeterminadas mediante el isomorfismo dado en el teorema 15 con grado menor o igual que r. ◂

Nota 17. Por convenio, escribiremos que $\mathcal{RM}(\ell, m) = \{0 \dots 0\}$ para todo entero $\ell < 0$. Además, por la definición, es inmediato que $\mathcal{RM}(m, m) = \mathbb{F}_2^n$ y $\mathcal{RM}(0, m) = \{0 \dots 0, 1 \dots 1\}$. ◂

Observación 18. Esta definición es independiente del orden fijado sobre $\mathbb{F}_2^m$. En otras palabras, dado un código de Reed-Muller binario construido bajo un cierto orden en $\mathbb{F}_2^m$ fijado, al cambiar dicho orden, obtenemos otro código de Reed-Muller binario, con los mismos parámetros que el inicial (estos solo se diferencian en que se han permutado entre sí las letras de un número finito de posiciones fijadas en todas las palabras del código). Cuando dos códigos están relacionados tal y como se acaba de explicar, se dice que son *códigos equivalentes por permutación*. ◂

Nota 19. Esta construcción fue generalizada a cualquier cuerpo finito $\mathbb{F}_q$ en 1968. El lector interesado en ello puede consultar el trabajo de De Arriba De La Hera [1, capítulo 2] y las referencias ahí recogidas para comprender esta generalización. ◂

Proposición 20. *El código de Reed-Muller binario $\mathcal{RM}(r,m)$ es un código lineal de longitud n sobre $\mathbb{F}_2$. Manteniendo las notaciones introducidas en el teorema 15, una matriz generadora para este es*

$$G(r,m) \underset{not.}{\equiv} \begin{pmatrix} \psi_m(1) \\ \psi_m(x_1) \\ \vdots \\ \psi_m(x_m) \\ \psi_m(x_1x_2) \\ \vdots \\ \psi_m(x_{m-r+1}\dots x_m) \end{pmatrix}. \tag{4}$$

En consecuencia, la dimensión como $\mathbb{F}_2$-subespacio vectorial de $\mathbb{F}_2^n$ viene dada por la fórmula

$$\dim(\mathcal{RM}(r,m)) = \sum_{k=0}^{r} \binom{m}{k}. \tag{5}$$

Además, no resulta complicado comprobar que $\mathcal{RM}(r,m)^\perp = \mathcal{RM}(m-r-1,m)$.

Demostración. Basta probar que todas las palabras de $\mathbb{F}_2^n$ asociadas a los monomios reducidos de $\mathcal{P}_m(r)$ constituyen una base de $\mathcal{RM}(r,m)$. Nos es suficiente ver que estas generan un sistema linealmente independiente argumentando por reducción al absurdo y haciendo uso del lema 14. Una vez se tiene esto, con un simple argumento combinatorio, obtenemos la fórmula dada para la dimensión. Finalmente, la igualdad entre códigos lineales se deduce de la inclusión $\mathcal{RM}(m-r-1,m) \subseteq \mathcal{RM}(r,m)^\perp$ dado que la suma de las dimensiones de $\mathcal{RM}(m-r-1,m)$ y $\mathcal{RM}(r,m)$ es 2^m. ■

Ejemplo 21. Vamos a construir empleando la definición dada el código $\mathcal{RM}(1,3)$. Para ello, vamos a fijar el orden $\mathbb{F}_2^3 = \{000, 001, 010, 011, 100, 101, 110, 111\}$. Hecho esto, tenemos que obtener $\mathcal{P}_3(1)$, esto es, los polinomios booleanos de 3 indeterminadas que tengan grado menor o igual que 1. Estos son los siguientes:

$$\begin{array}{cccccccc} 0, & x_1, & x_2, & x_3, & x_1+x_2, & x_1+x_3, & x_2+x_3, & x_1+x_2+x_3, \\ 1, & 1+x_1, & 1+x_2, & 1+x_3, & 1+x_1+x_2, & 1+x_1+x_3, & 1+x_2+x_3, & 1+x_1+x_2+x_3. \end{array}$$

Así, ya podemos enumerar las palabras del código de Reed-Muller deseado, y tenemos que estas son

$$\left\{ \begin{array}{cccccccc} 00000000 & 00001111 & 00110011 & 01010101 & 00111100 & 01011010 & 01100110 & 01101001 \\ 11111111 & 11110000 & 11001100 & 10101010 & 11000011 & 10100101 & 10011001 & 10010110 \end{array} \right\}.$$

A partir de esto, podemos obtener una matriz generadora para el código inmediatamente, y tenemos por la fórmula (5) que la dimensión de este es 4. ◂

3.2.2. Construcción recursiva de Plotkin

Nos basamos en la construcción de Plotkin recogida en el ejemplo 7 para obtener una construcción recursiva de los códigos de Reed-Muller binarios, solo válida para un cierto orden en los elementos de $\mathbb{F}_2^m$.

Definición 22. Dada la expansión binaria $i_0 + 2i_1 + 2^2 i_2 + \cdots + 2^{m-2} i_{m-2} + 2^{m-1} i_{m-1}$ de un cierto entero i, con $i_0, i_1, \dots, i_{m-2}, i_{m-1} \in \{0,1\}$, asociamos a $i+1$ el elemento $(i_{m-1}, i_{m-2}, \dots, i_2, i_1, i_0) \in \mathbb{F}_2^m$. Esto es,

$$\begin{array}{ccc} 1 & \longrightarrow & (0,0,\dots,0,0,0); \\ 2 & \longrightarrow & (0,0,\dots,0,0,1); \\ 3 & \longrightarrow & (0,0,\dots,0,1,0); \\ 4 & \longrightarrow & (0,0,\dots,0,1,1); \\ 5 & \longrightarrow & (0,0,\dots,1,0,0); \\ \vdots & \vdots & \vdots \\ 2^m-2 & \longrightarrow & (1,1,\dots,1,0,1); \\ 2^m-1 & \longrightarrow & (1,1,\dots,1,1,0); \\ 2^m & \longrightarrow & (1,1,\dots,1,1,1). \end{array}$$

Al orden inducido por esta asociación sobre $\mathbb{F}_2^m$ lo llamaremos *orden canónico* de $\mathbb{F}_2^m$. ◂

Supondremos a lo largo de todo este apartado que tenemos fijado el orden canónico para los elementos de $\mathbb{F}_2^m$. Además, mantendremos la notación ψ_m introducida en el teorema 15 para la aplicación (3). Nuestro primer objetivo pasa por estudiar las propiedades fundamentales que se tienen bajo este orden, las cuales se recogen en los resultados siguientes, y nos permiten escribir $\mathcal{RM}(r,m)$ como construcción de Plotkin entre $\mathcal{RM}(r,m-1)$ y $\mathcal{RM}(r-1,m-1)$. Hecho esto, podremos obtener expresiones dependientes de r y m tanto para la distancia mínima como para una matriz generadora de todos los códigos de Reed-Muller binarios (esto último solo para cuando se tiene el orden canónico) de manera recursiva.

Lema 23. *Se tiene la siguiente tabla de verdad:*

$$\begin{array}{rcl} \psi_m(x_m) & = & 0101010101010101\dots010101; \\ \psi_m(x_{m-1}) & = & 001100110011001\dots0110011; \\ \psi_m(x_{m-2}) & = & 00001111000011\dots00001111; \\ \vdots & \vdots & \vdots \\ \psi_m(x_2) & = & 00\dots011\dots100\dots011\dots11; \\ \psi_m(x_1) & = & \underbrace{00000\dots0000}_{2^{m-1}}\underbrace{11111\dots1111}_{2^{m-1}}. \end{array}$$

Lema 24. *Sea F un polinomio booleano en cuya expresión no aparece la indeterminada x_1 y que, por tanto, podemos ver como polinomio de $m-1$ indeterminadas. Si la palabra binaria asociada en este caso es $\boldsymbol{f} = f_1\dots f_{2^{m-1}}$, entonces se tiene que la palabra binaria asociada a F como polinomio booleano de $\mathcal{P}_m$ es la concatenación $\boldsymbol{f} * \boldsymbol{f} = f_1\dots f_{2^{m-1}} f_1 \dots f_{2^{m-1}}$.*

Estos dos resultados son fundamentales para probar el que sigue a continuación, que es característico de esta construcción. Las demostraciones para ambos son inmediatas: para el primero, basta construir las tablas de verdad para el orden canónico de $\mathbb{F}_2^m$ asociadas a los monomios booleanos x_i como elementos de $\mathcal{P}_m$, mientras que el segundo se sigue de construir las tablas de verdad para el orden canónico de $\mathbb{F}_2^m$ asociadas al polinomio dado en el enunciado, visto como elemento tanto de $\mathcal{P}_m$ como de $\mathcal{P}_{m-1}$, y, hecho esto, comparar estas dos palabras para ver que la segunda concatenada consigo misma es la primera.

Teorema 25. *Dado un número natural r tal que $0 < r < m$, se cumple que*

$$\mathcal{RM}(r,m) = \mathcal{RM}(r,m-1) \circledast \mathcal{RM}(r-1,m-1). \tag{6}$$

Demostración. Veamos primero que $\mathcal{RM}(r,m) \subseteq \mathcal{RM}(r,m-1) \circledast \mathcal{RM}(r-1,m-1)$. Sea $\boldsymbol{x} \in \mathcal{RM}(r,m)$ una palabra código, asociada al polinomio booleano F de m indeterminadas y grado menor o igual que r. Podemos expresar F en función de dos polinomios booleanos de $m-1$ indeterminadas G y H por

$$F(x_1,\dots,x_m) = G(x_2,\dots,x_m) + x_1 H(x_2,\dots,x_m), \tag{7}$$

donde G tiene grado menor o igual que r y H, grado menor o igual que $r-1$. Supongamos que $\boldsymbol{x}_G$ y $\boldsymbol{x}_H$ son las palabras binarias asociadas a estos polinomios, respectivamente, los cuales estamos viendo como si fuesen polinomios booleanos de $m-1$ indeterminadas. Es evidente, por definición, que $\boldsymbol{x}_G \in \mathcal{RM}(r,m-1)$ y $\boldsymbol{x}_H \in \mathcal{RM}(r-1,m-1)$. Ahora, por el lema 24, se tiene que $\boldsymbol{x}_G * \boldsymbol{x}_G$ y $\boldsymbol{x}_H * \boldsymbol{x}_H$ son las palabras binarias asociadas a nuestros polinomios booleanos, respectivamente, vistos como si tuvieran m indeterminadas. Aplicando entonces ψ_m a (7), por el lema 23 se tiene, por ser este un homomorfismo de álgebras, que

$$\boldsymbol{x} = \psi_m(F) = \psi_m(G) + \psi_m(x_1)\psi_m(H) = \boldsymbol{x}_G * \boldsymbol{x}_G + \boldsymbol{0} * \boldsymbol{x}_H \in \mathcal{RM}(r,m-1) \circledast \mathcal{RM}(r-1,m-1).$$

Para obtener la igualdad basta ver que ambos códigos tienen la misma dimensión. En efecto, por (5) y la fórmula de Pascal, por como viene dada la fórmula correspondiente a la dimensión en la construcción de Plotkin recogida en el ejemplo 7,

$$\begin{aligned} \dim(\mathcal{RM}(r,m)) &= \sum_{k=0}^{r}\binom{m}{k} = \binom{m}{0} + \sum_{k=1}^{r}\binom{m}{k} = \binom{m-1}{0} + \sum_{k=1}^{r}\binom{m-1}{k} + \sum_{k=1}^{r}\binom{m-1}{k-1} \\ &= \sum_{k=0}^{r}\binom{m-1}{k} + \sum_{k=0}^{r-1}\binom{m-1}{k} = \dim(\mathcal{RM}(r,m-1) \circledast \mathcal{RM}(r-1,m-1)) \end{aligned}$$

por las propiedades de los coeficientes binomiales y haciendo el correspondiente cambio de variable. ∎

Definición 26. Dado r un entero tal que $0 \leq r \leq m$, se define recursivamente $\mathcal{RM}(r,m)$ para el orden canónico de $\mathbb{F}_2^m$ teniendo en cuenta el caso base de la nota 17 y empleando la fórmula (6). ◂

Proposición 27. *Dado r natural tal que $0 \leq r \leq m$, la distancia mínima de $\mathcal{RM}(r,m)$ es 2^{m-r}.*

Demostración. Consecuencia del ejemplo 7, calculando primero esta a mano en el caso base. ∎

Lema 28. *Se tiene que $G(0,m) = (1\ 1\ \cdots\ 1)$ y $G(m,m) = \begin{pmatrix} G(m-1,m) \\ 0\ \cdots\ 0\ 1 \end{pmatrix}$.*

Demostración. Ambas igualdades son consecuencia inmediata de cómo viene dada la matriz generadora correspondiente en (4). La primera se debe a que la aplicación ψ_m verifica que $\psi_m(1) = \mathbf{1}$. La segunda se sigue de que ψ_m es un homomorfismo de álgebras y el lema 23, pues tenemos de esta manera que $\psi_m(x_1 \cdots x_m) = \psi_m(x_1) \cdots \psi_m(x_m) = 0 \ldots 01$ y el resto de polinomios que quedan tienen por palabras características a las del código $\mathcal{RM}(m-1,m)$, cuya matriz generadora es $G(m-1,m)$. ∎

Proposición 29. *Dado r natural tal que $0 \leq r \leq m$, se tiene que la matriz generadora de $\mathcal{RM}(r,m)$ viene dada de manera recursiva a partir de las matrices generadoras de $\mathcal{RM}(r,m-1)$ y $\mathcal{RM}(r-1,m-1)$ como sigue:*

$$G(r,m) = \begin{pmatrix} G(r,m-1) & G(r,m-1) \\ 0 & G(r-1,m-1) \end{pmatrix}. \tag{8}$$

Demostración. Consecuencia del ejemplo 7, teniendo en cuenta que tenemos por caso base los dos casos recogidos en el lema 28. ∎

Ejemplo 30. Usando la proposición 29 que acabamos de demostrar, vamos a construir $G(1,3)$. Esto es, vamos a dar la matriz generadora de $\mathcal{RM}(1,3)$ para el orden canónico de $\mathbb{F}_2^m$. De esta manera, podemos enumerar las palabras de $\mathcal{RM}(1,3)$ para este orden fácilmente. En virtud de lo que hemos visto, por inducción, se tiene que

$$G(1,3) = \begin{pmatrix} G(1,2) & G(1,2) \\ 0 & G(0,2) \end{pmatrix} = \begin{pmatrix} G(1,1) & G(1,1) & G(1,1) & G(1,1) \\ 0 \quad 0 & G(0,1) & 0 \quad 0 & G(0,1) \\ 0 \quad 0 & 0 \quad 0 & 1 \quad 1 & 1 \quad 1 \end{pmatrix}$$

$$= \begin{pmatrix} G(0,1) & G(0,1) & G(0,1) & G(0,1) \\ 0 \quad 1 & 0 \quad 1 & 0 \quad 1 & 0 \quad 1 \\ 0 \quad 0 & 1 \quad 1 & 0 \quad 0 & 1 \quad 1 \\ 0 \quad 0 & 0 \quad 0 & 1 \quad 1 & 1 \quad 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

Además, por la fórmula dada en la proposición 27, se tiene que la distancia mínima es $2^{3-1} = 2^2 = 4$. Así, hemos concluido que este es un código de Reed-Muller binario cuya longitud y dimensión vienen a ser 8 y 4, respectivamente, con distancia mínima 4 también. Por tanto, pese a las buenas capacidades para corregir errores, este código no resulta ser de bajo coste. Sin embargo, es posible dar con otro código que es lo bastante eficaz en estos dos aspectos, como se va a comprobar al final del artículo. ◂

Nota 31. Cabe resaltar que el orden dado en la definición 22 no es único. A saber, en las mismas condiciones enunciadas, también podemos tomar el orden inducido por la asignación

$$\begin{array}{ccl} 1 & \longrightarrow & (0,0,0,\ldots,0,0); \\ 2 & \longrightarrow & (1,0,0,\ldots,0,0); \\ 3 & \longrightarrow & (0,1,0,\ldots,0,0); \\ 4 & \longrightarrow & (1,1,0,\ldots,0,0); \\ 5 & \longrightarrow & (0,0,1,\ldots,0,0); \\ \vdots & \vdots & \qquad \vdots \\ 2^m-2 & \longrightarrow & (1,0,1,\ldots,1,1); \\ 2^m-1 & \longrightarrow & (0,1,1,\ldots,1,1); \\ 2^m & \longrightarrow & (1,1,1,\ldots,1,1). \end{array}$$

Este orden parece más «coherente» y los resultados se cumplen exactamente igual, cambiando algunos detalles. Sin embargo, se ha tomado el otro debido al funcionamiento interno del comando `Tuples` en Mathematica, ya que este genera automáticamente el orden dado en la definición 22. ◂

3.2.3. Construcción geométrica

Para terminar, vamos a obtener información adicional acerca de los códigos de Reed-Muller binarios desde un punto de vista geométrico. Para ello, se emplea el $\mathbb{F}_2$-espacio vectorial $\mathbb{F}_2^m$ como geometría finita, que se denota por $EG(m, 2)$. Obtenemos así una caracterización geométrica de $\mathcal{RM}(r, m)$.

Definición 32. Dados un subespacio vectorial V de $\mathbb{F}_2^m$ y un punto $\boldsymbol{a} \in EG(m, 2)$, una **variedad afín** que pasa por $\boldsymbol{a}$ y tiene dirección V es la clase de equivalencia $\boldsymbol{a} + V = \{\boldsymbol{a} + \boldsymbol{v} \mid \boldsymbol{v} \in V\}$. Llamaremos **dimensión** de la variedad $\boldsymbol{a} + V$ a la del subespacio vectorial V. Si esta es k, diremos que $\boldsymbol{a} + V$ es una k**-variedad**. ◂

Observación 33. Un subconjunto de $EG(m, 2)$ es una k-variedad si y solo si este es el conjunto de soluciones para un sistema de $m - k$ ecuaciones lineales sobre $\mathbb{F}_2$ en m variables con rango $m - k$. Esto se observa mediante equivalencias teniendo en cuenta que, dado V un $\mathbb{F}_2$-subespacio vectorial de $\mathbb{F}_2^m$ de dimensión k (esto es, lo que hemos definido por código lineal binario de longitud m y dimensión k) con H matriz de control, entonces $\boldsymbol{x}$ es palabra código si y solo si se tiene que $\boldsymbol{x}H^\top = 0$. ◂

Sea F un polinomio booleano de m indeterminadas. Fijado un orden $EG(m, 2) = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ en nuestra geometría finita, asociamos a este, además de la correspondiente palabra binaria $\boldsymbol{f}$, el subconjunto

$$S_F \underset{\text{not.}}{\equiv} \{\boldsymbol{v} \in \mathbb{F}_2^m \mid f(\boldsymbol{v}) = 1\} = \{\boldsymbol{v}_i \mid f_i = 1\} \subseteq EG(m, 2).$$

Mediante este proceso se obtienen todos los subconjuntos de la geometría finita $EG(m, 2)$. Así, diremos que F es el polinomio booleano asociado a S_F y $\boldsymbol{f}$ es la palabra característica asociada a S_F. Nuestro objetivo es describir los códigos de Reed-Muller binarios en términos de las variedades afines de $EG(m, 2)$.

Proposición 34. *Si $S \subseteq EG(m, 2)$ es una k-variedad, el correspondiente polinomio booleano asociado a esta tiene grado $m - k$. Aunque el recíproco no es cierto en general, sí que lo es para monomios booleanos.*

Demostración. La k-variedad S es el conjunto de soluciones para un sistema de $m - k$ ecuaciones lineales en m variables con rango $m - k$ por la observación 33, en las cuales podemos suponer sin pérdida de generalidad que en la parte derecha tenemos un 1 en todos los casos. Una solución de este sistema lo es también de la ecuación que resulta de multiplicar todos los polinomios de la parte izquierda e igualarlos a 1. El polinomio resultante tiene grado $m - k$. En conclusión, como S es el conjunto de soluciones para la ecuación que se sigue de igualar este a 1, el polinomio booleano asociado a S tiene grado $m - k$. El recíproco es trivialmente cierto si se tienen monomios booleanos. Para cualquier otro caso, basta dar un contraejemplo: existen polinomios booleanos F de grado $k \geq 2$ tales que el subconjunto $S_F \subseteq EG(m, 2)$ no es una $(m - k)$-variedad. Sea $F = x_1x_2 + x_3$ un polinomio booleano de 3 variables y grado 2. Por reducción al absurdo, si S_F es una variedad afín, como F es de grado 2, esta es una recta afín. Pero

$$S_F = \{(x_1, x_2, x_3) \in \mathbb{F}_2^3 \mid x_1x_2 + x_3 = 1\} = \{(1, 1, 0), (0, 0, 1), (1, 0, 1), (0, 1, 1)\},$$

lo cual es absurdo, ya que las rectas afines tienen exactamente dos puntos. ∎

Teorema 35. *El código de Reed-Muller binario $\mathcal{RM}(r, m)$ es el subespacio vectorial generado por todas las palabras características asociadas a las variedades afines de $EG(m, 2)$ con dimensión al menos $m - r$.*

Demostración. Sea $\boldsymbol{x}$ la palabra característica asociada a una variedad afín $S_{\boldsymbol{x}}$ de dimensión al menos $m - r$. Supongamos que $F_{\boldsymbol{x}}$ es el polinomio booleano que tiene asociada esta palabra característica. En virtud de la proposición 34, se tiene que este polinomio booleano tiene grado menor o igual que r. Así, por como se definen los códigos de Reed-Muller binarios, se tiene que $\boldsymbol{x} \in \mathcal{RM}(r, m)$. Recíprocamente, sea F el polinomio booleano asociado a una palabra $\boldsymbol{x}_F \in \mathcal{RM}(r, m)$. Sabemos que este tiene grado $s \leq r$. Supongamos que $F = \sum_{i=1}^{l} P_i$, siendo P_i con $i \in \{1, \dots, l\}$ los monomios booleanos en los que se descompone F. Obsérvese que estos tienen grado $\deg(P_i) \leq s$ para todo $i \in \{1, \dots, l\}$. Por la linealidad de nuestra aplicación ψ_m se tiene que $\boldsymbol{x}_F = \sum_{i=1}^{l} \boldsymbol{x}_{P_i}$ es la palabra característica asociada a F. Ahora bien, por la segunda parte de la proposición 34, cada $\boldsymbol{x}_{P_i}$ es la palabra característica asociada a una variedad afín de dimensión $m - \deg(P_i) \geq m - s$. Por tanto, la palabra $\boldsymbol{x}_F$ es suma de palabras características de variedades afines con dimensión al menos $m - s \geq m - r$. ∎

Definición 36. Dado r un entero tal que $0 \leq r \leq m$, se define geométricamente $\mathcal{RM}(r, m)$ como el subespacio vectorial generado por todas las palabras características asociadas a las variedades afines de $EG(m, 2)$ con dimensión al menos $m - r$. ◂

La siguiente es una consecuencia importante del teorema 35, empleando la parte final de la proposición 20, necesaria para entender la decodificación en los códigos de Reed-Muller binarios.

Corolario 37. *Todas las palabras características asociadas a conjuntos que sean* $(r+1)$*-variedades de* $EG(m, 2)$ *son elementos de* $\mathcal{RM}(r, m)^{\perp}$.

3.3. Métodos de codificación y decodificación en códigos de Reed-Muller binarios

Para terminar, vamos a estudiar los métodos de codificación y decodificación propios para códigos de Reed-Muller binarios. Recogemos en una tabla los dos algoritmos que describen estos dos procesos.

Empezamos estableciendo un procedimiento de codificación basado en el resultado general siguiente que permite obtener un orden para los elementos de $\mathbb{F}_2^m$ tal que $\mathcal{RM}(r, m)$ admite una matriz generadora dada en forma estándar.

Proposición 38. *Toda matriz binaria* $G \in Mat_{s\times n}(\mathbb{F}_2)$ *de rango máximo* $s \leq n$ *puede llevarse, realizando operaciones elementales por filas y permutaciones en las columnas, a una matriz dada en forma estándar.*

Demostración. Sea $G = (g_{ij})_{(i,j)\in(\{1,\dots,s\}\times\{1,\dots,n\})} \in \mathrm{Mat}_{s\times n}(\mathbb{F}_2)$. Como el rango de G coincide con el número de filas s, necesariamente en cada una de estas ha de existir al menos un elemento no nulo.

1. Si $g_{11} = 1$, para cada $i \in \{2, \dots, s\}$, sustituimos cada fila i-ésima de G por la fila i-ésima de G menos su primera fila. De esta forma, a través de operaciones elementales por filas, transformamos G en

$$\left(\begin{array}{c|ccc} 1 & * & \cdots & * \\ \hline 0 & & & \\ \vdots & & B & \\ 0 & & & \end{array}\right), \tag{9}$$

   que sigue teniendo rango s, siendo $B \in \mathrm{Mat}_{(s-1)\times(n-1)}(\mathbb{F}_2)$.
2. Si $g_{11} = 0$, ha de existir una columna j de G tal que $g_{1j} = 1$. Permutamos las columnas 1 y j de G entre sí. La matriz que así obtenemos está en las condiciones descritas en el paso 1 anterior.

