Miscelánea Matemática **34** (2001) 25–42 SMM

# La conjetura de Fermat

Felipe Zaldívar
Departamento de Matemáticas
Universidad Autónoma Metropolitana-I
09340 México, D.F.
México
fzc@xanum.uam.mx

La historia de la *conjetura de Fermat*, a veces también conocida como el *último teorema de Fermat*, ha sido contada tantas veces que es tentador omitir esta parte. Sin embargo lo irresistible de esta historia es la suposición de que Pierre de Fermat (1601-1665) tuvo alguna vez una demostración del teorema que luego se perdió con el transcurso del tiempo porque nunca la publicó. Esta historia, muy posiblemente apócrifa, ha contribuido sin duda a la popularidad de esta conjetura, aunada también a la aparente simplicidad de su formulación. Lo cierto de esta historia es que comienza con la nota que escribió Fermat en un margen de su ejemplar de la *Arithmetica* de Diofanto. Este ejemplar con las anotaciones de Fermat se ha perdido y sólo conocemos de su existencia por la publicación de una edición de la *Arithmetica* por Samuel de Fermat, hijo de Pierre de Fermat, y en esta edición el hijo de Fermat transcribió la nota de su padre bajo los textos griego y latino de la pregunta 8 del libro 2 de Diofanto. Esta nota dice textualmente:

Observatio Domini Petri de Fermat.

*Cubum autem in duos cubos, aut quadrato quadratum in duos quadrato quadratos, et generaliter nullam in infinitum ultra quadratum potestatem in duos eiusdem nominis fas est dividere: cuius rei demostrationem mirabilem sane detexi. Hanc marginis exiguitas non caparet.*

O, en nuestro rudo español:

Observación del Señor Pierre de Fermat.

*Es imposible separar un cubo en dos cubos, o una cuarta potencia en dos cuartas potencias o, en general, cualquier potencia mayor que la*

*segunda en dos potencias similares. He descubierto una demostración verdaderamente maravillosa de esto, pero este margen es demasiado pequeño y no cabe.*

En notación matemática esta conjetura es: *si* $n \geq 3$ *es un entero, entonces la ecuación* $x^n + y^n = z^n$ *no tiene soluciones enteras con* $x, y, z \neq 0$.

**Introducción.** En este artículo, bosquejaremos para un público interesado, algunas ideas involucradas en la demostración reciente de la conjetura de Fermat. Comenzamos con algunas consideraciones históricas relevantes para después introducir los conceptos necesarios de curvas en general, enfatizando el caso de curvas elípticas por el papel relevante que juegan en la demostración de la conjetura de Fermat, después recordamos algunas generalidades sobre problemas diofantinos de nuevo concentrándonos en el caso de curvas elípticas, en particular explicando los conceptos relacionados con la reducción módulo un primo de curvas elípticas y el de curvas elípticas semiestables; también se recuerdan las ideas relativas a la acción de Galois sobre una curva y sobre sus puntos de torsión. Finalmente se introducen algunas ideas sobre formas modulares para culminar con la conjetura de Taniyama-Shimura-Weil. Al final se discuten algunas ideas sobre la demostración de la conjetura de Fermat usando todos los conceptos introducidos previamente.

**Un poco de historia.** La nota marginal trascrita arriba estaba junto a la pregunta VIII del libro 2 de la Arithmetica de Diofanto. Esta pregunta trata del caso de las ternas pitagóricas: la ecuación $x^2+y^2 = z^2$ es la fórmula de Pitágoras que relaciona los catetos con la hipotenusa de un triángulo rectángulo. Siglos antes de Diofanto, desde la época de Euclides, ya se sabía de la existencia de una infinidad de enteros no nulos que satisfacen la ecuación pitagórica. Es obvio también que para $n = 1$, la ecuación $x + y = z$ tiene una infinidad de soluciones enteras: la suma de dos enteros $x, y$ es otro entero $z$.

El salto cualitativo de $n = 1$ y 2 al caso $n \geq 3$, donde la conjetura de Fermat afirma que *no* se tienen soluciones enteras con los $x, y, z$ tales que $xyz \neq 0$, es bastante sorprendente.

Fermat mismo pudo probar su conjetura para el caso $n = 4$ (no en un margen). De hecho, Fermat probó una proposición un poco más general: la ecuación $x^4 + y^4 = z^2$ no tiene soluciones no triviales, y de aquí se deduce fácilmente que $x^4 + y^4 = z^4$ no tiene soluciones no triviales. El método introducido por Fermat en su demostración de la no existencia de soluciones no triviales de $x^4 + y^4 = z^2$, llamado *descenso*

*infinito*, es como sigue: supone primero que hay una solución no trivial, la cual (como los exponentes son pares) podemos suponer positiva, y luego por medio de una serie de operaciones aritméticas genera una solución entera positiva *menor*; procediendo de esta forma genera una sucesión infinita de soluciones enteras no nulas con valor absoluto cada vez más pequeño. Claramente esto no es posible, así que la suposición inicial sobre la existencia de una solución no trivial debe ser falsa.

Después de esto, tendremos que esperar hasta el siguiente siglo (XVIII) cuando Euler demuestra la conjetura de Fermat para el exponente $n = 3$, usando también descenso infinito, pero ahora con un argumento un poco más delicado. Es interesante observar que Euler también reconstruyó una demostración del caso $n = 4$ de la conjetura de Fermat, y por esa misma época, Euler estaba estudiando un problema de cálculo relacionado con la longitud de arco de la lemniscata. La integral involucrada tiene la forma

$$\int \frac{dx}{\sqrt{1-x^4}}\,,$$

la cual no es integrable por medio de funciones elementales y, como Jacobi observa después, uno puede suponer que ambos problemas (la no existencia de soluciones no triviales de la ecuación diofantina $x^4 + y^4 = z^4$ y la no integrabilidad por medio de funciones elementales de la integral elítica anterior) de alguna manera estaban relacionados en la mente de Euler y "esta coincidencia digna de notarse difícilmente se le haya escapado a" Euler. André Weil observa en [5] que, en los casos anteriores, Fermat y Euler sólo habían tratado con lo que ahora llamaríamos curvas algebraicas de género $\leq 1$, y el hecho de que no encontraron demostraciones para exponentes mayores que 4 tendría que ver con el salto cualitativo de pasar a curvas de género $> 1$. Por supuesto que el concepto de género (y de hecho, el de curva algebraica) no existía en tiempos de Fermat y Euler; tendremos que esperar a los grandes matemáticos de fines del siglo XIX para descubrir estos conceptos.

