# Modular Forms and Moonshine
# A Brief introduction

By David Padvorac

Abstract

Modular forms arise in primarily in algebra and number theory; as is well known, algebra and number theory are intricately interwoven. It is well known that algebra and number theory arise often in the natural world; likewise for modular forms, whose base is in both algebra and number theory. Elliptic curves arise in both algebra and number theory, and are tightly connected with modular forms. My goal will be to develop most of the necessary content related to modular forms to sketch the motivating observation behind the Monstrous Moonshine conjecture. The actual conjecture involves module theory and representation theory, which are both rather out of the scope of this project.

# Introduction

As noted in the abstract, the end motivation here, in simple, is to develop the definitions needed to actually understand the very peculiar key observation that brought about "moonshine" theory, a recent development interweaving number theory, algebra, and to a lesser extent, physics. By taking this admittedly very shallow approach to introducing modular forms, it is sure that a large portion of their depth will be lost; as an undergraduate research project, this is nearly bound to happen regardless, as many needed ideas are in the toolboxes of graduate algebra. That is the downside of what I'll be presenting here. The upshot is that it will be an undergraduate-accessible introduction to these topics, and a small glimpse into the motivation of the Monstrous Moonshine theory and conjectures.

# Lattices

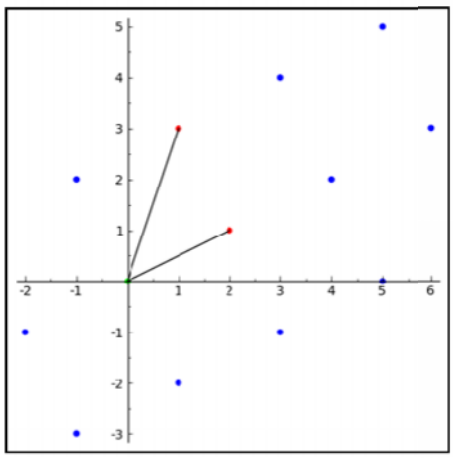
The lattice is a tremendously important tool; for starters, modular forms can be viewed as analytic functions on lattices in the complex plane. This will give us a simple definition to begin looking at modular forms with. A lattice $\lambda$ in the complex plane is completely defined by two points in the plane with a non-real ratio:

$$\omega_1, \omega_2, \tau \in C; \frac{\omega_1}{\omega_2} = \tau \notin H$$

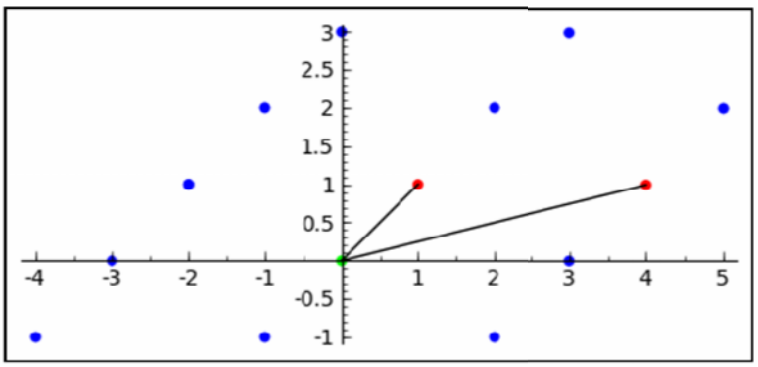
$$\lambda = L(\omega_1, \omega_2) = \{a\omega_1 + b\omega_2 : a, b \in Z\}$$

where $H$ is the upper half plane, $C$ is the complex plane, and in a shocking twist of events, we call the integers $Z$. This pair of points, the basis of the lattice, can be thought of as a pair of vectors in $R^2$. If the ratio of the two points is real, than equivalently the pair of vectors in $R^2$ is collinear, and will generate a boring lattice. Below are several examples of lattices, $L(1 + i, 4 + i)$ and $L(2 + i, 1 + 3i)$, plotted in Sage. The two generating points, $\omega_1$ and $\omega_2$ are red and the other lattice points are blue.

$L(2+i, 1+3i)$



$L(1+i, 4+i)$



It is worth noting that if one of the basis points is in the lower half plane, it can be "shifted" to the upper half plane by an addition of an integer multiple of the other basis point, and doing this yields a basis that generates the same lattice. For simplicity's sake, we'll only use bases with both points in the upper half plane. Additionally, the lattice base does not have to be integer; integer bases were chosen for ease of visualization/calculation.

