The Way to the Proof of Fermat's Last Theorem

Gerhard Frey

1 Fermat's Claim

About 350 years ago *Pierre de Fermat* stated on the margin of a copy of Diophant's work
Fermat's claim : *There are no natural numbers* $n \geq 3, x, y, z$ *such that*

$$x^n + y^n = z^n \qquad (\mathbf{FLT}).$$

1993 *Andrew Wiles* announced the

Theorem: *Semistable elliptic curves over* $\mathbb{Q}$ *are modular.*

It is the aim of the lecture [1] to explain the meaning of Wiles' theorem, his strategy to prove it and why it settles Fermat's claim .

For this we have to browse through 350 years of mathematics facing the fact that the density of research increases exponentially. One main attraction of Fermat's claim is that everyone can understand it. This is certainly not true for Wiles' result and the conjecture of Taniyama which lies behind it. Their explanation needs a thorough training in number theory and algebra. But they open a whole new landscape for our understanding of the structural background of diophantine problems and so Wiles' theorem would be one of the greatest achievements in mathematics in our century even without Fermat's claim as consequence. Contrary to this Fermat's Last Theorem has *no* consequence but it stimulated the research of Wiles , and this is the typical role Fermat's claim played during the last centuries again and again and which made it so important for mathematics.

Some people are unhappy because of the fact that Wiles' proof of FLT is not "elementary". In fact this proof is not a sum of "ingenious" algebraic manipulations and it does not use giant computations: It just uses everything what we have learned during 350 years of research in number theory, algebra and calculus. It is not "naive" but natural and beautiful. We should recall that Fermat as one of the founders of calculus and number theory used the best mathematics of his time, too, to get his results. The same can be said about Euler, Dirichlet,

[1] This paper is based on a talk at the ISIT meeting 1997. The author wants to thank the organizers for the invitation and the warm hospitality.

Legendre and above all, Kummer who developed a great part of algebraic number theory and applied it to Fermat's claim . We should be glad that Fermat's claim was finally proved by using the best mathematics at our disposal and that it is not true because of an accident but because of a reason derived from general principles concerning the Galois group of the rational numbers and its geometric and automorphic representations which are fundamental for contemporary research in number theory.

2 Diophantine Problems

Our base in arithmetic is the set of natural number $\mathbb{N}$ with addition, multiplication and its natural order. It is well known how to get more algebraic structure by extending it to the integers $\mathbb{Z}$ (in which one can subtract and so it is a ring) and to the field of rational fractions $\mathbb{Q}$ (in which one can divide through all numbers $\neq 0$) . A typical *diophantine problem* consists of the following tasks:

Let $\mathbb{Z}[X_1, \cdots, X_m]$ be the ring of polynomials in m variables and integer coefficients, and let $f_1, \cdots, f_n \in \mathbb{Z}[X_1, \cdots, X_m]$.

- a) Find *all* solutions of the system of equations

$$f_1 = f_2 = \cdots = f_n = 0$$

  or prove that there are only finitely (or infinitely) many solutions which lie in the chosen arithmetical domain $\mathbb{N}, \mathbb{Z}$ or $\mathbb{Q}$.

- Describe the "arithmetical structure" of solutions by properties like "large, small, prime to ..., divisible by ..., with high prime powers" and so on.

Questions of the second type lead in a natural way to diophantine questions over residue rings $\mathbb{Z}/n\mathbb{Z}$ consisting of congruence classes of numbers: Two numbers z_1, z_2 are congruent modulo n ($z_1 \equiv z_2 \mod n$) if their difference is divisible by n. So $\mathbb{Z}/n\mathbb{Z}$ can be identified with the residues $\{0, \cdots, n-1\}$ and the set of solutions of polynomial equations $f_i(X_1, \cdots, X_m)$ mod n consists of (congruence classes) of residues $(r_1, \cdots, r_m)$ such that $f_i(r_1, \cdots, r_m)$ is divisible by n. An important special case is that n is equal to a prime number p. Then $\mathbb{Z}/p\mathbb{Z}$ is a finite field.

The nature and the difficulty of the answer to diophantine questions depends on the arithmetic domain one allows. As a rule the questions become easier if we change from $\mathbb{Z}$ to $\mathbb{Q}$

or even better, to $\mathbb{Z}/n\mathbb{Z}$ or $\mathbb{Z}/p\mathbb{Z}$. So a method often used in number theory is to "localize" the problem by studying it over residue rings $\mathbb{Z}/n\mathbb{Z}$ with $n = p^k$ and p a prime. The next and usually most difficult step is then to exploit the *local* data to get *global* information.