En cualquier caso, obtenemos una matriz del tipo (9). Estudiamos ahora la posición $(2, 2)$ de esta nueva matriz. Sin pérdida de generalidad, podemos suponer que $g_{22} = 1$. Así, realizando una vez más operaciones elementales por filas, podemos llevar esta matriz a una de la forma

$$\left(\begin{array}{cc|ccc} 1 & 0 & * & \cdots & * \\ 0 & 1 & * & \cdots & * \\ \hline 0 & 0 & & & \\ \vdots & \vdots & & D & \\ 0 & 0 & & & \end{array}\right),$$

que vuelve a ser una matriz de rango s, donde $D \in \mathrm{Mat}_{(s-2)\times(n-2)}(\mathbb{F}_2)$. Reiterando este proceso un total de s veces, obtenemos finalmente una matriz dada en forma estándar. ∎

Corolario 39. *Dado el código de Reed-Muller binario* $\mathcal{RM}(r, m)$*, existe un orden para los elementos de* $\mathbb{F}_2^m$ *tal que* $\mathcal{RM}(r, m)$ *admite una matriz generadora dada en forma estándar.*

Demostración. Basta hacer uso de la proposición 38 con una matriz generadora, teniendo en cuenta que permutar dos columnas de esta entre sí equivale a intercambiar la posición de dos letras en todas las palabras del código $\mathcal{RM}(r, m)$. Por la observación 18, esto no cambia el código de Reed-Muller binario. ∎

Definición 40. Resulta natural referirnos al orden de $\mathbb{F}_2^m$ obtenido por aplicación del corolario 39 como *orden estándar* de $\mathbb{F}_2^m$ respecto de $\mathcal{RM}(r, m)$. Este dista mucho de ser único en general. ◂

La mayor ventaja que tienen los códigos de Reed-Muller binarios es su fácil decodificación por el llamado *algoritmo de Reed*. Este toma como base un método muy práctico y eficiente de decodificación para cierto tipo de códigos lineales, del cual hemos comentado algo al comienzo. Su principal característica es que no emplea síndromes, pues detecta directamente las posiciones donde se han producido los errores.

3.3.1. Algoritmo de Reed

Hay muchas formas de presentar el algoritmo de Reed, pero la mejor manera de hacerlo es en términos de la geometría finita $EG(m, 2)$. Supongamos que, fijado un orden $EG(m, 2) = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$, se ha enviado una palabra código $\boldsymbol{c} \in \mathcal{RM}(r, m)$, a partir de la cual recibimos $\boldsymbol{y} = \boldsymbol{c} + \boldsymbol{e}$. Asumiendo que $\omega(\boldsymbol{e}) \leq 2^{m-r-1} - 1$, este tiene que ser capaz de determinar las posiciones $i \in \{1, 2, \dots, n\}$ en las que se hayan cometido errores durante la transmisión (puesto que, conocidas estas, dado que estamos en el caso binario, la decodificación será trivial). Para ello, reformularemos el problema empleando las 0-variedades de $EG(m, 2)$.

Definición 41. Sea S una k-variedad con palabra característica asociada $\boldsymbol{x}_S$. Recibida $\boldsymbol{y} = \boldsymbol{c} + \boldsymbol{e}$ tal que $\boldsymbol{c} \in \mathcal{RM}(r, m)$, con $\omega(\boldsymbol{e}) \leq 2^{m-r-1} - 1$, la *paridad* de S respecto de $\boldsymbol{y}$ no es más que la paridad, en el sentido binario usual (donde 0 representa par, mientras que 1, impar), de $\langle \boldsymbol{x}_S, \boldsymbol{e} \rangle \equiv \omega(\boldsymbol{x}_S \boldsymbol{e}) \in \mathbb{F}_2$. ◂

En estas circunstancias, por como vienen dadas las palabras características de las 0-variedades, determinar si la i-ésima coordenada de $\boldsymbol{y}$ es correcta o no para cada $i \in \{1, 2, \dots, n\}$ equivale a calcular la paridad correspondiente a la 0-variedad $S = \{\boldsymbol{v}_i\}$. Desafortunadamente, esta no puede ser evaluada directamente. Afortunadamente, tenemos el siguiente resultado, que es consecuencia inmediata del corolario 37.

Proposición 42. *Bajo las condiciones enunciadas, si S es una $(r+1)$-variedad con palabra característica asociada $\boldsymbol{x}_S$, se tiene que la paridad de S respecto de $\boldsymbol{y}$ coincide con la de $\omega(\boldsymbol{x}_S \boldsymbol{y})$.*

La idea consiste en utilizar el conocimiento de las paridades respecto de $\boldsymbol{y}$ con todas las $(r+1)$-variedades para determinar la paridad respecto de $\boldsymbol{y}$ del resto de k-variedades, con $k \leq r$. Para ello, se procede por «lógica mayoritaria». Esto es, dada una k-variedad S para la que conocemos todas las paridades respecto de $\boldsymbol{y}$ en las $(k+1)$-variedades que la contienen, diremos que su paridad respecto de $\boldsymbol{y}$ coincide con aquella que tienen la mayoría de estas variedades afines. El resultado que demuestra la veracidad de este mecanismo es la segunda clave del algoritmo de Reed y requiere de otros dos resultados técnicos.

Lema 43. *Para cada k-variedad $S = \boldsymbol{a} + V$ de $EG(m, 2)$ y cada punto $\boldsymbol{b} \in EG(m, 2) - S$ que consideremos, existe una única variedad afín de dimensión $k+1$ que contiene tanto a S como a $\boldsymbol{b}$.*

Lema 44. *Cada k-variedad de $EG(m, 2)$, con $k < m$, está contenida exactamente en $2^{m-k} - 1$ variedades afines de dimensión $k+1$.*

El primero de estos dos es un análogo al quinto postulado de Euclídes para nuestra geometría finita. Es, por tanto, un resultado de existencia y unicidad, cuya prueba es similar a la que se da para este en un curso de geometría elemental. El segundo, por otro lado, se sigue del anterior haciendo un rápido argumento de combinatoria entre variedades afines.

Teorema 45 (criterio de la lógica mayoritaria, CLM). *Bajo estas condiciones, dada una k-variedad S, con $k \leq r$, se tiene que la paridad de S respecto de $\boldsymbol{y}$ coincide con la que tienen la mayoría de las $(k+1)$-variedades que contienen a S.*

Demostración. Por el lema 44, tenemos que S está contenida en $2^{m-k} - 1$ variedades afines de dimensión $k+1$, donde cada una de estas viene determinada de forma unívoca dando un punto exterior a S en virtud del lema 43. Por hipótesis, dado que el número de errores en $\boldsymbol{y}$ no supera los $2^{m-r-1} - 1$, existen a lo más $2^{m-r-1} - 1$ variedades afines de dimensión $k+1$ que contienen a S determinadas por los puntos exteriores correspondientes a una coordenada incorrecta de $\boldsymbol{y}$. El resto de las $(k+1)$-variedades tienen la propiedad de que no contienen puntos exteriores a S correspondientes a coordenadas erróneas de $\boldsymbol{y}$ por la unicidad probada en el lema 43. En efecto, si alguna de estas variedades afines contiene algún punto exterior a S correspondiente a una coordenada errónea de $\boldsymbol{y}$, necesariamente debería ser una de las anteriores debido a esta unicidad. Así, por la definición 41, estas tienen la misma paridad respecto de $\boldsymbol{y}$ que S. En resumen, por todo lo mencionado, el número de $(k+1)$-variedades con la misma paridad respecto de $\boldsymbol{y}$ que S es al menos de $(2^{m-k} - 1) - (2^{m-r-1} - 1)$. El resultado a partir de aquí se debe a que trivialmente se cumple la desigualdad $2^{m-k} - 2^{m-r-1} \geq 2^{m-r-1}$, pues $k \leq r$. ∎

Ejemplo 46. Supongamos recibida la palabra 01100001, codificada mediante el orden canónico de $\mathbb{F}_2^3$ en $\mathcal{RM}(1, 3)$, donde se ha producido un error. Vamos a decodificarla mediante el algoritmo de Reed. Primero calculamos la paridad de los planos afines de $EG(3, 2) = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_4, \boldsymbol{v}_5, \boldsymbol{v}_6, \boldsymbol{v}_7, \boldsymbol{v}_8\}$ empleando la proposición 42. Se tiene la tabla siguiente:

plano	palabra	paridad	plano	palabra	paridad
$\{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_4\}$	11110000	par	$\{\boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_5, \boldsymbol{v}_8\}$	01101001	impar
$\{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_5, \boldsymbol{v}_6\}$	11001100	impar	$\{\boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_6, \boldsymbol{v}_7\}$	01100110	par
$\{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_7, \boldsymbol{v}_8\}$	11000011	par	$\{\boldsymbol{v}_2, \boldsymbol{v}_4, \boldsymbol{v}_5, \boldsymbol{v}_7\}$	01011010	impar
$\{\boldsymbol{v}_1, \boldsymbol{v}_3, \boldsymbol{v}_5, \boldsymbol{v}_7\}$	10101010	impar	$\{\boldsymbol{v}_2, \boldsymbol{v}_4, \boldsymbol{v}_6, \boldsymbol{v}_8\}$	01010101	par
$\{\boldsymbol{v}_1, \boldsymbol{v}_3, \boldsymbol{v}_6, \boldsymbol{v}_8\}$	10100101	par	$\{\boldsymbol{v}_3, \boldsymbol{v}_4, \boldsymbol{v}_5, \boldsymbol{v}_6\}$	00111100	impar
$\{\boldsymbol{v}_1, \boldsymbol{v}_4, \boldsymbol{v}_5, \boldsymbol{v}_8\}$	10011001	impar	$\{\boldsymbol{v}_3, \boldsymbol{v}_4, \boldsymbol{v}_7, \boldsymbol{v}_8\}$	00110011	par
$\{\boldsymbol{v}_1, \boldsymbol{v}_4, \boldsymbol{v}_6, \boldsymbol{v}_7\}$	10010110	par	$\{\boldsymbol{v}_5, \boldsymbol{v}_6, \boldsymbol{v}_7, \boldsymbol{v}_8\}$	00001111	impar

Ahora, por el criterio de la lógica mayoritaria, se calculan las paridades correspondientes a las rectas afines. Estas son las siguientes:

recta	paridad	recta	paridad	recta	paridad	recta	paridad
$\{\boldsymbol{v}_1, \boldsymbol{v}_2\}$	par	$\{\boldsymbol{v}_2, \boldsymbol{v}_3\}$	par	$\{\boldsymbol{v}_3, \boldsymbol{v}_5\}$	impar	$\{\boldsymbol{v}_4, \boldsymbol{v}_8\}$	par
$\{\boldsymbol{v}_1, \boldsymbol{v}_3\}$	par	$\{\boldsymbol{v}_2, \boldsymbol{v}_4\}$	par	$\{\boldsymbol{v}_3, \boldsymbol{v}_6\}$	par	$\{\boldsymbol{v}_5, \boldsymbol{v}_6\}$	impar
$\{\boldsymbol{v}_1, \boldsymbol{v}_4\}$	par	$\{\boldsymbol{v}_2, \boldsymbol{v}_5\}$	impar	$\{\boldsymbol{v}_3, \boldsymbol{v}_7\}$	par	$\{\boldsymbol{v}_5, \boldsymbol{v}_7\}$	impar
$\{\boldsymbol{v}_1, \boldsymbol{v}_5\}$	impar	$\{\boldsymbol{v}_2, \boldsymbol{v}_6\}$	par	$\{\boldsymbol{v}_3, \boldsymbol{v}_8\}$	par	$\{\boldsymbol{v}_5, \boldsymbol{v}_8\}$	impar
$\{\boldsymbol{v}_1, \boldsymbol{v}_6\}$	par	$\{\boldsymbol{v}_2, \boldsymbol{v}_7\}$	par	$\{\boldsymbol{v}_4, \boldsymbol{v}_5\}$	impar	$\{\boldsymbol{v}_6, \boldsymbol{v}_7\}$	par
$\{\boldsymbol{v}_1, \boldsymbol{v}_7\}$	par	$\{\boldsymbol{v}_2, \boldsymbol{v}_8\}$	par	$\{\boldsymbol{v}_4, \boldsymbol{v}_6\}$	par	$\{\boldsymbol{v}_6, \boldsymbol{v}_8\}$	par
$\{\boldsymbol{v}_1, \boldsymbol{v}_8\}$	par	$\{\boldsymbol{v}_3, \boldsymbol{v}_4\}$	par	$\{\boldsymbol{v}_4, \boldsymbol{v}_7\}$	par	$\{\boldsymbol{v}_7, \boldsymbol{v}_8\}$	par

De la misma forma, si obtenemos la paridad de cada punto, se observa que el único impar es el $\boldsymbol{v}_5$. En definitiva, concluimos que el único error dado durante la transmisión de información se encuentra en la quinta posición. Por tanto, decodificamos la palabra recibida como 01101001. ◂

Dados los parámetros r y m del código de Reed-Muller binario, calculamos la matriz generadora dada en forma estándar con la que vamos a trabajar. Hecho esto, para cada palabra $\boldsymbol{x}$ a codificar, procedemos a su codificación tal y como ya se explicó en su momento al comienzo del artículo.

Algoritmo 1 (Cálculo de la matriz estándar).

subrutina MATRIZ ESTÁNDAR(r,m)
  calcular matriz generadora usando la proposición 29
  si la matriz ya está en forma estándar, **entonces**
    no hacer nada
  en caso contrario
    aplicar procedimiento de la proposición 38
  fin si
  devolver la matriz G dada en forma estándar
fin subrutina

Algoritmo 2 (Codificación).

subrutina CODIFICAR(G,$\boldsymbol{x}$)
  devolver el producto $\boldsymbol{x}G$
fin subrutina

Dado el código de Reed-Muller binario $\mathcal{RM}(r, m)$, recibida una palabra $\boldsymbol{y}$ arbitraria, se describen los pasos a seguir para su decodificación mediante el algoritmo de Reed.

Algoritmo 3 (Decodificación).

subrutina DECODIFICAR($\boldsymbol{y}$,r,m)
  si la palabra $\boldsymbol{y}$ tiene más de $2^{m-r-1} - 1$ errores, **entonces**
    no se puede decodificar de manera única
  en caso contrario
    calcular la paridad respecto de $\boldsymbol{y}$ para las $(r+1)$-variedades de $EG(m, 2)$ por la proposición 42
    para k desde r hasta 0 **hacer**
      calcular la paridad respecto de $\boldsymbol{y}$ para todas las k-variedades de $EG(m, 2)$ usando el CLM
    fin para
    corregir las coordenadas de $\boldsymbol{y}$ correspondientes a todas las 0-variedades impares respecto de $\boldsymbol{y}$
    devolver la palabra corregida
  fin si
fin subrutina

Referencias

[1] De Arriba De La Hera, Andoni. *Códigos de Reed-Muller*. Trabajo de Fin de Grado. Zientzia eta Teknologia Fakultatea-Facultad de Ciencia y Tecnología (ZTF-FCT), Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU), 2016. URL: `http://hdl.handle.net/10810/20121`.

[2] Iranzo Aznar, María Jesús y Pérez Monasor, Francisco. *LECCIONES de Elementos de Álgebra. Aplicaciones*. Facultad de Matemáticas, Universidad de Valencia. URL: `http://www.uv.es/iranzo/lecciones_de_codigos.pdf`.

[3] Muller, David Eugene. «Application of Boolean algebra to switching circuit design and to error detection». En: *Transactions of the IRE Professional Group on Electronic Computers* EC-3.3 (sep. de 1954), págs. 6-12. ISSN: 2168-1740. `https://doi.org/10.1109/IREPGELC.1954.6499441`.

[4] Reed, Irving Stoy. «A class of multiple-error-correcting codes and the decoding scheme». En: *Transactions of the IRE Professional Group on Information Theory* 4.4 (1954), págs. 38-49. ISSN: 2168-2690. `https://doi.org/10.1109/TIT.1954.1057465`.

[5] Shannon, Claude Elwood. «A mathematical theory of communication». En: *The Bell System Technical Journal* 27 (1948), págs. 379-423, 623-656. ISSN: 0005-8580. `https://doi.org/10.1002/j.1538-7305.1948.tb01338.x`.

A. Anexo: algunos programas en Mathematica

Para terminar, se incluye el archivo con los programas en Mathematica (junto con un PDF generado a partir de este) que son útiles para llevar a cabo la codificación y decodificación en códigos de Reed-Muller binarios. En particular, estamos interesados en el código $\mathcal{RM}(1,5)$ utilizado por la NASA cuando se llevo a cabo la misión del Mariner 9 (véase el ejemplo 47). Todos estos programas tienen como objetivo realizar los procedimientos de codificación y decodificación de manera eficaz para dicho código en especial, aunque resulten válidos para todos los de la forma $\mathcal{RM}(1,m)$ con m número natural arbitrario. Ambos archivos se pueden encontrar aquí:

Programa en Mathematica: `https://temat.es/articulo/2019-p45/a-anexocodrm-nb` (`anexocodrm.nb`)

Código en PDF: `https://temat.es/articulo/2019-p45/a-anexocodrm-pdf` (`anexocodrm.pdf`)

Ejemplo 47. Vamos a aplicar estos algoritmos para comprobar que $\mathcal{RM}(1,5)$ es el código de Reed-Muller binario que empleó la sonda Mariner 9 para obtener las primeras fotografías de la superficie marciana. Lo primero es observar que este es un código lineal de longitud 32. Calculamos una matriz generadora escribiendo en Mathematica la instrucción correspondiente, obteniendo así

$$\begin{pmatrix}
1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1\\
0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1\\
0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1\\
0&0&0&0&1&1&1&1&0&0&0&0&1&1&1&1&0&0&0&0&1&1&1&1&0&0&0&0&1&1&1&1\\
0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1
\end{pmatrix}$$

que tiene rango máximo 6 (coincide con el número de filas). Por consiguiente, esta es también la dimensión del código. La distancia mínima se obtiene a partir de la fórmula dada en la proposición 27, y es 16. En definitiva, este es el código de Reed-Muller binario mencionado al comienzo del artículo. Vamos a hacer un ejemplo de decodificación. Supongamos que se está enviando información codificada en $\mathcal{RM}(1,5)$ mediante el orden estándar de $\mathbb{F}_2^5$ y que recibimos la palabra 10011101010001101110010010001001 que no tiene más de siete errores. Vamos a decodificar esta mediante el algoritmo de Reed. Escribiendo la instrucción correspondiente en Mathematica, obtenemos la palabra 10010001111001100110011010011001 (obsérvese que hemos sido capaces de corregir los siete errores que se habían dado durante la transmisión de información). Así, como se sabe que esta ha sido codificada mediante una matriz dada en forma estándar y nuestro código es de dimensión 6, como ya hemos comprobado, la palabra original enviada es 100100. ◂

TEMat

Distribuciones tipo fase en un estudio de fiabilidad

Christian José Acal González
Departamento de Estadística e Investigación Operativa, Universidad de Granada, e Instituto de Matemáticas de la Universidad de Granada (IEMath-GR)
chracal@correo.ugr.es

Juan Eloy Ruiz Castro
Departamento de Estadística e Investigación Operativa, Universidad de Granada, e Instituto de Matemáticas de la Universidad de Granada (IEMath-GR)
jeloy@ugr.es

Ana María Aguilera del Pino
Departamento de Estadística e Investigación Operativa, Universidad de Granada, e Instituto de Matemáticas de la Universidad de Granada (IEMath-GR)
aaguiler@ugr.es

Resumen: Hoy en día, el análisis de fiabilidad está presente en cualquier área de conocimiento donde se esté interesado en comprobar el tiempo de vida (o, análogamente, el tiempo de fallo) de un sistema dado. A modo de ejemplo, esta disciplina es ampliamente utilizada en estudios de ingeniería donde el objetivo fundamental es garantizar la calidad y el buen funcionamiento de los aparatos. La rama encargada de estudiar y analizar los tiempos de fallos de estos sistemas es la estadística. La estadística pretende ajustar los datos experimentales medidos en estos sistemas mediante el uso de distribuciones de probabilidad. Un enfoque relativamente nuevo, y que cada vez está teniendo más importancia en la actualidad, es a través de las distribuciones tipo fase, cuya flexibilidad y propiedades hacen que esta clase de distribuciones sea una buena candidata para obtener un ajuste más riguroso.

Abstract: Nowadays, the reliability analysis is present in any knowledge area where one is interested in checking the lifetime (or, similarly, the failure time) of a given system. As an example, this discipline is widely used in engineering studies where the main objective is to ensure the quality and the appropriate operation of the devices. The branch of knowledge responsible for studying and analyzing the failure times of these systems is statistics. The purpose of statistics is to fit the experimental data measured in these systems through the use of probability distributions. A relatively new approach, which is having more significance in currently, is through phase type distributions, whose flexibility and properties make them good candidates to obtain a better fit.

Palabras clave: fiabilidad, tiempos de fallo, distribuciones de probabilidad, análisis gráfico, distribución tipo fase, ajuste estadístico.

MSC2010: 62J05, 62N05, 62P99.

Recibido: 28 de noviembre de 2018.
Aceptado: 26 de abril de 2019.

Agradecimientos: Los autores agradecen el apoyo del Ministerio de Economía y Competitividad de España en el proyecto MTM2017-87708-P, también respaldado por el programa FEDER.

Referencia: Acal González, Christian José; Ruiz Castro, Juan Eloy y Aguilera del Pino, Ana María. «Distribuciones tipo fase en un estudio de fiabilidad». En: *TEMat*, 3 (2019), págs. 63-74. issn: 2530-9633. url: https://temat.es/articulo/2019-p63.

1. Introducción

El análisis de fiabilidad es una materia multidisciplinar que ha sido desarrollada en diferentes contextos dentro de la ingeniería (eléctrica, mecánica, química, civil) y que tiene múltiples aplicaciones, entre otras áreas, en supervivencia y medicina. Desde el punto de vista general, la fiabilidad comprende el conjunto de operaciones utilizadas para el buen funcionamiento y seguridad de los sistemas (un sistema puede ser una red eléctrica, cualquier electrodoméstico, un coche, un ser vivo, etc.). Estos sistemas funcionan un determinado tiempo influenciados por ciertas condiciones ambientales específicas que no se pueden controlar y sometidos a un continuo desgaste. El conjunto de todas estas condiciones o variables afecta al funcionamiento de los sistemas, provocando que dichos sistemas fallen de manera aleatoria. A raíz de esto, y haciendo hincapié en que la ocurrencia de fallos se produce aleatoriamente, la teoría de la probabilidad juega un papel determinante en el cálculo de la fiabilidad de componentes y sistemas. Está claro que todas las unidades de un cierto tipo, fabricadas y operando en condiciones similares, no fallarán a la vez, sino que habrá diferencias entre los tiempos de fallo. Consecuentemente, estos tiempos de fallo obedecen a distribuciones de probabilidad que pueden ser o no conocidas, y que permiten calcular la probabilidad de fallo de las unidades. Se debe destacar que en un estudio de fiabilidad existen múltiples ocasiones en las que el objetivo no es estudiar los tiempos de fallo en sí, sino otra variable que, aunque esté altamente relacionada con los tiempos de fallo, trate un problema distinto. Por ejemplo, en los artículos de Long *et al.* [8] y Luo *et al.* [9] se realiza un estudio de los datos experimentales correspondientes al voltaje de fallo de un tipo de memorias concretas (estos aparatos dejan de funcionar tras aplicarles un cierto voltaje, pero paralelamente han estado funcionando un determinado tiempo).

En consecuencia, la estadística y, más concretamente, la probabilidad proporcionan una serie de herramientas para el cálculo y la mejora de la fiabilidad y suministran una definición cuantitativa de la misma. Aunque existen distintas definiciones para aclarar este concepto, la definición usual es la que sigue: «Probabilidad de que un dispositivo efectuará la función para la que fue construido hasta un momento dado bajo condiciones específicas de uso» [4]. Por lo tanto, podemos concluir que la fiabilidad es la probabilidad de que un sistema se comporte adecuadamente durante un tiempo establecido y, en cierto sentido, podemos verla como una medida de calidad: cuanto más tiempo funcione, mejor. Sin embargo, dejando a un lado los conceptos típicos de ingeniería para centrarnos en la teoría de la probabilidad, que es lo que realmente nos compete, saber qué distribución probabilística siguen los tiempos de fallo nos permite, en primer lugar, determinar las principales características de operación de dicho sistema y, en segundo lugar, conocer cómo funcionará el sistema en el futuro.

En muchos campos donde se hace uso de la estadística aplicada, la distribución normal es el punto de partida natural para modelizar cualquier variable aleatoria de interés. La razón fundamental de su uso se debe a las buenas propiedades que presenta esta distribución (es simétrica, unimodal, asintótica, etc.) y por proporcionar la base para la estadística inferencial clásica por su relación con el teorema central del límite. Sin embargo, en el ámbito de la fiabilidad, donde es habitual trabajar con variables que toman valores positivos, la distribución normal tiene menor interés. En la década de los 50, Epstein y Sobel empezaron a trabajar con la distribución exponencial como modelo probabilístico para analizar el tiempo de vida de unos dispositivos [6]. Una razón fundamental de la popularidad e importancia de la distribución exponencial para su uso en el ámbito de la fiabilidad (probablemente la más usada) es su simplicidad y versatilidad. Por ejemplo, esta distribución es muy útil cuando los datos que se utilizan son escasos, cuando se estudian sistemas en régimen estacionario (el tiempo de funcionamiento crece indefinidamente) o cuando se tienen sistemas complejos en los cuales no es fácil aplicar técnicas analíticas. Sin embargo, la distribución exponencial se ha quedado «corta» en la actualidad y cada vez es más común utilizar otros tipos de distribuciones, como pueden ser la distribución de Weibull, la distribución gamma, la distribución de valores extremos, la distribución log-normal, etc. No obstante, y a pesar de que estas distribuciones son bien conocidas y ampliamente utilizadas de manera exitosa en diversas ramas de la ciencia, la ingeniería y la medicina (por ejemplo, y para seguir con los dispositivos mencionados por Long *et al.* [8] y Luo *et al.* [9], en el artículo de Pan *et al.* [12] se emplea la distribución de Weibull para modelizar los datos experimentales medidos en estos tipos de aparatos), a veces el ajuste que se obtiene no es del todo preciso y, en consecuencia, es necesario plantearse un enfoque diferente que mejore la rigurosidad del ajuste. Bajo este contexto, en los últimos años se ha venido utilizando con bastante frecuencia un enfoque basado en las distribuciones tipo fase, cuya flexibilidad y propiedades hacen que sean unas buenas candidatas para obtener un mejor ajuste de los datos experimentales. A modo de ejemplo, y continuando con el tipo de memorias mencionadas a lo largo de la introducción, en el artículo de Acal *et al.* [1] se demuestra tras un estudio pormenorizado que el ajuste que se obtiene con las distribuciones tipo fase es más riguroso que el que se logra con la distribución de Weibull.

