BOOK REVIEWS

A first course in modular forms, by Fred Diamond and Jerry Shurman, Graduate
  Texts in Mathematics, vol. 228, Springer-Verlag, New York, 2005, xvi + 436 pp.,
  US$69.95, ISBN 0-387-23229-X

It is fair to say that the development of algebraic number theory over the past
thirty years has been profoundly influenced by attempts to understand the arith-
metic of modular forms. One of the beauties of the subject is the immense breadth
of mathematics that has been applied in this pursuit. Even the very definition of
a modular form can take a different shape whether seen from the perspective of a
Klein, Hardy, Eichler or Katz. (The novice might be forgiven for imagining number
theorists as blind men studying an elephant: "It's a complex analytic function,"
says one. "It's a function on lattices," says another. "No, it's a rule on a moduli
space," says the third.) In 1972, when the Antwerp Summer School on modular
functions was held (on "a central topic intermediate between all the then extant
math fields")[1] it was possible at a single conference to broadly sketch the entire
theory of modular forms in one variable. Nowadays, partly as a result of the efforts
of that conference, the area has blossomed in several distinct (if not disparate) di-
rections. Consequently any book on modular forms must have a definite viewpoint
in mind to avoid becoming an extended survey article. The authors take as their
inspiration the spectacular work of Wiles towards the Taniyama–Shimura conjec-
ture (naturally enough, as the first author is partly responsible for completing the
final cases of this fundamental result).

A *modular form* of *weight k* is a holomorphic function f on the upper half plane
$\mathcal{H} = \{z | z \in \mathbf{C}, \mathrm{Im}(z) > 0\}$ satisfying the following two conditions:

(1) The function f satisfies

$$f\left(\frac{az+b}{cz+d}\right) = (cz+d)^k f(z) \quad \text{for all elements} \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbf{Z}).$$

(2) The function $f(z)$ is bounded as z approaches $i\infty$.

This definition, familiar to many, is supremely unenlightening at first glance. To
understand it there are two natural questions that should be discussed: what is the
nature of the action of $\mathrm{SL}_2(\mathbf{Z})$ on $\mathcal{H}$, and what is the reason for the $(cz+d)^k$ term
(the so-called *automorphy factor*). To answer these questions requires a discussion
of lattices in $\mathbf{C}$.

---

[1] According to the organizer Willem Kuijk (see [1]).

A *lattice* $\Lambda \subset \mathbf{C}$ is a free abelian group generated by two complex numbers ω_1 and ω_2 such that ω_1 and ω_2 are linearly independent over $\mathbf{R}$ (equivalently, two complex numbers that generate $\mathbf{C}$ as an $\mathbf{R}$-vector space). The lattice Λ does not uniquely determine the pair $[\omega_1, \omega_2]$, but given such a pair, one may consider the ratio $z = \omega_1/\omega_2$. After reordering if necessary, we may assume that $\mathrm{Im}(z) > 0$, and so $z \in \mathcal{H}$. The extent to which z is determined by Λ is the extent to which Λ determines a choice of basis up to sign. As a $\mathbf{Z}$-module the automorphism group of Λ is simply $\mathrm{GL}_2(\mathbf{Z})$. Since we are insisting that the ratio ω_1/ω_2 has positive imaginary real part, one should take the corresponding index two subgroup $\mathrm{SL}_2(\mathbf{Z})$. In other words, we have a map

$$\text{Lattices} \longrightarrow \mathcal{H}/\mathrm{SL}_2(\mathbf{Z}).$$

This map is not injective, since the image depends only on the ratio ω_1/ω_2. If Λ and Λ' are two lattices such that $[\omega_1, \omega_2] = [\lambda\omega_1', \lambda\omega_2']$, then we say that Λ and Λ' are *homothetic*. Since any lattice is homothetic to one of the form $[z, 1]$ for some $z \in \mathcal{H}$, it follows that there is a natural bijection of sets

$$\text{Lattices up to homothety} \longleftrightarrow \mathcal{H}/\mathrm{SL}_2(\mathbf{Z}).$$

A modular form of weight zero is therefore a function on lattices up to homothety. It turns out that holomorphicity of f together with condition (2) (boundedness at infinity) force all modular forms of weight zero to be constant. There are, however, many interesting functions on lattices up to homothety that correspond to *meromorphic* functions on $\mathcal{H}$; we call such objects *modular functions*. If one is considering functions f on lattices it is natural to relax the condition that f depend only on the homothety type of Λ, and simply ask that f transform via a simple rule under scaling by $\lambda \in \mathbf{C}$. For example, one could consider functions f on lattices such that

$$f([\lambda\omega_1, \lambda\omega_2]) = \lambda^{-k} f([\omega_1, \omega_2]).$$

