MATHEMATICS FROM EXAMPLES, SPRING 2023

INSTRUCTOR: YU-WEI FAN

Course Description. Examples in mathematics are like phenomena in physics. They play a vital role in the historical development of mathematics and are the driving force behind profound mathematical concepts and methods. Important theorems in modern mathematics often come from the understanding and research of some basic examples. The goal of this course is to provide the motivation and intuition behind abstract mathematical concepts by introducing some interesting examples.

Contents

1. Overview of the course

Lecture 1

Example 1.1. Let $x \in (0,1)\backslash\mathbb{Q}$ be an irrational number. It can be written uniquely as a continued fraction

$$x = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{\ddots}}}}$$

where $a_1, a_2, \ldots$ are positive integers.

How often does a positive integer k appear in such an expression? It turns out that for any given k, the frequency of k appearing in the continued fraction expression of x is the same for *almost every* $x \in (0,1)\backslash\mathbb{Q}$, and is given by the following formula

$$\lim_{n\to\infty} \frac{\#\{i \mid a_i = k, 1 \le i \le n\}}{n} = \frac{1}{\log 2} \log\left(\frac{(k+1)^2}{k(k+2)}\right).$$

In order to prove this, we will introduce some basic ideas of *measure theory* and *ergodic theory*.

Example 1.2. Consider the following *necklace-splitting problem*. Two thieves have stolen a precious necklace (opened, with two ends), on which there are d kinds of stones (diamonds, sapphires, rubies, etc.), an even number of each kind. The thieves do not know the values of stones of various kinds, so they want to divide the stones of each kind evenly. They would like to achieve this by as few cuts as possible. The question is, what is the minimum amount of cuts needed to divide the stones of each kind evenly?

It is not hard to show that at least d cuts is necessary: place the stones of the first kind first, then the stones of the second kind, and so on. The *necklace theorem* shows that d cuts is always sufficient. Surprisingly, all known proofs of this theorem are *topological*.

Example 1.3. Let $C \subseteq \mathbb{R}^2$ be a simple closed curve. One considers the following *Rectangular Peg Problems*.

- Does there always exist four points on C such that they form the vertices of a rectangle?
- A much harder question: Fix a rectangle R. Does there always exist four points on C such that they form the vertices of a rectangle which is similar to R?

The first question was answered positively by Vaughan in 1981, which uses some basic *topology*. The second question was also answered positively quite recently by Greene and Lobb [9]; their proof involves more advanced tools from *symplectic geometry*, which is beyond the scope of this course.

Example 1.4. Which positive integers n can be written as the sum of two squares $n = x^2 + y^2$?

To answer this question, it is natural to introduce the *ring* of *Gaussian integers* $\mathbb{Z}[i]$, since one has the factorization $x^2 + y^2 = (x + iy)(x - iy)$. The question then reduced to studying the properties of the ring $\mathbb{Z}[i]$.

Example 1.5. One can consider a more refined question: How many ways can a positive integer n be written as the sum of two (or more) squares?

The problem is closely related to the *theta function*, which is a function defined for a complex variable $\tau \in \mathbb{H}$ on the upper half plane:

$$\theta(\tau) = \sum_{n=-\infty}^{\infty} e^{2\pi i n^2 \tau} = \sum_{n=-\infty}^{\infty} q^{n^2}, \qquad \text{where} \quad q = \exp(2\pi i \tau).$$

Let us define $r_2(n)$ to be the number of ways that n can be written as the sum of two squares

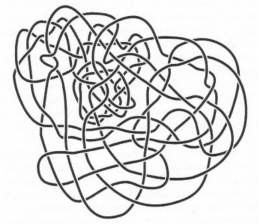
$$r_2(n) = \#\{(x, y) \in \mathbb{Z}^2 \mid x^2 + y^2 = n\}.$$

It is not hard to see that

$$\theta(\tau)^2 = \sum_{n=0}^{\infty} r_2(n) q^n.$$

The problem then reduces to understand $\theta(\tau)^2$. It turns out that $\theta(\tau)^2$ is a *modular form of weight* 1 *for the congruence subgroup* $\Gamma_1(4) \subseteq \mathrm{SL}(2, \mathbb{Z})$, and we can use the theory of modular forms to obtain an explicit formula of $r_2(n)$. In fact, the same method also applies to the sum of $2k$ square numbers for any positive integer k, where we can use modular forms to obtain explicit formula of $r_{2k}(n)$.

Example 1.6. Is the rope in the following figure knotted? Motivated by this sort of questions, we will introduce various *knot invariants* and their *categorifications*, and discuss what kinds of information are encoded by them. The construction of the categorification involves ideas including *cobordism categories* and *topological quantum field theory*, which are of independent interests.



Example 1.7. In 1696, Johann Bernoulli posed the problem of the brachistochrone (from ancient Greek, which means "shortest time") as a challenge to the mathematicians of his day: Given two points A and B in a plane, where B is lower and not directly below A, what is the curve traced out by a point acted on only by gravity, which starts from A and reaches B in the *shortest time*?

This problem is widely regarded as the founding problem of the *calculus of variations*, which study ways of finding the curve, or surface, minimizing a given integral. We will discuss the approach developed by Euler (in 1736) and Lagrange (in 1755) to deal with general problems of this kind.

Example 1.8. Let $P = (p_1, \ldots, p_n) \colon \mathbb{C}^n \to \mathbb{C}^n$ be a polynomial function, i.e. each coordinate $p_1, \ldots, p_n$ is a polynomial in $\mathbb{C}^n$. It was proved independently by Grothendieck (1966) and Ax (1968) that if P is injective then it is bijective. In fact, this theorem can be generalized to any *algebraic variety* over an *algebraically closed field*.

The method of proof is really noteworthy: it showcases the idea that *finitely many* algebraic relations in fields of characteristic 0 can be translated into algebraic relations over finite fields with large characteristics. Thus, one can use the arithmetic of finite fields to prove a statement about $\mathbb{C}$ even though there is no homomorphism from any finite field to $\mathbb{C}$. This is a great example of applications of techniques from *model theory* in *mathematical logic*.

Example 1.9. Let a and m be integers that are relatively prime. Are there infinitely many primes in the sequence

$$a, a + m, a + 2m, \ldots?$$

This was conjectured to be true by Legendre, and later proved by Dirichlet in 1837 with his *L-series*. This theorem is believed to represent the beginning of rigorous *analytic number theory*. In fact, Dirichlet shows a stronger result

that the "density" of the subset

$$\{\text{prime } p \mid p \equiv a \pmod{m}\} \subseteq \{\text{prime } p\}$$

is $1/\varphi(m)$. In other words, the prime numbers are equally distributed among different classes modulo m which are relatively prime to m.

Example 1.10. In 1657, Fermat wrote letters to his friend de Bessy, his Dutch correspondent van Schooten, and English mathematicians Wallis and Brouncker. In the letters, Fermat invited them to solve some curious mathematical problems. The central questions are concerned with certain quadratic equations of the form

$$x^2 - Ny^2 = 1, \qquad x, y \in \mathbb{Z}_{>0}.$$

To Wallis and Brouncker, he challenged them with the cases $N = 151$ and $N = 313$; but to his countryman de Bessy, he merely demanded answers for the cases $N = 61$ and $N = 109$, "so as not to give him too much trouble".

More generally, this problem can be interpreted as understanding the values of integral binary quadratic forms, such as $3x^2 + 6xy - 5y^2$. We will give a quick tour to the concept of *Conway's topograph*, with his *wells, rivers, lakes*, and *weirs*, and see how these help us with answering the problem.

Example 1.11. The dilogarithm function is defined by the power series

$$\mathrm{Li}_2(z) = \sum_{n=1}^{\infty} \frac{z^n}{n^2} \qquad \text{for } |z| < 1.$$

The definition (and the name) come from the analogy with the Taylor series of the ordinary logarithm around 1

$$-\log(1-z) = \sum_{n=1}^{\infty} \frac{z^n}{n} \qquad \text{for } |z| < 1,$$

which leads similarly to the definition of the *polylogarithm*

$$\mathrm{Li}_m(z) = \sum_{n=1}^{\infty} \frac{z^n}{n^m} \qquad \text{for } |z| < 1, \qquad m = 1, 2, \ldots.$$

The dilogarithm function is one of the simplest non-elementary functions one can imagine. It is also one of the strangest. Almost all of its appearances in mathematics, and almost all the formulas relating to it, have something of the

fantastical in them. We will discuss its relations with *hyperbolic* 3*-manifolds*, *quantum dilogarithm identity*, and *wall-crossing formula of stability conditions*.

Example 1.12. Let us consider the power series

$$\sum_{k=0}^{\infty}(-1)^k k! x^{k+1}.$$

Clearly, it is divergent for any $x \neq 0$, which makes it seems uninteresting. However, this power series and certain series of this sort, actually appears "in nature". For instance, the series could represent the solution of an ordinary differential equation, or gives the value of a physical quantity of interest, such as the energy. Many mathematicians and physicists have recently become interested in these series due to their appearance in numerous topics at the forefront of research, including: gauge theory of singular connections, quantization of symplectic and Poisson manifolds, Floer homology and Fukaya categories, knot invariants, wall-crossing and stability conditions in algebraic geometry, perturbative expansions in quantum field theory, etc.

We will discuss an approach toward making sense of this divergent issue, via the method of *Borel summation*. Along the way, we will see interesting phenomenons like *resurgence*, *Stokes phenomenon*, and relate it back to the *wall-crossing formula*.

2. Measure theory and ergodic theory

Recall our motivating question: Let $x \in (0,1)\backslash\mathbb{Q}$ be an irrational number. It can be written uniquely as a continued fraction

$$x = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{\ddots}}}}$$

where $a_1, a_2, \ldots$ are positive integers. How often does a positive integer k appear in this expression? Below is the sketch of ideas toward answering this question.

- Define the *continued fraction map* $T\colon [0,1] \to [0,1]$ by $T(0) = 0$ and

$$T(x) = \frac{1}{x} - \left\lfloor \frac{1}{x} \right\rfloor \text{ for } x \neq 0,$$

  where $\lfloor t \rfloor$ denotes the greatest integer less than or equal to t. In other words, $T(x)$ is the fractional part $\{\frac{1}{x}\}$ of $\frac{1}{x}$.

- Observe that $a_n = k$ if and only if $T^{n-1}(x) \in (\frac{1}{k+1}, \frac{1}{k}]$. Hence

$$\frac{\#\{i \mid a_i = k, 1 \le i \le n\}}{n} = \frac{1}{n}\sum_{i=0}^{n-1} \chi_{(\frac{1}{k+1},\frac{1}{k}]}(T^i(x))$$

  where χ is the characteristic function.
- Define the *Gauss measure* μ on $[0,1]$ to be

$$\mu(A) = \frac{1}{\log 2}\int_A \frac{1}{1+x}\,\mathrm{d}x \text{ for any measurable set } A \subseteq [0,1].$$

- Prove that the Gauss measure μ is *T-invariant* and *ergodic* .
- By *Birkhoff's pointwise ergodic theorem*, for almost every $x \in [0,1]\backslash\mathbb{Q}$ we have

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\chi_{(\frac{1}{k+1},\frac{1}{k}]}(T^i(x)) = \int \chi_{(\frac{1}{k+1},\frac{1}{k}]}\,\mathrm{d}\mu = \mu\left(\left(\frac{1}{k+1},\frac{1}{k}\right]\right).$$

- The conclusion then follows from a simple calculation

$$\frac{1}{\log 2}\int_{\frac{1}{k+1}}^{\frac{1}{k}} \frac{1}{1+x}\,\mathrm{d}x = \frac{1}{\log 2}\log\left(\frac{(k+1)^2}{k(k+2)}\right).$$

In order to understand this approach and appreciate the powerful tools provided by ergodic theory (in our case, the pointwise ergodic theorem), we will discuss the following topics in this section:

- basic measure theory;
- basic ergodic theory;
- ergodic theorems and applications.

Some references that might be helpful include [7] and [17].

2.1. **An outlook.** Consider a map $T\colon X \to X$. In ergodic theory, one studies how *typical* orbits $\{x, T(x), T^2(x), \ldots\}$ are distributed. We would be interested in properties like *frequencies of visits*, *equidistribution*, *mixing*, etc.

Here is a basic example. Let $A \subseteq X$ be a subset, and x be an element of X. The number of visits of orbit of x to the subset A up to time n is given by

$$\#\{0 \le k \le n-1 \mid T^k(x) \in A\}.$$

A convenient way to write this quantity is as follows. Let $\chi_A \colon X \to \mathbb{R}$ be the characteristic function of the subset A: $\chi_A(x) = 1$ if $x \in A$, and $\chi_A(x) = 0$ if $x \notin A$. Then we have

$$\sum_{k=0}^{n-1} \chi_A(T^k(x)) = \#\{0 \leq k \leq n-1 \mid T^k(x) \in A\}.$$

The *frequency* of visits up to time n is defined to be the average

$$\frac{1}{n} \sum_{k=0}^{n-1} \chi_A(T^k(x)) \in [0,1].$$

Question 2.1. *We are interested in the following questions.*

(a) Does the frequency of visits converge to a limit as n tends to infinity? (for all points of $x \in X$? or only for a typical point?)

(b) If the limit exists, what does the frequency converge to?

Another type of question concerns the equidistributioness. Let us consider specifically in the setting of the unit interval $[0,1]$. We say a sequence of points $\{x_n\}$ in $[0,1]$ is *equidistributed* if for all intervals $I \subseteq [0,1]$ we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \chi_I(x_k) = \text{length}(I).$$

An equivalent definition is for all continuous functions $f \colon [0,1] \to \mathbb{R}$ we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f(x_k) = \int_0^1 f(x)\,\mathrm{d}x.$$

So if we have a dynamical system $T \colon [0,1] \to [0,1]$ (or $T \colon S^1 \to S^1$, where $S^1 \cong \mathbb{R}/\mathbb{Z} \cong [0,1]/0 \sim 1$), we can ask whether orbits $\{x.T(x), T^2(x), \ldots\}$ are equidistributed or not.

Example 2.2. Consider the *rotation* map

$$R_\alpha \colon S^1 \to S^1; \qquad x \mapsto x + \alpha \pmod 1.$$

If $\alpha \in \mathbb{Q}$ is a rational number, then every orbit of R_α is periodic, therefore cannot be equidistributed. If $\alpha \notin \mathbb{Q}$ is irrational, then one can show that every orbit of R_α is equidistributed (this is often thought of as the first ergodic theorem to have been proved).

Example 2.3. Consider the *doubling map*

$$T_2 \colon S^1 \to S^1; \qquad x \mapsto 2x \pmod 1.$$

It is not hard to see that there is a dense subset of X for which the orbit of T_2 is periodic, therefore not equidistributed. However, it turns out that for *almost all* $x \in X$ the orbit of T_2 is equidistributed.

We may also have maps (e.g. the continued fraction map) where the orbits are not equidistributed for almost all $x \in X$.

To make these notions precise, we need to introduce some measure theory, which have the advantage of introducing a theory of integration that is suitable for our purposes.

2.2. **σ-algebras, measures, probability spaces.** Intuitively, a *measure* μ on a space X is a function on a collection of subsets of X, called *measurable sets*, which assigns to each measurable set A its *measure* $\mu(A) \geq 0$. You already know at least two natural examples of measures.

Example 2.4. Let $X = \mathbb{R}$. The Lebesgue measure λ on $\mathbb{R}$ assigns to each interval $[a, b]$ its length

$$\lambda([a, b]) = b - a = \int_a^b \mathrm{d}x.$$

Let $X = \mathbb{R}^2$. The Lebesgue measure λ on $\mathbb{R}^2$ assigns to each measurable set $A \subseteq \mathbb{R}^2$ its area

$$\lambda(A) = \int_A \mathrm{d}x\, \mathrm{d}y.$$

One might hope to assign a measure to all subsets of X. Unfortunately, if we want the measure to have reasonable and useful properties, this would lead to a contradiction in certain cases (we will see an example later). So we are forced to assign a measure only to a sub-collection of all subsets of X.

Let X be a set. Denote by $\mathbb{P}(X)$ the collection of all subsets of X.

Definition 2.5. A subset $\mathcal{B} \subseteq \mathbb{P}(X)$ is called a *σ-algebra* on X if

(a) the empty set $\emptyset \in \mathcal{B}$,
(b) $\mathcal{B}$ is closed under complementation: $A \in \mathcal{B}$ implies $X \backslash A \in \mathcal{B}$,
(c) $\mathcal{B}$ is closed under countable union: $A_1, A_2, \ldots \in \mathcal{B}$ implies $\cup_{n=1}^{\infty} A_i \in \mathcal{B}$.

Elements of the σ-algebra are called *measurable sets*.

Remark 2.6. Let $F \subseteq \mathbb{P}(X)$ be an arbitrary subset (may or may not be a σ-algebra). Then there exists a unique smallest σ-algebra which contains every set in F. It is called the *σ-algebra generated by F*.

An important example is the *Borel algebra* over any *topological space*: it is the σ-algebra generated by the *open sets*. For instance, the Borel algebra over $[0,1]$ is the σ-algebra generated by the collection of open sub-intervals of $[0,1]$.

Definition 2.7. Let X be a set and $\mathcal{B}$ be a σ-algebra on X. A function $\mu\colon \mathcal{B} \to \mathbb{R} \cup \{\infty\}$ is called a *measure* if

(a) $\mu(\emptyset) = 0$,
(b) (non-negativity) $\mu(E) \geq 0$ for all $E \in \mathcal{B}$,
(c) (countable additivity) for all countable collections $\{E_k\}_{k=1}^{\infty}$ of pairwise disjoint sets in $\mathcal{B}$, we have

$$\mu\left(\bigcup_{k=1}^{\infty} E_k\right) = \sum_{k=1}^{\infty} \mu(E_k).$$

The triple $(X, \mathcal{B}, \mu)$ is called a *measurable space*, and it is called a *probability space* if $\mu(X) = 1$.

Example 2.8. Let $X = [0,1]$ and let $\mathcal{B}$ be the Borel algebra on X, i.e. the σ-algebra generated by all open subintervals (a,b). There exists a measure (the *Lebesgue* measure) $\lambda\colon \mathcal{B} \to \mathbb{R}$ such that $\lambda((a,b)) = b - a$. The triple $(X, \mathcal{B}, \lambda)$ forms a probability space.

Remark 2.9. Given a probability space $(X, \mathcal{B}, \mu)$, one can regard X as the space of all possible *events*, and $\mu(A)$ gives the probability of an event occurs in a measurable subset $A \subseteq X$.

Example 2.10. Let us consider a discrete example. Let $X = \{1, \ldots, n\}$, and let $\mathcal{B} = \mathbb{P}(X)$ be the σ-algebra consists of all subsets of X. Choose any $0 \leq p_1, \ldots, p_n \leq 1$ such that $\sum p_i = 1$. Then one can define a measure $\mu\colon \mathcal{B} \to \mathbb{R}$ by

$$\mu(\{i_1, \ldots, i_k\}) = p_{i_1} + \cdots + p_{i_k}.$$

Remark 2.11. In this remark, we show that in general it is necessary to restrict the definition of measure on a subset $\mathcal{B} \subseteq \mathbb{P}(X)$, as opposed to defining it on the *whole* collection of subsets of X. Consider the Lebesgue measure $\lambda\colon \mathcal{B} \to \mathbb{R}_{\geq 0}$ on $X = \mathbb{R}$. It satisfies the following properties:

- λ has the countable additivity property in the definition of measure,
- if two subsets of A and B are related by a translation, then $\lambda(A) = \lambda(B)$,
- $\lambda([0,1]) = 1$.

We show that unfortunately it is not possible to extend the definition of λ to *all* subsets of $\mathbb{R}$ that still satisfy these three properties.

Let us consider the example constructed by Vitali in 1905. A *Vitali set* is a subset $V \subseteq [0,1]$ of real numbers such that, for each real number r, there is exactly one number $v \in V$ such that $v - r \in \mathbb{Q}$. Equivalently, V is constructed by choosing a representative in $[0,1]$ of each element of the quotient group $\mathbb{R}/\mathbb{Q}$.

Let $q_1, q_2, \ldots$ be an enumeration of the rational numbers in $[-1,1]$ (recall that $\mathbb{Q}$ is *countable*). Consider the translated sets $V_k = V + q_k$ for $k = 1, 2, \ldots$. It is not hard to show the following:

- V_k's are pairwise disjoint,
- $[0,1] \subseteq \cup_{k=1}^{\infty} V_k \subseteq [-1,2]$.

Assume the contrary that it is possible to extend the definition of Lebesgue measure to *all* subsets of $\mathbb{R}$ which satisfies the properties above. Then we have

$$1 \leq \sum_{k=1}^{\infty} \lambda(V_k) \leq 3.$$

Since Lebesgue measure is translation invariant, we have $\lambda(V_k) = \lambda(V)$, hence

$$1 \leq \sum_{k=1}^{\infty} \lambda(V) \leq 3.$$

But this is impossible: If $\lambda(V) = 0$ then $\sum_{k=1}^{\infty} \lambda(V) = 0$; if $\lambda(V) > 0$ then $\sum_{k=1}^{\infty} \lambda(V) = \infty$. Contradiction.

2.3. **Measure-preserving functions.**

Definition 2.12. Let $(X, \mathcal{B}, \mu)$ and $(Y, \mathcal{C}, \nu)$ be two probability spaces.

- A map $T\colon X \to Y$ is called *measurable* if $T^{-1}(A) \in \mathcal{B}$ for any $A \in \mathcal{C}$.
- Furthermore, a measurable function T is called *measure-preserving* if $\mu(T^{-1}(A)) = \nu(A)$ for any $A \in \mathcal{C}$.

- If $T\colon X \to X$ is measure-preserving, then we say $(X, \mathcal{B}, \mu, T)$ is a *measure-preserving system*.

Exercise. Let X be a topological space and $\mathcal{B}$ be the Borel σ-algebra on X (which is generated by open sets of X). Show that any *continuous* map $T\colon X \to X$ is measurable.

Exercise. To show a measurable map $T\colon X \to Y$ is measure-preserving, it is enough to check $\mu(T^{-1}(A)) = \nu(A)$ holds for a generating set of $\mathcal{C}$.

Example 2.13 (Rotation on S^1). Consider the circle $S^1 \cong \mathbb{R}/\mathbb{Z}$, which can be obtained by identifying the two endpoints of $[0, 1]$. One equips S^1 with the Lebesgue measure. It is easy to show that the rotation

$$R_\alpha\colon S^1 \to S^1; \qquad x \mapsto x + \alpha \pmod 1$$

is measure-preserving for any α.

Example 2.14 (Doubling map on S^1). Define the *doubling map*

$$T_2\colon S^1 \to S^1; \qquad x \mapsto 2x \pmod 1.$$

Let us show that it is measure-preserving. It is enough to check this on intervals: we have $\mu(T_2^{-1}(a, b)) = \mu(a, b)$ since

$$T_2^{-1}(a, b) = \left(\frac{a}{2}, \frac{b}{2}\right) \cup \left(\frac{a+1}{2}, \frac{b+1}{2}\right).$$

Note that the measure-preserving property cannot be seen by studying "forward iterates": $\mu(T_2(a, b)) \neq \mu(a, b)$ in general.

Example 2.15. Define the $(\frac{1}{2}, \frac{1}{2})$-measure $\mu_{(1/2,1/2)}$ on the finite set $\{1, 2\}$ by

$$\mu_{(1/2,1/2)}(\{1\}) = \mu_{(1/2,1/2)}(\{2\}) = \frac{1}{2}.$$

Consider the space of infinite product $X = \{1, 2\}^{\mathbb{N}}$, which models the set of possible outcomes of the infinitely repeated toss of a coin. Given a finite subset $I \subseteq \mathbb{N}$ and a map $a\colon I \to \{1, 2\}$, we define the *cylinder set* associated to I and a to be

$$I(a) = \{x \in X \mid x_j = a(j) \text{ for all } j \in I\},$$

i.e. one specifies the outcome of the j-th thows for all $j \in I$. We define $\mathcal{B}$ to be the σ-algebra generated by all cylinder sets, and define a measure $\mu\colon \mathcal{B} \to \mathbb{R}$ via

$$\mu(I(a)) = \left(\frac{1}{2}\right)^{\#|I|}.$$

Consider the *left shift map* $\sigma\colon X \to X$ defined by

$$\sigma(x_1, x_2, \ldots) = (x_2, x_3, \ldots).$$

It is easy to see that $(X, \mathcal{B}, \mu, \sigma)$ is a measure-preserving system.

In fact, this system is *measurably isomorphic* to the doubling map T_2 on S^1, which roughly means that they are identical except on a measure zero set. Indeed, consider the map $\phi\colon X \to S^1 \cong [0,1]/0 \sim 1$ where

$$\phi(x_1, x_2, \ldots) = \sum_{n=1}^{\infty} \frac{x_n}{2^n}.$$

Then we have $\phi \circ \sigma = T_2 \circ \phi$. Below is the precise definition of the notion of measurably isomorphic.

Definition 2.16. We say two measure-preserving systems $(X, \mathcal{B}, \mu, T)$ and $(Y, \mathcal{C}, \nu, S)$ are *measurably isomorphic* if there exists $X' \in \mathcal{B}$ and $Y' \in \mathcal{C}$ such that:

- $\mu(X') = \nu(Y') = 1$,
- $T(X') \subseteq X'$, $S(Y') \subseteq Y'$,
- there exists a bijective map $\phi\colon X' \to Y'$ such that both ϕ and ϕ^{-1} are measurable and measure-preserving, and
- $\phi \circ T(x) = S \circ \phi(x)$ for any $x \in X'$.

Example 2.17 (Bernoulli shift). Consider the two-sided infinite set

$$\begin{aligned} X &= \{1, \ldots, n\}^{\mathbb{Z}} \\ &= \{x = (\ldots, x_{-1}, x_0, x_1, \ldots) \mid x_i \in \{1, \ldots, n\} \text{ for all } i\}. \end{aligned}$$

which gives the sample space of the outcome of throwing an n-sided die (each appears with probabilities $p_1, \ldots, p_n$) infinitely many times. Let us define a σ-algebra and a measure on X. Given a finite subset $I \subseteq \mathbb{Z}$ and a map $a\colon I \to \{1, \ldots, n\}$, we define the *cylinder set* associated to I and a to be

$$I(a) = \{x \in X \mid x_j = a(j) \text{ for all } j \in I\},$$

i.e. one specifies the outcome of the j-th thows for all $j \in I$. We define $\mathcal{B}$ to be the σ-algebra generated by all cylinder sets, and define a measure $\mu\colon \mathcal{B} \to \mathbb{R}$ via

$$\mu(I(a)) = \prod_{j \in I} p_{a(j)}.$$

Now, consider the left shift map $\sigma\colon X \to X$ defined by $\sigma(x)_i = x_{i+1}$. It clearly preserves the measure of all cylinder sets, hence $(X, \mathcal{B}, \mu, \sigma)$ is a measure-preserving system. The map σ is called the *Bernoulli shift*.

2.4. **Recurrence.** One of the central themes in ergodic theory is *recurrence*, which concerns how points in measurable dynamical systems return close to themselves under iterations.

Theorem 2.18 (Poincaré recurrence)**.** *Let $T\colon X \to X$ be a measure-preserving transformation on a probability space $(X, \mathcal{B}, \mu)$, and let $E \in \mathcal{B}$ be a measurable set with $\mu(E) > 0$. Then almost every point $x \in E$ returns to E infinitely many often under iterations of T. More precisely, there exists a measurable set $F \subseteq E$ such that $\mu(F) = \mu(E)$, and for every point $x \in F$ the sequence of points $\{T^n(x)\}_{n=1}^{\infty}$ returns to E infinitely many times.*

Proof. Let

$$B = \{x \in E \mid T^n(x) \notin E \text{ for all } n \geq 1\}.$$

It is an easy exercise to show that B is measurable. Using the definition of B, one can show that the sets $B, T^{-1}B, T^{-2}B, \ldots$ are pairwise disjoint. Hence

$$\sum_{k=0}^{\infty} \mu(T^{-k}B) = \mu\left(\bigcup_{k=0}^{\infty} T^{-k}B\right) \leq \mu(X) = 1.$$

Therefore we have $\mu(B) = 0$, since T is measure-preserving.

Observe that the points of the union

$$\bigcup_{k=0}^{\infty} (T^{-k}B \cap E)$$

are precisely those points of E which do not return to E infinitely many often. Therefore, it suffices to show that the measure of the above union is zero.

$$\mu\left(\bigcup_{k=0}^{\infty} (T^{-k}B \cap E)\right) \leq \mu\left(\bigcup_{k=0}^{\infty} T^{-k}B\right) = \sum_{k=0}^{\infty} \mu(T^{-k}B) = 0$$

since $\mu(B) = 0$ and T is measure-preserving. □

Remark 2.19. The key step of the proof is to show that $\mu(B) = 0$, which is essentially the pigeon-hole principle: the sets $B, T^{-1}B, T^{-2}B, \dots$ are disjoint and with the same measure, so they can not fit into a space of finite measure ($\mu(X) = 1$) unless $\mu(B) = 0$. The recurrence property does not hold for spaces of infinite measure (can you give an example?).

Remark 2.20. If one further assumes that the map $T\colon X \to X$ is *ergodic*, then one can show that the *frequency* of return to the set E is precisely $\mu(E) > 0$.

2.5. **Lebesgue integral.**

Definition 2.21. Let $(X, \mathcal{B}, \mu)$ be a probability space. A function $f\colon X \to \mathbb{R}$ is called *measurable* if $f^{-1}(A) \in \mathcal{B}$ for any (Borel) measurable set $A \subseteq \mathbb{R}$.

We would like to define the *(Lebesgue) integral* $\int f \,\mathrm{d}\mu$ of measurable functions f. First, a function $g\colon X \to \mathbb{R}$ is called *simple* if

$$g(x) = \sum_{j=1}^{m} c_j \chi_{A_j}(x)$$

for some constants $c_j \in \mathbb{R}$ and *disjoint* measurable sets $A_j \in \mathcal{B}$. In this case, the integral of g is defined to be

$$\int g \,\mathrm{d}\mu = \sum_{j=1}^{m} c_j \mu(A_j).$$

Second, one can show that for any *non-negative* measurable function $f\colon X \to \mathbb{R}_{\geq 0}$, there exists a pointwise increasing sequence of simple functions $(g_n)_{n\geq 1}$ which converges to g_n pointwisely converges to f. This allows us to define

$$\int f \,\mathrm{d}\mu = \lim_{n\to\infty} \int g_n \,\mathrm{d}\mu.$$

A non-negative measurable function $f\colon X \to \mathbb{R}_{\geq 0}$ is called *integrable* if $\int f \,\mathrm{d}\mu < \infty$.

Finally, for a general measurable function $f\colon X \to \mathbb{R}$, one can decompose it into $f = f^+ - f^-$ where $f^+(x) = \max\{f(x), 0\}$. Both f^+, f^- are non-negative measurable functions. The function f is called *integrable* if both f^+, f^- are

integrable, and its integral is defined to be

$$\int f \,\mathrm{d}\mu = \int f^+ \,\mathrm{d}\mu - \int f^- \,\mathrm{d}\mu.$$

Notation. Let $(X, \mathcal{B}, \mu)$ be a measurable space. Define

$$L^1_\mu = \left\{ f \colon X \to \mathbb{R} : f \text{ is measurable and } \|f\|_1 \coloneqq \int |f| \,\mathrm{d}\mu < \infty \right\}.$$

Similarly, define

$$L^2_\mu = \left\{ f \colon X \to \mathbb{R} : f \text{ is measurable and } \|f\|_2 \coloneqq \left(\int |f|^2 \,\mathrm{d}\mu \right)^{1/2} < \infty \right\}.$$

The following theorem provides an important characterization of measure-preserving maps.

Theorem 2.22. *Let $(X, \mathcal{B}, \mu)$ be a probability space. A map $T \colon X \to X$ is measure-preserving if and only if*

$$\int f \,\mathrm{d}\mu = \int f \circ T \,\mathrm{d}\mu \qquad \textit{for all } f \in L^1_\mu.$$

Proof. First, we prove the "if" part. Take $f = \chi_B$ for any $B \in \mathcal{B}$, one gets

$$\mu(T^{-1}B) = \int \chi_{T^{-1}B} \,\mathrm{d}\mu = \int \chi_B \circ T \,\mathrm{d}\mu = \int \chi_B d\mu = \mu(B).$$

Conversely, if T is measure-preserving, then the integral equality holds for any simple functions. For any $f \in L^1_\mu$, one can take an increasing sequence (f_n) of simple functions such that $\lim f_n = f$ pointwise. Hence we also have $\lim f_n \circ T = f \circ T$. By dominated convergence theorem,

$$\int f \,\mathrm{d}\mu = \lim_{n\to\infty} \int f_n \,\mathrm{d}\mu = \lim_{n\to\infty} \int f_n \circ T \,\mathrm{d}\mu \int f \circ T \,\mathrm{d}\mu$$

□

Remark 2.23. The Lebesgue integral is more general than the *Riemann integral*: The Lebesgue integral allows a countable infinity of discontinuities, while Riemann integral allows only a finite number of discontinuities. As an example, consider the set $A = \mathbb{Q} \cap [0, 1]$ of rational numbers in $[0, 1]$. It is an easy exercise of Riemann integral to show that the characteristic function

$\chi_A\colon [0,1] \to \mathbb{R}$ is not integrable. On the other hand, the set A is measurable and its Lebesgue measure is $\lambda(A) = 0$. Therefore, χ_A is Lebesgue measurable and

$$\int \chi_A \, \mathrm{d}\lambda = \lambda(A) = 0.$$

Lecture 2

2.6. **Ergodicity.**

Definition 2.24. Let $(X, \mathcal{B}, \mu)$ be a probability space. A measure-preserving transformation $T\colon X \to X$ is said to be *ergodic* if for any $B \in \mathcal{B}$,

$$T^{-1}B = B \implies \mu(B) = 0 \text{ or } \mu(B) = 1.$$

In words, it is impossible to split X into T-invariant subsets of positive measures.

Non-example. Consider the rotation map $R_\alpha(x) = x + \alpha \pmod 1$ on the circle S^1. It is not hard to show that if α is rational then R_α is not ergodic. For instance, when $\alpha = \frac{1}{2}$, the set $B = (0, \frac{1}{4}) \cup (\frac{1}{2}, \frac{3}{4})$ satisfies $R_\alpha^{-1}B = B$ but $\mu(B) = \frac{1}{2}$. We will see later that if α is irrational then R_α is ergodic.

Example 2.25. Let us show that the *Bernoulli shifts* σ are ergodic. First, we claim that the Bernoulli shifts are *mixing*, i.e.

$$\lim_{n\to\infty} \mu(B \cap \sigma^{-n}B') = \mu(B)\mu(B') \qquad \text{for all } B, B' \in \mathcal{B}.$$

It is easy to see that the statement is true if B and B' are both finite unions of cylinder sets. By Kolmogorov extension theorem (which we will not discuss here), for any measurable set B and any $\epsilon > 0$, there exists a finite union of cylinder sets A such that $\mu(A \Delta B) < \epsilon$. (Here $A \Delta B \coloneqq (A \backslash B) \cup (B \backslash A)$.) It is then an easy exercise to show the mixing property.

Second, we claim that mixing implies ergodic. Let $B = \sigma^{-1}B$ be a measurable σ-invariant set. By the mixing property, we have

$$\mu(B) = \lim_{n\to\infty} \mu(B \cap \sigma^{-n}B) = \mu(B)^2.$$

Hence $\mu(B) \in \{0, 1\}$.

Remark 2.26. As the proof above suggests, the concept of ergodicity is closely related to the idea of *mixing*, meaning, given a measurable set $A \subseteq X$, how the set $T^{-n}A$ is spread around the whole space X under large iterations n? It

can be proved that a measure-preserving system $(X, \mathcal{B}, \mu, T)$ is ergodic if and only if it is *weak-mixing* (a weaker condition than *mixing*), i.e.

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mu(A \cap T^{-n}B) = \mu(A)\mu(B) \text{ for all } A, B \in \mathcal{B}.$$

A proof of this fact can be found in [7, Section 2.7].

The following theorem is very useful for proving a system is ergodic (or non-ergodic).

Theorem 2.27. *For a measure-preserving system $(X, \mathcal{B}, \mu, T)$, the following are equivalent.*

(a) T is ergodic.
(b) For any $f\colon X \to \mathbb{R}$ measurable, if $f \circ T = f$ almost everywhere, then f is constant almost everywhere.

Proof. It is easy to see that (b) implies (a): Suppose $T^{-1}B = B$. Take $f = \chi_B$. Then we have χ_B is constant almost everywhere, thus $\mu(B) \in \{0, 1\}$. A proof of (a) implies (b) can be found in [7, Proposition 2.14]. □

Remark 2.28. One can show that in the characterization theorem above, instead of considering all measurable functions, it is enough to consider only the integrable functions $f \in L^1_\mu$ or the square-integrable functions $f \in L^2_\mu$. More precisely, for a measure-preserving system $(X, \mathcal{B}, \mu, T)$, the following statements are all equivalent:

(a) T is ergodic.
(b) For any $f\colon X \to \mathbb{R}$ measurable, if $f \circ T = f$ almost everywhere, then f is constant almost everywhere.
(c) For any $f \in L^1_\mu$, if $f \circ T = f$ almost everywhere, then f is constant almost everywhere.
(d) For any $f \in L^2_\mu$, if $f \circ T = f$ almost everywhere, then f is constant almost everywhere.

Using this remark and some basic knowledge of *Fourier series*, one can easily show that the rotation maps and the doubling map of S^1 are ergodic. Let $f\colon S^1 \cong \mathbb{R}/\mathbb{Z} \to \mathbb{R}$ be a square-integrable function, i.e. $f \in L^2(S^1)$. Results of

Fourier series imply that there exists a *unique* collection of complex numbers $\ldots, c_{-2}, c_{-1}, c_0, c_1, c_2, \ldots$, called the *Fourier coefficients* of f, such that

$$f(x) = \sum_{n\in\mathbb{Z}} c_n e^{2\pi i n x} \qquad \text{for a.e. } x \in \mathbb{R}/\mathbb{Z}.$$

Moreover, we have $\|f\|_2 = \sum_{n\in\mathbb{Z}} |c_n|^2 < \infty$.

