

Modular Forms

Bas Edixhoven, Gerard van der Geer and Ben Moonen

*There are five fundamental operations in mathematics: addition,
 subtraction, multiplication, division and modular forms*

—M. Eichler¹

Modular functions played a prominent role in the mathematics of the 19th century, where they appear in the theory of elliptic functions, i.e., elements of the function field of an elliptic curve, but also in the theory of binary quadratic forms. The term seems to stem from Dirichlet, but the functions are clearly present in the works of Gauss, Abel and Jacobi. They play an important role in the work of Kronecker, Eisenstein and Weierstrass, and later in that century they appear as central themes in the work of Poincaré and Klein. The theory of Riemann surfaces developed by Riemann became an important tool, and Klein and Fricke studied and popularized the Riemann surfaces defined by congruence subgroups of the modular group $SL(2, \mathbb{Z})$.

Modular forms appear as theta functions in the work of Jacobi in the 1820's, and, up to a factor $q^{1/24}$, already in Euler's identity

$$\prod_{n \geq 1} (1 - q^n) = \sum_{k \in \mathbb{Z}} (-1)^k q^{k(3k-1)/2}.$$

They show up in a natural way in the expansions of elliptic functions and as such they were studied by Eisenstein, but the concept of modular forms was formalized only later. Apparently, it was Klein who introduced the term “Modulform”, cf. page 144 of Klein-Fricke [12].

One had to wait till Hecke for the first systematic study of modular forms on $SL(2, \mathbb{Z})$ and its congruence subgroups. The first appearance of the word “Modulform” in Hecke's work seems to be in [11].

¹ Apocryphal statement ascribed to Martin Eichler, March 29, 1912–October 7, 1992.

A crucial point in our story came when Hecke introduced the “averaging” operators that bear his name and that give essential arithmetic information on modular forms. Given (in modern terminology) a Hecke eigenform f on $\Gamma_1(N)$ with Fourier series $\sum a(n)q^n$, normalised by the condition $a(1) = 1$, Hecke could interpret the Fourier coefficient $a(n)$ as the eigenvalue of his operator $T(n)$. This also enabled him to express the Dirichlet series $L(f, s) = \sum_{n \geq 1} a(n)n^{-s}$ as an Euler product $\prod_p (1 - a(p)p^{-s} + \varepsilon(p)p^{k-1-2s})^{-1}$, where k is the weight of f and $\varepsilon: (\mathbb{Z}/N\mathbb{Z})^\times \rightarrow \mathbb{C}^\times$ its character. Thus he generalized a result of Mordell, who had proved in 1917 the multiplicativity of the Ramanujan τ -function that gives the Fourier coefficients of the weight 12 cusp form Δ . (This property of the τ -function had been observed by Ramanujan in 1916.) Though the eigenvalues of eigenforms showed a definite arithmetic flavour, it remained at that time a mystery why there should be arithmetic information in the Fourier coefficients of eigenforms. Hecke did not know, at that time, that the space of cusp forms of a given weight and level possesses a basis of eigenforms for the Hecke operators $T(n)$ with n prime to the level. But a little later Petersson defined an inner product with respect to which these $T(n)$ are normal, and with this it followed that such a basis exists. Hecke also proved, using the Mellin transform, that the Dirichlet series $L(f, s)$ associated to a cusp form f of weight k on $\Gamma_1(N)$ has an analytic continuation to a holomorphic function on the whole complex plane and satisfies a functional equation relating $L(f, s)$ to $L(g, k - s)$, where $g(\tau) = \tau^k f(-1/N\tau)$.

The second important step that Hecke made was to characterise the Dirichlet series $\sum_{n > 0} a(n)n^{-s}$ of the form $L(f, s)$ with f a cusp form of weight k on $\mathrm{SL}(2, \mathbb{Z})$ by regularity conditions and a functional equation relating $L(f, s)$ to $L(f, k - s)$. Indeed, a Fourier series $f = \sum_{n \geq 1} a(n)q^n$ that is holomorphic on the upper half plane is a cusp form of weight k on $\mathrm{SL}(2, \mathbb{Z})$ precisely when $f(-1/\tau) = \tau^k f(\tau)$. This so-called converse theorem generalized a theorem of Hamburger, saying that a sufficiently regular Dirichlet series that satisfies the functional equation of the Riemann zeta function is in fact a multiple of the Riemann zeta function.