Regresando a la historia temprana de la conjetura de Fermat, al principio del siglo XIX el matemático francés Adrien-Marie Legendre y el matemático alemán P.G. Lejeune Dirichlet hallaron demostraciones de la conjetura para el exponente $n = 5$. Dirichlet también intentó el caso $n = 7$ pero sólo pudo probarla para $n = 14$. El caso $n = 7$ fue resuelto después por el matemático francés Gabriel Lamé.

Después de estos intentos iniciales con exponentes pequeños, el primer intento hacia una demostración de la conjetura de Fermat en toda su generalidad fue hecho por el matemático alemán Ernest E. Kummer.

El trabajo de Kummer, de importancia capital para el desarrollo de la Aritmética y Álgebra actuales, se centró en el estudio de la factorización de la ecuación

$$z^p = x^p + y^p = (x+y)(x+\omega y)\cdots(x+\omega^{p-1}y),$$

donde $\omega$ es una raíz primitiva $p$-ésima de la unidad. En lenguaje actual, lo que Kummer consideró fue el anillo formado por los números complejos de la forma:

$$a_0 + a_1\omega + a_2\omega^2 + \cdots + a_n\omega^n$$

con los $a_j \in \mathbb{Z}$. Uno de los resultados más importantes obtenido por Kummer es el hecho de que si este anillo fuera un *dominio de factorización única*, es decir, si como en el anillo de los enteros $\mathbb{Z}$, los elementos de este anillo se pudieran factorizar en producto de primos de manera esencialmente única, entonces, después de clasificar las unidades de este anillo, Kummer prueba que la conjetura de Fermat es cierta para el exponente primo $p$. Kummer supo, casi de inmediato, que en general los anillos anteriores no son de factorización única. A los primos $p$ para los cuales los anillos anteriores son de factorización única se les llama *primos regulares* y así, Kummer probó que la conjetura de Fermat es cierta para estos primos. Kummer obtuvo varias caracterizaciones de estos primos pudiendo listar los primos regulares menores de 100. Después de Kummer se han obtenido varios criterios para decidir si un primo es o no regular, destacando los trabajos del matemático norteamericano Vandiver a principios de este siglo. Algunos de estos criterios se pueden algoritmizar y con el uso de computadoras se han verificado exponentes primos hasta órdenes de alrededor de 4 millones. Resulta un poco desanimante que después se haya podido probar que existen infinitos primos irregulares y no se sabe todavía si hay infinitos primos regulares.

Un criterio más teórico para la regularidad de un primo fue encontrado por el matemático francés Jacques Herbrand en 1932 y es interesante notar que en 1976 Ken Ribet pudo demostrar el recíproco del criterio de Herbrand. Ken Ribet aparecerá después en conexión con la demostración reciente de la conjetura de Fermat.

Antes de continuar con la historia reciente de la conjetura de Fermat, recordemos algunos conceptos aritmético-geométricos pertinentes:

**Curvas.** Dado un polinomio $f(x,y)$ en dos variables (digamos, con coeficientes racionales), este polinomio define una curva en el plano $XY$,

a saber, el lugar geométrico de todos los puntos $(x, y)$ que satisfacen la ecuación $f(x, y) = 0$. La curva se llama *lisa* o *no singular* si en cada punto de la curva se tiene una tangente bien definida. Por razones más profundas, la curva anterior debe compactarse añadiendo a los puntos $(x, y)$ que satisfacen la ecuación $f(x, y) = 0$, "puntos al infinito" y así, la curva como objeto geométrico vivirá ahora en el plano proyectivo. Tiene entonces sentido hablar del *género* de la curva compacta: si la curva definida por el polinomio $f(x, y)$ es lisa y $n$ es el grado del polinomio, entonces el género $g$ de esta curva es $g = (1/2)(n-1)(n-2)$. Así, las *rectas* (dadas por polinomios de grado $n = 1$) tienen género 0; similarmente, las *cónicas* (dadas por polinomios cuadráticos) tienen género 0.

Desde un punto de vista topológico, el género de una curva se visualiza como sigue: si consideramos los *puntos complejos* de la curva $C$, denotados $C(\mathbb{C})$, i.e., los puntos con coordenadas complejas que satisfacen la ecuación $f(x, y) = 0$ que define a $C$, entonces $C(\mathbb{C})$ es una *superficie de Riemann compacta* con $g$ agujeros, y se ve como en la figura 1:

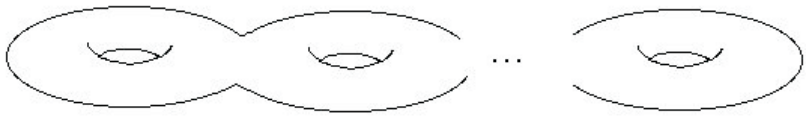


Figura 1

El caso de género $g = 0$ corresponde a una esfera sin hoyos.