The corresponding function on $\mathcal{H}$ given by $f([z, 1])$ is no longer invariant under $\mathrm{SL}_2(\mathbf{Z})$, but rather transforms exactly as we defined modular forms of weight k to transform, and so gives a more intrinsic explanation of what modular forms are. This definition also allows us to write some explicit examples. Let

$$G_k(\Lambda) = \sum_{\gamma \in \Lambda'} \frac{1}{\gamma^k},$$

where Λ' denotes $\Lambda \setminus \{0\}$. This sum is absolutely convergent for $k > 2$, and not zero (for all Λ) for even k. Moreover, $G_k(\lambda \cdot \Lambda) = \lambda^{-k} G_k(\Lambda)$, and one finds (after checking condition (2)) that for even $k > 2$ the function $G_k(z) := G_k([z, 1])$ is a modular form of weight k. Modular forms of any weight are invariant under the element

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

which sends z to $z + 1$, and thus they can be expanded in terms of Fourier series:

$$f(z) = \sum_{-\infty}^{\infty} a_n q^n, \qquad q = \exp(2\pi i z).$$

The boundedness condition (2) now becomes the condition that $a_n = 0$ for all $n < 0$. For example $G_k(z)$ (for $k > 2$ even) can be written as

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

where $\sigma_k(n)$ is the sum $\sum_{d|n} d^k$ of k-th powers of positive divisors of n, and $\zeta(k)$ is the Riemann-Zeta function evaluated at k. One may write other such functions. For example

$$\sum_{n=1}^{\infty} \tau(n)q^n := \Delta(z) = q\prod_{n=1}^{\infty}(1-q^n)^{24}$$
$$= q - 24q^2 + 252q^3 - 1472q^4 + 4830q^5 - 6048q^6 + \dots$$

turns out to be a modular form of weight 12. Moreover, the coefficients $\tau(n)$ satisfy the remarkable condition (noted by Ramanujan and proved by Mordell) that $\tau(ab) = \tau(a)\tau(b)$ for $(a,b) = 1$, and $\tau(p^n) = \tau(p)\tau(p^{n-1}) - p^{11}\tau(p^{n-2})$ for all primes p. Hecke realized that Mordell's argument could be vastly generalized— in general, modular forms admit the action of certain "Hecke operators" T_n for $n \in \mathbf{N}$—that spaces of modular forms decompose into eigenforms under this action and that all modular forms could be written as linear combinations of so-called *Hecke eigenforms* whose coefficients were multiplicative for $(a,b) = 1$ and satisfied a relation similar to τ for prime powers.

A third and perhaps the most significant way of thinking about modular forms involves the moduli of elliptic curves. Given a lattice Λ in $\mathbf{C}$ we may form the quotient $E = \mathbf{C}/\Lambda$. The natural complex structure on $\mathbf{C}$ makes E a smooth Riemann surface of genus one, and given Λ one may canonically write a Weierstrass equation

$$y^2 = 4x^3 - g_2 x - g_3$$

defining E. Moreover, there is a natural group structure on E coming from the obvious group structure on $\mathbf{C}$. Since all Riemann surfaces are algebraizable, E can also be thought of as an algebraic curve of genus one together with a distinguished point, the "origin" of the group law. We should note that from this description it is not at all obvious that the group law on E is also algebraically defined, although this is true. The correspondence between elliptic curves and lattices is particularly clean: two lattices define isomorphic elliptic curves if and only if the lattices are homothetic. Thus we obtain a natural bijection

$$\text{Elliptic curves over } \mathbf{C} \text{ up to isomorphism} \longleftrightarrow \mathcal{H}/\mathrm{SL}_2(\mathbf{Z}).$$