A function $G: \lambda \rightarrow C$ is simply a complex function restricted to the given set of lattice points; for example, taking the trivial lattice of Gaussian integers, $\lambda = L(1, i)$, $G(\lambda)$ is simply the range of $G$ when evaluated at integer points in $C$. An important concept is the homogeneous function:

$$G: \lambda \rightarrow C \text{ is homogeneous of degree –k, if:}$$
$$G(a\lambda) = a^{-k}G(\lambda);\ \ k \in Z, a \in C$$

Examples:

1. $k = 0$: $G(a\lambda) = G(\lambda)$
2. $k = 1$: $G(a\lambda) = \frac{1}{a}G(\lambda)$
3. $k = 2$: $G(a\lambda) = \frac{1}{a^2}G(\lambda)$

It is worth noting that in these examples, that by intuition there seems to be a periodicity of some sort. Unfortunately a straight-up "normal" periodicity, as observed with sine or cosine, or any other similar function, only appears in a special case where $k = 0$; the other cases are this peculiar "scaling" sort of periodicity.

The set of homogenous functions of degree –k is in direct bijective correspondence with the set of modular forms of weight k; the proof is in Lang's text, and will not be copied here. This simply gives an additional context for modular forms to be viewed in, which plays a key role later in the more detailed development of modular forms omitted in this paper.

## The Modular Group

The modular group is a name for the group of invertible $n\ x\ n$ integer matrices, otherwise known as $SL_2(Z)$. Each element of the modular group is assigned to a function, a Linear Fractional Transformation, and operates on the upper half plane as follows:

$$\tau \in H\ ;\ \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \sigma \in SL_2(Z)$$
$$\sigma(\tau) = \frac{a\tau + b}{c\tau + d}$$

The proof that $\tau \in H\ \Rightarrow\ \sigma(\tau) \in H$, i.e. that the modular group maps the upper half plane to itself, is an elementary calculation; the only significant step is the clearing of complex numbers from the denominator of $\sigma(\tau)$, and this is done by multiplying the numerator and denominator

by the complex conjugate of the denominator. The modular group is generated by two elements, $\sigma_1$and$\sigma_2$:

$$\sigma_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix};\ \sigma_1(\tau) = \tau + 1$$
$$\sigma_2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix};\ \sigma_2(\tau) = -\frac{1}{\tau}$$

The proof of this fact is a somewhat involved calculation but does not provide much deep insight, and so is omitted.

## Some Complex Analysis Terminology

- A complex function is holomorphic if it is complex-differentiable (similar to differentiability of a function from $R^2$ to $R^2$) on some neighborhood of every point in its domain.
- Meromorphicity is a very slightly generalized version of holomorphicity. A function with isolated poles, which is holomorphic everywhere except at those poles, is called "meromorphic."

A nice intuitive geometry way to think of a meromorphic function is as the top of an old style camping tent; generally speaking, it's a nice and smooth surface except for where the tent poles poke the material up. Those points aren't too nice, but they are well isolated.

## Weak Modularity

A meromorphic function $f: H \rightarrow C$ is considered weakly modular of weight k if the following condition holds:

$$f\big(\sigma(\tau)\big)(c\tau + d)^k = f(\tau), \sigma(\tau) = \frac{a\tau+b}{c\tau+d}$$

where $\sigma$ is an element of the modular group and $\tau$ is an element of the upper half plane. Because the modular group is generated by the afore mentioned two elements $\sigma_1$ and $\sigma_2$, we can check if any function is weakly modular of some weight by checking these two inequalities, which come from plugging $\sigma_1$ and $\sigma_2$ into our definition:

1. $\sigma_1(\tau) = \tau + 1$: $\quad f(\tau) = f(\tau + 1)$
2. $\sigma_2(\tau) = -\frac{1}{\tau}$: $\quad f(\tau) = \tau^k f\left(-\frac{1}{\tau}\right)$

There are several important facts we can observe.