Notemos ahora que polinomios $f(x, y)$ de grado 3, si dan lugar a curvas lisas, entonces estas curvas son de género 1, y sus puntos complejos forman un *toro*:

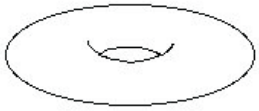
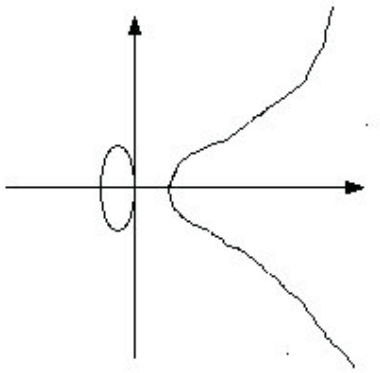


Figura 2

Estas curvas de género 1 se llaman *curvas elípticas*, y vale la pena observar que como su nombre *no* lo indica, las curvas elípticas no son

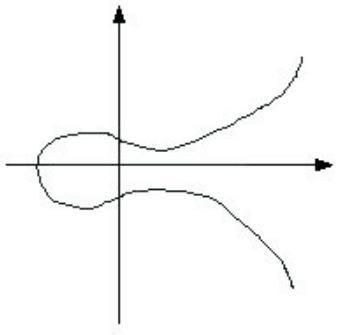
elipses (las elipses son cónicas, y como vimos antes, estas tienen género 0), aún cuando hay una relación entre curvas elípticas y funciones que aparecen al calcular la longitud de arco de una elipse.

Nótese que aún cuando el polinomio cúbico $f(x, y)$ que define una curva elíptica tenga coeficientes racionales, los puntos $(x, y)$ que lo satisfacen son números complejos en general y así sólo podemos graficar sus partes reales. Las figuras 3 y 4 muestran las gráficas de las partes reales de unas curvas elípticas típicas:



$y^2 = x^3 - x$

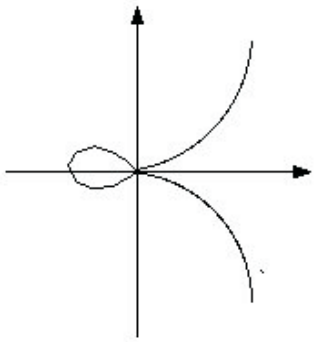
FIGURA 3



$y^2 = x^3 - 3x + 3$

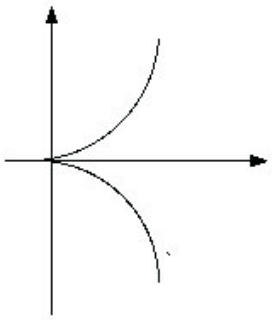
FIGURA 4

y las figuras 5 y 6 muestran las gráficas de las partes reales de curvas cúbicas que *no* son elípticas, i.e., no son lisas:



$y^2 = x^3 + x^2$

Figura 5



$y^2 = x^3$

Figura 6

(En el primer caso, la curva tiene dos tangentes distintas en el punto $(0, 0)$, y en el segundo caso tiene una tangente doble en el punto $(0, 0)$).

Este es el lugar adecuado para observar que una curva cúbica $f(x, y) = 0$ con coeficientes racionales se puede escribir de la forma

$$y^2 = F(x)$$

con $F(x)$ un polinomio cúbico en $x$. Desde un punto de vista algebraico la curva definida por $y^2 = F(x)$ es lisa si y sólo si $F(x)$ tiene sus tres raíces distintas. Nótese que en los ejemplos dados en las figuras 5 y 6

se tienen, en el primer caso dos raíces iguales, y en el segundo caso las tres raíces iguales.

**Problemas diofantinos.** Regresando a la aritmética, los problemas que interesan son los siguientes: dada una ecuación polinomial $f(x, y) = 0$ (con coeficientes enteros o racionales) queremos saber: (i) ¿Tiene soluciones enteras, i.e., con $x, y \in \mathbb{Z}$? ¿o racionales, i.e., con $x, y \in \mathbb{Q}$?, (ii) si es así, ¿cuántas?, (iii) ¿hay un algoritmo para generar o detectar estas soluciones?. Todos estos problemas los englobamos bajo el nombre de *problemas diofantinos*. Así, el problema aritmético se enlaza con la geometría de la curva definida por el polinomio $f(x, y)$.

Resumiendo lo que se conoce hasta ahora tenemos: para el caso de género $g = 0$ (i.e., para curvas $C$ dadas por polinomios de grado 1 ó 2 con coeficientes en $\mathbb{Q}$) es fácil probar que si la curva $C$ tiene un punto racional, i.e., con coordenadas en $\mathbb{Q}$, entonces $C$ tiene una infinidad de puntos racionales. Esto es obvio para una recta y para una cónica se tiene una biyección entre los puntos racionales de la cónica y los puntos racionales de una recta (con la excepción de uno).

Si ahora saltamos al caso de curvas lisas $C$ de género $g \geq 2$ definidas por polinomios con coeficientes racionales, y consideramos el conjunto de puntos racionales $C(\mathbb{Q})$ de $C$, en 1922 el matemático inglés L.J. Mordell había conjeturado que $C(\mathbb{Q})$ debería ser un conjunto finito. Esta intrigante conjetura de Mordell fue estudiada y relacionada con otros resultados y conjeturas en geometría diofantina por matemáticos de la escuela de Shafarevich en Rusia en los años sesenta, y de una forma sorprendente fue demostrada por el matemático alemán Gerd Faltings en 1983. Este resultado marcó un avance fundamental en la aritmética de curvas y tiene relación con la conjetura de Fermat si observamos que la ecuación de Fermat $x^n + y^n = z^n$ con $n \geq 5$ define la curva $x^n + y^n = 1$, lisa de género $g \geq 2$; y así, $x^n + y^n = 1$ sólo tiene un número finito de puntos racionales y por lo tanto la ecuación $x^n + y^n = z^n$ sólo tiene un número finito de soluciones enteras. Esto todavía no es la conjetura de Fermat ya que ésta afirma que sólo debe tener soluciones triviales.