The L -function that Hecke associates to a cusp form has its roots in earlier work of Gauss, Dirichlet and Riemann. But although Hecke was working at the same mathematics department (in Hamburg) as Artin, who was then working on his Artin L -series for representations of the Galois group of a number field, it seems that neither of them appreciated the link between the two types of L -functions. This may seem odd to us, but it is good to realize that the moment that the link was recognized in its full conjectural setting represents a second turning point in our history. Indeed, looking from a large distance

one may distinguish two turning points for the history of modular forms in the 20th century: Hecke's introduction of the Hecke operators and his converse theorem, and Langlands's letter of January 1967 to Weil, in which he laid out a grand program in which modular forms are an incarnation of non-abelian class field theory. Langlands's letter pointed out the common source for the L -series of Hecke and Artin, and brought the two types of L -functions together in a larger framework. We will come to that later.

But at the time that Hecke revolutionized the topic, it also lost its prominence, as novel developments in topology and algebra started to attract more attention. This was a time when many new concepts appeared, like the notions of algebraic topology and homology theory, and when new algebraic structures like rings and algebras were studied. These notions completely changed the face of mathematics at the time. Klein writes in this connection: "Es hat sich hier ein merkwürdiger Umschwung vollzogen. Als ich studierte, galten die Abelschen Funktionen—in Nachwirkung der Jacobischen Tradition—als der unbestrittene Gipfel der Mathematik, und jeder von uns hatte den selbstverständlichen Ehrgeiz, hier selbst weiter zu kommen. Und jetzt? Die junge Generation kennt die Abelschen Funktionen kaum mehr." (Vorlesungen über die Entwicklung der Mathematik, VII.)

In retrospect these developments, like the construction of homology and cohomology, the emergence of new algebraic structures and the development of an algebraic foundation for algebraic geometry, were the necessary ingredients for the later growth of the theory of modular forms.

The fact that there was a shift of focus to new topics in mathematics does not mean that the theory of modular forms came to a standstill. Throughout the 20th century there have been new ideas and generalizations, broadening but also deepening the subject. Some of these generalizations dealt with the extension of the notion of modular forms to other groups. An example of this is the step from $SL(2, \mathbb{Z})$ to the group $SL(2, O_K)$ with O_K the ring of integers of a totally real field, the Hilbert modular group. Hilbert was inspired by Kronecker's "Jugendtraum" about generating abelian extensions of imaginary quadratic fields. The Kronecker-Weber theorem says that all abelian extensions of \mathbb{Q} are contained in the field generated, over \mathbb{Q} , by all roots of unity, i.e., by the torsion points of the circle group. It was also found that for an imaginary quadratic field K , the values of a suitable elliptic function at the torsion points of an elliptic curve with complex multiplication by O_K could be used to generate abelian extensions of K . Hilbert envisioned an analogue of the Kronecker-Weber theorem and the theory of complex multiplication for abelian extensions of CM-fields (totally imaginary quadratic extensions of totally real number fields). He devoted to this the 12th of his famous

Mathematische Probleme, presented at the ICM 1900 in Paris.² As part of his investigations, Hilbert had worked out a theory of modular functions for totally real fields, more precisely for modular functions for the action of $SL(2, \mathcal{O}_K)$ on the product of $n = [K : \mathbb{Q}]$ upper half planes. He wrote an unpublished manuscript about it, and under his guidance his student Blumenthal wrote his *Habilitationschrift* about the basics of the theory. Hecke, also a student of Hilbert, wrote his thesis about it, this time with the purpose of setting up a theory of abelian extensions of quartic CM-fields. After these beginnings this development seemed to dry up, and though impressive progress has been made, Hilbert's 12th problem is to date unsolved. But recently two new ideas have been launched: Manin's "Altersträum", and Darmon's "Stark-Heegner points".