Here are two examples for diophantine problems related to Fermat's claim : We look at the "trivial" cases with exponent $n = 1$ and $n = 2$ in the equation (FLT). We begin with $n = 2$ and have to solve the equation

$$X^2 + Y^2 = Z^2.$$

From the shape of the equation (it is homogeneous of degree 2) one sees immediately that it is enough to determine solutions (x, y, z) which are relatively prime and non negative. We are only interested in non-trivial solutions which means that all coordinates have to be different from 0. We get at once a local information: Since the sum and the difference of two odd numbers is even exactly one of the coordinates of the solution is even, and a computation modulo 4 yields that z is odd. So we can assume without loss of generality that $y = 2y_1$ with $y_1 \in \mathbb{N}$. This is a global information. Hence $(x - z)(x + z) = 4y_1^2$. Now we use another global information: Every element in $\mathbb{N}$ can be written in a unique way as the product of powers of prime numbers. We apply this basic property of $\mathbb{Z}$ and conclude : $x = m^2-n^2, y = 2mn, z = m^2+n^2$ with $m > n, m-n$ odd and m, n natural numbers without a proper common divisor. We see that there are infinitely many solutions with relatively prime coordinates, and these triples are called "Pythagorean triples". (But this is the only connection of Fermat's claim with the theorem of Pythagoras!)

Now we make things even more trivial and take the exponent 1. Our equation is

$$X + Y = Z.$$

Again we are interested only in solution triples with coordinates in $\mathbb{N}$ which are relatively prime. The description of solutions is extremely easy: For any relatively prime natural numbers x, z with $z > x$ we take $y = z-x$ to get a solution. But there is a very interesting question related to the second kind of diophantine problems. The so-called **ABC-conjecture** predicts:
For every $\epsilon > 0$ there is a constant c depending only on ϵ such that for all relatively prime natural numbers x, z and $y = z - x$ we get:

$$z < c \cdot (\prod_{p|xyz} p)^{1+\epsilon}.$$

This conjecture and generalized versions of it play a central role in arithmetical geometry. They could give an alternative approach for the proof of Fermat's claim even in a more general frame and in asymptotic versions (cf.[F2]).

3 Cyclotomic Numbers

We continue to study the equation (FLT). Fermat's claim predicts that there are no non trivial solutions. To prove such a negative assertion one often uses the following strategy: Against our conviction we assume that we have found a solution. We use all our knowledge and try to find consequences built on the existence of the solution which contradict known facts. If we succeed to find such a contradiction and if we can exclude the possibility of a mistake in our arguments there remains only one other possibility: The assumption about the existence of a solution was wrong, and hence there is no solution!

Fermat himself developed such a machinery, the descente infinie. It works as follows: A solution of an equation is associated with a natural number measuring its size. Then one shows: There is an algorithm which produces for any given solution another one with a smaller size. Since the size is a natural number this can happen only finitely often, and we get our contradiction. This method is very elegant but it has the disadvantage that for each equation one has to invent its own machinery.

Fermat applied this method to the exponent 4. This allows us to assume in future that $n = p$ is an odd prime number.

100 years after Fermat Euler, (with a small gap) and even later Gauss (in his notebook) used descente infinie to settle the exponent 3, the exponent 5 was solved by Dirichlet and Legendre in 1820 and Lamé proved the case $p = 7$. All these proofs used descente infinie. But they became more and more involved and there was no perspective for a general approach. In addition to these results for small exponents there were remarkable results excluding "special" solutions (Sophie Germain). So 200 years after Fermat stated his claim the state of the art was not encouraging and in fact the methods of elementary number theory had come to their limits.

But one idea was hidden already in the proofs for $p = 3$. In fact the number

$$\zeta_3 := -1/2 + 1/2\sqrt{-3}$$

played an essential role. ζ_3 is a complex number whose third power is equal to 1. More generally complex numbers ζ different from 1 with $\zeta^p = 1$ are called $p-$th roots of unity. They are the zeroes of the polynomial $X^p - 1$ which are different from 1. One possible choice for such a number is $\zeta_p := e^{2\pi i/p}$ and all the other p-th roots of unity are powers of ζ_p. One can construct these numbers geometrically by dividing the unit circle in the Gaussian number plane into p pieces of equal length.