Para el caso restante, el de curvas de género $g = 1$, si se tiene una curva lisa $E$ dada por un polinomio con coeficientes racionales y con un punto racional $0 \in E(\mathbb{Q})$, diremos que $E$ es una *curva elíptica* definida sobre $\mathbb{Q}$. Consideremos entonces el conjunto $E(\mathbb{Q})$ de puntos racionales de $E$. A diferencia de los casos $g = 0$ y $g \geq 2$ discutidos anteriormente, el conjunto $E(\mathbb{Q})$ puede ser *finito* o *infinito*, dependiendo de la curva elíptica en consideración. Sin embargo, hay una propiedad adicional de

$E$ (y de $E(\mathbb{Q})$) que no tienen las otras curvas de género $\neq 1$. Dicho rápidamente: el conjunto de puntos de $E$ tiene estructura de grupo conmutativo, es decir, existe una operación en $E$ que es asociativa, tiene neutro, inversos y es conmutativa. Geométricamente esta operación se define como sigue: dados dos puntos $P, Q$ en $E$ consideremos la recta (cuerda) que une $P$ con $Q$. Esta recta corta a $E$ en un tercer punto $R$ (esto, aunque intuitivamente claro ya que $E$ está definido por un polinomio de grado 3 es, de hecho, una consecuencia de un teorema básico de geometría, el teorema de Bezout, que afirma que dos curvas de grados $d_1$ y $d_2$ se intersectan, contando multiplicidades, en $d_1 d_2$ puntos; aquí la recta tiene grado 1 y $E$ es de grado 3). Después, consideramos el punto $R'$ de $E$ obtenido al intersectar $E$ con la recta "vertical" que pasa por $R$ y el punto al infinito (ver la figura siguiente). Se define entonces $P + Q := R'$. Nótese que en esta construcción el neutro aditivo es el punto racional $0 \in E$.

También, en el caso extremo cuando $P = Q$ es el mismo punto de $E$, entonces la cuerda por $P$ y $Q$ querrá decir la recta tangente por el punto dado; esta recta también corta a $E$ en un tercer punto $R$ y aplicamos la misma construcción anterior. Otro caso extremo a considerar es cuando $P$ y $Q$ son puntos de $E$ tales que la recta $\mathcal{L}$ por $P$ y $Q$ es vertical. Entonces, recordando que $E$ incluye un "punto al infinito", este punto al infinito es el tercer punto donde la recta $\mathcal{L}$ intersecta a $E$.

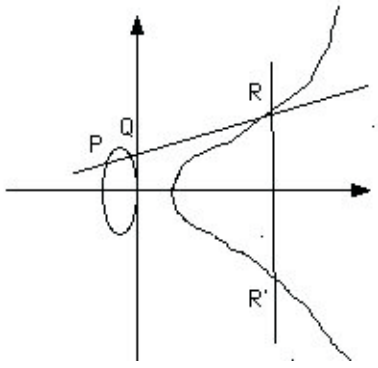



Figura 7

Considerando ahora el conjunto $E(\mathbb{Q})$ de puntos racionales de $E$, usando la construcción de la operación de grupo de $E$ se puede probar que si $P$ y $Q$ son puntos racionales de $E$, entonces $P + Q$ también es un punto racional y, consecuentemente, $E(\mathbb{Q})$ es un grupo abeliano también. De hecho, por la definición de la operación de grupo de $E$,

tenemos un método para generar a los puntos racionales de $E$: dados dos puntos racionales de $E$ trazando la cuerda (o tangente, si los puntos son iguales) que pasa por ellos obtenemos otro punto racional. Más aún, un teorema de Mordell (generalizado por A. Weil) nos dice que el grupo de puntos racionales $E(\mathbb{Q})$ es un *grupo abeliano finitamente generado*, es decir, existe un conjunto finito de puntos racionales de $E$ a partir de los cuales trazando cuerdas y tangentes se generan todos los puntos racionales de $E$. Recordemos en este punto que, como todo grupo abeliano finitamente generado, $E(\mathbb{Q})$ se puede separar en un *subgrupo libre* (de rango finito) y un *subgrupo finito* (llamado también de *torsión*). Puede suceder que el rango del subgrupo libre sea cero y en este caso $E(\mathbb{Q})$ resulta un grupo finito. Sin embargo, hay casos en los cuales el rango es mayor que cero y entonces $E(\mathbb{Q})$ es infinito.

Recolectando lo que tenemos hasta ahora: las curvas elípticas sobre $\mathbb{Q}$ ocupan un lugar intermedio entre las curvas de género 0 y las curvas de género $\geq 2$. Sin embargo, en cierto sentido son más ricas: son grupos abelianos y sus conjuntos de puntos racionales son grupos abelianos también. Un problema típico a considerar es calcular el grupo $E(\mathbb{Q})$, o su rango, o su parte de torsión. Hay una intensa actividad matemática alrededor de estos problemas y es hasta recientemente que se han tenido ciertos progresos hacia la solución de algunos de estos problemas. Ya que estamos hablando de esto, recordemos uno de los enfoques usados en este contexto: como nuestra curva $E$ está definida sobre los racionales, podemos eliminar denominadores del polinomio que la define y pensar que la curva está dada por un polinomio cúbico $f(x, y)$ con coeficientes enteros. Lo que se ocurre naturalmente es reducir estos coeficientes módulo un entero $m$ y considerar ahora al polinomio con coeficientes en el anillo de enteros módulo $m$. Denotemos este polinomio por $\widetilde{f}(x, y)$. Por ejemplo, si

