An insight into the world of modular forms

Daniel Fretwell

Mathematicians love the notion of invariance. It appears in many guises throughout maths. Not only is invariance a helpful tool to have in solving problems but psychologically there is just something satisfying about knowing a given property of something can never change. In the famous Erlangen program Felix Klein is often credited with the radical idea of studying or defining geometries of spaces via their invariants. As we shall see, modular forms are a nicely behaved set of functions that have "almost invariance" under a particular matrix group action on the domain.

Modular forms have become a well used tool in number theory, but they are actually objects belonging to analysis. They provide many identities of number theoretical significance as well as being important objects in studying elliptic curves. In fact one of the steps in the proof of Fermat's Last Theorem was to move into the world of modular forms and to derive a contradiction there. In this short article I will give a brief introduction to classical modular forms.

Recall the set of complex numbers $\mathbb{C} = \{a + ib \,|\, a, b \in \mathbb{R}\}$, where i is a special element satisfying $i^2 = -1$. This set is actually a field; essentially this means that we can add, subtract, multiply and divide complex numbers and get other complex numbers (as long as we do not divide by 0). Geometrically these numbers describe a plane, meaning that complex numbers can often help to describe geometrical situations. Algebraically they are also special since the field of complex numbers is algebraically closed; this means that if we look only in $\mathbb{C}$ we are guaranteed to be able to solve polynomial equations if we allow all parameters to lie in $\mathbb{C}$. This was not possible over $\mathbb{R}$; the polynomial equation $x^2 + 1 = 0$ had no real number solutions, we added the "number" i to the pot to make this equation solvable. The importance of $\mathbb{C}$ being algebraically closed is that no more extensions need to be made; in adding i to the pot we automatically allowed all complex number polynomial equations to be solved.

An important subset of the complex numbers is the set of complex numbers with positive imaginary part, i.e. $\mathcal{H} = \{a + ib \in \mathbb{C} \,|\, b > 0\}$. These numbers form the upper half plane. Geometrically this can be described as the part of the plane lying above the x-axis.

We may define a group action of the matrix group $SL_2(\mathbb{R})$ on $\mathcal{H}$. This works as follows. Take any $\tau \in \mathcal{H}$, then a matrix $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{R})$ may act on τ by sending it to the complex number $\gamma\tau := \frac{a\tau+b}{c\tau+d}$. It is easily checked that this complex number does indeed lie in $\mathcal{H}$, so that the action is well defined. Moreover the action is transitive, meaning that given any two complex numbers $\tau_1, \tau_2 \in \mathcal{H}$, there always exists $\gamma \in SL_2(\mathbb{R})$ such that $\gamma\tau_1 = \tau_2$.

For reasons provided by number theory it makes sense to consider instead the action of $SL_2(\mathbb{Z})$ on $\mathcal{H}$. This is a more interesting group arithmetically,

relating to a notion of equivalence of lattices.

So what are modular forms? They are certain nicely behaved complex-valued functions on the upper half plane. Now what do I mean by nicely behaved? Well it would be nice to consider functions which have values invariant under the group action of the so called modular group $\Gamma = SL_2(\mathbb{Z})$ on $\mathcal{H}$. Also we would like to be able to use analysis to help, so we had better consider holomorphic functions (i.e. complex differentiable). Unfortunately these conditions are a little restrictive and all such functions have been described. We weaken the invariance property a little and settle on the following:

Definition 0.1. A *weight k modular form* for Γ is a function $f : \mathcal{H} \longrightarrow \mathbb{C}$ with the following properties:

1. f is holomorphic on $\mathcal{H}$;
2. $f(\gamma z) = (cz + d)^k f(z)$ for any $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma$;
3. f is "holomorphic at the cusps".

The third point in the above need not concern us much and is just necessary in order for nice things to follow. The important part is point 2. Here we are not demanding invariance under the action mentioned above but "weak invariance" upto a multiplier depending on a fixed integer k, which we have called the "weight" of the modular form. Having weight 0 is the same as being invariant under the action.