Las distribuciones tipo fase (PHD), las cuales fueron introducidas y analizadas en detalle por Neuts [10, 11], constituyen una clase de distribuciones no negativas que hacen posible modelizar problemas complejos con resultados bien estructurados gracias a su forma algebraico-matricial. Debido a sus valiosas propiedades, muchas variedades de esta clase de distribuciones han sido consideradas en diversas ramas del conocimiento y aplicadas en estudios de fiabilidad, procesos de renovación, teoría de colas y análisis de supervivencia [13, 14]. Casos particulares de las distribuciones tipo fase son las distribuciones exponencial, Erlang, Erlang generalizada, hipergeométrica y coxiana, entre otras. De hecho, no solo algunas distribuciones de probabilidad muy conocidas son distribuciones tipo fase, sino que cualquier distribución de probabilidad no negativa puede ser aproximada tanto como se desee mediante una PHD.

Finalmente, una vez que tenemos controlada la distribución, podemos extraer las principales propiedades del sistema a partir de las funciones que caracterizan a una distribución (función de densidad, función de distribución, etc.). En un análisis de fiabilidad habitual destaca el papel que juegan la función de fiabilidad y la función de riesgo (también conocida como función razón de fallo), y la aparición de datos censurados, que son aquellos tiempos de fallo de algunas unidades experimentales que se desconocen por algún motivo (por ejemplo, en un estudio de medicina donde estamos interesados en ver cómo afecta la dosis de un nuevo medicamento en el tiempo de vida de pacientes con cáncer, si un paciente deja el estudio, ya sea porque se ha mudado de ciudad o porque no quiere seguir perteneciendo a dicho estudio, este valor será desconocido y, por tanto, censurado). Aunque en el presente artículo no vamos a tratar los datos censurados, hay que tener en cuenta que es un concepto muy común en estos tipos de estudios y que deben ser tratados de manera correcta.

Este artículo está estructurado como sigue. En la sección 2 se describen las funciones típicas de un estudio de fiabilidad y las expresiones que adoptan estas funciones en cada una de las distribuciones más comunes, además de detallar una técnica ampliamente utilizada para ajustar distribuciones de probabilidad a los tiempos de fallo. La sección 3 está centrada en detallar y en definir brevemente las propiedades más importantes de las distribuciones tipo fase. Un ejemplo de simulación con el paquete estadístico R, donde se discute la modelización de distintas distribuciones a los datos simulados, se puede ver en la sección 4. Finalmente, tras las secciones mencionadas, figura un apartado dedicado a las conclusiones obtenidas en el presente trabajo.

2. Metodología básica en fiabilidad

En esta sección se definen una serie de medidas que se utilizan en el contexto de la fiabilidad y análisis de supervivencia. Se debe notar que para estas definiciones se va a considerar el tiempo de fallo del sistema, pero son extensibles para cualquier otra variable que esté relacionada con el tiempo de fallo.

2.1. Función de distribución, fiabilidad y razón de fallo en fiabilidad

La *función de distribución* de una variable aleatoria T, definida sobre el eje real positivo, se define como la función $F(t) = \mathbb{P}[T \leq t]$. Asimismo, su *función de densidad* se define como

$$f(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}[t \leq T < t + \Delta t]}{\Delta t} = \frac{\mathrm{d}F(t)}{\mathrm{d}t}.$$

Dado que la variable aleatoria T representa el tiempo de fallo (tiempo de vida), la *función de fiabilidad* representa la probabilidad de que no ocurra un fallo en el intervalo $(0, t)$ o, lo que es lo mismo, la probabilidad de que un sistema sobreviva en el intervalo $(0, t)$, de ahí que también sea conocida como *función de supervivencia*. Se representa como sigue:

$$R(t) = \mathbb{P}[T > t] = 1 - F(t).$$

Cabe recalcar que la gráfica de esta función es decreciente, ya que es la complementaria de la función de distribución; se inicia en $R(0) = 1$, y tiende a cero cuando $t \to \infty$. Otra medida que es muy utilizada en este ámbito es el *tiempo medio de fallo*, que se define como la esperanza de la variable,

$$\mathbb{E}[T] = \int_0^\infty t f(t)\,\mathrm{d}t = \int_0^\infty R(t)\,\mathrm{d}t.$$

Esta relación se obtiene aplicando el teorema de Fubini, como se puede ver a continuación:

$$\int_0^\infty R(t)\,\mathrm{d}t = \int_0^\infty \int_t^\infty f(u)\,\mathrm{d}u\,\mathrm{d}t = \int_0^\infty \int_0^u f(u)\,\mathrm{d}t\,\mathrm{d}u = \int_0^\infty f(u) \int_0^u \mathrm{d}t\,\mathrm{d}u = \int_0^\infty u f(u)\,\mathrm{d}u.$$

Por otro lado, la *función razón de fallo* o *razón de riesgo* se interpreta como la razón instantánea de fallo, y se define a partir de la siguiente relación:

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}[t \leq T < t + \Delta t \mid T > t]}{\Delta t} = \frac{f(t)}{R(t)}.$$

Asimismo, dado que la función de densidad se puede expresar en términos de la función de fiabilidad, la expresión anterior se puede escribir

$$h(t) = -\frac{\mathrm{d}R(t)/\,\mathrm{d}t}{R(t)},$$

e integrando esta ecuación diferencial en el intervalo $(0, t)$ con la condición inicial $R(0) = 1$ se tiene que

$$R(t) = \mathrm{e}^{-H(t)},$$

siendo

$$H(t) = \int_0^t h(x)\,\mathrm{d}x$$

la *tasa de fallo acumulada* hasta el tiempo t. Esta relación muestra que la función de fallo caracteriza la distribución.

2.2. Expresiones de algunas distribuciones en fiabilidad

En el cuadro 1 se pueden observar las expresiones de las medidas definidas en el apartado anterior que adoptan algunas de las distribuciones más comunes en un estudio de fiabilidad. Además, se indican los límites de cada uno de los parámetros de los que depende cada una de las distribuciones definidas, donde $t > 0$ en todas las distribuciones salvo en la log-normal, que satisface $t \in \mathbb{R}$.

Como se comentó en la introducción del presente artículo, además de estas distribuciones, existen otras distribuciones de probabilidad que tienen un rol importante en el campo de la fiabilidad. Sin embargo, como el objetivo no es estudiar todas las distribuciones que se emplean para modelizar los tiempos de fallo, simplemente se procede a mencionar el nombre de estas distribuciones: Erlang, gamma, chi-cuadrado, Pareto, etc.

2.3. Análisis gráfico

Cuando la estimación de los parámetros utilizando el método de máxima verosimilitud presenta serias dificultades de cálculo, como es el caso de la distribución de Weibull, se emplea un análisis gráfico que facilita el ajuste de un conjunto de datos observados. Esta técnica gráfica es una técnica paramétrica que se basa en el principio de mínimos cuadrados, y suele ser aplicada debido a su sencillez y porque permite una primera idea gráfica del ajuste.

Grosso modo, lo que se hace es construir una nube de puntos a partir de los tiempos de fallo observados y a este conjunto de datos se le ajusta una recta por el criterio de mínimos cuadrados. La forma final de la nube de puntos dependerá de la distribución de probabilidad que se considere. Si el ajuste es bueno, lo que se valora en función del coeficiente de determinación, se acepta que el conjunto de valores observados sigue la distribución considerada y se calculan los parámetros de dicha distribución a partir de la recta estimada por mínimos cuadrados.

Cuadro 1: Expresiones que adoptan las distribuciones consideradas.

	Exponencial	Weibull	Valores extremos (Gumbel)	Log-normal
$F(t)$	$1 - \mathrm{e}^{-\lambda t}$	$1 - \mathrm{e}^{-(\lambda t)^{\beta}}$	$1 - \mathrm{e}^{-\gamma \mathrm{e}^{\beta t}}$	$\phi\left(\frac{\ln(t) - \mu}{\sigma}\right)$, con ϕ f. distr. $N(0,1)$
$f(t)$	$\lambda \mathrm{e}^{-\lambda t}$	$\beta\lambda(\lambda t)^{\beta-1}\mathrm{e}^{-(\lambda t)^{\beta}}$	$(\beta\gamma)\mathrm{e}^{\beta t - \gamma \mathrm{e}^{\beta t}}$	$\frac{1}{\sigma t\sqrt{2\pi}}\mathrm{e}^{-\frac{(\ln(t)-\mu)^2}{2\sigma^2}}$
$R(t)$	$\mathrm{e}^{-\lambda t}$	$\mathrm{e}^{-(\lambda t)^{\beta}}$	$\mathrm{e}^{-\gamma \mathrm{e}^{\beta t}}$	$1 - \phi\left(\frac{\ln(t) - \mu}{\sigma}\right)$
$\mathbb{E}[T]$	$1/\lambda$	$\frac{1}{\lambda}\Gamma\left(\frac{1}{\beta} + 1\right)$, con Γ f. gamma	$\int_0^{\infty} \mathrm{e}^{-\gamma \mathrm{e}^{\beta t}}\,\mathrm{d}t$	$\mathrm{e}^{\mu + \frac{\sigma^2}{2}}$
$h(t)$	λ	$(\beta\lambda)(\lambda t)^{\beta-1}$	$\beta\gamma \mathrm{e}^{\beta t}$	$f(\ln(t))/R(\ln(t))$
Parám.	$\lambda > 0$	$\lambda > 0, \beta > 0$	$\gamma = \lambda^{\beta}, \lambda > 0, \beta > 0$	$\mu \in \mathbb{R}, \sigma > 0$

2.3.1. Análisis gráfico en la distribución exponencial

Puesto que el análisis estadístico habitual que se utiliza en datos experimentales correspondientes a los tiempos de fallo de un sistema es mediante la distribución exponencial, se va a explicar cómo se aplica esta técnica y cómo se obtienen los parámetros de esta distribución en la práctica. Esto se puede extender para el caso de considerar otra distribución cualquiera.

Se parte de la función de fiabilidad de la distribución exponencial,

$$R(t) = \mathrm{e}^{-\lambda t}, \quad t \geq 0, \quad \lambda > 0,$$

y se opera en esta expresión hasta conseguir una relación lineal. En consecuencia, si se toman logaritmos neperianos, queda

$$\ln(R(t)) = -\lambda t, \quad t \geq 0, \quad \lambda > 0,$$

o lo que es lo mismo,

$$\ln(1 - F(t)) = -\lambda t, \quad t \geq 0, \quad \lambda > 0.$$

Teniendo en cuenta esta última expresión, si despejamos el signo negativo que precede a t, la nube de puntos que hay que representar tiene la siguiente forma:

$$\{(t_i, -\ln(1 - p_i)), i = 1, 2, \dots, n\},$$

siendo t_i los tiempos de fallo observados (ordenados de menor a mayor) y p_i los valores que toma la función de distribución empírica en los t_i. En la literatura se han introducido varias elecciones posibles para estos valores p_i, como, por ejemplo, cualquiera de las tres aproximaciones siguientes, siendo la segunda opción la más utilizada:

$$p_i = \frac{i}{n}, \qquad p_i = \frac{i - 0{,}5}{n}, \qquad p_i = \frac{i}{n+1}.$$

A este conjunto de datos se le ajusta una recta del tipo $y = ax + b$ por el principio de mínimos cuadrados y se estudia el ajuste. Si el ajuste es bueno, se acepta que el conjunto de valores observados sigue una distribución exponencial y, finalmente, a partir de la ecuación de la recta, se calcula la estimación del parámetro de la distribución, el cual en este caso coincide con la pendiente de la recta, es decir,

$$\hat{\lambda} = a.$$

2.3.2. Análisis gráfico en otras distribuciones

Extrapolando la teoría explicada en el anterior apartado, se pueden estimar los parámetros de cualquier distribución que se desee. En la cuadro 2 figuran las estimaciones de los parámetros de las distribuciones detalladas en el cuadro 1 aplicando la técnica del *análisis gráfico*, así como las nubes de puntos que hay que suponer en cada caso para calcular dichas estimaciones.

Cuadro 2: Estimaciones y nubes de puntos para cada distribución considerada.

	Nube de puntos	Estimación de parámetros a partir de $y = ax + b$
Weibull	$\{(\ln(t_i), \ln(-\ln(1-p_i)))\}$	$\hat{\beta} = a;\ \hat{\lambda} = \mathrm{e}^{b/a}$
Valores extremos	$\{(t_i, \ln(-\ln(1-p_i)))\}$	$\hat{\beta} = a;\ \hat{\gamma} = \mathrm{e}^{b}$
Log-normal	$\{(\ln(t_i), \phi^{-1}(p_i))\}$	$\hat{\sigma} = \dfrac{1}{a};\ \hat{\mu} = -\dfrac{b}{a}$

3. Distribuciones tipo fase

Como se ha comentado en la introducción del presente artículo, las distribuciones tipo fase fueron introducidas por Neuts en 1975. Estas distribuciones pueden ser definidas en el ámbito de cadenas de Markov y poseen propiedades interesantes (como, por ejemplo, la falta de memoria parcial o las propiedades de clausura) que hacen que dichas distribuciones sean consideradas en diversas ramas de la ciencia o ingeniería.

Una PHD se define como la distribución del tiempo hasta la absorción en una cadena de Markov con un estado o clase absorbente. El espacio de estados viene dado por un número general m de estados transitorios, donde la probabilidad de estar inicialmente en el estado i es α_i, y un estado absorbente, $m+1$ (se considera que la cadena inicialmente no está en el estado absorbente, $\alpha_{m+1} = 0$). Además, la intensidad de transición del estado i al estado j viene dada por q_{ij} para $i \neq j$, y si $i = j$ entonces $q_{ii} = -\sum_{j=1, i\neq j}^{m+1} q_{ij}$. Una distribución tipo fase se representa, considerando solo los estados transitorios de la cadena de Markov asociada, a través del par $(\boldsymbol{\alpha}, \boldsymbol{T})$, siendo $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)$ y $\boldsymbol{T} = (q_{ij})_{i,j=1,\dots,m}$.

La función de distribución de una PHD está dada por

$$F(t) = 1 - \boldsymbol{\alpha}\mathrm{e}^{\boldsymbol{T}t}\boldsymbol{e}, \quad t \geq 0,$$

donde $\boldsymbol{e}$ es un vector columna de unos con el orden apropiado. A partir de esta definición se pueden obtener de forma inmediata las distintas expresiones definidas en la sección 2.1. Sin embargo, y como se utilizarán en la siguiente sección, se procede a mostrar seguidamente la expresión de la función de fiabilidad y de la función razón de fallo, respectivamente, para mayor claridad:

$$R(t) = \boldsymbol{\alpha}\mathrm{e}^{\boldsymbol{T}t}\boldsymbol{e}, \quad t \geq 0, \qquad \text{y} \qquad h(t) = \frac{\boldsymbol{\alpha}\mathrm{e}^{\boldsymbol{T}t}\boldsymbol{T}^0}{\boldsymbol{\alpha}\mathrm{e}^{\boldsymbol{T}t}\boldsymbol{e}}, \quad t \geq 0,$$

siendo $\boldsymbol{T}^0 = -\boldsymbol{T}\boldsymbol{e}$ el vector de intensidad de transición desde un estado transitorio hasta un estado de absorción.

Las distribuciones tipo fase son importantes no solo por su estructura, sino también por las buenas propiedades que permiten la aplicabilidad e interpretación de los resultados de manera sencilla. Las propiedades de estas distribuciones han sido estudiadas exhaustivamente en los últimos años. De hecho, en el libro de He [7] puede verse un estudio actual en el que se revisan las propiedades más esenciales e importantes de las PHD. Sin embargo, la razón fundamental por la cual esta clase de distribuciones es tan atractiva en estos tipos de datos es debido a que la clase PHD es densa en el conjunto de distribuciones de probabilidad no negativas [2]. Este resultado implica que cualquier distribución de probabilidad no negativa puede ajustarse tanto como se desee mediante una PHD, por lo que dicha flexibilidad hace que las PHD sean unas buenas candidatas para obtener un mejor ajuste.

No obstante, a continuación se enumeran otras propiedades interesantes e importantes:

1. Constituyen una clase de distribuciones no negativas que permite describir las principales medidas asociadas en una forma algorítmica.
2. La clase de PHD es cerrada bajo una serie de operaciones: mínimo, máximo, suma, etc.
3. Generalizan un gran número de distribuciones conocidas:
   - Distribución exponencial: $F(t) = 1 - \mathrm{e}^{-\lambda t}$ para $t \geq 0$, $\boldsymbol{\alpha} = 1$, $\boldsymbol{T} = -\lambda$ y $m = 1$.
   - Distribución de Erlang: $F(t) = 1 - \sum_{j=0}^{m-1} \mathrm{e}^{-\lambda t}(\lambda t)^j/j!$ para $t \geq 0$, $m \geq 1$ y $\lambda > 0$,

$$\boldsymbol{\alpha} = (1, 0, \dots, 0, 0), \quad \boldsymbol{T} = \begin{pmatrix} -\lambda & \lambda & & \\ & -\lambda & \ddots & \\ & & \ddots & \lambda \\ & & & -\lambda \end{pmatrix}_{m\times m}.$$

   - Distribución hipoexponencial: $F(t) = 1 - \sum_{x=0}^{t} \sum_{i=1}^{m} \mathrm{e}^{-\lambda_i x} \left(\prod_{j=1;j\neq i}^{m} \frac{\lambda_j}{\lambda_j - \lambda_i} \right)$ para $t \geq 0$, $\lambda_i \neq \lambda_j$ con $i \neq j$,

$$\boldsymbol{\alpha} = (1, 0, \dots, 0, 0), \quad \boldsymbol{T} = \begin{pmatrix} -\lambda_1 & \lambda_1 & & \\ & -\lambda_2 & \ddots & \\ & & \ddots & \lambda_{m-1} \\ & & & -\lambda_m \end{pmatrix}_{m\times m}.$$

   - Distribución hiperexponencial: $F(t) = 1 - \sum_{i=1}^{m} \alpha_i (1 - \mathrm{e}^{-\lambda_i t})$ para $t \geq 0$,

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_m), \quad \boldsymbol{T} = \begin{pmatrix} -\lambda_1 & & & \\ & -\lambda_2 & & \\ & & \ddots & \\ & & & -\lambda_m \end{pmatrix}_{m\times m}.$$

   - Distribución coxiana: $F(t) = 1 - \sum_{x=0}^{t} \sum_{i=1}^{m} \mathrm{e}^{-\lambda_i x} \left(\prod_{j=1;j\neq i}^{m} \frac{g_j \lambda_j}{\lambda_j - \lambda_i} \right)$ para $t \geq 0$, $\lambda_i \neq \lambda_j$ con $i \neq j$,

$$\boldsymbol{\alpha} = (1, 0, \dots, 0, 0), \quad \boldsymbol{T} = \begin{pmatrix} -\lambda_1 & g_1\lambda_1 & & \\ & -\lambda_2 & \ddots & \\ & & \ddots & g_{m-1}\lambda_{m-1} \\ & & & -\lambda_m \end{pmatrix}_{m\times m}.$$

   - Distribución coxiana generalizada: $F(t) = 1 - \sum_{i=1}^{m} \alpha_i \left(\sum_{x=0}^{t} \sum_{i=1}^{m} \mathrm{e}^{-\lambda_i x} \left(\prod_{j=1;j\neq i}^{m} \frac{g_j \lambda_j}{\lambda_j - \lambda_i} \right) \right)$ para $t \geq 0$, $\lambda_i \neq \lambda_j$ con $i \neq j$,

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_m), \quad \boldsymbol{T} = \begin{pmatrix} -\lambda_1 & g_1\lambda_1 & & \\ & -\lambda_2 & \ddots & \\ & & \ddots & g_{m-1}\lambda_{m-1} \\ & & & -\lambda_m \end{pmatrix}_{m\times m}.$$

Finalmente, en lo que a las PHD se refiere, cabe decir que para la estimación de los parámetros de estas distribuciones no se puede utilizar el análisis gráfico definido en la sección anterior, porque las PHD no se pueden linealizar. En consecuencia, se recurre a un método iterativo denominado *algoritmo esperanza-maximización*, más conocido como algoritmo EM (desarrollado por Asmussen, Nerman y Olsson [3] y presentado en el libro de Buchholz, Kriege y Felko [5]), que alterna dos pasos, esperanza y maximización, para obtener la estimación de los parámetros por máxima verosimilitud. En la actualidad existen paquetes implementados en Matlab o en R (véase el paquete `mapfit`), e incluso aplicaciones disponibles para ordenador (por ejemplo, EMpht), tanto para determinar la estructura de las distribuciones tipo fase como para la estimación de sus parámetros.

4. Simulación con R

El objetivo de la presente sección es demostrar el poder de ajuste de las distribuciones tipo fase. Para ello, se han simulado cien valores de una distribución no negativa (más concretamente, cien valores de una distribución uniforme en el intervalo [0,3]) y, seguidamente, se ha procedido a ajustar las distribuciones de Erlang, Weibull, valores extremos y log-normal a los datos simulados (ya ordenados) mediante la técnica del análisis gráfico. La bondad del ajuste de cada una de las distribuciones mencionadas se ha comparado por medio del coeficiente de determinación (R^2) obtenido en cada caso, eligiendo como óptima aquella cuyo valor R^2 esté más cercano a 1. Seleccionada la distribución más precisa, se procede, en primer lugar, a estimar la distribución tipo fase que mejor se ajusta a estos datos y, posteriormente, a comparar el ajuste obtenido en ambas distribuciones.

En la figura 1 aparece el ajuste de los datos simulados con las distribuciones descritas en la sección 2.2 utilizando el análisis gráfico: la gráfica (A) representa el ajuste mediante la distribución exponencial, la gráfica (B) denota el ajuste a través de la distribución de Weibull, la gráfica (C) muestra el ajuste por medio de la distribución de valores extremos, y la gráfica (D) revela el ajuste mediante la distribución log-normal.

A tenor del gráfico y apoyándose en el cuadro 3, en el cual figura la estimación de mínimos cuadrados de los parámetros de cada distribución y el coeficiente de determinación alcanzado en cada caso, se concluye que la distribución que mejor funciona para los datos simulados es la distribución de Weibull. En consecuencia, esta será la distribución elegida para compararla con la distribución tipo fase al final de la presente sección.

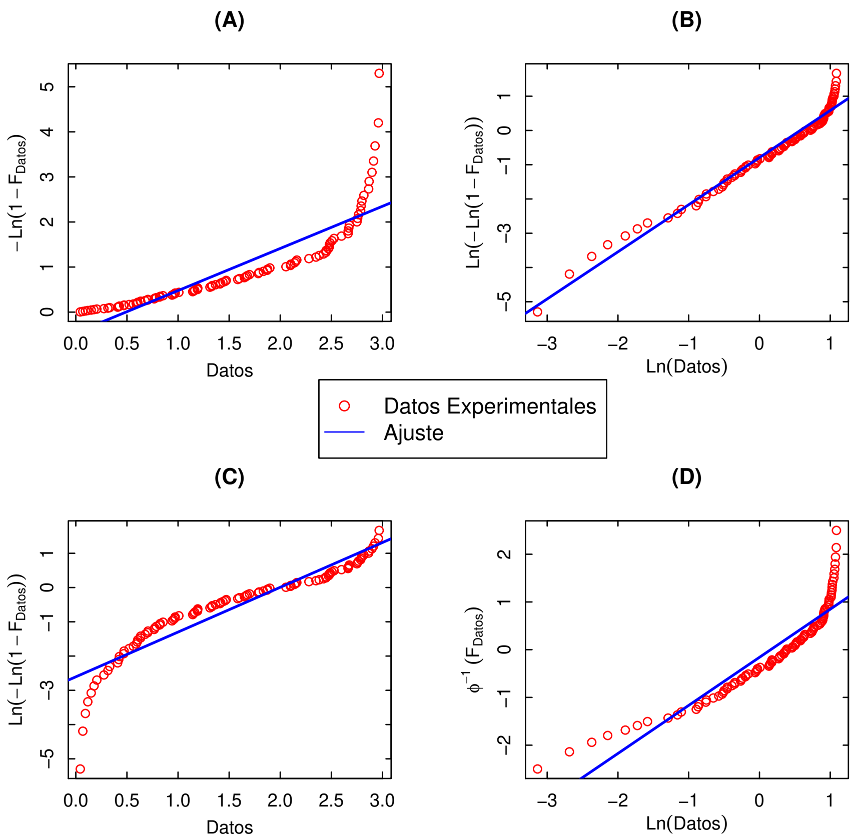

Figura 1: Ajuste por mínimos cuadrados de los datos simulados a través del método de análisis gráfico: (A) distribución exponencial, (B) distribución de Weibull, (C) distribución de valores extremos, (D) distribución log-normal. Para más claridad se indica la nube de puntos considerada en cada caso.

Seguidamente, se han ajustado distintas distribuciones tipo fase a los datos simulados usando el algoritmo EM, asumiendo que la matriz que contiene las intensidades transitorias hasta la absorción, es decir, la matriz $\boldsymbol{T}$, tiene estructura interna general. Esto da lugar a que se estimen $m(m+1)+m$ parámetros, siendo

Cuadro 3: Coeficiente de determinación alcanzado y parámetros estimados por mínimos cuadrados tras aplicar el método de análisis gráfico a los datos simulados considerando distintas distribuciones.