It is also important to consider the RHS of this correspondence geometrically. The quotient $\mathcal{H}/\mathrm{SL}_2(\mathbf{Z})$ is (topologically) a punctured sphere. Completing at $i\infty$ gives a sphere that can be given a natural complex structure. We understand the bijection above as saying that the (open) complex curve $\mathcal{H}/\mathrm{SL}_2(\mathbf{Z})$ is a coarse moduli space for elliptic curves (over $\mathbf{C}$) up to isomorphism. When considering coarse moduli spaces, it is natural to rigidify the problem somewhat by considering auxiliary data. For example, one could consider elliptic curves $\mathbf{C}/\Lambda$ along with a cyclic subgroup H of order N. Now two pairs (E, H) and (E', H') are isomorphic if and only if there is an isomorphism $E \simeq E'$ *sending H to H'*. Determining the automorphism

group of a lattice together with a fixed subgroup is an elementary variation on the analysis above, and we find that there is a correspondence

$$\begin{array}{c}\text{Elliptic curves over } \mathbf{C} \text{ up to isomorphism} \\ \text{together with } H \subset E \text{ cyclic of order } N\end{array} \quad \longleftrightarrow \mathcal{H}/\Gamma_0(N),$$

where $\Gamma_0(N) \subset \mathrm{SL}_2(\mathbf{Z})$ is the group

$$\Gamma_0(N) = \left\{ \gamma \in \mathrm{SL}_2(\mathbf{Z}) \,|\, c \equiv 0 \mod N \right\}.$$

Once more this gives an identification of coarse moduli spaces. The right-hand side as an algebraic curve is usually denoted by $Y_0(N)$, and its completion by $X_0(N)$. One defines modular forms for $\Gamma_0(N)$ as holomorphic functions on $\mathcal{H}$ satisfying (1) for all $\gamma \in \Gamma_0(N)$, and satisfying the technical extension of (2) (one needs boundedness at the "cusps", the finite set of points $X_0(N) \setminus Y_0(N)$). The theory of Fourier expansions and Hecke operators applies in this context as well, with some technical modifications at primes dividing N (see [2]). We also have another interpretation of a modular form: it is a function of an elliptic curve *together* with a Weierstrass equation. The elliptic curve depends only on the homothety type of the corresponding lattice, but a Weierstrass equation is affected by a scaling of this lattice and so "remembers" this extra information. For example, the function $\Delta(z)$ described above is equal to the discriminant of the Weierstrass equation corresponding to z, and it follows that Δ does not vanish for any $z \in \mathcal{H}$ (although it does vanish at the cusp of $X_0(1)$).

Once we have the curves $X_0(N)$, there is a natural interpretation of modular forms of weight two. Consider the differential dz on the upper half plane $\mathcal{H}$. It is not invariant under $\mathrm{SL}_2(\mathbf{Z})$ or $\Gamma_0(N)$, but a simple computation shows that

$$d\gamma z = d\left(\frac{az+b}{cz+d} \right) = \frac{1}{(cz+d)^2} \cdot dz.$$

Thus if f is a modular form of weight two for $\Gamma_0(N)$, then $\omega = f\,dz$ is *invariant* under the action of $\Gamma_0(N)$, and ω descends to a holomorphic differential on $Y_0(N)$. The condition of boundedness at the cusps ensures that ω has at worst logarithmic singularities at the cusps, and if we know *a priori* that f vanishes at the cusps (a so-called *cusp form*), then ω extends to a holomorphic differential over the entire curve $X_0(N)$. The disparity between vanishing and holomorphicity comes from the fact that the differential dz has a pole at $i\infty$. If Ω^1 is the usual sheaf of differentials on $X_0(N)$, then this gives an isomorphism

$$H^0(X_0(N), \Omega^1) \simeq \text{cuspidal Modular forms of weight 2 for } \Gamma_0(N) =: S_2(\Gamma_0(N)).$$

We shall consider what happens when $N = 50$, an example nicely explained by Birch in [3]. The curve $X_0(50)$ has genus two, and the functions

$$u = q^{-1} \prod_{n=1}^{\infty} \frac{(1-q^{25n})(1-q^{2n})}{(1-q^n)(1-q^{50n})} = q^{-1} + 1 + q + 2q^2 + 2q^3 + \dots,$$

$$v = q^{-3} \prod_{n=1}^{\infty} \frac{(1-q^{25n})(1-q^{2n})^2}{(1-q^n)(1-q^{50n})^2} = q^{-3} + q^{-2} + 1 + q^3 + q^7 + \dots$$

are both meromorphic functions on this curve,[2] and moreover they generate the function field of $X_0(50)$. Naturally, as $X_0(50)$ is a curve, they must satisfy some