Lamé had the idea to use these numbers and to factorize the equation (FLT) to get

$$X^p + Y^p = \prod_{j=1}^{p}(X + \zeta_p^j Y) = Z^p.$$

The reader will remember a corresponding manipulation occurring in the case of exponent 2. But we paid a price: We have left our natural arithmetical domain $\mathbb{Z}$ and we have to do computations in the *cyclotomic integers*

$$\mathbb{Z}[\zeta_p] = \{z_0 + z_1\zeta + \cdots + z_{p-2}\zeta_p^{p-2};\ z_i \in \mathbb{Z}\}.$$

This set is a subring of the complex numbers and it turns out that many arithmetical rules one is used to in $\mathbb{Z}$ are valid in $\mathbb{Z}[\zeta_p]$ too. But there is a crucial difference: There are not enough prime elements to guarantee that we can factorize cyclotomic integers into powers of such elements! This is the flaw of many of the wrong "proofs" of Fermat's claim and Lamé himself did not realize this trap.

It was E.E. Kummer who laid the foundation to Algebraic Number Theory by an exact study of the arithmetic of the cyclotomic integers. His ingenious invention was the introduction of *ideal numbers* which we call nowadays *ideals*. They can be multiplied in a way which generalizes the multiplication of numbers. The role of prime numbers is taken over by prime ideals and they remedy the lack of a prime factorization for numbers: Every ideal is in a unique way the product of powers of prime ideals. But the transition from numbers to ideals causes a loss of information. Kummer saw how one can overcome this difficulty when looking for solutions of (FLT) in the case that this loss of information in $\mathbb{Z}[\zeta_p]$ only mildly affects the $p-$th powers. Such primes p were called "regular" primes. Kummer was able to invent a descente machinery for solutions of (FLT) for such primes and so he could prove Fermat's claim for regular primes. There are many regular primes. For instance up to 100 only 37 is not regular. But there are irregular primes, too, and ironically till today we can only prove that there that there are infinitely many irregular primes and not that there are infinitely many regular primes. Nevertheless Kummer's work and his beautiful criterion for regularity using Bernoulli numbers and refinements of these results made it possible to attack Fermat's claim one hundred years after Kummer's publications with the help of computers and so to prove that a counterexample would belong to astronomically high exponents and coordinates (cf.[R]). Essentially this was the state of art 340 years after Fermat 's note.

To understand the dramatic change afterwards we have to leave Fermat's claim for quite a while and to sketch the development of number theory during the 150 years after Kummer.

4 Algebraic Numbers with Galois Action

Nowadays the cyclotomic numbers are important but special elements in the *ring of algebraic integers*

$$\bar{\mathbb{Z}} = \{\alpha \in \mathbb{C}; \alpha \text{ is a zero of a polynomial } \underline{\text{with highest coefficient equal to 1}} \text{ in } \mathbb{Z}[X]\}.$$

These numbers are closed under addition and multiplication, they form a ring. Algebraic number theory is the study of arithmetical properties of $\bar{\mathbb{Z}}$. For instance one develops a modulo-arithmetic with regard to prime ideals $\mathfrak{p}$ instead of prime elements. Each $\mathfrak{p}$ contains exactly one prime number p and the congruence classes are infinite fields of characteristic p obtained as algebraic closure of $\mathbb{Z}/p\mathbb{Z}$. If we want to be able to divide we have to enlarge $\bar{\mathbb{Z}}$ to the field of algebraic numbers $\bar{\mathbb{Q}}$ defined in the same way as $\bar{\mathbb{Z}}$ but dropping the condition about the highest coefficients of the polynomials. By construction every non constant polynomial has zeroes in $\bar{\mathbb{Q}}$ and so diophantine questions seem to become trivial. But we have an all important additional structure: the operation of the *Galois group* $G_{\mathbb{Q}}$ of $\mathbb{Q}$!

$G_{\mathbb{Q}}$ is the set of maps σ from $\bar{\mathbb{Q}}$ to $\bar{\mathbb{Q}}$ with the property:

$$\text{For all} \quad x, y \in \bar{\mathbb{Q}} \text{ we have: } \sigma(x+y) = \sigma(x) + \sigma(y),\ \sigma(x \cdot y) = \sigma(x) \cdot \sigma(y).$$

It is easy to see that elements of $G_{\mathbb{Q}}$ permute zeroes of polynomials $f(X) \in \mathbb{Z}[X]$. Not so easy but important is the fact:
An algebraic number x lies in $\mathbb{Q}$ if and only if for all $\sigma \in G_{\mathbb{Q}}$ we have: $\sigma(x) = x$.