In the years after Hecke the number of mathematicians involved in modular forms shrank to a small group, including Eichler, Maass, Petersson and Rankin, but they continued to contribute. In 1946 Maass, working under difficult circumstances in postwar Germany, showed that one could sacrifice holomorphicity by considering eigenfunctions of the Laplacian $y^2(\partial^2/\partial x^2 + \partial^2/\partial y^2)$ that are invariant under the modular group.

In another direction, Siegel generalized the notion of the modular group inspired by his quantitative theory of representations of quadratic forms by quadratic forms, and also by the theory of period matrices of Riemann surfaces; see [19]. He made a detailed study of the symplectic group and its geometry, thus picking up a thread left by Riemann and neglected by many, Scorza being one of the exceptions. In his groundbreaking paper of 1857, Riemann had introduced the period matrix of a Riemann surface of genus g , and had shown that it can be normalized in the form of a complex symmetric g by g matrix with positive definite imaginary part. Siegel considered the so-called Siegel upper half space \mathbb{H}_g of all such period matrices, on which the symplectic group acts by fractional linear transformations. He determined a fundamental domain and its natural volume, studied the function field of the quotient space $Sp(2g, \mathbb{Z}) \backslash \mathbb{H}_g$, and he introduced the notion of a (Siegel) modular form. Siegel's main motivation was his desire to describe in a quantitative way the representations of integral quadratic forms by other quadratic forms. His central result can be expressed as an equality of a theta series with an Eisenstein series for the Siegel modular group.

In the 1950's and 1960's another vast generalization of the theory of modular forms was conceived by the introduction of the general notion of automorphic form, and of the subsequent adèlisation of this concept. According to Borel

² See [15] for further historical information on Hilbert's 12th problem and Kronecker's "Jugendtraum".

and Jacquet in [3] and [4], it had first been observed by Gelfand and Fomin that modular forms and other automorphic forms on the upper half plane and other bounded symmetric domains can be viewed as smooth vectors in representations of the ambient Lie group G on suitable spaces of functions on G that are invariant under the discrete subgroup Γ . A general definition was given by Harish-Chandra in [9] for a semisimple connected Lie group G , a discrete subgroup Γ and a maximal compact subgroup K . An automorphic form on G with respect to K and Γ is then a left- Γ -invariant and right- K -finite smooth function $f: G \rightarrow \mathbb{C}$, finite under the center of the enveloping algebra of the Lie algebra of G , and satisfying a certain growth condition. The consideration of the system of all congruence subgroups of a connected reductive group G over \mathbb{Q} then led to the notion of automorphic forms on the group $G(\mathbb{A})$ of adèlic points of G , and this notion was then further generalised to connected reductive groups over global fields F . An important consequence of this point of view is that the space of automorphic forms on $G(\mathbb{A})$ can be studied as a representation of the group $G(\mathbb{A}_f)$, as well as of K and the Lie algebra of $G(\mathbb{R})$. Irreducible representations thus obtained are called automorphic representations; they can be decomposed as restricted tensor products of irreducible representations of the local groups $G(F_v)$. In a precise way, these local representations generalise the systems of eigenvalues of a Hecke eigenform. Especially the Russian school contributed to the early development in this direction (Gelfand, Graev, Piatetskii-Shapiro, ...). The necessary theory of algebraic groups, arithmetic subgroups and adèle groups had been developed in the meantime, see for example [5].

On another stage but also during the 1950's and 1960's, weight two modular forms for congruence subgroups of $SL(2, \mathbb{Z})$ were related to differential forms on modular curves, and hence to the Jacobian varieties of modular curves, also in positive characteristic. Advances in algebraic geometry made it possible to study the reduction of curves and Jacobians at almost all primes. This led to the identification of the (partial) Hasse-Weil zeta functions of modular curves with a product of L -functions of such modular forms, at least at almost all primes (Eichler, Shimura; see [17]), thus proving the meromorphic continuation and the existence of a functional equation for these zeta functions. Kuga and Shimura were even able to do the same for forms of higher weight on the unit group of a quaternion algebra (see [13]).