1. For all weights:
    a. $f(\tau)$ is $Z$-periodic, i.e. it repeats itself on a unit interval: $f(\tau) = f(\tau + n)$ for any integer $n$. This can be proven by a trivial application of induction to (1).

   b. The zeroes and poles of $f(\tau)$ are invariant under operation by $SL_2(Z)$ on $\tau$. This follows by comparing the right and left hand sides of the definition above: clearly the factor $(c\tau + d)^k$ will not have zeroes or poles, so $f(\tau)$ has zeroes and/or poles if corresponding exactly to those of $f(\sigma(\tau))$.
2. For weight $k = 0$, $f(\tau)$ is invariant under the modular group, $SL_2(Z)$. This follows from setting $k = 0$ in equation (2).
3. For any odd weight $f(\tau) \equiv 0$. We can see this by setting $\sigma = -I$ in the definition. This gives us $f(\tau)(-1) = f(\tau)$, clearly only true when $f(\tau) \equiv 0$.
4. For weight $k = 2$, we get a result familiar from path integrals of complex analysis. There is a bit of legwork to do to see this result, however.
   a. First we manipulate our definition to get that $\frac{f(\sigma(\tau))}{f(\tau)} = \frac{1}{(c\tau+d)^2}$
   b. Next, we note that $\frac{d\sigma}{d\tau} = \frac{1}{(c\tau+d)^2}$; this derivative can be calculated easily using the quotient rule and the definition of $\sigma(\tau)$.
   c. Finally, equating the right hand sides of the two equations and redistributing the fractions, we have $f(\sigma(\tau))d\sigma(\tau) = f(\tau)d\tau$, and so a path integral in the complex plane is invariant under $SL_2(Z)$.

## Modular Forms

A function $f: H \to C$ is called a modular form of weight k if the following conditions satisfied:

1. $f$ is holomorphic on $H$
2. $f$ is holomorphic "at infinity"
3. $f$ is weakly modular of weight k

For a function $f(\tau)$ to be holomorphic "at infinity" simply means that it must be bounded as $Im(\tau) \to \infty$. An easy way to show this would be by showing that $\lim_{Im(\tau)\to\infty} f(\tau)$ exists and is finite. This condition is necessary for some further results on the set of all modular forms of some given weight; this set is denoted $M_k(SL_2(Z))$.

An important example of specific type of modular form is the cusp form of weight k, a modular form of weight k, with the addition that in its Fourier expansion, the first terms coefficient is zero.

$$f(\tau) = \sum_{n=1}^{\infty} a_n q^n, \quad q = e^{2\pi i\tau}$$

$f(\tau)$ a cusp form $\Rightarrow a_n = 0$

Also, from this definition we also have that a modular form is a cusp form when $\lim_{Im(\tau)\to\infty} f(\tau) = 0$. The significance of cusp forms to this work will be in defining a very

important modular form in terms of them; this function plays a key role in the question that motivated the moonshine conjectures.

## Eisenstein Series

An Eisenstein series is essentially a 2-dimensional analog of the Riemann zeta function. In good tradition (and it also happens to be the generally accepted standard definition), we first use a definition that looks just like Riemann's original definition (this definition and it's immediate corollaries are directly from Diamond and Shurman).

$$G_k(\tau) = \sum_{(c,d) \in I} \frac{1}{(c\tau + d)^k} ;$$

$$\tau \in \mathcal{H}$$
$$k \in \mathbb{Z}$$
$$I = \mathbb{Z}^2 - \{(0,0)\}$$

$G_k$ is absolutely convergent $\forall k$ and uniformly convergent on compact subsets of $\mathcal{H}$. By definition than, $G_k$ is holomorphic on $\mathcal{H}$. By applying an arbitrary element $\sigma$ of $SL_2(Z)$ to $\tau$ and using some tricky manipulation of the summation of $G_k(\sigma(\tau))$, it falls out that $G_k$ is weakly modular with weight $k$. Lastly, $G_k$ is bounded as $Im(\tau) \to \infty$, so it is indeed a modular form.