The composition of maps makes $G_{\mathbb{Q}}$ to a group, and a good part of modern diophantine mathematics can be described as the study of algebraic solutions of polynomial equations *with* Galois action. The importance of the Galois group for solutions of polynomials was discovered in the simplest cases already by Galois, Abel and Gauss. They solved famous classical problems with this new structure:

- A general angle cannot be trisected by using circle and rule.
- It is not possible to construct a cube of volume 2 from the unit cube by circle and rule.
- A regular p-gone (p a prime) is constructible by circle and rule if and only if $p = 2^{2^k} + 1$.

And most important:

- There is no general formula for the solutions of polynomials of degree ≥ 5.

Though these results are beautiful and should be part of everyone's education they need nearly no arithmetic. The key ingredient to relate arithmetic with group theory are the *Frobenius automorphisms:*
Let p be a prime. We recall the simple polynomial identity

$$(X+Y)^p \equiv X^p + Y^p \mod p$$

which follows from elementary properties of binomial coefficients. By evaluating this identity with x and y in $\bar{\mathbb{Z}}$ we get that exponentiation with p is compatible with addition (and of course with multiplication) in $\bar{\mathbb{Z}}$ modulo every prime ideal $\mathfrak{p}$ containing p. (The error term is in $\bar{\mathbb{Z}}$ divisible by p.)

Definition: *Let p be a prime number. $\sigma \in G_{\mathbb{Q}}$ is a Frobenius automorphism to p if there is a prime ideal $\mathfrak{p}$ of $\bar{\mathbb{Z}}$ containing p such that for all $x \in \bar{\mathbb{Z}}$ holds: $\sigma(x) - x^p \in \mathfrak{p}$.*

For fixed p there are (infinitely many) different Frobenius automorphisms but they are closely related. We choose one Frobenius automorphism for each prime p but we have to make sure that our assertions do not depend on this choice. The importance of Frobenius automorphisms is emphasized by "density theorems" going back to *Chebotarev* . They have to be understood as "local-global"- principles: The *global* Galois action of $G_{\mathbb{Q}}$ on $\bar{\mathbb{Z}}$ is determined by the Frobenius automorphisms which belong to the *localization* at the prime p (more precisely: to the action of the Galois group of the field of $p-$adic numbers).

5 Galois Representations

The assertions at the end of the last section can be made more precise by using the concept of Galois representations. The group $G_{\mathbb{Q}}$ is too huge to be studied directly and so we have to "linearize" by mapping it into matrix groups i.e. we look for homomorphisms (maps which are compatible with the composition law in $G_{\mathbb{Q}}$ resp. matrix multiplication)

$$\rho : G_{\mathbb{Q}} \to M_n(R)$$

where M_n are the $n \times n-$matrices with R as coefficient domain.

Here n is the dimension of the representation. For us $n = 2$ is the most important case. As coefficients we mostly use $R = \mathbb{Z}/n\mathbb{Z}$ with n a power of a prime number. But one

crucial ingredient in Wiles' proof uses a very deep result about two dimensional Galois representations with $R = \mathbb{C}$.

We begin with $n = 1$ and give an important example:
As above take $\zeta_p = e^{2\pi i/p}$. For all $\sigma \in G_{\mathbb{Q}}$ we get:

$$\sigma(\zeta_p) = \zeta_p^{k_\sigma}$$

where k_σ is a uniquely determined number between 1 and $p-1$. It is not difficult to see that

$$\chi_p : G_{\mathbb{Q}} \to (\mathbb{Z}/p\mathbb{Z}) \setminus \{0\} \text{ with } \sigma \mapsto \chi_p(\sigma) = k_\sigma \mod p$$

is a one-dimensional representation. It is called the *cyclotomic character* and plays together with its generalizations, the Dirichlet characters, the leading role in *class field theory* which is one of the most advanced parts of algebraic number theory developing since Kummer's time and central till today.

Remark (Warning:For mathematicians): One can interpret Kummer's results on Fermat's claim in this context as follows: Fermat's claim is true if one-dimensional Galois representations of the absolute Galois group of the field of cyclotomic numbers with a very special arithmetical property (to be unramified) do not exist. It turned out to be very difficult to generalize "class field theory" to higher dimensional Galois representations. Since the sixties of our century we have "Langland's philosophy" as ingenious guideline and the importance of the relation between Galois representations and automorphic functions is evident. The result of Wiles is a breakthrough in this philosophy for **two dimensional representations**

$$\rho : G_{\mathbb{Q}} \to \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \right\}$$

with $a, b, c, d \in \mathbb{Z}/p^k\mathbb{Z}$.