$$f(x, y) = y^2 - x^3 - 7x^2 + 144x\,,$$

y si $m = 5$, entonces

$$\widetilde{f}(x, y) = y^2 - x^3 + 3x^2 + 4x\,,$$

y si $m = 2$, entonces

$$\widetilde{f}(x, y) = y^2 - x^3 + x^2\,.$$

Es natural preguntarse si los nuevos polinomios $\widetilde{f}(x, y)$ definen curvas lisas o no. Si recordamos ahora que un polinomio cúbico de la forma $y^2 = f(x)$ define una curva lisa si y sólo si las tres raíces de $f(x)$ son

distintas, y como el discriminante de un polinomio está dado por el producto de las diferencias de sus raíces, este discriminante decide si las raíces son o no distintas. Así, para que $y^2 = f(x)$ defina una curva elíptica se requiere que el discriminante $\Delta$ de $f(x)$ sea distinto de cero (y que la curva tenga un punto racional). En el ejemplo de arriba

$$y^2 = x^3 + 7x^2 - 144x = x(x-9)(x+16),$$

sus tres raíces son distintas y así $\Delta \neq 0$ y por lo tanto sí define una curva elíptica (sobre $\mathbb{Q}$).

Ahora, si pedimos que $f(x)$ tenga coeficientes enteros, su discriminante $\Delta$ resulta un entero también y al considerar la reducción módulo un entero $m$ del polinomio $y^2 = f(x)$ que define a la curva elíptica $E$, el discriminante $\widetilde{\Delta}$ del polinomio reducido $\widetilde{f}(x)$ es precisamente la reducción módulo $m$ del discriminante $\Delta$ de $f(x)$. Si ahora nos preguntamos si los polinomios $y^2 = \widetilde{f}(x)$ definen o no curvas elípticas, es necesario ser un poco más precisos y preguntarnos dónde consideraremos estas curvas. Para esto recordemos que los enteros módulo $m$ forman un *campo* (objeto algebraico con operaciones análogas a las de $\mathbb{Q}$) si y sólo si $m$ es un *primo*; y así, en lugar de considerar reducciones módulo cualquier entero nos restringiremos a considerar reducciones módulo un entero primo $p$ de tal forma que el polinomio reducido $\widetilde{f}$ tendrá ahora coeficientes en el campo $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ de enteros módulo $p$.

Ahora, dada una curva elíptica $E$ sobre $\mathbb{Q}$ a la cual podemos considerar, sin perder generalidad, que está definida por un polinomio con coeficientes enteros, de los polinomios que definen a $E$, existe uno con coeficientes enteros tal que su discriminante (el cual es un entero positivo) es *mínimo*. Este *discriminante mínimo* es un invariante más sutil que el discriminante de cualquier otro polinomio que defina a $E$. Diremos que este polinomio con discriminate mínimo define un *modelo mínimo* de $E$.

Si ahora reducimos la curva $E$ módulo $p$ y si $p \nmid \Delta$ (el discriminante de $E$), entonces la curva reducida $\widetilde{E}$ tiene discriminante $\widetilde{\Delta} \neq 0 \bmod p$, y por lo tanto es una curva elíptica definida sobre el campo finito $\mathbb{F}_p$. Aquí es donde se usa la existencia de un modelo mínimo de la curva elíptica $E$: de todos los polinomios con coeficientes enteros que definen a $E$, algunos al reducirlos mod $p$ pueden tener discriminantes que sean $0 \bmod p$ y otros no; de todos esos modelos escogemos, para no tener ambigüedad, al modelo mínimo. Si al reducir mod $p$ el modelo mínimo de la curva $E$, la curva reducida $\widetilde{E}$ resulta una curva elíptica, diremos que $E$ *tiene buena reducción en* $p$. Si la curva $\widetilde{E}$ no es elíptica,

se tienen los casos ilustrados en las figuras (5) y (6) previas, y en el primer caso se dice que $E$ tiene *reducción multiplicativa en* $p$ y en el segundo caso se dice que $E$ *tiene reducción aditiva en* $p$. Diremos que $E$ es una *curva elíptica semiestable* si para todos los primos $p$ su reducción es buena o multiplicativa. Nótese que los únicos primos $p$ donde $E$ puede tener mala reducción son aquellos primos que dividen a su discriminante mínimo y por lo tanto $E$ sólo tiene un número finito de primos donde tiene mala reducción.

Supongamos ahora que la curva reducida $\widetilde{E}$ mod $p$ es elíptica. El análogo de los puntos racionales para $\widetilde{E}$ son los puntos de $\widetilde{E}$ que tienen coordenadas en el campo finito $\mathbb{F}_p$. Es claro que el conjunto de puntos $\mathbb{F}_p$-racionales $\widetilde{E}(\mathbb{F}_p)$ es finito ya que $\mathbb{F}_p$ lo es y así $\widetilde{E}(\mathbb{F}_p)$ es más manejable que $E(\mathbb{Q})$. De hecho, por un teorema de A. Weil se tiene la fórmula siguiente para el orden del grupo $\widetilde{E}(\mathbb{F}_p)$:

$$\#\widetilde{E}(\mathbb{F}_p) = p + 1 - a_p,$$

donde $a_p$ es un cierto entero que depende del primo $p$. Este entero $a_p$ tiene la interpretación siguiente que se usa en una de las primeras reducciones de la conjetura de Fermat: cuando consideramos los ceros de un polinomio $f(x, y)$ con coeficientes en un campo (en nuestro caso $\mathbb{Q}$ ó $\mathbb{F}_p$), estas soluciones viven, en principio, en un campo más grande: una *cerradura algebraica* del campo en consideración. En el caso de $\mathbb{Q}$ se tiene una cerradura algebraica contenida en $\mathbb{C}$. Se tienen inclusiones del campo dado en la correspondiente cerradura algebraica: $\mathbb{Q} \subseteq \overline{\mathbb{Q}}$ y $\mathbb{F}_p \subseteq \overline{\mathbb{F}_p}$. Estas inclusiones son llamadas *extensiones de campos*, y asociado a cualquier extensión de campos como las anteriores, se tiene el grupo de automorfismos del campo grande que fijan al campo chico. Este es el *grupo de Galois* de la extensión dada, y en los ejemplos anteriores se suele denotar como $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ y $\mathrm{Gal}(\overline{\mathbb{F}_p}/\mathbb{F}_p)$. Estos grupos de Galois tienen, de entrada, un papel importante en la determinación de los puntos racionales de una curva elíptica dada, ya que, por ejemplo, $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ opera en forma natural sobre los puntos de $E$ de la forma siguiente: si $\sigma \in \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ y $(x, y) \in E$, entonces se define: $\sigma \cdot (x, y) := (\sigma x, \sigma y)$, donde $x, y \in \overline{\mathbb{Q}}$. Similarmente $\mathrm{Gal}(\overline{\mathbb{F}_p}/\mathbb{F}_p)$ opera sobre $\widetilde{E}$. Observemos ahora que los puntos racionales de $E$ (respectivamente, $\widetilde{E}$) son los puntos invariantes bajo la acción del grupo de Galois correspondiente. Ahora, como estos grupos de Galois son muy grandes, conviene restringir su acción a conjuntos más chicos que las curvas $E$ ó $\widetilde{E}$. Veamos cómo se hace esto: para el grupo abeliano $E$, y dado un entero $m$, consideremos los puntos de $E$ que tengan orden divisible por $m$ y denotemos este conjunto por $E[m]$; resulta que este conjunto es un

subgrupo de $E$ y, para $E$ definida sobre $\mathbb{Q}$, el grupo $E[m]$ es isomorfo a $\mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/m\mathbb{Z}$. En particular, para $m = p$ un primo se tiene que:

$$E[p] \simeq \mathbb{F}_p \times \mathbb{F}_p \,.$$

Notemos ahora que la acción del grupo $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ sobre $E$ se restringe a una acción en $E[m]$, y para $m = p$ un primo, esta acción nos da un automorfismo de $E[p]$ para cada $\sigma \in \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. Estos automorfismos de $E[p]$ son transformaciones lineales (sobre $\mathbb{F}_p$), y así, para cada $\sigma \in \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ se tiene un isomorfismo $\mathbb{F}_p$-lineal

$$\rho(\sigma) : E[p] \to E[p] \,,$$

y como $E[p] \simeq \mathbb{F}_p \times \mathbb{F}_p$ tiene dimensión 2 sobre $\mathbb{F}_p$, la matriz asociada a $\rho(\sigma)$ es de tamaño $2 \times 2$ y es invertible ya que $\rho(\sigma)$ es isomorfismo. Hemos asociado a cada $\sigma \in \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ una matriz $\rho(\sigma) \in \mathrm{GL}(2, \mathbb{F}_p)$ y de hecho esta asociación es un homomorfismo de grupos y decimos entonces que se tiene una *representación* (de grado 2) del grupo de Galois $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ en $\mathbb{F}_p$. Este es un procedimiento típico en matemáticas: a un objeto abstracto (en este caso, un grupo $G$) le asociamos objetos más manejables (matrices en este caso). Podemos ahora enunciar una relación más profunda entre estas representaciones del grupo de Galois $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ y el orden del grupo de puntos $\mathbb{F}_p$-racionales $\widetilde{E}(\mathbb{F}_p)$: existe un elemento $\mathrm{Fr}_p \in \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ tal que para casi todos (es decir, todos excepto un número finito) los primos $p$, la traza de la matriz $\rho(\mathrm{Fr}_p)$ es el entero $a_p = p + 1 - \#\widetilde{E}(\mathbb{F}_p)$. Para los primos $p$ donde $E$ tiene mala reducción, hay una forma especial de definir los números $a_p$ asociados a $E$.

En general, para cada campo finito $\mathbb{F}_{p^n}$, con $n \in \mathbb{N}$, podemos considerar la curva reducida $\widetilde{E}$ sobre $\mathbb{F}_{p^n}$ y tomar la función generadora asociada a la sucesión de enteros $\#\widetilde{E}(\mathbb{F}_{p^n})$:

$$Z(E, u) := \exp\left(\sum_{n=1}^{\infty} \frac{\#\widetilde{E}(\mathbb{F}_{p^n})}{n} \cdot u^n\right)$$

Se prueba que esta función zeta es de la forma siguiente:

$$Z(E, u) = \begin{cases} \frac{1 - a_p u + p u^2}{(1-u)(1-pu)} & \text{si } E \text{ tiene buena reducción en } p \\ \\ \frac{1 - a_p u}{(1-u)(1-pu)} & \text{si } E \text{ tiene mala reducción en } p \end{cases}$$

donde $a_p$ son los números definidos anteriormente. Notamos que en esta función zeta sólo el entero $a_p$ depende de la curva $E$ por lo que sólo los numeradores nos dan información sobre la curva.

Variando los primos $p$ y substituyendo $u = p^{-s}$, para $s \in \mathbb{C}$ se define la función $L$ de Hasse-Weil de la curva elíptica $E$, con discriminante $\Delta$, como:

$$L(E,s) := \prod_{p|\Delta} \frac{1}{1 - a_p p^{-s}} \prod_{p\nmid\Delta} \frac{1}{1 - a_p p^{-s} + p^{1-2s}} \,.$$

Desde mediados del siglo XX ya se sabía que $L(E,s)$ es holomorfa en el semiplano de los números complejos $s$ con parte real $\Re(s) > 3/2$ y se conjeturaba que tenía una *continuación analítica a todo* $\mathbb{C}$ *y debía satisfacer una ecuación funcional*. Esta importante conjetura, de Hasse y Weil, fue demostrada en los últimos años del siglo XX como consecuencia de otra conjetura de la cual esbozaremos algunos detalles a continuación.