Example 2.29. Consider the rotation map $R_\alpha(x) = x + \alpha \pmod 1$ on the circle S^1 where α is irrational. By Remark 2.28, it suffices to show that for any $f \in L^2(S^1)$, if $f \circ R_\alpha = f$ almost everywhere, then f is constant almost everywhere. Let the Fourier series of f be $\sum_{n\in\mathbb{Z}} c_n e^{2\pi i n x}$. Then

$$\left(\sum_{n\in\mathbb{Z}} c_n e^{2\pi i n x}\right) \circ R_\alpha = \sum_{n\in\mathbb{Z}} c_n e^{2\pi i n(x+\alpha)} = \sum_{n\in\mathbb{Z}} c_n e^{2\pi i n\alpha} e^{2\pi i n x}.$$

By the uniqueness of the Fourier coefficients, we have

$$c_n \left(1 - e^{2\pi i n\alpha}\right) = 0 \qquad \text{for all } n \in \mathbb{Z}.$$

Suppose α is irrational, then $1 - e^{2\pi i n\alpha} \neq 0$ for all $n \in \mathbb{Z}\backslash\{0\}$, thus we have $c_n = 0$ for all $n \in \mathbb{Z}\backslash\{0\}$. Hence $f(x) = \sum_{n\in\mathbb{Z}} c_n e^{2\pi i n x} = c_0$ is constant almost everywhere. (Can you identify at where this argument fails for α rational?)

Example 2.30. We show that the doubling map $T_2 \colon S^1 \to S^1$ is ergodic. Let $f \in L^2(S^1)$ with $f \circ T = f$ almost everywhere. Let $\sum_{n\in\mathbb{Z}} c_n e^{2\pi i n x}$ be the Fourier series of f, where $\|f\|_2^2 = \sum_{i\in\mathbb{Z}} |a_i|^2 < \infty$. Then

$$\left(\sum_{n\in\mathbb{Z}} c_n e^{2\pi i n x}\right) \circ T_2 = \sum_{n\in\mathbb{Z}} c_n e^{2\pi i n(2x)} = \sum_{n\in\mathbb{Z}} c_n e^{2\pi i (2n) x}$$

By the uniqueness of the Fourier coefficients, we have $c_n = c_{2n}$ for all $n \in \mathbb{Z}$. This implies that $c_n = 0$ for all $n \neq 0$. Hence f is a constant function almost everywhere.

2.7. **Ergodic theorems.** Let X be the *phase space* of a physical system (e.g. the points of X can represent configurations of positions and velocities of particles in a box). A measurable function $f \colon X \to \mathbb{R}$ represents an *observable* of the system, i.e. a quantity that can be measured (e.g. velocity, temperature, position, etc.). The value $f(x)$ is the measurement of the observable f that

one gets when the system is in the state x. *Time evolution* of the system, if measured by discrete time units, can be given by a transformation $T\colon X \to X$, so that if $x \in X$ is the initial state of the system, then $T(x)$ is the state of the system after one time unit. The map T is measure-preserving if the system is in equilibrium.

In order to measure a physical quantity, one usually measures repeatedly in time and consider their average. The average of the first n measurements is given by

$$\frac{1}{n}\sum_{j=0}^{n-1} f(T^j x).$$

This quantity is called the *time average*. On the other hand, the *space average* of the observable f is simply

$$\int f \,\mathrm{d}\mu.$$

In physics, one would like to know the space average of the observable; but since experimentally it is easier to compute the time average, it is natural to ask whether the time average gives a good approximation of the space average as $n \to \infty$.

Boltzmann's Hypothesis was that for almost every initial state $x \in X$ the time averages of any observable f converge to the space average as time tends to infinity. Unfortunately, this is not true for general measure-preserving map T. On the other hand, *under the assumption that T is ergodic*, the conclusion of Boltzmann's Hypothesis is true, and this is exactly the content of Birkhoff's ergodic theorem. Finding the right condition under which Boltzmann's Hypothesis holds motivated the definition of ergodicity, and gave birth to the study of ergodic theory.

Theorem 2.31. *Let $(X, \mathcal{B}, \mu, T)$ be a measure-preserving system on a probability space, and let $f\colon X \to \mathbb{R}$ be an integrable function.*

(a) The limit

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} f(T^j x) = f^*(x)$$

converges almost everywhere to a T-invariant integrable function f^*, where

$$\int f^* \, \mathrm{d}\mu = \int f \, \mathrm{d}\mu.$$

(b) Moreover, if T is ergodic, then

$$f^*(x) = \int f \, \mathrm{d}\mu$$

almost everywhere.

A proof of the theorem can be found in [7, Section 2.6]. Note that the second part of the statement is an easy corollary of the first part using Theorem 2.27.

Remark 2.32. Note that for an ergodic system $(X, \mathcal{B}, \mu, T)$ and a measurable function $f\colon X \to \mathbb{R}$, the ergodic theorem only guarantees the limit

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} f(T^j x) = \int f \, \mathrm{d}\mu$$

almost everywhere; the equality may not be satisfied by *every* points of X. For instance, consider the doubling map $T_2\colon S^1 \to S^1$ which is ergodic. Choose any measurable function $f\colon S^1 \to \mathbb{R}$ such that $\int f \, \mathrm{d}\mu \neq f(0)$. Then the above equality is not satisfied at the point $x = 0 \in S^1$.

Example 2.33 (Frequency of visits). Let $(X, \mathcal{B}, \mu, T)$ be a measure-preserving ergodic system, and let $A \subseteq X$ be a measurable set with $\mu(A) > 0$. We would like to understand the frequency of visits:

$$\frac{\#\{0 \leq k \leq n-1 \mid T^k(x) \in A\}}{n} = \frac{1}{n} \sum_{k=0}^{n-1} \chi_A(T^k(x)).$$

Applying Birkhoff's pointwise ergodic theorem to $f = \chi_A$, one gets

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \chi_A(T^k(x)) = \int \chi_A \, \mathrm{d}\mu = \mu(A).$$

2.8. **Back to continued fractions.**

Definition 2.34. A *continued fraction* is an expression of the form

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}},$$

denotes alternatively by $[a_0; a_1, a_2, a_3, \ldots]$, where $a_0 \in \mathbb{Z}_{\geq 0}$ and $a_n \in \mathbb{Z}_{>0}$ for all $n \geq 1$. This expression can be finite (when the represented number is rational) or infinite (when the represented number is irrational).

Exercise. Fix a sequence $(a_n)_{n\geq 0}$ where $a_0 \in \mathbb{Z}_{\geq 0}$ and $a_n \in \mathbb{Z}_{>0}$ for all $n \geq 1$. Denote the partial expressions as

$$\frac{p_n}{q_n} = [a_0; a_1, \ldots, a_n]$$

where p_n, q_n are coprime positive integers. Then they satisfy the recursive relation

$$\begin{pmatrix} p_n & p_{n-1} \\ q_n & q_{n-1} \end{pmatrix} = \begin{pmatrix} a_0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a_1 & 1 \\ 1 & 0 \end{pmatrix} \cdots \begin{pmatrix} a_n & 1 \\ 1 & 0 \end{pmatrix}.$$

Therefore, we have

$$p_{n+1} = a_{n+1}p_n + p_{n-1}, \qquad q_{n+1} = a_{n+1}q_n + q_{n-1}.$$

Also, by taking the determinants of the matrix equation, we get

$$p_n q_{n-1} - p_{n-1}q_n = (-1)^{n+1}.$$

Hence

$$\begin{aligned} \frac{p_n}{q_n} &= \frac{p_{n-1}}{q_{n-1}} + (-1)^{n+1}\frac{1}{q_{n-1}q_n} \\ &= a_0 + \frac{1}{q_0 q_1} - \frac{1}{q_1 q_2} + \cdots + (-1)^{n+1}\frac{1}{q_{n-1}q_n} \end{aligned}$$

by induction, and show that

$$x = \lim_{n\to\infty}[a_0; a_1, \ldots, a_n] = \lim_{n\to\infty}\frac{p_n}{q_n} = a_0 + \sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{q_{n-1}q_n}.$$

Moreover, we have

$$\frac{p_0}{q_0} < \frac{p_2}{q_2} < \cdots < \frac{p_{2n}}{q_{2n}} < \cdots < x < \cdots < \frac{p_{2n+1}}{q_{2n+1}} < \cdots < \frac{p_3}{q_3} < \frac{p_1}{q_1}.$$

The rational numbers $\frac{p_n}{q_n}$ are called the *convergents* of the continued fraction for x, and they provide very rapid rational approximation to x. We have

$$\left| x - \frac{p_n}{q_n} \right| < \frac{1}{q_n q_{n+1}}.$$

The numbers q_n and p_n grow exponentially as $n \to \infty$: using the recursive relation, one can show that both p_n and q_n are greater than $2^{(n-2)/2}$.

In fact, the continued fraction convergents provide the *optimal* rational approximants of an irrational number in the following sense.

Proposition 2.35. *Let $x > 0$ be an irrational number, $[a_0; a_1, \ldots]$ be its associated continued fraction, and $\frac{p_n}{q_n}$ be its convergents defined above. For any $1 \leq q < q_n$ and any $p_n > 0$, we have*

$$\left| x - \frac{p_n}{q_n} \right| < \left| x - \frac{p}{q} \right|.$$

Definition 2.36. Define the *continued fraction map* $T\colon [0,1] \to [0,1]$ by $T(0) = 0$ and

$$T(x) = \frac{1}{x} - \left\lfloor \frac{1}{x} \right\rfloor \text{ for } x \neq 0,$$

where $\lfloor t \rfloor$ denotes the greatest integer less than or equal to t. In other words, $T(x)$ is the fractional part $\{\frac{1}{x}\}$ of $\frac{1}{x}$.

For our purpose, we would like to find a measure on $[0,1]$ such that the continued fraction map T is measure-preserving. Unfortunately, the usual Lebesgue measure on $[0,1]$ does not work. For instance,

$$T^{-1}\left(0, \frac{1}{2}\right) = \left(\frac{2}{3}, 1\right) \cup \left(\frac{2}{5}, \frac{1}{2}\right) \cup \left(\frac{2}{7}, \frac{1}{3}\right) \cup \cdots,$$

which has measure strictly greater than $1/2$ with respect to the standard Lebesgue measure.

Definition 2.37. Define the *Gauss measure* μ on $[0,1]$ to be

$$\mu(A) = \frac{1}{\log 2} \int_A \frac{1}{1+x} dx \text{ for any measurable set } A \subseteq [0,1].$$

Exercise. The Gauss measure is "comparable" with the standard Lebesgue measure λ on $[0,1]$: Show that

$$\frac{\lambda(B)}{2\log 2} \leq \mu(B) \leq \frac{\lambda(B)}{\log 2} \qquad \text{for any measurable set } B \subseteq [0,1].$$

Proposition 2.38. *The continued fraction map T preserves the Gauss measure μ.*

Proof. It suffices to show it for $A = [0,b]$ for all $b > 0$. Observe that

$$T^{-1}[0,b] = \bigcup_{n=1}^{\infty} \left[\frac{1}{b+n}, \frac{1}{n}\right].$$

It is an easy exercise to show that

$$\begin{aligned}\mu(T^{-1}[0,b]) &= \frac{1}{\log 2} \sum_{n=1}^{\infty} \int_{\frac{1}{b+n}}^{\frac{1}{n}} \frac{1}{1+x} dx \\ &= \frac{1}{\log 2} \int_0^b \frac{1}{1+x} dx \\ &= \mu([0,b]).\end{aligned}$$

□

We now move on to prove the *ergodicity* of the continued fraction map T with respect to the Gauss measure. Notice that in terms of the continued fraction expansion, T behaves similar to the shift map in that

$$T([a_1, a_2, \ldots]) = [a_2, a_3, \ldots].$$

We therefore would like to pursue a method of proof similar to the proof of the ergodicity of Bernoulli shifts: we want to control the size of the *cylinder sets* and their *intersections*.

Exercise. Given an n-tuple $a = (a_1, \ldots, a_n) \in \mathbb{Z}_{>0}^n$ of positive integers, define the cylinder set

$$I(a) = \{[x_1, x_2, \ldots] \mid x_i = a_i \text{ for } 1 \leq i \leq n\} \subseteq [0,1].$$

- $I(a)$ is a subinterval of $[0,1]$ with length $\frac{1}{q_n(q_n+q_{n-1})}$, where $\frac{p_n}{q_n}$ is the convergent of $[a_1, \ldots, a_n]$.

- Since $q_n \geq 2^{(n-2)/2}$, the length of $I(a) = I([a_1, \ldots, a_n])$ shrinks to zero as $n \to \infty$. Use this to show that the cylinder sets $I(a)$ for all possible strings of positive integers generate the Borel σ-algebra on $[0, 1]$.

Proposition 2.39. *The continued fraction map T on $[0, 1]$ is ergodic with respect to the Gauss measure μ.*

Proof. The key step of the proof is to show that

$$\mu(T^{-n}A \cap I(a)) \asymp \mu(A)\mu(I(a)) \qquad \text{for any measurable set } A, \tag{2.1}$$

i.e. there exist constants $C_1, C_2 > 0$ which are independent of the choice of A (but may depend on $I(a)$), such that

$$C_1\mu(T^{-n}A \cap I(a)) \leq \mu(A)\mu(I(a)) \leq C_2\mu(T^{-n}A \cap I(a)).$$

We first prove that T is ergodic assuming (2.1). Let $B \subseteq [0, 1]$ be a measurable set with $T^{-1}B = B$. By (2.1) we have

$$\mu(B \cap I(a)) \asymp \mu(B)\mu(I(a)).$$

Since the cylinder sets generate the Borel σ-algebra of A, we have

$$\mu(B \cap A) \asymp \mu(B)\mu(A) \qquad \text{for any measurable set } A.$$

By applying this to $A = X\backslash B$, we obtain $\mu(B)\mu(X\backslash B) = 0$, which concludes the proof.

We now proceed to prove (2.1). Recall that the Gauss measure μ is comparable with the Lebesgue measure λ, thus it suffices to show

$$\lambda(T^{-n}A \cap I(a)) \asymp \lambda(A)\lambda(I(a)) \qquad \text{for any measurable set } A$$

As usual, it suffices to show it for any interval $A = [d, e]$. It is an exercise to show that $T^{-n}A \cap I(a)$ is an interval with endpoints given by

$$\frac{p_n + p_{n-1}d}{q_n + q_{n-1}d} \qquad \text{and} \qquad \frac{p_n + p_{n-1}e}{q_n + q_{n-1}e}.$$

Therefore

$$\begin{aligned}
\lambda(T^{-n}A \cap I(a)) &= \frac{e - d}{(q_n + q_{n-1}d)(q_n + q_{n-1}e)} \\
&= \lambda(A)\lambda(I(a))\frac{q_n(q_n + q_{n-1})}{(q_n + q_{n-1}d)(q_n + q_{n-1}e)} \\
&\asymp \lambda(A)\lambda(I(a)).
\end{aligned}$$

□

Example 2.40. This answers our motivating question: By applying Birkhoff's pointwise ergodic theorem, we have

$$\begin{aligned}\lim_{n\to\infty}\frac{\#\{i \mid a_i = k, 1 \le i \le n\}}{n} &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\chi_{(\frac{1}{k+1},\frac{1}{k}]}(T^i(x)) \\ &= \int \chi_{(\frac{1}{k+1},\frac{1}{k}]}\,\mathrm{d}\mu \\ &= \mu\left(\left(\frac{1}{k+1},\frac{1}{k}\right]\right) \\ &= \frac{1}{\log 2}\int_{\frac{1}{k+1}}^{\frac{1}{k}}\frac{1}{1+x}\,\mathrm{d}x = \frac{1}{\log 2}\log\left(\frac{(k+1)^2}{k(k+2)}\right)\end{aligned}$$

for almost every $x \in (0,1)$.

Example 2.41. The following result also is an application of the pointwise ergodic theorem: for almost every $x \in (0,1)$, the rate of approximation of the continued fractions is given by

$$\lim_{n\to\infty}\frac{1}{n}\log\left|x - \frac{p_n(x)}{q_n(x)}\right| = \frac{-\pi^2}{6\log 2}.$$

3. Topology

3.1. **The Borsuk–Ulam theorem.** Let us consider the following *continuous* version of the necklace splitting problem. We say a probability measure μ on $[0,1]$ is *continuous* if $\int_0^x \mathrm{d}\mu$ is continuous in x.

Question 3.1. *Let $\mu_1, \dots, \mu_n$ be continuous probability measures on $[0,1]$. Does there exist a partition of $[0,1]$ into $n+1$ intervals $I_0, \dots, I_n$ and signs $\epsilon_0, \dots, \epsilon_n \in \{\pm 1\}$ such that*

$$\sum_{j=0}^{n}\epsilon_j \cdot \mu_i(I_j) = 0 \quad \text{for all } 1 \le i \le n \text{ ?}$$

Remark 3.2. In the original necklace splitting problem, the n measures μ_i corresponds to the n kinds of precious stones, the interval $[0,1]$ is separated into $n+1$ subintervals by n cuts, and the signs ± 1 determine the corresponding

portion of the necklace belongs to which one of the two thieves. An affirmative answer to the above continuous version would imply an affirmative answer to the original necklace splitting problem. For more details, cf. [14].

There is a clever way to encode the divisions of the necklace by points of the *n-dimensional sphere* S^n. With every point of the sphere

$$S^n = \left\{(x_0, \dots, x_n) \in \mathbb{R}^{n+1} \mid x_0^2 + \cdots + x_n^2 = 1\right\}$$

we associate a division of the interval $[0, 1]$ into $n+1$ parts, of lengths $x_0^2, \dots, x_n^2$; i.e. we cut the interval at the points $0 = z_0 \le z_1 \le \cdots z_n \le z_{n+1} \le 1$. The sign ϵ_j for the j-th interval $[z_{j-1}, z_j]$ is chosen as $\operatorname{sign}(x_j)$. This defines a continuous map $g\colon S^n \to \mathbb{R}^n$, where its i-th component is given by

$$g_i(x) = \sum_{j=0}^{n} \operatorname{sign}(x_j) \cdot \mu_i\left([z_{j-1} - z_j]\right).$$

The function g clearly satisfies $g(-x) = -g(x)$ for all $x \in S^n$. We would like to show that $g(x) = 0$ for some $x \in S^n$. It follows directly from the *Borsuk–Ulam theorem*.

Theorem 3.3 (Borsuk–Ulam)**.** *Let $f\colon S^n \to \mathbb{R}^n$ be a continuous map. Then there exists an $x \in S^n$ such that $f(-x) = f(x)$.*

For instance, the case $n = 2$ can be illustrated by saying that at any moment, there is always a pair of antipodal points on the Earth's surface with equal temperatures and equal pressures.

Lecture 3

Exercise. For any $n \ge 1$, the following statements are equivalent:

- For every continuous map $f\colon S^n \to \mathbb{R}^n$ there exists a point $x \in S^n$ such that $f(-x) = f(x)$.
- For every *antipodal* continuous map $f\colon S^n \to \mathbb{R}^n$ (antipodal means $f(-x) = -f(x)$ for all $x \in S^n$), there exists $x \in S^n$ such that $f(x) = 0$.
- There is no antipodal map $f\colon S^n \to S^{n-1}$.
- There is no continuous map $f\colon B^n \to S^{n-1}$ that is antipodal on the boundary, i.e. satisfies $f(-x) = -f(x)$ for all $x \in S^{n-1} = \partial B^n$.

Remark 3.4. As a direct corollary, there is no continuous map $f\colon B^n \to S^{n-1}$ that is the *identity* on the boundary $\partial B^n = S^{n-1}$, which implies the *Brouwer fixed point theorem*.

As an another corollary of the Borsuk–Ulam theorem, one can show the following *ham sandwich theorem*. The informal statement that gave the ham sandwich theorem its name is this: "For every sandwich made of ham, cheese, and bread, there is a planar cut that simultaneously halves the ham, the cheese, and the bread."

Theorem 3.5 (Ham sandwich theorem)**.** *For any compact sets $A_1, \dots, A_n \subseteq \mathbb{R}^n$, there exists a hyperplane dividing each of them into two subsets of equal measure.*

One can prove a more general version of ham sandwich theorem in terms of measures. We say a measure on $\mathbb{R}^n$ is a *finite Borel measure* if all open subsets of $\mathbb{R}^n$ are measurable and $0 < \mu(\mathbb{R}^n) < \infty$. For instance, for any compact set $A \subseteq \mathbb{R}^n$, one can define a finite Borel measure μ_A by $\mu_A(X) \coloneqq \lambda(X \cap A)$.

Theorem 3.6 (Ham sandwich theorem for measures)**.** *For any finite Borel measures $\mu_1, \dots, \mu_n$ on $\mathbb{R}^n$, there exists a hyperplane h such that*

$$\mu_i(h^+) = \frac{1}{2}\mu_i(\mathbb{R}^n) \quad \text{for } 1 \le i \le n$$

where h^+ denotes one of the half-spaces defined by h.

Proof. Let $u = (u_0, \dots, u_n)$ be a point of the sphere $u \in S^n$. If at least one of the components $u_1, \dots, u_n$ is nonzero, we assign u the half-space

$$h^+(u) = \{(x_1, \dots, x_n) \in \mathbb{R}^n \mid u_1x_1 + \cdots + u_nx_n \le u_0\}.$$

It is clear that antipodal points of S^n correspond to opposite half-spaces. For $u = (\pm 1, 0, \dots, 0) \in S^n$, we have by the same formula

$$h^+((+1, 0, \dots, 0)) = \mathbb{R}^n,$$
$$h^+((-1, 0, \dots, 0)) = \emptyset.$$

Define a continuous function $f\colon S^n \to \mathbb{R}^n$ where the i-th component is

$$f_i(u) \coloneqq \mu_i(h^+(u)).$$

By the Borsuk–Ulam theorem, there exists $x \in S^n$ such that $f(-x) = f(x)$. Then the boundary of the half space $h^+(x)$ is the desired hyperplane. □

Let us discuss the proof of the Borsuk–Ulam theorem. For $n = 1$, the theorem follows easily from the intermediate value theorem. One can prove

the $n = 2$ case using some basic knowledge of *fundamental groups* of topological spaces. We will be discussing this in more details in later subsections.

For $n \geq 3$, the proofs usually are more involved (we will only discuss the case of $n = 2$ later); let us sketch a proof here.

- Assume the contrary that there exists an antipodal map $f\colon S^n \to S^{n-1}$. This descends to a continuous map $g\colon \mathbb{RP}^n \to \mathbb{RP}^{n-1}$. Here $\mathbb{RP}^n \cong S^n/\mathbb{Z}_2$ is the n-dimensional *real projective space*.
- One can show that such g induces an isomorphism $g_*\colon \pi_1(\mathbb{RP}^n) \to \pi_1(\mathbb{RP}^{n-1})$ between the *fundamental groups*.
- By the *Poincaré–Hurewicz theorem*, we have an isomorphism $g_*\colon H_1(\mathbb{RP}^n, \mathbb{Z}) \to H_1(\mathbb{RP}^{n-1}, \mathbb{Z})$ between the *homology groups*.
- By the *universal coefficient theorem*, we have an induced *ring homomorphism* between the *cohomology rings*

  $$\mathbb{F}_2[b]/b^n \cong H^*(\mathbb{RP}^{n-1}, \mathbb{F}_2) \xrightarrow{g^*} H^*(\mathbb{RP}^n, \mathbb{F}_2) \cong \mathbb{F}_2[a]/a^{n+1}$$

  which sends $b \mapsto a$. But then we get that $b^n = 0$ is sent to $a^n \neq 0$, a contradiction.

Remark 3.7. The real projective space $\mathbb{RP}^n$ is the topological space that parametrizes the 1-dimensional subspaces of $\mathbb{R}^{n+1}$. It can be defined by quotienting the scaling action:

$$\mathbb{RP}^n = (\mathbb{R}^{n+1}\backslash\{0\})/\mathbb{R}^*.$$

Thus $\mathbb{RP}^n$ can also be formed by identifying antipodal points of S^n. It is a smooth compact manifold, and is a special case of *Grassmannians* $\mathrm{Gr}(k, n+1)$ which parametrizes the k-dimensional subspaces of $\mathbb{R}^{n+1}$.

In the following, we will introduce the notion of *fundamental groups* of topological spaces, and prove the Borsuk–Ulam theorem for $n = 2$. A nice reference in which you can find all these notions mentioned above is a book by Hatcher [10].

3.2. **Fundamental groups.** Let us start with recalling the definition of *topological spaces* and *continuous maps* between them.

Definition 3.8. A *topology* on a set X is a collection τ of subsets of X satisfying the following axioms:

- The empty set and X itself belong to τ.

- Any arbitrary (finite or infinite) union of members of τ belongs to τ.
- The intersection of any finite number of members of τ belongs to τ.

Members of τ are called *open subsets* of X (with respect to this topology).

Definition 3.9. A map $f\colon X \to Y$ between topological spaces is called *continuous* if

$$U \subseteq Y \text{ is an open subset } \implies f^{-1}(U) \subseteq X \text{ is an open subset.}$$

The map f is called a *homeomorphism* if it is bijective, and both f and f^{-1} are continuous. In this case, X and Y are said to be *homeomorphic*.

The *fundamental groups* of topological spaces will be defined in terms of *loops* and their deformations.

Definition 3.10. Let X be a topological space.

- A *path* in X is a continuous map $\gamma\colon I \to X$ where $I = [0,1]$.
- Its *inverse path* $\gamma^{-1}\colon I \to X$ is defined by $\gamma^{-1}(t) = \gamma(1-t)$.
- A path is called a *loop* if $\gamma(0) = \gamma(1)$. It can be considered as a map $\gamma\colon S^1 \to X$, with *basepoint* $x_0 = \gamma(0) = \gamma(1)$.
- If $\gamma(t) = x_0 \in X$ for all $t \in [0,1]$, then such γ is called a *constant path*, and denoted by i_{x_0}.
- If γ_1 and γ_2 are two loops satisfying $\gamma_1(1) = \gamma_2(0)$, we define their *composition* or *product path* to be

$$(\gamma_1 \cdot \gamma_2)(s) = \begin{cases} \gamma_1(2s), & 0 \le s \le 1/2 \\ \gamma_2(2s-1), & 1/2 \le s \le 1 \end{cases}$$

Definition 3.11. Two paths γ_0, γ_1 with the same endpoints x_0, x_1 are called *homotopic* if there exists a continuous map $F\colon I \times I \to X$ such that

- $F(s,0) = \gamma_0(s)$ and $F(s,1) = \gamma_1(s)$ for all $s \in [0,1]$.
- $F(0,t) = x_0$ and $F(1,t) = x_1$ for all $t \in [0,1]$.

In this case, we will denote $\gamma_0 \simeq \gamma_1$.

Example 3.12. Any two paths γ_0, γ_1 in $\mathbb{R}^n$ having the same endpoints x_0, x_1 are homotopic via the linear homotopy $F(s,t) = (1-t)\gamma_0(s) + t\gamma_1(s)$.

Exercise. The relation of homotopy on paths with fixed endpoints is an *equivalence relation*, i.e.

- $\gamma \simeq \gamma$.

- If $\gamma_1 \simeq \gamma_2$, then $\gamma_2 \simeq \gamma_1$.
- If $\gamma_1 \simeq \gamma_2$ and $\gamma_2 \simeq \gamma_3$, then $\gamma_1 \simeq \gamma_3$.

We denote the homotopy class of γ as $[\gamma]$.

Exercise. Let $\gamma_1, \gamma_2, \beta_1, \beta_2$ be paths in X. Suppose $\gamma_1 \simeq \gamma_2$, $\beta_1 \simeq \beta_2$, and $\gamma_1(1) = \gamma_2(1) = \beta_1(0) = \beta_2(0)$. Prove that $\gamma_1 \cdot \beta_1 \simeq \gamma_2 \cdot \beta_2$.

This shows that the *composition* (or *product*) can be defined on homotopy classes:

$$[\gamma] \cdot [\beta] \coloneqq [\gamma \cdot \beta].$$

Exercise. This exercise shows that the product on homotopy classes has *associativity*. Let $\gamma_1, \gamma_2, \gamma_3$ be paths in X satisfying $\gamma_1(1) = \gamma_2(0)$ and $\gamma_2(1) = \gamma_3(0)$. Prove that

$$([\gamma_1] \cdot [\gamma_2]) \cdot [\gamma_3] = [\gamma_1] \cdot ([\gamma_2] \cdot [\gamma_3]).$$

Note that the equality is not true without considering their homotopy classes: $(\gamma_1 \cdot \gamma_2) \cdot \gamma_3 \neq \gamma_1 \cdot (\gamma_2 \cdot \gamma_3)$ in general.

Exercise. Let γ be a path from x_0 to x_1 in X. Prove that

$$[\gamma] \cdot [\gamma^{-1}] = [i_{x_0}], \quad [\gamma^{-1}] \cdot [\gamma] = [i_{x_1}], \quad [\gamma] \cdot [i_{x_1}] = [\gamma] = [i_{x_0}] \cdot [\gamma].$$

We are now ready to define the fundamental group.

Definition 3.13. The *fundamental group* of X at the basepoint x_0, denoted by $\pi_1(X, x_0)$, is defined to be the set of all homotopy classes $[\gamma]$ of loops $\gamma \colon I \to X$ with basepoint x_0, where

- the group structure given by the product $[\gamma_1] \cdot [\gamma_2] = [\gamma_1 \cdot \gamma_2]$,
- the identity element is $[i_{x_0}]$,
- the inverse of an element $[\gamma]$ is given by $[\gamma^{-1}]$.

Example 3.14. Hold a mug in your hand. Now, without letting go of the mug and without spilling the coffee, see if you can rotate the mug *two full turns* and return your hand, arm, and cup to their original positions. If you can do that, can you do the same trick with only *one* full turn? (*No!*)
Continuously rotating a mug is equivalent to following a path in SO(3), the space of rotations in $\mathbb{R}^3$, and if you start and end the mug in the same orientation, you have traced a loop in SO(3). The reason this trick works for 2 twists but not 1 twist could be explained by the fact that $\pi_1(\mathrm{SO}(3)) \cong \mathbb{Z}/2\mathbb{Z}$.

Proposition 3.15. *Suppose X is path-connected, i.e. for any two points $x_0, x_1 \in X$, there exists a path $\gamma\colon I \to X$ such that $\gamma(0) = x_0$ and $\gamma(1) = x_1$. Then the isomorphic class of the fundamental group $\pi_1(X, x_0)$ is independent of the choice of the basepoint x_0, i.e. for any two points $x_0, x_1 \in X$ we have $\pi_1(X, x_0) \cong \pi_1(X, x_1)$.*

Proof. Let γ be a path connecting x_0 and x_1. It is easy to check that

$$\pi_1(X, x_0) \to \pi_1(X, x_1); \qquad [\beta] \mapsto [\gamma^{-1}] \cdot [\beta] \cdot [\gamma]$$

and

$$\pi_1(X, x_1) \to \pi_1(X, x_0); \qquad [\beta] \mapsto [\gamma] \cdot [\beta] \cdot [\gamma^{-1}]$$

are group homomorphisms inverse with each other. Thus $\pi_1(X, x_0) \cong \pi_1(X, x_1)$. □

Proposition 3.16. *A continuous map $f\colon X \to Y$ induces a group homomorphism*

$$f_*\colon \pi_1(X, x_0) \to \pi_1(Y, f(x_0)); \qquad [\gamma] \mapsto [f \circ \gamma].$$

Proof. One can verify that the map preserves homotopy equivalences and compositions. The proposition then follows easily. □

3.3. **Fundamental group of a circle and applications.** Consider the circle

$$S^1 = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\} = \{(\cos(2\pi s), \sin(2\pi s)) \in \mathbb{R}^2 \mid s \in \mathbb{R}\}$$

and choose a basepoint $x_0 = (1, 0) \in S^1$.

Theorem 3.17. *The fundamental group $\pi_1(S^1, x_0) \cong \mathbb{Z}$ is an infinite cyclic group generated by the homotopy class of the loop $\omega(s) = (\cos(2\pi s), \sin(2\pi s))$.*

Note that $[\omega]^n = [\omega_n]$ where $\omega_n(s) = (\cos(2\pi ns), \sin(2\pi ns))$ for all $n \in \mathbb{Z}$. The theorem is therefore equivalent to the statement that every loop in S^1 based at $(1, 0)$ is homotopic to ω_n for a unique $n \in \mathbb{Z}$.

The main idea is to compare paths in S^1 with paths in $\mathbb{R}$ via the map

$$p\colon \mathbb{R} \to S^1; \qquad s \mapsto (\cos(2\pi s), \sin(2\pi s)).$$

Consider the path $\widetilde{\omega_n}(s) = ns$ in $\mathbb{R}$, which starts at 0 and ends at ns. The relation $\omega_n = p\widetilde{\omega_n}$ is expressed by saying that $\widetilde{\omega_n}$ is a *lift* of ω_n.

Definition 3.18. Let X be a topological space. A *covering space* of X consists of a space $\widetilde{X}$ and a map $p\colon \widetilde{X} \to X$ such that: for each point $x \in X$ there is an open neighborhood U of x such that $p^{-1}(U)$ is a union of disjoint open sets each of which is mapped homeomorphically onto U by p.

Example 3.19. Here are some basic examples of covering spaces of S^1.

- The map $p\colon \mathbb{R} \to S^1$ where $s \mapsto (\cos(2\pi s), \sin(2\pi s))$ is a covering map.
- The map $S^1 \to S^1$ where $(\cos(2\pi s), \sin(2\pi s)) \mapsto (\cos(2\pi n s), \sin(2\pi n s))$ is a covering map for any nonzero integer n. In terms of complex numbers, the map can be expressed as $z \mapsto z^n$.

Exercise. Below are two basic (yet important) facts about covering spaces $p\colon \widetilde{X} \to X$.

(a) For each path $f\colon I \to X$ starting at a point $x_0 \in X$ and each $\widetilde{x_0} \in p^{-1}(x_0)$, there is a unique lift $\widetilde{f}\colon I \to \widetilde{X}$ of f starting at $\widetilde{x_0}$.

(b) For each homotopy $F\colon I \times I \to X$ starting at a point $x_0 \in X$ and each $\widetilde{x_0} \in p^{-1}(x_0)$, there is a unique lifted homotopy $\widetilde{F}\colon I \times I \to \widetilde{X}$ of f starting at $\widetilde{x_0}$.

Proof of Theorem 3.17. Let $f\colon I \to S^1$ be a loop at the basepoint $x_0 = (1, 0)$. We would like to show that it is homotopic to ω_n for a unique $n \in \mathbb{Z}$. By (a) there is a unique lift $\widetilde{f}$ of the loop f starting at 0. Note that the path $\widetilde{f}$ ends at some integer n since $p\widetilde{f}(1) = f(1) = x_0$. Recall that $\widetilde{f}$ and $\widetilde{\omega_n}$ are homotopic since they can be linearly homotopic with each other in $\mathbb{R}$. Thus $[f] = [\omega_n]$.

To show that n is uniquely determined by $[f]$, suppose there is $\omega_m \simeq \omega_n$ for some $m, n \in \mathbb{Z}$. Let F be a homotopy from ω_m to ω_n. By (b) it lifts to a homotopy $\widetilde{F}$ starting at 0, therefore the endpoints of $\widetilde{\omega_m}$ and $\widetilde{\omega_n}$ coincide. Hence $m = n$. □

Remark 3.20. For a covering space $p\colon \widetilde{X} \to X$, a homeomorphism $d\colon \widetilde{X} \to \widetilde{X}$ is called a *deck transformation* if $p \circ d = p$. Together with the composition of maps, the set of deck transformation forms a group $\mathrm{Deck}(p)$. For instance, for the n-sheeted covering space $S^1 \to S^1$ given by $z \mapsto z^n$, the deck transformations are the rotations of S^1 through angles that are multiples of $2\pi/n$, so the deck transformation group is $\mathbb{Z}/n\mathbb{Z}$. Similarly, the deck transformation group of the covering space $\mathbb{R} \to S^1$ is isomorphic to $\mathbb{Z} \cong \pi_1(S^1)$.

The covering space $p\colon \mathbb{R} \to S^1$ where $s \mapsto (\cos(2\pi s), \sin(2\pi s))$ is the *universal cover* of S^1: any covering space of S^1 can be covered by the universal cover. For instance, the covering space $S^1 \xrightarrow{z^n} S^1$ can be covered by $p_n\colon \mathbb{R} \to S^1$ where $s \mapsto (\cos(2\pi s/n), \sin(2\pi s/n))$; we have $z^n \circ p_n = p$. The deck transformation group of p_n is given by $n\mathbb{Z}$. In general, there is a one-to-one correspondence:

$$\{\text{covering space of } X\} \leftrightarrow \{\text{subgroups of } \pi_1(X)\}$$

where a covering space $p\colon \widetilde{X} \to X$ corresponds to the subgroup $p_*(\pi_1(\widetilde{X}))$ of $\pi_1(X)$. Moreover, the deck transformation group of p is isomorphic to $N(p_*(\pi_1(\widetilde{X})))/p_*(\pi_1(\widetilde{X}))$, where $N(p_*(\pi_1(\widetilde{X})))$ is the normalizer subgroup of $p_*(\pi_1(\widetilde{X}))$ in $\pi_1(X)$.

Theorem 3.21 (Borsuk–Ulam in dimension 2)**.** *There is no antipodal map* $f\colon S^2 \to S^1$.

Proof. Assume the contrary that such map f exists. Define a loop η circling the equator

$$\eta\colon I \to S^2; \qquad s \mapsto (\cos(2\pi s), \sin(2\pi s), 0),$$

and consider the loop $g = f \circ \eta\colon I \to S^1$.

On the one hand, the loop η in S^2 is homotopic to a constant map, thus so is the loop g in S^1. In other words, $[g] = 0$ in $\pi_1(S^1) \cong \mathbb{Z}$.