In particular, the Hasse-Weil L -functions of elliptic curves over \mathbb{Q} occurring as isogeny factor of the Jacobian of a modular curve were identified (up to finitely many Euler factors) with L -functions of modular forms. Deuring proved in 1955 that the L -function of an elliptic curve with complex multiplication is a product of two Hecke L -functions associated to

“Größencharaktere”. In the same year, Taniyama [20] raised the question whether the meromorphicity and functional equation of the Hasse-Weil zeta functions of elliptic curves over number fields could be proved by finding suitable automorphic forms (see [18], where Shimura evokes a vivid portrait of their interaction in that time). Taniyama’s idea was that the expected functional equation should imply modularity for the associated Fourier series along the lines of Hecke who characterized modular forms on $SL(2, \mathbb{Z})$ by the functional equation of their associated Dirichlet series.

In [24] Weil extended Hecke’s argument by showing that if for sufficiently many Dirichlet characters χ the Dirichlet series $\sum \chi(n)a(n)n^{-s}$ associated to a function f on the upper half plane given by a Fourier series $\sum a(n)q^n$ have a suitable continuation to \mathbb{C} and satisfy an explicitly given functional equation then f is a modular form on a congruence subgroup $\Gamma_0(N)$ (with N determined by the functional equations). At the end Weil states the modularity question for an elliptic curve E over \mathbb{Q} in a precise form: the complete Hasse-Weil L -function is defined, as well as the conductor of E , and the expected functional equations. It was this paper of Weil that drew renewed attention to the modularity question for elliptic curves over \mathbb{Q} .

In January 1967 Langlands wrote a letter [14] to Weil that marked the start of the “Langlands program”. The main idea of this program is that the L -function associated to a Galois representation should coincide with the L -function that can be associated to some “algebraic” automorphic representation (generalising algebraic Hecke characters on idèle groups), and therefore has an analytic continuation and satisfies a functional equation. For example, the Artin L -function for an irreducible continuous n -dimensional complex representation of the Galois group of a number field F should be the L -function associated to an automorphic cuspidal representation of $GL(n, \mathbb{A}_F)$. This leads to conjectural correspondences, both global and local, between Galois representations and automorphic representations, characterised by being compatible with suitable L -factors and ε -factors. Also compatible systems of l -adic representations can be taken into account, and general reductive groups G over number fields F are considered. The Langlands dual group ${}^L G$ is introduced in order to formulate the natural (conjectural) relations between automorphic representations on different reductive groups: the functoriality principle. In collaboration with Jacquet, Langlands gave support for the functoriality principle by working it out and establishing the Jacquet-Langlands correspondence for the group $GL(2)$ and its inner twists (unit groups of quaternion algebras). Here, trace formulas (Selberg) play the main role. The Langlands program constitutes a grand framework for number theory, representation

theory and algebraic geometry, and has become one of the focal points in pure mathematics.

By that time the new methods of algebraic geometry, after the revolution in that field led by Grothendieck, came to play their role in the theory of modular forms. Eichler and Shimura had shown that the space of modular forms of weight $k \geq 2$ and level N can be interpreted as the $(k - 1, 0)$ -part of the Hodge decomposition of the cohomology of a suitable local system on the modular curve $X_1(N)$: the $k - 2$ symmetric power of the rank two local system given by the fiberwise cohomology of the universal family of elliptic curves. In 1968 Deligne showed that the l -adic étale cohomology of a non-singular projective model of the $k - 2$ power of the universal elliptic curve over the j -line provides the Galois representations then conjecturally associated to modular forms. Here Deligne had to deal with the technical difficulties caused by the presence of cusps. As a consequence of his results, the Ramanujan conjecture on the absolute value of the Fourier coefficients of these modular forms would follow from the Weil conjectures on the cohomology of non-singular projective varieties over finite fields. Six years later Deligne himself proved the last open part of these conjectures, and Ramanujan's conjecture followed. A clear link between modular forms and Galois representations was established.

