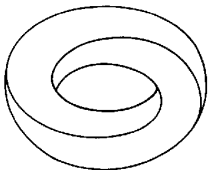
Krzywe Freya, formy modularne, hipoteza Shimury-Taniyamy i Wielkie Twierdzenie Fermata

Mariusz Gromada
styczeń 2003
mariusz.gromada@wp.pl
http://multifraktal.net



„Cubum autem in duos cubos, aut quadrato-quadratum in duos quadrato-quadratos, et generaliter nullam in infinitum ultra quadratum potestatem in duas ejusdem nominis fas est dividere; cujus rei demonstrationem mirabilem sane detexi. Hanc marginis exiguitas non caperet."

Pierre de Fermat

1 Wstęp

„Przeciwnie, nie można rozłożyć ani sześcianu na dwa sześciany, ani bikwadratu na dwa bikwadraty, i w ogóle żadnej potęgi większej niż druga na dwie potęgi z takim samym wykładnikiem. Odkryłem naprawdę zadziwiający dowód tego. Margines jest na to za mały."

Powyższe słowa odnaleziono na marginesie egzemplarza *Arytmetyki Diofantosa*[1] należącego do Fermata. Podejrzewa się, że dopisek pojawił się w 1630 roku.

Tekst poświęcony jest związkom krzywych Freya z dowodem Wielkiego Twierdzenia Fermata.

[1] Arytmetyka Diofantosa z komentarzami i uzupełnieniami (sześć znanych z trzynastu napisanych rozdziałów) wydana w greckiej i łacińskiej wersji w 1621 roku (G. C. Bachet de Meziriac).

1.1 Niezmienniki

Niezmiennikiem obiektu matematycznego w określonej klasie przekształceń (transformacji) nazywamy pewną wielkość charakteryzującą ten obiekt, która jest niezmiennicza względem tych przekształceń.

Przykłady:

1. *Niezmienniki topologiczne* w klasie przekształceń zwanych homeomorfizmami (np. *genus powierzchni*, czyli z grubsza liczba „dziur" w tej powierzchni.)
2. *Niezmienniki algebraiczne*, takie jak *wyróżnik wielomianu*, w klasie pewnych przekształceń algebraicznych (np. izomorfizmów).

1.2 Krzywe algebraiczne

Definicja 1.1 *Krzywą algebraiczną C nad ciałem K nazywamy krzywą zadaną równaniem $f(x, y) = 0$, gdzie $f(x, y)$ jest wielomianem zmiennych x i y o współczynnikach należących do ciała K.*

$$f(x) = a_{nm}x^n y^m + \cdots + a_{22}x^2y^2 + a_{21}x^2y + a_{12}xy^2 + a_{11}xy + a_{10}x + a_{01}y + a_{00}$$

$$a_{ij} \in K \qquad i = 0, \ldots, n \qquad j = 0, \ldots, m$$

Mówimy, że krzywa algebraiczna jest *nieosobliwa*, jeżeli krzywa ta nie posiada punktów osobliwych.

Definicja 1.2 *Punkt (x, y) leżący na krzywej algebraicznej C nad ciałem K nazywamy K-wymiernym punktem krzywej C, jeżeli x i y należą do ciała K.*

Definicja 1.3 *Zbiór:*

$$C(K) = \{\ (x, y) \in K^2 \mid f(x, y) = 0\ \} \tag{1}$$

nazywamy zbiorem K-wymiernych punktów krzywej C.

Definicja 1.4 *Jeżeli $g(x)$ jest wielomianem stopnia k, a $r_1, \ldots, r_k$ są jego pierwiastkami to wyróżnik wielomianu g definiujemy następująco:*

$$\Delta(g) = \prod_{1 \leqslant i < j \leqslant k} (r_i - r_j)^2 \tag{2}$$

2 Krzywe eliptyczne

Ogólnie powiedzieć można, że krzywe eliptyczne są pewnym rodzajem krzywych algebraicznych stopnia 3 (krzywych kubicznych). Krzywe te ogranicza część przestrzeni topologicznie równoważna torusowi. Genus takich krzywych równy jest 1 (podobnie jak i genus torusa). Przejście z torusa do postaci algebraicznej krzywej eliptycznej umożliwia *eliptyczna funkcja Weierstrass'a.*

Definicja 2.1 *Krzywą eliptyczną nad ciałem K nazywamy każdą nieosobliwą, jednorodną krzywą algebraiczną stopnia 3 i genus 1, z co najmniej jednym K-wymiernym punktem, który to punkt może być nieskończonością (przy uwzględnieniu odpowiednich przestrzeni rzutowych).*

Zazwyczaj za K przyjmuje się ciało liczb zespolonych C, rzeczywistych R, wymiernych Q, lub ciała skończone.