On the other hand, since $f(-x) = -f(x)$, we have

$$g\left(s + \frac{1}{2}\right) = -g(s) \qquad \text{for all } s \in \left[0, \frac{1}{2}\right].$$

Let $\widetilde{g}\colon I \to \mathbb{R}$ be a lift of g. Then for each $s \in \left[0, \frac{1}{2}\right]$ we have

$$\widetilde{g}\left(s + \frac{1}{2}\right) = \widetilde{g}(s) + \frac{q}{2} \qquad \text{for some odd integer } q.$$

Note that q depends continuously on $s \in \left[0, \frac{1}{2}\right]$, so it must be a constant for all $s \in \left[0, \frac{1}{2}\right]$ since it is of integer value. In particular, we have

$$\widetilde{g}(1) = \widetilde{g}(0) + q.$$

Thus $[g] \neq 0$ in $\pi_1(S^1) \cong \mathbb{Z}$ since q is odd. Contradiction. □

Theorem 3.22 (Fundamental theorem of algebra)**.** *Every non-constant polynomial with complex coefficients has a root in* $\mathbb{C}$*.*

Proof. Consider a complex polynomial $p(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_0$. Assume the contrary that $p(z)$ has no roots in $\mathbb{C}$, then for each $r \geq 0$

$$f_r(s) = \frac{p(re^{2\pi is})/p(r)}{|p(re^{2\pi is})/p(r)|}$$

defines a loop in S^1 based at 1. As r varies, f_r is a homotopy of loops in S^1 based at 1. Since f_0 is the trivial loop, we have $[f_r] = 0$ in $\pi_1(S^1)$ for all $r \geq 0$.

On the other hand, for r sufficiently large, on the circle $|z| = r$ we have

$$|z^n| > (|a_0| + \cdots + |a_{n-1}|)|z^{n-1}| \geq |a_{n-1}z^{n-1} + \cdots + a_0|.$$

Thus the polynomial $p_t(z) = z^n + t(a_{n-1}z^{n-1} + \cdots + a_0)$ has no zero on the circle $|z| = r$ when $0 \leq t \leq 1$. Replacing p by p_t in the formula above and letting t go from 1 to 0, one obtains a homotopy from the loop f_r to the loop $\omega_n(s) = e^{2\pi ins}$, thus $[f_r] = [\omega_n]$ in $\pi_1(S^1)$. We then conclude that $n = 0$. □

3.4. **The rectangular peg problem.** Let $C \subseteq \mathbb{R}^2$ be a continuous simple closed curve. Does there always exist four points on C such that they form the vertices of a rectangle? Below is the sketch of ideas toward answering this question (affirmatively).

Lecture 4

- Denote M the *moduli space* of unordered pairs of points in C: each (unordered) pair of points c_1, c_2 in C corresponds to a unique point in M.
- Observe that M is naturally topologically equivalent to a Möbius strip, where its boundary can be identified with the curve C.
- Define a continuous function $f_C \colon M \to \mathbb{R}^3$ which sends a pair of points $c_1 = (x_1, y_1), c_2 = (x_2, y_2)$ on the curve C to the point

$$\left(\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2}, \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}\right) \in \mathbb{R}^3$$

  where the first two coordinates give the midpoint of c_1, c_2, and the third coordinate is the distance between c_1 and c_2.
- Observe that the rectangular peg problem has an affirmative answer for a curve C if and only if f_C is not injective.

- Observe that one gets the *real projective plane* $\mathbb{RP}^2$ by gluing the Möbius strip with a disk along their boundaries.
- Assume the contrary that there exists a curve C such that f_C is injective. Then one gets an embedding of the real projective plane $\mathbb{RP}^2$ into $\mathbb{R}^3$.
- Use topological tools to show that there is no embedding of $\mathbb{RP}^2$ into $\mathbb{R}^3$. This concludes the proof.

One way to show the last statement, namely there is no embedding of $\mathbb{RP}^2$ into $\mathbb{R}^3$, is by consider the *orientability* of the real projective plane $\mathbb{RP}^2$. It is known that $\mathbb{RP}^2$ is *non-orientable*: this can be rigorously proved by computing the homology groups of $\mathbb{RP}^2$. On the other hand, assume the contrary that there exists an embedding of $\mathbb{RP}^2$ into $\mathbb{R}^3$, then the image would bound a compact region in $\mathbb{R}^3$ (by the *generalized Jordan curve theorem*). The outward-pointing normal vector field would then give an orientation of $\mathbb{RP}^2$. Contradiction.

4. Algebra

Which positive integers n can be written as the sum of two squares? To answer this question, it is convenient to consider the factorization in the *ring* of *Gaussian integers* $\mathbb{Z}[i]$:

$$n = x^2 + y^2 = (x + iy)(x - iy).$$

One would also like to study other number rings; for instance, to understand the Diophantine equation $n = x^2 - 5y^2$, one would like to do factorizations in the ring $\mathbb{Z}[\sqrt{5}]$.

It is important to be aware that not all number rings have the same properties. For instance, the ring of Gaussian integers $\mathbb{Z}[i]$ is a *Unique Factorization Domain (UFD)*, but the ring $\mathbb{Z}[\sqrt{5}]$ is not: there are factorizations

$$(3 + \sqrt{5})(3 - \sqrt{5}) = 4 = 2 \cdot 2$$

where $3 \pm \sqrt{5}$ and 2 are all *irreducible* elements of $\mathbb{Z}[\sqrt{5}]$, so there are two truly different factorizations of 4 in $\mathbb{Z}[\sqrt{5}]$.

We will begin our discussions with the general notion of *rings*, then gradually specialized to commutative rings, integral domains, unique factorization domains, principal ideal domains, Euclidean domains. It turns out that the

ring of Gaussian integers $\mathbb{Z}[i]$ is an *Euclidean domain* (a condition stronger than UFD), which will allow us to completely classify the integers that can be written as the sum of two squares. A nice reference for this part (and abstract algebra in general) is a book of Artin [1].

4.1. **Rings.**

Definition 4.1. A *ring* is a set R equipped with two binary operations $+$ (addition) and $\cdot$ (multiplication) satisfying:

(1) R is an abelian group under addition, namely:
   - $(a+b)+c = a+(b+c)$ for all $a, b, c \in R$.
   - $a+b = b+a$ for all $a, b \in R$.
   - There is an element $0 \in R$ such that $a+0 = a$ for all $a \in R$.
   - For each $a \in R$ there exists $-a \in R$ such that $a+(-a) = 0$.

(2) R is a monoid under multiplication, namely:
   - $(a \cdot b) \cdot c = a \cdot (b \cdot c)$.
   - There is an element $1 \in R$ such that $a \cdot 1 = a = 1 \cdot a$ for all $a \in R$.

(3) Multiplication is distributive with respect to addition, namely:
   - $a \cdot (b+c) = a \cdot b + a \cdot c$ for all $a, b, c \in R$.
   - $(b+c) \cdot a = b \cdot a + c \cdot a$ for all $a, b, c \in R$.

Note that the multiplication symbol $\cdot$ is often omitted: for instance, ab means $a \cdot b$.

Definition 4.2. A ring R is said to be *commutative* if $ab = ba$ for all $a, b \in R$.

Non-example. The set of 2×2 real matrices forms a ring under the standard matrix additions and multiplications. It is not commutative.

Remark 4.3. Whether a ring is commutative has profound implications on its behavior. *Commutative algebra*, the theory of commutative rings, is a major branch of ring theory. Its development has been greatly influenced by problems and ideas of *algebraic number theory* and *algebraic geometry*. If you are interested, a standard textbook on commutative algebra is [2].

Commutative rings resemble familiar number systems, and various definitions for commutative rings are designed to formalize properties of the integers.

Definition 4.4. A nonzero commutative ring R is called an *integral domain* if the product of any two nonzero elements is nonzero.

Non-example. The quotient ring $\mathbb{Z}/6\mathbb{Z}$ is a commutative ring, but is not an integral domain.

Non-example. The quotient ring $\mathbb{Z}[x]/(x^2-1)$ is a commutative ring, but is not an integral domain.

In order to introduce the definition of unique factorization domain, we need to define the notion of *units*.

Definition 4.5. An element $u \in R$ is called a *unit* if there exists $v \in R$ such that $uv = vu = 1$. In other words, a unit is an invertible element for the multiplication of the ring.

Example 4.6. Here are some basic examples:

- The units of $\mathbb{Z}$ are 1 and -1.
- The units of $\mathbb{Z}[i]$ are 1, -1, i, and $-i$.
- The units of $M_2(\mathbb{R})$ are all invertible matrices.
- The ring $\mathbb{Z}[\sqrt{3}]$ has infinitely many units: for instance, $(2+\sqrt{3})$ and its powers are units of the ring. In general, the ring of integers in a number field can be determined by the *Dirichlet's unit theorem*.

Definition 4.7. An element of an integral domain R is called *irreducible* if it is not a unit, and is not the product of two non-unit elements.

Remark 4.8. An element of an integral domain R is called *prime* if, whenever $a \mid bc$ (i.e. $bc = ax$ for some $x \in R$), then $a \mid b$ or $a \mid c$. In an integral domain, every prime element is irreducible, but the converse is not true in general. For instance, in the ring $\mathbb{Z}[\sqrt{-5}]$, it can be shown that 3 is irreducible. However, it is not a prime element since

$$3 \mid (2+\sqrt{-5})(2-\sqrt{-5}) = 9$$

but 3 does not divide either of the two factors.

Definition 4.9. An integral domain R is said to be a *unique factorization domain* (or UFD for short) if every nonzero element $x \in R$ can be written as a product

$$x = up_1 \cdots p_n$$

where u is a unit and p_i's are irreducible, and this representation is unique in the following sense: If we also have

$$x = vq_1 \cdots q_m$$

where v is a unit and q_i's are irreducible, then $m = n$, and there exists a bijective map $\sigma\colon \{1, \dots, n\} \to \{1, \dots, n\}$ such that $p_i = w_i q_{\sigma(i)}$ for some units w_i.

Non-example. The quadratic ring $\mathbb{Z}[\sqrt{-5}]$ is an integral domain, but is not a UFD:

$$2 \cdot 3 = 6 = (1 + \sqrt{-5})(1 - \sqrt{5}).$$

One can show that $2, 3, 1 + \sqrt{-5}, 1 - \sqrt{-5}$ are all irreducible, and the only units of $\mathbb{Z}[\sqrt{-5}]$ is ± 1, therefore these truly are two different factorizations.

One important class of examples of UFDs are given by *principal ideal domains* (PID).

Definition 4.10. An ideal I of a commutative ring R is an additive subgroup of R which is closed under multiplications: more precisely,

- $(I, +)$ is a subgroup of $(R, +)$.
- For every $r \in R$ and $x \in I$, the product rx is in I.

An ideal is called *principal* if it can be generated by a single element, i.e. it is of the form $xR = \{xr \mid r \in R\}$.

Definition 4.11. An integral domain R is called a *principal ideal domain* (PID) if every ideal of R is principal.

Non-example. $\mathbb{Z}[x]$ is a UFD, but is not a PID: for instance, the ideal $\langle 2, x \rangle$ can not be generated by a single polynomial.

Theorem 4.12. *Every PID is a UFD.*

Proof. Let R be a PID. First, we show that R satisfies the *ascending chain condition* (ACC) on ideals; namely, whenever there are ideals

$$I_1 \subseteq I_2 \subseteq \cdots \subseteq I_n \subseteq \cdots$$

then there is some $N > 0$ such that $I_n = I_N$ for all $n \geq N$. Consider the union

$$I = \bigcup_{n \geq 1} I_n$$

which is also an ideal of R. Thus $I = (a)$ for some $a \in I$, and there exists $N > 0$ such that $a \in I_N$. This shows that R satisfies ACC.

Second, we show that every irreducible elements of R is prime. Let $a \in R$ be an irreducible element. Suppose $a \mid bc$ for some $b, c \in R$. We would like

to show that $a \mid b$ or $a \mid c$ holds. Let us consider the ideal (a, b). Since R is PID, there exists $x \in R$ such that $(x) = (a, b)$. In particular, $a = xy$ for some $y \in R$. Since a is irreducible, x or y has to be a unit.

- If y is a unit, then $(a) = (x) = (a, b)$, thus $a \mid b$ as desired.
- If x is a unit, then $(1) = (x) = (a, b)$, so there exists $c, d \in R$ such that $ac+bd = 1$. Multiplying both sides with c, one gets $ac^2+bcd = c$. Note that the left hand side is a multiple of a since $a \mid bc$, thus we obtain $a \mid c$.

Now we are ready to show that R is a UFD. First, we show that any nonzero nonunit element of R can be written as a product of irreducible elements. Assume the contrary that there exists nonzero nonunit element of R that cannot be written as a product of irreducibles. Denote the collection of such elements by S. Since R satisfies ACC, there exists $r \in S$ such that $(r) \nsubseteq (s)$ for any $s \in S\backslash\{r\}$. In particular, r is not irreducible, so it can be written as $r = xy$ for some nonunit elements $x, y \in R$. Since $(r) \subseteq (x)$ and $(r) \subseteq (y)$, we have $x, y \notin S$, therefore x and y both can be written as a product of irreducibles. But then we get $r = xy$ can also be written as a product of irreducibles. Contradiction.

Finally, we show that the factorization is unique. Suppose

$$a = up_1 \cdots p_n = vq_1 \cdots q_m$$

where u, v are units and p_i, q_i's are irreducibles (therefore are primes by what we proved earlier). Then $p_1 \mid vq_1 \cdots q_m$, thus it must divide some q_j. Since p_1 and q_j are both primes, they are the same up to a unit. We may continue this process and match each prime factor on both sides. □

Definition 4.13. An integral domain R is said to be a *Euclidean domain* if there exists a function $N \colon R\backslash\{0\} \to \mathbb{Z}_{\geq 0}$ (called a *norm function*) such that:

- For all nonzero elements $a, b \in R$, there exists $q, r \in R$ such that $a = qb + r$ and either $r = 0$ or $N(r) < N(b)$.
- For all nonzero elements $a, b \in R$ we have $N(a) \leq N(ab)$.

Non-example. The ring $\mathbb{Z}\left[\frac{1+\sqrt{-19}}{2}\right]$ is a PID, but is not a Euclidean domain.

Example 4.14. Here are some basic examples of Euclidean domains.

- The ring of integers $\mathbb{Z}$, with $N(a) = |a|$.

- The ring of Gaussian integers $\mathbb{Z}[i]$, with $N(a+ib) = a^2+b^2$ (we will discuss more details later).
- The ring of polynomials $\mathbb{R}[x]$ over $\mathbb{R}$ (can be replaced by any *field*), with $N(P) = \deg(P)$.

Theorem 4.15. *Every Euclidean domain is a PID.*

Proof. Let R be a Euclidean domain. Let $I \subseteq R$ be a nonzero ideal. Then there exists a nonzero element $a \in I$ such that $N(a)$ is minimal among all elements of the ideal. We claim that $I = (a)$. For any $b \in I$, there exists $q, r \in R$ such that $b = qa + r$ where $r = 0$ or $N(r) < N(a)$. Since $a, b \in I$, we have $r \in I$, thus $N(r) \geq N(a)$ by the minimality. Therefore we have $r = 0$ and $b \in (a)$. □

4.2. **Ring of Gaussian integers.**

Definition 4.16. The norm function on the ring of Gaussian integers $\mathbb{Z}[i]$ is defined to be

$$N(a+ib) = (a+ib)(a-ib) = a^2+b^2.$$

Exercise. Here are some basic properties of the norm function.

- $N(\alpha) = 0$ if and only if $\alpha = 0$.
- $N(\alpha\beta) = N(\alpha)N(\beta)$ for all $\alpha, \beta \in \mathbb{Z}[i]$.
- $N(\alpha) = 1$ if and only if α is a unit of $\mathbb{Z}[i]$.
- $\{1, -1, i, -i\}$ are the only units of $\mathbb{Z}[i]$.

Theorem 4.17. $\mathbb{Z}[i]$ *is a Euclidean domain.*

Proof. Let a, b be nonzero elements of $\mathbb{Z}[i]$. Observe that the set $b\mathbb{Z}[i]$ forms a lattice of squares with side length $|b| = \sqrt{N(b)}$. Then the distance between a and the lattice point closest to it (say bq) is no bigger than $|b|/\sqrt{2}$. Let $r = a - bq \in \mathbb{Z}[i]$. Then

$$N(r) = |r|^2 \leq \frac{|b|^2}{2} = \frac{N(b)}{2} < N(b).$$

□

Lemma 4.18. *If $\pi \in \mathbb{Z}[i]$ is such that $N(\pi)$ is a prime number, then π is a prime in $\mathbb{Z}[i]$.*

Proof. If $\pi = \alpha\beta$ in $\mathbb{Z}[i]$, then $N(\pi) = N(\alpha)N(\beta)$. So either $N(\alpha)$ or $N(\beta)$ is 1, which means that either α or β is a unit. □

Lemma 4.19. *Let q be a prime number with $q = 3$ (mod 4). Then q is a prime in $\mathbb{Z}[i]$.*

Proof. If $q = \alpha\beta$ in $\mathbb{Z}[i]$, then $q^2 = N(\alpha)N(\beta)$. Note that $q = N(\alpha) = a^2 + b^2$ is impossible since $q = 3 \pmod 4$. Thus either $N(\alpha)$ or $N(\beta)$ is 1. □

Lemma 4.20. *Let p be a prime number with $p = 1$ (mod 4). Then there exists a Gaussian prime π such that $p = \pi\overline{\pi}$.*

Proof. First, we claim that there exists an integer $c \in \mathbb{Z}$ such that $c^2 = -1 \pmod p$. This can be easily proved by assuming the fact that the multiplicative group $\mathbb{Z}_p^*$ of the finite field $\mathbb{Z}_p$ is cyclic. Let a be a generator of the multiplicative group $\mathbb{Z}_p^*$ (which has $p-1$ elements), i.e.

$$\mathbb{Z}_p^* = \{1, a, a^2, \ldots, a^{p-2}\}.$$

Observe that -1 is the unique order two element of $\mathbb{Z}_p^*$, thus $a^{\frac{p-1}{2}} = -1 \pmod p$. The claim then follows from the assumption that $p = 1 \pmod 4$.

By the claim, we have $p \mid (c+i)(c-i)$ in $\mathbb{Z}[i]$. It is easy to show that p does not divide $c+i$ or $c-i$. Therefore p is *not* a Gaussian prime. Hence there exists nonunit elements $\alpha, \beta \in \mathbb{Z}[i]$ such that $p = \alpha\beta$. By comparing the norms on both sides, we obtain $N(\alpha) = N(\beta) = p$. Therefore both α and β are Gaussian primes. It is then easy to check that they are complex conjugate with each other. □

Lecture 5

Proposition 4.21. *Up to multiplying by units, all the Gaussian primes are the following:*

- *$1+i$ (which is of norm 2),*
- *π and $\overline{\pi}$, where $p = \pi\overline{\pi}$ is a prime number with $p = 1$ (mod 4) (the norms of π and $\overline{\pi}$ are both p),*
- *q, where q is a prime number with $q = 3$ (mod 4) (which is of norm q^2).*

Proof. Let α be a Gaussian prime. Then we can find a Gaussian prime π in the above list so that $\pi \mid N(\alpha) = \alpha\overline{\alpha}$. So either π or $\overline{\pi}$ divides α. Thus α is also in the above list. □

4.3. **Applications.** Let us apply the arithmetic of $\mathbb{Z}[i]$ to solve a classic problem: finding all *Pythagorean triples*. A Pythagorean triples is $(x, y, z) \in \mathbb{Z}^3_{>0}$ where $x^2 + y^2 = z^2$. It suffices to only look for *primitive* Pythagorean triples, i.e. $\gcd(x, y, z) = 1$. Also, observe that x and y cannot both be odd, so may assume that x is odd and y is even.

Theorem 4.22. *Let $(x, y, z) \in \mathbb{Z}^3_{>0}$ be a primitive Pythagorean triples with x odd and y even. Then there exists coprime integers a, b with $a > b > 0$ and $a \neq b \pmod 2$ such that*

$$x = a^2 - b^2, \quad y = 2ab, \quad z = a^2 + b^2.$$

Proof. Let $\alpha = x + iy \in \mathbb{Z}[i]$, so $N(\alpha) = x^2 + y^2 = z^2$. The idea is to show that α is a *square* in $\mathbb{Z}[i]$; writing $\alpha = (a + ib)^2$ gives the desired result. We have

$$z^2 = N(\alpha) = (x + iy)(x - iy).$$

We claim that $x + iy$ and $x - iy$ are coprime in $\mathbb{Z}[i]$. Assume the contrary that there exists a Gaussian integer π that divides both $x + iy$ and $x - iy$. Then it also divides $2x$ and $2y$. Since x, y are coprime, π has to divide 2. Therefore $\pi = 1 + i$ (up to a unit). But $1 + i$ does not divide $x + iy$ since $x \neq y \pmod 2$. Contradiction.

Hence $x + iy$ and $x - iy$ are coprime in $\mathbb{Z}[i]$. As their product is a square, unique factorization in $\mathbb{Z}[i]$ implies that each of them is a square (up to a unit). Using $-1 = i^2$, each of them must be a square or i times a square.

If $x + iy = i(a + ib)^2$, then $x = -2ab$ which contradicts with the assumption that x is odd. Therefore $x + iy$ is a square. □

Next, we solve the sum of two squares problem.

Theorem 4.23. *Let $n = a \cdot b^2$ be an integer with a square-free. Then n can be written as a sum of two squares if and only if no prime $q = 3 \pmod 4$ divides a.*

Proof. The "if" part: For each prime p dividing a, there is a Gaussian prime π_p such that $p = \pi_p \overline{\pi}_p$. Let $x + iy = b \cdot \prod_{p|a} \pi_p$. Then $x^2 + y^2 = n$.

The "only if" part: Suppose $n = x^2 + y^2 = (x + iy)(x - iy)$. If a prime $q = 3 \pmod 4$ divides n, as it is a Gaussian prime, it divides $x + iy$ or $x - iy$, which implies that q divides both $x + iy$ and $x - iy$. Thus q^2 divides n. The statement can then be proved by induction on b. □

In the upcoming section, we will use the theory of *modular forms* to count the number

$$r_2(n) = \#\{(x_1, x_2) \in \mathbb{Z}^2 \mid x_1^2 + x_2^2 = n\}.$$

Here is a sketch of the main idea. One can show that

$$E_1^\chi(q) = \frac{1}{4} + \sum_{n=1}^{\infty} \left(\sum_{d|n} \chi(d) \right) q^n \in M_1(\Gamma_1(4)), \qquad \text{where } \chi(d) = \begin{cases} 1 & \text{if } d = 1 \pmod 4 \\ -1 & \text{if } d = 3 \pmod 4 \\ 0 & \text{if } d \text{ is even} \end{cases}$$

and the space $M_1(\Gamma_1(4))$ of modular form of weight 1 for the group $\Gamma_1(4) \subseteq \mathrm{SL}(2, \mathbb{Z})$ is one-dimensional, therefore is generated by the function $E_1^\chi(q)$. On the other hand, one can also show that

$$\theta(q)^2 = \sum_{n=0}^{\infty} r_2(n) q^n \in M_1(\Gamma_1(4)).$$

Thus $\theta(q)^2$ is a scalar multiple of $E_1^\chi(q)$. The coefficient of the constant term of $\theta(q)^2$ is $r_2(0) = 1$, while the coefficient of the constant term of $E_1^\chi(q)$ is $1/4$. Hence one obtains

$$\theta(q)^2 = 4E_1^\chi(q).$$

By comparing the coefficients on both sides, we get an explicit formula for $r_2(n)$:

$$r_2(n) = 4 \sum_{d|n} \chi(d).$$

Let us give another proof of the formula using the properties of the ring of Gaussian integers. The number $r_2(n)$ can also be interpreted as the number of Gaussian integers with norm n. Thus

$$\sum_{n \geq 1} \frac{r_2(n)}{n^s} = \sum_{0 \neq \alpha \in \mathbb{Z}[i]} \frac{1}{N(\alpha)^s}.$$

Denote the set of all Gaussian primes (up to units) by $\mathcal{P}$. Then we have

$$\begin{aligned}\sum_{0\neq\alpha\in\mathbb{Z}[i]}\frac{1}{N(\alpha)^s} &= 4\prod_{\pi\in\mathcal{P}}\frac{1}{1-N(\pi)^{-s}}\\ &= 4\cdot\frac{1}{1-2^{-s}}\cdot\prod_{p=1(\text{mod }4)}\frac{1}{(1-p^{-s})^2}\prod_{q=3(\text{mod }4)}\frac{1}{1-q^{-2s}}\\ &= \zeta(s)\cdot L(s,\chi).\end{aligned}$$

Here $\zeta(s)$ is the Riemann zeta function

$$\zeta(s)=\sum_{n\geq 1}\frac{1}{n^s}=\prod_{p\in\mathbb{Z}\text{ prime}}\frac{1}{1-p^{-s}}$$

and $L(s,\chi)$ is the Dirichlet L-series

$$L(s,\chi)=\sum_{n=1}^{\infty}\frac{\chi(n)}{n^s}=\prod_{p\in\mathbb{Z}\text{ prime}}\frac{1}{1-\chi(p)p^{-s}}.$$

So we have

$$\frac{1}{4}\sum_{n\geq 1}\frac{r_2(n)}{n^s}=\left(\sum_{m\geq 1}\frac{1}{m^s}\right)\left(\sum_{d=1}^{\infty}\frac{\chi(d)}{d^s}\right).$$

Thus

$$\frac{1}{4}r_2(n)=\sum_{md=n}\chi(d)=\sum_{d|n}\chi(d).$$

5. Complex analysis, elliptic functions, and modular forms

The German mathematician Martin Eichler once stated that there were five fundamental operations of mathematics: addition, subtraction, multiplication, division, and *modular forms*. In this unit, we will start with discussing the basic concepts of complex analysis, then move on to the discussions of *elliptic functions* and *modular forms*. We will mention many applications along the way, and solve the sums of four squares problem at the end. Some references that might be helpful include [6], [15], and [16].

5.1. **Some applications of modular forms.** Let us discuss the *j-invariant* first. Classically, the j-invariant was studied as a parameterization of *elliptic curves* over $\mathbb{C}$. Every elliptic curve E over $\mathbb{C}$ is a complex torus, and thus can be identified with a rank 2 lattice. This lattice can be rotated and scaled (which preserve the isomorphism class), so that it is generated by 1 and $\tau \in \mathbb{H}$. This lattice corresponds to the elliptic curve

$$y^2 = 4x^3 - g_2(\tau)x - g_3(\tau),$$

where

$$g_2(\tau) = \frac{4\pi^4}{3}E_4(\tau), \qquad g_3(\tau) = \frac{8\pi^6}{27}E_6(\tau),$$

and

$$E_4(\tau) = 1 + 240\sum_{r\geq 1}\sigma_3(r)q^r, \qquad E_6(\tau) = 1 - 504\sum_{r\geq 1}\sigma_5(r)q^r$$

are *Eisenstein series* (which are *modular forms* of weight 4 and 6, respectively), where $q = e^{2\pi i\tau}$ and $\sigma_k(r) = \sum_{d|r} d^k$. The isomorphic class of elliptic curves is uniquely determined by the j-invariant

$$j(\tau) = 1728\frac{g_2(\tau)^3}{g_2(\tau)^3 - 27g_3(\tau)^2}.$$

It is the *unique* (up to scalar multiplication) holomorphic function on $\mathbb{H}$ that is invariant under the $\mathrm{SL}(2,\mathbb{Z})$-action and has a simple pole at infinity. In fact, any meromorphic modular function (i.e. invariant under $\mathrm{SL}(2,\mathbb{Z})$-action) on $\mathbb{H}$ is a rational function of $j(\tau)$.

The j-invariant has many interesting and surprising applications. For instance, let us consider

$$e^{\pi\sqrt{163}} = 262537412640768743.99999999999925...$$

which is very close to an integer. This remarkable phenomenon can be easily deduced using the fact that

$$j\left(\frac{1+\sqrt{-163}}{2}\right) \in \mathbb{Z}.$$

together with the q-expansion of the j-function

$$j(\tau) = \frac{1}{q} + 744 + 196884q + 21493760q^2 + O(q^3), \qquad \text{where } q = e^{2\pi i\tau}.$$

Consider primitive positive-definite quadratic forms $Q(x,y) = ax^2+bxy+cy^2$, where $a,b,c \in \mathbb{Z}$, $\gcd(a,b,c) = 1$, $a > 0$, and $D = b^2 - 4ac < 0$. There is a natural notion of *equivalence* between two such quadratic forms, essentially given by change of variables. One can show that two such quadratic forms are equivalent if and only if $D = D'$ and

$$j\left(\frac{b+\sqrt{-D}}{2a}\right) = j\left(\frac{b'+\sqrt{-D}}{2a'}\right).$$

For each possible discriminant D there are only finitely many equivalence classes, thus we get a finite set of j-values for each discriminant. The big theorem is that these values are the solutions of a monic algebraic equation with integer coefficients. In particular, when there is only one equivalence class for D, the j-invariant of the corresponding quadratic form must be an integer. The above phenomenon then follows from the fact that all positive-definite integer quadratic forms of discriminant $D = -163$ are equivalent (to $x^2 - xy + 41y^2$). In fact, 163 is the largest number satisfying this property; other numbers are: $1, 2, 3, 7, 11, 19, 43, 67$; for instance, we also have

$$e^{\pi\sqrt{67}} \approx \mathbb{Z} + 0.0000013; \qquad e^{\pi\sqrt{43}} \approx \mathbb{Z} + 0.00022.$$

These results on the j-function are one of the starting points of the theory of *complex multiplications* of elliptic curves.

Another surprising result is a connection between the j-function and the *monster group*.

Theorem 5.1. *Every finite simple group is isomorphic to one of the following groups:*

- *a member of one of three infinite classes of:*
  - *the cyclic groups of prime order,*
  - *the alternating groups A_n for $n \geq 5$,*
  - *the groups of Lie type*
- *one of the 27 sporadic groups.*

Among the 27 sporadic groups, the *monster group* M has the largest order of roughly 8×10^{53}. The minimal dimension of a faithful complex representation of the monster group is 196883, which happens to be very close to one of the coefficients in the q-expansion

$$j(\tau) = \frac{1}{q} + 744 + 196884q + 21493760q^2 + 864299970q^3 + 20245856256q^4 + \cdots.$$

In fact, the dimensions of the irreducible representations of M are: $r_1 = 1$, $r_2 = 196883$, $r_3 = 21296876$, $r_4 = 842609326$, $r_5 = 18538750076$, etc., and the coefficients of the q-expansion of j-function satisfies

$$\begin{aligned} 196884 &= r_1 + r_2 \\ 21493760 &= r_1 + r_2 + r_3 \\ 864299970 &= 2r_1 + 2r_2 + r_3 + r_4 \\ 20245856256 &= 3r_1 + 3r_2 + r_3 + 2r_4 + r_5 \\ &\dots \end{aligned}$$

Very roughly, this can be explained by the fact that there exists a *vertex operator algebra* which admits an infinite-dimensional graded representation of the monster group, whose graded dimensions are the coefficients of the j-function. The precise content of this statement and their detailed properties (Conway–Norton conjecture) are proved by Borcherds, who won the Fields Medal in 1998 in part for his solution of the conjecture.

Let us consider a more elementary application of modular forms. Consider the functions

$$\sigma_3(r) = \sum_{d|r} d^3 \quad \text{and} \quad \sigma_7(r) = \sum_{d|r} d^7.$$

They satisfy a relation

$$\sigma_7(r) = \sigma_3(r) + 120 \sum_{p+q=r} \sigma_3(p)\sigma_3(q).$$

This is not an easy statement to prove. Using the fact that

$$E_4(\tau) = 1 + 240 \sum_{r\geq 1} \sigma_3(r) q^r \quad \text{and} \quad E_8(\tau) = 1 + 480 \sum_{r\geq 1} \sigma_7(r) q^r$$

are modular forms of weight 4 and 8, respectively; together with the fact the space of modular forms of weight 8 is one-dimensional, one deduces $E_4(\tau)^2 = E_8(\tau)$. The above relation then follows from comparing the coefficients of both sides of the equation.

5.2. **A crash course on complex analysis.** We recall in this subsection some theorems of complex analysis that will be useful and necessary for our discussions of modular forms later. The proofs of these theorems can be found in any textbook on complex analysis, for instance [16].

Let $U \subseteq \mathbb{C}$ be an open subset of the complex plane. A function $f\colon U \to \mathbb{C}$ is called *holomorphic* if for every $z_0 \in U$, the limit

$$\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} \text{ exists.}$$

In other words, it is holomorphic if the derivative in the "complex sense" exists. If the limit exists, it will be denoted by $f'(z_0) \in \mathbb{C}$. This is exactly the complex analogue of the *differentiable* functions over $\mathbb{R}$. However, holomorphic functions possess many nicer properties than differentiable functions.

Example 5.2. Holomorphic functions satisfy the "local determine global" principle. Namely, suppose there are two holomorphic functions f, g on a (connected) open set $U \subseteq \mathbb{C}$ such that their values agree on an open subset $V \subseteq U$, i.e. $f(z) = g(z)$ for all $z \in V$. Then, no matter how small the open subset V is, we would have $f(z) = g(z)$ for all $z \in U$.

This is not true for smooth functions over $\mathbb{R}$. For instance, the smooth function

$$f(x) = \begin{cases} e^{-1/x^2} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

is identical with the zero function on $\mathbb{R}_{<0}$, but they are obvious not identical on the whole real line.

Example 5.3. Another important result is that if $f\colon U \to \mathbb{C}$ is holomorphic, then its derivative $f'\colon U \to \mathbb{C}$ is automatically holomorphic as well. This implies that any holomorphic is infinitely differentiable, i.e. $f, f', f'', f''', \ldots$ exist. Moreover, for any $z_0 \in U$ the power series

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(z_0)}{n!}(z - z_0)^n$$

converges in a neighborhood of z_0, and the limit coincides with $f(z)$. These are again not true for differentiable functions over $\mathbb{R}$.

These results, together with other basic theorems in complex analysis, including Liouville's theorem, Morera's theorem, residue formula, argument principle, etc., essentially all are corollaries of a single theorem, the *Cauchy integral theorem*. To state the theorem, we need to define the notion of *path integrals*.

Definition 5.4. A *parametrized smooth curve* in $U \subseteq \mathbb{C}$ is a map

$$\gamma\colon [a,b] \to U; \qquad \gamma(t) = x(t) + iy(t)$$

such that

- $x(t), y(t)$ are differentiable, and $x'(t), y'(t)$ are continuous,
- $\gamma'(t) = (x'(t), y'(t)) \neq (0,0)$ for all $t \in (a,b)$.

Example 5.5. $\gamma\colon [0,\pi] \to \mathbb{C}$ where $\gamma(t) = e^{it} = \cos(t) + i\sin(t)$ parametrizes the upper half of the unit circle (going counterclockwise). Note that there are infinitely many ways to represent a curve. For instance, $\gamma'\colon [0,2\pi] \to \mathbb{C}$ where $\gamma'(s) = e^{is/2}$ also parametrizes the upper half of the unit circle with the same orientation.

Definition 5.6. Two parametrized smooth curves $\gamma\colon [a,b] \to \mathbb{C}$ and $\gamma'\colon [c,d] \to \mathbb{C}$ are said to be *equivalent* if there exists a smooth bijective map $\varphi\colon [c,d] \to [a,b]$ so that $\gamma(\varphi(s)) = \gamma'(s)$ and $\varphi'(s) > 0$ for all $s \in [c,d]$.

Note that the condition $\varphi'(s) > 0$ guarantees that the two curves have the same orientations (going in the same direction).

Definition 5.7. A *piecewise parametrized smooth curve* in $U \subseteq \mathbb{C}$ is a continuous map

$$\gamma\colon [a,b] \to U$$

such that there exists $a < p_1 < \cdots < p_n < b$ so that

$$\gamma|_{[a,p_1]}, \ldots, \gamma_{[p_n,b]}$$

are parametrized smooth curves.

Definition 5.8. Let $\gamma\colon [a,b] \to \mathbb{C}$ be a piecewise parametrized smooth curve on an open set $U \subseteq \mathbb{C}$, and let $f\colon U \to \mathbb{C}$ be a continuous function. The *integral of f along γ* is defined to be

$$\int_\gamma f(z)\,\mathrm{d}z \coloneqq \int_a^b f(\gamma(t)) \cdot \gamma'(t)\,\mathrm{d}t.$$

Exercise. Show that if γ and γ' are equivalent, then

$$\int_\gamma f(z)\,\mathrm{d}z = \int_{\gamma'} f(z)\,\mathrm{d}z \quad \text{for any } f.$$

In other words, the integral depends only on the underlying curve (and its orientation). (Hint: This essentially follows from the change of variables of integrals.)

Exercise. Show that if γ and γ' parametrizes the same curve but with opposite orientations, then

$$\int_\gamma f(z)\,\mathrm{d}z = -\int_{\gamma'} f(z)\,\mathrm{d}z \quad \text{for any } f.$$

The following is perhaps the most important (yet simple) example of path integrals.

Example 5.9. Consider the unit circle parametrizes counterclockwisely $\gamma\colon [0,2\pi] \to \mathbb{C}$ where $\gamma(t) = e^{it}$. The function $f(z) = \frac{1}{z}$ is continuous (in fact, holomorphic) on $\mathbb{C}\backslash\{0\}$, so it makes sense to compute the path integral of f along the unit circle.

$$\int_\gamma f(z)\,\mathrm{d}z = \int_0^{2\pi} \frac{1}{e^{it}} \cdot ie^{it}\,\mathrm{d}t = 2\pi i.$$

Remark 5.10. In general, let γ be a curve, not necessarily simple (i.e. may have self-intersections), that does not pass through the origin. Then the integral

$$\frac{1}{2\pi i}\int_\gamma \frac{1}{z}\,\mathrm{d}z \in \mathbb{Z}$$

is always an integer, which gives the *winding number* of γ around the origin.