As is well known, the characteristic polynomial carries much of the coordinate-invariant information of a matrix. So we associate to $\rho(\sigma)$

$$\chi_{\rho(\sigma)}(T) = T^2 - \mathrm{Tr}(\rho(\sigma))T + \det(\rho(\sigma))$$

with $\mathrm{Tr}(\rho(\sigma)) = a_\sigma + d_\sigma$ and $\det(\rho(\sigma)) = a_\sigma d_\sigma - c_\sigma b_\sigma$.

Definition: *ρ is semisimple if ρ is determined (up to equivalence) by $\{\chi_{\rho(\sigma)}(T)\}$.*

For such representations we can state

Chebotarev's Density Theorem: *If ρ is semisimple then ρ is determined by* $\{\chi_{\rho(\sigma_p)}(T); \quad p \text{ runs over all primes}\}$. *It is even allowed to omit arbitrary finite sets of primes.*

As said before, Chebotarev's density theorem has to be interpreted as a local-global principle, and it turns out to be one of the most powerful amongst these principles. In the next section we shall study an important family of representations to which this principle is applicable.

6 Galois Representations Attached to Elliptic Curves

An **elliptic curve** E can be given as a plane cubic curve without singularities in the $2-$dimensional projective space. It is possible to choose an affine equation for E (missing just one point ∞) of the form

$$Y^2 = X^3 + AX^2 + BX + C =: f_3(X)$$

with A, B, C in $\mathbb{Z}$ and $f_3(X)$ without multiple zeroes. Hence Δ_E, the discriminant of $f_3(X)$ (which is up to sign equal to the product of the differences of the zeroes of $f_3(X)$) is different from 0. Over $\mathbb{C}$ such equations are well known as differential equation between elliptic Weierstrass functions $\wp$ and $\wp'$.

We assume that p is an odd prime. By definition E is semistable at p if $f_3(X)$ has at least 2 different zeroes modulo p. E has good reduction at p if f_E has 3 different zeroes modulo p. In this case $E \mod p$ is an elliptic curve over $\mathbb{Z}/p\mathbb{Z}$. The conductor N_E is a number which describes the primes at which E has bad reduction. If E is semistable at all primes p then $N_E = \prod_{p\, divides\, \Delta_E} p$. The set of algebraic points of E is

$$E(\bar{\mathbb{Q}}) := \{(x, y) \in \bar{\mathbb{Q}} \times \bar{\mathbb{Q}};\ \ y^2 = f_3(x)\} \cup \{\infty\}.$$

This set is an abelian group in which the addition law $\oplus$ is defined as follows: Take two points $P, Q \in E(\bar{\mathbb{Q}})$. Take the line through these two points. It intersects E in a third point. Reflecting this point at the X-axis gives the point $R := P \oplus Q$. It is easy to see that ∞ is the neutral element with respect to this addition. The addition law as a function of the coordinates is given by polynomials with coefficients in $\mathbb{Z}$ and so for all natural numbers n $G_{\mathbb{Q}}$ is acting on

$$E_n := E(\bar{\mathbb{Q}})_n := \{P \in E(\bar{\mathbb{Q}}); n \cdot P = \infty\}.$$

From the classical theory of elliptic functions it follows that there are points $P_1, P_2 \in E_n$ such that every $P \in E_n$ can be written as

$$P = \lambda_1 P_1 \oplus \lambda_2 P_2 \quad \text{with uniquely determined} \quad \lambda_i \in \{0, \cdots, n-1\}.$$

Now take $\sigma \in G_{\mathbb{Q}}$ and write $\sigma(P_1) = a_\sigma P_1 \oplus c_\sigma P_2; \sigma(P_2) = b_\sigma P_1 \oplus d_\sigma P_2$. The map

$$\sigma \mapsto \begin{pmatrix} a_\sigma & b_\sigma \\ c_\sigma & d_\sigma \end{pmatrix}$$

defines a two dimensional representation

$$\rho_{E,n} : G_{\mathbb{Q}} \to M_2(\mathbb{Z}/n).$$

We have the important

Theorem: *$\rho_{E,n}$ is semisimple for almost all natural numbers n.*

Hence we can apply Chebotarevs density theorem and it becomes important to compute the characteristic polynomials of the images of Frobenius automorphisms σ_l corresponding to prime numbers l. We are allowed to exclude "bad" primes dividing $n \cdot N_E$.

The result is due to H. Hasse and nearly a miracle:

Let a_l be the number of points of $E \mod l$. Then

$$\mathrm{Tr}(\rho_{E,n}(\sigma_l)) \equiv l + 1 - a_l \mod n \text{ and } \det(\rho_{E,n}(\sigma_l)) \equiv \chi_n(\sigma_l) \equiv l \mod n.$$

Consequence: *The polynomials*

$$\{\chi_l(T) = T^2 + (a_l - l - 1)T + l;\ \ l \text{ prime numbers not dividing } N_E\}$$

determine all the representations $\rho_{E,n}$.