We now define two important cusp forms. They do seem entirely arbitrary, and neither their purpose nor significance readily meets the eye. This is understandable, considering that they are part of a much more complicated thing, the Weierstrass $\wp$-function: they are the 2nd and 3rd coefficients of its Laurent expansion. The $\wp$-function won't be discussed here, although it is part of Koblitz's presentation of elliptic curves, and can be studied in full depth in his text.

$$g_2(\tau) = 60G_4(\tau)$$
$$g_3(\tau) = 140G_6(\tau)$$

Now, we define the elliptic discriminant function $\Delta(\tau)$, and the modular invariant $j(\tau)$. Again, like with $g_2$ and $g_3$, a full development of the nature and contexts of these functions isn't included here. They both play key roles in the theory of elliptic curves and modular forms.

$$\Delta: \mathcal{H} \to \mathbb{C}$$
$$\Delta(\tau) = (g_2(\tau))^3 + 27(g_3(\tau))^2$$

$$j: \mathcal{H} \to \mathbb{C}$$
$$j(\tau) = 1728\frac{(g_2(\tau))^3}{\Delta(\tau)}$$

There is a trivially-proved result on modular forms that the weight of the product of two forms is the product of their weights. From this, we can note that the numerator and denominator of $j(\tau)$ both have the same weight; it follows that $j(\tau)$ is invariant under $SL_2(Z)$. Finally, writing $j(\tau)$ as a Fourier series in terms of $q = e^{2\pi i\tau}$:

$$j(\tau) = q^{-1} + 744 + 196{,}884q + 21{,}493{,}760q^2 + 864{,}299{,}970q^3 + \cdots$$

Note: the four definitions given here are taken straight from the corresponding section of the text by Diamond and Shurman.

This is the end of the good definitions I'll give here. The rest of the needed content is either from entry level algebra or upper level algebra, or just simply very inaccessible in any rigor to most typical undergraduates, myself included. For the inaccessible content, I'll give the hand-waving definitions and descriptions.

## Some Light Definitions

1. Finite Simple Groups
   Gannon very succinctly describes finite simple groups in his text; I'll summarize it here:
   "Finite simple groups are to finite groups what the primes are to integers—they are their elementary building blocks."

   All finite simple groups have been classified, this monumental work was wrapped up, give or take, in the late 90's. There are 18 infinite families (for example, cyclic groups of prime order, $\mathbb{Z}/p\mathbb{Z}$) and 26 exceptional groups. The largest of these groups is the so called "monster" group of Fischer and Griess. The order of the monster group is approximately $8 \times 10^{53}$.

2. Representation theory
   Again, we'll use Gannon's description:
   A representation of a group is an assignment of an $n \times n$ matrix $A(g)$ to each element of that group, so that the matrix product respects the group product:
   $$A(g_1)A(g_2) = A(g_1g_2)$$
   In essence, a representation of a group is a full description of its structure in the language of linear algebra. The dimension of a representation is the size $n$ of the matrices.
   An representation can be reduced in terms of subspaces of the vector space $V$ acted on by the matrices $A$. A subspace of $V$ fixed under the group action matrices $A$ is essentially a building block for the representation.

## The Observation!

And finally, we've come to the point where we can look at the observation made by John McKay, detailed in Gannon's text. Mckay first noticed the following:

$$196{,}884 \approx 196{,}883$$
$$21{,}493{,}760 \approx 21{,}296{,}876$$
$$864{,}299{,}970 \approx 842{,}609{,}326$$

and shortly came to a more precise resulting numerology:

$$196{,}884 = 196{,}883 + 1$$
$$21{,}493{,}760 = 21{,}296{,}876 + 196{,}883 + 1$$
$$864{,}299{,}970 = 842{,}609{,}326 + 21{,}296{,}876 + 2 \times 196{,}883 + 2 \times 1$$

The numbers on the right side of these equalities are the first coefficients of the modular invariant, $j(\tau)$. The numbers on the right hand side are the dimensions of smallest "building blocks" of the group representation of the Monster group.

This striking set of equalities continues, for all coefficients of the modular invariant. The conjecture this observation lead to is, again, far beyond the means of this paper. Regardless, I do believe that this seemingly-out-of-nowhere correspondence is simply another example of the relations that appear in seemingly completely unrelated fields of mathematics. There are implications of this observation in physics, more specifically string theory; even rudimentary explanations of these implications are far beyond my current grasp.

## Bibliography


Diamond, F., & Shurman, J. (2005). A First Course in Modular Forms. Springer.

Gannon, T. (2006). Moonshine Beyond the Monster. Cambridge, UK: Cambridge University Press.

Koblitz, N. (1993). Introduction to Elliptic Curves and Modular Forms. Springer.

Lang, S. (1976). An Introduction to Modular Forms. New York City, NY: Springer.