**Formas modulares.** Habiendo de alguna forma arribado al universo de las funciones complejas (o quizá estábamos en este universo desde el principio), es tiempo de presentar a los otros actores que intervienen en la saga: de todas las funciones complejas, quizá las que tienen más simetrías son las funciones modulares. Estas son funciones definidas en el semiplano complejo superior y se comportan casi invariantes bajo la acción de ciertos subgrupos del grupo de matrices $2 \times 2$ con entradas enteras y determinante 1; las que nos interesan corresponden al subgrupo $\Gamma_0(N)$ dado por las matrices $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ con $a, b, c, d \in \mathbb{Z}$, con determinante 1 y además $c \equiv 0 \mod N$, para $N > 1$ un entero dado. Con respecto a este subgrupo, una función compleja $f(z)$ definida en el semiplano complejo superior, se dice que es modular de peso $k$ (un entero positivo) y nivel $N > 1$ si

$$f\left(\frac{az+b}{cz+d}\right) = (cz+d)^k f(z)\,.$$

Debido a esta invarianza, se sigue que $f(z)$ tiene una expansión en serie de Fourier de la forma

$$f(z) = \sum_{n=-\infty}^{\infty} c_n q^n, \qquad \text{con} \quad q := e^{2\pi\, iz}\,.$$

Las formas modulares que nos interesan son aquellas para las cuales en la expansión de Fourier anterior se tiene que $c_n = 0$ para todos

los $n < 0$, con algunas otras otras condiciones, y a estas formas se les llama *formas modulares parabólicas*. Es fácil ver que el conjunto $S(k, N)$ de formas modulares parabólicas de peso $k$ y nivel $N$ es un espacio vectorial sobre $\mathbb{C}$, y aunque esto ya no es tan sencillo, se sabe que existe una familia de operadores lineales $T_n$ en $S(k, N)$, para cada natural $n$, llamados operadores de Hecke y, como el álgebra lineal nos tiene acostumbrados, lo primero que piensa uno es en los vectores propios de estos operadores. Para peso $k = 2$, si $f \in S(2, N)$ es un vector propio de todos los operadores de Hecke (y es una forma nueva en un sentido técnico), entonces la expansión de Fourier de $f(z)$ se puede normalizar, i.e., se tiene que $c_1 = 1$, y si sucede que una vez normalizada se tiene además que todos los coeficientes de Fourier $c_n$ de $f(z)$ son *enteros*, entonces un teorema de Eichler nos dice que existe una curva elíptica $E_f$ definida sobre $\mathbb{Q}$, para la cual su función $L$ asociada $L(E_f, s) = \sum_{n=1}^{\infty} c_n/n^s$ es tal que los enteros $c_n$ son los mismos $c_n$ de la expansión de Fourier de $f(z)$. El hecho de que esta curva elíptica $E_f$ provenga de una forma modular $f(z)$ tan bien comportada tiene como consecuencia que la función $L(E_f, s)$ también se comporte bien, es decir, satisface la conjetura de Hasse-Weil. A mediados de la década de los años 1950, el matemático japonés Yutaka Taniyama fue el primero en sugerir que todas las curvas elípticas $E$ deberían provenir de una forma modular mediante la construcción de Eichler. En los años 1960, Goro Shimura precisó y dio sustento teórico a la sugerencia de Taniyama, resaltando en particular el hecho de que la curva debería estar definida sobre $\mathbb{Q}$ y no sobre cualquier campo de números. A finales de esa misma década, André Weil, probó un teorema que precisaba la conjetura y le daba más evidencia teórica mostrando el papel que jugaba el conductor de la curva. Por eso a esta conjetura, que en forma resumida nos dice que *toda curva elíptica definida sobre $\mathbb{Q}$ es modular (i.e., proviene de una forma modular)* se le conoce como la *conjetura de Shimura-Taniyama-Weil*, y hemos recordado cómo esta conjetura implica la conjetura de Hasse-Weil.

**La curva de Frey y la conjetura de Fermat.** Los matemáticos que trabajan en la aritmética de curvas elípticas consideran las dos conjeturas anteriores muy importantes y habían estado trabajando en ellas por varias décadas, sin embargo el aliciente final que llevó a la demostración de estas conjeturas, provino de nuestra antigua acompañante, la *conjetura de Fermat*, gracias a la penetración del matemático alemán Gerhard Frey, quien alrededor del año 1985, dio a conocer las siguientes ideas: supongamos por un momento que la conjetura de Fermat es falsa, i.e., que existen enteros no cero $a, b, c$ tales que $a^n + b^n = c^n$. Entonces,

Frey considera la curva elíptica cuya ecuación es:

$$E_{abc}: \qquad y^2 = x(x-a^n)(x+b^n).$$

Nótese que aunque $c^n$ no aparece en la ecuación de $E_{abc}$, en realidad está implícito por la igualdad $a^n + b^n = c^n$. La idea de Frey es que la curva $E_{abc}$ es bastante extraña: tiene coeficientes en $\mathbb{Q}$ y sin embargo por ciertas razones técnicas no debiera ser modular, es decir, violaría la conjetura de Shimura-Taniyama-Weil. Esta contradicción sólo se salva si reconocemos que los enteros $a^n, b^n, c^n$ no pueden existir, es decir, si la conjetura de Fermat es cierta. Esencialmente este es el programa esbozado por Frey en 1985 y 1986.