We have found the *local factors of the L-series of E!*

Following Hasse we *globalize:*
For the bad primes we use an explicit recipe to define a rational function $f_E^*(s)$ and we form the infinite product

$$L_E(s) := f_E^*(s) \cdot \prod_{l \text{ prime to } N_E} (1 - (l + 1 - a_l)l^{-s} + l^{1-s})^{-1}$$

.

This product has to be seen as an analogue of the Riemann Zeta-function and it is called the L-series of E. Formally we write it as Dirichlet series

$$L_E(s) = \sum_{n=1}^{\infty} b_n n^{-s}.$$

It turns out that both the product and the sum converge for complex numbers s with real part $> 3/2$ and so $L_E(s)$ is an analytic function in a complex half plane.
We expect: This analytic function determines the arithmetical properties of E.

We know:

- L_E "encodes"the points of finite order of E with Galois structure.
- L_E determines the curve E (to be precise: up to isogeny). This is a famous theorem due to G. Faltings.

For the next step we again refer to the Riemann Zeta-function: It is well known that this function which is a priori defined only for complex numbers with real part > 1 can be extended to a meromorhic function of the whole set $\mathbb{C}$ such that this function satisfies a simple functional equation relating values at s with those at $1 - s$. That the same should be true for L_E is predicted by the central

Conjecture (Hasse):*L_E has an analytic continuation to $\mathbb{C}$ and satisfies a functional equation relating values at s with those at $2 - s$.*

So the local data used to define L_E are tied together in such a way that a very special analytic function is created.

7 Modular Elliptic Curves

Finally we can explain the **conjecture of Taniyama** stated 1955. In the last section we formulated Hasse's conjecture and said rather vaguely that the L-series of E is expected to be a very special function. Taniyama made this precise:
In the classical (i.e. 19th century) theory of elliptic functions it became already clear that complex functions behaving nicely under linear transformations play an important role. Especially **cusp forms of level N and weight k** are interesting. They are functions

$$f(z) = \sum_{n=1}^{\infty} b_n e^{2\pi i z} \quad \text{with } b_n \in \mathbb{C}$$

which satisfy:
For all $a, b, c, d \in \mathbb{Z}$ with $ad - Nbc = 1$ we have

$$f(\frac{az+b}{Ncz+d}) = (Ncz+d)^{-2} f(z).$$

For fixed N these functions form a finite dimensional $\mathbb{C}-$vector space which can be interpreted geometrically as space of holomorphic differentials of so-called modular curves (denoted by $X_0(N)$ in the literature) whose dimension can be calculated easily.

Example: *There is no non trivial cuspform of weight* 2 *and level* 2.

Conjecture (Taniyama): *Assume that Hasse's conjecture holds for the L-series*

$$L_E(s) = \sum_{n=1}^{\infty} b_n n^{-s}.$$

Then

$$f_E(z) := \sum_{n=1}^{\infty} b_n e^{2\pi i n z}$$

is a cusp form.

This conjecture has been made more precise by work of A. Weil, H. Carayol and especially G. Shimura who cleared the geometrical background. He showed that Taniyama's and Hasse's conjecture imply that there is a non trivial map from the modular curve $X_0(N_E)$ to the elliptic curve E. It has become common use to call such elliptic curves **modular**.

Conjecture (Taniyama-Shimura): *Every elliptic curve defined over* $\mathbb{Q}$ *is modular.*

We see that Theorem of A. Wiles on the first page proves part of this conjecture. (New results of F. Diamond give even more: E is modular if 27 does not divide N_E.)

For us the relations of cusp forms with Galois representations is most important. In fact the theory of modular curves and cusp forms has become one of the main streams in arithmetical geometry since the seventies of our century. Because of ground breaking ideas of Langlands and deep results of Deligne, Weil, Serre, Tate, Ribet, Mazur, Faltings and many other mathematicians, we have understood how geometry, representation theory of $G_{\mathbb{Q}}$ and complex function theory are interwoven inseparably in the modular theory.

We need the following technically rather complicated

Definition: *A representation*

$$\rho : G_{\mathbb{Q}} \to M_2(\mathbb{Z}/p^k)$$

is modular of level N if there is a cusp form of weight 2 and level N given by

$$f(z) = \sum_{n=1}^{\infty} b_n e^{2\pi i n z} \text{ with } b_n \in \bar{\mathbb{Z}},\ b_1 = 1$$

such that for all prime numbers l outside of a finite exceptional set we have:

$$Tr(\rho(\sigma_l)) \equiv b_l \mod \mathfrak{p}_k$$

where $\mathfrak{p}_k$ is an ideal of $\bar{\mathbb{Z}}$ different from $\bar{\mathbb{Z}}$ containing p^k and σ_l is a Frobenius automorphism to l.