	R^2	Parámetros estimados
Exponencial	0,75	$\hat{\lambda} = 0{,}935$
Weibull	0,97	$\hat{\lambda} = 0{,}562$; $\hat{\beta} = 1{,}375$
Valores extremos	0,87	$\hat{\gamma} = 0{,}07$; $\hat{\beta} = 1{,}31$
Log-normal	0,86	$\hat{\mu} = 0{,}16$; $\hat{\sigma} = 0{,}99$

m el número de estados transitorios. Después de este análisis, se ha observado que la estructura interna y el vector que contiene las probabilidades iniciales no adoptan una expresión concreta para un m fijo, sino que van variando según el número de estados que se considere. En base a ello, no se puede asumir ninguna distribución tipo fase conocida a los datos.

Los parámetros de la distribución tipo fase han sido estimados aplicando el algoritmo EM con la aplicación EMpht y con el paquete `mapfit` de R. Se empieza con un número pequeño de estados y se va aumentando hasta que se consigue el ajuste óptimo, el cual es alcanzado para los datos simulados en veinte fases. La representación de dicha distribución tipo fase se puede apreciar en el apéndice A que figura al final del artículo. Una vez estimados los parámetros de la distribución tipo fase, se procede a observar gráficamente la precisión del ajuste de esta distribución a los datos simulados. La tasa de fallo acumulada experimental estimada por la distribución tipo fase se muestra en la figura 2. Cabe notar que en este caso, y con el objetivo de utilizar de nuevo la técnica del análisis gráfico, en el eje de ordenadas se representa $-\ln(1-F)$ y en el eje de abscisas, los datos ordenados (esta transformación es una generalización de la realizada en el caso exponencial).

Finalmente, se procede a comparar los ajustes obtenidos a través de la distribución de Weibull y de la distribución tipo fase. Para ello, se utilizan la función de fiabilidad y la función tasa de riesgo. En la figura 3 se muestra, por un lado, la función de fiabilidad experimental, además del ajuste por Weibull y tipo fase (A) y, por otro lado, la tasa de riesgo experimental frente al ajuste por Weibull y tipo fase.

A raíz de los gráficos de la figura 3, el mejor resultado se logra cuando se considera la distribución tipo fase con veinte estados transitorios (fases) y puede apreciarse que la precisión es bastante notable.

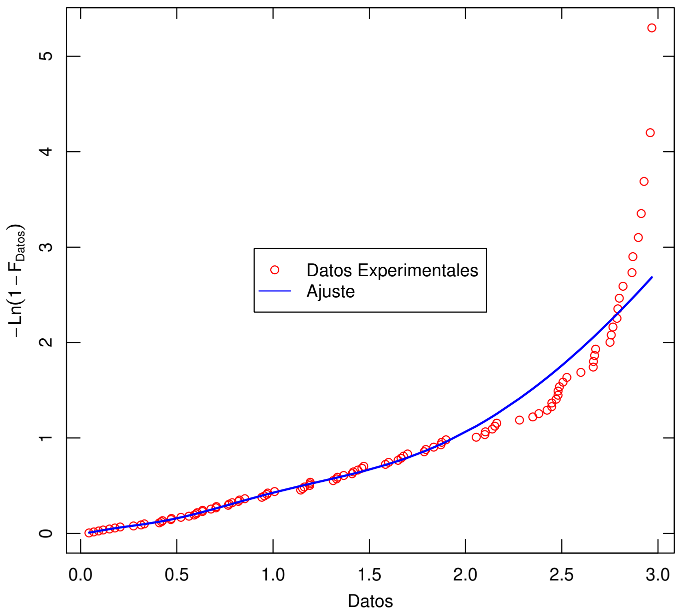

Figura 2: Tasa de fallo acumulada de los datos simulados y el correspondiente ajuste de la distribución tipo fase con veinte fases.

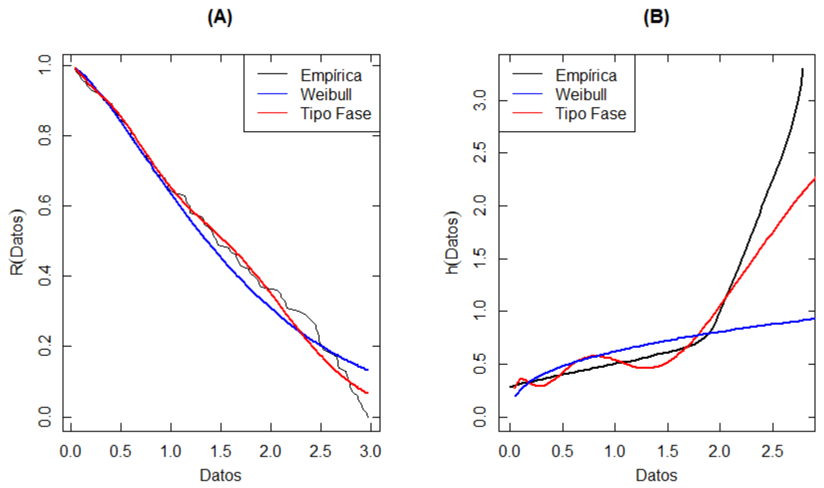

Figura 3: (A) Función de fiabilidad $R(t)$ frente a los datos simulados y las distribuciones de Weibull y tipo fase. (B) Tasa de riesgo $h(t)$ frente a los datos simulados y las distribuciones de Weibull y tipo fase.

5. Conclusiones

El análisis estadístico habitual realizado con datos experimentales en estudios de fiabilidad para caracterizar las principales razones de fallo de cualquier sistema es a través de distribuciones continuas, como pueden ser la distribución de Weibull, exponencial, gamma, log-normal, etc. Sin embargo, a veces el ajuste obtenido a los datos medidos no es del todo riguroso. Este hecho sugiere que otras distribuciones estadísticas podrían funcionar de una mejor manera y el ajuste sería más preciso. A este respecto, en los últimos años es cada vez más común utilizar un enfoque basado en las distribuciones tipo fase, ya que cualquier distribución no negativa puede aproximarse tanto como sea necesario mediante una distribución tipo fase. Cabe destacar también las buenas propiedades algebraico-matricial y de clausura que posee la clase de distribuciones tipo fase y su fácil algoritmización que permite obtener expresiones de las cantidades de interés del modelo en forma bien estructurada, de modo que puedan ser tratadas computacionalmente de manera rápida y eficaz, facilitando así la estimación de los parámetros.

Para poner de manifiesto el poder de precisión de estas distribuciones, en el presente trabajo se han simulado cien valores de una distribución uniforme en el intervalo $[0, 3]$, y se han ajustado y comparado distintas distribuciones. Después de un estudio exhaustivo y pormenorizado se ha concluido que la distribución que mejor funciona para los datos simulados es la distribución tipo fase. Por lo tanto, y en consecuencia de todo lo comentado a lo largo del artículo, parece evidente concluir que las distribuciones tipo fase deben ser tenidas en cuenta cuando se quiera analizar datos experimentales referidos al fallo de cualquier sistema.

Referencias

[1] Acal, Christian; Ruiz-Castro, Juan Eloy; Aguilera, Ana María; Jiménez-Molinos, Francisco, y Roldán, Juan B. «Phase-type distributions for studying variability in resistive memories». En: *Journal of Computational and Applied Mathematics* 345 (2019), págs. 23-32. issn: 0377-0427. https://doi.org/10.1016/j.cam.2018.06.010.

[2] Asmussen, Søren. *Ruin Probabilities*. Hong Kong: World Scientific, 2000. https://doi.org/10.1142/7431.

[3] Asmussen, Søren; Nerman, Olle, y Olsson, Marita. «Fitting Phase-Type Distributions via the EM Algorithm». En: *Scandinavian Journal of Statistics* 23.4 (1996), págs. 419-441. issn: 03036898. url: http://www.jstor.org/stable/4616418.

[4] Bazovsky, Igor. *Reliability theory and practice*. Prentice-Hall, 1961.

[5] Buchholz, Peter; Kriege, Jan, y Felko, Iryna. *Input Modeling with Phase-Type Distributions and Markov Models. Theory and Applications.* 1.ª ed. SpringerBriefs in Mathematics. An optional note. Cham: Springer, 2014. https://doi.org/10.1007/978-3-319-06674-5.

[6] Epstein, Benjamin y Sobel, Milton. «Life Testing». En: *Journal of the American Statistical Association* 48.263 (1953), págs. 486-502. https://doi.org/10.1080/01621459.1953.10483488.

[7] He, Qi-Ming. *Fundamentals of Matrix-Analytic Methods.* New York: Springer, 2014. https://doi.org/10.1007/978-1-4614-7330-5.

[8] Long, Shibing; Cagli, Carlo; Ielmini, Daniele; Liu, Ming, y Suñé, Jordi. «Analysis and modeling of resistive switching statistics». En: *Journal of Applied Physics* 111.7, 074508 (2012). https://doi.org/10.1063/1.3699369.

[9] Luo, Wun-Cheng; Liu, Jen-Chieh; Feng, Hsien-Tsung; Lin, Yen-Chuan; Huang, Jiun-Jia; Lin, Kuan-Liang, y Hou, Tuo-Hung. «RRAM SET speed-disturb dilemma and rapid statistical prediction methodology». En: *2012 International Electron Devices Meeting.* Dic. de 2012, págs. 9.5.1-9.5.4. https://doi.org/10.1109/IEDM.2012.6479012.

[10] Neuts, Marcel F. «Probability distributions of phase type». En: *Liber Amicorum Professor Emeritus Dr. H. Florin* (1975), págs. 173-206.

[11] Neuts, Marcel F. *Matrix-geometric solutions in stochastic models: an algorithmic approach.* Courier Corporation, 1994. isbn: 978-0-486-68342-3.

[12] Pan, Feng; Gao, Shuang; Chen, Chao; Song, Chen, y Zeng, Fei. «Recent progress in resistive random access memories: Materials, switching mechanisms, and performance». En: *Materials Science and Engineering: R: Reports* 83 (2014), págs. 1-59. issn: 0927-796X. https://doi.org/10.1016/j.mser.2014.06.002.

[13] Ruiz-Castro, Juan Eloy y Dawabsha, Mohammed. «A discrete MMAP for analysing the behaviour of a multi-state complex dynamic system subject to multiple events». En: *Discrete Event Dynamic Systems* (dic. de 2018). issn: 1573-7594. https://doi.org/10.1007/s10626-018-0274-0.

[14] Ruiz-Castro, Juan Eloy; Dawabsha, Mohammed, y Alonso, Francisco Javier. «Discrete-time Markovian arrival processes to model multi-state complex systems with loss of units and an indeterminate variable number of repairpersons». En: *Reliability Engineering & System Safety* 174 (2018), págs. 114-127. issn: 0951-8320. https://doi.org/10.1016/j.ress.2018.02.019.

A. Apéndice

La representación de la distribución tipo fase lograda en la sección 4 es la siguiente:

1. El vector $\boldsymbol{\alpha}$ toma los valores

$$(0{,}5416, 0{,}1237, 0{,}2346, 0, 0{,}0515, 0{,}0003, 0{,}0001, 0, 0{,}0026, 0, 0, 0{,}0121, 0, 0, 0{,}0215, 0{,}0001, 0, 0, 0{,}0119, 0).$$

2. La matriz $\boldsymbol{T}$ adopta la expresión que se ve en la página siguiente, en la figura 4.

$$
\begin{pmatrix}
-5.9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5.9 & 0 & 0 & 0 \\
0 & -8 & 0 & 0 & 0.18 & 4.17 & 0 & 0 & 0 & 0 & 0 & 0.87 & 0 & 0 & 0.02 & 2.77 & 0 & 0 & 0 & 0 \\
0 & 0.85 & -18.91 & 0 & 3.67 & 2.85 & 0 & 0 & 0.02 & 0 & 0 & 2.64 & 0 & 0 & 0.5 & 0.87 & 0 & 7.43 & 0.08 & 0 \\
0 & 0 & 0 & -5.9 & 0 & 0 & 5.9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.44 & 0 & 0 & -6.17 & 3.9 & 0 & 0 & 0 & 0 & 0 & 0.3 & 0 & 0 & 0.05 & 1.47 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -6.51 & 0 & 0 & 0 & 3.94 & 0 & 0 & 0 & 0.85 & 0 & 1.72 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -5.9 & 0 & 1.87 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4.02 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -5.9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5.9 \\
0 & 0.77 & 0 & 0 & 1.39 & 0 & 0 & 0 & -6.15 & 0 & 0 & 0.28 & 0 & 0 & 3.39 & 0 & 0 & 0 & 0.32 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -7.07 & 2.38 & 0 & 0 & 4.69 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -6.08 & 0 & 0 & 0 & 0 & 0 & 0 & 6.08 & 0 & 0 \\
0 & 0.06 & 0 & 0 & 0.18 & 0.87 & 0 & 0 & 0 & 0.02 & 0 & -7.58 & 0 & 0.02 & 0 & 6.42 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 5.9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -5.9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.06 & 6.69 & 0 & 0 & -6.75 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 2.78 & 0 & 0 & 0.17 & 1.94 & 0 & 0 & 0 & 0 & 0 & 1.47 & 0 & 0 & -6.9 & 0.53 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.01 & 0 & 0 & 0 & 2.57 & 0 & 0 & 0 & 3.94 & 0 & -6.53 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 5.9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -5.9 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -6.77 & 0 & 0 \\
0 & 0.33 & 0 & 0 & 3.01 & 0 & 0 & 0 & 0.07 & 0 & 0 & 0.18 & 0 & 0 & 2.41 & 0 & 0 & 0 & -6 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5.9 & 0 & 0 & 0 & 0 & 0 & 0 & -5.9
\end{pmatrix}
$$

Figura 4: Expresión de la matriz $\boldsymbol{T}$.

TEMat

Álgebras de Boole y la dualidad de Stone

Clara María Corbalán Mirete
Universidad de Murcia
claramaria.corbalan@um.es

Resumen: Este artículo tiene como objetivo introducir los conceptos de álgebra de Boole y espacio de Stone, así como presentar la dualidad existente entre ambos. Para ello, comenzamos presentando este tipo de álgebras, algunas de sus propiedades y sus elementos y subconjuntos más destacables: átomos, ideales, filtros y ultrafiltros. Gracias a ellos seremos capaces de demostrar el teorema de Stone, el cual cuenta con dos versiones y establece que toda álgebra de Boole $\mathfrak{B}$ es isomorfa al álgebra de los clopen sobre el espacio de los ultrafiltros de $\mathfrak{B}$. Además de esto, y ya para finalizar, probaremos que todo espacio de Stone X es homeomorfo al espacio de los ultrafiltros del álgebra de los clopen sobre X.

Abstract: The purpose of this paper is to introduce the concepts of Boolean algebra and Stone space, and to present the duality between them. In order to achieve this, we begin by presenting some of the Boolean algebra's properties and its most fundamental elements and subsets: atoms, ideals, filters and ultrafilters. Then, we will be able to prove Stone's theorem, which has two versions and states that every Boolean algebra $\mathfrak{B}$ is isomorphic to the clopen algebra on the space of ultrafilters of $\mathfrak{B}$. Finally, we will prove that every Stone space X is homeomorphic to the space of ultrafilters of the clopen algebra on X.

Palabras clave: álgebra de Boole, espacio de Stone, clopen, dual, ultrafiltro.

MSC2010: 03G05, 06E15.

Recibido: 1 de octubre de 2018.
Aceptado: 25 de febrero de 2019.

Agradecimientos: Me gustaría mostrar mi gratitud a Antonio Avilés López y a Gonzalo Martínez Cervantes, directores de mi Trabajo Final de Máster, por su ayuda, paciencia y dedicación, ya que este artículo se basa, en parte, en dicho trabajo. A los revisores por ayudarme a mejorar la calidad del mismo. A José Luis, mi amigo, por el apoyo prestado ahora y siempre. Y también a Paco porque siempre he podido contar con él.

Referencia: Corbalán Mirete, Clara María. «Álgebras de Boole y la dualidad de Stone». En: *TEMat*, 3 (2019), págs. 75-86. ISSN: 2530-9633. URL: https://temat.es/articulo/2019-p75.

1. Introducción

Llamadas así en honor al matemático inglés autodidacta George Boole, las álgebras de Boole aparecieron por primera vez como estructura algebraica en un pequeño panfleto [1] publicado en 1847 en respuesta a la controversia generada entre el profesor Augustus De Morgan y *sir* William Hamilton acerca de la llamada «cuantificación del predicado» y, más tarde, en 1854, como parte de su trabajo más importante, *An investigation of the laws of thought, on which are founded the mathematical theories of logic and probabilities* [2]. Boole intentó reducir la lógica a un álgebra sencilla que solo utilizara dos cantidades (0, 1) y tres operaciones básicas (y, o, no).

Sin embargo, en la actualidad estos trabajos únicamente poseen interés histórico, pues, a pesar de los pequeños avances que realizaron Schröder, Löwenheim y Huntington a principios de siglo, se considera que la teoría moderna de álgebras de Boole se inició en 1930 con las aportaciones de Marshall Stone y Alfred Tarski. Desde entonces ha habido un desarrollo constante de este campo. Así, en 1948, el matemático estadounidense Claude Shannon demostró que las álgebras de Boole se podían aplicar para optimizar el diseño de los sistemas de conmutación telefónica y que los circuitos con relés eran capaces de resolver problemas relacionados con ellas. De este modo, Boole se convirtió, con la ayuda de Shannon, en uno de los fundadores de la era digital.

Por otra parte, la dualidad topológica de Stone tiene su origen a la vez que la teoría moderna de álgebras de Boole y los espacios de Hausdorff compactos cero-dimensionales, también llamados espacios de Stone, pues existe una correspondencia entre los homomorfismos entre álgebras de Boole y las aplicaciones entre espacios de Stone. Como consecuencia, las cuestiones algebraicas en las álgebras de Boole se traducen en topológicas en los espacios de Stone y viceversa.

2. Álgebras de Boole

Las álgebras de Boole se definen a partir de una lista de axiomas algebraicos. Ahora nos dedicaremos a presentar las leyes aritméticas que se derivan de dichos axiomas y algunas nociones básicas acerca de estas estructuras y sobre algunos de sus subconjuntos. Las principales fuentes de referencia para la elaboración de este artículo han sido los libros de Koppelberg [5] y Jané [4]. Se recomienda su consulta para una mayor profundización en el tema.

Definición 1. Un **álgebra de Boole** es un álgebra $\mathfrak{B} = (\mathcal{B}, \vee^{\mathfrak{B}}, \wedge^{\mathfrak{B}}, \neg^{\mathfrak{B}}, 0^{\mathfrak{B}}, 1^{\mathfrak{B}})$, donde $\mathcal{B} \neq \emptyset$, que satisface los siguientes axiomas:

I) $\forall x, y \in \mathcal{B},\ x \vee y = y \vee x \quad \text{y} \quad x \wedge y = y \wedge x.$

II) $\forall x, y, z \in \mathcal{B},\ x \vee (y \vee z) = (x \vee y) \vee z \quad \text{y} \quad x \wedge (y \wedge z) = (x \wedge y) \wedge z.$

III) $\forall x, y, z \in \mathcal{B},\ x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z) \quad \text{y} \quad x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z).$

IV) $\forall x \in \mathcal{B},\ x \vee \neg x = 1 \quad \text{y} \quad x \wedge \neg x = 0.$

V) $\forall x \in \mathcal{B},\ x \vee 0 = x \quad \text{y} \quad x \wedge 1 = x.$ ◂

Observación 2. Conviene tener en cuenta que los axiomas dados en la definición anterior dependen del autor, pues existen varias versiones equivalentes. Esto puede comprobarse consultando el libro de Koppelberg [5], en donde se enuncian unos axiomas similares que pueden resultar útiles para probar algunos de los resultados cuya demostración no se ha realizado. ◂

Un álgebra de Boole se dice que es **propia** si se cumple que $0^{\mathfrak{B}} \neq 1^{\mathfrak{B}}$; en otro caso, diremos que es impropia.

Si $a, b \in \mathfrak{B}$, decimos que $a \vee b$ es la unión o disyunción de a y b; que $a \wedge b$ es la intersección o conjunción de a y b, y que $\neg a$ es el complemento de a.

Definición 3. Una estructura $\mathfrak{A} = (\mathcal{A}, \vee^{\mathfrak{A}}, \wedge^{\mathfrak{A}}, \neg^{\mathfrak{A}}, 0^{\mathfrak{A}}, 1^{\mathfrak{A}})$ es una **subálgebra** de un álgebra de Boole $\mathfrak{B} = (\mathcal{B}, \vee^{\mathfrak{B}}, \wedge^{\mathfrak{B}}, \neg^{\mathfrak{B}}, 0^{\mathfrak{B}}, 1^{\mathfrak{B}})$ si $\mathcal{A} \subseteq \mathcal{B}$, $0_{\mathfrak{A}} = 0_{\mathfrak{B}}$, $1_{\mathfrak{A}} = 1_{\mathfrak{B}}$ y las operaciones $\vee^{\mathfrak{A}}, \wedge^{\mathfrak{A}}, \neg^{\mathfrak{A}}$ son las restricciones de $\vee^{\mathfrak{B}}, \wedge^{\mathfrak{B}}, \neg^{\mathfrak{B}}$ a $\mathcal{A}$. ◂

Para todo conjunto A, el **álgebra potencia** $P(A) = (\mathcal{P}(A), \cup, \cap, \,^{-}, \varnothing, A)$, donde para cada $X \subseteq A$, $\overline{X} = A \setminus X$, es un álgebra de Boole. $P(A)$ es propia si y solo si $A \neq \varnothing$.

Más generalmente, dado un conjunto A, un **álgebra de conjuntos** sobre A es una subálgebra de $P(A)$. Así, una colección $\mathcal{B}$ de subconjuntos de A es el universo de un álgebra de conjuntos sobre A si y solo si $\varnothing, A$ son elementos de $\mathcal{B}$ y para todo $X, Y \in \mathcal{B}$ se cumple que $X \cup Y, X \cap Y, A \setminus X \in \mathcal{B}$.

Ejemplo 4 (álgebras finitas/cofinitas). Sea X un conjunto cualquiera. Un subconjunto a de X se dice que es cofinito en X si $X \setminus a$ es finito. Consideremos

$$A = \{a \subseteq X \,:\, a \text{ es finito o cofinito}\}.$$

Entonces, A es un álgebra de conjuntos sobre X, denominada álgebra finita/cofinita sobre X. Para comprobar que $a \cup b$ y $a \cap b$ pertenecen a A para $a, b \in A$, nótese que $a \cup b$ es finita si a y b son finitos, y cofinita en otro caso. Que $a \cap b \in A$ se sigue de las leyes de De Morgan $a \cap b = X \setminus ((X \setminus a) \cup (X \setminus b))$ puesto que A es cerrado bajo complementarios y uniones. ◂

Si X tiene una cardinalidad infinita $\varkappa$, entonces tiene un álgebra finita/cofinita. Puesto que X tiene exactamente $\varkappa$ subconjuntos finitos, todo cardinal infinito se corresponde con la cardinalidad de un álgeba de Boole. Un entero no negativo k, sin embargo, cumple esto si y solo si $k = 2^n$ para algún $n \in \mathbb{N}$, como se sigue del corolario 28.

Ejemplo 5 (álgebra de los clopen). Sea X un espacio topológico. Un subconjunto de X es un clopen si es abierto y cerrado. El conjunto de los subconjutos clopen, $\mathrm{Clop}(X)$, es un álgebra de conjuntos sobre X, denominada álgebra de los clopen sobre X. ◂

He aquí algunos primeros resultados en relación a las álgebras booleanas cuyas demostraciones, como se puede apreciar, son sencillas y rutinarias.

Lema 6. *Toda álgebra booleana cumple que*

I) $x \vee x = x$ *y* $x \wedge x = x$.
II) $x \wedge 0 = 0$ *y* $x \vee 1 = 1$.
III) $\neg x$ *es el único elemento* $y \in \mathcal{B}$ *tal que* $x \wedge y = 0$ *y* $x \vee y = 1$.
IV) $\neg\neg x = x$.
V) *si* $\neg x = \neg y$*, entonces* $x = y$.
VI) $\neg 0 = 1$ *y* $\neg 1 = 0$.
VII) $\neg(x \vee y) = \neg x \wedge \neg y$ *y* $\neg(x \wedge y) = \neg x \vee \neg y$.
VIII) $x \vee (x \wedge y) = x$ *y* $x \wedge (x \vee y) = x$.
IX) $x \vee (\neg x \vee y) = 1$ *y* $x \wedge (\neg x \wedge y) = 0$.

Demostración. I) Haciendo uso de los axiomas de la definición 1 es sencillo ver que

$$x \vee x = (x \vee x) \wedge 1 = (x \vee x) \wedge (x \vee \neg x) = x \vee (x \wedge \neg x) = x \vee 0 = x.$$

De forma similar se obtiene el otro resultado.

II) En efecto,

$$x \wedge 0 = x \wedge (x \wedge \neg x) = (x \wedge x) \wedge \neg x = x \wedge \neg x = 0.$$

De forma similar se obtiene que $x \vee 1 = 1$.