Definiuje się również pojęcie *izomorfizmu* krzywych eliptycznych. Każda krzywa eliptyczna jest więc reprezentantem swojej klasy abstrakcji względem relacji izomorfizmu. Dalej mówiąc o krzywej eliptycznej będziemy mieli na myśli całą klasę krzywych izomorficznych.

Jeżeli charakterystyka ciała K jest różna od 2 i 3 to równanie krzywej eliptycznej można zapisać w postaci:

$$y^2 = x^3 + ax + b \tag{3}$$

Mamy do czynienia ze znacznie gorszą sytuacją gdy charakterystyka ciała K jest równa 2 lub 3. Na szczęście ciała R, C, Q mają charakterystykę 0.

2.1 Wyróżnik krzywej eliptycznej

Niech E będzie krzywą eliptyczną nad ciałem K daną w postaci równania:

$$E : y^2 = g(x)$$

Definicja 2.2 *Wyróżnik krzywej eliptycznej E definiujemy następująco:*

$$\Delta = k\Delta(g) = k(r_1 - r_2)^2(r_1 - r_3)^2(r_2 - r_3)^2 \tag{4}$$

gdzie $r_1, \ldots, r_3$ są pierwiastkami $g(x)$.

Obecność stałej k w definicji wynika z faktu, że tak naprawdę definiujemy wyróżnik dla całej klasy krzywych izomorficznych.

Wyróżnik krzywej eliptycznej nad ciałem K, którego charakterystyka jest różna do 2 i 3 przyjmuje więc postać:

$$\Delta = -16(4a^3 + 27b^2) \tag{5}$$

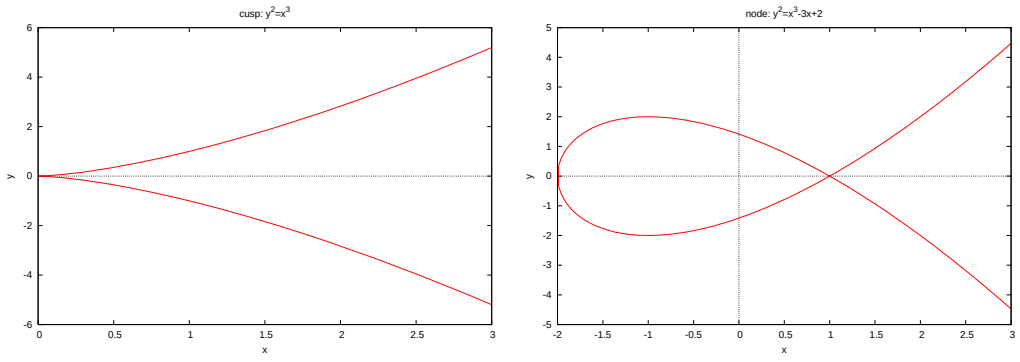
Tak zdefiniowany wyróżnik jest niezmiennikiem krzywej eliptycznej. Zamieniając założenie eliptyczności na algebraiczność otrzymujemy ciekawą interpretację geometryczną wyróżnika krzywej algebraicznej:

- $\Delta \neq 0$ to krzywa algebraiczna jest nieosobliwa i posiada genus 1 (jest więc to krzywa eliptyczna);
- $\Delta = 0$ to krzywa algebraiczna posiada co najmniej 1 punkt osobliwy (pierwiastki wielokrotne, nie jest to więc krzywa eliptyczna).

Wyróżnia się dwa typy osobliwości: *ostrza* (ang. cusp) i *węzły* (ang. node). Jeżeli jesteśmy w ciele o charakterystyce różnej od 2 i 3, to typ osobliwości zależy od współczynnika a (3). Dokładnie jeśli:

- $a = 0$ to punkty osobliwe są ostrzami;
- $a \neq 0$ to punkty osobliwe są węzłami.

wykres przedstawiający dwa typy osobliwości



Rysunek 1: Ostrze, węzeł

2.2 Krzywe eliptyczne nad ciałem Q

Niech E oznacza krzywą eliptyczną nad ciałem Q. Po odpowiedniej zamianie zmiennych możemy przejść do równości, w której występują całkowite współczynniki.