(To get a feeling for this definition it is sufficient to think that the traces of the images of Frobenius automorphisms under ρ are congruent modulo p^k to integers coming from *one* modular form.)

Example: *Let E be a modular elliptic curve. Then ρ_{E,p^k} is modular for all primes p and all natural numbers k. As modular form we can take $f_E(z)$.*

Theorem: (Characterization of modular elliptic curves) *Let $E/\mathbb{Q}$ be an elliptic curve with $L_E(s) = \sum_{n=1}^{\infty} b_n n^{-s}$. The following properties are equivalent:*

1) E is modular.

2) Hasse's conjecture holds for E.

3) For all primes l and all $k \in \mathbb{N}$ the representation ρ_{E,l^k} is modular.

4) For <u>one</u> prime l and all $k \in \mathbb{N}$ the representation ρ_{E,l^k} is modular.

8 A Conditional Proof of Fermat's Claim

The development described in the last section was very fruitful for the study of arithmetical properties of points of finite order of elliptic curves. When doing this the author became aware that nearly unavoidably one was led to Fermat's claim and that one could link potential solutions of (FLT) to Galois representations attached to points of order p of a rather "exotic" elliptic curve and that this representations would have so few ramifications (here a parallel to Kummer's approach is obvious) that they should contradict properties of cusp forms. Hence either Fermat's claim should be true or Taniyama's conjecture should be

wrong (cf. [F1]). These reasonings were made precise by *K. Ribet*. The key ingredient is the phenomenon that cusp forms to different levels can induce the same representations. So for a given ρ one can look for forms of minimal level related to ρ. A recipe for this minimal level was given by J.P. Serre in principle already in the seventies and precisely formulated 1986. In the relevant example this recipe was proved by Ribet in the same year. We state his result for the example we need:
Let E be given by the equation

$$Y^2 = X(X-A)(X-B) \text{ with } A, B \in \mathbb{Z} \text{ relatively prime.}$$

It follows that E is semisimple.
Assume that E is modular and that p divides $AB(A-B)$ exactly with a power divisible by p.
Then $\rho_{E,p}$ is modular of level

$$N_p = 2\prod l$$

where the product runs over those primes l which divide $AB(A-B)$ exactly with a power not divisible by p.

Now take $A = x^p, B = y^p$ and assume that $A - B = z^p$.
Take the corresponding elliptic curve E and assume that E is modular.
Then $\rho_{E,p}$ is modular of level 2. But we know already that there are no non trivial cusp forms of this level and so we get a **contradiction**.

Conclusion: *The conjecture of Taniyama-Shimura for semistable elliptic curves implies Fermat's claim.*

This was the state of the art in 1986 and it was A.Wiles who had the courage to take this conclusion seriously and to prove Fermat's claim by proving Theorem 1.1.

9 The Theorem of Wiles

In 1994, Andrew Wiles published a paper [W] in the Annals of Mathematics (together with a joint work with Richard Taylor ([T-W])) in which he proved:

Every elliptic curve defined over $\mathbb{Q}$ which is semistable at the primes 3 and 5 is modular.

As explained above he got as a corollary Fermat's claim .

It is not possible to give Wiles' proof in detail. So I shall restrict myself to give some hints for his strategy. But the proof of Wiles' result is accessible to mathematicians with good education in arithmetical geometry since Wiles' original paper, the treatment in [DDT] and the expositions in [MF] give a fair guidance to it.

Wiles had to show one of the criteria for modularity we listed in section 7. His choice was criterion 4 and for the prime l he took $l = 3$. His starting point was $\rho_{E,3}$ so all his input information was the action of $G_{\mathbb{Q}}$ on 8 points. The reason for this choice was one additional deep information: $\rho_{E,3}$ can be interpreted as representation into matrices with complex entries, too, and the best result (due to Langlands and Tunnell) we have for such representations is that for this complex representation Artin's conjecture (which has the same flavor as Hasse's conjecture) is true and hence $\rho_{E,3} =: \rho_0$ is modular. So the beginning of the induction is done.

