12
Modular Forms and Elliptic Curves

Let E be an elliptic curve over $\mathbb{Q}$, so E is defined by an equation $y^2 = x^3 + ax + b$ with $a, b \in \mathbb{Q}$. It was recently proved by Andrew Wiles and others that all such elliptic curves are "modular", a result which provides a huge number of tools for studying elliptic curves over $\mathbb{Q}$. Two important consequences are that Fermat's Last Theorem is true, and that the conjecture of Birch and Swinnerton-Dyer about the rank of $E(\mathbb{Q})$ (see Chapter 14) involves objects that are defined.

In Section 12.1 we define modular forms, and in Section 12.2 we give a definition of what it means for an elliptic curve to be modular. Section 12.3 contains a brief discussion of how modularity of elliptic curves implies the truth of Fermat's Last Theorem.

There is a huge amount of literature about modular forms. See [Ser73, Ch. 7] for a first introduction, then look at the modern survey paper [DI95] for an excellent overview of the most important basic facts about modular forms along with a fairly complete bibliography.

12.1 Modular Forms

The complex *upper half plane* is the set

$$\mathfrak{h} = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}.$$

A *holomorphic function* $f : \mathfrak{h} \to \mathbb{C}$ is a function such that for all $z \in \mathfrak{h}$ the derivative

$$f'(z) = \lim_{h \to 0} \frac{f(z+h) - f(z)}{h}$$

exists. Let $\mathcal{H}$ denote the complex vector space of holomorphic functions on $\mathfrak{h}$.

Holomorphicity is a very strong condition because $h \in \mathbb{C}$ can approach 0 in many ways. For example, if $f(z)$ is holomorphic, then all derivatives $f^{(n)}(z)$ automatically exist, and $f(z)$ converges to its Taylor expansion in a neighborhood of any point.

Recall that $\mathrm{SL}_2(\mathbb{Z})$ denotes the group of 2×2 integer matrices with determinant 1. The linear fractional transformation induced by $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$ is

$$\gamma(z) = \frac{az + b}{cz + d}.$$

The group $\mathrm{SL}_2(\mathbb{Z})$ acts on the right on $\mathcal{H}$ by pre-composition:

$$f(z) \mapsto f(\gamma(z)).$$

The space of *holomorphic differentials on* $\mathfrak{h}$ is the complex vector space of expressions

$$\Omega = \{f(z)dz : f \text{ is a holomorphic function on } \mathfrak{h}\}.$$

There is a bijection between the holomorphic functions on $\mathfrak{h}$ and the holomorphic differentials on $\mathfrak{h}$ given by $f(z) \mapsto f(z)dz$ (the inverse is $\omega \mapsto \omega/dz$). The group $\mathrm{SL}_2(\mathbb{Z})$ acts on Ω in a different and more interesting way than it acts on $\mathcal{H}$. For $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ and $f(z)dz \in \Omega$, let

$$(f(z)dz)|_\gamma = f(\gamma(z))d(\gamma(z)).$$

Remark 12.1.1. The quotient rule from calculus and that $\det(\gamma) = 1$ imply that $d(\gamma(z)) = (cz + d)^{-2}dz$. Thus under the bijection between Ω and $\mathcal{H}$, the action of $\mathrm{SL}_2(\mathbb{Z})$ on Ω corresponds to the action of $\mathrm{SL}_2(\mathbb{Z})$ on $\mathcal{H}$ given by

$$f(z)|_\gamma = f(\gamma(z))(cz + d)^{-2}.$$

For any positive integer N, consider the subgroup

$$\Gamma_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \,:\, N \mid c \right\} \subset \mathrm{SL}_2(\mathbb{Z}).$$

Let $\Omega(\Gamma_0(N))$ be the subspace of Ω of functions f such that $f(z)dz$ is fixed by every element of $\Gamma_0(N)$. If $f(z)dz \in \Omega(\Gamma_0(N))$, then since $\left(\begin{smallmatrix} 1 & h \\ 0 & 1 \end{smallmatrix}\right) \in \Gamma_0(N)$ for some integer $h > 0$ (in fact, $h = 1$), we have $f(z + h)dz = f(z)dz$, so $f(z + h) = f(z)$.

Proposition 12.1.2. *The holomorphic function $q_h(z) = e^{2\pi i z/h}$ maps the vertical strip*

$$V = \{z \in \mathfrak{h} : 0 \leq \mathrm{Re}(z) < h\}$$

bijectively onto the punctured open unit disk $D = \{z \in \mathbb{C} : 0 < |z| < 1\}$. If $f : \mathfrak{h} \to \mathbb{C}$ is a function that satisfies $f(z+h) = f(z)$, then there is a unique function $F : D \to \mathbb{C}$ such that $f(z) = F(q_h(z))$.