Exercise. Let $F\colon U \to \mathbb{C}$ be a holomorphic function on an open set U, and let γ be a piecewise smooth curve in U, starting at w_1 and ending at w_2. Then

$$\int_\gamma F'(z)\,\mathrm{d}z = F(w_2) - F(w_1).$$

In particular, $\int_{S^1} z^n\,\mathrm{d}z = 0$ unless $n = -1$.

Remark 5.11. The previous example and exercise suggest that the log function $\log z$ is not well-defined on $\mathbb{C}\backslash\{0\}$. Indeed, it is only possible to define $\log z$ on the *universal cover* of $\mathbb{C}\backslash\{0\}$.

We now state the Cauchy integral theorem.

Theorem 5.12 (Cauchy integral theorem)**.** *Let γ be a simple closed curve in $\mathbb{C}$. Suppose f is holomorphic on an open set containing γ* and its interior*, then*

$$\int_\gamma f(z)\,\mathrm{d}z = 0.$$

Corollary 5.13. *Let γ be a simple closed curve in $\mathbb{C}$ (oriented counterclockwisely). Suppose f is holomorphic on an open set containing γ and its interior, except at the points $z_1, \dots, z_k$ in the interior of γ where f is not defined. Choose any small loops $\gamma_1, \dots, \gamma_k$ (oriented counterclockwisely) that lie in the interior of γ, so that γ_i contains only one of the z_i. Then*

$$\int_\gamma f(z)\,\mathrm{d}z = \sum_{i=1}^{k} \int_{\gamma_i} f(z)\,\mathrm{d}z.$$

In other words, to compute $\int_\gamma f(z)\,\mathrm{d}z$, it suffices to compute the integrals $\int_{\gamma_i} f(z)\,\mathrm{d}z$ around the *singularities* (where f is not defined) $z_1, \dots, z_k$. These integrals are completely determined by the local behavior of f near the singular points.

Theorem 5.14 (Laurent series expansion)**.** *Let $z_0 \in \mathbb{C}$ and $R > 0$. Suppose f is a holomorphic function on the open set $0 < |z - z_0| < R$. For each $n \in \mathbb{Z}$, define*

$$a_n = \frac{1}{2\pi i} \int_\gamma \frac{f(z)}{(z - z_0)^{n+1}}\,\mathrm{d}z$$

where γ is counterclockwise around a simple closed curve enclosing z_0 inside the open set $0 < |z - z_0| < R$. Then the series

$$\sum_{n=-\infty}^{\infty} a_n (z - z_0)^n$$

converges and coincides with $f(z)$ for any $0 < |z - z_0| < R$.

Remark 5.15. In particular, if f is holomorphic on the whole neighborhood $|z - z_0| < R$, then $a_{-n} = 0$ for any $n > 0$ by the Cauchy integral theorem, so the series above gives the power series expansion near z_0. In particular, for each $n \geq 0$ the n-th derivative of f at z_0 is

$$f^{(n)}(z_0) = \frac{n!}{2\pi i} \int_\gamma \frac{f(z)}{(z - z_0)^{n+1}}\,\mathrm{d}z. \tag{5.1}$$

Lecture 6

This is the *Cauchy integral formula*.

Remark 5.16. Logically speaking, the theorem on Laurent series expansion is a consequence of the Cauchy integral formula, which, is ultimately a consequence of the Cauchy integral theorem that we started with. Let us sketch the proof of

$$f(z_0) = \frac{1}{2\pi i} \int_\gamma \frac{f(z)}{z - z_0} \, \mathrm{d}z \tag{5.2}$$

assuming the Cauchy integral theorem (here f is holomorphic on z_0 and its neighborhood); and in the next remark, we sketch the proof of the Cauchy integral theorem. The idea is to write

$$\frac{1}{2\pi i} \int_\gamma \frac{f(z)}{z - z_0} \, \mathrm{d}z = \frac{1}{2\pi i} \int_\gamma \frac{f(z) - f(z_0)}{z - z_0} \, \mathrm{d}z + \frac{1}{2\pi i} \int_\gamma \frac{f(z_0)}{z - z_0} \, \mathrm{d}z.$$

By Cauchy integral theorem, the second term is $f(z_0)$, so it suffices to show that the first term is zero. This follows from the following two observations. First, the function $(f(z) - f(z_0))/(z - z_0)$ is bounded (say, by $M > 0$) near z_0 since f is holomorphic at z_0. Second, again by Cauchy integral theorem, we have

$$I_1 = \frac{1}{2\pi i} \int_\gamma \frac{f(z) - f(z_0)}{z - z_0} \, \mathrm{d}z = \frac{1}{2\pi i} \int_{\gamma_\epsilon} \frac{f(z) - f(z_0)}{z - z_0} \, \mathrm{d}z$$

for any circle γ_ϵ of radius $\epsilon > 0$ centered at z_0. Thus

$$|I_1| \leq \frac{1}{2\pi} \cdot M \cdot \mathrm{length}(\gamma_\epsilon) = M \cdot \epsilon \quad \text{for any} \quad \epsilon > 0.$$

Hence $I_1 = 0$. The fact that holomorphic functions are indefinitely differentiable, and the general Cauchy integral formula (5.1) are both easy consequences of (5.2).

Remark 5.17. In this remark, we sketch the proof of the Cauchy integral theorem. Let us discuss only the case where the curve γ is a *triangle*. The general case would follow from this case together with certain limiting process, which we omit here. Let us denote the interior of γ, which is a triangle, by $T^{(0)}$. One can divide $T^{(0)}$ into four sub-triangles, so that the path integral of f along γ equals to the sum of the path integrals along the boundary of these

four sub-triangles. Therefore, at least one of the four sub-triangles, say $T^{(1)}$, satisfies

$$\left|\int_{\gamma=\partial T^{(0)}} f(z)\,\mathrm{d}z\right| \leq 4\left|\int_{\partial T^{(1)}} f(z)\,\mathrm{d}z\right|.$$

Continue this process indefinitely, we obtain a sequence of triangles

$$\cdots \subseteq T^{(2)} \subseteq T^{(1)} \subseteq T^{(0)}$$

where the diameter $d^{(n)}$ and perimeter $p^{(n)}$ is decreased by half in each step, and

$$\left|\int_{\gamma=\partial T^{(0)}} f(z)\,\mathrm{d}z\right| \leq 4^n\left|\int_{\partial T^{(n)}} f(z)\,\mathrm{d}z\right|.$$

Since each triangle is a compact subset, the sequence would converge to a unique point, say z_0. Using the condition that f is holomorphic, for any $\epsilon > 0$ there exists a $\delta > 0$ so that

$$|f(z) - f(z_0) - (z - z_0)f'(z_0)| < \epsilon \cdot (z - z_0) \quad \text{for all} \quad z \in B_\delta(z_0).$$

Choose n large enough so that $T^{(n)} \subseteq B_\delta(z_0)$, then we have

$$\begin{aligned}\left|\int_{\partial T^{(n)}} f(z)\,\mathrm{d}z\right| &= \left|\int_{\partial T^{(n)}} (f(z) - f(z_0) - (z - z_0)f'(z_0))\,\mathrm{d}z\right| \text{ (why?)}\\ &\leq p^{(n)} \cdot \sup_{z\in\partial T^{(n)}} |f(z) - f(z_0) - (z - z_0)f'(z_0)|\\ &< p^{(n)} \cdot \epsilon \cdot d^{(n)} = \epsilon \cdot \frac{p^{(0)}}{2^n} \cdot \frac{d^{(0)}}{2^n}.\end{aligned}$$

Thus

$$\left|\int_{\gamma} f(z)\,\mathrm{d}z\right| \leq 4^n \cdot \epsilon \cdot \frac{p^{(0)}}{2^n} \cdot \frac{d^{(0)}}{2^n} = \epsilon \cdot p^{(0)} \cdot d^{(0)} \quad \text{for any} \quad \epsilon > 0.$$

Hence $\int_\gamma f(z)\,\mathrm{d}z = 0$.

Definition 5.18. The residue of f at a singular point z_0 is defined to be

$$\mathrm{Res}(f, z_0) \coloneqq a_{-1} = \frac{1}{2\pi i}\int_\gamma f(z)\,\mathrm{d}z.$$

Notation. Let n be a positive integer. Let f be a holomorphic function on $0 < |z - z_0| < R$, and a_n be the coefficients of its Laurent series expansion defined earlier. We say

- f has a *zero of order* n at z_0 if $a_n \neq 0$ and $a_m = 0$ for all $m < n$.
- f has a *pole of order* n at z_0 if $a_{-n} \neq 0$ and $a_m = 0$ for all $m < -n$.

Example 5.19. Suppose f has a *simple pole* at z_0, i.e. $(z - z_0)f(z)$ can be extended to a holomorphic function on the whole neighborhood $|z - z_0| < R$. Then a_{-n} for any $n \geq 2$ by the Cauchy integral theorem, so the Laurent series expansion of f near the point z_0 is given by

$$f(z) = \frac{a_{-1}}{z - z_0} + a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + \cdots .$$

Therefore a_{-1} can be computed by the limit

$$a_{-1} = \lim_{z \to z_0} f(z)(z - z_0).$$

Similarly, suppose f has a pole of order n at z_0 (i.e. $(z - z_0)^n f(z)$ can be extended to a holomorphic function on the whole neighborhood $|z - z_0| < R$, but $(z - z_0)^{n-1} f(z)$ cannot), then its residue can be computed by

$$\operatorname{Res}(f, z_0) = \frac{1}{(n-1)!} \lim_{z \to z_0} \frac{d^{n-1}}{dz^{n-1}} \left((z - z_0)^n f(z) \right).$$

Example 5.20. Let $a > 0$ be a positive real number. The function

$$f(z) = \frac{e^{iz}}{z^2 + a^2}$$

is holomorphic except at $\pm ia$. It is clear that both $\pm ia$ are simple poles of f.

$$\operatorname{Res}(f, ia) = \lim_{z \to ia} \frac{e^{iz}}{z^2 + a^2}(z - ia) = \lim_{z \to ia} \frac{e^{iz}}{z + ia} = \frac{e^{-a}}{2ia}.$$

Example 5.21. How to compute the integral

$$\int_{-\infty}^{\infty} \frac{\cos x}{x^2 + a^2} \, \mathrm{d}x \ = \ ?$$

Let $R \gg 0$ and let γ_R parametrizes the upper half of the circle $|z| = R$ going counterclockwisely. Then

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{\cos x}{x^2+a^2}\,\mathrm{d}x &= \lim_{R\to\infty}\int_{-R}^{R} \frac{\cos x}{x^2+a^2}\,\mathrm{d}x \\ &= \operatorname{Re}\left(\lim_{R\to\infty}\int_{-R}^{R} \frac{e^{iz}}{z^2+a^2}\,\mathrm{d}z\right) \\ &= \operatorname{Re}\left(2\pi i\cdot \operatorname{Res}\left(\frac{e^{iz}}{z^2+a^2}, ia\right) - \lim_{R\to\infty}\int_{\gamma_R} \frac{e^{iz}}{z^2+a^2}\,\mathrm{d}z\right) \\ &= \frac{\pi e^{-a}}{a} - \operatorname{Re}\left(\lim_{R\to\infty}\int_{\gamma_R} \frac{e^{iz}}{z^2+a^2}\,\mathrm{d}z\right).\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\left|\int_{\gamma_R} \frac{e^{iz}}{z^2+a^2}\,\mathrm{d}z\right| &= \left|\int_0^{\pi} \frac{e^{iRe^{i\theta}}}{R^2e^{2i\theta}+a^2}\cdot iRe^{i\theta}\,\mathrm{d}\theta\right| \\ &\le \int_0^{\pi} \frac{1}{|R^2-a^2|}\cdot R\,\mathrm{d}\theta \longrightarrow 0 \quad \text{as } R\to\infty.\end{aligned}$$

Thus we get

$$\int_{-\infty}^{\infty} \frac{\cos x}{x^2+a^2}\,\mathrm{d}x = \frac{\pi e^{-a}}{a}.$$

Theorem 5.22 (Liouville)**.** *Let f be a bounded ($|f(z)| < M$ for all z) and entire (holomorphic on the whole complex plane $\mathbb{C}$) function. Then f is a constant function.*

Proof. It suffices to show that the derivative $f'(z_0)$ is zero for all $z_0 \in \mathbb{C}$. Let $\gamma_R(z_0)$ be the circle of radius R centered at the point z_0. By the Cauchy integral formula, we have

$$\begin{aligned}|f'(z_0)| &= \frac{1}{2\pi}\left|\int_{\gamma_R(z_0)} \frac{f(z)}{(z-z_0)^2}\,\mathrm{d}z\right| \\ &< \frac{1}{2\pi}\cdot\frac{M}{R^2}\cdot 2\pi R = \frac{M}{R}.\end{aligned}$$

The inequality $|f'(z_0)| < \frac{M}{R}$ holds for all $R > 0$. Thus $f'(z_0) = 0$. □

The fundamental theorem of algebra is a simple corollary of the Liouville theorem.

Corollary 5.23 (Fundamental theorem of algebra)**.** *Any non-constant complex polynomial $p(z)$ has a root in $\mathbb{C}$.*

Proof. Assume the contrary that $p(z)$ has no roots in $\mathbb{C}$. Then $\frac{1}{p(z)}$ is an entire function. It is not hard to show that $\frac{1}{p(z)}$ is a bounded function on $\mathbb{C}$. By Liouville theorem, it can only be the constant function. □

Finally, we state the *argument principle*, which claims that the integral

$$\frac{1}{2\pi i}\int_\gamma \frac{f'(z)}{f(z)}\,\mathrm{d}z$$

counts the number of zeros minus the number of poles in the interior of γ. To illustrate this, let us start with two basic examples.

Example 5.24. Consider $f(z) = z^n$, which has a zero of order n at the point 0. Let γ be any simple closed curve enclosing 0 (oriented counterclockwisely). Then

$$\int_\gamma \frac{f'(z)}{f(z)}\,\mathrm{d}z = \int_\gamma \frac{nz^{n-1}}{z^n}\,\mathrm{d}z = n\int_\gamma \frac{1}{z}\,\mathrm{d}z = 2\pi i\cdot n.$$

Example 5.25. Consider $f(z) = z^{-n}$, which has a pole of order n at the point 0. Let γ be any simple closed curve enclosing 0 (oriented counterclockwisely). Then

$$\int_\gamma \frac{f'(z)}{f(z)}\,\mathrm{d}z = \int_\gamma \frac{-nz^{-n-1}}{z^{-n}}\,\mathrm{d}z = (-n)\int_\gamma \frac{1}{z}\,\mathrm{d}z = 2\pi i\cdot(-n).$$

In general, we have the following theorem.

Theorem 5.26 (Argument principle)**.** *Let γ be a simple closed curve. Suppose f is a holomorphic function on an open set containing γ and its interior, except at finitely many poles. Then*

$$\frac{1}{2\pi i}\int_\gamma \frac{f'(z)}{f(z)}\,\mathrm{d}z = (\#\ \textit{zeros of}\ f(z)\ \textit{inside}\ \gamma) - (\#\ \textit{poles of}\ f(z)\ \textit{inside}\ \gamma)\,.$$

Here the numbers are counted with multiplicities, i.e. an order n zero is counted as n zeros, and an order n pole is counted as n poles.

5.3. **Elliptic functions.** We discuss the *elliptic functions* in this subsection; some of the aspects of elliptic functions are closely related to modular forms and will be useful later.

Definition 5.27. Let $\omega_1, \omega_2 \in \mathbb{C}$ be two complex numbers that are linearly independent over $\mathbb{R}$ (i.e. they span the vector space $\mathbb{C} \cong \mathbb{R}^2$). We say a function f on $\mathbb{C}$ is *elliptic* (with respect to ω_1, ω_2) if

$$f(z) = f(z + \omega_1) = f(z + \omega_2) \quad \text{for all} \quad z \in \mathbb{C}.$$

The parallelogram with vertices $0, \omega_1, \omega_2, \omega_1 + \omega_2$ is called the *fundamental domain*. It is easy to see that the values of an elliptic function on $\mathbb{C}$ is determined by its value on the fundamental domain.

Exercise. Show that the only *holomorphic* elliptic functions are the constant functions. (Hint: Liouville's theorem.)

Therefore, it is more interesting to consider the *meromorphic* elliptic functions. (A function on $U \subseteq \mathbb{C}$ is called *meromorphic* if for any $z_0 \in U$, the function f either is holomorphic at z_0 or has a pole at z_0.) Here is a rough idea of a way to construct such functions. Let g be a meromorphic function on $\mathbb{C}$. Then

$$f(z) = \sum_{m,n \in \mathbb{Z}} g(z + m\omega_1 + n\omega_2)$$

must be elliptic, provided that the series on the right hand side converges. Suppose we have $|g(z)| < \frac{C}{|z|^\alpha}$ for $|z| \gg 0$. Observe that for a fix $z \in \mathbb{C}$, the number of points of the form $z + m\omega_1 + n\omega_2$ in the annulus $R \leq |z + m\omega_1 + n\omega_2| < R + 1$ is roughly (constant)$\cdot R$. Thus

$$\begin{aligned} \sum_{m,n \in \mathbb{Z}} |g(z + m\omega_1 + n\omega_2)| &= \sum_{R=0}^{\infty} \sum_{\substack{m,n \in \mathbb{Z} \\ R \leq |z + m\omega_1 + n\omega_2| < R+1}} |g(z + m\omega_1 + n\omega_2)| \\ &\approx \sum_{R=0}^{\infty} \frac{C}{R^\alpha} \cdot R \end{aligned}$$

Hence, if $\alpha > 2$, then the series $\sum_{m,n \in \mathbb{Z}} g(z + m\omega_1 + n\omega_2)$ converges absolutely. Let us summarize this as the following example.

Example 5.28. Let $C > 0$ be a constant and $\alpha > 2$. If $|g(z)| < \frac{C}{|z|^\alpha}$ for all $|z| \gg 0$, then

$$f(z) = \sum_{m,n\in\mathbb{Z}} g(z + m\omega_1 + n\omega_2)$$

is a meromorphic elliptic function.

For instance, one can take $g(z) = \frac{1}{(z-\alpha)(z-\beta)(z-\gamma)}$ for some $\alpha, \beta, \gamma \in \mathbb{C}$. Then

$$f(z) = \sum_{m,n\in\mathbb{Z}} g(z + m\omega_1 + n\omega_2)$$

is a meromorphic elliptic function, which has 3 poles in the fundamental domain.

Question 5.29. *Do there exist meromorphic elliptic functions with only* 1 *or* 2 *poles in the fundamental domain?*

Notation. We denote the lattice

$$\Lambda = \{m\omega_1 + n\omega_2 \mid m, n \in \mathbb{Z}\} \subseteq \mathbb{C}.$$

To answer this question, let us first establish the following basic (yet important) fact about elliptic functions.

Theorem 5.30. *Let f be a meromorphic elliptic function with respect to Λ. Assume that f has no zeros or poles on the boundary of the fundamental domain. Then*

(a) The number of zeros of f in the fundamental domain coincides with the number of poles of f in the fundamental domain.
(b) The sum of the zeros of f (which is a complex number) minus the sum of the poles of f in the fundamental domain is an element in Λ.

Here the zeros and poles are counted with multiplicities.

Proof. The first statement follows directly from the argument principle. To show the second the statement, we first claim a general statement that the sum of the zeros of f minus the sum of the poles of f in an area enclosed by a loop γ (oriented counterclockwisely) is given by

$$\frac{1}{2\pi i}\int_\gamma z \cdot \frac{f'(z)}{f(z)}\,\mathrm{d}z.$$

The computation of the integral boils down to computing the integral around the zeros and poles of f. As an example, say z_0 is a zero of order n of f. Then $f(z) = (z - z_0)^n h(z)$ where h is holomorphic in a neighborhood of z_0 with $h(z_0) \neq 0$. Let γ_0 be a small loop centered at the zero z_0. Then we have

$$\begin{aligned}\frac{1}{2\pi i}\int_{\gamma_0} z \cdot \frac{f'(z)}{f(z)}\,\mathrm{d}z &= \frac{1}{2\pi i}\int_{\gamma_0} z \cdot \frac{n(z-z_0)^{n-1}h(z) + (z-z_0)^n h'(z)}{(z-z_0)^n h(z)}\,\mathrm{d}z \\ &= \frac{1}{2\pi i}\int_{\gamma_0} z \cdot \frac{n}{z-z_0}\,\mathrm{d}z \\ &= \frac{1}{2\pi i}\int_{\gamma_0} (z-z_0)\cdot\frac{n}{z-z_0}\,\mathrm{d}z + \frac{1}{2\pi i}\int_{\gamma_0} z_0 \cdot \frac{n}{z-z_0}\,\mathrm{d}z \\ &= 0 + nz_0 = nz_0.\end{aligned}$$

Similar computations work for the poles. Therefore, to show the second statement, one needs to show that

$$\begin{aligned}&\frac{1}{2\pi i}\int_0^{\omega_1} z\cdot\frac{f'(z)}{f(z)}\,\mathrm{d}z + \frac{1}{2\pi i}\int_{\omega_1}^{\omega_1+\omega_2} z\cdot\frac{f'(z)}{f(z)}\,\mathrm{d}z \\ &\qquad + \frac{1}{2\pi i}\int_{\omega_1+\omega_2}^{\omega_2} z\cdot\frac{f'(z)}{f(z)}\,\mathrm{d}z + \frac{1}{2\pi i}\int_{\omega_2}^{0} z\cdot\frac{f'(z)}{f(z)}\,\mathrm{d}z \in \Lambda.\end{aligned}$$

Observe that

$$\frac{1}{2\pi i}\int_{\omega_1}^{\omega_1+\omega_2} z\cdot\frac{f'(z)}{f(z)}\,\mathrm{d}z - \frac{1}{2\pi i}\int_0^{\omega_2} z\cdot\frac{f'(z)}{f(z)}\,\mathrm{d}z = \frac{\omega_1}{2\pi i}\int_0^{\omega_2}\frac{f'(z)}{f(z)}\,\mathrm{d}z$$

by the periodicity of f. Define $\eta(t) = f(\omega_2 t)$ for $t \in [0,1]$, which parametrizes a closed curve (not necessarily simple) in $\mathbb{C}\backslash\{0\}$. Then

$$\frac{1}{2\pi i}\int_0^{\omega_2}\frac{f'(z)}{f(z)}\,\mathrm{d}z = \frac{1}{2\pi i}\int_0^1 \frac{f'(\omega_2 t)}{f(\omega_2 t)}\cdot\omega_2\,\mathrm{d}t = \frac{1}{2\pi i}\int_0^1\frac{\eta'(t)}{\eta(t)}\,\mathrm{d}t = \frac{1}{2\pi i}\int_\eta \frac{1}{z}\,\mathrm{d}z \in \mathbb{Z}.$$

Hence

$$\frac{1}{2\pi i}\int_{\omega_1}^{\omega_1+\omega_2} z\cdot\frac{f'(z)}{f(z)}\,\mathrm{d}z - \frac{1}{2\pi i}\int_0^{\omega_2} z\cdot\frac{f'(z)}{f(z)}\,\mathrm{d}z \in \omega_1\mathbb{Z}.$$

Similarly, one can show that

$$\frac{1}{2\pi i}\int_{\omega_2}^{\omega_1+\omega_2} z\cdot\frac{f'(z)}{f(z)}\,\mathrm{d}z - \frac{1}{2\pi i}\int_0^{\omega_1} z\cdot\frac{f'(z)}{f(z)}\,\mathrm{d}z \in \omega_2\mathbb{Z}.$$

This concludes the proof. □

Remark 5.31. In fact, one can show that given $z_1, \dots, z_n, p_1, \dots, p_n$ in the fundamental domain satisfying $\sum z_i = \sum p_i$, there exists a meromorphic elliptic function f with zeros at $z_1, \dots, z_n$ and poles at $p_1, \dots, p_n$.

Corollary 5.32. *There is no meromorphic elliptic function with exactly* 1 *pole in the fundamental domain (counted with multiplicity).*

Proof. Assume the contrary that f is a meromorphic elliptic function with exactly 1 pole in the fundamental domain, say at p_0. By the first part of the theorem, there is exactly 1 zero in the fundamental domain as well, say at z_0. The second part of the theorem then implies that $z_0 = p_0$, which is impossible since a point cannot be a zero and a pole of f simultaneously. □

It turns out that there exist meromorphic elliptic functions with exactly 2 poles in the fundamental domain. One of such functions is the *Weierstrass $\wp$-function*. Recall that the naive construction using

$$\sum_{\lambda \in \Lambda} \frac{1}{(z+\lambda)^2}$$

fails, because the series does not converge. One way to get around this is to consider

$$\frac{1}{(z+\lambda)^2} - \frac{1}{\lambda^2} = \frac{-z^2 - 2z\lambda}{(z+\lambda)^2\lambda^2}$$

which is now of degree -3 in λ. Indeed, one can show that the series

$$\wp(z) = \frac{1}{z^2} + \sum_{\lambda \in \Lambda \setminus \{0\}} \left(\frac{1}{(z+\lambda)^2} - \frac{1}{\lambda^2} \right)$$

converges, and is defined to be the *Weierstrass $\wp$-function*. Now, because the right hand side is not symmetric with respect to all $\lambda \in \Lambda$, we have to show that it is indeed an elliptic function.

Proposition 5.33. *The function $\wp(z)$ is elliptic with respect to Λ.*

Proof. It is clear that the derivative $\wp'(z)$ is elliptic (the asymmetry of $\wp$ is caused by the terms $\frac{1}{\lambda^2}$, which will be annihilated by the derivative in z), so

$$\wp'(z) = \wp'(z + \omega_1) = \wp'(z + \omega_2).$$

Therefore, the function $\wp(z) - \wp(z+\omega_1)$ is a constant function in z, say

$$\wp(z) - \wp(z+\omega_1) = C.$$

Using the fact that $\wp(z)$ is an even function ($\wp(-z) = \wp(z)$), we have

$$C = \wp(-\omega_1/2) - \wp(\omega_1/2) = 0.$$

Thus $\wp(z) = \wp(z+\omega_1)$. Similarly, one can show that $\wp(z) = \wp(z+\omega_2)$. □

The proposition shows that $\wp(z)$ is an elliptic meromorphic function. It has exactly two poles in the fundamental domain (which is given by the double pole at the origin). By the theorem we proved earlier, $\wp(z)$ should also have two zeros in the fundamental domain.

Question 5.34. *What are the zeros of $\wp(z)$ (in the fundamental domain)?*

It turns out that the answer to this simple question is much harder than it appears to be.

Theorem 5.35 (Eichler–Zagier)**.** *Let Λ_τ be the lattice generated by 1 and $\tau \in \mathbb{H} = \{x+iy \mid y > 0\}$. Then the zeros of $\wp(z,\tau)$ in the fundamental domain are given by*

$$\frac{1}{2} \pm \left(\frac{\log(5+2\sqrt{6})}{2\pi i} + 144\pi i\sqrt{6} \int_\tau^{i\infty} (\sigma - \tau) \frac{E_4(\sigma)^3}{E_6(\sigma)^{3/2} j(\sigma)} \, \mathrm{d}\sigma \right)$$

where E_4, E_6, j are the modular forms and functions we will discuss further.

Remark 5.36. Recall that the trick to make the series $\sum_{\lambda\in\Lambda} \frac{1}{(z+\lambda)^2}$ converges was to consider

$$\frac{1}{(z+\lambda)^2} - \frac{1}{\lambda^2} = \frac{-z^2 - 2z\lambda}{(z+\lambda)^2\lambda^2},$$

which becomes of degree -3 in λ. One can apply the same method to the series $\sum_{\lambda\in\Lambda} \frac{1}{z-\lambda}$. Since

$$\frac{1}{z-\lambda} = \frac{-1}{\lambda}\left(1 + \frac{z}{\lambda} + \frac{z^2}{\lambda^2} + \cdots\right),$$

the expression

$$\frac{1}{z-\lambda} + \frac{1}{\lambda} + \frac{z}{\lambda^2} \quad \text{is of degree 3 in } \lambda.$$

This gives the *Weierstrass ζ-function*

$$\zeta(z) = \frac{1}{z} + \sum_{\lambda \in \Lambda \setminus \{0\}} \left(\frac{1}{z-\lambda} + \frac{1}{\lambda} + \frac{z}{\lambda^2} \right).$$

Since the Weierstrass ζ-function has only one simple pole in the fundamental domain of Λ, so it cannot be an elliptic function. On the other hand, its derivative

$$\zeta'(z) = -\wp(z)$$

is elliptic. Hence $\zeta'(z) = \zeta'(z+\omega_1)$, thus $\zeta(z+\omega_1) - \zeta(z)$ is a constant function. Similarly, $\zeta(z+\omega_2) - \zeta(z)$ also is a constant function.

Remark 5.37. By the last property, for any $a, b \in \mathbb{C}$, the function $\zeta(z-a) - \zeta(z-b)$ is an elliptic function, which has exactly two poles in the fundamental domain (a, b modulo Λ).

Let us compute the Laurent series expansion of the Weierstrass $\wp$-function near $z = 0$. First, since for any $|w| < 1$ we have

$$\frac{1}{1-w} = \sum_{n=0}^{\infty} w^n, \quad \text{thus} \quad \frac{1}{(1-w)^2} = \sum_{n=0}^{\infty} (n+1) w^n.$$

Thus

$$\frac{1}{(z-\lambda)^2} = \frac{1}{\lambda^2 \left(1 - \frac{z}{\lambda}\right)^2} = \frac{1}{\lambda^2} + \frac{1}{\lambda^2} \sum_{n=1}^{\infty} (n+1) \left(\frac{z}{\lambda}\right)^n.$$

Therefore

$$\begin{aligned} \wp(z) &= \frac{1}{z^2} + \sum_{\lambda \neq 0} \left(\frac{1}{\lambda^2} \sum_{n=1}^{\infty} (n+1) \left(\frac{z}{\lambda}\right)^n \right) \\ &= \frac{1}{z^2} + \sum_{n=1}^{\infty} \left(\left(\sum_{\lambda \neq 0} \frac{1}{\lambda^{n+2}} \right) (n+1) z^n \right) \end{aligned}$$

For each $n \geq 3$, define the *Eisenstein series* of Λ as

$$\widetilde{E_n}(\Lambda) = \sum_{\lambda \in \Lambda \setminus \{0\}} \frac{1}{\lambda^n}.$$

Note that $\widetilde{E_n}(\Lambda) = 0$ if n is odd. Thus we have

$$\wp(z) = \frac{1}{z^2} + 3\widetilde{E_4}z^2 + 5\widetilde{E_6}z^4 + \cdots .$$

Remark 5.38. Using the fact that the only holomorphic elliptic functions are constant functions, one can deduce many identities about $\wp(z)$. For instance, one can prove the following proposition.

Proposition 5.39. *$\wp'(z)^2$ can be expressed as a cubic polynomial of $\wp(z)$.*

Proof. Compute the first few terms of the Laurent series of:

$$\wp'(z) = \frac{-2}{z^3} + 6\widetilde{E_4}z + 20\widetilde{E_6}z^3 + \cdots$$

$$\wp'(z)^2 = \frac{4}{z^6} - \frac{24\widetilde{E_4}}{z^2} - 80\widetilde{E_6} + \cdots$$

$$\wp(z)^3 = \frac{1}{z^6} + \frac{9\widetilde{E_4}}{z^2} + 15\widetilde{E_6} + \cdots$$

Thus

$$\wp'(z)^2 - 4\wp(z)^3 + 60\widetilde{E_4}\wp(z) = -140\widetilde{E_6} + \cdots$$

is a holomorphic elliptic function, therefore is a constant. Hence

$$\wp'(z)^2 = 4\wp(z)^3 - 60\widetilde{E_4}\wp(z) - 140\widetilde{E_6}.$$

□

Remark 5.40. The proposition is closely related to the cubic equation of *elliptic curves*. There is a map

$$\mathbb{C}/\Lambda \longrightarrow \{y^2 = 4x^3 - 60\widetilde{E_4}x - 140\widetilde{E_6}\} \subseteq \mathbb{C}^2; \quad z \mapsto (\wp(z), \wp'(z)).$$

Remark 5.41. In fact, one can show that

$$\wp'(z)^2 = 4\left(\wp(z) - \wp\left(\frac{\omega_1}{2}\right)\right)\left(\wp(z) - \wp\left(\frac{\omega_2}{2}\right)\right)\left(\wp(z) - \wp\left(\frac{\omega_1+\omega_2}{2}\right)\right).$$

Theorem 5.42. *Any meromorphic elliptic function can be expressed as a rational polynomial in $\wp(z)$ and $\wp'(z)$.*

Proof. Let f be a meromorphic elliptic function. By considering

$$f(z) = \left(\frac{f(z)+f(-z)}{2}\right) + \left(\frac{f(z)-f(-z)}{2}\right),$$

it suffices to prove the theorem for *even* meromorphic elliptic functions and *odd* meromorphic elliptic functions. Up to multiplying $\wp'(z)$ (which is an odd function), it suffices to prove the theorem only for *even* meromorphic elliptic functions.

We claim that any even meromorphic elliptic function f is a rational polynomial in the Weierstrass $\wp$-function $\wp(z)$.

- Up to multiplying $\wp(z)-\wp(z_0)$, one reduces to the case where the poles of f are at Λ.
- Let $f(z) = \frac{a_{-2n}}{z^{2n}} + \cdots$ be the Laurent series expansion near $z = 0$. Then

  $$f(z) - a_{-2n}\wp(z)^n = \frac{\star}{z^{2n-2}} + \cdots$$

  is also an even meromorphic elliptic function.
- Continue this process inductively, one finds $a_{-2}, a_{-4}, \ldots, a_{-2n}$ so that

  $$f(z) - a_{-2n}\wp(z)^n - a_{-2(n-1)}\wp(z)^{n-1} - \cdots - a_{-2}\wp(z)$$

  is a holomorphic elliptic function, therefore is a constant function. Thus $f(z)$ can be expressed as a polynomial in $\wp(z)$.

□

Exercise. Let $\tau \in \mathbb{H}$ be an element in the upper half-plane $\mathbb{H}$. Denote the lattice $\langle 1, \tau\rangle$ as Λ_τ. The Weierstrass $\wp$-function depends on the choice of the lattice. We denote

$$\wp(z,\tau) = \frac{1}{z^2} + \sum_{\lambda\in\Lambda_\tau\setminus\{0\}} \left(\frac{1}{(z+\lambda)^2} - \frac{1}{\lambda^2}\right).$$

Prove that for any integers $a, b, c, d \in \mathbb{Z}$ with $ad - bc = 1$, we have

$$\wp\left(\frac{z}{c\tau+d}, \frac{a\tau+b}{c\tau+d}\right) = (c\tau+d)^2\wp(z,\tau).$$

Lecture 7

5.4. **Modular functions and modular forms.** Informally, modular functions (resp. modular forms) are functions (resp. differential forms) define on the *moduli space of complex torus*, or equivalently, on the *moduli space of lattices in* $\mathbb{C}$ up to the equivalence $\Lambda_1 \sim \Lambda_2$ if $\Lambda_1 = c\Lambda_2$ for some $c \in \mathbb{C}\backslash\{0\}$. Since any lattice is equivalent to a lattice of the form $\Lambda_\tau = \langle 1, \tau\rangle$ for some $\tau \in \mathbb{H}$, the modular functions or forms can be regarded as functions on $\mathbb{H}$ that behaves nicely under the $\mathrm{SL}(2,\mathbb{Z})$-action (since the $\mathrm{SL}(2,\mathbb{Z})$-actions preserve lattices). Recall that $g = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}(2,\mathbb{Z})$ acts on $\tau \in \mathbb{H}$ by

$$g \cdot \tau = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \tau \coloneqq \frac{a\tau + b}{c\tau + d}.$$

Remark 5.43. One can check easily that

$$\mathrm{Im}(g \cdot \tau) = \frac{\mathrm{Im}(\tau)}{|c\tau + d|^2}.$$

Therefore the $\mathrm{SL}(2,\mathbb{Z})$ action preserves the set $\mathbb{H}$. Note that the element $-\mathrm{id} \in \mathrm{SL}(2,\mathbb{Z})$ acts trivially on $\mathbb{H}$, so one can also consider the $\mathrm{PSL}(2,\mathbb{Z}) \cong \mathrm{SL}(2,\mathbb{Z})/\{\pm\mathrm{id}\}$ action on $\mathbb{H}$.

Let

$$T = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \qquad \text{and} \qquad S = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

One has:

$$T(\tau) = \tau + 1; \qquad S(\tau) = -1/\tau; \qquad S^2 = (ST)^3 = I.$$

Consider the set

$$D = \left\{ z \in \mathbb{H} \colon |z| \geq 1 \text{ and } -\frac{1}{2} \leq \mathrm{Re}(z) \leq \frac{1}{2} \right\}.$$

We will show that D is a *fundamental domain* for the action of $\mathrm{PSL}(2,\mathbb{Z})$ on the upper half plane $\mathbb{H}$.

Theorem 5.44. *More precisely, we have:*

(a) For every $\tau \in \mathbb{H}$, there exists $g \in \mathrm{PSL}(2,\mathbb{Z})$ such that $g \cdot \tau \in D$.

(b) Suppose $\tau' = g\tau$ for some $\tau, \tau' \in D$ and $g \in \mathrm{PSL}(2,\mathbb{Z})\backslash\{I\}$, then:

- *either $\mathrm{Re}(\tau) = \pm 1/2$ and $\tau = \tau' \pm 1$,*

- *or* $|\tau| = 1$ *and* $\tau' = -1/\tau$.

(c) *Let* $\tau \in D$ *and let* $H_\tau = \{g \in \mathrm{PSL}(2,\mathbb{Z}) \mid g\tau = \tau\}$ *be the stabilizer of* τ. *Then:*

- $H_\tau = \langle S \rangle \cong \mathbb{Z}_2$ *if* $\tau = i$.
- $H_\tau = \langle ST \rangle \cong \mathbb{Z}_3$ *if* $\tau = e^{2\pi i/3} (= \omega)$.
- $H_\tau = \langle TS \rangle \cong \mathbb{Z}_3$ *if* $\tau = e^{\pi i/3} = (-1/\omega)$.
- $H_\tau = \{I\}$ *otherwise.*

Theorem 5.45. *The group* $\mathrm{PSL}(2,\mathbb{Z})$ *is generated by* S *and* T.