---

[2]That u^{12} and v^{12} are functions on $X_0(N)$ follows directly from the fact that $\Delta(z)$ is modular. To deduce that u and v are themselves meromorphic requires an analysis on the poles and zeros

polynomial relation, which turns out to be

$$v + 5u^3/v = u^3 - 2u^2 - 2u + 1.$$

This provides an explicit uniformization of $X_0(50)$. Now the curve $X_0(50)$ admits an involution W_2 sending u to $-1/u$ and v to v/u. Thus as $s = v(u+1)/u^2$ and $t = (u+1)^2/u$ generate the invariants of $\mathbf{C}(u, v)$ under W_2, they define the function field of the quotient curve, which is

$$s^2 - t(t-5)s + 5t = 0.$$

This is an elliptic curve, and further fiddling with the variables $x = -5/t$ and $y = 5(2s - t(t-5))/t^2$ leads to a more familiar looking equation,

$$E : y^2 = 4x^3 + 25x^2 + 50x + 25,$$

where

$$x = -5q + 15q^2 - 35q^3 + 80q^4 - 175q^5 + 380q^6 + \ldots,$$
$$y = 5 - 25q + 75q^2 - 225q^3 + 600q^4 - 1525q^5 + 3700q^6 + \ldots.$$

The curve E is an elliptic curve with the property that it is *uniformized by modular functions*; by definition we call such a curve *modular*. From the perspective of the particular algebraic curve E, this seems a particularly serendipitous property to have. Examples of such curves were already known to Klein, who found an explicit parameterization of the modular curve $X_0(11)$ of genus one. It is hard to imagine from this calculation that *all* elliptic curves over $\mathbf{Q}$ have this property. On the other hand, it is not clear what advantage is conferred on the elliptic curve E for being modular; the definition is purely analytic and seems to relate to modular functions rather than modular forms. First we show how to construct a canonical weight two modular form from such a uniformization. Elliptic curves have a unique holomorphic differential ω_E up to scalar, and thus given a map

$$\pi : X_0(N) \to E,$$

we can pull back ω_E to $H^0(X_0(N), \Omega^1)$ to obtain a cusp form of weight two. In the case of $N = 50$ above, we may choose

$$\omega_E = -\frac{dx}{y}$$

and then

$$2\pi i f dz := \pi^* \omega_E = -\frac{dx}{y} = \left(q - q^2 + q^3 + q^4 - q^6 + 2q^7 - q^8 - 2q^9 + \ldots\right)\frac{dq}{q}.$$

(Recall that $q = e^{2\pi i z}$ and so $dq/q = 2\pi i dz$.) Thus f is a cusp form of level $\Gamma_0(50)$ and weight two. Crucially, however, it not only has integral coefficients, but it is a Hecke eigenform. One of Shimura's fundamental results is that this is no accident, and that given a cusp form $f \in S_2(\Gamma_0(N))$ with integral coefficients that is a Hecke eigenform, one can construct a surjective map

$$X_0(N) \to E_f$$

to some elliptic curve E_f that depends only on f. The construction is as follows. Given a cusp form f, one has a corresponding differential ω_f. The curve $X_0(N)$ comes along with a fixed cusp given by the image of $i\infty$. There is a map $X_0(N) \to$

---

of u^{12} and v^{12}. Since Δ only has a zero at the cusp, one need check only the poles and cusps of $X_0(50)$, which has 12 cusps.

$\mathbf{C}$ defined by taking a point P and integrating ω_f from $i\infty$ to P. This map is well defined only up to the periods of ω_f, which, if $X_0(N)$ has genus g, is a $2g$-dimensional $\mathbf{Z}$ module, which we call Λ. If $g = 1$, then Λ will be a lattice and this map will be exactly the Abel–Jacobi map from $X_0(N)$ to its Jacobian (which is isomorphic to $X_0(N)$ in the case of genus one). When the genus of $X_0(N)$ is greater than one, however, the space $\mathbf{C}/\Lambda$ will in general be non-Hausdorff, and thus not even a Riemann surface. Yet when f is a Hecke eigenform with coefficients over $\mathbf{Q}$, Λ miraculously turns out to be a lattice and $E = \mathbf{C}/\Lambda$ is an elliptic curve. This is the resulting map $X_0(N) \to E$. A main feature of the construction of Shimura, of course, is that this *a priori* complex construction can be done algebraically over $\mathbf{Q}$, and the resulting elliptic curve E_f is an elliptic curve over $\mathbf{Q}$. The required algebraic background is essential for anyone in the field; one must understand all the concepts we have considered so far (the moduli space of elliptic curves, Hecke operators, etc.) on an intrinsically algebraic level. It is a key feature of this book that many of the details of this theorem are included. There is an important consequence of Shimura's construction, which describes the more usual definition of modularity, namely that the coefficients of f record arithmetic properties of the elliptic curve E_f. For example, suppose that p is a prime that does not divide N. Then the elliptic curve E_f turns out to have good reduction at the prime p, and it makes sense to consider it over the finite field $\mathbf{F}_p$. One can then count the number of points over this finite field, and one finds the remarkable relation