We can immediately notice a few things. Firstly there are no modular forms of negative weight for Γ. The weak invariance condition in these cases will contradict the holomorphy condition at the cusps. Also since $-I_2 \in \Gamma$ we see that by definition any modular form of weight k for Γ has to satisfy $f(z) = (-1)^k f(z)$ for all $z \in \mathcal{H}$. From this we deduce that there are no non-zero modular forms of odd weight for Γ.

Ok, so we have defined the concept of a modular form but how do we know one even exists? Well a big source of examples can be given by special infinite series called Eisenstein series. Given any even integer $k \geq 2$ we may construct the weight k Eisenstein series:

$$G_k(z) = \sum_{\substack{m,n \in \mathbb{Z} \\ (m,n) \neq (0,0)}} \frac{1}{(mz+n)^k}$$

This series is in some sense a "2-dimensional" version of the series defining the Riemann zeta function on $\mathrm{Re}(s) > 1$:

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}.$$

It is easily checked that each G_k satisfies the weak invariance property. Also almost all of the G_k are holomorphic everywhere (there is a problem with absolute convergence when $k = 2$). Thus for $k > 2$ we have that G_k is a modular form of weight k for Γ.

The Eisenstein series G_4 and G_6 have connections with the theory of lattices, after all the sum itself is taken over the rank 2 lattice $\mathbb{Z}^2$.

For reasons that will become clear in a minute, we often like to normalize G_k to define new series:

$$E_k(z) = \frac{G_k(z)}{2\zeta(k)}.$$

Now that we have given basic examples of modular forms, we note another interesting property of modular forms for Γ. The fact that the matrix $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ lies in Γ tells us that any modular form must satisfy:

$$f(z) = f(z+1)$$

for all $z \in \mathcal{H}$. Thus in particular f is periodic of period 1, so must have a complex Fourier series of the form:

$$f(z) = \sum_{n=-\infty}^{\infty} a_n e^{2n\pi iz},$$

for complex numbers a_n. We often substitute $q = e^{2\pi iz}$ to get a power series:

$$f(z) = \sum_{n=-\infty}^{\infty} a_n q^n.$$

This series if often referred to as the q-expansion of f. The "holomorphy at the cusps" part of the definition of modular forms translates into a property of the q-expansion, namely that $a_n = 0$ for $n < 0$ (i.e. we get a true power series and not a Laurent series, there are no singularities "at infinity"). Any modular form with $a_0 = 0$ is called a cusp form. These have a zero at the cusp, i.e. "at infinity".

It is relatively easy to get the q-expansion of G_k for $k > 2$. It turns out that it is the following:

$$G_k(z) = 2\zeta(k) + \frac{2(2\pi i)^k}{(k-1)!} \sum_{n=1}^{\infty} \sigma_{k-1}(n) q^n,$$

where $\sigma_{k-1}(n) = \sum_{d|n} d^k$ is the kth power divisor sum. This expansion gives the reason behind the normalizing factor mentioned earlier. It makes the constant term equal to 1 as well as clearing most of the coefficient of the sum.

Now it has been known for centuries how to evaluate the zeta function at even integers. The formula is as follows:

$$\zeta(2n) = (-1)^{n+1} \frac{B_{2n}(2\pi)^{2n}}{2(2n)!},$$

where B_{2n} is the $2n$-th Bernoulli number. Thus after dividing by the normalizing factor we get the q-expansion for E_k:

$$E_k(z) = 1 - \frac{2k}{B_k} \sum_{n=1}^{\infty} \sigma_{k-1}(n) q^n.$$

This q-expansion fills in a small amount of the picture as to why modular forms are interesting to number theorists. Here is a modular form, E_k. It is an

object of complex analysis, yet somehow encoded within this function is something number theoretical, a divisor sum. If we study more about the structure of modular forms we might be able to learn more about the number theoretic information encoded within. This is essentially the philosophy of modular forms.