The new methods of algebraic geometry were also needed strongly to overcome the hurdles in extending results for $GL(2)$ to other groups, like the symplectic group. The main reason for this is that the associated modular varieties are of higher dimension. Moreover, the fact that the spaces that are considered are usually not complete presents serious obstacles. Satake showed how the quotient space $Sp(2g, \mathbb{Z}) \backslash \mathbb{H}_g$ can be compactified by adding the orbits of the rational boundary components \mathbb{H}_i with $0 \leq i \leq g$ in $\overline{\mathbb{H}}_g$, thus obtaining a normal analytic space, which however for $g > 1$ is very singular. Baily and Borel generalized his construction to the so-called Baily-Borel compactification where the quotient of a bounded symmetric domain under an arithmetic subgroup is compactified to a projective variety that contains the original as a quasi-projective open subvariety. The embedding in projective space is given by modular forms of an appropriate weight. In other words, the homogeneous coordinate rings of these compactifications are the graded algebras of modular forms. These Baily-Borel compactifications are in general very singular. Igusa constructed a smooth compactification of $Sp(2g, \mathbb{Z}) \backslash \mathbb{H}_g$ for $g \leq 3$ by blowing up the Satake compactification along the ideal of the boundary. Mumford launched a big program to construct smooth compactifications by toroidal methods. A drawback of these compactifications is that they are not canonical, but depend on combinatorial data (cone decompositions).

Around the same time Hirzebruch discovered how to resolve the singularities of Hilbert modular surfaces. The singularities were resolved by cycles of rational curves. This provided a lot of information about Hilbert modular forms for real quadratic fields. In particular, it led to a geometric interpretation of the inverse of the Doi-Naganuma lifting from elliptic modular forms to Hilbert modular forms and the Fourier coefficients of the elliptic modular forms were interpreted by Hirzebruch and Zagier as the intersection numbers of modular curves on these modular surfaces.

It is interesting to note the parallel to the situation of a century earlier, when the modern theory of Riemann surfaces was brought into play in order to understand the spaces on which the modular functions live. But for a mature arithmetic theory of modular forms the full force of the newly developed algebraic geometry was needed. Here we think of Grothendieck's theory of moduli functors and their representability and Mumford's results in geometric invariant theory for the moduli spaces of abelian varieties, the compactification theory of Mumford c.s. and a version over the integers that was provided by Chai and Faltings.

Another instance where the power of algebraic geometry was brought to bear is the beautiful theorem of Gross and Zagier relating derivatives of L -functions of modular forms at the center of the critical strip to the heights of Heegner points on modular curves.

Around 1985, Frey came up with the idea and some arguments that the modularity conjecture should contradict the marvelous properties of the “ ABC -elliptic curve” over \mathbb{Q} associated by Hellegouarch to a hypothetical solution of the Fermat equation. Hence, Fermat's last theorem should be a consequence of the modularity conjecture, which therefore attracted much attention. Soon thereafter, in [16], Serre formulated a conjecture on irreducible odd continuous representations $\rho: \text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \rightarrow \text{GL}(2, \overline{\mathbb{F}}_p)$, where odd means that the determinant of complex conjugation equals -1 . The precise form of this conjecture was to make clear the “epsilon” that was needed apart from the modularity conjecture to prove Fermat's last theorem. Serre conjectured that every such ρ can be obtained from a normalised eigenform $f = \sum a(n)q^n$ of weight $k(\rho)$ on $\Gamma_1(N(\rho))$, with $k(\rho)$ and $N(\rho)$ given in terms of the ramification of ρ . The pair $(N(\rho), k(\rho))$ was intended to be the minimal possible. After a first step by Mazur, Ribet was able to establish the “epsilon”, and this motivated Wiles to set out to prove a form of the modularity conjecture that would suffice to prove Fermat's last theorem. The realization of this by Wiles in 1994, with the help of Taylor, is certainly one of the triumphs of 20th century mathematics and of the theory of modular forms.

Wiles's breakthrough, now about fifteen years ago, was based on the study of deformations of Galois representations, a theory initiated by Mazur. The most striking of his results is that completions of Hecke algebras can often be interpreted as universal deformation rings, where one considers deformations whose ramification is suitably restricted at all primes.