Una vez conocidas estas ideas de Frey, Jean-Pierre Serre, pudo precisar lo que se necesitaba para probar que la existencia de la curva de Frey violaba la conjetura de Shimura-Taniyama-Weil. En una carta a J.-F. Mestre, Serre formula unas conjeturas precisas sobre representaciones modulares del grupo $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ de tal forma que una (pequeña) parte de estas conjeturas implica que Frey estaba en lo correcto: la curva $E_{abc}$ no puede ser modular. La idea de Frey es observar que la representación módulo $p$ asociada a los puntos de $p$-torsión de $E_{abc}$ parece provenir de una forma modular parabólica de peso $k = 2$, sin embargo al buscar el nivel $N$ de esta forma se debería tener que $N = 2$. El problema es que ¡no hay formas modulares parabólicas diferentes de cero de peso 2 y nivel 2!. Se sigue que $E_{abc}$ no puede ser modular y por lo tanto contradice a la conjetura de Shimura-Taniyama-Weil. Ken Ribet, en 1990, usando la herramienta de la aritmética moderna, y considerando los puntos de torsión $E_{abc}[n]$, demuestra un teorema de descenso que implica la parte de la conjetura de Serre necesaria, completando el programa esbozado por Serre y Frey. Así, en 1990 ya se sabía que la conjetura de Shimura-Taniyama-Weil implica la conjetura de Fermat.

Al conocer lo anterior, Andrew Wiles de la Universidad de Princeton, comienza un proyecto con el objetivo de probar la conjetura de Shimura-Taniyama-Weil, al menos en el caso cuando la curva $E$ es semiestable (la curva de Frey es semiestable, un hecho que ya habían notado Frey y Serre). De entrada, Wiles usa las mismas herramientas que Ribet, es decir considera las representaciones de $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ asociadas a los puntos de torsión $E[n]$ de la curva $E$; la idea es mostrar que éstas son modulares para un conjunto infinito de enteros $n$. El punto de arranque fue considerar la familia de representaciones asociadas a $E[3^m]$, $m \geq 1$, ya que si la representación asociada a $E[3]$ es irreducible (i.e., si el espacio vectorial correspondiente no se descompone en suma directa de subespacios no triviales), entonces es modular como conse-

cuencia de los trabajos de Robert Langlands y Jerrold Tunnell de principios de los años 1980. En este momento Wiles ya tiene varias familias de representaciones, algunas de las cuales vienen de formas modulares (a estas representaciones les llamamos modulares) y otras provienen directamente de la curva elíptica $E$ y lo que se quiere probar es que estas últimas también son modulares. Aquí es donde entra en juego la teoría de deformación de representaciones de Galois desarrollada por Barry Mazur a mediados de la década de 1980, y la estrategia de Wiles es mostrar que las representaciones asociadas a $E$ son deformaciones de las representaciones modulares que ya tiene. En última instancia, el argumento final es un argumento de conteo: la idea es mostrar que hay más deformaciones que representaciones modulares, todo esto mediante un argumento algebraico (desarrollado en colaboración con R. Taylor) para calcular el tamaño del grupo de Selmer asociado. Por supuesto que podría suceder que $E[3]$ no es irreducible, y en este caso Wiles demuestra que existe otra curva elíptica $E'$ para la cual $E'[3]$ es irreducible y la representación asociada a $E'[5]$ es la misma que la de $E[5]$. Se sigue que $E'$ es modular y por lo tanto $E'[5]$ también lo es; usando teoría de deformación se deduce la modularidad de $E[3]$. El resultado final es la conclusión de que *todas las curvas elípticas semiestables definidas sobre* $\mathbb{Q}$*, son modulares.* Y como Frey ya había mostrado que la curva $E_{abc}$ es semiestable, el teorema de Wiles implica que es modular, lo cual junto con el teorema de Ribet implica que la conjetura de Fermat es cierta. ¡Ciertamente no un argumento que cupiera en el margen de cualquier libro!.

**Desarrollos posteriores.** O tal vez debiera escribir: *La vie dáprés Fermat.* ¿Qué ha sucedido después de que Wiles demostró la conjetura de Shimura-Taniyama-Weil para curvas elípticas semiestables?. Para comenzar, casi inmediatamente se obtuvieron simplificaciones de algunas partes de los argumentos, por G. Faltings y H. W. Lenstra y algunas generalizaciones a familias más grandes de curvas elípticas por Fred Diamond, culminando con la demostración de la conjetura completa en el año 1999 por F. Diamond, B. Conrad y C. Breuil.

Como consecuencia de esto, se tiene que la conjetura de Hasse es verdadera: para toda curva elíptica $E/\mathbb{Q}$, la función $L(E, s)$ se extiende a una función entera. Tiene entonces sentido hablar de si $s = 1$ es o no cero de $L(E, s)$, es decir tiene sentido ya una parte de la formulación de un grupo de conjeturas hechas a principios de la década de 1960 por Birch y Swinnerton-Dyer que esencialmente nos dicen que $s = 1$ es un cero de $L(E, s)$ de orden el rango del subgrupo libre del grupo

de puntos $\mathbb{Q}$-racionales de $E$. Pero estas conjeturas siguen bastante abiertas y el estímulo para trabajar en ellas se ha incrementado al anunciarse en Mayo del 2000 una serie de premios por la Fundación Clay para unos 6 problemas importantes de la matemática, que incluyen las conjeturas de Birch y Swinnerton-Dyer. En alguno de sus libros, Serge Lang afirma que es posible escribir sin fin sobre curvas elípticas, y como no tomaremos esto como una amenaza, mejor nos detenemos en este punto.

# Referencias

[1] P. Ribenboim, *13 Lectures on Fermat's Last Theorem*. Springer Verlag, New York 1979.

[2] K. Ribet, *A modular construction of unramified p-extensions of* $\mathbb{Q}(\mu_p)$. Inv. Math. **34** (1976), 151–162.

[3] K. Ribet, *On modular representations of* $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ *arising from modular forms*. Inv. Math. **100** (1990), 431–476.

[4] J.-P. Serre, *Sur les represéntations modulaires de degré 2 de* $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. Duke Math. J. **54** (1987), 179–230.

[5] A. Weil, *Number theory: An approach through history from Hammurapi to Legendre*. Birkhäuser Verlag, Boston 2000.

[6] A. Wiles, *Modular elliptic curves and Fermat's last theorem*. Ann. of Math. **142** (1995), 443–551.