What can be said about this structure then? Well the set of modular forms of any given weight k for Γ forms a vector space over $\mathbb{C}$, denoted $M_k(\Gamma)$. This, in a nutshell, means that the addition and scalar multiplication of such functions is well defined and nicely behaved. Further the set of cusp forms of weight k form a subspace of $M_k(\Gamma)$, denoted $S_k(\Gamma)$. The extremely nice thing here is that all of these spaces turn out to be finite dimensional, in other words the space of weight k modular forms can really be built by using a finite number of weight k modular forms. This is not so easy to prove but can be done using sophisticated tools in algebraic geometry.

There is a formula for the dimension of these spaces (for even $k \geq 0$):

$$\dim(M_k(\Gamma)) = \begin{cases} \left[\frac{k}{12}\right] + 1 & \text{if } k \not\equiv 2 \bmod 12 \\ \left[\frac{k}{12}\right] & \text{if } k \equiv 2 \bmod 12 \end{cases}$$

Even more can be said about the structure of these spaces. In fact there is a multiplication on modular forms that turns the entire set of modular forms (for all weights) into a structure known as a graded algebra. The details of this need not concern us but just the fact that if $f \in M_k(\Gamma)$ and $g \in M_{k'}(\Gamma)$ then $fg \in M_{k+k'}(\Gamma)$. This fact is important for being able to construct modular forms of higher weights from modular forms that we already know. It is actually possible to show that this algebra is generated by G_4 and G_6, meaning that every modular form can be written as some complex polynomial of these two Eisenstein series.

The above dimension formulae are quite a huge help. To demonstrate how they may be of use, consider the space $M_8(\Gamma)$. The dimension formula above tells us that this space is one dimensional. But we know two particular modular forms belonging to this space, E_8 and E_4^2. These modular forms must therefore be linearly dependent, i.e. $E_8 = aE_4^2$ for some $a \in \mathbb{C}$. We can now switch to q-expansions, so that we may formally compare coefficients. We see that $a = 1$ by comparison of constant terms (since the normalized Eisenstein series have constant term 1). Thus $E_8 = E_4^2$. By comparing the other coefficients we recover a strange identity (for $n > 1$):

$$\sigma_7(n) = \sigma_3(n) + 120 \sum_{k=1}^{n-1} \sigma_3(n)\sigma_3(n-k).$$

This identity tells us that the 7th power divisor sum can really be defined in terms of the 3rd power divisor sum. It would be extremely difficult to prove using elementary methods, in fact it would be tough to even imagine such an identity existing!

Other similar identities exist between power divisor sums, each being a consequence of linear dependence. This is why finite dimensionality is important. Another use of finite dimensionality produces this mysterious result, stemming back to the work of Ramanujan:

$$\tau(n) \equiv \sigma_{11}(n) \bmod 691,$$

where τ is the Ramanujan tau function, which takes its values as the coefficients of the q-expansion of the weight 12 modular form:

$$\Delta(z) = q \prod_{n \geq 1} (1 - q^n)^{24}.$$

The function Δ is called the discriminant function and also has many important connections to lattices and elliptic curves.

We find the above congruence by studying the 2-dimensional space $M_{12}(\Gamma)$ and studying certain modular forms in this space; E_{12}, E_4^3, E_6^2 and Δ itself.

This congruence is the first of many congruences between the coefficients of modular forms and tells us quite a bit of information about τ. In particular it tells us that for most n, $\tau(n) \neq 0$. This provides a huge amount of evidence for Lehmer's conjecture, that $\tau(n) \neq 0$ for all n.

The τ function was also conjectured by Ramanujan to satisfy the following properties:

$$\tau(mn) = \tau(m)\tau(n) \qquad \text{if } m, n \text{ are coprime,}$$

$$\tau(p^m) = \tau(p)\tau(p^{m-1}) - p^{11}\tau(p^{m-2}) \qquad \text{for any prime } p \text{ and integer } m > 2$$

These relations would show that the τ function is completely described by its values at primes.