Since then, these deformation theoretic methods have been generalised and have led to spectacular progress. The full modularity conjecture for elliptic curves over \mathbb{Q} was proved in [6]. Here, the formulation of the restrictions on the ramification uses the local Langlands correspondence for $GL(2)$, as well as Fontaine's theory of p -adic Galois representations. Fontaine and Mazur have conjectured that all irreducible continuous p -adic representations of the Galois group of \mathbb{Q} that are unramified at almost all primes and are everywhere potentially semi-stable should come from geometry, and, according to the Langlands program, from automorphic representations. Breuil started investigating the possibility of a p -adic local Langlands correspondence for p -adic Galois representations of p -adic fields; here the question is what one should put on the automorphic side. Taylor obtained potential modularity results for two-dimensional p -adic Galois representations over totally real fields, thereby proving meromorphic continuation and functional equation for the associated L -functions; see [22] and [23], and [21] (the long version). It came as a surprise to many that these methods could be applied in the $GL(n)$ -case, when in March of 2006 Taylor, Clozel, Harris and Shepherd-Barron announced their proof of the Sato-Tate conjecture for elliptic curves over \mathbb{Q} with multiplicative reduction at at least one prime (see the preprints on Taylor's home page). Dieulefait and Wintenberger noticed that Taylor's potential modularity results made it possible to construct compatible systems of l -adic representations even in cases where modularity was not known. This led to the proof, in 2007, by Khare and Wintenberger, using important results of Kisin, of Serre's conjecture that is mentioned above (see the preprints on Khare's home page). All this is more than many had been inclined to hope.

Of course there has been progress on the subject of modular forms and the Langlands program that is independent of Wiles's breakthrough. In this respect we should mention the work of Drinfeld, who proved the global Langlands correspondence for $GL(2)$ in the function field case in the 1970's, and Lafforgue, who generalised that to $GL(n)$ in 2002. In the 1970's, Mazur did groundbreaking work concerning rational points on modular curves and their Jacobians. Harris and Taylor proved the local Langlands correspondence for $GL(n)$ over p -adic fields around 2000, using the geometry of certain Shimura varieties. And there must be much more, that we, the editors of this volume, are not aware of because of our own limited background. For example, it is clear that

in the results described above, base change results are used, trace formulas, properties of L -functions of pairs, fundamental lemmas and what not. There are important developments that we have not even mentioned, like the work of Borcherds. We hope that readers will enjoy this introduction nevertheless, and will excuse us for any omissions.

For an algebraic geometer the main lure of modular forms may come from the fact that algebraic varieties defined over a number field are a natural source for modular forms. Indeed, according to Langlands the corresponding Galois representations should all come from automorphic representations. The developments of the recent years have thus tied modular forms very closely to arithmetic algebraic geometry and this has been fruitful to both algebraic geometry and the theory of modular forms. But further progress certainly requires a better understanding of modular forms on other groups than $GL(2)$. The groups that correspond to modular varieties parametrizing algebro-geometric objects offer maybe the best hopes, as algebraic geometry may bring further clues. But even well-studied moduli spaces still seem far beyond our grasp. For example, what automorphic forms occur in the cohomology of the moduli space M_g of curves?

The goals as formulated by the Langlands conjectures may seem very distant, but recent developments as in the work of Laumon-Ngô and Ngô on the “fundamental lemma” yield the prospect of rapid advances in the near future. Apart from that modular forms appear again and again at unexpected places, for example in new developments in mathematical physics like string theory, showing that the topic is still full of life.

Bibliography

- [1] A. Ash, D. Mumford, M. Rapoport, Y. Tai: Smooth compactification of locally symmetric varieties. *Lie Groups: History, Frontiers and Applications*, Vol. IV. Math. Sci. Press, Brookline, Mass., 1975.
- [2] W. Baily, A. Borel: Compactification of arithmetic quotients of bounded symmetric domains. *Ann. of Math.* **84** (1966), pp. 442–528.
- [3] A. Borel: Introduction to automorphic forms. *Algebraic Groups and Discontinuous Subgroups (Proc. Sympos. Pure Math., Boulder, Colo., 1965)*, pp. 199–210, Amer. Math. Soc., Providence, R.I., 1966.
- [4] A. Borel, H. Jacquet: Automorphic forms and automorphic representations. With a supplement “On the notion of an automorphic representation” by R. P. Langlands. *Proc. Sympos. Pure Math., XXXIII, Automorphic forms, representations and L-functions (Proc. Sympos. Pure Math., Oregon State Univ., Corvallis, Ore., 1977)*, Part 1, pp. 189–207, Amer. Math. Soc., Providence, R.I., 1979.