III) Supongamos que $x \wedge y = 0$ y $x \vee y = 1$. Entonces, aplicando los axiomas, se tiene que

$$\begin{aligned} y &= y \wedge 1 = y \wedge (x \vee \neg x) = (y \wedge x) \vee (y \wedge \neg x) = (x \wedge y) \vee (\neg x \wedge y) = 0 \vee (\neg x \wedge y) \\ &= (x \wedge \neg x) \vee (\neg x \wedge y) = (\neg x \wedge x) \vee (\neg x \wedge y) = \neg x \wedge (x \vee y) = \neg x \wedge 1 = \neg x. \end{aligned}$$

IV) Aplicando los axiomas se sabe que $\neg x \vee x = x \vee \neg x = 1$ y que $\neg x \wedge x = x \wedge \neg x = 0$. Por la propiedad anterior, se tiene que $x = \neg\neg x$.

v) Si $\neg x = \neg y$, entonces $x = \neg\neg x = \neg\neg y = y$.

vi) Veamos que $\neg 0 = 1$. Sabemos que $x \vee \neg x = 1$; por tanto, $0 \vee \neg 0 = 1$ y debemos probar que $0 \vee 1 = 1$:

$$0 \vee 1 = 1 \vee 0 = (0 \vee \neg 0) \vee 0 = (0 \vee 0) \vee (\neg 0 \vee 0) = 0 \vee (\neg 0 \vee 0) = (\neg 0 \vee 0) \vee 0 = \neg 0 \vee 0 = 0 \vee \neg 0 = 1.$$

El otro caso se prueba de forma similar.

vii) Se tiene que

$$\begin{aligned}(x \vee y) \wedge \neg x \wedge \neg y &= ((x \wedge \neg x) \vee (y \wedge \neg x)) \wedge \neg y = (0 \vee (y \wedge \neg x)) \wedge \neg y = (0 \wedge \neg y) \vee (y \wedge \neg x \wedge \neg y)\\ &= 0 \vee (y \wedge \neg y \wedge \neg x) = 0 \vee (0 \wedge \neg x) = 0 \vee 0 = 0.\end{aligned}$$

La segunda igualdad se obtiene por un argumento análogo.

viii) Es fácil comprobar que

$$x \vee (x \wedge y) = (x \wedge 1) \vee (x \wedge y) = x \wedge (1 \vee y) = x \wedge (y \vee 1) = x \wedge 1 = x.$$

De forma análoga se obtiene el segundo resultado.

ix) Por último,

$$x \vee (\neg x \vee y) = (x \vee \neg x) \vee y = 1 \vee y = 1.$$

La otra igualdad se prueba de forma similar. ■

Proposición 7. *En toda álgebra de Boole se cumple que*

i) $(x \vee y) \vee (\neg x \wedge \neg y) = 1$.

ii) $(x \vee y) \wedge (\neg x \wedge \neg y) = 0$.

Demostración. i) Tenemos que

$$\begin{aligned}(x \vee y) \vee (\neg x \wedge \neg y) &= ((x \vee y) \vee \neg x) \wedge ((x \vee y) \vee \neg y)\\ &= (y \vee (x \vee \neg x)) \wedge (x \vee (y \vee \neg y)) = (y \vee 1) \wedge (x \vee 1) = 1 \wedge 1 = 1.\end{aligned}$$

ii) Se procede de manera similar. ■

Corolario 8 (leyes de De Morgan). *Toda álgebra booleana cumple que*

i) $\neg(x \vee y) = \neg x \wedge \neg y$.

ii) $\neg(x \wedge y) = \neg x \vee \neg y$.

Demostración. Basta con aplicar la proposición anterior y el punto ii) del lema 6. ■

A continuación enunciaremos un principio general que, en lo que a computación se refiere, ahorra una gran cantidad de trabajo.

Definición 9. Si E es una expresión en un lenguaje cuyos únicos símbolos no lógicos son $\vee, \wedge, \neg, 0, 1$, entonces E', la **expresión dual**, se obtiene a partir de E al reemplazar los símbolos $\vee, \wedge, 0, 1$ por $\wedge, \vee, 1, 0$, respectivamente, dejando fijo $\neg$.

Una forma alternativa de interpretar E es como función de $\mathcal{B}^n$ en $\mathcal{B}$ dada por múltiples composiciones de $\vee^{\mathfrak{B}}, \wedge^{\mathfrak{B}}, \neg^{\mathfrak{B}}$. ◂

Obsérvese que $(E')'$ es E y que la expresión dual de cada uno de los axiomas anteriores es un axioma.

Definición 10. Si $\mathfrak{B} = (\mathcal{B}, \vee^{\mathfrak{B}}, \wedge^{\mathfrak{B}}, \neg^{\mathfrak{B}}, 0^{\mathfrak{B}}, 1^{\mathfrak{B}})$ es un álgebra de Boole, denotaremos como $\mathfrak{B}'$ al álgebra dual de $\mathfrak{B}$, cuyo universo también es $\mathcal{B}$ pero tal que $\vee^{\mathfrak{B}'} = \wedge^{\mathfrak{B}}$, $\wedge^{\mathfrak{B}'} = \vee^{\mathfrak{B}}$, $\neg^{\mathfrak{B}'} = \neg^{\mathfrak{B}}$, $0^{\mathfrak{B}'} = 1^{\mathfrak{B}}$ y $1^{\mathfrak{B}'} = 0^{\mathfrak{B}}$. ◂

Está claro que $(\mathfrak{B}')' = \mathfrak{B}$ y que todo enunciado E es verdadero en $\mathfrak{B}$ si y solo si E' lo es en $\mathfrak{B}'$. En consecuencia, $\mathfrak{B}'$ satisface los axiomas y es, por tanto, un álgebra de Boole. A partir de estas consideraciones podemos establecer el siguiente principio:

Proposición 11 (principio de dualidad). *Si E es un enunciado verdadero en toda álgebra de Boole, su dual E' también lo es.*

Demostración. Sea E tal enunciado y sea $\mathfrak{B}$ un álgebra de Boole. Veamos que E' es verdadero en $\mathfrak{B}$.

Dado que $\mathfrak{B}'$ es un álgebra de Boole, se tiene que E es verdadero en $\mathfrak{B}'$. Por tanto, como sabemos que todo enunciado E es verdadero en $\mathfrak{B}$ si y solo si E' lo es en B', se tiene que E' es verdadero en $(\mathfrak{B}')'$. Pero $(\mathfrak{B}')' = \mathfrak{B}$. Por tanto, se tiene que, tal y como queríamos probar, E' es verdadero en $\mathfrak{B}$. ■

Seguidamente, definiremos una relación de orden en las álgebras de Boole, pero antes debemos probar un primer resultado.

Lema 12. *En toda álgebra de Boole se cumple que*

$$x \wedge y = x \iff x \vee y = y.$$

Demostración. Supongamos que $x \wedge y = x$. Haciendo uso de los axiomas se tiene que

$$\begin{aligned} y &= y \wedge 1 = y \wedge (x \vee \neg x) = (y \wedge x) \vee (y \wedge \neg x) = (x \wedge y) \vee (y \wedge \neg x) \\ &= x \vee (y \wedge \neg x) = (x \vee y) \wedge (x \vee \neg x) = (x \vee y) \wedge 1 = x \vee y. \end{aligned}$$

Para la implicación contraria deberemos utilizar esto mismo haciendo uso del principio de dualidad. Así, sabemos que, si $y \wedge x = y$, entonces $y \vee x = x$. Por dualidad, se tiene que, si $y \vee x = y$, entonces $y \wedge x = x$. Y, finalmente, por el axioma I) de la definición 1, se tiene que, si $x \vee y = y$, entonces $x \wedge y = x$. ■

Definición 13. Dada un álgebra de Boole $\mathfrak{B}$, definimos el **orden canónico** de $\mathfrak{B}$, al que denotaremos como $\leq^{\mathfrak{B}}$ o simplemente $\leq$, como

$$x \leq y \iff x \wedge y = x.$$

O, equivalentemente,

$$x \leq y \iff x \vee y = y.$$ ◂

Lema 14. *El orden canónico de un álgebra de Boole es un orden parcial, pues*

I) $x \leq x$;
II) *si* $x \leq y$ *e* $y \leq x$, *entonces* $x = y$;
III) *si* $x \leq y$ *e* $y \leq z$, *entonces* $x \leq z$.

Demostración. I) Como vimos en el lema 6, se tiene que $x \wedge x = x$. Por tanto, se tiene que $x \leq x$ por la definición de orden.

II) Si $x \leq y$ e $y \leq x$, por la definición, se tiene que $x = x \wedge y$ e $y = y \wedge x$. Entonces, por el axioma I), se tiene que $x = y$.

III) Si $x \leq y$ e $y \leq z$, por definición, $x = x \wedge y$ e $y = y \wedge z$. Por tanto,

$$x \wedge z = (x \wedge y) \wedge z = x \wedge (y \wedge z) = x \wedge y = x.$$

Así, $x \leq z$. ■

Proposición 15. *Si $\mathfrak{B}$ es un álgebra de Boole,*

I) $(\mathcal{B}, \leq)$ *es un retículo, es decir, para cada par de elementos* $\{x, y\}$ *existen un supremo,* $x \vee y$, *y un ínfimo,* $x \wedge y$.
II) 1 *es el elemento máximo y* 0, *el mínimo de* $(\mathcal{B}, \leq)$.

Demostración. I) Veamos que el supremo de $\{x, y\}$ es $x \vee y$. Se tiene que $x \leq x \vee y$, pues

$$x \vee (x \vee y) = (x \vee x) \vee y = x \vee y.$$

Sea z una cota superior arbitraria de x e y, es decir, $y \leq z$ y $x \leq z$. Por tanto,

$$(x \vee y) \vee z = x \vee (y \vee z) = x \vee z = z.$$

Se tiene, pues, que $x \vee y \leq z$, lo que prueba que $x \vee y$ es la menor cota superior de $\{x, y\}$.

El caso del ínfimo resulta análogo por dualidad.

II) Dado que $x \vee \neg x = 1$, 1 es el supremo de $\{x, \neg x\}$. En particular, $x \leq 1, \forall\, x$. ■

Lema 16. *Si $x \leq x'$ e $y \leq y'$, entonces $x \vee y \leq x' \vee y'$, $x \wedge y \leq x' \wedge y'$ y, además, $\neg x' \leq \neg x$.*

Lema 17. *En toda álgebra de Boole $\mathfrak{B}$ se cumple que*

I) *$x \leq y$ si y solo si $\neg y \leq \neg x$ si y solo si $x \wedge \neg y = 0$.*

II) *$z \wedge x \leq y$ si y solo si $x \leq \neg z \vee y$.*

Definición 18. Un álgebra de Boole se dirá **completa** si todo conjunto tiene un supremo. ◂

Definición 19. Un **homomorfismo** entre dos álgebras de Boole $\mathfrak{A}$ y $\mathfrak{B}$ es una aplicación $f : \mathfrak{A} \to \mathfrak{B}$ tal que, para todo $x, y \in \mathcal{A}$,

$$f(x \vee y) = f(x) \vee f(y), \quad f(x \wedge y) = f(x) \wedge f(y), \quad f(\neg x) = \neg f(x), \quad f(0) = 0 \quad \text{y} \quad f(1) = 1.$$

Un homomorfismo $f : \mathfrak{A} \to \mathfrak{B}$ entre dos álgebras de Boole es un monomorfismo o embebimiento si es inyectivo. Y diremos que es epimorfismo si es suprayectivo. ◂

Definición 20. Un isomorfismo $f : \mathfrak{A} \to \mathfrak{B}$ entre álgebras de Boole es un monomorfismo que además es epimorfismo. ◂

Proposición 21. *Supongamos que $\mathfrak{A}$ y $\mathfrak{B}$ son dos álgebras de Boole y que f es una biyección entre ellas. Entonces, f es un isomorfismo entre las álgebras $\mathfrak{A}$ y $\mathfrak{B}$ si y solo si f es un isomorfismo entre los órdenes $(\mathcal{A}, \leq^{\mathfrak{A}})$ y $(\mathcal{B}, \leq^{\mathfrak{B}})$.*

Demostración. Por la proposición 15 tenemos definidos los símbolos lógicos en términos del orden. La demostración se deja para el lector. ■

2.1. Átomos, ultrafiltros y el teorema de Stone

Esta sección está destinada a enunciar teorema de Stone, el cual tiene dos versiones, la versión conjuntista y la topológica, cuyas demostraciones veremos más adelante.

Notación 22. Habitualmente se identifica $\mathcal{B}$ con $\mathfrak{B}$; por ello, a partir de ahora escribiremos $x \in \mathfrak{B}$ cuando un elemento x pertenezca al álgebra. ◂

Definición 23. Llamaremos **átomo** en un álgebra de Boole a todo elemento $a \in \mathfrak{B}$ distinto de 0 tal que $\{b \in \mathfrak{B} : b \leq a\} = \{0, a\}$. Denotaremos como $\mathrm{At}\,\mathfrak{B}$ al conjunto de los átomos de $\mathfrak{B}$.

Diremos que $\mathfrak{B}$ es un álgebra atómica si para todo elemento $x \in \mathfrak{B}$ distinto de 0 existe algún átomo a tal que $a \leq x$. ◂

Ejemplo 24. Un álgebra potencia $P(X)$ y el álgebra finita/cofinita sobre X son atómicas, siendo los átomos los conjuntos unipuntuales $\{x\}$, con $x \in X$.

Otro ejemplo de álgebra atómica es cualquier álgebra de Boole finita, dado que si $x > 0$ en un álgebra de Boole y no existiese ningún átomo por debajo de x, habría una sucesión infinita estrictamente decreciente $x_0 = x > x_1 > x_2 > \dots$ en $\mathfrak{A}^+ = \{x \in \mathfrak{A} : 0 < x\} = \mathfrak{A} \setminus \{0\}$. ◂

Lema 25. *Para todo elemento $a \in \mathfrak{B}$ las siguientes afirmaciones son equivalentes:*

I) *a es un átomo de $\mathfrak{B}$;*

II) *para todo $x \in \mathfrak{B}$, o bien $a \leq x$ o bien $a \leq \neg x$, pero no ambos;*

III) *$a > 0$ y, para todo $x, y \in \mathfrak{B}$, $a \leq x \vee y$ si y solo si $a \leq x$ o $a \leq y$.*

Proposición 26. *Para toda álgebra de Boole, la aplicación $f : \mathfrak{B} \to P(\mathrm{At}\,\mathfrak{B})$ definida por el conjunto $f(x) = \{a \in \mathrm{At}\,\mathfrak{B} : a \leq x\}$ es un homomorfismo. Será un embebimiento si $\mathfrak{B}$ es atómica y un epimorfismo si es completa.*

Demostración. Evidentemente, $f(0) = \varnothing$ y $f(1) = \operatorname{At}\mathfrak{B}$. Por el lema 25 II) se tiene que

$$f(\neg x) = \{a \in \operatorname{At}\mathfrak{B} : a \leq \neg x\} = \operatorname{At}\mathfrak{B} \setminus \{a \in \operatorname{At}\mathfrak{B} : a \leq x\} = \operatorname{At}\mathfrak{B} \setminus f(x),$$

y que $f(x \vee y) = f(x) \cup f(y)$ se obtiene de manera similar aplicando el lema 25 III). Además, se tiene que $f(x \wedge y) = f(x) \cap f(y)$ puesto que, por ser $x \wedge y$ el ínfimo de $\{x, y\}$,

$$a \leq x \wedge y \text{ si y solo si } a \leq x \text{ y } a \leq y,$$

lo cual es cierto para todo $a \in \mathfrak{B}$. Por tanto, f es un homomorfismo.

Consideremos ahora $\mathfrak{B}$ un álgebra atómica y sean $x \neq y$ en $\mathfrak{B}$. Sin pérdida de generalidad podemos suponer que $x \nleq y$. En tal caso, $x \wedge \neg y \neq 0$ por el lema 17 I), y existirá un átomo $a \in \mathfrak{B}$ tal que $a \leq x \wedge \neg y$. Se tiene que $a \in f(x)$ y $a \notin f(y)$ y, por tanto, que $f(x) \neq f(y)$, siendo f inyectiva.

Por último, f es un epimorfismo si $\mathfrak{B}$ es completa pues, dado un conjunto de átomos $A \subseteq \operatorname{At}\mathfrak{B}$, se tiene que $f(\sup A) = A$. En efecto, obviamente $A \subseteq f(\sup A)$. Y si $A \neq f(\sup A)$, existe un átomo $a \notin A$ tal que $a \leq \sup A$. En tal caso, como $a \notin A$, $b \leq \neg a$ para todo $b \in A$. De ello se tiene que $a \leq \sup A \leq \neg a$, es decir, $a \leq \neg a$, lo cual es absurdo. ∎

Esta proposición no solo es una versión más débil del teorema de Stone, sino que es una descripción de las álgebras de Boole atómicas completas.

Corolario 27. *Toda álgebra de Boole atómica es isomorfa a un álgebra de conjuntos. Toda álgebra de Boole completa y atómica es isomorfa a un álgebra potencia.*

Corolario 28. *Las álgebras booleanas finitas son, salvo isomorfismos, las álgebras potencia de los conjuntos finitos.*

Demostración. Si $\mathfrak{B}$ es un álgebra booleana finita, entonces $\operatorname{At}\mathfrak{B}$ es finita y $\mathfrak{B}$ es completa y atómica. Por la proposición 26, $\mathfrak{B}$ es isomorfa a $P(\operatorname{At}\mathfrak{B})$. ∎

En particular, un álgebra de Boole tendrá por cardinal un número natural si y solo si es una potencia de 2.

Corolario 29. *Dos álgebras booleanas finitas son isomorfas si y solo si tienen la misma cardinalidad.*

Demostración. Si $\mathfrak{A}$ y $\mathfrak{B}$ tienen la misma cardinalidad $\varkappa$, entonces, como $\mathfrak{A} \simeq P(\operatorname{At}\mathfrak{A})$ y $\mathfrak{B} \simeq P(\operatorname{At}\mathfrak{B})$, se tiene que $\varkappa = 2^n$, donde $n = |\operatorname{At}\mathfrak{A}| = |\operatorname{At}\mathfrak{B}|$. Recíprocamente, toda biyección entre $\operatorname{At}\mathfrak{A}$ y $\operatorname{At}\mathfrak{B}$ da lugar a un isomorfismo entre $P(\operatorname{At}\mathfrak{A})$ y $P(\operatorname{At}\mathfrak{B})$ y, por ende, entre $\mathfrak{A}$ y $\mathfrak{B}$. ∎

Pasemos ahora a ver algunas definiciones y propiedades referentes a los filtros y ultrafiltros en álgebras de Boole.

Los ultrafiltros son la principal herramienta en la prueba del teorema de Stone. Estos surgen de manera natural a partir de los embebimientos de álgebras de Boole en álgebras potencia. Veamos, primeramente, una definición más sencilla de lo que es un ultrafiltro en un álgebra booleana.

Sea $e\colon \mathfrak{B} \to P(X)$ un embebimiento o, para mayor generalidad, un homomorfismo. Entonces, para cualquier punto $x \in X$, el subconjunto

$$F = \{a \in \mathfrak{B} : x \in e(a)\}$$

de $\mathfrak{B}$ tiene las siguientes propiedades:

- $1 \in F$, $0 \notin F$,
- $a \wedge b \in F$ si y solo si $a, b \in F$,
- $a \vee b \in F$ si y solo si $a \in F$ o $b \in F$,
- $\neg a \in F$ si y solo si $a \notin F$.

Los subconjuntos de $\mathfrak{B}$ con estas propiedades son exactamente los ultrafiltros de $\mathfrak{B}$. Pasemos ahora a la definición en base a los filtros del álgebra.

Definición 30. Un **ideal** en un álgebra de Boole $\mathfrak{B}$ es un subconjunto, I, de $\mathfrak{B}$ tal que para cualesquiera $a, b \in \mathfrak{B}$

- $0 \in I$;
- si $a, b \in I$, entonces $a \vee b \in I$;
- si $a \in I$ y $b \leq a$, entonces $b \in I$. ◂

Un ideal I es **propio** si y solo si $I \neq \mathfrak{B}$. En otro caso se dirá impropio.

Definición 31. Un **filtro** en un álgebra de Boole $\mathfrak{B}$ es un ideal en el álgebra dual $\mathfrak{B}'$. Así, un subconjunto F de $\mathfrak{B}$ es un filtro en $\mathfrak{B}$ si y solo si para cualesquiera $a, b \in \mathfrak{B}$

- $1 \in F$;
- si $a, b \in F$, entonces $a \wedge b \in F$;
- si $a \in F$ y $a \leq b$, entonces $b \in F$. ◂

Un filtro F es **propio** si y solo si $F \neq \mathfrak{B}$. En otro caso se dirá impropio. Un filtro F es **maximal** si es propio y no existe otro filtro en $\mathfrak{B}$ que lo contenga como subconjunto propio.

Observación 32. Para todo ideal I se cumple que $0 \in I$ y para todo filtro F, que $1 \in F$. Además, un ideal es propio si y solo si no contiene al 1 y un filtro lo es si y solo si no contiene al 0. ◂

Lema 33. *Sea $\mathfrak{B}$ un álgebra de Boole. Si $X \subset \mathcal{B}$, entonces*

- *X es ideal de $\mathfrak{B}$ si y solo si $\{\neg a \,:\, a \in X\}$ es filtro en $\mathfrak{B}$.*
- *X es filtro de $\mathfrak{B}$ si y solo si $\{\neg a \,:\, a \in X\}$ es ideal de $\mathfrak{B}$.*

Si $\mathfrak{B}$ es un álgebra de Boole, la intersección de todo conjunto de filtros en $\mathfrak{B}$ es también un filtro en $\mathfrak{B}$. Así, dado X un subconjunto de $\mathfrak{B}$, hay un filtro mínimo, $F(X)$, en $\mathfrak{B}$ que incluye a X, a saber, la intersección de todos los filtros en $\mathfrak{B}$ que incluyen a X. $F(X)$ es el **filtro generado** por X. Si X es vacío, $F(X) = \{1^{\mathfrak{B}}\}$; si X es no vacío, $F(X)$ es el conjunto de los elementos $b \in \mathfrak{B}$ tales que $x_1 \wedge \ldots \wedge x_n \leq b$ para $n \geq 1$ y $x_1, \ldots, x_n \in X$.

Del mismo modo, dado que la intersección de todo conjunto de ideales en $\mathfrak{B}$ es un ideal en $\mathfrak{B}$, si $X \subseteq \mathfrak{B}$, hay un menor ideal, $I(X)$, que incluye a X. $I(X)$ es el **ideal generado** por X. Si X es vacío, $I(X) = \{0\}$; si es no vacío, $I(X)$ es el conjunto de los elementos $b \in \mathfrak{B}$ tales que $b \leq x_1 \vee \ldots \vee x_n$ para $n \geq 1$ y $x_1, \ldots, x_n \in X$.

Definición 34. Si $\mathfrak{B}$ es un álgebra de Boole, decimos que un subconjunto $X \subseteq \mathfrak{B}$ tiene la **propiedad de la intersección finita** (PIF) si y solo si el ínfimo de todo subconjunto finito de X es distinto de 0.

Decimos que X tiene la **propiedad de la unión finita** (PUF) si y solo si el supremo de todo subconjunto finito de X es distinto de 1. ◂

Definición 35. Decimos que un ideal I en un álgebra de Boole $\mathfrak{B}$ es **primo** si y solo si es propio y, para cualesquiera $a, b \in \mathfrak{B}$, si $a \wedge b \in I$ se tiene que $a \in I$ o $b \in I$.

Dualmente, un filtro F en un álgebra $\mathfrak{B}$ es un **ultrafiltro** si y solo si es propio y, para cualesquiera $a, b \in \mathfrak{B}$, si $a \vee b \in F$ se tiene que $a \in F$ o $b \in F$. ◂

Así, un ideal es primo si y solo si su filtro dual es un ultrafiltro, y un filtro es un ultrafiltro si y solo si su ideal dual es primo.

Observación 36. Otra forma de definirlos y que resulta mucho más práctica a la hora de comprobar si un conjunto es un ultrafiltro es la siguiente:

Un filtro F en $\mathfrak{B}$ es un ultrafiltro si para cada $x \in \mathfrak{B}$ se tiene que $x \in F$ o $\neg x \in F$, pero no ambos. ◂

Proposición 37. *Un filtro es maximal si y solo si es un ultrafiltro.*

Demostración. Para un filtro F y un elemento $a \notin F$, el conjunto $\{b \in \mathfrak{B} \,:\, \exists z \in F, z \wedge \neg a \leq b\}$ es un filtro que extiende a F y contiene a $\neg a$. ■

Teorema 38. *Un subconjunto de un álgebra de Boole está contenido en un ultrafiltro si y solo si posee la propiedad de intersección finita (PIF).*

Demostración. Por tener la propiedad PIF, el filtro generado es propio, luego se tiene que el conjunto $\{F \subseteq \mathfrak{B} : X \subseteq F$ y F filtro propio$\}$ es no vacío. Aplicando el lema de Zorn y la proposición 37 se obtiene el resultado. ■

Corolario 39. *Un elemento a de un álgebra de Boole está contenido en un ultrafiltro si y solo si $a > 0$.*

Definición 40. Para un álgebra de Boole $\mathfrak{B}$,

$$\operatorname{Ult}\mathfrak{B} = \{F \subseteq \mathfrak{B} : F \text{ es un ultrafiltro de } \mathfrak{B}\}$$

es el conjunto de los ultrafiltros de $\mathfrak{B}$.

La aplicación $s\colon \mathfrak{B} \to P(\operatorname{Ult}\mathfrak{B})$ definida por

$$s(x) = \{F \in \operatorname{Ult}\mathfrak{B} : x \in F\}$$

se denomina **aplicación de Stone**. ◂

Teorema 41 (de representación de Stone, versión conjuntista). *Toda álgebra de Boole es isomorfa a un álgebra de conjuntos.*

Demostración. La prueba de que la aplicación de Stone s es un homomorfismo de $\mathfrak{B}$ en $P(\operatorname{Ult}\mathfrak{B})$ es análoga a la que aparece en la proposición 26.