$$a_p = 1 + p - \#E_f(\mathbf{F}_p), \quad \text{where} \quad f = \sum_{n=1}^{\infty} a_n q^n.$$

This relation for all p not dividing N is the more familiar way in which modularity of elliptic curves is defined.

The book of Diamond and Shurman is centered around explaining the various concepts of modularity and proving the so-called "easy direction" (the theorem of Shimura alluded to above). The first half of the book is devoted to the necessary background, in particular the analytic theory of modular forms. The second half works towards a proof of the Eichler–Shimura relation, which in many ways is the raison d'être of the text. There are many fine books on closely related subjects, such as Silverman's books on elliptic curves, [7], [8]. Knapp's book [5] even has a nice description of Shimura's construction over $\mathbf{C}$. Yet this is the first introductory book that tackles the Eichler–Shimura relation and tries to make it as accessible as possible, and for that it should prove useful. Note that the authors are not afraid in the latter stages of the book to suppress some of the algebraic-geometric details (or at least assign them to references), undoubtedly a necessary step to keep their goal of being accessible to students "without previous background in algebraic number theory and algebraic geometry." Looming everywhere, of course, is the omnipresence of Wiles' theorem. Readers should be aware of the excellent volume [4] from the Boston University conference on Fermat's last theorem, which is a wonderful advanced introduction to Wiles' arguments. The current book could be seen as a "prequel" to this volume, and the final chapters of the former dovetail nicely into some of the early sections of the latter (particularly Rohrlich's chapter). Last, we must mention Shimura's "Introduction to the arithmetic theory of automorphic functions" [6]. With today's iPod generation more likely to study elliptic curves and modular forms before learning any class field theory, Shimura's book by itself is

no longer apposite as an introduction to modular forms. Nonetheless, it remains an invaluable source of ideas and perspectives. More recent works such as the current book contrast and complement [6] rather than replace it.

References

[1] `http://modular.fas.harvard.edu/antwerp_photo/`
[2] A.O.L. Atkin and J. Lehner, *Hecke operators on* $\Gamma_0(m)$, Math. Ann. **185** 1970 134–160. MR0268123 (42:3022)
[3] B. Birch *Some calculations of modular relations*, Modular functions of one variable, I (Proc. Internat. Summer School, Univ. Antwerp, 1972), pp. 175–186. Lecture Notes in Mathematics, Vol. 320, Springer, Berlin, 1973. MR0332658 (48:10984)
[4] *Modular forms and Fermat's last theorem.* Papers from the Instructional Conference on Number Theory and Arithmetic Geometry held at Boston University, Boston, MA, August 9–18, 1995. Edited by Gary Cornell, Joseph H. Silverman and Glenn Stevens. Springer-Verlag, New York, 1997. MR1638473 (99k:11004)
[5] A. Knapp, *Elliptic curves*, Mathematical Notes, 40. Princeton University Press, Princeton, NJ, 1992. MR1193029 (93j:11032)
[6] G. Shimura, *Introduction to the arithmetic theory of automorphic functions*, Kanã Memorial Lectures, No. 1. Publications of the Mathematical Society of Japan, No. 11. Iwanami Shoten, Publishers, Tokyo; Princeton University Press, Princeton, NJ, 1971. MR0314766 (47:3318)
[7] J. Silverman, *The arithmetic of elliptic curves*, Graduate Texts in Mathematics, 106. Springer-Verlag, New York, 1986. MR0817210 (87g:11070)
[8] J. Silverman, *Advanced topics in the arithmetic of elliptic curves*, Graduate Texts in Mathematics, 151. Springer-Verlag, New York, 1994. MR1312368 (96b:11074)

Frank Calegari

Harvard University

E-mail address: `fcale@math.harvard.edu`