Let us prove both theorems together.

Proof. Let $G \subseteq \mathrm{PSL}(2,\mathbb{Z})$ be the subgroup of $\mathrm{PSL}(2,\mathbb{Z})$ generated by S and T. Let $\tau \in \mathbb{H}$. We will show that there exists $g \in G = \langle S, T \rangle$ so that $g\tau \in D$, which proves the first statement. Recall that

$$\mathrm{Im}\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \tau\right) = \frac{\mathrm{Im}(\tau)}{|c\tau + d|^2}.$$

Since c, d are integers, the number of pairs (c, d) such that $|c\tau + d|$ is less than a given number is *finite*. Therefore, there exists $g \in G$ such that $\mathrm{Im}(g\tau)$ is maximum. Now, choose an integer n so that $T^n g\tau$ has real part between $-\frac{1}{2}$ and $\frac{1}{2}$. Then the element $\tau' = T^n g\tau \in D$: indeed, it suffices to show that $|\tau'| \geq 1$; but if $|\tau'| < 1$, then the element $-1/\tau'$ would have imaginary part strictly greater than $\mathrm{Im}(\tau')$, contradiction. Thus $T^n g \in G$ has the desired property.

We now prove the second and third statements of the first theorem. Let $\tau \in D$ and $g \in \mathrm{PSL}(2,\mathbb{Z})$ so that $g\tau \in D$. By replacing (τ, g) by $(g\tau, g^{-1})$ if necessary, one may assume that $\mathrm{Im}(g\tau) \geq \mathrm{Im}(\tau)$, i.e. $|c\tau + d| \leq 1$. This is clearly impossible if $|c| \geq 2$, leaving the cases $c = 0, \pm 1$. If $c = 0$, then $d = \pm 1$ and g is the translation by $\pm b$. This is only possible for $\mathrm{Re}(\tau) = \pm 1/2$ and $g = T^{\pm 1}$. (Also, note that $T^{\pm 1}$ do not fix any point on D.)

If $c = 1$, then we have $|\tau + d| \leq 1$.

- If $d = 0$, then $|\tau| \leq 1$ hence $|\tau| = 1$ since $\tau \in D$. On the other hand, $ad - bc = 1$ implies $b = -1$, hence $g\tau = a - \frac{1}{\tau} \in D$. This is only if:
  - $a = 0$; in which case $g = S$, which sends $\{|\tau| = 1\} \cap D$ to itself. (Note that S has a unique fixed point $i \in D$.)

  - $a = 1$ and $\tau = -1/\omega$, which gives rise to the element $TS \in H_{-1/\omega}$.
  - $a = -1$ and $\tau = \omega$, which gives rise to the element $ST \in H_\omega$.
- If $d \neq 0$, then the only d and τ that satisfies $|\tau + d| \leq 1$ are:
  - $d = 1$ and $\tau = \omega$, which gives rise to the element $(ST)^2 \in H_\omega$.
  - $d = -1$ and $\tau = -1/\omega$, which gives rise to the element $(TS)^2 \in H_{-1/\omega}$.

This concludes the proof of the first theorem.

It remains to prove that $G = \mathrm{PSL}(2, \mathbb{Z})$. Let $g \in \mathrm{PSL}(2, \mathbb{Z})$. Choose any interior point of D, say $z_0 = 2i$. Consider the element $gz_0 \in \mathbb{H}$. By (a), there exists an element $g' \in G$ such that $g'gz_0 \in D$. By (b), we must have $g'g = I$. Thus $g \in G$. □

Remark 5.46. One can show that

$$\mathrm{PSL}(2, \mathbb{Z}) = \left\langle S, T \mid S^2 = (ST)^3 = 1 \right\rangle,$$

or equivalently, G is the *free product* of the cyclic group of order 2 generated by S and the cyclic group of order 3 generated by ST.

Remark 5.47. By the second theorem, to check the invariance of $\mathrm{SL}(2, \mathbb{Z})$, it suffices to check the invariance under the actions by T and S. For instance, a function $f\colon \mathbb{H} \to \mathbb{C}$ satisfies $f(\frac{a\tau+b}{c\tau+d}) = f(\tau)$ for all $\tau \in \mathbb{H}$ and $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}(2, \mathbb{Z})$ is equivalent to it satisfies $f(\tau) = f(\tau + 1) = f(\frac{-1}{\tau})$ for all $\tau \in \mathbb{H}$.

Remark 5.48. It turns out that the j-function we saw earlier, which can be written as

$$j(\tau) = \frac{\left(1 + 240 \sum_{n=1}^{\infty} \left(\sum_{d|n} d^3\right) q^n\right)^3}{q \prod_{n=1}^{\infty} (1 - q^n)^{24}} \quad \text{where} \quad q = e^{2\pi i \tau}$$

is the *simplest* non-constant holomorphic function on $\mathbb{H}$ invariant under $\mathrm{SL}(2, \mathbb{Z})$-action! It is really the simplest in the sense that *any* holomorphic function $f\colon \mathbb{H} \to \mathbb{C}$ satisfying $f(\tau) = f(\tau + 1) = f(\frac{-1}{\tau})$ for all $\tau \in \mathbb{H}$ can be written as a polynomial in $j(\tau)$.

Because of the fact stated in the previous remark, there are not many interesting functions that are $SL(2, \mathbb{Z})$-invariant on the nose. On the other hand,

there are many interesting functions on $\mathbb{H}$ (such as the Eisenstein series) that satisfies a slightly modified condition. These are the *modular forms*.

Definition 5.49 (non-precise version)**.** Let k be a positive integer. A holomorphic function $f\colon \mathbb{H} \to \mathbb{C}$ is called a *modular form of weight k* if

$$f\left(\frac{a\tau+b}{c\tau+d}\right) = (c\tau+d)^k f(\tau) \quad \text{for all } \tau \in \mathbb{H} \text{ and } \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}(2,\mathbb{Z}).$$

Or equivalently, $f(\tau+1) = f(\tau)$ and $f(-1/\tau) = \tau^k f(\tau)$ hold for all $\tau \in \mathbb{H}$.

Exercise. Let k be an odd integer. Show that the only modular form of weight k is the zero function. (Hint: Consider the action by $-\mathrm{id} \in \mathrm{SL}(2,\mathbb{Z})$.)

Remark 5.50. The notion of modular "*forms*" comes from the following observation. Consider the differential form $f(\tau)d\tau$ on $\mathbb{H}$. One can ask whether it is invariant under the $\mathrm{SL}(2,\mathbb{Z})$-action. Since

$$f\left(\frac{a\tau+b}{c\tau+d}\right) d\left(\frac{a\tau+b}{c\tau+d}\right) = f\left(\frac{a\tau+b}{c\tau+d}\right) \frac{d\tau}{(c\tau+d)^2},$$

we find that $f(\tau)d\tau$ is $\mathrm{SL}(2,\mathbb{Z})$-invariant if and only if $f\left(\frac{a\tau+b}{c\tau+d}\right) = (c\tau+d)^2 f(\tau)$, which is equivalent to $f(\tau)$ is a modular form of weight 2. In general, the differential form $f(\tau)(d\tau)^{k/2}$ is $\mathrm{SL}(2,\mathbb{Z})$-invariant if and only if $f(\tau)$ is a modular form of weight k.

Remark 5.51. One way to produce non-trivial modular function is that, if one can find two modular forms f_1, f_2 of the same weight which are linearly independent, then their ratio f_1/f_2 would give a (meromorphic) modular function. The j-function actually arises from the quotient of two modular forms of weight 12, which, as we will see later, is the smallest weight whose space of modular forms has dimension greater than one.

Recall the following facts about the Weierstrass $\wp$-function, with respect to a lattice $\Lambda_\tau = \langle 1, \tau \rangle$:

- We have the following expansion

$$\wp(z,\tau) = \frac{1}{z^2} + 3\widetilde{E_4}(\tau)z^2 + 5\widetilde{E_6}(\tau)z^4 + 7\widetilde{E_8}(\tau)z^6 + \cdots$$

  where $\widetilde{E_{2k}}(\tau) = \sum_{\lambda \in \Lambda \setminus \{0\}} \frac{1}{\lambda^{2k}}$ is the Eisenstein series.

- For any $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}(2,\mathbb{Z})$ we have

$$\wp\left(\frac{z}{c\tau+d}, \frac{a\tau+b}{c\tau+d}\right) = (c\tau+d)^2 \wp(z,\tau).$$

It is then not hard to deduce that

$$\widetilde{E_{2k}}(\tau) \text{ is a modular form of weight } 2k \text{ for any } k \geq 2.$$

Remark 5.52. Like modular functions, modular forms are also rare. In fact, we will show later that any modular form can be written as a polynomial in $\widetilde{E_4}$ and $\widetilde{E_6}$! In particular, the space of modular forms of weight k is always finite dimensional for any k.

Definition 5.53 (Precise version)**.** Let k be a positive integer. A holomorphic function $f\colon \mathbb{H} \to \mathbb{C}$ is called a *modular form of weight k* if

- $f\left(\frac{a\tau+b}{c\tau+d}\right) = (c\tau+d)^k f(\tau)$ for all $\tau \in \mathbb{H}$ and $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}(2,\mathbb{Z})$.
- $f(\tau)$ is bounded as $\mathrm{Im}(\tau) \to \infty$.

Remark 5.54. Since f is invariant under $\tau \mapsto \tau+1$, it is convenient to introduce the variable $q = \exp(2\pi i\tau)$, where f can be considered as a function in q on the punctured unit disk $\mathbb{D}_1^\times(0) = \{q\colon 0 < |q| < 1\}$. Then the second condition that$f(\tau)$ is bounded as $\mathrm{Im}(\tau) \to \infty$ is equivalent to $f(q)$ is bounded near $q = 0$. This actually is equivalent to f can be extended holomorphic to the whole unit disk $\mathbb{D}_1(0) = \{q\colon |q| < 1\}$.

Let us compute the *q-expansion* (i.e. the expansion near $q = 0$) of the Eisenstein series.

Proposition 5.55. *Let $k \geq 4$ even and $Im(\tau) > 0$. We have*

$$\widetilde{E_k}(\tau) = 2\zeta(k) + \frac{2(-1)^{k/2}(2\pi)^k}{(k-1)!} \sum_{r=1}^{\infty} \sigma_{k-1}(r) e^{2\pi i \tau r}.$$

Here $\sigma_{k-1}(r) = \sum_{d|r} d^{k-1}$ is the divisor function.

Proof. Recall that

$$\widetilde{E_k}(\tau) = \sum_{(m,n)\neq(0,0)} \frac{1}{(m+n\tau)^k} = 2\zeta(k) + 2\sum_{n\geq 1}\sum_{m\in\mathbb{Z}} \frac{1}{(m+n\tau)^k}.$$

The right hand side can be computed by the *Poisson summation formula*. Let f be a function with certain appropriate regularity and decay conditions, one can define its Fourier transform as

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i x \xi} \, \mathrm{d}x,$$

and the *Poisson summation formula* states that

$$\sum_{n \in \mathbb{Z}} f(n) = \sum_{n \in \mathbb{Z}} \hat{f}(n).$$

By applying the Poisson summation formula to the function $f(z) = (\tau + z)^{-k}$, one obtains

$$\sum_{n \in \mathbb{Z}} \frac{1}{(\tau + n)^k} = \frac{(-2\pi i)^k}{(k-1)!} \sum_{m=1}^{\infty} m^{k-1} e^{2\pi i m \tau}.$$

Thus

$$\begin{aligned} \widetilde{E_k}(\tau) &= 2\zeta(k) + 2 \sum_{n \geq 1} \sum_{m \in \mathbb{Z}} \frac{1}{(m + n\tau)^k} \\ &= 2\zeta(k) + \frac{2(-1)^{k/2}(2\pi)^k}{(k-1)!} \sum_{n \geq 1} \sum_{\ell=1}^{\infty} \ell^{k-1} e^{2\pi i \ell (n\tau)} \\ &= 2\zeta(k) + \frac{2(-1)^{k/2}(2\pi)^k}{(k-1)!} \sum_{r \geq 1} \sigma_{k-1}(r) e^{2\pi i \tau r}. \end{aligned}$$

□

Exercise. Show that the Eisenstein series $\widetilde{E_{2k}}(\tau)$ is a modular form of weight $2k$ for all $k \geq 2$. (One needs to check its boundedness near $q = 0$.)

Notation. The *Bernoulli numbers* B_n are a sequence of rational numbers, which can be defined via

$$\frac{x}{e^x - 1} = \sum_{k \geq 0} \frac{B_k x^k}{k!}.$$

The first few Bernoulli numbers are:

k	0	1	2	3	4	5	6	7	8	9	10	$\cdots$
B_k	1	$-\frac{1}{2}$	$\frac{1}{6}$	0	$-\frac{1}{30}$	0	$\frac{1}{42}$	0	$-\frac{1}{30}$	0	$\frac{5}{66}$	$\cdots$

We have

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

Hence, for $k \geq 4$ even we have

$$\widetilde{E_k}(\tau) = 2\zeta(k)\left(1 - \frac{2k}{B_k}\sum_{r\geq 1}\sigma_{k-1}(r)q^r\right), \qquad \text{where } q = e^{2\pi i\tau}.$$

One can normalized the series $\widetilde{E_k}(\tau)$ as

$$E_k(\tau) \coloneqq 1 - \frac{2k}{B_k}\sum_{r\geq 1}\sigma_{k-1}(r)q^r$$

which still is a modular form of weight k for any $k \geq 4$ even. By plugging in the first few Bernoulli numbers, one obtains the formula we saw earlier

$$E_4(\tau) = 1 + 240\sum_{r\geq 1}\sigma_3(r)q^r; \qquad E_6(\tau) = 1 - 504\sum_{r\geq 1}\sigma_5(r)q^r;$$

$$E_8(\tau) = 1 + 480\sum_{r\geq 1}\sigma_7(r)q^r; \qquad E_{10}(\tau) = 1 - 264\sum_{r\geq 1}\sigma_9(r)q^r.$$

The following theorem is crucial for understanding the space of modular forms.

Theorem 5.56 (Valence formula)**.** *Let $f\colon \mathbb{H} \to \mathbb{C}$ be a nonzero modular form of weight k. Then*

$$v_\infty(f) + \frac{1}{2}v_i(f) + \frac{1}{3}v_\omega(f) + \sum_{\tau\in\mathbb{H}'/\mathrm{SL}(2,\mathbb{Z})} v_\tau(f) = \frac{k}{12}.$$

Here the notion $v_z(f)$ denotes the order of zero at z; the summation runs through the orbits in $\mathbb{H}/SL(2,\mathbb{Z})$ other than those of i and ω.

Before proving this theorem, let us demonstrate some of its applications. Denote M_k the (complex vector) space of modular forms of weight k.

Corollary 5.57. *We have*

(a) $M_k = \{0\}$ if $k < 0$ or $k = 2$.
(b) $M_0 \cong \mathbb{C}$ consists only of constant functions.

(c) $M_4 = \langle E_4 \rangle$ *is a one-dimensional vector space generated by the Eisenstein series* E_4*, which has simple zeros at the orbit of* ω *and has no other zeros.*
(d) $M_6 = \langle E_6 \rangle$*, which has simple zeros at the orbit of* i *and has no other zeros.*
(e) $M_8 = \langle E_8 \rangle$*, which has double zeros at the orbit of* ω *and has no other zeros. In particular, we have* $E_8 = E_4^2$*.*
(f) $M_{10} = \langle E_{10} \rangle$*, which has simple zeros at the orbits of* ω *and* i *and has no other zeros. In particular, we have* $E_{10} = E_4 E_6$*.*

Proof. Part (a) follows directly from the valence formula. To prove (b), let f be a modular form of weight 0. Since the constant function $g = f(2i)$ is a modular form of weight 0 (the point $2i$ can be chosen arbitrarily), so is $f - g \in M_0$. But $f - g$ now has a zero at $2i$, by the valence formula, one must have $f = g$.

To prove (c), observe that any modular form of weight 4 has simple zeros at the orbit of ω and has no other zeros. Therefore, given any two modular forms f_1, f_2 of weight 4, the ratio f_1/f_2 is a modular form of weight zero, therefore a constant function. The remaining statements can be proved similarly. □

Proposition 5.58. *The smallest weight* k *that admits linearly independent modular forms is* $k = 12$*. In fact,* M_{12} *is of two dimension, generated by* $M_{12} = \langle E_4^3, E_6^2 \rangle$*.*

Proof. Let us denote

$$\Delta = \frac{E_4^3 - E_6^2}{1728} \in M_{12}$$

which is a modular form of weight 12, and its q-expansion has vanishing constant term (in fact, $\Delta(q) = q + (\text{higher order terms})$). Therefore $v_\infty(\Delta) = 1$. By the valence formula, we find that Δ has no other zeros except at $\tau = \infty$ (equivalently, at $q = 0$).

Let $f \in M_{12}$, with its q-expansion given by

$$f(q) = a_0 + a_1 q + \cdots .$$

Observe that $f - a_0 E_4^3 \in M_{12}$, which has a zero at $q = 0$. By the same argument as in the previous corollary, one deduces that $f - a_0 E_4^3$ is a constant multiple of Δ. Thus $f \in \langle E_4^3, E_6^2 \rangle$.

Finally, it is clear that E_4^3 and E_6^2 are linearly independent since the locations of their zeros are different. □

Theorem 5.59. *Any modular form is a polynomial in E_4 and E_6. In other words, the space M_k has a basis given by*

$$M_k = \left\langle M_4^a M_6^b \mid a, b \geq 0,\ 4a + 6b = k \right\rangle.$$

Proof. We prove the statement by induction on k. The statement is true for $k \leq 12$ by our previous discussions. Now, let $f \in M_k$ where $k > 12$ an even integer. Choose $a, b \geq 0$ so that $4a + 6b = k$. Then

$$f - f(\infty) \cdot E_4^a E_6^b \in M_k$$

has a zero at $\tau = \infty$. Therefore we have

$$\frac{f - f(\infty) \cdot E_4^a E_6^b}{\Delta} \in M_{k-12},$$

which can be written as a polynomial in E_4 and E_6 by induction hypothesis, thus so can f. □

Definition 5.60. The j-function is defined to be

$$j(\tau) = \frac{E_4(\tau)^3}{\Delta(\tau)} = 1728 \frac{E_4(\tau)^3}{E_4(\tau)^3 - E_6(\tau)^2}.$$

It satisfies the following properties:

- $j(\tau)$ is a modular function, i.e. invariant under the $\mathrm{SL}(2, \mathbb{Z})$-action.
- $j(\tau)$ is holomorphic on $\mathbb{H}$, and has a simple pole at $\tau = \infty$.
- $j(\tau)$ has zeros of order 3 at ω and its $\mathrm{SL}(2, \mathbb{Z})$-orbit.
- $j(\tau) - 1728 = \frac{E_6(\tau)^2}{\Delta(\tau)}$ has zeros of order 2 at i and its orbit.

A perhaps unexpected application of the j-function is a proof of the *little Picard theorem*.

Theorem 5.61 (Little Picard theorem)**.** *Let $f : \mathbb{C} \to \mathbb{C}$ be an entire function. Suppose f omits (at least) two values, i.e. there exists z_1, z_2 so that $f^{-1}(z_1) = f^{-1}(z_2) = \emptyset$. Then f is a constant function.*

Proof. First, we claim that the j-function defines a bijection between $\mathbb{H}/\mathrm{PSL}(2, \mathbb{Z})$ onto $\mathbb{C}$. To see this, one has to show that for all $\lambda \in \mathbb{C}$, the modular form $E_4(\tau)^3 - \lambda\Delta(\tau)$ has a unique zero $\tau \in \mathbb{H}$ up to the $\mathrm{PSL}(2, \mathbb{Z})$-action. This

follows directly from the valence formula.

Suppose f is an entire function which omits two values. Up to composing f with an invertible linear map, one can assume that it omits $\{0, 1728\}$. The map $\mathbb{H}'/\mathrm{PSL}(2,\mathbb{Z}) \to \mathbb{C}\backslash\{0,1728\}$ is biholomorphic (which admits an inverse). The composition

$$\mathbb{C} \xrightarrow{f} \mathbb{C}\backslash\{0,1728\} \to \mathbb{H}'/\mathrm{PSL}(2,\mathbb{Z}) \xrightarrow{1/z} \mathbb{D}_1(0)$$

is therefore a bounded entire function, thus is a constant map by Liouville's theorem. □

Theorem 5.62. *Any meromorphic modular function is a rational polynomial of $j(\tau)$.*

Proof. Let f be a meromorphic modular function. By suitably multiplying $j(\tau)-j(z_0)$, one can assume that f is holomorphic on $\mathbb{H}$. Write the q-expansion of f near $q = 0$ as

$$f(q) = a_{-n}q^{-n} + \cdots .$$

Then $f(q) - a_{-n}j(q)^n$ is also holomorphic on $\mathbb{H}$, and with pole at $q = 0$ of order strictly less than n:

$$f(q) - a_{-n}j(q)^n = b_{-(n-1)}q^{-(n-1)} + \cdots .$$

By continuing this process, one deduces that there exist constants $a_{-n}, \ldots, a_{-1}$ so that

$$f(q) - a_{-n}j(q)^n - \cdots - a_{-1}j(q)$$

is modular and holomorphic on $\mathbb{H} \cup \{\infty\}$, therefore is a constant function. □

Remark 5.63. A more natural way to understand the above theorem is that, the j-function defines an isomorphism of the *compactification* $\overline{\mathbb{H}/\mathrm{PSL}(2,\mathbb{Z})}$ onto the Riemann sphere $\mathbb{CP}^1 = \mathbb{C} \cup \{\infty\}$. A meromorphic modular function is nothing but a meromorphic function on $\overline{\mathbb{H}/\mathrm{PSL}(2,\mathbb{Z})}$. The above theorem amounts to the well-known fact that the only meromorphic functions on $\mathbb{CP}^1$ are the rational functions.

Lecture 8

Now we return to the proof of the valence formula. Before that, let us state the following "arc version" of the Cauchy integral formula.

Exercise. Let f be a holomorphic function on a neighborhood of z_0. Let $0 < \theta_r \leq 2\pi$ be a number depending on $r > 0$, which satisfies $\lim_{r\to 0} \theta_r = \theta$ where

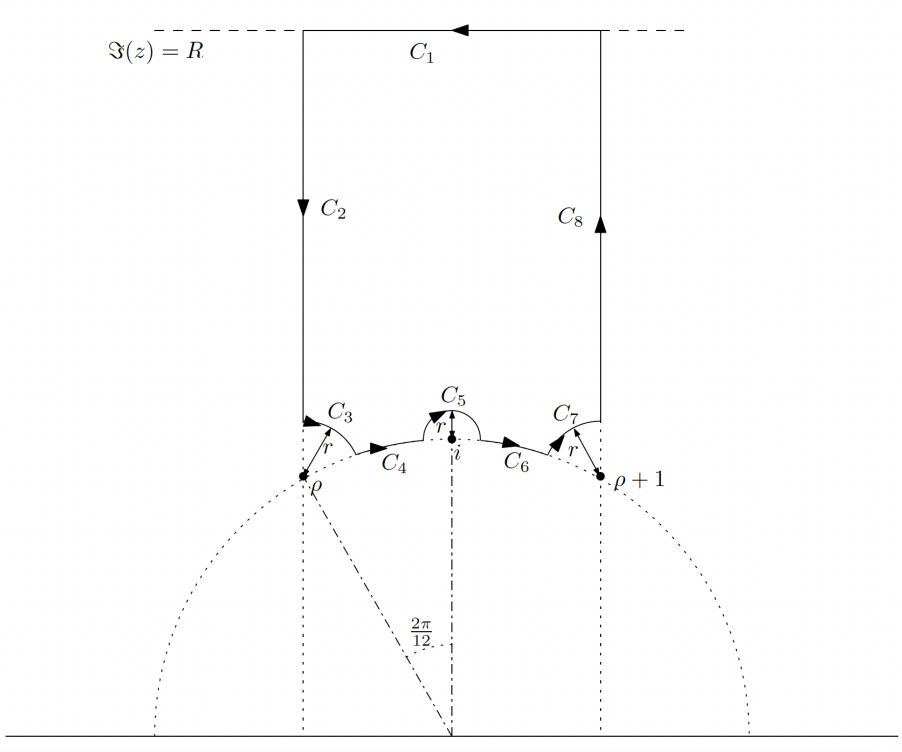
$0 < \theta \leq 2\pi$. Let $C(z_0, r, \theta_r)$ be an arc of a circle, of radius r and angle θ_r around the point z_0. Then we have

$$\lim_{r\to 0} \int_{C(z_0,r,\theta_r)} \frac{f(z)}{z - z_0}\,\mathrm{d}z = \theta i f(z_0).$$

Similarly, we also have the "arc version" of the argument principle

$$\lim_{r\to 0} \int_{C(z_0,r,\theta_r)} \frac{f'(z)}{f(z)}\,\mathrm{d}z = \theta i v_{z_0}(f).$$

Proof of the valence formula. Consider the following closed loop in the fundamental domain. (The ρ in the figure is our ω.) Let C be the union of



these paths, which forms a closed loop. Let $f\colon \mathbb{H} \to \mathbb{C}$ be a modular form of weight k. There exists $R > 0$ large enough so that f has no zeros in $\{z \in \mathbb{H}\colon |\mathrm{Re}(z)| < \frac{1}{2},\ \mathrm{Im}(z) > R\}$ (this follows from the stronger form of the

local determine global principle). By the argument principle, we have

$$\frac{1}{2\pi i}\int_C \frac{f'(z)}{f(z)}\,\mathrm{d}z = \sum_{\tau\in\mathbb{H}'/\mathrm{SL}(2,\mathbb{Z})} v_\tau(f).$$

Thus, to prove the valence formula, it suffices to show that

$$\frac{1}{2\pi i}\int_C \frac{f'(z)}{f(z)}\,\mathrm{d}z = \frac{k}{12} - v_\infty(f) - \frac{1}{2}v_i(f) - \frac{1}{3}v_\omega(f).$$

(a) Integration along C_1: Consider the change of variable $q = \exp(2\pi i z)$. Then, in terms of the q-coordinate, the path C_1 becomes a loop around $q = 0$ of radius $\exp(-2\pi R)$ traveling *clockwisely*. By the argument principle, we have

$$\frac{1}{2\pi i}\int_{C_1} \frac{f'(z)}{f(z)}\,\mathrm{d}z = -v_\infty(f).$$

(b) Integrations along C_2 and C_8: Since f is a modular form, it satisfies $f(z) = f(z+1)$. Therefore the integrations along C_2 and C_8 canceled with each other.

(c) Integration along C_5: By the arc version of the argument principle, we have

$$\frac{1}{2\pi i}\int_{C_5} \frac{f'(z)}{f(z)}\,\mathrm{d}z \xrightarrow{r\to 0} -\frac{1}{2}v_i(f).$$

(d) Integrations along C_3 and C_7: By the arc version of the argument principle, we have

$$\frac{1}{2\pi i}\int_{C_3} \frac{f'(z)}{f(z)}\,\mathrm{d}z + \frac{1}{2\pi i}\int_{C_7} \frac{f'(z)}{f(z)}\,\mathrm{d}z \xrightarrow{r\to 0} 2\cdot\left(-\frac{1}{6}v_\omega(f)\right) = -\frac{1}{3}v_\omega(f).$$

(e) Integrations along C_4 and C_6: This is the most interesting part of the computation, where the weight k of the modular form gets involved. Observe that the map $S\colon z \mapsto -\frac{1}{z}$ sends C_4 to $-C_6$. The modularity of f implies that $f(z) = z^{-k}f(S(z))$. Thus

$$f'(z) = -kz^{-k-1}f(S(z)) + z^{-k}f'(S(z))S'(z),$$

hence

$$\frac{f'(z)}{f(z)} = -\frac{k}{z} + \frac{f'(S(z))S'(z)}{f(S(z))}.$$

Therefore

$$\frac{1}{2\pi i}\int_{C_4}\frac{f'(z)}{f(z)}\,\mathrm{d}z = \frac{1}{2\pi i}\int_{C_4}-\frac{k}{z}\,\mathrm{d}z - \frac{1}{2\pi i}\int_{C_6}\frac{f'(z)}{f(z)}\,\mathrm{d}z.$$

Again by the arc version of the argument principle, we have

$$\frac{1}{2\pi i}\int_{C_4}\frac{f'(z)}{f(z)}\,\mathrm{d}z + \frac{1}{2\pi i}\int_{C_6}\frac{f'(z)}{f(z)}\,\mathrm{d}z = -\frac{1}{2\pi i}\int_{C_4}\frac{k}{z}\,\mathrm{d}z = \frac{k}{12}.$$

This completes the proof. □

5.5. **Sum of four squares.** Let us return to our motivating problem, the sum of squares problem. We would like to understand the counting

$$r_k(n) = \#\{(x_1,\dots,x_k)\in\mathbb{Z}^2 \mid x_1^2+\cdots+x_k^2 = n\}.$$

Consider the *theta function*

$$\theta(\tau) = \sum_{n=-\infty}^{\infty} e^{2\pi i\tau n^2} = \sum_{n=-\infty}^{\infty} q^{n^2} = 1 + 2\sum_{n=1}^{\infty} q^{n^2}.$$

It is not hard to see that

$$\theta(\tau)^k = \sum_{n=0}^{\infty} r_k(n) q^n.$$

The problem then reduces to understanding the coefficients of powers of the theta function. It turns out that the theta function satisfies certain modular properties, which will allow us to write down an explicit formula for $r_k(n)$. The key fact is that θ satisfies the following transformation formula.

Lemma 5.64.

$$\theta\left(\frac{-1}{4\tau}\right) = \sqrt{-2i\tau}\,\theta(\tau).$$

Proof. The proof uses again the *Poisson summation formula*. Consider the function $f(x) = e^{2\pi i\tau x^2}$. Then we have $\theta(\tau) = \sum_{n\in\mathbb{Z}} f(n)$. Its Fourier transform

is

$$\begin{aligned}\hat{f}(n) &= \int_{\mathbb{R}} f(x)e^{-2\pi ixn}\,\mathrm{d}x \\ &= \int_{\mathbb{R}} \exp\left(2\pi i\tau\left(x-\frac{n}{2\tau}\right)^2 - \frac{\pi i}{2\tau}n^2\right)\mathrm{d}x \\ &= e^{-\frac{\pi i}{2\tau}n^2}\frac{1}{\sqrt{-2i\tau}}.\end{aligned}$$

By the Poisson summation formula, we have

$$\theta(\tau) = \sum_{n\in\mathbb{Z}} f(n) = \sum_{n\in\mathbb{Z}} \hat{f}(n) = \theta\left(\frac{-1}{4\tau}\right)\frac{1}{\sqrt{-2i\tau}}.$$

□

Corollary 5.65.

$$\theta\left(\frac{\tau}{4\tau+1}\right) = \sqrt{4\tau+1}\theta(\tau).$$

Proof.

$$\begin{aligned}\theta\left(\frac{\tau}{4\tau+1}\right) &= \theta\left(-\frac{1}{4\left(\frac{-1}{4\tau}-1\right)}\right) = \sqrt{2i\left(\frac{1}{4\tau}+1\right)}\theta\left(\frac{-1}{4\tau}-1\right) \\ &= \sqrt{2i\left(\frac{1}{4\tau}+1\right)}\theta\left(\frac{-1}{4\tau}\right) = \sqrt{2i\left(\frac{1}{4\tau}+1\right)}\sqrt{-2i\tau}\theta(\tau) \\ &= \sqrt{4\tau+1}\theta(\tau).\end{aligned}$$

□

Thus, the function $\theta(\tau)^{2k}$ satisfies the following:

- $\theta(\tau+1)^{2k} = \theta(\tau)^{2k}$.
- $\theta\left(\frac{\tau}{4\tau+1}\right)^{2k} = (4\tau+1)^k\theta(\tau)^{2k}$.

Definition 5.66. Let $\Gamma \subseteq \mathrm{PSL}(2,\mathbb{Z})$ be a subgroup. We say a holomorphic function $f\colon \mathbb{H}\to\mathbb{C}$ is a *modular form of weight k with respect to Γ* if

$$f\left(\frac{a\tau+b}{c\tau+d}\right) = (c\tau+d)^k f(\tau) \quad \text{for all } \tau\in\mathbb{H} \text{ and } \begin{bmatrix}a & b\\ c & d\end{bmatrix}\in\Gamma.$$

The lemma above shows that the function $\theta(\tau)^{2k}$ is a modular form of weight k with respect to the group

$$\Gamma_1(4) \coloneqq \left\langle \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 4 & 1 \end{bmatrix} \right\rangle.$$

Let us focus on the case of $2k = 4$, i.e. the sum of *four* squares problem. The following theorem involves more detailed study of the modular curve $\mathbb{H}/\Gamma_1(4)$, for which the proof we omit. (Essentially, one has to do the same analysis as we did in the last subsection, with the group $\mathrm{PSL}(2,\mathbb{Z})$ replaced by its subgroup $\Gamma_1(4)$.)

Theorem 5.67. *The space $M_2(\Gamma_1(4))$ of modular forms of weight 2 with respect to $\Gamma_1(4)$ is 2-dimensional, with basis given by*

$$M_2(\Gamma_1(4)) = \mathrm{Span}\{E_2(\tau) - 2E_2(2\tau), E_2(\tau) - 4E_2(4\tau)\},$$

where

$$E_2(\tau) = -\frac{1}{24} + \sum_{n=1}^{\infty} \sigma_1(n) q^n, \quad \sigma_1(n) = \sum_{d|n} d.$$

Now, by comparing the first two coefficients of the q-expansions of $\theta(\tau)^4$ and the basis functions $E_2(\tau) - 2E_2(2\tau)$ and $E_2(\tau) - 4E_2(4\tau)$, one obtains

$$\theta(\tau)^4 = 8(E_2(\tau) - 4E_2(4\tau)).$$

Therefore,

$$r_4(n) = 8\left(\sum_{d|n} d - 4\sum_{d|\frac{n}{4}} d\right) = 8\sum_{\substack{d|n \\ 4\nmid d}} d.$$

This is the *Jacobi four-square theorem.*

6. Knot invariants and categorification

We discuss certain knot invariants and their "categorification" in this section. Some references that might be helpful: [12, 13].

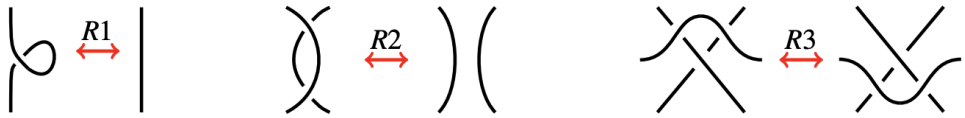
6.1. **Jones polynomial.**

Definition 6.1. An *oriented knot* $K \subseteq \mathbb{R}^3$ is a subset of the form $K = f(S^1)$ where $f\colon S^1 \to \mathbb{R}^3$ is a smooth embedding. We say two knots K_0, K_1 are *equivalent* if there is a smooth map $F\colon S^1 \times [0,1] \to \mathbb{R}^3$ so that $K_0 = F|_{S^1\times\{0\}}$, $K_1 = F|_{S^1\times\{1\}}$, and $K_t = F|_{S^1\times\{t\}}$ is a knot for each t.

One can generalize the notion of oriented knots to oriented links. An *oriented n-component link* in $L \subseteq \mathbb{R}^3$ is a subset of the form $L = f(\coprod^n S^1)$ where $f\colon \coprod^n S^1 = S^1 \coprod \cdots \coprod S^1 \to \mathbb{R}^3$ is a smooth embedding. The notion of equivalence between links can be defined in the same way.

To draw pictures of a link, we consider its image under a linear projection $\mathbb{R}^3 \to \mathbb{R}^2$. Note that any given link can be represented by various different

planar diagrams. We say two diagrams are related by *Reidemeister moves* if we can obtain the second diagram by applying the three moves (R1, R2, or R3) to some small regions of the first diagram. It is easy to see that if two



diagrams are related by Reidemeister moves, then they represent equivalent links.

Example 6.2. This example shows how a diagram of a knot with 3 crossings can be transformed into the *unknot* by a sequence of Reidemeister moves.

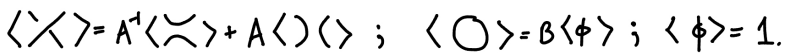
Theorem 6.3 (Reidemeister, 1932)**.** *Two diagrams represent the equivalent link if and only if they can be related by a sequence of Reidemeister moves.*

It would be nice to have some ways of telling when two diagrams represent different links. Here, the idea is to define certain *link invariants*: one would like to associate certain invariants (numbers, polynomials, or other objects) to each planar diagram, so that two diagrams have the same invariant if they are related by Reidemeister moves.

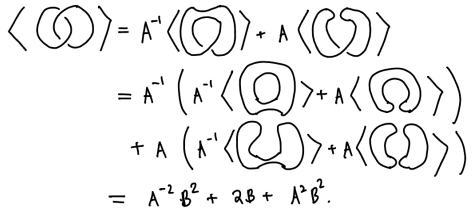
Let us try to define a link invariant with the help of the *Kauffman bracket*, which is a function

$$\langle \cdot \rangle : \mathcal{D} \to \mathbb{Z}[A^{\pm 1}, B]$$

(let $\mathcal{D}$ denote the set of all link diagrams up to isotopy), satisfying the local relations:



Example 6.4. One can compute the Kauffman bracket of the *Hopf link* as follows.



Remark 6.5. The procedure above can be applied to any planar diagram D. Each crossing has two *resolutions*, which we call the 0 *resolution* and 1 *resolution*. If D has n crossings, then there will be 2^n ways to resolve all of them, so

we can express $\langle D \rangle$ as the sum of 2^n terms which involve the bracket of diagrams with no crossings (i.e. disjoint union of circles). These 2^n diagrams are

in bijection with the vertices of the n-dimensional cube $[0,1]^n$. If D_v denotes the planar diagram corresponds to vertex v, then

$$\langle D \rangle = \sum_v A^{n-2|v|} B^{|D_v|}$$

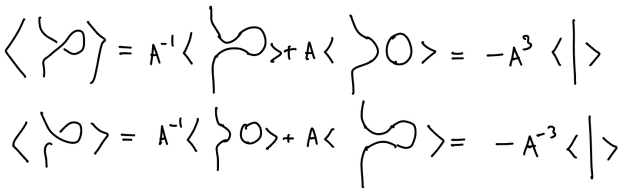
where $|v|$ denotes the sum of the coefficients of v (which is a string of $\{0,1\}$ of length n), and $|D_v|$ denotes the number of components of D_v.