Veamos que s es inyectiva. Consideremos $x \neq y$ en $\mathfrak{B}$; sin pérdida de generalidad podemos considerar $x \nleq y$. Por tanto, $x \wedge \neg y > 0$, por el lema 17 i). Por el corolario 39, sea F un ultrafiltro que contenga a $x \wedge \neg y$, de modo que, por una definición de ultrafiltro equivalente a la dada, contendrá a x y a $\neg y$. De ese modo, se tiene que $x \in F$ e $y \notin F$, siendo esto último por la observación 36, lo cual implica que $F \in s(x) \setminus s(y)$. ■

Recordemos que un álgebra de conjuntos es una subálgebra del álgebra potencia sobre un conjunto dado. Esta versión del teorema de Stone establece que toda álgebra de Boole $\mathfrak{B}$ es isomorfa a un álgebra de conjuntos, y gracias a la versión topológica que estudiaremos a continuación veremos que se trata del álgebra de los clopen sobre un espacio topológico.

3. Espacios de Stone

Esta sección establece la dualidad fundamental entre las álgebras de Boole y unos espacios topológicos especiales, los espacios de Stone. Así, el álgebra dual de un espacio de Stone X es $\operatorname{Clop}(X)$, el álgebra de los subconjuntos clopen de X, y el espacio dual de un álgebra de Boole $\mathfrak{B}$ es el conjunto $\operatorname{Ult}\mathfrak{B}$ de los ultrafiltros de $\mathfrak{B}$, equipado con la llamada topología de Stone. El resultado de toda esta teoría es que toda álgebra de Boole $\mathfrak{B}$ es isomorfa a $\operatorname{Clop}(\operatorname{Ult}\mathfrak{B})$ —para ser exactos, la aplicación $s\colon \mathfrak{B} \to P(\operatorname{Ult}\mathfrak{B})$ definida anteriormente es un isomorfismo entre $\mathfrak{B}$ y $\operatorname{Clop}(\operatorname{Ult}\mathfrak{B})$—. En particular, obtendremos una versión más fuerte del teorema de representación de Stone.

El álgebra $\operatorname{Clop}(X)$ de los subespacios clopen de un espacio topológico arbitrario es uno de los ejemplos estándar de álgebra de conjuntos. Para un espacio conexo X, sin embargo, este álgebra se reduce a $\{\varnothing, X\}$.

Recordemos algunos conceptos topológicos antes de entrar en materia.

Definición 42. Diremos que $\beta \subset \tau$ es una **base** de la topología τ si y solo si para todo punto p contenido en un abierto U existe $B \in \beta$ tal que $p \in B \subset U$. ◂

Equivalentemente, una familia β no vacía de subconjuntos de X formará base de alguna topología si se cumple:

- $\bigcup\{B : B \in \beta\} = X$.
- Para cualesquiera $B, B' \in \beta$ se verifica que $B \cap B' = \cup\beta_i$, para ciertos $\beta_i \in \beta$.

Ahora introduciremos algunos **axiomas de separación** en topología, ya que haremos uso de ellos más adelante. Un espacio topológico X se dice que es

- T_2 o **Hausdorff** si para todo $x, y \in X$ con $x \neq y$ existen U, V abiertos tales que $x \in U$, $y \in V$ y $U \cap V = \varnothing$.
- **normal** si para todo par de cerrados disjuntos $C, C' \subset X$ existen U, V abiertos disjuntos tales que $C \subset U$, $C' \subset V$ y $U \cap V = \varnothing$.
- T_4 si es normal y T_2.

Definición 43. Un espacio topológico X es **compacto** si todo cubrimiento por abiertos de X admite un subrecubrimiento finito. ◂

Proposición 44. *Si K es compacto y T_2, entonces K es T_4.*

Definición 45. Sea X un espacio topológico. Diremos que X es **cero-dimensional** si $\mathrm{Clop}(X)$ es una base de la topología de X. ◂

Definición 46. X es un **espacio de Stone** si es Hausdorff, compacto y cero-dimensional. ◂

Que un espacio sea cero-dimensional es equivalente a que exista una base o subbase para X formada por conjuntos clopen. Por ejemplo, el espacio de los números irracionales con la topología heredada de $\mathbb{R}$ es cero-dimensional, teniendo $(a, b) \cap (\mathbb{R} \setminus \mathbb{Q})$ con $a, b \in \mathbb{Q}$ como base de clopens.

Ejemplo 47. Veamos algunos ejemplos de espacios de Stone:

i) Todo espacio finito y discreto es de Stone. En particular, denotaremos como 2 al espacio de Stone $2 = \{0, 1\}$ con la topología discreta.
ii) Por el teorema de Tychonoff, el espacio producto de cualquier familia de espacios de Stone es de Stone. En particular, también lo es el espacio 2^I para cualquier conjunto de índices I, el cual es conocido como el espacio de Cantor de peso $|I|$, para un I infinito.
iii) Todo subespacio cerrado de un espacio de Stone es un espacio de Stone. ◂

Definición 48. Un espacio X diremos que es **conexo** si $\mathrm{Clop}(X) = \{\varnothing, X\}$, es decir, si X no es la unión de dos subconjuntos cerrados disjuntos no vacíos.

Un espacio topológico X es **totalmente disconexo** si ningún subespacio con al menos dos elementos es conexo. ◂

Teorema 49. *Un espacio de Hausdorff compacto es cero-dimensional, y por tanto de Stone, si y solo si es totalmente disconexo.*

Demostración. Sea X un compacto Hausdorff. Si X es cero-dimensional e Y un subespacio de X con dos puntos distintos y e y', entonces existe un clopen A tal que $y \in A$ e $y' \notin A$. Así que $Y \cap A$ es un subconjunto clopen propio no vacío de Y e Y no es conexo. Por consiguiente, X es totalmente disconexo.

Recíprocamente, supongamos que X es totalmente disconexo. Sean pues $x \in X$ y U un entorno abierto de x; debemos encontrar un clopen F de X tal que $x \in F \subseteq U$, probando así que $\mathrm{Clop}(X)$ es una base. Para ello definimos

$$\mathcal{F} = \{F \in \mathrm{Clop}(X) : x \in F\}, \quad q = \bigcap_{F \in \mathcal{F}} F.$$

Basta probar que $q \subseteq U$ y que q es clopen. En efecto, dado que X es compacto, U abierto y todo $F \in \mathcal{F}$ cerrado, para algún subconjunto finito F' de $\mathcal{F}$ se tiene que $\bigcap F' \subseteq U$ y podemos considerar $F = \bigcap F'$. Veamos que $q = \{x\}$; para ello, supongamos que q tiene al menos dos puntos, por lo que no es conexo al ser X totalmente disconexo. De este modo,

$$q = q_1 \cup q_2,$$

donde q_1, q_2 son cerrados disjuntos no vacíos de q. Ahora tenemos que q es un subconjunto cerrado de X; por tanto, cada q_i es cerrado en X. Por la compacidad y, por tanto, normalidad de X, podemos escoger dos abiertos disjuntos U_1, U_2 tales que $q_i \subseteq U_i$. Entonces, $q \subseteq U_1 \cup U_2$ y, nuevamente por compacidad,

$F \subseteq U_1 \cup U_2$ para algún $F \in \mathcal{F}$. De este modo, tanto $U_1 \cap F$ como $U_2 \cap F$ son clopen en F y, por tanto, como F es clopen, también en X. Dado que $x \in F$, podemos asumir que $x \in U_1 \cap F$. Entonces, $U_1 \cap F \in \mathcal{F}$ y $q \subseteq U_1 \cap F$. Esto implica que $q_2 \subseteq q \subseteq U_1$, lo cual contradice que $q_2 \subseteq U_2$, que los U_i fuesen disjuntos y que q_i fuese no vacío. ∎

A esto se debe que a los compactos Hausdorff en ocasiones se les denote como compactos totalmente disconexos o cero-dimensionales de manera indistinta.

Lema 50. *Sea X un espacio de Stone.*

i) *Si $B \subseteq \mathrm{Clop}(X)$ es una base para X cerrada bajo uniones finitas, entonces $B = \mathrm{Clop}(X)$.*
ii) *Si Y es un subespacio cerrado de X, entonces $\mathrm{Clop}(Y) = \{a \cap Y : a \in \mathrm{Clop}(X)\}$.*
iii) *Si y, z son subconjuntos cerrados y disjuntos de X, existe un subconjunto clopen de X que separa y de z, es decir, tal que $y \subseteq a$ y $z \subseteq X \setminus a$.*

Ahora estamos preparados para probar la versión topológica del teorema de representación de Stone.

Recordemos que $\mathrm{Ult}\,\mathfrak{B}$ es el conjunto de todos los ultrafiltros de un álgebra de Boole $\mathfrak{B}$ y que la aplicación de Stone es un monomorfismo de álgebras de Boole. En particular, $s(\mathfrak{B})$ es una familia de subconjuntos de $\mathrm{Ult}\,\mathfrak{B}$ cerrada bajo intersecciones finitas y, por tanto, base de una topología.

Definición 51. Para un álgebra de Boole $\mathfrak{B}$ se define la topología de Stone como la única topología en $\mathrm{Ult}\,\mathfrak{B}$ que tiene como base $s(\mathfrak{B})$.

$\mathrm{Ult}\,\mathfrak{B}$ dotado de la topología de Stone es el espacio de Stone o espacio dual o el espacio de los ultrafiltros de $\mathfrak{B}$. ◂

Teorema 52 (de representación de Stone, versión topológica). *Toda álgebra de Boole es isomorfa al álgebra de los clopen de un espacio de Stone. Más concretamente, el espacio dual $\mathrm{Ult}\,\mathfrak{B}$ de un álgebra de Boole $\mathfrak{B}$ es un espacio de Stone y la aplicación de Stone $s : \mathfrak{B} \to P(\mathrm{Ult}\,\mathfrak{B})$ es un isomorfismo entre $\mathfrak{B}$ y $\mathrm{Clop}(\mathrm{Ult}\,\mathfrak{B})$.*

Demostración. Sea $\mathfrak{B}$ un álgebra de Boole y $X = \mathrm{Ult}\,\mathfrak{B}$ su espacio dual. X es cero-dimensional puesto que todos los elementos de la base $s(a)$ son clopen, por ser $X \setminus s(a) = s(\neg a)$. Además, X es Hausdorff, pues si suponemos que F, G son dos ultrafiltros distintos de $\mathfrak{B}$, por la maximalidad de F podemos tomar $a \in F \setminus G$. Entonces, $s(a)$ y $s(\neg a)$ son entornos disjuntos de F y G.

Veamos que X es compacto. Para ello, sea U un recubrimiento abierto de X. Bastará considerar el caso en el que todo elemento de U es un elemento básico; por ello, consideremos $U = \{s(a) : a \in A\}$, con $A \subseteq \mathfrak{B}$. Supongamos que X no posee subrecubrimiento finito; entonces, para $n \in \mathbb{N}$ y $a_1, \dots, a_n \in A$ se tiene que

$$s(a_1) \cup \dots \cup s(a_n) \neq X = s(1),$$

y, por tanto, $a_1 \vee \dots \vee a_n \neq 1$ y $\neg a_1 \wedge \dots \wedge \neg a_n \neq 0$. De esto, se sigue que el conjunto $\neg A = \{\neg a : a \in A\}$ tiene la PIF. Por el teorema 38, consideremos ahora F un ultrafiltro de $\mathfrak{B}$ que contiene a $\neg A$. Entonces, para cada $a \in A$ se tiene que $\neg a \in F$, $a \notin F$ y $F \notin s(a)$, lo cual contradice que U es un cubrimiento de X.

Por tanto, X es un espacio de Stone. Dado que la aplicación de Stone s es un monomorfismo entre $\mathfrak{B}$ y $\mathrm{Clop}(X)$, por el lema 50 i) se tiene que $\mathrm{Clop}(X) = s(\mathfrak{B})$ considerando $B = s(\mathfrak{B})$ la base de X. ∎

Definición 53. Para un espacio de Stone X, llamamos álgebra dual de X a $\mathrm{Clop}(X)$. Para cada $x \in X$, el conjunto

$$t(x) = \{A \in \mathrm{Clop}(X) : x \in A\}$$

es un ultrafiltro de $\mathrm{Clop}(X)$. Esto define la aplicación de Stone

$$t : X \to \mathrm{Ult}\,\mathrm{Clop}(X).$$ ◂

De este modo, vemos que el teorema 52 puede enunciarse de forma que establezca que toda álgebra de Boole es isomorfa a su bidual, esto es, al álgebra dual de su espacio dual. Recíprocamente, todo espacio de Stone es homeomorfo a su bidual.

Teorema 54. *Todo espacio de Stone es homeomorfo al espacio de Stone de un álgebra de Boole. Más concretamente, para un espacio de Stone X, la aplicación $t\colon X \to \operatorname{Ult}\operatorname{Clop}(X)$ es un homeomorfismo entre X y $\operatorname{Ult}\operatorname{Clop}(X)$.*

Demostración. Basta probar que t es continua y biyectiva puesto que tanto X como $\operatorname{Ult}\operatorname{Clop}(X)$ son espacios Hausdorff y compactos.

Consideremos $\mathfrak{A} = \operatorname{Clop}(X)$. La continuidad de t se sigue del hecho de que las preimágenes de los conjuntos básicos $s(a)$, $a \in \mathfrak{A}$, son abiertos: para $a \in \mathfrak{A}$ y $x \in X$, se tiene $x \in t^{-1}(s(a))$ si y solo si $t(x) \in s(a)$ si y solo si $a \in t(x)$ si y solo si $x \in a$, de modo que $t^{-1}(s(a)) = a$ es clopen.

Para probar que t es inyectiva, consideremos x e y puntos distintos de X. Dado que X es un espacio de Stone, podemos tomar $a \in \operatorname{Clop}(X)$ tal que $x \in a$ e $y \notin a$. Por tanto, $a \in t(x) \setminus t(y)$.

Finalmente, para demostrar que t es suprayectiva debemos considerar un ultrafiltro $F \in \operatorname{Ult}\operatorname{Clop}(X)$. Como X es compacto y F es una familia de subconjuntos cerrados de X que posee la PIF, podemos tomar $x \in X$ de modo que $x \in a$ para cada $a \in F$. Entonces, $F \subseteq t(x)$ y, por la maximalidad del ultrafiltro F, se tiene que $F = t(x)$. ∎

Referencias

[1] Boole, George. *The mathematical analysis of logic*. Philosophical Library, 1847.

[2] Boole, George. *An investigation of the laws of thought, on which are founded the mathematical theories of logic and probabilities*. Dover Publications, Inc., New York, 1957, xi+424 pp. (1 plate).

[3] Corbalán Mirete, Clara María. *Sumas torcidas de espacios de Banach*. Trabajo de Fin de Máster. Universidad de Murcia, 2018. url: https://www.um.es/documents/118351/7120416/Clara+Corbal%C3%A1n+Mirete+-+TFM.pdf/7150b8f5-7023-4679-9a9c-2fe41f74157a.

[4] Jané, Ignacio. *Álgebras de Boole y lógica*. Vol. 5. Materials Docents. Edicions Universitat Barcelona, 1989. isbn: 978-84-7875-040-5.

[5] Koppelberg, Sabine. *Handbook of Boolean algebras*. Vol. 1. Edited by J. Donald Monk and Robert Bonnet. North-Holland Publishing Co., Amsterdam, 1989, págs. xx+312. isbn: 978-0-444-70261-6.

TEMat

ENEM '17 SEVILLA

Este trabajo colaboró con una conferencia plenaria durante el XVIII *Encuentro Nacional de Estudiantes de Matemáticas*, celebrado en Sevilla en julio de 2017.

Buscando triángulos en grafos muy grandes: un ejemplo de *property testing*

Alberto Espuny Díaz
University of Birmingham
axe673@bham.ac.uk

Resumen: El *property testing* hace referencia a un conjunto de técnicas algorítmicas que se han desarrollado para conseguir algoritmos decisionales que trabajen en tiempo sublineal en el tamaño de la entrada a cambio de perder precisión en la respuesta del algoritmo. En este artículo presentamos una breve discusión sobre el *property testing*, explicando el porqué de su utilidad y cómo valorar su eficiencia. En particular, nos centramos en algoritmos que comprueban propiedades en grafos, presentando los tres modelos más utilizados para comprobar estas propiedades y un ejemplo de cómo tratar una propiedad, la ausencia de triángulos, en cada uno de ellos.

Abstract: Property testing refers to a group of algorithmic techniques developed to achieve decision algorithms that work in sublinear time at the expense of some accuracy in the algorithms' output. In this paper, we present a brief discussion about property testing, explaining its usefulness and how to evaluate the efficiency of property testing algorithms. In particular, we consider algorithms that test graph properties. We present the three most common property testing models for graphs and, as an example, show how to test one property, triangle-freeness, in each of these models.

Palabras clave: algoritmos, grafos, *property testing*, libre de triángulos, regularidad, complejidad.

MSC2010: 05C85, 05D40, 68R10.

Recibido: 11 de septiembre de 2017.
Aceptado: 9 de septiembre de 2018.

Referencia: Espuny Díaz, Alberto. «Buscando triángulos en grafos muy grandes: un ejemplo de *property testing*». En: *TEMat*, 3 (2019), págs. 87-100. ISSN: 2530-9633. URL: https://temat.es/articulo/2019-p87.

1. Introducción

El desarrollo de los ordenadores y de los algoritmos que estos ejecutan ha permitido que a día de hoy podamos resolver en unos pocos segundos problemas que hace cien años habrían llevado meses de trabajo a un gran equipo de personas. Los problemas que se pueden resolver son muchos y muy variados, al igual que lo son las ideas subyacentes en los algoritmos. Sin embargo, y al contrario de una noción bastante extendida en la sociedad, no todos los problemas se pueden resolver rápido con nuestros ordenadores, y no es cuestión solo de conseguir ordenadores más potentes.

En general, podemos decir que un algoritmo es un proceso que, dada una entrada de datos (un conjunto de datos para el que queremos resolver un problema), sigue una serie de pasos hasta que resuelve el problema. En general, cuanto más grande sea el conjunto de entrada, mayor será el tiempo que tarda el algoritmo en resolver el problema, aunque sea solo porque necesita leer toda la entrada. Así, si el conjunto de entrada tiene tamaño n (en número de bits, aunque se acostumbra a utilizar el tamaño natural de la entrada), el tiempo que tarda en resolver el problema es una función de n, y esta función es a lo que llamamos la *complejidad temporal* del algoritmo. En general, nos interesa que los algoritmos resuelvan problemas para entradas de datos muy grandes (porque esas son las que nos cuesta mucho resolver a nosotros), de modo que nos interesa conocer cómo crece esta complejidad cuando n tiende a infinito.

Vamos a considerar un par de ejemplos. Supongamos que la entrada de datos es un grafo con n vértices; es muy razonable suponer que queremos conocer una propiedad del grafo de entrada. Por ejemplo, supongamos que queremos saber si el grafo es *bipartito*. En ese caso, es fácil demostrar que hay un algoritmo lineal que permite resolver el problema (lineal en el tamaño de la entrada, que en este caso no es n sino n más el número de aristas del grafo). Los algoritmos que tienen una complejidad lineal (y, en general, polinómica) se consideran «eficientes» (o «rápidos»).

El problema de saber si un grafo es bipartito es equivalente a saber si es 2-coloreable. En general, consideremos el problema de la k-*coloreabilidad* (es decir, queremos saber si los vértices del grafo se pueden colorear con k colores de manera que no haya dos vértices del mismo color unidos por una arista). Para cualquier valor de $k \geq 3$ se sabe que este problema es NP-completo [17], lo que quiere decir que los mejores algoritmos que conocemos para resolverlo tienen una complejidad *superpolinomial* (es decir, que crece más rápido que cualquier polinomio) en el tamaño de la entrada. De hecho, los algoritmos exactos que se conocen tienen complejidad exponencial. En general, se considera que los algoritmos que tienen una complejidad superpolinomial son «ineficientes». Los ejemplos de problemas que se comportan de este modo son muy variados, de manera que hay muchos problemas para los que un ordenador tardará mucho tiempo en dar una respuesta exacta si el conjunto de datos de la entrada es grande.

Consideremos ahora un algoritmo eficiente (cuya complejidad sea polinomial), pero supongamos que queremos estudiar una entrada de datos enorme. Por ejemplo, supongamos que queremos estudiar el grafo de internet, o de Facebook, o estudiar una propiedad en una base de datos enorme de una empresa. En estos casos, aunque el algoritmo sea eficiente, el simple hecho de leer toda la entrada de datos y procesarla ya supone un coste absurdo, además de que solo el almacenamiento de los datos ya supondría un problema (por el espacio de memoria que tienen los ordenadores).

Así pues, la pregunta natural es «¿qué podemos hacer en estos casos en los que queremos resolver un problema pero no tenemos suficientes recursos temporales?».

2. Complejidad y notación asintótica

El análisis de algoritmos es una tarea imprescindible para saber si su implementación es adecuada o si los recursos que necesitan para resolver un problema son excesivos. De este análisis se desprende la idea de que los algoritmos sean eficientes o no, y a partir de este estudio se han desarrollado muchos problemas en el área de la teoría de la computación.

En general, dado un algoritmo que resuelve un problema, el objetivo del análisis de algoritmos es determinar los recursos (temporales, espaciales o de cualquier otro tipo) que el algoritmo necesita para resolver el problema. La cantidad de recursos que necesita depende del tamaño de la entrada, pero distintas entradas con el mismo tamaño pueden necesitar distintos recursos. En ese caso, se toma una cota superior sobre los recursos que necesita considerando el peor caso de entre todas las entradas con el mismo tamaño. Los recursos más estudiados habitualmente son el tiempo y el espacio, aunque en este artículo no nos centraremos en el espacio.

Definición 1. La **complejidad temporal** de un algoritmo para cada valor de n se define como el máximo número total de operaciones atómicas que realiza el ordenador sobre todas las entradas de tamaño n. ◂

Nótese que esta definición no tiene que tener una equivalencia exacta con el tiempo real que tarda el algoritmo. Este tiempo depende de muchos otros factores, como el ordenador empleado, la distribución de los datos en la memoria o el tipo de operaciones que el ordenador realiza. Sin embargo, por simplicidad, se suele trabajar en el *modelo de coste uniforme*, según el cual todas las operaciones que realiza el ordenador tienen el mismo coste constante (operaciones atómicas). Este modelo es bastante adecuado si se puede garantizar que el tiempo de todas las operaciones está acotado por alguna constante.

A la hora de estudiar la complejidad de algoritmos (y también el comportamiento de muchas estructuras combinatorias que dependen de un cierto parámetro de tamaño n) muchas veces es interesante saber cuál es su crecimiento asintótico en función del parámetro n cuando este tiende a infinito. Esto es especialmente útil cuando solo interesa saber cómo se comparan las complejidades de dos algoritmos diferentes. Para hablar de estos comportamientos asintóticos es útil emplear la *notación asintótica*.

Definición 2. Sean f y g dos funciones de n que toman solo valores no negativos.

- Decimos que $f(n) = \mathcal{O}(g(n))$ (y se lee que «f es **o grande** de g») si y solo si existen constantes $M \in \mathbb{R}$ y $n_0 \in \mathbb{N}$ tales que $f(n) \leq Mg(n)$ para todo $n \geq n_0$.
- Decimos que $f(n) = o(g(n))$ (y se lee que «f es **o pequeña** de g») si y solo si para toda constante $\delta > 0$ existe una constante $N \in \mathbb{N}$ tal que $f(n) \leq \delta g(n)$ para todo $n \geq N$. Equivalentemente, si g no se anula a partir de un cierto punto, podemos decir que $f(n) = o(g(n))$ si $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$.
- Decimos que $f(n) = \Omega(g(n))$ (y se lee que «f es **omega grande** de g») si y solo si existen constantes $M \in \mathbb{R}$ y $n_0 \in \mathbb{N}$ tales que $f(n) \geq Mg(n)$ para todo $n > n_0$. Equivalentemente, podemos decir que $f(n) = \Omega(g(n))$ si y solo si $g(n) = \mathcal{O}(f(n))$.
- Decimos que $f(n) = \Theta(g(n))$ (y se lee que «f es **theta** de g») si y solo si existen constantes $k_1, k_2 \in \mathbb{R}$ y $n_0 \in \mathbb{N}$ tales que $k_1 g(n) \leq f(n) \leq k_2 g(n)$ para todo $n \geq n_0$. Equivalentemente, decimos que $f(n) = \Theta(g(n))$ si y solo si $f(n) = \mathcal{O}(g(n))$ y $f(n) = \Omega(g(n))$. ◂

Nótese que aquí el uso de «=» es un abuso de notación. Se puede utilizar $\mathcal{O}(g(n))$ para denotar el conjunto de todas las funciones f que cumplen la definición 2 (y lo mismo se puede hacer para las otras notaciones presentadas). De este modo, el símbolo «=» indica que f pertenece a ese conjunto, no la igualdad.

Al trabajar con la complejidad (temporal, digamos) de un algoritmo es muy habitual utilizar este tipo de notación. Así, por ejemplo, se pueden programar algoritmos para determinar si un grafo $G = (V, E)$ es bipartito o no con complejidad $\mathcal{O}(|V| + |E|)$. En general, la notación o grande se utiliza para dar cotas superiores sobre la complejidad.

Es habitual hablar de la «complejidad» de un problema para referirnos a la complejidad de un algoritmo que resuelve el problema de manera óptima. Si sabemos que ningún algoritmo podrá resolver el problema con una complejidad menor que una determinada función, esta constituye una cota inferior, y podemos utilizar la notación omega grande para denotar esta cota.