Twierdzenie 2.1 *Każdą krzywą eliptyczną E nad ciałem Q można przedstawić w postaci równania o całkowitych współczynnikach*

$$E : y^2 = g(x) = x^3 + a_2x^2 + a_1x + a_0 \tag{6}$$

gdzie a_1, a_2, a_3 są tak dobrane, aby wyróżnik $\Delta(E)$ był możliwie najmniejszy.

Tak przekształcona krzywa nazywana jest *modelem minimalnym* krzywej eliptycznej. Model minimalny można traktować jako krzywą algebraiczną nad ciałem Z. Nie musi to być krzywa eliptyczna nad Z (nad Q jest to nadal krzywa eliptyczna).

Stwierdzenie 2.1 *Całkowite współczynniki krzywej w postaci (6) implikują całkowitość jej wyróżnika.*

2.3 Redukcja modulo p krzywej eliptycznej

Niech będzie dana postać minimalna krzywej eliptycznej E nad ciałem Q

$$E : y^2 = g(x)$$

oraz liczba pierwsza p. Wprowadzamy kongruencję:

$$E_p : y^2 \equiv g(x) \ (mod\ p)$$

poprzez redukcję modulo p współczynników (ich obraz w ciele Z_p) krzywej E.

Łatwo jest zauważyć, że taka kongruencja określa nam pewną krzywą nad ciałem skończonym Z_p (Z_p ciało o charakterystyce p). Krzywą tę oznaczmy przez E_p i nazywamy *redukcją modulo p* krzywej eliptycznej E.

Definicja 2.3 *Mówimy, że krzywa eliptyczne E ma dobrą redukcję (mod p) jeżeli krzywa E_p jest krzywą nieosobliwą nad ciałem Z_p. W przeciwnym wypadku mamy do czynienia ze złą redukcją (mod p) krzywej eliptycznej E.*

2.3.1 Semi-stabilna redukcja modulo p

Zauważmy, że jeżeli liczba pierwsza p nie dzieli Δ, to E_p jest nieosobliwa, a więc i eliptyczna. Jeżeli p zaś dzieli Δ to mamy do czynienia z osobliwością. Wyróżniamy podtyp „złej redukcji".

Definicja 2.4 *Jeżeli osobliwość krzywej E_p jest węzłem, to mówimy o semi-stabilnej redukcji krzywej eliptycznej E.*

Zaznaczmy, że osobliwość może jeszcze być ostrzem.

2.4 Semi-stabilne krzywe eliptyczne

Definicja 2.5 *Mówimy, że krzywa eliptyczna E jest semi-stabilna, jeżeli posiada dobrą lub semi-stabilną redukcję dla każdej liczby pierwszej p.*

Semi-stabliność krzywej E oznacza, że dla każdej liczby pierwszej p tylko dwa z jej trzech pierwiastków są w relacji kongruencji (mod p).

2.5 Przewodnik (konduktor) krzywej eliptycznej

Zakładamy, że E nadal jest krzywą eliptyczną w postaci minimalnej.

Definicja 2.6 *Przewodnik N krzywej eliptycznej E określamy następująco:*

$$N = \prod_p p^{v_p} \qquad p \ - \ liczba\ pierwsza \tag{7}$$

$$v_p = \begin{cases} 0 & \text{jeżeli E ma dobrą redukcję mod p} \\ 1 & \text{jeżeli E ma semi} - \text{stabilną redukcję mod p} \\ 2 + \lambda_p & w.p.p. \end{cases}$$

Gdzie λ_p jest całkowite i nieujemne. Jedynie dla p=2,3 λ_p może być dodatnie.

Przewodnik N niesie informację o rodzaju osobliwości redukcji (mod p) krzywej E. Określa to dokładnie wykładnik liczby pierwszej p, z jakim dzieli ona przewodnik N. Ogólnie można powiedzieć, że

$$p|\Delta \Leftrightarrow p|N$$

Twierdzenie 2.2 *Krzywa eliptyczna E jest semi-stabilna jeżeli jej przewodnik N jest bezkwadratowy.*

3 Krzywe Freya

Dla danych liczb naturalnych A i B, względnie pierwszych, gdzie A jest podzielne przez 16, Frey rozpatrywał krzywą o równaniu:

$$y^2 = x(x - A)(x + B)$$

Frey wpadł na pomysł powiązania tego typu krzywych z ewentualnie istniejącymi rozwiązaniami równania Fermata. Dokładnie mówiąc jeżeli WTF nie zachodzi dla pewnego wykładnika pierwszego $q \geqslant 5$, to niech liczby naturalne a, b, c, parami względnie pierwsze, gdzie a jest parzyste, spełniają równanie:

$$a^q + b^q = c^q$$

Przyjmijmy, ze $A = a^q (\Rightarrow A$ jest podzielne przez 16) $B = b^q$. Dostajemy więc krzywą Freya o równaniu:

$$y^2 = x(x - a^q)(x + b^q) = x^3 + (b^q - a^q)x^2 - a^q b^q x = g(x)$$

oznaczmy ją przez E.