Proof. We will not prove that e is holomorphic, because this is a standard part of any complex analysis course. If $z = x + iy \in V$, then

$$e^{2\pi i z/h} = e^{2\pi i(x+iy)/h} = e^{-2\pi y/h} e^{2\pi i x/h}$$

is in D since $y > 0$. Every element of D is uniquely of the form $e^{-2\pi y/h} e^{2\pi i x/h}$ for $y/h > 0$ and $0 \leq x/h < 1$, so q_h is a bijection.

For $w \in D$ let $F(w) = f(q_h^{-1}(w))$, where $q_h^{-1} : D \to V$ is the inverse of q_h. Then for $z \in V$, we have $F(q_h(z)) = f(q_h^{-1}(q_h(z))) = f(z)$ as required. Since $f(z) = f(z+1)$, we have $F(q_h(z)) = f(z)$ for all $z \in \mathfrak{h}$. □

Suppose $f \in \mathcal{H}$ satisfies $f(z+h) = f(z)$ for some positive integer h. Then $f(z)$ is *holomorphic at infinity* if the function $F(q_h)$ of Proposition 12.1.2 on $D \subset \mathbb{C}$ extends to a holomorphic function at 0. If this extension (which is necessarily unique) is 0 at 0, we say that f *vanishes at infinity*. If f is holomorphic at infinity, then $F(q_h)$ has a Taylor expansion (it is a theorem in complex analysis that holomorphic functions converge to their Taylor expansions) and there is a neighborhood of infinity such that

$$f = \sum_{n=0}^{\infty} a_n q_h^n$$

for complex numbers a_n. This expansion is called the *q-expansion of $f(z)$ at infinity*.

Definition 12.1.3 (Modular Forms). The vector space of *modular forms (of weight 2) for* $\Gamma_0(N)$ is the subspace $M_2(\Gamma_0(N))$ of $\mathcal{H}$ of holomorphic function $f : \mathfrak{h} \to \mathbb{C}$ such that

1. $f(z)dz \in \Omega(\Gamma_0(N))$
2. For every $\alpha \in \mathrm{SL}_2(\mathbb{Z})$, the *function* $(f(z)dz)|_\alpha/dz$ is holomorphic at infinity.

It takes some work to see that the second condition in the definition makes sense.

Lemma 12.1.4. *If $\alpha \in \mathrm{SL}_2(\mathbb{Z})$ and $\omega \in \Omega(\Gamma_0(N))$ then $\omega|_\alpha$ is fixed by $\alpha^{-1}\gamma\alpha$ for any $\gamma \in \Gamma_0(N)$.*

Proof. We have
$$(\omega|_\alpha)|_{\alpha^{-1}\gamma\alpha} = \omega|_{\gamma\alpha} = \omega|_\alpha.$$
□

If $\alpha = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ then there exists h such that if $t_h = \left(\begin{smallmatrix} 1 & h \\ 0 & 1 \end{smallmatrix}\right)$, then
$$\gamma = \alpha t_h \alpha^{-1} = \begin{pmatrix} -ach+1 & ha^2 \\ -hc^2 & hac+1 \end{pmatrix} \in \Gamma_0(N).$$
Thus $t_h = \alpha^{-1}\gamma\alpha$ with $\gamma \in \Gamma_0(N)$, so by Lemma 12.1.3, $(\omega|_\alpha)|_{t_h} = \omega|_\alpha$ so $g(z) = \omega|_\alpha/dz$ satisfies $g(z+h) = g(z)$. Thus g has a Fourier expansion at ∞ and condition 2 makes sense.

Remark 12.1.5. It is unfortunate that modular forms are actually functions instead of differential forms, but this is standard in the literature.

Definition 12.1.6. The subspace $S_2(\Gamma_0(N))$ of *cusp forms* is the subspace of elements $f \in M_2(\Gamma_0(N))$ such that the function $(f(z)dz)|_\alpha/dz$ vanishes at infinity for all $\alpha \in \mathrm{SL}_2(\mathbb{Z})$.

The cusp forms correspond to the differentials that are holomorphic even at the points at infinity, in the following sense. Letting $q = e^{2\pi i z}$, we have $\frac{dq}{q} = dz$, so if $f(q) = \sum_{n=0}^{\infty} a_n q^n$, then the differential
$$f(z)dz = f(q)\frac{dq}{q} = \left(\frac{a_0}{q} + a_1 + a_2 q + a_3 q^2 + \cdots\right) dq$$
is "holomorphic at infinity" if and only if $a_0 = 0$.

Remark 12.1.7. The condition that $(f(z)dz)|_\alpha/dz$ have a nice property at infinity for all $\alpha \in \mathrm{SL}_2(\mathbb{Z})$ probably seems ad hoc. It is motivated by the following geometric observation. The quotient of $\mathfrak{h}$ by the action of $\Gamma_0(N)$ is a noncompact Riemann surface $Y_0(N)$ (it is missing a finite set set of points). Elements of $\Omega(\Gamma_0(N))$ correspond to differentials on $Y_0(N)$. Differentials on noncompact Riemann surfaces are not well behaved; for example, the space of holomorphic differentials will not be finite dimensional. The differentials on $Y_0(N)$ that extend to holomorphic differentials on the compactification $X_0(N)$ are exactly the elements of $S_2(\Gamma_0(N))$. This is a finite dimensional space with dimension equal to the genus (number of holes) of the Riemann surface $X_0(N)$.