3. *Property testing*

La idea básica del *property testing* surge como una respuesta natural a la pregunta final de la introducción: «¿Qué podemos hacer si queremos resolver un problema pero no tenemos suficientes recursos temporales?». Digamos que tenemos una cierta estructura de datos y queremos saber si cumple una cierta propiedad (por ejemplo, dada una función f como un conjunto de pares (x, y), donde $y = f(x)$, podemos querer saber si f es lineal o no). Nuestro objetivo es desarrollar algoritmos que nos permitan saber esto y cuya complejidad sea mejor que la de los algoritmos exactos. Como nuestros algoritmos exactos no son suficientemente eficientes para nuestros objetivos, a cambio estamos dispuestos a sacrificar parte del conocimiento que queremos obtener; en particular, en lugar de saber si la propiedad se cumple o no, nos puede interesar saber si se cumple o está «lejos» de cumplirse, y nos podemos conformar con saber esto no de manera exacta, pero sí con una probabilidad suficientemente grande. Con esta premisa, nos gustaría desarrollar algoritmos *ultraeficientes*.

Para ser precisos, un algoritmo de *property testing* toma una decisión aproximada sobre si la estructura de datos presenta o está «lejos» de presentar la propiedad estudiada, donde «lejos» quiere decir que la entrada se debe modificar de un modo no negligible para que llegue a tener la propiedad. Para ello, al algoritmo se le proporciona un parámetro de *distancia*, ε, y el algoritmo deberá aceptar la entrada con probabilidad alta si presenta la propiedad, rechazarla con probabilidad alta si está ε-lejos de presentarla, y puede contestar cualquier cosa en el caso restante. Por probabilidad alta queremos decir probabilidad mayor que $2/3$, aunque esta es una convención, y se podría tomar cualquier probabilidad mayor que $1/2$. La definición de distancia se debe especificar en cada problema concreto.

Podemos observar que el solo hecho de leer toda la entrada ya supone un coste lineal para el algoritmo, de modo que no vamos a poder mejorar esto. En su lugar, lo que suponemos es que el algoritmo no lee la entrada, sino que los datos de entrada están almacenados en algún sitio y nuestro algoritmo tiene acceso a un *oráculo* que conoce estos datos. El algoritmo tiene permitido hacer *consultas* a este oráculo, de manera que podemos leer solo una parte pequeña de los datos, ahorrando así mucho tiempo al algoritmo. Nuestro objetivo general, ahora, es encontrar algoritmos con complejidad *sublineal*; en particular, el uso de esta teoría ha permitido desarrollar algoritmos que tienen una complejidad *independiente del tamaño de la entrada*, es decir, podríamos decir que tienen complejidad *constante*.

El *property testing* se comenzó a estudiar en los años 90, siendo utilizado implícitamente por Blum, Luby y Rubinfeld [7] y definido formalmente por primera vez por Rubinfeld y Sudan [25] en 1996. Más tarde, Goldreich, Goldwasser y Ron [12] extendieron la definición a un contexto más general y empezaron a estudiar propiedades de grafos. El estudio del *property testing* se ha hecho muy generalizado desde entonces y se han obtenido aplicaciones muy variadas, como algoritmos para trabajar con diferentes propiedades algebraicas de funciones, con funciones lógicas, propiedades geométricas, de lenguajes restringidos, de distribuciones, etc. Pero quizá el área en la que más énfasis se ha dado al *property testing* es en combinatoria y, en particular, en teoría de grafos. Nosotros nos vamos a concentrar en una de estas aplicaciones; para leer sobre otros muchos problemas, técnicas y aplicaciones se puede acudir al estudio de Ron [24], aunque también existen otros estudios menos enfocados a la parte combinatoria y numerosos artículos.

Dado un algoritmo cualquiera, el estudio de su complejidad es siempre necesario para saber si implementarlo vale la pena. En el caso del *property testing* vamos a considerar especialmente dos tipos de recursos: el tiempo y el número de consultas al oráculo. Podemos hablar así de complejidad de consultas.

Definición 3. La **complejidad de consultas** para cada valor de n de un algoritmo con acceso a un oráculo se define como el máximo número total de consultas que realiza el algoritmo sobre todas las entradas de tamaño n. ◂

En este artículo nos centramos en trabajar con grafos. Recordamos que un **grafo** se define como un par $G = (V, E)$ donde V es un conjunto de **vértices** y $E \subseteq V \times V$ es un conjunto de pares de vértices, que se denominan **aristas**. Por simplicidad, denotaremos una arista (u, v) como uv. Los vértices de un grafo se suelen representar mediante puntos, y las aristas, mediante líneas que unen los puntos. Dada una arista, decimos que es **incidente** a un vértice si uno de sus dos extremos es dicho vértice. El número de aristas incidentes a un vértice es su **grado**. Para un vértice v, decimos que un vértice u es **vecino** de v si $uv \in E$. Decimos que un grafo es **simple** si no hay aristas múltiples (dos o más aristas uniendo los mismo vértices) ni aristas que unan un vértice consigo mismo. Aquí nos centraremos solo en grafos simples.

Sea $G = (V, E)$ un grafo simple con $|V| = n$. A la hora de dar este grafo como una entrada para un algoritmo, los vértices se deben etiquetar, es decir, se les da un orden $v_1, v_2, \dots, v_n$. Las formas de dar las aristas pueden ser diferentes, y los algoritmos que se desarrollan dependen de esta estructura de datos.

Para trabajar con problemas de grafos en *property testing* se han definido varios modelos. Cada uno de ellos tiene sus particularidades y su utilidad, y es interesante en general estudiar problemas en los distintos modelos.

3.1. El modelo de grafos densos

Una de las estructuras de datos tradicionales para almacenar grafos es la que se denomina la **matriz de adyacencia**, una matriz cuadrada $A = (a_{i,j})_{i,j=1}^{n}$ de tamaño $n \times n$ en la que las filas y columnas se numeran con las etiquetas de los vértices, y la entrada $a_{i,j}$ toma el valor 1 si $v_i v_j \in E(G)$, y 0 en caso contrario.

Dada esta estructura de datos, las consultas que se le pueden hacer al oráculo son del tipo «¿hay una arista entre los vértices v_i y v_j del grafo?», y la respuesta del oráculo deberá ser «sí» o «no». A estas consultas las denominaremos **consultas de par de vértices**. La idea de distancia que emplearemos también tiene que ver con esta representación del grafo. Diremos que un grafo G está ε-lejos de tener una propiedad si hay que modificar por lo menos εn^2 entradas de la matriz de adyacencia para que G presente dicha propiedad, es decir, si hay que modificar $\varepsilon n^2/2$ aristas del grafo. Nótese que el número máximo de aristas que un grafo puede tener es $\binom{n}{2} \approx n^2/2$. Este modelo fue introducido por Goldreich, Goldwasser y Ron [12].

El motivo por el que este modelo se conoce como el de grafos densos tiene que ver con esta noción de distancia. Sea $\{G_n : n \in \mathbb{N}\}$ una secuencia de grafos tales que G_n tiene n vértices. Decimos que los grafos de la secuencia son **densos** (a partir de un índice n_0) si existe alguna constante δ tal que todos los grafos G_n con $n \geq n_0$ tienen al menos δn^2 aristas. Por comodidad, no hablaremos de secuencias de grafos, sino solo de grafos densos, pero siempre con la idea subyacente de que el tamaño de los grafos tiende a infinito. Es en este tipo de grafos en los que es más interesante estudiar propiedades en este modelo, ya que, por ejemplo, cualquier grafo no denso nunca estará ε-lejos de una propiedad que se pueda conseguir eliminando aristas, de modo que nuestros algoritmos no podrán determinar si la tiene o no.

3.2. El modelo de grado acotado

La otra estructura de datos tradicional para presentar los grafos se conoce como la **lista de incidencia**, y consiste en una lista de los vértices vecinos a v_i para cada $i \in \{1, \dots, n\}$. El orden en que aparecen estos vecinos puede ser arbitrario, e induce un doble etiquetado de las aristas del grafo. Si el grado de todos los vértices está uniformemente acotado por una cota d, esta estructura de datos presenta toda la información sobre las aristas del grafo de una forma bastante compacta.

Si tenemos esta estructura de datos, para cada vértice u las consultas que se le pueden hacer al oráculo son del tipo «¿cuál es el i-ésimo vecino de u?». La respuesta del oráculo deberá ser o bien un vértice v, que aparece en la i-ésima posición en la lista de incidencia de u, o bien un símbolo especial si el grado de u es menor que i. A estas consultas las denominaremos **consultas de vecindad**. La distancia que definimos también tiene que ver con la estructura de datos: diremos que un grafo está ε-lejos de tener una propiedad si hay que modificar más de εnd entradas de la lista de incidencia para que obtenga la propiedad. Nótese que el máximo número de aristas que puede tener el grafo, dada la cota sobre el grado, es $nd/2$.

Este modelo, introducido por Goldreich y Ron [13], es especialmente útil para grafos en los que el número de aristas es del orden de nd, es decir, aquellos en los que el grado medio del grafo es del mismo orden que el grado máximo. En particular, como veremos, esto es cierto si d es una constante.

3.3. El modelo de grafos generales

El modelo de grafos generales fue introducido por Kaufman, Krivelevich y Ron [18] después de que Parnas y Ron [22] propusieran separar el modelo de la estructura de datos utilizada e introdujesen un modelo intermedio. Así, sin conocer la estructura de datos empleada, el oráculo puede responder tanto a consultas de par de vértices como de vecindad. Además, a veces también se le permite responder a la consulta «¿cuál es el grado del vértice u?», aunque nosotros no utilizaremos este tipo de consultas.

Dado que la estructura de datos no se conoce, la distancia en este modelo no se mide respecto al máximo número de aristas que el grafo podría tener, sino respecto al número de aristas que tiene. Así, decimos que un grafo $G = (V, E)$ está ε-lejos de presentar una propiedad si hay que modificar al menos $\varepsilon|E|$ aristas para que la propiedad se dé.

El modelo de grafos generales es hasta ahora el menos estudiado de los tres. Presenta la ventaja de que se puede trabajar con todo tipo de grafos, pero el análisis de algoritmos en él puede llegar a ser muy complejo. En cualquier caso, cabe insistir en que los modelos no son equivalentes. En este sentido, un mismo problema estudiado en modelos diferentes puede tener soluciones también diferentes. Veremos un ejemplo de esto más adelante.

4. El problema de la ausencia de triángulos

Sea $G = (V, E)$ un grafo. Sean $u, v, w \in V(G)$ tres vértices cualesquiera. Si $uv, uw, vw \in E(G)$, decimos que estas tres aristas forman un **triángulo**. Si G no tiene ningún triángulo, decimos que G está **libre de triángulos**. A la propiedad de que G esté libre de triángulos la llamamos **ausencia de triángulos**.

Supongamos que tenemos un grafo grande y queremos saber si está libre de triángulos. Un algoritmo exacto que comprueba si el grafo está libre de triángulos mediante una búsqueda en anchura tiene una complejidad lineal ($\mathcal{O}(|V| + |E|)$), de modo que es un problema que se resuelve de manera muy eficiente, pero ¿qué pasa si el tamaño del grafo es tal que un tiempo lineal no es suficiente, o que nuestro ordenador no tiene suficiente memoria para almacenar los datos? Vamos a intentar resolver este problema como un ejemplo de *property testing* en cada uno de los modelos.

4.1. Ausencia de triángulos en el modelo de grado acotado

El problema de la ausencia de triángulos en el modelo de grado acotado fue resuelto por Goldreich y Ron [13] en el mismo artículo en que introdujeron dicho modelo. La solución que dieron se basa en dar directamente un algoritmo bastante sencillo:

Algoritmo 1 (comprueba la ausencia de triángulos para grafos en el modelo de grado acotado).

1: **Entrada:** tamaño del grafo n, grado máximo d y parámetro de distancia ε
2: seleccionar uniforme e independientemente $s = \Theta(1/\varepsilon)$ vértices de G
3: etiquetar los vértices seleccionados como $v_1, v_2, \dots, v_s$
4: **para** $i = 1$ hasta s **hacer**
5:     consultar todos los vecinos de v_i
6:     **para** cada vecino u de v_i **hacer**
7:         consultar todos los vecinos de u
8:         comprobar si algún vecino de u es vecino de v_i
9:         **si** sí, **entonces**
10:             rechazar
11:         **fin si**
12:     **fin para**
13: **fin para**
14: aceptar

Lo que hace este algoritmo es buscar triángulos de los que forme parte alguno de los vértices v_i seleccionados. Para cada $i \in \{1, \dots, s\}$, lo que hace el algoritmo es seleccionar todos los vecinos de v_i y, para cada uno de estos, comprobar si cierra un triángulo con alguno de los otros vecinos (línea 8). Si el algoritmo consigue encontrar un solo triángulo, entonces rechaza la entrada (es decir, el algoritmo afirma que el grafo está ε-lejos de la ausencia de triángulos); si no encuentra ningún triángulo, entonces acepta la entrada.

Teorema 4. *El algoritmo 1 es un algoritmo que comprueba la ausencia de triángulos en un grafo en el modelo de grado acotado con $\mathcal{O}(d^2/\varepsilon)$ consultas en tiempo $\mathcal{O}(d^3/\varepsilon)$.*

Demostración. Para demostrar este teorema hay que comprobar varias cosas: que el algoritmo es correcto (es decir, que dada una entrada cualquiera contesta de forma correcta con las probabilidades pertinentes) y que se cumplen las dos cotas de complejidad dadas.

En primer lugar, es muy fácil observar que si el grafo sobre el que se hacen consultas no tiene triángulos, el algoritmo nunca podrá encontrar ninguno y aceptará la entrada con probabilidad 1. Ahora concentrémonos en el bucle que empieza en la línea 6. En cada iteración el algoritmo realiza como mucho d consultas para obtener los vecinos de v_i, y después d consultas para cada uno de ellos, lo que supone un total de como máximo $d^2 + d$ consultas. Dado que al bucle se entra $\Theta(1/\varepsilon)$ veces, está claro que el número total de consultas es $\mathcal{O}\left(d^2/\varepsilon\right)$. En cuando al tiempo que tarda el algoritmo en total, en el mismo bucle tenemos que comprobar si se cierran triángulos. Para esto, para cada uno de los vecinos u de v_i (de los que hay como máximo d) comparamos cada vecino de u con cada uno de los vecinos de v_i; esto supone un máximo de como mucho d^3 comparaciones. De este modo, la cota $\mathcal{O}\left(d^3/\varepsilon\right)$ sobre la complejidad temporal del algoritmo está clara.

Finalmente, queda comprobar que si el grafo que estudiamos está ε-lejos de la ausencia de triángulos, entonces el algoritmo 1 rechazará la entrada con probabilidad al menos 2/3. Por definición, si el grafo está ε-lejos de la ausencia de triángulos, entonces hay que realizar al menos εnd modificaciones sobre las aristas para convertirlo en un grafo libre de triángulos. Como añadir aristas nunca nos ayuda a eliminar triángulos, esto quiere decir que debemos eliminar al menos εnd aristas para eliminar todos los triángulos del grafo, lo cual significa que hay al menos εnd aristas del grafo que forman parte de triángulos. Ahora bien, como cada vértice tiene grado como máximo d, eso quiere decir que hay al menos εn vértices del grafo que forman parte de triángulos. Así, al elegir los s vértices se puede tomar una constante (implícita en la notación Θ) suficientemente grande como para que la probabilidad de que al menos uno de los s vértices seleccionados pertenezca a un triángulo sea al menos 2/3. Y, entonces, el algoritmo encontrará ese triángulo y rechazará la entrada. ■

En particular, podemos calcular el valor de esa constante implícita. Como la muestra se toma de manera independiente, tenemos que la probabilidad de que cada uno de los vértices elegidos pertenezca a un triángulo es mayor o igual que ε, de modo que la probabilidad de que no pertenezca a ningún triángulo es menor o igual que $1-\varepsilon$. Después de tomar s vértices, la probabilidad de que ninguno pertenezca a un triángulo es menor o igual que $(1-\varepsilon)^s$, así que la probabilidad de que alguno de los vértices pertenezca a un triángulo es

$$\Pr(\text{algún vértice de la muestra pertenezca a un triángulo}) \geq 1-(1-\varepsilon)^s \geq 1-\mathrm{e}^{-\varepsilon s}.$$

Como queremos que esta probabilidad sea de al menos 2/3, tenemos que

$$1-\mathrm{e}^{-\varepsilon s} \geq \frac{2}{3} \iff e^{-\varepsilon s} \leq \frac{1}{3} \iff \varepsilon s \geq \log 3 \iff s \geq \frac{\log 3}{\varepsilon},$$

de modo que tomar $s = \log 3/\varepsilon$ es suficiente para que el algoritmo funcione.

Es interesante remarcar el hecho de que la cota que hemos obtenido sobre el número de consultas *no depende de n*, como tampoco lo hace la complejidad temporal del algoritmo. Así, podemos ir aumentando el tamaño del grafo tanto como queramos, y nuestro algoritmo siempre tardará el mismo tiempo en decidir.

Por otra parte, si queremos afinar más nuestra decisión, nos interesará variar el parámetro ε. Si el parámetro es muy pequeño podremos distinguir grafos que están más cerca según nuestra distancia, con lo cual podríamos decir que la respuesta del algoritmo es mejor en este sentido (distingue grafos que no distinguiría para valores de ε más grandes). Como queremos variar este parámetro, también es bueno saber cómo varía la complejidad del algoritmo respecto a él. Y lo que tenemos en este caso, en virtud del teorema anterior, es que la complejidad (tanto temporal como del número de consultas) es lineal en el inverso de ε, es decir, el algoritmo es bastante eficiente también en este sentido.

4.2. Ausencia de triángulos en el modelo de grafos densos: regularidad

Vamos a considerar ahora el mismo problema, pero trabajando en el modelo de grafos densos. Así, podemos suponer que los grafos que queremos estudiar tienen un número cuadrático de aristas. Para poder resolver bien el problema en este contexto vamos a necesitar una serie de definiciones y un resultado central en la combinatoria extremal de los últimos cincuenta años.

Dado un grafo $G = (V, E)$, sean A y B dos conjuntos no vacíos de vértices tales que $A \cap B = \varnothing$. Supongamos que $E(A, B)$ representa el conjunto de aristas cuyos vértices están uno en A y otro en B.

Definición 5. La **densidad** del par A, B se define como

$$d(A,B) = \frac{|E(A,B)|}{|A||B|}.$$ ◂

Definición 6. Se dice que el par A, B es γ**-regular**, para algún $\gamma \in [0,1]$, si para todo par de conjuntos $A' \subseteq A$, $B' \subseteq B$ con $|A'| \geq \gamma|A|$ y $|B'| \geq \gamma|B|$ se cumple que $|d(A',B') - d(A,B)| < \gamma$. ◂

Una de las herramientas centrales en combinatoria extremal hoy en día, aunque también tiene muchas aplicaciones en otros campos, es el siguiente lema, debido a Szemerédi [26].

Lema 7 (lema de regularidad de Szemerédi). *Para cualquier $\ell_0 \in \mathbb{N}$ y cualquier $\gamma \in (0,1]$ existe un entero $M = M(\ell_0, \gamma)$ tal que todo grafo $G = (V,E)$ con $n \geq M$ vértices tiene una partición $\mathcal{A} = \{V_1, \dots, V_\ell\}$ de V con $||V_i| - |V_j|| \leq 1$ para todo $i, j \in \{1, \dots, \ell\}$, donde $\ell_0 \leq \ell \leq M$, tal que todos los pares (V_i, V_j) excepto como mucho $\gamma\binom{\ell}{2}$ son γ-regulares.*

La demostración de este resultado no es complicada, pero nos interesa más ver una de sus aplicaciones. Vamos a presentar un algoritmo muy simple que comprueba si un grafo cualquiera está libre de triángulos o está ε-lejos de estarlo, y vamos a utilizar el lema 7 para analizarlo.

Algoritmo 2 (comprueba la ausencia de triángulos para grafos en el modelo de grafos densos).

```
Entrada: tamaño del grafo n y parámetro de distancia ε
M ← M(8/ε, ε/8)
seleccionar uniforme e independientemente s = Θ(M²/ε³) vértices de G
consultar todos los pares de vértices
si encuentra un triángulo, entonces
    rechazar
en caso contrario
    aceptar
fin si
```

Teorema 8. *El algoritmo 2 es un algoritmo que comprueba la ausencia de triángulos en un grafo en el modelo de grafos densos con complejidad temporal y de consultas $\mathcal{O}\left(M^4/\varepsilon^6\right)$.*

Demostración. En el caso de que el grafo de la entrada no tenga ningún triángulo, está claro que el algoritmo 2 no podrá encontrar ninguno, así que dirá que el grafo está libre de triángulos con probabilidad 1. De este modo, debemos comprobar que el algoritmo es correcto cuando el grafo $G = (V,E)$ de la entrada está ε-lejos de la ausencia de triángulos.

Por el lema 7, tomando $\ell_0 = 8/\varepsilon$ y $\gamma = \varepsilon/8$, sabemos que existe una partición $\mathcal{A} = \{V_1, \dots, V_\ell\}$ de los vértices de G en ℓ partes, $\ell_0 \leq \ell \leq M = M(\varepsilon)$, con las propiedades dadas por el lema. Cada una de las partes tendrá tamaño $\lfloor n/\ell \rfloor$ o $\lceil n/\ell \rceil$; como trabajamos con problemas asintóticos, tenemos que n tiende a infinito y podemos despreciar los errores de redondeo, y trabajar solo con n/ℓ. Alternativamente, también se puede pensar que tomamos n múltiplo de ℓ para este desarrollo.

Si nos fijamos en una cualquiera de las partes de la partición, el número máximo de aristas que puede haber en su interior es $\binom{n/\ell}{2} \leq \frac{1}{2}(n/\ell)^2$, lo que supone un total de como máximo $\frac{\ell}{2}(n/\ell)^2 = \frac{1}{2\ell}n^2 \leq \frac{1}{2\ell_0}n^2 = \frac{\varepsilon}{16}n^2$ aristas. Si definimos un grafo G_1 como el grafo que se obtiene de G después de eliminar todas estas aristas, tenemos que G_1 está al menos $(\frac{15}{16}\varepsilon)$-lejos de estar libre de triángulos.

Consideremos ahora las parejas de la partición de G que no son regulares. Por el lema 7, hay como máximo $\frac{\varepsilon}{8}\binom{\ell}{2} \leq \frac{\varepsilon}{16}\ell^2$ parejas de estas, y cada pareja puede formar un total de como máximo $(n/\ell)^2$ aristas, de modo que en total hay como máximo $\frac{\varepsilon}{16}\ell^2(n/\ell)^2 = \frac{\varepsilon}{16}n^2$ aristas entre las parejas no regulares. Si definimos un grafo G_2 como el que se obtiene de G_1 después de eliminar todas las aristas que hay entre las parejas no regulares, tenemos que G_2 está al menos $(\frac{7}{8}\varepsilon)$-lejos de la ausencia de triángulos.

Consideremos, finalmente, las parejas (V_i, V_j) de la partición con $d(V_i, V_j) < \varepsilon/2$, es decir, aquellas en las que $|E(V_i, V_j)| < \frac{\varepsilon}{2}(n/\ell)^2$. El número de parejas con esta propiedad es como máximo $\binom{\ell}{2} \leq \ell^2/2$ (en el caso de que todas tengan la propiedad), de modo que el número total de aristas que hay entre todas estas parejas de la partición es de como máximo $\frac{\varepsilon}{4}n^2$. Si definimos un grafo G_3 como el que se obtiene de G_2 después de eliminar todas estas aristas, tenemos que G_3 está al menos $(\frac{5}{8}\varepsilon)$-lejos de la ausencia de triángulos. En particular, como nuestro grafo aún tiene triángulos, tienen que existir tres índices i, j y k tales que $d(V_i, V_j), d(V_j, V_k), d(V_i, V_k) \geq \varepsilon/2$. Todo este proceso de eliminación de aristas se muestra en la figura 1.

Vamos a demostrar ahora que el hecho de que exista esta terna (V_i, V_j, V_k) tal que todas las parejas son $\frac{\varepsilon}{8}$-regulares y tienen densidad al menos $\varepsilon/2$ implica que en el grafo va a haber muchos triángulos, y que una muestra suficientemente grande conseguirá capturar alguno de estos. Sea $v \in V_i$ un vértice cualquiera. Sea $\Gamma_j(v)$ el conjunto de vecinos de v en V_j, y sea $\Gamma_k(v)$ el conjunto de vecinos de v en V_k. Vamos a decir que v es *útil* si $|\Gamma_j(v)| \geq \frac{\varepsilon}{4}\frac{n}{\ell}$ y $|\Gamma_k(v)| \geq \frac{\varepsilon}{4}\frac{n}{\ell}$. Como (V_j, V_k) es $\frac{\varepsilon}{8}$-regular, si v es útil entonces

$$|E\left(\Gamma_j(v), \Gamma_k(v)\right)| \geq \left(d\left(V_j, V_k\right) - \frac{\varepsilon}{8}\right)\left(\frac{\varepsilon}{4}\right)^2\left(\frac{n}{\ell}\right)^2 \geq \frac{3\varepsilon^3}{256\ell^2}n^2,$$

ya que $d\left(V_j, V_k\right) \geq \varepsilon/2$ en G_3. Teniendo en cuenta que $\ell \leq M$, esto quiere decir que si tenemos un vértice útil $v \in V_i$ y tomamos una muestra adicional de $\mathcal{O}\left(M^2/\varepsilon^3\right)$ pares de vértices, con una constante implícita suficientemente grande, entonces encontraremos uno de estos triángulos con probabilidad al menos 2/3. Y para esto basta, claramente, con tomar una muestra de $\mathcal{O}\left(M^2/\varepsilon^3\right)$ vértices y considerar todas las posibles parejas.