In order to obtain a link invariant from the Kauffman bracket, one needs to consider how the bracket changes under Reidemeister moves.

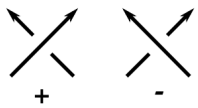
Exercise. In order for the bracket to be invariant under R2 and R3 (the second and third Reidemeister moves), we must set

$$B = -A^{-2} - A^2.$$

It remains to consider the R1 move. Disappointingly, we find that the bracket



is *not* invariant under the R1 move, and hence *not* a link invariant. But it is very close, and indeed there is a fix for our problem. The fix involves paying attention to the *orientations*. There are two possible crossings, which we refer to as *positive* and *negative* crossings. We denote $n_\pm(D)$ the number



of positive/negative crossings of a planar diagram D, and define the *writhe* of D to be

$$w(D) = n_+(D) - n_-(D).$$

It is clear that the writhe is invariant under R2 and R3 moves. On the other hand, an R1 move will either increase or decrease the writhe by 1. We can use it to counteract the change in the Kauffman bracket under an R1 move, and therefore obtain a link invariant.

Definition 6.6. Let D be an oriented link diagram. Its *Jones polynomial* is defined to be

$$V(D) = (-A^3)^{-w(D)} \langle D \rangle \in \mathbb{Z}[A^{\pm 1}].$$

It is an invariant of oriented links.

Example 6.7. With this definition, we have $V(\emptyset) = 1$, and $V(\bigcirc) = -A^{-2} - A^2$. More generally, $V(\bigcirc^n) = (-A^{-2} - A^2)^n$, where $\bigcirc^n$ denotes the n-component unlink. The Jones polynomial of the (positively oriented) Hopf link is $1 + A^{-4} + A^{-8} + A^{-12}$.

Remark 6.8. Using the formula $\langle D \rangle = \sum_v A^{n-2|v|} B^{|D_v|}$ we saw earlier, one can see that $V(D) \in \mathbb{Z}[A^{\pm 2}]$. Therefore, it is more common to use the variable $q = -A^{-2}$ for Jones polynomial. Then we have $V(\bigcirc) = q + q^{-1}$ and $V(\text{Hopf link}) = 1 + q^2 + q^4 + q^6$.

Remark 6.9. $V(D)$ satisfies the *skein relation*:

In fact, one can define the Jones polynomial using the skein relation.

Remark 6.10. There are also other variants of the Jones polynomial defined by replacing $q^{\pm 2}$ on the left hand side of the above formula by $q^{\pm n}$, denoted by $P_n(q) \in \mathbb{Z}[q, q^{-1}]$.

- $P_0(q)$ is the *Alexander polynomial* of the link.
- $P_1(q) \equiv 1$.
- $P_2(q)$ is the *Jones polynomial* of the link.

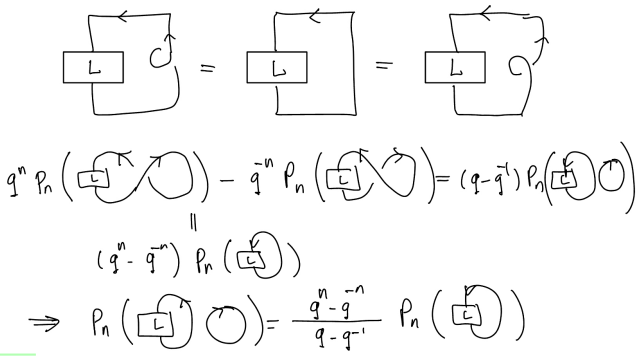
From the computational perspective, for $n \geq 2$ the polynomial becomes harder to compute; while the Alexander polynomial can be computed in polynomial time.

Remark 6.11. Our strategy of checking invariance under the Reidemeister moves is an effective way of proving that some quantity is a link invariant, but it is a terrible way of finding such invariants to start with. The definition of the Jones polynomial we have given is completely elementary, but remained undiscovered for 100 years after mathematicians first started thinking about

knots. Jones arrived at his original definition by thinking about something entirely different – representations of von Neumann algebras.

After Jones's discovery, Witten realized that the Jones polynomial should fit into a much broader theory of invariants of 3-manifolds defined using Chern–Simons theory. His work launched an entire industry devoted to the study of these *quantum invariants*, both in physics and mathematics. It is worth knowing that Witten's approach assigns a polynomial invariant of knots to each complex Lie algebra $\mathfrak{g}$ equipped with a representation; the Jones polynomial corresponds to the vector representation of $\mathfrak{sl}_2$.

Remark 6.12. This computation suggests that, if we want P_n to be a *tensor*



functor (don't worry about what this means for now), we should set

$$P_n(\bigcirc) = \frac{q^n - q^{-n}}{q - q^{-1}}.$$

In the case of $n = 2$ (Jones polynomial), this is precisely what we did: $V(\bigcirc) = q + q^{-1}$. The polynomial

$$\frac{q^n - q^{-n}}{q - q^{-1}} = q^{n-1} + q^{n-2} + \cdots + q^{-(n-1)}$$

is also referred to as the *quantum integers*. We will encounter them again in later sections.

Lecture 9

6.2. **Categorification.** The moral of *categorification* is to consistently convert integers into vector spaces (or free abelian groups). For instance, to natural numbers, we can assign to them vector spaces with the corresponding dimensions. Then, the operations on integers are "upgraded" into:

$\mathbb{N}$	Categorification
$n \in \mathbb{N}$	V_n, where $\dim(V_n) = n$
$n + m$	$V_n \oplus V_m$
$n \cdot m$	$V_n \otimes V_m$
$n - m$	??

To categorify "$n - m$", we are forced to introduce *chain complexes* of vector spaces, whose *Euler characteristic* χ is the alternating sum of dimensions. For instance,

$$\chi(0 \to V_n \xrightarrow{d_0} V_m \to 0) = n - m.$$

In general, a *chain complex* is a sequence of vector spaces connected by linear maps

$$\cdots \xrightarrow{d_{n-1}} V^n \xrightarrow{d_n} V^{n+1} \xrightarrow{d_{n+1}} \cdots, \qquad \text{where } d_n \circ d_{n-1} = 0 \text{ for all } n.$$

Moreover, tensor products of complexes can be defined: suppose we have two complexes $V^\bullet = (\cdots \to V^i \xrightarrow{d} V^{i+1} \cdots)$ and $W^\bullet = (\cdots \to W^i \xrightarrow{d} W^{i+1} \cdots)$, then their tensor product is defined to be the complex $T^\bullet$ where

$$T^p = \bigoplus_{k \in \mathbb{Z}} \left(V^k \otimes W^{p-k}\right)$$

with differential given by

$$d(v^i \otimes w^j) = (dv^i) \otimes w^j + (-1)^i v^i \otimes (dw^j).$$

It also satisfies that

$$\chi(V^\bullet \otimes W^\bullet) = \chi(V^\bullet)\chi(W^\bullet).$$

Example 6.13. Here is a nice example of categorification. Let X be a topological space. One can associate to it the *Euler characteristic* $\chi(X) \in \mathbb{Z}$. For instance, when X is a polytope in $\mathbb{R}^3$, then $\chi(X)$ equals to (the number of vertices) $-$ (the number of edges) $+$ (the number of faces), which is $\chi(X) = 2$ by Euler's formula.

The Euler characteristic admits an "upgrade" as follows: For any topological space X, there exists a *chain complex* of real vector spaces

$$\cdots \xrightarrow{d_{n+1}} C_n(X, \mathbb{R}) \xrightarrow{d_n} C_{n-1}(X, \mathbb{R}) \xrightarrow{d_{n-1}} \cdots$$

where each $d_\bullet$ is a linear map and satisfies $d_n \circ d_{n+1} = 0$ for all n (this is the *singular chain complex*). The *homology groups* of X is then defined to be

$$H_n(X, \mathbb{R}) = \frac{\mathrm{Ker}(d_n)}{\mathrm{Im}(d_{n+1})}.$$

The Euler characteristic of X can be recovered by

$$\chi(X) = \sum_{n \in \mathbb{Z}} (-1)^n \dim(H_n(X, \mathbb{R})).$$

The chain complex and the homology groups certainly carry much finer topological information than the Euler characteristic.

By applying certain 1+1 *dimensional topological quantum field theory (TQFT)* to a link L, one obtains a categorification of the Jones polynomial, the *Khovanov homology*. It is a *bi-graded* abelian group $\mathrm{Kh}^{i,j}(L)$, whose graded Euler characteristic recovers the Jones polynomial

$$\chi(\mathrm{Kh}(L)) = \sum_{i,j} (-1)^i q^j \dim \mathrm{Kh}^{i,j}(L) = V(L).$$

The Khovanov homology is closely related to HOMFLY homology, Floer homology, Fukaya categories, etc., which are the central objects of current study in low-dimensional (especially 3 and 4) geometric topology.

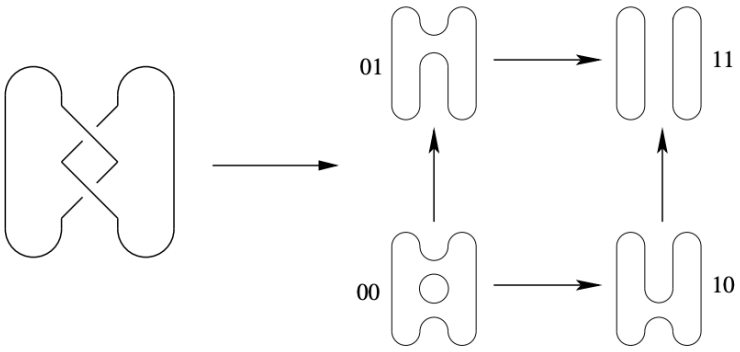
To define $\mathrm{Kh}(L)$, we first represent L by a planar diagram D. To such a diagram, Khovanov assigns a bi-graded chain complex $\mathrm{CKh}(D)$, and $\mathrm{Kh}(D)$ is its homology. Khovanov shows that if D_1 and D_2 represent the same link, then $\mathrm{CKh}(D_1)$ and $\mathrm{CKh}(D_2)$ would have the same homology, therefore $\mathrm{Kh}(D)$ gives a link invariant.

We aim to explain the following slogan.

Slogan. $\mathrm{CKh}(D)$ is obtained by applying a certain $1+1$ *dimensional TQFT* $\mathcal{A}$ to the *cube of resolutions* of D.

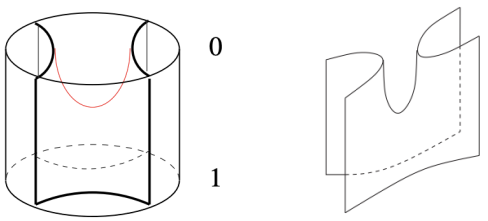
Each crossing in a diagram can be resolved in two ways, which we call the 0- and 1-resolutions. If D has n crossings, there are 2^n ways to resolve all of them,

which (after ordering the crossings) bijectively correspond to the vertices of the cube $[0,1]^n$. The figure illustrates this process for the Hopf link. If v is a

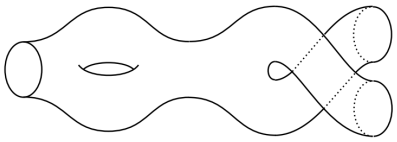


vertex of the cube, we write D_v the diagram of the corresponding resolution, which is always a collection of disjoint circles.

Along each edge e of the cube, one coordinate varies from 0 to 1 while all the other coordinates are fixed. We orient the edge from the vertex where the varied coordinate is 0 to the vertex with varied coordinate 1. To each edge $e\colon v_0 \to v_1$, we assign a *surface* S_e with boundary $\partial S_e = D_{v_0} \cup D_{v_1}$ as follows. The diagrams D_{v_0} and D_{v_1} are identical away from a neighborhood of a single crossing, and we define S_e to be the product $D_{v_0} \times [0,1]$ away from this neighborhood. Inside the neighborhood, S_e is given by the following saddle shape cobordism.



We now introduce the concept of *cobordism category*, which will be used to decorate the vertices and edges of the cube of resolutions, and will be the input of certain *topological quantum field theory (TQFT)*. Roughly speaking, the cobordism category encodes "closed spaces" (*compact manifolds*) and time-evolutions between them (called *cobordism*). These cobordisms can be thought of as a model for *space-time*. The TQFT assigns to such spaces certain algebraic data, which can be interpreted as some measurement of physical quantities.



Let us digress a bit to discuss the notions of *categories* and *functors* in general. A *category* $\mathcal{C}$ consists of

- a class $\mathrm{Ob}(\mathcal{C})$ of *objects*,
- a class $\mathrm{Hom}(\mathcal{C})$ of *morphisms*,
- a *domain* class function $\mathrm{dom}\colon \mathrm{Hom}(\mathcal{C}) \to \mathrm{Ob}(\mathcal{C})$,
- a *codomain* class function $\mathrm{codom}\colon \mathrm{Hom}(\mathcal{C}) \to \mathrm{Ob}(\mathcal{C})$,
- for every three objects a, b, c, there is a binary operation $\mathrm{Hom}(a,b) \times \mathrm{Hom}(b,c) \to \mathrm{Hom}(a,c)$, called the *composition* of morphisms,

such that:

- (associativity) $h \circ (g \circ f) = (h \circ g) \circ g$,
- (identity) for every object x, there exists a morphism $1_x \in \mathrm{Hom}(x,x)$, called the *identity* morphism, such that $1_x \circ f = f$ for any $f \in \mathrm{Hom}(-,x)$ and $g \circ 1_x = g$ for any $g \in \mathrm{Hom}(x,-)$.

Exercise. Find examples of categories.

Let $\mathcal{C}$ and $\mathcal{D}$ be two categories. A *functor* $F\colon \mathcal{C} \to \mathcal{D}$ is a mapping that:

- associate each object $x \in \mathrm{Ob}(\mathcal{C})$ to an object $F(x) \in \mathrm{Ob}(\mathcal{D})$,
- associate each morphism $f \in \mathrm{Hom}_{\mathcal{C}}(x,y)$ to a morphism $F(f) \in \mathrm{Hom}_{\mathcal{D}}(F(x), F(y))$,

such that:

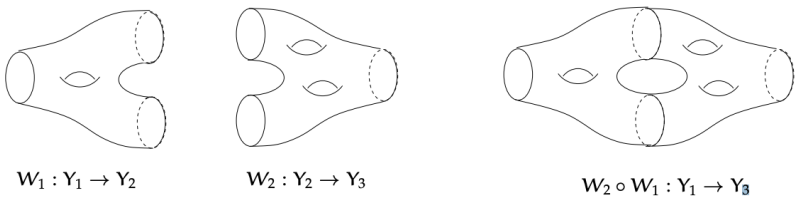
- $F(1_x) = 1_{F(x)}$ for every object $x \in \mathrm{Ob}(\mathcal{C})$,
- $F(g \circ f) = F(g) \circ F(f)$ for every morphisms $f\colon a \to b$ and $g\colon b \to c$ in $\mathcal{C}$.

Example 6.14. The *fundamental group* gives a functor from the category of *pointed topological spaces* (topological space with a base point) to the category of groups.

Remark 6.15. Many of the concepts can be extended to *higher categories*. For instance, a 2-*category* is a category with "morphisms between morphisms", i.e. each Hom-set itself carries the structure of a category. Similarly, one can extend it further and define the notion of ∞-*category*. This notion turned out

to be crucial in the study of *symplectic geometry*, where it is important to study certain A_∞-*category* (the *Fukaya category*) of a symplectic manifold.

Let Y_1 and Y_2 be compact oriented n-manifolds. (Here, we interpret "manifolds" topologically: so, Y is an n-manifold simply means that locally Y is homeomorphic to an open subset of $\mathbb{R}^n$.) We define a *cobordism* W from Y_1 to Y_2 to be a compact oriented $(n+1)$-manifold with $\partial W = -Y_1 \coprod Y_2$, and denote it by $W\colon Y_1 \to Y_2$. Two cobordisms W, W' are called *equivalent* if there is a homeomorphism $W \to W'$ whose restriction to ∂W is the identity. If $W_1\colon Y_1 \to Y_2$ and $W_2\colon Y_2 \to Y_3$ are cobordisms, their *composition* $W_2 \circ W_1 \coloneqq W_1 \cup_{Y_2} W_2$ is a cobordism from $Y_1 \to Y_3$. The $(n+1)$-*dimensional*



cobordism category Cobor_{n+1} is the category whose

- objects are compact oriented n-manifolds, and
- morphisms are equivalence classes of cobordisms between them.

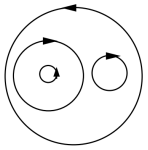
Exercise. Show that this indeed is a category. In particular, given any compact oriented n-manifold Y, what is the identity morphism $1_Y \in \mathrm{Hom}(Y, Y)$?

Note that the objects and morphisms of the cobordism category Cobor_{n+1} admit the structure of a *coproduct* $\coprod$, which is simply taking the disjoint union:

- For $Y_1, Y_2 \in \mathrm{Ob}(\mathrm{Cobor}_{n+1})$, the disjoint union $Y_1 \coprod Y_2$ is again an object of Cobor_{n+1}.
- For $W_1 \in \mathrm{Hom}(Y_1, Y_1')$ and $W_2 \in \mathrm{Hom}(Y_2, Y_2')$, the disjoint union $W_1 \coprod W_2$ (as an oriented $(n+1)$-manifold with boundary) is a morphism which lies in $\mathrm{Hom}(Y_1 \coprod Y_2, Y_1' \coprod Y_2')$.

We would like to view the vertices of the cube of resolutions as being decorated by objects of the $1+1$ dimensional cobordism category, and its edges as being decorated by morphisms. In order to do this, we need to orient the objects involved. It can be done explicitly, as the objects D_v are nothing but a collections of disjoint circles in $\mathbb{R}^2$.

Definition 6.16. Let D_v be a collection of disjoint circles in $\mathbb{R}^2$. The *canonical orientation* on D_v is defined by giving the i-th circle C_i $(-1)^{n_i}$-times the standard orientation, where n_i is the number of circles separating C_i from infinity in $\mathbb{R}^2$.



Exercise. Let $e\colon v_0 \to v_1$ be an edge in the cube of resolutions. If we give D_{v_0} and D_{v_1} the canonical orientations, then the surface S_e can be given an orientation so that it gives an oriented cobordism from D_{v_0} to D_{v_1}.

Exercise. Each 2-dimensional face of the cube resolutions (vertices denoted by $v_{00}, v_{01}, v_{10}, v_{11}$) corresponds to a square of morphisms in the cobordism category. The square of morphisms commutes: the composition $D_{00} \to D_{01} \to D_{11}$ coincides with the composition $D_{00} \to D_{10} \to D_{11}$. (This amounts to the fact that "1-handles" can be added in any order without changing the homeomorphism type of the resulting surface.)

Let us summarize what we have discussed so far in the following table.

Cube of resolutions	Cobor_{1+1}
vertex v	complete resolution D_v
edge $e\colon v_0 \to v_1$	cobordism $S_e\colon D_{v_0} \to D_{v_1}$
2-dimensional face	commuting square of morphisms

We now introduce the concept of *topological quantum field theory (TQFT)*, which has been widely used in recent advancements of various fields of mathematics: geometric topology, algebraic topology, symplectic geometry, complex geometry, mathematical physics, etc. just to name a few.

An $(n+1)$ dimensional TQFT is a *monoidal* functor

$$\mathcal{A}\colon \left(\mathrm{Cobor}_{n+1}, \coprod\right) \to (\mathrm{Vect}_{\mathbb{R}}, \otimes).$$

Here, *monoidal* means that it behaves well under disjoint unions:

- $\mathcal{A}(Y \coprod Y') = \mathcal{A}(Y) \otimes \mathcal{A}(Y')$,

- For $W_1 \in \mathrm{Hom}(Y_1, Y_1')$ and $W_2 \in \mathrm{Hom}(Y_2, Y_2')$, we have $\mathcal{A}(W_1 \coprod W_2) = \mathcal{A}(W_1) \otimes \mathcal{A}(W_2)$.

Exercise. Try to define the tensor product of two morphisms in $\mathrm{Vect}_{\mathbb{R}}$.

Remark 6.17. Let us mention some important examples of TQFTs.

- For a symplectic manifold (X, ω), one can associate to it a (2+1) dimensional TQFT, which encodes the information of *pseudo-holomorphic maps* from a Riemann surface Σ into X. This is based on a series of work by Floer, Gromov, Witten, among others.
- Certain $(3+1)$ dimensional TQFT is used to study the *Chern–SImons theory* of three-dimensional manifolds.
- Certain $(4+1)$ dimensional TQFT can be used to define the *Donaldson invariants* of four-dimensional manifolds, which is one of the most fundamental invariants in low dimensional topology.

This is far from a complete list; there are also other important variants of TQFT, which will lead to, for instance, Yang–Mills theory, cohomologicla field theories, mirror symmetry, stability conditions on triangulated categories, Gromov–Witten invariants, quantum cohomology, geometric Langlands, etc.

By applying a TQFT (which we will specified later) on the $(1+1)$ dimensional cobordism category, we obtain the following table.

Cube of resolutions	Cobor_{1+1}	$\mathrm{Vect}_{\mathbb{R}}$
vertex v	complete resolution D_v	vector space $\mathcal{A}(D_v)$
edge $e\colon v_0 \to v_1$	cobordism $S_e\colon D_{v_0} \to D_{v_1}$	linear map $\mathcal{A}(S_e)\colon \mathcal{A}(D_{v_0}) \to \mathcal{A}(D_{v_1})$
2-dimensional face	commuting square	commuting square

We can now define the *Khovanov complex*. As a vector space we define

$$\mathrm{CKh}(D) = \bigoplus_v \mathcal{A}(D_v)$$

where the sum runs over all vertices of the cube of resolutions $[0,1]^n$. For $x \in \mathcal{A}(D_v)$, the differential on $\mathrm{CKh}(D)$ is defined by

$$\mathrm{d}x = \sum_{e\colon v \to v'} (-1)^{\sigma(e)} \mathcal{A}(S_e)(x)$$

where σ is a map from the set of edges to $\{0,1\}$, which we will define to make $\mathrm{d}^2 = 0$. Indeed, we have:

Lemma 6.18. *If σ is chosen so that each two-dimensional face of the cube has an odd number of edges with $\sigma(e) = 1$, then $\mathrm{d}^2 = 0$.*

Proof. If v'' is a vertex of the cube obtained by changing two 0 coordinates of v to 1's, then the component of $\mathrm{d}^2(x)$ which lies in the summand in $\mathcal{A}(v'')$ is

$$(-1)^{\sigma(e_1)+\sigma(e_2)}\mathcal{A}(S_{e_2}) \circ \mathcal{A}(S_{e_1})(x) + (-1)^{\sigma(e_3)+\sigma(e_4)}\mathcal{A}(S_{e_3}) \circ \mathcal{A}(S_{e_4})(x)$$

where e_1, e_2, e_3, e_4 are the edges of the two-dimensional face containing v and v'', labeled clockwise starting from v. The lemma then follows from the commutativity for each two-dimensional face

$$\mathcal{A}(S_{e_2}) \circ \mathcal{A}(S_{e_1}) = \mathcal{A}(S_{e_3}) \circ \mathcal{A}(S_{e_4}).$$

□

Remark 6.19. One can show that such σ always exists, and that any two such choices of σ give rise to isomorphic chain complexes. This involves some general machinery concerning homotopy of chain complexes, which we omit here.

Up until this point, the construction we described works for any TQFT, but the resulting homology depends on the planar diagram D, rather than its underlying link L. To get a chain complex $(\mathrm{CKh}(D), \mathrm{d})$ whose *homology* is a *link invariant*, we will use a particular TQFT $\mathcal{A}$ for which

$$\mathcal{A}(S^1) = \langle \mathbf{1}, \mathbf{x} \rangle =: V$$

is a two-dimensional vector space with a basis denoted by $\mathbf{1}$ and $\mathbf{x}$.

Exercise. Show that in order for the homology of the Khovanov complex $(\mathrm{CKh}(D), \mathrm{d})$ to be a link invariant, the dimension of $V := \mathcal{A}(S^1)$ has to be 2. (Consider the unknot and its one-crossing diagram.)

Since $\mathcal{A}$ is a monoidal functor, we must have

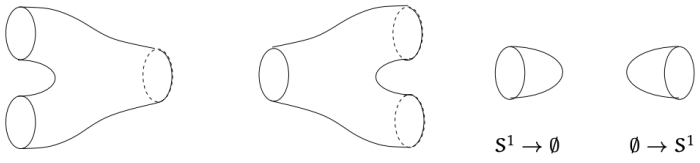
$$\mathcal{A}(\coprod^n S^1) = V^{\otimes n}.$$

This completely specifies the functor $\mathcal{A}$ at the level of objects.

If D is a closed 1-manifold (i.e. a disjoint union of circles), we define a *state* of D to be a labeling of each component of D by either $\mathbf{1}$ or $\mathbf{x}$. The vector space $\mathcal{A}(D)$ has a basis consisting of states of D. More generally, if D is a planar diagram, we define a *state* of D to be a choice of a complete resolution

of D, together with a state of the complete resolution. Then, as a vector space, $\mathrm{CKh}(D)$ has a basis consisting of states of D.

The functor $\mathcal{A}$ is monoidal, so to understand how it acts on morphisms, it is enough to describe it for the following "elementary" cobordisms: merge, split, death, and birth.



We define the corresponding four linear maps as follow.

$$\begin{aligned}
\text{(merge) } m\colon V \otimes V &\to V \\
\mathbf{1} \otimes \mathbf{1} &\mapsto \mathbf{1} \\
\mathbf{1} \otimes \mathbf{x},\ \mathbf{x} \otimes \mathbf{1} &\mapsto \mathbf{x} \\
\mathbf{x} \otimes \mathbf{x} &\mapsto 0
\end{aligned}$$

$$\begin{aligned}
\text{(split) } \Delta\colon V &\to V \otimes V \\
\mathbf{1} &\mapsto \mathbf{1} \otimes \mathbf{x} + \mathbf{x} \otimes \mathbf{1} \\
\mathbf{x} &\mapsto \mathbf{x} \otimes \mathbf{x}
\end{aligned}$$

$$\begin{aligned}
\text{(death) } \epsilon\colon V &\to \mathbb{R} \\
\mathbf{1} &\mapsto 0 \\
\mathbf{x} &\mapsto 1
\end{aligned}$$

$$\begin{aligned}
\text{(birth) } i\colon \mathbb{R} &\to V \\
1 &\mapsto \mathbf{1}
\end{aligned}$$

This completes the definition of the chain complex $\mathrm{CKh}(D)$.

Exercise. Let D be the zero-crossing diagram of the unknot, and let D' be an one-crossing diagram of the unknot. Compute the chain complex $\mathrm{CKh}(D)$ and $\mathrm{CKh}(D')$, and show that they have the same homology.

The complex $\mathrm{CKh}(D)$ can be equipped with a natural bigrading

$$\mathrm{CKh}(D) = \bigoplus_{i,j} \mathrm{CKh}^{i,j}(D)$$

which we now describe. The first grading is called the *homological* grading, which will be increase by 1 under d. If v is a vertex of $[0,1]^n$, we write $|v|$ for the sum of the coefficients of v. In particular, any element of $\mathcal{A}(D_v)$ has homological grading $|v|$.

To define the second grading, which we call the *q-grading*, we first define a grading $\widetilde{q}$ on V by setting

$$\widetilde{q}(\mathbf{1}) = 1 \qquad \text{and} \qquad \widetilde{q}(\mathbf{x}) = -1,$$

and extend it to $V^{\otimes n}$ by setting $\widetilde{q}(a \otimes b) = \widetilde{q}(a) + \widetilde{q}(b)$. Since each cobordism S_e is a union of a pair of pants with some cylinders, one can verify that

$$\widetilde{q}(\mathrm{d}x) = \widetilde{q}(x) - 1$$

for any $x \in \mathcal{A}(D_v)$. (It suffices to check that the merge and split maps both decrease the $\widetilde{q}$-grading by 1.) We now define the *q-grading* for $x \in \mathcal{A}(D_v)$ to be

$$q(x) = \widetilde{q}(x) + |v|.$$

Then we have $q(\mathrm{d}x) = q(x)$, i.e. the differential d preserves the q-grading. Thus, $\mathrm{CKh}(D)$ decomposes as a direct sum of chain complexes

$$(\mathrm{CKh}(D), \mathrm{d}) = \bigoplus_{j} (\mathrm{CKh}^{\star,j}(D), \mathrm{d}).$$

Finally, to pin down the exact normalization of the Jones polynomial $V(L)$ of a link, we need to fix an orientation on L, and shift the gradings (both homological and q-gradings) of the Khovanov complex. For $n, m \in \mathbb{Z}$, we define $t^m q^n \mathrm{CKh}(D)$ to be the bi-graded chain complex whose (i,j)-th vector space is $\mathrm{CKh}^{i-m,j-n}(D)$. Let o be an orientation of a diagram D, and let $n_\pm(D,o)$ be the number of positive/negative crossings in D. It turns out that the correct normalization is

$$\mathrm{CKh}(D,o) = t^{-n_-(D,o)} q^{n_+(D,o)-2n_-(D,o)} \mathrm{CKh}(D).$$

Theorem 6.20 (Khovanov)**.** *If (D,o) and (D',o') are related a Reidemeister move, then $\mathrm{CKh}(D,o)$ and $\mathrm{CKh}(D',o')$ have the same homology.*

Exercise. Compute the Khovanov complex for the unknot, and its one-crossing diagrams (with a positive or negative crossing). Justify the correction term "$t^{-n_-(D,o)}q^{n_+(D,o)-2n_-(D,o)}$" in this case.

Definition 6.21. Let L be an oriented link represented by an oriented diagram (D, o). Its *Khovanov homology* $\mathrm{Kh}(L)$ is defined to be the homology of the complex $(\mathrm{CKh}(D, o), \mathrm{d})$, which is a bi-graded vector space.

The *graded Euler characteristic* of $\mathrm{Kh}(L)$ is defined to be

$$\chi(\mathrm{Kh}(L)) = \sum_{i,j}(-1)^i q^j \dim \mathrm{Kh}^{i,j}(L).$$

Theorem 6.22 (Khovanov)**.** *For any oriented link as above, we have*

$$\chi(\mathrm{Kh}(L)) = V(L).$$

Exercise. Compute and verify this theorem in the cases of the unknot and the Hopf link.

Lecture 10

7. Calculus of variations

We discuss in this section some concrete problems that can be answered by the method of *calculus of variations*.

7.1. **Brachistochrone problem.** Let us formulate the Brachistochrone problem in a precise manner. Consider two points $A = (0, 0)$ and $B = (a, -b)$ in $\mathbb{R}^2$, where $a, b > 0$. We would like to consider paths connecting A and B, say (its mirror along the x-axis) parametrized by

$$y\colon [0, a] \to [0, b]; \qquad y(0) = 0 \text{ and } y(a) = b.$$

The first thing to do is to express the time of descent as a function of $y(x)$. Let $v = \frac{\mathrm{d}s}{\mathrm{d}t}$ be the speed of the object, where $\mathrm{d}s$ is the arc length along the graph of $y(x)$. From *conservation of energy*, at height $y(x)$ we have:

$$\frac{1}{2}mv(x)^2 = mgy(x)$$

where m is the mass of the descending object, and g is the gravitational constant. The time of descent is therefore given by

$$\int \mathrm{d}t = \int \frac{\mathrm{d}s}{v} = \frac{1}{\sqrt{2g}} \int \frac{\sqrt{\mathrm{d}x^2 + \mathrm{d}y^2}}{\sqrt{y}} = \frac{1}{\sqrt{2g}} \int_0^a \frac{\sqrt{1 + y'(x)^2}}{\sqrt{y(x)}}\, \mathrm{d}x.$$

The problem then is to find, among all functions $y(x)$ satisfying the boundary conditions ($y(0) = 0$ and $y(a) = b$), the one which minimizes

$$T[y] = \int_0^a \frac{\sqrt{1 + y'(x)^2}}{\sqrt{y(x)}} \, \mathrm{d}x.$$

The method of Euler and Lagrange applies to "variational problems" of the following kind. Given a function of three variables $F(x, y, p)$, find the function $y(x)$ (satisfying given boundary conditions) that minimize the integral

$$T[y] = \int_0^a F(x, y(x), y'(x)) \, \mathrm{d}x.$$

Clearly, brachistochrone is of this form.

Proposition 7.1 (Euler, Lagrange)**.** *The function $y(x)$ achieving the minimum of $T[y]$ (if exists) must satisfy a second order differential equation, called the* Euler–Lagrange equation

$$\frac{\partial F}{\partial y}\Big|_{(x,y(x),y'(x))} = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial F}{\partial p}\Big|_{x,y(x),y'(x)}\right).$$

Here the right-hand side is understood as follows: take partial derivatives of F with respect to the argument p, evaluate the resulting function of (x, y, p) at $(x, y(x), y'(x))$ to obtain a function of x only, then take its (one-variable) derivative with respect to x.

Proof. Suppose $y(x)$ minimizes the integral $T[y]$. Let $w(x)$ be an arbitrary function with $w(0) = w(a) = 0$, and let $\epsilon \in \mathbb{R}$ be a small real number. Then

$$y_\epsilon(x) = y(x) + \epsilon w(x)$$

can be considered as small perturbation of $y(x)$. Therefore

$$g(\epsilon) \coloneqq T[y_\epsilon] = \int_0^a F(x, y_\epsilon(x), y'_\epsilon(x)) \, \mathrm{d}x$$

has a minimum at $\epsilon = 0$. Thus $g'(0) = 0$.

$$\frac{\mathrm{d}g}{\mathrm{d}\epsilon} = \int_0^a \left(\frac{\partial F}{\partial y} \cdot w(x) + \frac{\partial F}{\partial p} \cdot w'(x)\right) \mathrm{d}x = \int_0^a w(x) \left(\frac{\partial F}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial F}{\partial p}\right)\right) \mathrm{d}x.$$

By evaluating at $\epsilon = 0$, one finds that

$$\frac{\partial F}{\partial y}|_{(x,y(x),y'(x))} - \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial F}{\partial p}\right)|_{(x,y(x),y'(x))} = 0$$

as desired. □

Before solving the brachistochrone problem, let us use the Euler–Lagrange method to solve a simpler problem: Suppose we want to connect $(0,0)$ and (a,b) by a curve of least possible *length* (we know the answer to this, right?); i.e. we want to find $y(x)$ satisfying the same boundary conditions that minimize the integral

$$L[y] = \int_0^a \sqrt{1 + y'(x)^2}\,\mathrm{d}x.$$

in this case, we consider $F(x,y,p) = \sqrt{1+p^2}$. Thus

$$\frac{\partial F}{\partial y} = 0 \qquad \text{and} \qquad \frac{\partial F}{\partial p} = \frac{p}{\sqrt{1+p^2}}.$$

The Euler–Lagrange equation reads

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{y'(x)}{\sqrt{1+y'(x)^2}} = 0, \qquad \text{or equivalently,} \qquad \frac{y''(x)}{(1+y'(x)^2)^{3/2}} = 0.$$

Therefore $y''(x) = 0$, i.e. $y(x)$ is linear function as expected.

Let us return to the brachistochrone problem. in this case, we consider $F(x,y,p) = \sqrt{\frac{1+p^2}{y}}$. Thus

$$\frac{\partial F}{\partial y} = -\frac{\sqrt{1+p^2}}{2y^{3/2}} \qquad \text{and} \qquad \frac{\partial F}{\partial p} = \frac{p}{\sqrt{y(1+p^2)}}.$$

After some calculations, one finds

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial F}{\partial p}|_{x,y(x),y'(x)}\right) = \frac{1}{\sqrt{y(1+(y')^2)}}\left(\frac{y''}{1+(y')^2} - \frac{(y')^2}{2y}\right).$$

The Euler–Lagrange equation, after some simplifications, read

$$2y(x)y''(x) + (y'(x))^2 + 1 = 0.$$

This becomes a *second order ordinary differential equation*. We would like to find a function $y(x)$ satisfying this differential equation, together with the

boundary conditions $y(0) = 0$ and $y(a) = b$. By multiplying both sides by $y'(x)$, we obtain

$$0 = 2yy'y'' + (y')^3 + y' = \frac{\mathrm{d}}{\mathrm{d}x}\left(y(x)y'(x)^2 + y(x)\right).$$

Thus $y(x)y'(x)^2 + y(x) \equiv C$ is a constant function. Hence

$$y'(x)^2 = \frac{C - y(x)}{y(x)}, \qquad \text{or equivalently,} \qquad \frac{\mathrm{d}y}{\mathrm{d}x} = \sqrt{\frac{C-y}{y}}.$$

Therefore

$$x = \int \mathrm{d}x = \int \sqrt{\frac{y}{C-y}}\,\mathrm{d}y + (\text{constant}).$$

One can compute the integral by a substitution like $y = C\sin^2 t$, and obtains

$$x = 2C\left(\frac{t}{2} - \frac{1}{4}\sin(2t)\right) + (\text{constant}).$$

Hence we have

$$\begin{cases} x(t) = Ct - \frac{C}{2}\sin(2t) + D \\ y(t) = C\sin^2(2t) \end{cases}$$

Plug in $t = 0$, one finds $D = 0$. Thus the brachistochrone path can be parametrized by

$$t \mapsto \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = C\begin{bmatrix} t - \frac{1}{2}\sin(2t) \\ \frac{1}{2} - \frac{1}{2}\cos(2t) \end{bmatrix} = \frac{C}{2}\begin{bmatrix} 2t - \sin(2t) \\ 1 - \cos(2t) \end{bmatrix}$$

where the constant C is chosen so that the curve passes through the endpoint (a, b).