Now one has to show that for all $k \in \mathbb{N}$ the representation $\rho_{E,3^k}$ is modular. This cannot be done directly. Wiles has to look at all representations ρ' of $G_{\mathbb{Q}}$ with image in $M_2(\mathbb{Z}/3^k)$ which become equal to ρ_0 modulo 3 and to show that "enough" of them, including $\rho_{E,3^k}$, are modular. This leads to a deformation problem for Galois representations which can be described by a "tangent space". This space is computable if appropriate local conditions $\mathcal{D}$ (i.e. conditions for the restrictions of ρ' to the Galois group of $l-$adic fields) are imposed. In the beginning $\mathcal{D}$ has to be chosen such that $\rho_{E,3^k}$ satisfies the conditions. Deformations of "type" $\mathcal{D}$ are studied and it is shown that there exists a universal deformation represented by a ring $R_{\mathcal{D}}$: Deformations of ρ_0 of type $\mathcal{D}$ correspond one-one to homomorphisms of $R_{\mathcal{D}}$.

But we are interested in *modular* deformations only. The theory of Hecke operators is used to describe these deformations of type $\mathcal{D}$ by a ring $H_{\mathcal{D}}$ and the beginning of the induction implies that there is a homomorphism

$$\eta_{\mathcal{D}} : R_{\mathcal{D}} \to H_{\mathcal{D}}$$

which is surjective because of Chebotarev's density theorem.

One has to prove: $\eta_{\mathcal{D}}$ is a one-to-one map.

By using algebraic number theory Wiles can describe the algebraic properties of $R_{\mathcal{D}}$ and he can control how this ring changes if one replaces the type $\mathcal{D}$ be another (less or more restrictive) type. By using variants of Ribet's theorem a similar computation can be done for $H_{\mathcal{D}}$. This establishes the first important step (based on the "numerical criterion" of Wiles) of the proof: It is sufficient to show the injectivity of $\eta_{\mathcal{D}}$ for minimal types where minimality is determined by ρ_0 (and not by $\rho_{E,3^k}$).

For the proof in the minimal case Wiles uses another criterion which has a more geometrical flavor. It is written up in [T-W] and has been simplified by Faltings, Schoof, Diamond and others. To apply it Wiles adds carefully chosen auxiliary primes to $\mathcal{D}$ which make the structure of $R_{\mathcal{D}}$ and $H_{\mathcal{D}}$ so "easy" (Gorenstein property, complete intersection) that commutative algebra finally gives the result.

References

[MF] Modular forms and Fermat's Last Theorem, ed. G.Cornell, J.H.Silverman, G.Stevens, New York 1997

[DDT] H.Darmon,F.Diamond,R.Taylor, Fermat's Last Theorem; in Current Developments in Mathematics, Cambridge 1995

[F1] G.Frey, Links between stable elliptic curves and certain Diophantine equations, Ann. Univ. Saraviensis, 1(1986), 1-40

[F2] G.Frey, On ternary equations of Fermat type and relations with elliptic curves, 527-548, in [MF]

[R] P.Ribenboim, 13 Lectures on Fermat's Last Theorem, New York 1982

[T-W] R.Taylor, A.Wiles, Ring theoretic properties of certain Hecke algebras, Annals of Math. 141(1995), 553-572

[W] A.Wiles, Modular elliptic curves and Fermat's Last Theorem; Annals of Math. 142(1995), 443-551

Gerhard Frey
Institute for Experimental Mathematics
University of Essen
Ellernstraße 29
D-45326 Essen, Germany
e-mail: frey@exp-math.uni-essen.de

Gerhard Frey was born in Bensheim, Germany, on June 1st, 1944. In 1967 he graduated in mathematics and physics at the University of T"ubingen. He continued his postgraduate

studies in Heidelberg where he received the Ph.D. degree in 1970 and his "Habilitation" in 1973.

He was assistant professor at the University of Heidelberg from 1969-1973, professor at the University of Erlangen (1973-1975) and at the University of Saarbr"ucken (1975-1990) and has currently a chair for number theory at the Institute for Experimental Mathematics at the University of Essen. His research areas are number theory and arithmetical geometry as well as coding theory and cryptography as application. He was a visiting scientist at several universities and research institutions, e.g., at OSU in Columbus, Ohio, Harvard University, U.C. and MSRI at Berkeley, the Inst. f. Adv. Stud. at the Hebrew Univ. Jerusalem and at IMPA in Rio de Janeiro.

Prof. Frey is co-editor of the "manuscripta mathematica". He has been awarded the Gauss medal of the "Braunschweigische Wissenschaftliche Gesellschaft" in 1996 for his work on Fermat's Theorem. Since 1998 he is a member of the Academy of Sciences of G"ottingen, Germany.