Wyróżnik tej krzywej wynosi:

$$\Delta = \Delta(g) = a^{2q} b^{2q} (a^q + b^q)^2 = (abc)^{2q}$$

i jest całkowity (zgodnie z poczynionym wcześniej stwierdzeniem.)

Jej minimalny wyróżnik to:

$$\Delta_{min} = \frac{(abc)^{2q}}{2^8} \tag{8}$$

Łatwo zauważyć, że $\Delta \neq 0$. Czyli tak zadana krzywa jest krzywą eliptyczną.

Frey doszedł do wniosku, że tego typu krzywa ma własności znacznie odbiegające od własności innych krzywych eliptycznych. Nabrał on przekonania, że tak być nie może. To było właśnie bodźcem do szukania dowodu prawdziwości WTF. Bo gdyby rzeczywiście nie istniały krzywe eliptyczne o takich własnościach jak ta krzywa Freya, to nie istniałyby również takie liczby a, b, c, które z założenia są rozwiązaniem równania Fermata.

Twierdzenie 3.1 *Krzywe Freya są semi-stabilne.*

Dowód (szkic). Niech E będzie krzywą Freya. Dla każdej liczby pierwszej p, rozpatrujemy krzywą E_p będącą redukcją (mod p) krzywej Freya E.

$$E_p : y^2 \equiv x^3 + (b^q - a^q)x^2 - a^q b^q x \quad (mod\ p)$$

Niech p będzie liczbą pierwszą.

1. Jeżeli p nie dzieli Δ, to E_p jest krzywą eliptyczną nad ciałem Z_p. W tym przypadku mamy do czynienia z dobrą redukcją $(mod\ p)$ krzywej Freya E.
2. Jeżeli p zaś dzieli Δ, to otrzymujemy osobliwą krzywą E_p. W osobliwości tej mamy tylko węzły. Mowa jest więc o semi-stabilnej redukcji krzywej Freya E (patrz wyróżnik krzywej eliptycznej).

$\Rightarrow$ krzywe Freya są semi-stabilne.

Policzmy przewodnik krzywej Freya.

$$N = \prod_p p^{v_p} = \prod_{p|\Delta} p \quad \text{gdzie p - liczba pierwsza}$$

Przewodnik ten jest oczywiście bezkwadratowy.

4 Liczby a_p

Dla każdej liczby pierwszej nie dzielącej Δ obliczamy liczbę punktów na krzywej eliptycznej E modulo p. Zwiększamy tę liczbę o 1, aby uwzględnić punkt w nieskończoności na odpowiedniej krzywej rzutowej. Definiujemy:

$$a_p = p + 1 - |E_p(Z_p)|$$

Liczby a_p nie muszą być dodatnie.

Głównymi charakterystykami dla krzywych eliptycznych są: wyróżnik, przewodnik liczby a_p. Do wyznaczania liczb a_p służą *formy modularne.*

5 Formy modularne

Niech $N \geqslant 1$ będzie liczbą całkowita i niech $\Gamma_0(N)$ będzie zbiorem takich macierzy:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

że a, b, c, d są całkowite, gdzie $N|c$ i $ad - bc = 1$. Wtedy $\Gamma_0(N)$ jest grupa multiplikatywną, zwaną *grupą kongruencji poziomu* N.

Niech H będzie górną półpłaszczyzną. $H = \{\ z = x + iy \in C \mid y > 0\ \}$

Określamy działanie grupy $\Gamma_0(N)$ na H:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} z = \frac{az + b}{cz + d}$$

Grupie $\Gamma_0(N)$ przyporządkowujemy jeszcze skończony zbiór punktów zwanych *ostrzami* (def. nie zamieszczamy). Są to:

- punkt w nieskończoności,
- pewne punkty na półprostej,
- pewne punkty z Q.

Niech $H^* = H \cup \{ostrza\}$

Podamy definicje formy modularnej tylko dla ciężaru 2, gdyż dalsze rozumowanie ogranicza się jedynie do tego przypadku.