12.1.1 Examples

The following theorem is proved using techniques from algebraic geometry:

Theorem 12.1.8. *The complex vector space $S_2(\Gamma_0(N))$ has finite dimension:*
$$\dim S_2(\Gamma_0(N)) = 1 + \frac{\mu}{12} - \frac{\nu_2}{4} - \frac{\nu_3}{3} - \frac{\nu_\infty}{2},$$

where

$$\begin{aligned}
\mu &= N \prod_{p \mid N} (1 + 1/p) \\
\nu_2 &= \begin{cases} 0 & \textit{if } 4 \mid N \\ \prod_{p \mid N} \left(1 + \left(\frac{-4}{p}\right)\right) & \textit{otherwise} \end{cases} \\
\nu_3 &= \begin{cases} 0 & \textit{if } 2 \mid N \textit{ or } 9 \mid N \\ \prod_{p \mid N} \left(1 + \left(\frac{-3}{p}\right)\right) & \textit{otherwise} \end{cases} \\
\nu_\infty &= \sum_{d \mid N} \varphi(\gcd(d, N/d)).
\end{aligned}$$

For example,

$$\dim_{\mathbb{C}} S_2(\Gamma_0(2)) = 1 + \frac{3}{12} - \frac{1}{4} - \frac{0}{3} - \frac{2}{2} = 0,$$

and

$$\dim_{\mathbb{C}} S_2(\Gamma_0(11)) = 1 + \frac{12}{12} - \frac{0}{4} - \frac{0}{3} - \frac{2}{2} = 1.$$

For the rest of this section, let $q(z) = e^{2\pi i z}$. The following basis were computed using algorithms implemented in [BCP97] that are beyond the scope of this book.

Example 12.1.9. The vector space $M_2(\Gamma_0(11))$ has basis

$$\begin{aligned}
f_1 &= 5 + 12q + 36q^2 + 48q^3 + 84q^4 + 72q^5 + 144q^6 + 96q^7 + \cdots \\
f_2 &= q - 2q^2 - q^3 + 2q^4 + q^5 + 2q^6 - 2q^7 + \cdots .
\end{aligned}$$

and the subspace $S_2(\Gamma_0(11))$ of cusp forms has basis f_2.

The smallest N such that $S_2(\Gamma_0(N))$ has dimension bigger than 1 is $N = 22$. A basis for this space is

$$\begin{aligned}
f_1 &= q - q^3 - 2q^4 + q^5 - 2q^7 + \cdots , \\
f_2 &= q^2 - 2q^4 - q^6 + \cdots ,
\end{aligned}$$

Example 12.1.10. The space $S_2(\Gamma_0(43))$ has dimension 3 and basis

$$\begin{aligned}
f_1 &= q + 2q^5 - 2q^6 - 2q^7 + \cdots , \\
f_2 &= q^2 + q^3 - q^4 + 3q^5 - 3q^6 - q^7 + \cdots , \\
f_3 &= 2q^3 - q^4 + 4q^5 - 3q^6 - 2q^7 + \cdots
\end{aligned}$$

12.2 Modular Elliptic Curves

Let E be an elliptic curve defined by a Weierstrass equation $y^2 = x^3 + ax + b$ with $a, b \in \mathbb{Z}$. (If $a, b \in \mathbb{Q}$, then the equation can be transformed into one with $a, b \in \mathbb{Z}$; see Exercise 2.) For each prime $p \nmid \Delta = -16(4a^3 + 27b^2)$, set

$$a_p = p + 1 - \#E(\mathbb{Z}/p\mathbb{Z}).$$

Definition 12.2.1 (Modular). Let $N = |\Delta|$ be the absolute value of the discriminant of $y^2 = x^3 + ax + b$ (with $a, b \in \mathbb{Z}$). Then the elliptic curve E defined by $y^2 = x^3 + ax + b$ is *modular* if there exists a cuspidal modular form

$$f(z) = \sum_{n=1}^{\infty} b_n q^n \in S_2(\Gamma_0(N))$$

such that $b_p = a_p$ for all $p \nmid \Delta$.

At first glance, modularity appears to be a bizarre and unlikely property for an elliptic curve to have. Yutaka Taniyama and Goro Shimura first suggested in 1955 that every elliptic curve is modular, but mathematicians were initially dubious. Andre Weil later gave significant theoretical evidence for the conjecture. Motivated by a deep connection between this conjecture and Fermat's last theorem, Andrew Wiles proved enough of the conjecture to deduce Fermat's last theorem. A full proof of the conjecture was finally completed in 1999, and it is one of the crowning achievements of number theory.

Theorem 12.2.2 (Breuil, Conrad, Diamond, Taylor, Wiles). *Every elliptic curve over $\mathbb{Q}$ is modular.*

Wiles