His conjectures were correct and were proved by the use of Hecke operators, certain linear maps on the space of cusp forms of weight k. I will not define these operators but will explain the idea. There are a bunch of linear maps on $S_k(\Gamma)$, one for each natural number n, denoted T_n. It turns out that these linear maps commute with each other and are self adjoint with respect to a certain inner product on this space.

Why is this important? Well the spectral theory of linear operators now tells us that there must be an orthogonal basis of $S_k(\Gamma)$ consisting of eigenvectors for all of the T_n operators. This means we are guaranteed the existence of a special basis for the spaces of cusp forms, one consisting of eigenforms, modular forms that are eigenvectors for the Hecke operators. These eigenforms have special properties. It turns out that their L-series are expressible in terms of an Euler product.

What has this got to do with Ramanujan's conjectures mentioned above? Well it turns out that $\dim(S_{12}(\Gamma)) = 1$, this space being spanned by the discriminant function, mentioned earlier. It follows that this function must automatically be an eigenform for all of the Hecke operators. The eigenvalues correspond to the coefficients of the q-expansion, i.e. the values of the τ function. But the Hecke operators can be shown to behave in exactly the same way as the conjectural results from Ramanujan, giving proof for these results.

Finally I wish to mention that the study of modular forms does not end here. There are many ways in which this world can expand. Firstly we may replace the modular group $\Gamma = SL_2(\mathbb{Z})$ by certain interesting subgroups and consider functions that satisfy the modular form properties but for these smaller groups. Of course, any modular form for $SL_2(\mathbb{Z})$ is automatically a modular form for such a subgroup, but we may discover more functions that behave well for this smaller group. Most of the theory that we have discussed here for $SL_2(\mathbb{Z})$ carries over with no problem.

One such interesting group we like to consider is (for positive integer N):

$$\Gamma_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z}) \,|\, c \equiv 0 \bmod N \right\}.$$

Just as the Eisenstein series contained objects of number theoretical significance in their q-expansions, modular forms for this group contain similar coefficients of importance to number theorists. Consider the function on $\mathcal{H}$ defined by:

$$\theta(z) = \sum_{n \in \mathbb{Z}} e^{2\pi i n^2 z} = 1 + 2\sum_{n=1}^{\infty} q^{n^2}.$$

This function turns out to be a modular form of "weight $\frac{1}{2}$" for $\Gamma_0(4)$. It is clear that taking powers of θ will tell us things about sums of squares. For example the coefficient of q^m in $\theta^2(z)$ will tell us the number of ways of writing m as a sum of two squares (counting rearrangements and sign changes). Similar dimension and linear dependence arguments let us find formulae for these numbers. This is part of a whole study in modular forms on the representation numbers of arbitrary quadratic forms (of which $x^2 + y^2$ is one specific example). One can define these "theta series" for any quadratic form and discover similar results to the above.

The spaces $S_2(\Gamma_0(N))$ were also important in establishing a proof of Fermat's Last Theorem. The reason is as follows. Supposing a non-trivial solution to Fermat's equation exists, it would then be possible to create a specific elliptic curve defined over the rational numbers with strange properties. But it turns out that given any such elliptic curve we can create a weight 2 modular form for the group $\Gamma_0(N)$, where N is a special number depending on the elliptic curve. This modular form would turn out to be a cusp form, in fact it would also be an eigenform for almost all Hecke operators (there are problems with the theory for primes dividing N). This is known as the modularity conjecture. Now we can use the theory of modular forms along with a few extra theorems. Basically we shift focus into a space of modular forms that is empty, creating a contradiction. Notice again the importance of knowing the dimension of such spaces.

There are many other ways of extending the theory of modular forms. One in particular is to move up in dimension. In this way we end up considering functions on a more general upper half plane that satisfy weak invariance properties under an action of the symplectic group $Sp_{2g}(\mathbb{Z})$ (which for $g = 1$ gives $SL_2(\mathbb{Z})$ so really does extend the previous theory). These kinds of modular forms are called Siegel modular forms and have other significances in number theory. However less is known about them than classical modular forms.