Definicja 5.1 *Formą modularną poziomu N i ciężaru 2 nazywamy odwzorowanie*

$$f : H^* \to C$$

spełniające warunki:

1. *dla dowolnej macierzy* $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(N)$ *i* $z \in H^*$ *mamy:*

$$f\left(\frac{az + b}{cz + d}\right) = (cz + d)^2 f(z)$$

2. *funkcja f jest holomorficzna w każdym punkcie zbioru H^* (wymaga to określenia tego pojęcia dla ostrzy).*

Definicja 5.2 *Formę modularną znikającą we wszystkich ostrzach nazywamy formą paraboliczną.*

Oznaczenia:

$M_2(N)$ - zbiór form modularnych poziomu N i ciężaru 2
$S_2(N)$ - zbiór form parabolicznych poziomu N i ciężaru 2

5.1 Własności form modularnych

1. $M_2(N)$ jest przestrzenią liniową nad C;
2. $S_2(N)$ jest podprzestrzenią liniową przestrzeni liniowej $M_2(N)$;
3. $S_2(2)$ składa się tylko z formy zerowej;
4. $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \in \Gamma_0(N) \Rightarrow f(z+1) = f(z)$, czyli każda forma modularna ma rozwinięcie na szereg Fouriera postaci:

$$f(z) = \sum_{n=0}^{\infty} c_n e^{2\pi n z}$$

Dla form parabolicznych mamy $c_0 = 0$.

5.2 Krzywe eliptyczne a formy modularne

Dla danej krzywej eliptycznej, liczby a_p (gdzie p jest liczbą pierwszą nie dzielącą wyróżnika tej krzywej) niosą bardzo ważne informacje „lokalne" o krzywej. Zasadniczą sprawą jest powiązanie tych danych lokalnych za pomocą pewnego niezmiennika globalnego.

Idea ta jest bardzo daleko idącym uogólnieniem faktu, że każda liczba naturalna jest iloczynem potęg liczb pierwszych i to tylko na jeden sposób.

Jakiś czas temu zauważono na podstawie przeprowadzonych obliczeń, że dla wielu specjalnych krzywych eliptycznych liczby a_p są równe współczynnikom rozwinięcia w szereg Fouriera pewnej formy modularnej.

Definicja 5.3 *Krzywe eliptyczne o powyższej własności nazywamy krzywymi eliptycznymi modularnymi.*

5.3 Hipoteza Shimury-Taniyamy

Hipoteza 5.1 *Każda krzywa eliptyczna jest modularna.*

Jest to zwięzła forma pewnego stwierdzania, którego przedstawienie wymagałoby zdefiniowania wielu nieelementarnych pojęć.

5.4 Prace Ribeta

Ribet pokazał, że przy założeniu prawdziwości hipotezy Shimury-Taniyamy (dla krzywych eliptycznych semi-stabilnych) krzywe Freya nie mogą istnieć.

Skrót rozumowania Ribeta:

- zakładamy, że WTF nie zachodzi dla pewnego wykładnika q;
- konstruujemy krzywą Freya E dla tego wykładnika;
- krzywa E jest semi-stabilna o przewodniku N;
- zakładamy prawdziwość hipotezy Shimury-Taniyamy;
- dochodzimy do wniosku, że istnieje niezerowa forma paraboliczna f ciężaru 2 i poziomu 2;
- sprzeczność, gdyż $S_2(2)$ składa się tylko z formy zerowej.

Wniosek 5.1 *Z założenia prawdziwości hipotezy Shimury-Taniyamy dla semi-stabilnych krzywych eliptycznych, wynika niemożność istnienia krzywych Freya, co implikuje prawdziwość WTF.*

5.5 Dowód Wielkiego Twierdzenia Fermata

Andrew Wiles udowodnił hipotezę Shimury-Taniyamy dla krzywych eliptycznych semi-stabilnych.

Twierdzenie 5.1 *Każda semi-stabilna krzywa eliptyczna jest modularna.*

Wniosek 5.2 *Wielkie Twierdzenie Fermata jest prawdziwe!*

Literatura

[1] J.S. Milne: *Elliptic Curves*

[2] Ezra Brown: *Three Fermat Trails to Elliptic Curves*

[3] William F. Hammond: *Fermat's Last Theorem After 356 Years*

[4] Paulo Ribenboim: *Wielkie Twierdzenie Fermata dla laików*