Exercise. Show that the curve parametrized above is a *cycloid*: The cycloid is the path described by a fixed point on a circle, as the circle rolls on a fixed line.

Exercise. Here is one remarkable property of the cycloid: If we release an object from *any* point on the path, the time of descent to the lowest point will be the *same*, regardless of where on the path we release it.

7.2. **Isoperimetric problem.** Another (an even older) problem which can be solved by the method of calculus of variations, is the *Dido's problem*. The Roman poet Publius Vergilius Maro (70-19 B.C.) tells in his epic passage, *Aeneid*, the story of queen Dido, the daughter of the Phoenician king of the 9th century B.C.

"The Kingdom you see is Carthage, the Tyrians, the town of Agenor;
But the country around is Libya, no folk to meet in war.
Dido, who left the city of Tyre to escape her brother,
Rules here–a long and labyrinthine tale of wrong
Is hers, but I will touch on its salient points in order....Dido, in great disquiet, organised her friends for escape.
They met together, all those who harshly hated the tyrant
Or keenly feared him: they seized some ships which chanced to be ready...
They came to this spot, where to-day you can behold the mighty
Battlements and the rising citadel of New Carthage,
And purchased a site, which was named 'Bull's Hide' after the bargain
By which they should get as much land as they could enclose with a bull's hide."

After the assassination of her husband by her brother, Dido fled to a haven near Tunis. There she asked the local leader, Yarb, for as much land as could be enclosed by the hide of a bull. Since the deal seemed very modest, he agreed. Dido cut the hide into narrow strips, tied them together and encircled a large tract of land which became the city of Carthage.

Dido faced the following mathematical problem, which is also known as an *isoperimetric problem*:

Find among all curves of given length the one which encloses maximal area.

Dido found intuitively the right answer – the circle.

Let us formulate this problem mathematically. Consider a parametrization of the curve

$$t \in [a,b] \quad \mapsto \quad (x(t),y(t)) \in \mathbb{R}^2, \qquad x(a)=x(b),\ y(a)=y(b).$$

We would like to maximaize the area which, by Green's theorem, can be written as

$$A[t,x,y,x',y'] = \frac{1}{2}\int_a^b (x(t)y'(t) - y(t)x'(t))\,\mathrm{d}t;$$

under the constrain that the length expressed below is fixed

$$L[t,x,y,x',y'] = \int_a^b \sqrt{x'(t)^2+y'(t)^2}\,\mathrm{d}t.$$

By the method of *Lagrange multiplier*, we are led to finding the extremals of

$$\int_a^b H(t,x,y,x',y')\,\mathrm{d}t$$

where

$$H(t,x,y,x',y') = \frac{1}{2}\left(x(t)y'(t) - y(t)x'(t)\right) + \lambda\sqrt{x'(t)^2+y'(t)^2}.$$

By the Euler–Lagrange method (now we have two functions $x(t), y(t)$ instead of one), at the extremum we have

$$\frac{\partial H}{\partial x} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial H}{\partial x'}\right) \quad \text{and} \quad \frac{\partial H}{\partial y} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial H}{\partial y'}\right),$$

which are equivalent to

$$\frac{y'(t)}{2} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{-y(t)}{2} + \frac{\lambda x'(t)}{\sqrt{x'(t)^2+y'(t)^2}}\right) \text{ and } \frac{-x'(t)}{2} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{x(t)}{2} + \frac{\lambda y'(t)}{\sqrt{x'(t)^2+y'(t)^2}}\right).$$

Integrating on both sides, we get

$$\frac{\lambda x'(t)}{\sqrt{x'(t)^2+y'(t)^2}} = y(t)+A \quad \text{and} \quad \frac{\lambda y'(t)}{\sqrt{x'(t)^2+y'(t)^2}} = -x(t)-B$$

for some constants A and B. Now square both equations and add them to get

$$(x(t)+B)^2 + (y(t)+A)^2 = \lambda^2.$$

This is indeed a circle!

The isoperimetric problem in higher dimensions consists in finding among all domains of given *surface area* the one with maximal volume. The solution is the ball. Its proof is much more delicate than in the plane. One reason is that the convex hull has not necessarily a smaller perimeter. It was solved in the most elegant way by means of an inequality derived by Brunn (1887) and Minkowski (1896) for convex sets and then generalized to nonconvex sets by L. A. Lyusternik (1935).

Let A and B be subsets of $\mathbb{R}^n$. Their *sum* is defined to be

$$A + B = \{a + b \mid a \in A,\ b \in B\} \subseteq \mathbb{R}^n.$$

Theorem 7.2 (Brunn–Minkowski)**.** *If $A, B \subseteq \mathbb{R}^n$ are bounded open subsets, then*

$$\mu(A)^{\frac{1}{n}} + \mu(B)^{\frac{1}{n}} \leq \mu(A+B)^{\frac{1}{n}}.$$

Here μ denotes the standard Lebesgue measure on $\mathbb{R}^n$.

Proof. Let us prove it only for the case where A and B are both rectangular

$$A = I_1 \times \cdots \times I_n \qquad \text{and} \qquad B = J_1 \times \cdots \times J_n.$$

The general case can be (quite non-trivially) deduced from this case. Let $a_1, \ldots, a_n, b_1, \ldots, b_n$ be the lengths of the intervals $I_1, \ldots, I_n, J_1, \ldots, J_n$. Then

$$\begin{aligned}
\frac{\mu(A)^{\frac{1}{n}} + \mu(B)^{\frac{1}{n}}}{\mu(A+B)^{\frac{1}{n}}} &= \frac{(\prod_i a_i)^{1/n} + (\prod_i b_i)^{1/n}}{(\prod_i (a_i + b_i))^{1/n}} \\
&= \left(\prod_i \frac{a_i}{a_i + b_i}\right)^{1/n} + \left(\prod_i \frac{b_i}{a_i + b_i}\right)^{1/n} \\
&\leq \frac{1}{n}\left(\sum_i \frac{a_i}{a_i + b_i}\right) + \frac{1}{n}\left(\sum_i \frac{b_i}{a_i + b_i}\right) = 1.
\end{aligned}$$

□

Let $\Omega \subseteq \mathbb{R}^n$ be a region with boundary $\partial\Omega$. The *surface area* of $\partial\Omega$ defined to be

$$S(\partial\Omega) = \lim_{\epsilon \to 0} \frac{\mu(\Omega + \epsilon B) - \mu(\Omega)}{\epsilon}$$

where $B = B_0(1) \subseteq \mathbb{R}^n$ denotes the unit ball in $\mathbb{R}^n$. The isoperimetric problem concerns maximizing the volume $\mu(\Omega)$ while fixing the surface area $S(\partial\Omega)$;

equivalently, we would like to find Ω that maximize

$$\frac{\mu(\Omega)^{1/n}}{S(\partial\Omega)^{1/(n-1)}}.$$

Theorem 7.3 (Isoperimetric inequality)**.** *For any region $\Omega \subseteq \mathbb{R}^n$ we have*

$$\frac{\mu(\Omega)^{1/n}}{S(\partial\Omega)^{1/(n-1)}} \leq \frac{\mu(B)^{1/n}}{S(\partial B)^{1/(n-1)}}.$$

Proof. By Brunn–Minkowski inequality, we have

$$\frac{\mu(\Omega+\epsilon B)-\mu(\Omega)}{\epsilon} \geq \frac{\left(\mu(\Omega)^{1/n}+\epsilon\mu(B)^{1/n}\right)^n-\mu(\Omega)}{\epsilon} \geq n\mu(\Omega)^{\frac{n-1}{n}}\mu(B)^{\frac{1}{n}}.$$

Thus

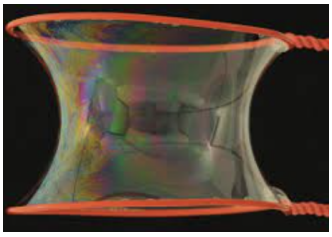
$$\frac{\mu(\Omega)^{\frac{n-1}{n}}}{S(\partial\Omega)} \leq \frac{1}{n\mu(B)^{\frac{1}{n}}}.$$

The isoperimetric inequality then follows from the fact that

$$n\mu(B) = S(\partial B).$$

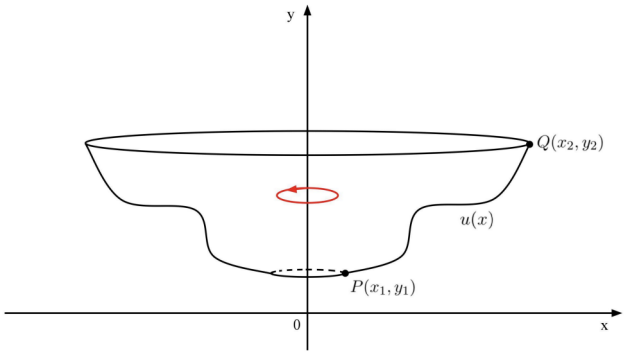
□

7.3. **Minimal surface of revolution.** Stretching a soap film between two parallel circular wires: the soap film naturally takes on the shape with least surface area, such surface is called a *minimal surface*.



Depending on the positions of the two wires, the surface can be connected (when they are near) or disconnected (when they are far from each other).

Given two points $P = (x_1, y_1)$ and $Q = (x_2, y_2)$ where $x_1, x_2, y_1, y_2 > 0$. Consider a curve $u(x)$ connecting P and Q, and the *surface of revolution* generated by rotating the curve with respect to the y-axis.



The problem here is to find such $u(x)$ that minimize the area of the surface of revolution

$$A[x, u, u'] = \int_{x_1}^{x_2} 2\pi x \sqrt{1 + u'(x)^2}\, \mathrm{d}x$$

with the constrains $u(x_1) = y_1$ and $u(x_2) = y_2$. The Euler–Lagrange equation reads:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{xu'(x)}{\sqrt{1 + u'(x)^2}}\right) = 0.$$

Integrate on both sides, one gets

$$u'(x) = \frac{C}{\sqrt{x^2 - C^2}} \qquad \text{for some constant } C,$$

therefore

$$u(x) = C \cdot \cosh^{-1}\left(\frac{x}{C}\right) + D \qquad \text{for constants } C, D.$$

If there exists C and D so that the curve $(x, u(x))$ passes through the two given points P, Q, then we get a continuous surface of revolution, which is generated by rotation of the *catenary* curve.

When such C and D do not exist (for instance, when the two circular wires are too far apart), then there is no continuous minimal surface of revolution; in this case, one simply gets two disjoint disks bounded by the two circles.

8. Analytic number theory

8.1. **Prime number theorem.** Let us start with Euler's viewpoint on the fact, originally due to Euclid, that there are infinitely many prime numbers.

Euclid's original proof was quite simple, and entirely *algebraic*: assume there are only finitely many primes, multiply them together, add 1, then gets a contradiction.

Euler realized instead that a basic fact from *analysis* also leads to the infinitude of primes. This fact is the divergence of the harmonic series

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots$$

which follows for instance from the fact that

$$\sum_{n=1}^{N} \frac{1}{n} \geq \sum_{n=1}^{N} \frac{1}{2^{\lceil \log_2 n \rceil}} \geq \frac{1}{2} \lfloor \log_2 N \rfloor$$

and the right hand side tends to ∞ as $N \to \infty$. On the other hand, if there were only finitely many primes, then the unique factorization of integers would imply that

$$\sum_{n=1}^{\infty} \frac{1}{n} = \prod_{p} \left(1 + \frac{1}{p} + \frac{1}{p^2} + \cdots\right)^{-1} = \prod_{p} \left(1 - \frac{1}{p}\right)^{-1} < \infty.$$

Contradiction.

Remark 8.1. One would often want a better estimation of the harmonic series. It can be proved that there exists a constant $\gamma > 0$ such that

$$\sum_{i=1}^{n} \frac{1}{i} - \log n = \gamma + \mathrm{O}(n^{-1}).$$

The constant γ is called the *Euler's constant*, and is one of the most basic constants in analytic number theory. However, since it is defined purely analytically, we remain astonishingly ignorant about it; for instance, γ is most likely irrational (even transcendental) but no proof is known.

The famous English mathematician G. H. Hardy is alleged to have offered to give up his Savilian Chair position at Oxford to anyone who proved γ to be irrational. Hilbert mentioned the irrationality of γ as an unsolved problem that seems "unapproachable" and in front of which mathematicians stand helpless. Conway and Guy are "prepared to bet that it is transcendental", although they do not expect a proof to be achieved within their lifetimes.

Euler's idea turns out to be quite fruitful: the introduction of analysis into the study of prime numbers allows us to prove distribution statements about

primes in a much more flexible fashion than is allowed by algebraic techniques. For instance, we will see later how Dirichlet adapted this idea to prove that every arithmetic progression whose terms do not all share a common factor contains infinitely many primes.

Let us now introduce Riemann's idea which fits Euler's idea into the theory of complex functions of one variable. Riemann considered the series

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

for all $s \in \mathbb{C}$ with $\mathrm{Re}(s) > 1$. Note that for $\mathrm{Re}(s) > 1$, the series is absolutely convergent; moreover, it converges uniformly in any region of the form $\mathrm{Re}(s) \geq 1 + \epsilon$ for $\epsilon > 0$. Consequently, it gives a holomorphic function in the half-plane $\mathrm{Re}(s) > 1$. The boundary $\mathrm{Re}(s) = 1$ is sometimes called the *critical line*.

We now show that the function ζ can be extended beyond the domain of absolute convergent of the original series.

Theorem 8.2. *The function $f(s) = \zeta(s) - \frac{s}{s-1}$ on the domain $\mathrm{Re}(s) > 1$ extends (uniquely) to a holomorphic function on the domain $\mathrm{Re}(s) > 0$. Consequently, $\zeta(s)$ can be extended to a meromorphic function on $\mathrm{Re}(s) > 0$ with a simple pole at $s = 1$ of residue 1 and no other poles.*

Proof. This is an application of one of the basic tools in analytic number theory, *Abel's method of resummation*. Namely,

$$\sum_{n=1}^{N} a_n b_n = a_{N+1} B_N - \sum_{n=1}^{N} (a_{n+1} - a_n) B_n, \qquad \text{where } B_n = \sum_{i=1}^{n} b_i.$$

Apply this to $\sum_{n=1}^{N} \frac{1}{n^s}$ by taking $a_n = n^{-s}$ and $b_n = 1$, so $B_n = n$. Note that the error term $a_{N+1} B_N = (N+1)^{-s} N \to 0$ for $\mathrm{Re}(s) > 1$. Thus we have

$$\begin{aligned} \zeta(s) &= \sum_{n=1}^{\infty} n \left(\frac{1}{n^s} - \frac{1}{(n+1)^{-s}} \right) \\ &= s \sum_{n=1}^{\infty} n \int_{n}^{n+1} x^{-s-1} \, \mathrm{d}x \\ &= s \int_{1}^{\infty} \lfloor x \rfloor x^{-s-1} \, \mathrm{d}x. \end{aligned}$$

Therefore we can write

$$f(s) = -s \int_1^\infty \{x\} x^{-s-1} \, \mathrm{d}x$$

where $\{x\}$ denotes the fractional part of x. One can show that the integral converges absolutely for $\mathrm{Re}(s) > 0$, and uniformly for $\mathrm{Re}(s) \geq \epsilon$ for any $\epsilon > 0$. This proves the theorem. □

Theorem 8.3 (Hadamard, de la Vallée–Poussin)**.** *The function $\zeta(s)$ has no zeros on the line* $\mathrm{Re}(s) = 1$.

Let us defer the proof of this theorem at the moment. It turns out that many interesting properties of the prime numbers are encoded in the Riemann zeta function $\zeta(s)$ (and its meromorphic extension, which we still denoted by $\zeta(s)$). For instance, using the above theorem, Hadamard and de la Vallée–Poussin independently established the *prime number theorem* in 1897.

Lecture 11

For $x \in \mathbb{R}_{>0,}$, write

$$\pi(x) = \sum_{p \leq x} 1 \qquad \text{and} \qquad \vartheta(x) = \sum_{p \leq x} \log p.$$

Remark 8.4. An elementary property of $\vartheta(x)$ is that $\vartheta(x) = \mathrm{O}(x)$, i.e. there exists $C > 0$ such that $\vartheta(x) \leq Cx$ for all $x > 0$. Here is a proof: For any n we have

$$2^{2n} > \binom{2n}{n} > \prod_{n < p \leq 2n} p, \qquad \text{thus} \qquad 2n \log 2 > \vartheta(2n) - \vartheta(n).$$

Therefore

$$\vartheta(2^k) < (2^k + 2^{k-1} + \cdots + 1) \log 2 < 2^{k+1} \log 2.$$

One can then easily deduce $\vartheta(x) = \mathrm{O}(x)$ using the fact that $\vartheta(x)$ is non-decreasing.

Theorem 8.5 (Prime number theorem)**.**

$$\pi(x) \sim \frac{x}{\log x}, \qquad \textit{i.e.} \qquad \lim_{x \to \infty} \frac{\pi(x)}{x/\log x} = 1.$$

Remark 8.6. One can show that this is equivalent to

$$\vartheta(x) \sim x,$$

because we have

$$\vartheta(x) \leq \sum_{p \leq x} \log x = \pi(x) \log(x),$$

and for any $\epsilon > 0$ we have

$$\vartheta(x) \geq \sum_{x^{1-\epsilon} < p \leq x} \log x^{1-\epsilon} = \big(\pi(x) - \pi(x^{1-\epsilon})\big) \log x^{1-\epsilon} = (1-\epsilon) \log x \cdot \big(\pi(x) + \mathrm{O}(x^{1-\epsilon})\big) .$$

What we will prove is that the improper integral

$$\int_1^\infty \frac{\vartheta(x) - x}{x^2} \, \mathrm{d}x$$

converges, i.e. for every $\epsilon > 0$, there exists $N > 0$ such that for $y, z \geq N$ we have

$$\left| \int_y^z \frac{\vartheta(x) - x}{x^2} \, \mathrm{d}x \right| < \epsilon.$$

Assuming this, to show that $\vartheta(x) \sim x$, suppose that there exists $\lambda > 1$ such that $\vartheta(x) \geq \lambda x$ for arbitrarily large x. Since ϑ is non-decreasing, for any such x we have

$$\int_x^{\lambda x} \frac{\vartheta(t) - t}{t^2} \, \mathrm{d}t \geq \int_x^{\lambda x} \frac{\lambda x - t}{t^2} \, \mathrm{d}t = \int_1^\lambda \frac{\lambda - t}{t^2} \, \mathrm{d}t > 0.$$

Contradiction. Similarly, if there exists $\lambda < 1$ such that $\vartheta(x) \leq \lambda x$ for arbitrarily large x, then such x satisfies

$$\int_{\lambda x}^x \frac{\vartheta(t) - t}{t^2} \, \mathrm{d}t \leq \int_{\lambda x}^x \frac{\lambda x - t}{t^2} \, \mathrm{d}t = \int_\lambda^1 \frac{\lambda - t}{t^2} \, \mathrm{d}t < 0.$$

Contradiction. Consequently, we have reduced the prime number theorem to the convergence of the improper integral above. Let $x = e^t$. The improper integral can also be written as

$$\int_0^\infty \big(\vartheta(e^t) e^{-t} - 1\big) \, \mathrm{d}t.$$

Let us introduction the function

$$g(z) = \int_0^\infty \big(\vartheta(e^t) e^{-t} - 1\big) \, e^{-zt} \, \mathrm{d}t.$$

By the elementary fact that $\vartheta(x) = \mathrm{O}(x)$, we have $g(z)$ converges on $\mathrm{Re}(z) > 0$. Our goal is to show that the integral converges at $z = 0$.

First, we will show that $g(z)$ admits a holomorphic extension to an open neighborhood of $\mathrm{Re}(z) \geq 0$. Consider the function

$$\Phi(s) = \sum_p \frac{\log p}{p^s}$$

which converges absolutely in $\mathrm{Re}(s) > 1$. By Abel's resummation, we have

$$\Phi(s) = \sum_p \frac{\log p}{p^s} = s \int_1^\infty \vartheta(x) x^{-s-1} \, \mathrm{d}x = s \int_0^\infty e^{-st} \vartheta(e^t) \, \mathrm{d}t.$$

Then one can observe that

$$g(z) = \frac{\Phi(z+1)}{z+1} - \frac{1}{z} \qquad \text{for} \qquad \mathrm{Re}(z) > 0.$$

We will use the theorem that $\zeta(s)$ has no zeros on $\mathrm{Re}(s) = 1$ to show that the function $\frac{\Phi(s)}{s} - \frac{1}{s-1}$ can be extended holomorphically to an open neighborhood of $\mathrm{Re}(s) \geq 1$, therefore implying $g(z)$ can be extended holomorphically to an open neighborhood of $\mathrm{Re}(z) \geq 0$.

One crucial ingredient, as discovered by Euler's proof of the infinitude of prime numbers, is that the zeta function can be expressed as certain product over primes. In the domain of absolute convergence $\mathrm{Re}(s) > 1$ of $\zeta(s)$, we can write

$$\zeta(s) = \prod_p \left(1 - \frac{1}{p^s}\right)^{-1}$$

and this product converges absolutely, and uniformly for $\mathrm{Re}(s) \geq 1 + \epsilon$ where $\epsilon > 0$. (This follows from a complex analysis fact that a product $\prod_i (1 + a_i)$ converges absolutely if and only if $\sum_i a_i$ converges absolutely.) Also, note that from this expression we know that $\zeta(s) \neq 0$ for $\mathrm{Re}(s) > 1$.

By the theorem that $\zeta(s) \neq 0$ for $\mathrm{Re}(s) \geq 1$ (together with the fact that $\zeta(s)$ admits a meromorphic extension with a simple pole at $s = 1$ of residue 1), we find that $-\frac{\zeta'(s)}{\zeta(s)}$ also admits a meromorphic extension on $\mathrm{Re}(s) > 0$, which is holomorphic on a neighborhood of $\mathrm{Re}(s) \geq 1$ except a simple pole at $s = 1$ of residue 1.

From the product expression of $\zeta(s)$, we have

$$\log \zeta(s) = \sum_p -\log(1-p^{-s}) = \sum_p \sum_{n=1}^{\infty} \frac{p^{-ns}}{n}.$$

By differentiating this expression we obtain

$$-\frac{\zeta'(s)}{\zeta(s)} = \sum_p \sum_{n=1}^{\infty} (\log p) p^{-ns} = \Phi(s) + \sum_p \frac{\log p}{p^s(p^s-1)}.$$

The last term converges absolutely for $\mathrm{Re}(s) > \frac{1}{2}$. Therefore, by our previous discussion, we find that $\Phi(s)$ can be extended to a meromorphic function to an open neighborhood of $\mathrm{Re}(s) \geq 1$, which is holomorphic except a simple pole at $s = 1$ of residue 1. Thus, the function $\frac{\Phi(s)}{s}$ also admits a meromorphic extension to an open neighborhood of $\mathrm{Re}(s) \geq 1$, which is holomorphic except a simple pole at $s = 1$ of residue 1. This proves that the function $\frac{\Phi(s)}{s} - \frac{1}{s-1}$ can be extended holomorphically to an open neighborhood of $\mathrm{Re}(s) \geq 1$.

To summarize, we are in the following situation:

- Let $f(t) = \vartheta(e^t)e^{-t} - 1$ $(t \geq 0)$, which is a bounded and locally integrable function.
- The function $g(z) = \int_0^\infty f(t)e^{-zt}\,\mathrm{d}t$ converges on $\mathrm{Re}(z) > 0$, and can be extended holomorphically to a neighborhood of $\mathrm{Re}(z) \geq 0$.
- We would like to show that $\int_0^\infty f(t)\,\mathrm{d}t$ exists.

Right now, we know the integral defining $g(z)$ makes sense for $\mathrm{Re}(z) > 0$. We will deduce the convergence of the improper integral $\int_1^\infty \frac{\vartheta(x)-x}{x^2}\,\mathrm{d}x$ (after substituting $x = e^t$) and hence the prime number theorem if we can obtain convergence of $g(z)$ for $z = 0$.

The idea is to do this by leveraging complex function theoretic information about g; this sort of operation is known as a *Tauberian argument*. To be precise, we will prove the following complex analytic theorem.

Theorem 8.7 (Newman)**.** *Let $f\colon [0,\infty) \to \mathbb{R}$ be a bounded, locally integrable function, and define $g(z) = \int_0^\infty f(t)e^{-zt}\,\mathrm{d}t$. This integral converges absolutely for $\mathrm{Re}(z) \geq \epsilon$ for any $\epsilon > 0$. Suppose that $g(z)$ extends to a holomorphic*

function on a neighborhood of $\mathrm{Re}(z) \geq 0$. *Then* $\int_0^\infty f(t)\,\mathrm{d}t$ *exists and equals to* $g(0)$.

Proof. (Newman, Zagier) For $T > 0$, let $g_T(z) = \int_0^T f(t)e^{-zt}\,\mathrm{d}t$. Each function g_T is entire, and we would like to show that $\lim_{T\to\infty} g_T(0) = g(0)$.

Fix $R > 0$ large. Let C be the boundary of the region

$$\{z \in \mathbb{C}\colon |z| \leq R,\ \mathrm{Re}(z) \geq -\delta\}$$

for some $\delta > 0$ chosen so that C lies inside the domain on which g is holomorphic. By the Cauchy integral theorem,

$$g(0) - g_T(0) = \frac{1}{2\pi i}\int_C (g(z) - g_T(z))\, e^{zT}\left(1 + \frac{z^2}{R^2}\right)\frac{\mathrm{d}z}{z}.$$

To estimate the right hand side, let us separate the contour C into

$$C_+ = C \cap \{z \in \mathbb{C}\colon \mathrm{Re}(z) \geq 0\}$$
$$C_- = C \cap \{z \in \mathbb{C}\colon \mathrm{Re}(z) \leq 0\}$$

Remember that we assumed f is bounded; choose $B > 0$ so that $|f(t)| \leq B$ for all $t \geq 0$. For $\mathrm{Re}(z) > 0$ with $|z| = R$, we have

$$|g(z) - g_T(z)| = \left|\int_T^\infty f(t)e^{-zt}\,\mathrm{d}t\right| \leq B\int_T^\infty |e^{-zt}|\,\mathrm{d}t = \frac{Be^{-\mathrm{Re}(z)T}}{\mathrm{Re}(z)}$$

and

$$\left|e^{zT}\left(1 + \frac{z^2}{R^2}\right)\frac{1}{z}\right| = e^{\mathrm{Re}(z)T}\frac{2\mathrm{Re}(z)}{R^2}.$$

Since the length of the contour is at most $2\pi R$, the contribution over C_+ to the integral is bounded in absolute value by

$$\frac{1}{2\pi}(2\pi R)\frac{Be^{-\mathrm{Re}(z)T}}{\mathrm{Re}(z)}e^{\mathrm{Re}(z)T}\frac{2\mathrm{Re}(z)}{R^2} = \frac{2B}{R}.$$

Over C_-, we separate the integral into integrals involving g and g_T. Since g_T is entire, its integral over C_- can instead be calculated over the semicircle $C'_- = \{z \in \mathbb{C}\colon |z| = R,\ \mathrm{Re}(z) \leq 0\}$. Similar computation shows that this integral is bounded by $2B/R$.

Finally, we consider the contribution from g over C_- to the integral. We are going to show that this contribution tends to 0 as $T \to \infty$. The reason

is that the integrand is the product of the function $g(z)(1+\frac{z^2}{R^2})/z$, which is independent of T, and the function e^{zt}, which goes to 0 rapidly and uniformly on compact sets in the half plane $\mathrm{Re}(z) < 0$. □

Proof of $\zeta(s)$ has no zeros on $\mathrm{Re}(s) = 1$. Recall that

$$-\frac{\zeta'(s)}{\zeta(s)} = \Phi(s) + \sum_p \frac{\log p}{p^s(p^s-1)}.$$

Thus, on $\mathrm{Re}(s) = 1$, the function $\Phi(s)$ has poles only at $s = 1$ and at the zeros of $\zeta(s)$. Let $\alpha \in \mathbb{R}\backslash\{0\}$. Denote by μ the order of zero of $\zeta(s)$ at $s = 1 + i\alpha$, and by ν the order of zero of $\zeta(s)$ at $s = 1 + 2i\alpha$. We have $\mu, \nu \geq 0$, and would like to show that $\mu = 0$. Observe that

$$\lim_{\epsilon\to 0} \epsilon\Phi(1+\epsilon) = 1, \quad \lim_{\epsilon\to 0} \epsilon\Phi(1+\epsilon \pm i\alpha) = -\mu, \quad \lim_{\epsilon\to 0} \epsilon\Phi(1+\epsilon \pm 2i\alpha) = -\nu.$$

The inequality

$$\sum_{r=-2}^{2} \binom{4}{2+r} \Phi(1+\epsilon+ir\alpha) = \sum_p \frac{\log p}{p^{1+\epsilon}} \left(p^{i\alpha/2} + p^{-i\alpha/2}\right)^4 \geq 0$$

then implies $6 - 8\mu - 2\nu \geq 0$, so $\mu = 0$. □

Then we discuss Section 9 first

8.2. **Dirichlet series.** The Riemann zeta function $\zeta(s)$ is a special example of a type of series we will be considered. A *Dirichlet series* is a formal series of the form

$$\sum_{n=1}^{\infty} \frac{a_n}{n^s} \quad \text{where} \quad a_n \in \mathbb{C}.$$

One can consider it as a number-theoretic analogue of formal power series.

Exercise. There exists an extended real number $L \in \mathbb{R} \cup \{\pm\infty\}$ with the following property: the Dirichlet series $\sum_{n=1}^{\infty} \frac{a_n}{n^s}$ converges absolutely for $\mathrm{Re}(s) > L$, but not for $\mathrm{Re}(s) < L$. Moreover, for any $\epsilon > 0$, the convergence is uniform on $\mathrm{Re}(s) \geq L + \epsilon$, so the series represents a holomorphic function on $\mathrm{Re}(s) > L$.

The quantity L is called the *abscissa of absolute convergence* of the Dirichlet series; it is the analogue of the radius of convergence of a power series.

Among Dirichlet series, the Riemann zeta function had the unusual property that one could factor it as a product over primes

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_p \left(1 - \frac{1}{p^s}\right)^{-1}.$$

This property was crucial in our proof of the prime number theorem. In fact, a number of natural Dirichlet series also admit such factorizations. They are the ones corresponding to *completely multiplicative functions*.

We define an *arithmetic function* to simply be a function $f\colon \mathbb{N} \to \mathbb{C}$. Besides the obvious operations of addition and multiplications between such functions, another useful operation on arithmetic functions is the *(Dirichlet) convolution* $f \star g$, defined by

$$(f \star g)(n) = \sum_{d|n} f(d) g\left(\frac{n}{d}\right).$$

Just as one can think of formal power series as the generating functions for ordinary sequences, one may think of a formal Dirichlet series as the "arithmetic generating function" for the arithmetic function $n \mapsto a_n$. Under this correspondence, the convolution of arithmetic functions corresponds to ordinary multiplication of the Dirichlet series:

$$\sum_{n=1}^{\infty} \frac{(f \star g)(n)}{n^s} = \left(\sum_{n=1}^{\infty} \frac{f(n)}{n^s}\right)\left(\sum_{n=1}^{\infty} \frac{g(n)}{n^s}\right).$$

In particular, convolution is a commutative and associative operation.

Exercise. Show that the set of arithmetic functions $f\colon \mathbb{N} \to \mathbb{C}$ with $f(1) = 1$ forms a group, with the operation given by convolution. Moreover, the arithmetic functions taking only integral values (with the value 1 at $n = 1$) form a subgroup.

We say f is a *multiplicative function* if $f(1) = 1$ and $f(mn) = f(m)f(n)$ whenever $\gcd(m, n) = 1$. It is equivalent to its Dirichlet series factors as a product

$$\sum_{n=1}^{\infty} \frac{f(n)}{n^s} = \prod_p \left(\sum_{i=0}^{\infty} \frac{f(p^i)}{p^{is}}\right).$$

In particular, the property of being multiplicative is stable under convolution, and under taking the convolution inverse.

Exercise. Here are some examples of multiplicative functions, some of which you may already be familiar with. It is a fun exercise to verify their properties stated below.

- The unit function ϵ: $\epsilon(1) = 1$ and $\epsilon(n) = 0$ for $n > 1$. This is the identity under $\star$.
- The constant function 1: $1(n) = 1$ for all n.
- The *Möbius function* μ: if n is square-free with d distinct prime factors, then $\mu(n) = (-1)^d$; otherwise $\mu(n) = 0$. This is the inverse of 1 under $\star$.
- The identity function id: $\mathrm{id}(n) = n$
- The k-th power function id^k: $\mathrm{id}^k(n) = n^k$.
- The *Euler (totient) function* φ: $\varphi(n)$ counts the number of integers in $\{1, 2, \ldots, n\}$ coprime to n. Note that $1 \star \varphi = \mathrm{id}$, so $\mathrm{id} \star \mu = \varphi$.
- The divisor function d (or τ): $d(n)$ counts the number of integers in $\{1, 2, \ldots, n\}$ dividing n. Note that $1 \star 1 = d$.
- The divisor sum function σ: $\sigma(n)$ is the sum of the divisors of n. Note that $1 \star \mathrm{id} = d \star \varphi = \sigma$.
- The divisor power sum functions σ_k: $\sigma_k(n) = \sum_{d|n} d^k$. Note that $\sigma_0 = d$ and $\sigma_1 = \sigma$. Also note that $1 \star \mathrm{id}^k = \sigma_k$.

8.3. **Dirichlet characters.** The key ingredients in the proof of Dirichlet's theorem on primes in arithmetic progressions are the *Dirichlet characters* and their associated Dirichlet series.

For N a positive integer, a *Dirichlet character of level* N is an arithmetic function χ which factors through a homomorphism $(\mathbb{Z}/N\mathbb{Z})^* \to \mathbb{C}$ on integers $n \in \mathbb{N}$ coprime to N, and is zero on integers not coprime to N. Such a function is *completely multiplicative*, i.e. $\chi(mn) = \chi(m)\chi(n)$ holds for all $m, n \in \mathbb{N}$. Note that the nonzero values of such a function must all be roots of unity, and that the characters of level N form a group under termwise multiplication, called the *dual group* of $(\mathbb{Z}/N\mathbb{Z})^*$ and is denoted as $\widehat{(\mathbb{Z}/N\mathbb{Z})^*}$.

Remark 8.8. The notions of characters and dual group can be defined for a general *finite abelian group* G. It is a general fact that G is isomorphic to the dual of its dual group $\widehat{G}$: every $x \in G$ defines a function $\chi \mapsto \chi(x)$ which is a

character of $\widehat{G}$, and one can show that this defines an isomorphism between G and $\widehat{\widehat{G}}$.

Definition 8.9. For each level N, there is a Dirichlet character taking the value 1 at all n coprime to N; it is called the *principal (or trivial) character of level N*.

We state two basic *orthogonality relations*.

Lemma 8.10. *Let χ be a Dirichlet character of level N. Then*

$$\sum_{x \in (\mathbb{Z}/N\mathbb{Z})^*} \chi(x) = \begin{cases} \varphi(N) & \text{if } \chi \text{ is principal} \\ 0 & \text{if } \chi \text{ is non-principal} \end{cases}$$

Proof. The sum is invariant under multiplication by $\chi(m)$ for any $m \in \mathbb{N}$ coprime to N. Suppose χ is non-principal, then we can choose m with $\chi(m) \neq 1$ and conclude that the sum has to be zero. □

Lemma 8.11. *Let $x \in (\mathbb{Z}/N\mathbb{Z})^*$. Then*

$$\sum_{\chi} \chi(x) = \begin{cases} \varphi(N) & \text{if } x = 1 \\ 0 & \text{if } x \neq 1 \end{cases}$$

where we sum over all Dirichlet characters χ of $(\mathbb{Z}/N\mathbb{Z})^*$.

Proof. This is the dual statement of the previous lemma, and therefore follows from the fact that $G \cong \widehat{G} \cong \widehat{\widehat{G}}$. □

The Dirichlet series associated to a Dirichlet character χ of level N is called a *Dirichlet L-series* (or *Dirichlet L-function*) *of level N*, denoted by $L(s, \chi)$ (or sometimes $L_\chi(s)$). Since χ is completely multiplicative, we have

$$L(s, \chi) = \prod_p \left(1 - \frac{\chi(p)}{p^s}\right)^{-1}$$

for $\mathrm{Re}(s) > 1$. Also, $L(s, \chi) \neq 0$ for $\mathrm{Re}(s) > 1$.

Theorem 8.12. *Let χ be a Dirichlet character of level N. Then $L(s, \chi)$ extends to a meromorphic function on $\mathrm{Re}(s) > 0$ with no poles away from $s = 1$. If χ is principal, then $L(s, \chi)$ has a simple pole at $s = 1$ of residue $\prod_{p|N}(1 - \frac{1}{p})$; otherwise, $L(s, \chi)$ is holomorphic also at $s = 1$.*

Proof. If χ is principal, then

$$L(s,\chi) = \zeta(s) \prod_{p|N} \left(1 - \frac{1}{p^s}\right)$$

and the claims follow from what we already know about $\zeta(s)$. Now, suppose χ is non-principal. By Abel's resummation, we have

$$L(s,\chi) = \sum_{n=1}^{\infty} (\chi(1) + \cdots + \chi(n)) \left(\frac{1}{n^s} - \frac{1}{(n+1)^s}\right).$$

Since $\chi(1) + \cdots + \chi(N) = 0$, the quantity $\chi(1) + \cdots + \chi(n)$ is bounded for all n. On the other hand,

$$\frac{1}{n^s} - \frac{1}{(n+1)^s} = \frac{s}{n^{s+1}} + \mathrm{O}\left(\frac{1}{n^{s+2}}\right).$$

Thus, $L(s,\chi)$ converges uniformly for $\mathrm{Re}(s) \geq \epsilon$ for any $\epsilon > 0$. This proves the claim. □

As in the proof of the prime number theorem, we will need to understand the zeros of $L(s,\chi)$ on the line $\mathrm{Re}(s) = 1$. Let us define

$$f_N(s) \coloneqq \prod_{\chi} L(s,\chi)$$

the product of all of the Dirichlet L-series of level N.