Finalmente, queda demostrar que el número de vértices útiles es suficientemente grande, de modo que una muestra también suficientemente grande podrá encontrar uno de ellos con probabilidad al menos 2/3. Para ello, consideremos un vértice $v \in V_i$ que no es útil. Diremos que v es *inútil en j* si $|\Gamma_j(v)| < \frac{\varepsilon}{4}\frac{n}{\ell}$, y que es *inútil en k* si $|\Gamma_k(v)| < \frac{\varepsilon}{4}\frac{n}{\ell}$. Podemos suponer sin pérdida de generalidad que el número de vértices inútiles en j es mayor o igual que el de vértices inútiles en k. Supongamos que al menos la mitad de los vértices no son útiles. Entonces, al menos la cuarta parte son inútiles en j. Sea V_i' el conjunto de dichos vértices, de modo que $|V_i'| > \gamma|V_i|$ ya que $\gamma = \frac{\varepsilon}{8} \leq \frac{1}{8}$, y sea $V_j' = V_j$. Entonces,

$$d\left(V_i', V_j'\right) \leq \frac{|V_i'|\frac{\varepsilon}{4}\frac{n}{\ell}}{|V_i'||V_j'|} = \frac{\varepsilon}{4} < \frac{3}{8}\varepsilon \leq d\left(V_i, V_j\right) - \gamma,$$

lo que contradice la regularidad del par (V_i, V_j). Por lo tanto, el número de vértices útiles en V_i es de al menos $\frac{n}{2\ell}$, y una muestra independiente de $\mathcal{O}(M)$ vértices contendrá uno de estos con probabilidad al menos 2/3.

Así, se puede tomar una muestra de $\mathcal{O}(M) + \mathcal{O}\left(M^2/\varepsilon^3\right) = \mathcal{O}\left(M^2/\varepsilon^3\right)$ vértices, que contendrá al menos un vértice útil y un par de vértices que formen un triángulo con el anterior con probabilidad al menos 2/3. Al consultar todos los pares de vértices, el algoritmo descubrirá el triángulo y rechazará la entrada, lo que demuestra que el algoritmo es correcto. El número de consultas es $\mathcal{O}\left(M^4/\varepsilon^6\right)$, ya que el número de posibles aristas es cuadrático en el número de vértices seleccionados. La complejidad temporal es la misma, ya que se puede buscar un triángulo en el grafo inducido por la muestra con una búsqueda en anchura, que tiene un coste lineal en el tamaño del grafo que se estudia. ■

Como se desprende de la demostración, el teorema 8 establece que la complejidad, tanto temporal como de consultas, del algoritmo 2 es independiente del tamaño de la entrada, igual que en el caso del modelo de grado acotado. Sin embargo, el comportamiento con respecto a ε es mucho menos eficiente que en el caso anterior. Aunque el exponente de la cota sobre el número de consultas que da el teorema 8 se puede mejorar con un análisis más cuidadoso, la cota sobre M que da el lema de regularidad de Szemerédi es una torre de altura polinomial en el inverso de ε.

Definición 9. Denotamos por $T(n)$ la función **torre de altura** n, que se define de manera recursiva como $T(1) = 2$ y $T(n+1) = 2^{T(n)}$ para todo $n \geq 1$. ◂

La mejor cota superior sobre el valor de M que se obtiene del lema 7 es de la forma $T\left(\mathcal{O}(1/\varepsilon^2)\right)$, un número enorme que hace que este algoritmo no sea realmente aplicable para resolver este problema, ya que ningún ordenador tiene capacidad de memoria suficiente (aunque para grafos suficientemente grandes, sigue siendo mejor esta cota que cualquier otra que sea una función que tienda a infinito con n). Siendo esto una cota superior, nos podríamos plantear si no se puede mejorar utilizando otras técnicas. Sin embargo,

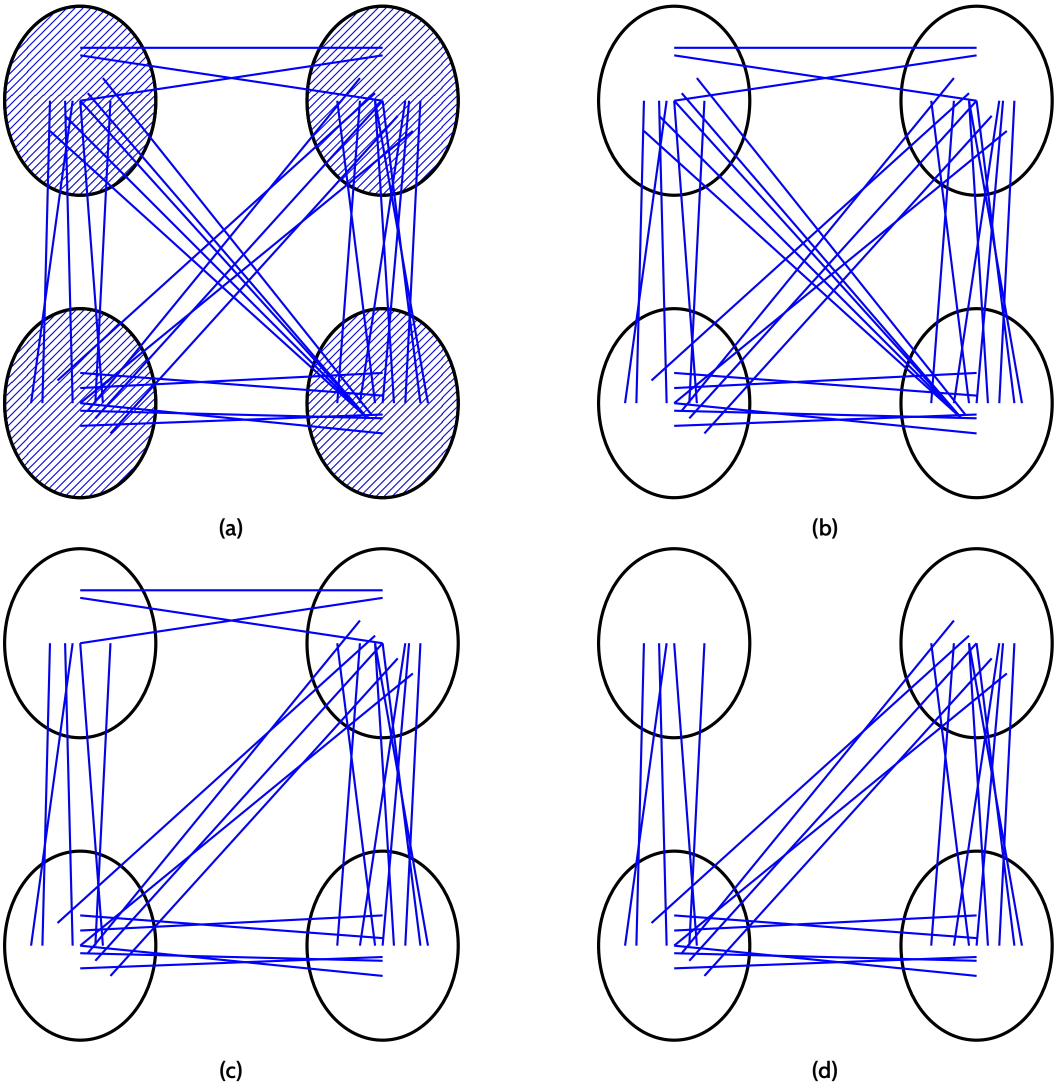

Figura 1: Representación del proceso de eliminación de aristas empleado en la demostración del teorema 8. La partición representada tiene cuatro conjuntos por simplicidad, pero el número de conjuntos dado por el lema 7 es mucho mayor. La figura 1a representa el grafo original G. La figura 1b representa el grafo G_1 obtenido tras eliminar las aristas dentro de cada conjunto de la partición. El grafo G_2, obtenido eliminando las aristas de pares no regulares, se representa en la figura 1c. La figura 1d representa el grafo G_3 final, obtenido tras eliminar también las aristas de parejas cuya densidad sea muy baja.

Alon [1] demostró que la complejidad de consultas de este problema tiene que ser superpolinómica, es decir, demostró que existen grafos que están ε-lejos de la ausencia de triángulos pero para los cuales no hay ningún algoritmo que trabaje con un número polinómico en $1/\varepsilon$ de consultas que pueda distinguirlos de grafos libres de triángulos. Este resultado fue completado más tarde por Alon y Shapira [5]. A pesar de la cota superpolinómica de Alon, sigue habiendo un salto muy grande entre las cotas inferior y superior conocidas. Cerrar este hueco es un problema abierto interesante.

El lema de regularidad de Szemerédi (lema 7) tiene muchas más aplicaciones en la teoría del *property testing* en el modelo de grafos densos. Se utilizó repetidamente para obtener resultados para algoritmos que comprueban diferentes propiedades en grafos, especialmente en la caracterización de propiedades de primer orden [2]. Por ejemplo, se demostró que la k-coloreabilidad es una propiedad que se puede comprobar con un número de consultas independiente del tamaño del grafo, a pesar de que este problema sea NP-completo en su versión exacta. Finalmente, Alon, Fischer, Newman y Shapira [3] demostraron un resultado muy general, estableciendo que una propiedad de grafos se puede comprobar en tiempo independiente del tamaño de la entrada en el modelo de grafos densos si y solo si comprobar dicha propiedad se puede reducir a comprobar si el grafo satisface una determinada partición de Szemerédi.

4.3. Ausencia de triángulos en el modelo de grafos generales: cotas

El problema de la ausencia de triángulos en el modelo de grafos generales fue estudiado por Alon, Kaufman, Krivelevich y Ron [4]. Lo que demostraron en este contexto es que no se puede resolver el problema con un número de consultas independiente de n. Vamos a ver aquí un resultado parcial.

Para poder trabajar con una distancia que depende del número de aristas del grafo, debemos tener la capacidad de trabajar con cada número posible de aristas o, al menos, con cada crecimiento asintótico de este número. Así, queremos encontrar cotas sobre la complejidad de los algoritmos que comprueban si el grafo de entrada está libre de triángulos o está ε-lejos de la ausencia de triángulos para cada crecimiento asintótico del número de aristas. Para hacer esto, vamos a trabajar con el grado medio del grafo y a establecer cotas para cada posible valor. El grado medio de un grafo $G = (V, E)$ con $|V| = n$ se define como $d = 2|E|/n$. Si consideramos valores de d como funciones de n, cada posible valor de d da lugar también a un único crecimiento asintótico de $|E|$ al tender n a infinito. Nótese que d es como máximo lineal en n, en cuyo caso estamos trabajando con grafos densos.

Teorema 10. *Cualquier algoritmo que compruebe la ausencia de triángulos en un grafo en el modelo de grafos generales debe realizar al menos $\Omega(\sqrt{n/d})$ consultas.*

Demostración. Para comprobar esto, basta con demostrar que existen dos familias de grafos $\mathcal{G}_1$ y $\mathcal{G}_2$ tales que todos los grafos de ambas tienen el mismo grado medio d y el mismo número de vértices, todos los grafos de $\mathcal{G}_1$ están libres de triángulos, todos los grafos de $\mathcal{G}_2$ están ε-lejos de la ausencia de triángulos, y cualquier algoritmo que realice $o(\sqrt{n/d})$ consultas sobre un grafo de cualquiera de las dos familias no será capaz de distinguir a cuál de las dos pertenece con probabilidad suficientemente alta.

Las dos familias que definimos son las siguientes. Cada familia se define con un único grafo sobre n vértices, y considerando todos los posibles etiquetados de los vértices. El grafo que define $\mathcal{G}_1$ consiste en un grafo bipartito completo sobre dos conjuntos de tamaño $\sqrt{nd/2}$, y el resto de vértices están aislados (es decir, no tienen ninguna arista incidente a ellos). El grafo que define $\mathcal{G}_2$ consiste en un grafo completo sobre $\sqrt{nd}$ vértices, y el resto de vértices están de nuevo aislados. Es muy fácil comprobar que el grado medio de cualquier grafo en $\mathcal{G}_1$ es exactamente d, mientras que el grado medio en $\mathcal{G}_2$ es $d(1 - \Theta(\sqrt{1/nd}))$, que asintóticamente crece igual que d, de modo que estas dos familias tienen un número de aristas comparable conforme n tiende a infinito. Además, es evidente que cualquier grafo en $\mathcal{G}_1$ está libre de triángulos, ya que es bipartito.

Hay que comprobar ahora que cualquier grafo en $\mathcal{G}_2$ está ε-lejos de la ausencia de triángulos, para algún $\varepsilon > 0$. Pero esto es una consecuencia del teorema de Mantel [19], que establece que el máximo número de aristas en un grafo libre de triángulos sobre k vértices es $\lfloor k^2/4 \rfloor$. Así, en nuestro caso, tenemos que eliminar casi $nd/4$ aristas, lo que supone que cualquier grafo en $\mathcal{G}_2$ está ε-lejos de la ausencia de triángulos para cualquier $\varepsilon < 1/4$.

Ahora, supongamos que un algoritmo trata de encontrar un triángulo en un grafo seleccionado uniformemente entre las dos familias con solo $o(\sqrt{n/d})$ consultas. Para ser capaz de distinguirlos necesitará encontrar uno de los vértices que tienen grado positivo, ya que en caso contrario solo habrá encontrado vértices aislados. Pero la probabilidad de encontrar uno cualquiera de estos vértices es $\Theta(\sqrt{d/n})$, de modo que al realizar todas las consultas la probabilidad de encontrar uno, por la desigualdad de Boole, es $o(\sqrt{n/d}\sqrt{d/n}) = o(1)$, es decir, tiende a cero conforme n tiende a infinito. Así pues, se deben realizar al menos $\Omega(\sqrt{n/d})$ consultas. ∎

La cota que da el teorema 10 es mejor para grafos poco densos, y va empeorando conforme d aumenta. En particular, si d es constante se observa una gran diferencia entre este resultado y el del teorema 4: en este caso tenemos una cota de $\Omega(\sqrt{n})$ sobre el número de consultas, mientras que en el modelo de grado acotado se podía resolver el problema con un número de consultas independiente de n.

Conforme d crece, la cota del teorema 10 es más débil. Sin embargo, Alon, Kaufman, Krivelevich y Ron [4] dan otras cotas para distintos valores de d que mejoran esta, utilizando técnicas más avanzadas que las presentadas aquí. En general, consiguen demostrar que para cualquier valor de $d = \mathcal{O}(n^{1-\nu(n)})$, con $\nu(n) = \frac{\log\log\log n + 4}{\log\log n} = o(1)$, cualquier algoritmo que compruebe la ausencia de triángulos deberá realizar $\Omega(n^{1/3})$ consultas.

Por otra parte, en el mismo artículo también se dan cotas superiores a la complejidad, es decir, se demuestra que existen algoritmos que resuelven el problema ejecutando un determinado número de consultas. La cota general que se obtiene es de $\mathcal{O}(n^{6/7}\operatorname{poly}\log(n))$. Así, aunque en el modelo de grafos generales no se pueden conseguir algoritmos que resuelvan el problema con un número de consultas independiente de n, sí se puede conseguir que lo hagan en tiempo sublineal, lo cual ya es una mejora considerable con respecto a los algoritmos exactos. Cerrar el espacio entre la cota superior y la inferior para determinar la complejidad real del problema es aún un problema abierto.

5. Conclusiones y otros problemas

Como se ha podido ver a lo largo del texto, el *property testing* permite desarrollar algoritmos para comprobar propiedades de estructuras de datos cuya complejidad es inferior a la de algoritmos exactos a cambio de sacrificar exactitud en la respuesta. Este sacrificio tiene una componente probabilística (la probabilidad de que el algoritmo dé una respuesta errónea no es nula, en la mayoría de casos) y determinista (la respuesta no se da sobre la propiedad, sino con respecto a una medida de distancia respecto a la propiedad). El desarrollo de estos algoritmos, además de un interés puramente teórico, también tiene muchas aplicaciones para resolver problemas en situaciones en las que los algoritmos tradicionales son demasiado lentos.

Hemos podido ver tres ejemplos de cómo tratar algoritmos en diferentes modelos para una determinada propiedad, y cómo el comportamiento de los algoritmos en cada modelo puede ser muy diferente. En general, las propiedades que se pueden estudiar son muy variadas. En el modelo de grafos densos existen resultados muy generales que engloban muchas de ellas, como ocurre con el caso del resultado de Alon, Fischer, Newman y Shapira [3]. También se puede considerar *property testing* aplicado a hipergrafos, y en este contexto destaca el resultado de Joos, Kim, Kühn y Osthus [16], que generaliza la clasificación de propiedades que se pueden estudiar con un número constante de consultas. En el caso del modelo de grado acotado, uno de los resultados más destacables se debe a Benjamini, Schramm y Shapira [6], quienes determinaron que toda propiedad cerrada por menores se puede estudiar con un número constante de consultas. Destacan también los resultados de Newman y Sohler [21]. El modelo de grafos generales es, sin duda, uno de los menos estudiados, y al trabajo de Alon, Kaufman, Krivelevich y Ron [4] se le pueden añadir resultados como el de Kaufman, Krivelevich y Ron [18], que estudiaron la propiedad de ser bipartito, o el de Espuny Díaz, Joos, Kühn y Osthus [8], que extendieron algunos resultados de Alon, Kaufman, Krivelevich y Ron [4] de la ausencia de triángulos a la ausencia de cualquier subgrafo fijo y también al caso de hipergrafos. En general, esta área aún tiene muchos problemas abiertos. Muchos de los problemas abiertos tienen que ver con determinar con exactitud la complejidad de los problemas decisionales en *property testing*.

Además de todo lo que se ha discutido hasta ahora, el *property testing* también se emplea para estudiar propiedades que no tienen nada que ver con grafos (se pueden ver muchos ejemplos en el estudio de

Ron [24]). Además de esto, existen algunas extensiones naturales del *property testing*. Una de ellas tiene que ver con la distribución subyacente a la hora de definir la distancia; en particular, si esta distribución es desconocida se habla de *distribution-free testing* [11, 12, 14, 15]. Otra de las extensiones es la que se conoce como *tolerant testing*, en la que el algoritmo recibe dos parámetros de distancia, ε_1 y ε_2, y debe distinguir entre el caso de que la entrada de datos esté ε_1-cerca de la propiedad o ε_2-lejos [9, 23]. Relacionado con lo anterior, también se pueden estudiar algoritmos que traten de aproximar la distancia de la entrada a una cierta propiedad [10, 20, 23]. En general, estudiar la relación entre los algoritmos de *property testing* y los algoritmos de aproximación clásicos (se sabe que son dos tipos de problemas completamente distintos, como se menciona en el estudio de Ron [24]) también puede ser un problema interesante.

Referencias

[1] Alon, Noga. «Testing subgraphs in large graphs». En: *Random Structures & Algorithms* 21.3-4 (2002). Random structures and algorithms (Poznan, 2001), págs. 359-370. issn: 1042-9832. https://doi.org/10.1002/rsa.10056.

[2] Alon, Noga; Fischer, Eldar; Krivelevich, Michael, y Szegedy, Mario. «Efficient testing of large graphs». En: *Combinatorica. An International Journal on Combinatorics and the Theory of Computing* 20.4 (2000), págs. 451-476. issn: 0209-9683. https://doi.org/10.1007/s004930070001.

[3] Alon, Noga; Fischer, Eldar; Newman, Ilan, y Shapira, Asaf. «A combinatorial characterization of the testable graph properties: it's all about regularity». En: *SIAM Journal on Computing* 39.1 (2009), págs. 143-167. issn: 0097-5397. https://doi.org/10.1137/060667177.

[4] Alon, Noga; Kaufman, Tali; Krivelevich, Michael, y Ron, Dana. «Testing triangle-freeness in general graphs». En: *SIAM Journal on Discrete Mathematics* 22.2 (2008), págs. 786-819. issn: 0895-4801. https://doi.org/10.1137/07067917X.

[5] Alon, Noga y Shapira, Asaf. «Testing subgraphs in directed graphs». En: *Journal of Computer and System Sciences* 69.3 (2004), págs. 353-382. issn: 0022-0000. https://doi.org/10.1016/j.jcss.2004.04.008.

[6] Benjamini, Itai; Schramm, Oded, y Shapira, Asaf. «Every minor-closed property of sparse graphs is testable». En: *Advances in Mathematics* 223.6 (2010), págs. 2200-2218. https://doi.org/10.1016/j.aim.2009.10.018.

[7] Blum, Manuel; Luby, Michael, y Rubinfeld, Ronitt. «Self-testing/correcting with applications to numerical problems». En: *Proceedings of the 22nd Annual ACM Symposium on Theory of Computing (Baltimore, MD, 1990)*. Vol. 47. 3. 1993, págs. 549-595. https://doi.org/10.1016/0022-0000(93)90044-W.

[8] Espuny Díaz, Alberto; Joos, Felix; Kühn, Daniela, y Osthus, Deryk. «Edge correlations in random regular hypergraphs and applications to subgraph testing». En: *ArXiv e-prints* (mar. de 2018). arXiv: 1803.09223 [math.CO].

[9] Fischer, Eldar y Fortnow, Lance. «Tolerant versus intolerant testing for Boolean properties». En: *Theory of Computing. An Open Access Journal* 2 (2006), págs. 173-183. issn: 1557-2862. https://doi.org/10.4086/toc.2006.v002a009.

[10] Fischer, Eldar y Newman, Ilan. «Testing versus estimation of graph properties». En: *SIAM Journal on Computing* 37.2 (2007), págs. 482-501. issn: 0097-5397. https://doi.org/10.1137/060652324.

[11] Glasner, Dana y Servedio, Rocco A. «Distribution-free testing lower bounds for basic Boolean functions». En: *Theory of Computing. An Open Access Journal* 5 (2009), págs. 191-218. issn: 1557-2862. https://doi.org/10.4086/toc.2009.v005a010.

[12] Goldreich, Oded; Goldwasser, Shafi, y Ron, Dana. «Property testing and its connection to learning and approximation». En: *Journal of the ACM* 45.4 (1998), págs. 653-750. issn: 0004-5411. https://doi.org/10.1145/285055.285060.

[13] Goldreich, Oded y Ron, Dana. «Property testing in bounded degree graphs». En: *Algorithmica. An International Journal in Computer Science* 32.2 (2002), págs. 302-343. issn: 0178-4617. https://doi.org/10.1007/s00453-001-0078-7.

[14] Halevy, Shirley y Kushilevitz, Eyal. «Distribution-free property-testing». En: *SIAM Journal on Computing* 37.4 (2007), págs. 1107-1138. issn: 0097-5397. https://doi.org/10.1137/050645804.

[15] Halevy, Shirley y Kushilevitz, Eyal. «Distribution-free connectivity testing for sparse graphs». En: *Algorithmica. An International Journal in Computer Science* 51.1 (2008), págs. 24-48. issn: 0178-4617. https://doi.org/10.1007/s00453-007-9054-1.

[16] Joos, Felix; Kim, Jaehoon; Kühn, Daniela, y Osthus, Deryk. «A characterization of testable hypergraph properties». En: *ArXiv e-prints* (jul. de 2017). arXiv: 1707.03303 [math.CO].

[17] Karp, Richard M. «Reducibility among combinatorial problems». En: *Complexity of computer computations (Proc. Sympos., IBM Thomas J. Watson Res. Center, Yorktown Heights, N.Y., 1972)*. Springer US, 1972, págs. 85-103. https://doi.org/10.1007/978-1-4684-2001-2_9.

[18] Kaufman, Tali; Krivelevich, Michael, y Ron, Dana. «Tight bounds for testing bipartiteness in general graphs». En: *SIAM Journal on Computing* 33.6 (2004), págs. 1441-1483. issn: 0097-5397. https://doi.org/10.1137/S0097539703436424.

[19] Mantel, W. «Problem 28 (solution by H. Gouwentak, W. Mantel, J. Teixeira de Mattes, F. Schuh and W.A. Wythoff». En: *Wiskundige Opgaven* 10 (1907), págs. 60-61.

[20] Marko, Sharon y Ron, Dana. «Distance approximation in bounded-degree and general sparse graphs». En: *Approximation, randomization and combinatorial optimization.* Vol. 4110. Lecture Notes in Comput. Sci. Springer, Berlin, 2006, págs. 475-486. https://doi.org/10.1007/11830924_43.

[21] Newman, Ilan y Sohler, Christian. «Every property of hyperfinite graphs is testable». En: *SIAM Journal on Computing* 42 (2013), págs. 1095-1112. issn: 0097-5397. https://doi.org/10.1137/120890946.

[22] Parnas, Michal y Ron, Dana. «Testing the diameter of graphs». En: *Random Structures & Algorithms* 20.2 (2002), págs. 165-183. issn: 1042-9832. https://doi.org/10.1002/rsa.10013.abs.

[23] Parnas, Michal; Ron, Dana, y Rubinfeld, Ronitt. «Tolerant property testing and distance approximation». En: *Journal of Computer and System Sciences* 72.6 (2006), págs. 1012-1042. issn: 0022-0000. https://doi.org/10.1016/j.jcss.2006.03.002.

[24] Ron, Dana. «Algorithmic and analysis techniques in property testing». En: *Foundations and Trends® in Theoretical Computer Science* 5.2 (2009), front matter, 73-205. issn: 1551-305X. https://doi.org/10.1561/0400000029.

[25] Rubinfeld, Ronitt y Sudan, Madhu. «Robust characterizations of polynomials with applications to program testing». En: *SIAM Journal on Computing* 25.2 (1996), págs. 252-271. issn: 0097-5397. https://doi.org/10.1137/S0097539793255151.

[26] Szemerédi, Endre. «Regular partitions of graphs». En: *Problèmes combinatoires et théorie des graphes (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976).* Vol. 260. Colloq. Internat. CNRS. CNRS, Paris, 1978, págs. 399-401.

TEMat, volumen 3. Mayo de 2019.
e-ISSN: 2530-9633

Publicado con la colaboración de la
Real Sociedad Matemática Española