Notation. For a prime p not dividing N, denote $f(p)$ the order of p in the group $(\mathbb{Z}/N\mathbb{Z})^*$, i.e. the smallest integer $f \geq 1$ such that $p^f \equiv 1 \pmod{N}$. Also, denote $g(p) \coloneqq \varphi(N)/f(p)$, which is the order of the quotient of $(\mathbb{Z}/N\mathbb{Z})^*$ by the subgroup generated by p.

Lemma 8.13. *If $p \nmid N$, then*

$$\prod_{\chi} (1 - \chi(p)T) = (1 - T^{f(p)})^{g(p)}.$$

Proof. Let W be the set of $f(p)$-th roots of unity. Then

$$\prod_{w \in W} (1 - wT) = 1 - T^{f(p)}.$$

The lemma then follows from the fact that for all $w \in W$ there exists $g(p)$ characters χ of $(\mathbb{Z}/N\mathbb{Z})^*$ such that $\chi(p) = w$. □

Proposition 8.14. *One has*

$$f_N(s) = \prod_{p \nmid N} \left(1 - \frac{1}{p^{f(p)s}}\right)^{-g(p)}.$$

This is a Dirichlet series, with positive integral coefficients, converging in the half plane $\mathrm{Re}(s) > 1$.

Proof. This follows immediate from the previous lemma. □

Theorem 8.15. $L(1, \chi) \neq 0$ *for all non-principal character* χ.

Proof. Suppose $L(1, \chi) = 0$ for some non-principal character χ. Then $f_N(s)$ would be holomorphic at $s = 1$, thus also for all s with $\mathrm{Re}(s) > 0$. This is not possible: one can show that the p-th factor of $f_N(s)$ dominates the series

$$1 + p^{-\varphi(N)s} + p^{-2\varphi(N)s} + \cdots.$$

Therefore $f_N(s)$ has all its coefficients greater than those of the series

$$\sum_{\gcd(n,N)=1} n^{-\varphi(N)s}$$

which diverges for $s = \frac{1}{\varphi(N)}$. Contradiction. □

Remark 8.16. In fact, with a bit more analysis on $f_N(s)$, one can show that $L(s, \chi) \neq 0$ for all $\mathrm{Re}(s) = 1$ and character χ. We skip the proof here.

8.4. **Density and Dirichlet theorem.**

Definition 8.17. For $S \subseteq T$ two sets of positive integers, with $|T| = \infty$, the *upper natural density* and *lower natural density* of $S \subseteq T$ are defined as

$$\limsup_{N\to\infty} \frac{\#\{n \in S : n \leq N\}}{\#\{n \in T : n \leq N\}} \quad \text{and} \quad \liminf_{N\to\infty} \frac{\#\{n \in S : n \leq N\}}{\#\{n \in T : n \leq N\}}.$$

If they coincide, we call the common value the *natural density* of $S \subseteq T$. Otherwise, we say $S \subseteq T$ has no natural density.

Many interesting examples fail to have a natural density (see homework problems). Below is a less restrictive notion of density by using Dirichlet series.

Definition 8.18. For $S \subseteq T$ two sets of positive integers, with $\sum_{n\in T} n^{-1}$ divergent, the *upper Dirichlet density* and *lower Dirichlet density* of $S \subseteq T$ are defined as

$$\limsup_{s\to 1^+} \frac{\sum_{n\in S} n^{-s}}{\sum_{n\in T} n^{-s}} \quad \text{and} \quad \liminf_{s\to 1^+} \frac{\sum_{n\in S} n^{-s}}{\sum_{n\in T} n^{-s}}.$$

If they coincide, we call the common value the *Dirichlet density* of $S \subseteq T$.

Example 8.19. Consider the case $T = \mathbb{N}$. Recall that $\zeta(s)$ has a simple pole of residue 1 at $s = 1$, so as $s \to 1^+$ we have

$$(s-1)\sum_{n=1}^{\infty} n^{-s} = 1 + o(1).$$

Hence the Dirichlet density of $S \subseteq \mathbb{N}$ is given by

$$\lim_{s\to 1^+} (s-1)\sum_{n\in S} n^{-s}$$

if the limit exists.

Example 8.20. Let T be the set of all primes. From the above expression, we have that

$$\log(s-1) + \log\zeta(s) = \mathrm{O}(1).$$

On the other hand, recall that

$$\log\zeta(s) = \sum_p -\log(1-p^{-s}) = \sum_p \sum_{n=1}^{\infty} \frac{p^{-ns}}{n} = \sum_p p^{-s} + \sum_p \sum_{n=2}^{\infty} \frac{p^{-ns}}{n}.$$

Since the last term is bounded, we have

$$\sum_p p^{-s} = -\log(s-1) + \mathrm{O}(1).$$

Hence the Dirichlet density of a subset of primes $S \subseteq T$ is given by

$$\lim_{s\to 1^+} \frac{\sum_{p\in S} p^{-s}}{-\log(s-1)}$$

if the limit exists.

Example 8.21. It is clear that density is *finitely additive*: If $S_1, \ldots, S_n$ are disjoint subsets of T with densities $\delta_1, \ldots, \delta_n$, then their union has density $\delta_1 + \cdots + \delta_n$. Here is a fun example: Let $\alpha, \beta > 0$ be two irrational numbers with $\frac{1}{\alpha} + \frac{1}{\beta} = 1$. Let

$$S_\alpha = \{\lfloor n\alpha \rfloor : n \in \mathbb{N}\} \quad \text{and} \quad S_\beta = \{\lfloor n\beta \rfloor : n \in \mathbb{N}\}.$$

Then S_α and S_β have natural densities $\frac{1}{\alpha}$ and $\frac{1}{\beta}$ in $\mathbb{N}$. The fact that they add up to 1 can be explained by the beautiful result (Rayleigh's theorem) that S_α and S_β are disjoint and their union is $\mathbb{N}$.

Exercise. Let $S \subseteq T$ be subsets of $\mathbb{N}$ such that S has natural density δ in T. Prove that S also has Dirichlet density δ in T. However, the converse is *not true* (see homework problems); so the existence of natural density is usually a stronger condition than that of Dirichlet density.

As we shall see later, proving that some set of primes has a non-zero *Dirichlet density* usually involves showing that certain L-functions do not vanish at the point $s = 1$, while showing that they have a *natural density* involves showing that the L-functions have no zeros on the line $\mathrm{Re}(s) = 1$.

Theorem 8.22 (Dirichlet)**.** *For any positive integers m, N with $\gcd(m, N) = 1$, the set of of primes congruent to m modulo N has Dirichlet density $\frac{1}{\varphi(N)}$ in the set of all primes. In particular, there are infinitely many such primes.*

Proof. As $s \to 1$ we have

$$\log L(s, \chi) = \sum_p -\log(1 - \chi(p)p^{-s}) = \sum_p \sum_{n=1}^{\infty} \frac{\chi(p^n)p^{-ns}}{n} = \sum_p \chi(p)p^{-s} + \mathrm{O}(1).$$

- For χ non-principal, $L(s, \chi)$ is holomorphic and *non-vanishing* at $s = 1$, therefore $\sum_p \chi(p)p^{-s} = \mathrm{O}(1)$.
- For χ principal, $\sum_p \chi(p)p^{-s} = \sum_{p \nmid N} p^{-s} = -\log(s-1) + \mathrm{O}(1)$.

It should be clear now how to proceed: form a certain linear combination of the $\log L(s, \chi)$ to isolate the sum $\sum_{p \equiv m(N)} p^{-s}$, and compare with the asymptotics of $-\log(s-1)$. We can do this by applying the *orthogonal relations* of Dirichlet characters.

$$\sum_\chi \chi(m^{-1}) \log L(s, \chi) = \sum_\chi \sum_p \chi(m^{-1})\chi(p)p^{-s} + \mathrm{O}(1) = \varphi(N) \sum_{p \equiv m(N)} p^{-s} + \mathrm{O}(1).$$

On the other hand, we have by the previous argument that

$$\sum_{\chi} \chi(m^{-1}) \log L(s, \chi) = -\log(s-1) + \mathrm{O}(1).$$

This concludes the proof. □

Remark 8.23. In fact, the stronger statement that the set of primes congruent to m modulo N has *natural density* $\frac{1}{\varphi(N)}$ in the set of all primes, is also correct. To prove this, one needs to use the fact that $L(s, \chi)$ is non-vanishing on the line $\mathrm{Re}(s) = 1$. (One can observe that the proof of Dirichlet density above uses only the properties of the L-function at the point $s = 1$.)

As in the proof of the prime number theorem, it suffices to show that the improper integral

$$\int_1^{\infty} \frac{\varphi(N)\vartheta_m(x) - x}{x^2} \, \mathrm{d}x$$

converges, where

$$\vartheta_m(x) = \sum_{p \leq x,\ p \equiv m(N)} \log p.$$

For χ a Dirichlet character of level N, one defines

$$\vartheta_\chi(x) = \sum_{p \leq x} \chi(p) \log p.$$

Then $\vartheta_m(x)$ can be written as

$$\vartheta_m(x) = \frac{1}{\varphi(N)} \sum_{\chi} \chi(m^{-1}) \vartheta_\chi(x).$$

It then suffices to show that:

- For χ principal, the following integral converges

$$\int_1^{\infty} \frac{\vartheta_\chi(x) - x}{x^2} \, \mathrm{d}x.$$

  (This follows immediately from the corresponding fact of $\vartheta(x)$, since they are only differ by finitely many terms.)
- For χ non-principal, the following integral converges

$$\int_1^{\infty} \frac{\vartheta_\chi(x)}{x^2} \, \mathrm{d}x.$$

(To prove this, one needs to use the fact that $L(s,\chi) \neq 0$ for $\mathrm{Re}(s) \geq 1$, therefore $-\frac{L'(s,\chi)}{L(s,\chi)}$ is holomorphic in an open neighborhood of $\mathrm{Re}(s) \geq 1$. Then apply the similar *Tauberian argument* we discussed in the proof of the prime number theorem.)

Lecture 13

8.5. **The functional equation for the zeta function.** In this subsection, we establish the *functional equation* for the zeta function $\zeta(s)$, which will imply it can be meromorphically extended to the whole complex plane $\mathbb{C}$. (Similar extensions can be done for the Dirichlet L-functions.) One of Riemann's key observations is that in the strip $0 < \mathrm{Re}(s) < 1$, the zeta function obeys a symmetry property relating $\zeta(s)$ and $\zeta(1-s)$; once we prove this, we will then be able to extend $\zeta(s)$ all the way across the complex plane $\mathbb{C}$.

Definition 8.24. The *Gamma function* $\Gamma(s)$ is defined as

$$\Gamma(s) = \int_0^\infty e^{-t} t^{s-1} \, \mathrm{d}t \qquad \text{for} \quad \mathrm{Re}(s) > 0.$$

Using integration by parts, we have

$$\Gamma(s+1) = s\Gamma(s) \qquad \text{for} \quad \mathrm{Re}(s) > 0.$$

Since $\Gamma(1) = 1$, we have $\Gamma(n+1) = n!$ for $n \in \mathbb{Z}_{\geq 0}$. Therefore, the Gamma function can be considered as a "complex version" of the factorial function. Also, using the equation $\Gamma(s+1) = s\Gamma(s)$, we can extend Γ to a meromorphic function on all of $\mathbb{C}$, with simple poles at $s = 0, -1, -2, \ldots$.

Substituting $t = \pi n^2 x$, we find

$$\pi^{-\frac{s}{2}} \Gamma\left(\frac{s}{2}\right) n^{-s} = \int_0^\infty x^{\frac{s}{2}-1} e^{-n^2 \pi x} \, \mathrm{d}x \quad \text{for} \quad \mathrm{Re}(s) > 0.$$

If we sum over $n \geq 1$, then we can interchange the sum and integral for $\mathrm{Re}(s) > 1$ since it converges absolutely. Hence

$$\pi^{-\frac{s}{2}} \Gamma\left(\frac{s}{2}\right) \zeta(s) = \int_0^\infty x^{\frac{s}{2}-1} \omega(x) \, \mathrm{d}x \quad \text{for} \quad \mathrm{Re}(s) > 1$$

where

$$\omega(x) = \sum_{n=1}^\infty e^{-n^2 \pi x}.$$

Note that this is related to a function that we are familiar with:

$$2\omega(x) + 1 = \theta(x) = \sum_{n=-\infty}^{\infty} e^{-n^2\pi x}.$$

The theta function here is the same as the one we considered in the *sum of four square problems* up to a change of variable. Recall that it satisfies a transformation formula

$$\theta(x^{-1}) = x^{1/2}\theta(x) \qquad \text{for} \quad x > 0.$$

One can then deduce that

$$\omega(x^{-1}) = -\frac{1}{2} + \frac{1}{2}x^{1/2} + x^{1/2}\omega(x).$$

Separate the integral $\int_0^\infty x^{\frac{s}{2}-1}\omega(x)\,\mathrm{d}x$ into two parts $\int_1^\infty$ and $\int_0^1$ (and do a change of variable $x \mapsto 1/x$, one finds

$$\pi^{-\frac{s}{2}}\Gamma\left(\frac{s}{2}\right)\zeta(s) = -\frac{1}{s(1-s)} + \int_1^\infty \left(x^{\frac{s}{2}-1} + x^{\frac{1-s}{2}-1}\right)\omega(x)\,\mathrm{d}x \quad \text{for } \operatorname{Re}(s) > 1.$$

Observe that

- The left hand side is a meromorphic function on $\operatorname{Re}(s) > 0$.
- The right hand side is a meromorphic function on *all of* $s \in \mathbb{C}$, since $\omega(x) = \mathrm{O}(e^{-\pi x})$ as $x \to \infty$, with simple poles at $s = 0, 1$.

Define

$$\xi(s) = \frac{1}{2}s(s-1)\pi^{-\frac{s}{2}}\Gamma\left(\frac{s}{2}\right)\zeta(s).$$

Then $\xi(s)$ can be holomorphically extended to an entire holomorphic function. Moreover, it satisfies the *functional equation*

$$\xi(s) = \xi(1-s).$$

Together with the facts that the Gamma function has no zeros, and $\zeta(s)$ admits a meromorphic extension to $\operatorname{Re}(s) > 0$ with only a simple pole at $s = 1$, one can then conclude that

- $\zeta(s)$ can be extended meromorphically to all of $\mathbb{C}$, with only a pole at $s = 1$.
- Except for the region $0 < \operatorname{Re}(s) < 1$, the zeta function has zeros only at $s = -2, -4, -6, \ldots$. These are called the *trivial zeros* of ζ.

Conjecture 8.25 (Riemann hypothesis)**.** *All non-trivial zeros of* $\zeta(s)$ *lies on the line* $\mathrm{Re}(s) = \frac{1}{2}$.

9. Model theory and first-order logic

Theorem 9.1 (Ax, Grothendieck)**.** *Let* $G = (g_1, \dots, g_n) \colon \mathbb{C}^n \to \mathbb{C}^n$ *be a polynomial function. If it is injective, then it is surjective as well.*

The goal of this section is to explain a very beautiful proof of this theorem, using *model theory* from *mathematical logic* (in particular, *Gödel's completeness theorem*). There are in fact other ways of proving it without model theory; other proofs were given by Borel, Rudin, Conrad, and Serre among others. However, it will be clear that if we are given Gödel's completeness theorem, then it would be extremely hard to imagine a simpler proof of the theorem without using model theory. Moreover, model theory provides a bridge between algebraic geometry of characteristic 0 and algebraic geometry in positive characteristic, which is very useful.

Below is a sketch of the argument. We will be filling in the details in the remaining of this section.

(1) Formulate the statement we want to prove as a *first-order statement* in the *language* of the *theory of algebraically closed fields of characteristic zero*.
(2) Here comes the highly non-trivial fact from *model theory*: the theory of algebraically closed fields of characteristic zero is *complete*: this means that any statement expressible in the first-order language of the field is either provable or disprovable in the first-order language. This is known as the *Gödel's completeness theorem*.
(3) Therefore, suppose the statement fails for $\mathbb{C}$, then there is a *disproof* in the *first-order logic* of algebraically closed fields of characteristic zero.
(4) By the *finiteness of proof*, it only uses finitely many of the assumptions $p \neq 0$, hence the theorem also fails in algebraically closed fields of sufficiently large characteristic, say $\overline{\mathbb{F}}_p$.
(5) Any specific counterexample lies in some finite extension of $\mathbb{F}_p$, which is a finite field.
(6) However, the statement is obvious *true* for finite fields. Contradiction.

Remark 9.2. The statement is *not* true if one replaces the field $\mathbb{C}$ with the field of rational numbers $\mathbb{Q}$: the map $x \mapsto x^3$ is injective, but not surjective. The

reason that the above proof fails in this case is because $\mathbb{Q}$ is *not* algebraically closed.

Remark 9.3. The statement is *not* true if one exchanges "injective" and "surjective" in the statement: the map $z \mapsto z^2$ is surjective, but not injective. In this case, the above argument breaks down at Step (5): a surjective map $\overline{\mathbb{F}}_p \to \overline{\mathbb{F}}_p$ is not necessarily surjective on finite extensions of $\mathbb{F}_p$.

9.1. **Preliminary on Fields.** We have discussed the notion of *rings* in earlier section; a *field* can be regarded as a very special kind of *ring*.

Definition 9.4. A *field* is a set F equipped with two binary operations $+$ (addition) and $\cdot$ (multiplication), so that

- $(F, +, \cdot)$ forms a *commutative* ring; in particular, there exist elements $0 \neq 1$ in F which are the additive and multiplicative identities.
- All nonzero elements $F^\times := F \backslash \{0\}$ are invertible under multiplication.

Example 9.5. Here are some (non-)examples of fields:

- $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ are fields; $\mathbb{Z}, \mathbb{N}$ are *not* fields.
- Let p be a prime number. Then $\mathbb{F}_p = \{0, 1, \ldots, p-1\}$ with addition and multiplication given by the standard ones modulo p, forms a field. If one replaces p by a non-prime number, then the construction does *not* give a field (why?).

Definition 9.6. A field F is said to have *characteristic* 0 if

$$n \cdot 1 = 1 + 1 + \cdots + 1 \neq 0 \quad \text{for any } n \geq 1.$$

Otherwise, if there is a positive integer n satisfying $n \cdot 1 = 0$, then the smallest such positive integer can be shown to be a prime number. It is usually denoted by p and the field is said to have *characteristic* p.

Remark 9.7. For a field F with characteristic $p > 0$, there is a *Frobenius map*

$$\mathrm{Frob}\colon F \to F; \quad x \mapsto x^p$$

which is a *field homomorphism*, i.e. it is compatible with the addition and multiplication of F. This map is crucially important in the study of fields in characteristic p, and therefore in *number theory*.

Remark 9.8. A field is called a *prime field* if it has no proper (i.e. strictly smaller) subfields. Any field F contains a prime field (the smallest subfield contains $0, 1 \in F$).

- If $\mathrm{char}(F) = p > 0$, then the prime field of F is isomorphic to $\mathbb{F}_p$.
- If $\mathrm{char}(F) = 0$, then the prime field of F is isomorphic to $\mathbb{Q}$.

Definition 9.9. Let $E \subseteq F$ be a subfield. We will say that F is a *field extension* (or just *extension*) of E. Note that F can be regarded as a *vector space* over E. The *degree* $[F : E]$ of this field extension is defined to be the dimension of F as an E-vector space.

Example 9.10. $\mathbb{C}$ is an extension of $\mathbb{R}$, with degree $[\mathbb{C} : \mathbb{R}] = 2$.

Exercise. Any finite field has $q = p^n$ elements, where p is prime and $n \geq 1$.

Exercise. Let $E \subseteq F \subseteq K$ be field extensions. Then $[K : E] = [K : F][F : E]$.

A central notion in the study of field extensions are *algebraic elements*.

Definition 9.11. Let $E \subseteq F$ be a field extension. An element $x \in F$ is *algebraic* over E if it is a root of a polynomial with coefficients in E. A field extension in which every element of F is algebraic over E is called an *algebraic extension*.

Exercise. Any finite extension (i.e. $[F : E] < \infty$) is algebraic. However, not every algebraic extension is finite.

Exercise. Let $E \subseteq F$ be a field extension. Suppose $x_1, \ldots, x_n \in F$ are algebraic over E. Then the extension $E \subseteq E(x_1, \ldots, x_n)$ is finite. (Here, $E(x_1, \ldots, x_n) \subseteq F$ is the smallest field containing E and $x_1, \ldots, x_n$.)

Definition 9.12. A field F is *algebraically closed* if it does not have any strictly bigger algebraic extensions, or equivalently, every non-constant polynomial in $F[x]$ has a root in F.

Example 9.13. Here are some (non-)examples of algebraically closed fields:

- $\mathbb{C}$ is algebraically closed by the fundamental theorem of algebra, which we proved several times previously.
- $\mathbb{Q}, \mathbb{R}$ are *not* algebraically closed.
- $\mathbb{F}_p$ is *not* algebraically closed.

Theorem 9.14. *Any field F has an* algebraic closure*, which is a field containing F, algebraic over F, and is algebraically closed. Moreover, the algebraic closure of F is unique up to (field) isomorphism, and is denoted by $\overline{F}$.*

Remark 9.15. The field $\overline{F}$ is usually rather implicit since its construction requires the *ultrafilter lemma*, which is a set-theoretic axiom that is weaker than the *axiom of choice*. In this regard, the algebraic closure of finite field $\mathbb{F}_p$ is exceptionally simple: it is simply the union of the finite fields containing $\mathbb{F}_p$ (i.e. the ones of order p^n). As another example, the algebraic closure $\overline{\mathbb{Q}}$ of $\mathbb{Q}$ is called the field of *algebraic numbers*, which is one of the fundamental objects in *algebraic number theory*.

Let us prove the following positive characteristic version of the Ax–Grothendieck theorem.

Lemma 9.16. *Let $G\colon \overline{\mathbb{F}}_p^n \to \overline{\mathbb{F}}_p^n$ be a polynomial map. If G is injective, then it is also surjective.*

Proof. Assume the contrary that G is not surjective, say $x \in \overline{\mathbb{F}}_p^n$ is not in the image of G. Let $\mathbb{F}_q$ be an extension of $\mathbb{F}_p$ containing all of the coefficients of G, as well as the coordinates of x. Then G induces a polynomial map $\widetilde{G}\colon \mathbb{F}_q^n \to \mathbb{F}_q^n$. Then $\widetilde{G}$ is injective and non-surjective by our assumption. However, this is not possible since $\mathbb{F}_q^n$ consists of finitely many elements. Contradiction. □

The proof of this $\overline{\mathbb{F}}_p$-version of the Ax–Grothendieck theorem is really simple: it simply boils down to the obvious fact that if a map from a finite set to itself is injective, then it must be bijective.

It turns out that powerful tools from *model theory* (in this case, *Gödel's completeness theorem*), directly show that the $\overline{\mathbb{F}}_p$-version of the Ax–Grothendieck theorem *implies* the $\mathbb{C}$-version which we would like to prove. This gives an extremely simple proof of the Ax–Grothendieck theorem.

Remark 9.17. Recall we mentioned that if one exchanges "injective" and "surjective" then the statement of Ax–Grothendieck theorem would be false. The model-theoretic step that reduces the $\mathbb{C}$-version to the $\overline{\mathbb{F}}_p$-version is still valid, but Lemma 9.16 would be untrue if one exchanges "injective" and "surjective" in the statement: There exists surjective polynomial map $\overline{\mathbb{F}}_p^n \to \overline{\mathbb{F}}_p^n$ which is not injective. The reason that the same proof does not work here is that the restriction of a surjective map may not be surjective (unlike the restriction of an injective map is always injective).

Lecture 12

9.2. **Model theory.** We explain some basic ideas of *model theory*, and complete the beautiful proof of the Ax–Grothendieck theorem. A nice reference is a survey by Clark [4].

Definition 9.18. A *language*

$$\mathcal{L} = \{c_i, f_j, r_k \mid i \in I, j \in J, k \in K\}$$

is a set of symbols. The symbols c_i are *constant symbols*, f_j are *function symbols*, and r_k are *relation symbols*. We require that associated to each function symbol f_j and relation symbol r_k is a unique natural number n_{f_j} or n_{r_k} respectively, defined as the the *arity* of the function or relation symbol. The arity is meant to represent the intended number of inputs of a function or relation represented by the function or relation symbol.

Example 9.19. Here are some basic examples of languages.

- The *language of rings* $\mathcal{L}_{\text{ring}} = \{0, 1, +, \cdot\}$, with constant symbols $0, 1$ and function symbols $+, \cdot$ with $n_+ = n_- = 2$.
- The *language of sets* $\mathcal{L}_{\text{set}} = \{\in\}$ with one relation symbol with $n_\in = 2$.
- The *language of orderings* $\mathcal{L}_{\text{order}} = \{\leq\}$ with one relation symbol with $n_\leq = 2$.

Languages are the lexicon of non-logical symbols from which sentences and formulas of *first-order* are built. The logical symbols are the standard symbols of *first-order logic*, which are:

- quantifiers symbols: $\forall$ and $\exists$.
- logical connectives: $\wedge$ (and), $\vee$ (or), $\to$ (imply), $\leftrightarrow$ (iff), $\neg$ (negation).
- parentheses: ().
- equality symbol: $=$.
- countable set of variable symbols: $x_1, x_2, \ldots$.

Regardless of choice of language, we allow logical symbols to be included as valid building blocks for any *formula*. Intuitively, a formula $\phi(x, y)$ with variables x, y is a "valid" string of symbols such that when when we replace (x, y) with a pair of elements (a, b) from a mathematical object M, the resulting expression $\phi(a, b)$ is something that can be evaluated as true or false in M.

For instance, in the language of rings, when $M = \mathbb{Z}$ and

$$\phi(x, y) \coloneqq x + y = 0.$$

Then $\phi(1,-1)$ is true, while $\phi(1,0)$ is false. The actual definition of formulas and sentences is by a tedious induction, so instead we give illustrating examples.

Example 9.20. The following are examples of *formulas* in their respective languages $\mathcal{L}$.

(1) $\mathcal{L} = \mathcal{L}_{\text{ring}}$; $\phi(x,y) \coloneqq x + y = 0$ (i.e. x and y sum to 0).
(2) $\mathcal{L} = \mathcal{L}_{\text{ring}}$; $\phi(x) \coloneqq ((x \cdot x) - 1) - 1 = 0$ (i.e. x is a square root of 2).
(3) $\mathcal{L} = \mathcal{L}_{\text{order}}$; $\phi \coloneqq \exists x \forall y (x \leq y)$ (i.e. there exists a lower bound).

Definition 9.21.
- A *bounded variable* is a variable referred to by quantifiers. The variables x, y in the third example are bounded variables.
- A *free variable* is a variable that is not bounded. The variables x, y in the first example are free variables.
- A formula with no free variables, like the third example, is called a *sentence*.

Definition 9.22. A *theory* of $\mathcal{L}$ is a set of sentences T each with symbols only from $\mathcal{L}$ and logical symbols. We say T is an *$\mathcal{L}$-theory*.

Example 9.23. Let $\mathcal{L} = \mathcal{L}_{\text{ring}}$ be the language of rings. Define

$$\phi_c \coloneqq \forall x \forall y \; xy = yx.$$

$$\phi_i \coloneqq \forall x \exists y \; xy = yx = 1.$$

These two sentences together give the *theory of fields*. In addition, for each $n \in \mathbb{N}$ we define

$$\phi_n \coloneqq \forall x_1 \forall x_2 \cdots \forall x_n \exists t \; t^n + x_1 t^{n-1} + \cdots + x_n = 0.$$

(Note that for each given n, this is a sentence with finitely many symbols from $\mathcal{L}$ and logical symbols.) Then $\{\phi_c, \phi_i\} \cup \{\phi_n\}_{n \geq 1}$ gives the *theory of algebraically closed field*, which we denoted by ACF.

Moreover, for a prime number p, we define

$$\text{ACF}_p \coloneqq \text{ACF} \cup \{p \cdot 1 = 0\}$$

to be the *theory of algebraically closed field of characteristic p*; and we define

$$\text{ACF}_0 \coloneqq \text{ACF} \cup \{n \cdot 1 \neq 0\}_{n \geq 1}$$

to be the *theory of algebraically closed field of characteristic* 0.

We now explain what is a *proof*. The formal definition of a proof is inductive and tedious, so we only give a working idea of it. Let σ be a sentence in $\mathcal{L}$ that we think of as a theorem to be proved, and let T be a $\mathcal{L}$-theory that we think of as a set of axioms. A *proof of σ from T* is a *finite* sequence of sentences $\sigma_1, \sigma_2, \ldots, \sigma_n \equiv \sigma$ such that each σ_i is an element of T, or a standard axiom of logic, or deduced from previous elements of the sequence. In particular, a proof is a *finite* and *first-order* process.

Definition 9.24. Let T be an $\mathcal{L}$-theory and let σ be an $\mathcal{L}$-sentence.

- We say T *proves* σ (denoted $T \vdash \sigma$) if there is a proof of σ from T.
- We say T is *complete* if it proves every $\mathcal{L}$-sentence or its negation, i.e. for every $\mathcal{L}$-sentence ϕ, either $T \vdash \phi$ or $T \vdash \neg\phi$.
- We say T is *consistent* if it does not prove a contradiction, i.e. if $T \vdash \phi$ then $T \nvdash \neg\phi$ for every $\mathcal{L}$-sentence ϕ.

A big theorem that we will be using is the following completeness theorem.

Theorem 9.25 (Gödel completeness theorem)**.** ACF_p *is complete for p prime or $p = 0$.*

In order to state the *transfer principle* that bridges characteristic 0 and characteristic p, we need to introduce the notion of *model*.

Definition 9.26. Let $\mathcal{L} = \{c_i, f_j, r_k \mid i \in I, j \in J, k \in K\}$ be a language. A *model* of $\mathcal{L}$ is a set M together with elements c_i', functions f_j', and relations r_k' for $i \in I$, $j \in J$, $k \in K$ such that:

- For every constant symbol $c_i \in \mathcal{L}$, we have $c_i' \in M$.
- For every function symbol $f_j \in \mathcal{L}$ with arity n, we have that f_j' is a function $M^n \to M$.
- For every relation symbol $r_k \in \mathcal{L}$ with arity s, we have that r_k' is a s-ary relation of M, i.e. $r_k' \subseteq M^s$.

When there is no confusion, we use the same symbols c_i, f_j, r_k to denote the actual constants, actual functions, and actual relations of M.

Example 9.27. The following are models of their respective languages:

- $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$ are models of $\mathcal{L}_{\mathrm{ring}}$.
- $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}$ with the usual ordering are models of $\mathcal{L}_{\mathrm{order}}$.

Definition 9.28. If an $\mathcal{L}$-sentence σ is *true* in a model M of $\mathcal{L}$, we say M *is a model of* σ or M *models* σ (denoted $M \models \sigma$). For an $\mathcal{L}$-theory T, we say M *models* T (denoted $M \models T$) if M models every sentence in T.

Remark 9.29. Note that the truth of a sentence is usually model-dependent: for instance, "-1 has a square root" is true in $\mathbb{C}$, but not in $\mathbb{R}$. In terms of symbols, we write

$$\mathbb{C} \models (\exists x) x^2 = -1, \qquad \mathbb{R} \not\models (\exists x) x^2 = -1.$$

Example 9.30. $\mathbb{C}$ models ACF_0; $\overline{\mathbb{F}}_p$ models ACF_p.

We now state the *transfer principle*.

Theorem 9.31 (Ax's transfer principle)**.** *Let ϕ be a sentence in $\mathcal{L}_{ring}$. The following statements are equivalent:*

(1) $\mathbb{C} \models \phi$.
(2) $\mathrm{ACF}_0 \vdash \phi$.
(3) There exists a constant N such that $\mathrm{ACF}_p \vdash \phi$ *for all* $p > N$.
(4) There exists a constant N such that $\overline{\mathbb{F}}_p \models \phi$ *for all* $p > N$.

Proof. We have (2) implies (1) since $\mathbb{C}$ models ACF_0. Conversely, assuming (1), i.e. ϕ is true in the field $\mathbb{C}$. By the completeness of ACF_0, we have either $\mathrm{ACF}_0 \vdash \phi$ or $\mathrm{ACF}_0 \vdash \neg\phi$, i.e. we can either prove or disprove the statement ϕ using ACF_0. Since ϕ is true in $\mathbb{C}$, therefore it is impossible to disprove ϕ using ACF_0. Thus we have (2).

Suppose (2), i.e. $\mathrm{ACF}_0 \vdash \phi$. By the *finiteness of proof*, there exists a finite subset T of ACF_0 such that $T \vdash \phi$. Since $T \subseteq \mathrm{ACF}_p$ for all but finitely many primes p, thus (3) follows. Conversely, suppose (3) holds and assume the contrary that $\mathrm{ACF}_0 \vdash \phi$ does not hold. Then the completeness of ACF_0 implies that $\mathrm{ACF}_0 \vdash \neg\phi$, and the above argument shows that $\mathrm{ACF}_p \vdash \neg\phi$ for all but finitely many primes p. Contradiction.

Finally, the equivalence between (3) and (4) can be proved similarly as (1) $\Leftrightarrow$ (2). $\square$

Remark 9.32. Note that the transfer principle works only when the statement can be expressed as a (first order) sentence, i.e. logic symbols and the language of rings. For instance, a statement such as "the number of elements in an

algebraic closed field is uncountable" can *not* be transferred: indeed, the field $\mathbb{C}$ consists of uncountably many elements, while $\overline{\mathbb{F}}_p$ has only countably many.

Proof of the Ax–Grothendieck theorem. For each fixed $n \in \mathbb{N}$, the statement of Ax–Grothendieck can be written as a sentence ϕ of $\mathcal{L}_{\text{ring}}$ (simple exercise), and we would like to show that $\mathbb{C} \models \phi$. This follows directly from Lemma 9.16 and the Ax's transfer principle. □

This works like a magic: One starts with the elementary fact that if a map from a finite set to itself is injective then it is bijective, and through the black box of Gödel completeness theorem, one obtains a highly non-trivial theorem in complex algebraic geometry.

Remark 9.33. Another great example of an elegant model-theoretic proof is the *Hilbert's 17th problem*. A polynomial $f \in \mathbb{R}[t_1, \dots, t_n]$ is said to be *positive semidefinite* if

$$f(x_1, \dots, x_n) \geq 0 \quad \text{for any } (x_1, \dots, x_n) \in \mathbb{R}^n.$$

Hilbert's 17th problem asks whether every positive semidefinite polynomial over $\mathbb{R}$ can be written as a sum of square of *rational* functions. The proof relies on the completeness of the *theory of real closed fields* in the *language of fields*, and the *extension of language* (from the language of fields to the language of ordered fields). It again showcases beautiful ideas from model theory, but the machinery is a bit more involved, so we refer the reader to the survey [4].

Then we return to Section 8.2

10. Conway's topograph

A nice reference of this section is the first chapter of a beautiful book written by Conway [5].

10.1. **Topograph and definite forms: The well.** Given integers $a, b, h \in \mathbb{Z}$, one can associate a quadratic form

$$Q(x, y) = ax^2 + hxy + by^2.$$

We would like to understand for which $n \in \mathbb{Z}$ does there exist $\vec{v} = (x, y) \in \mathbb{Z}^2$ such that $Q(\vec{v}) = Q(x, y) = n$.

Let us start with two simple observations.

- We have $Q(k\vec{v}) = k^2 Q(\vec{v})$ for any integer k. Therefore, it suffices to understand the values of Q for *primitive* vectors $\vec{v}$.
- We have $Q(-\vec{v}) = Q(\vec{v})$. So we will usually identify $\vec{v}$ with $-\vec{v}$, or denote them together as $\pm\vec{v}$.

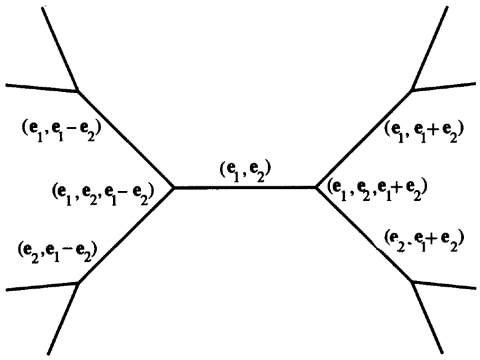
Definition 10.1. We say $\{\pm\vec{f_1}, \pm\vec{f_2}\} \subseteq \mathbb{Z}^2$ is a *basis* if for any $\vec{v} \in \mathbb{Z}^2$ there exists $k_1, k_2 \in \mathbb{Z}$ such that $\vec{v} = k_1\vec{f_1} + k_2\vec{v}_2$.

Example 10.2. Denote $\vec{e}_1$ and $\vec{e}_2$ the standard basis vectors of $\mathbb{R}^2$. Then $\{\pm\vec{e}_1, \pm\vec{e}_2\}$ and $\{\pm\vec{e}_1, \pm(\vec{e}_1 + \vec{e}_2)\}$ are bases, while $\{\pm\vec{e}_1, \pm 2\vec{e}_2\}$ is not.

Definition 10.3. We say $\{\pm\vec{f_1}, \pm\vec{f_2}, \pm\vec{f_3}\} \subseteq \mathbb{Z}^2$ is a *superbasis* if $\{\pm\vec{f_1}, \pm\vec{f_2}\}$ is a basis and $\vec{f_1} + \vec{f_2} + \vec{f_3} = 0$.

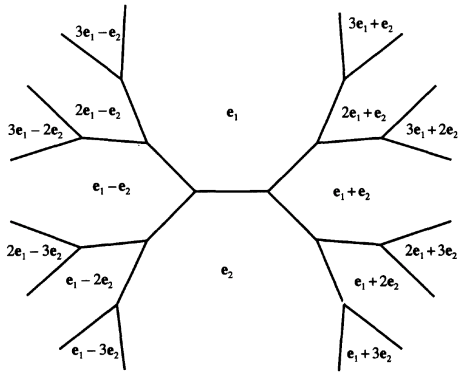
One can easily check the following facts:

- Any basis $\{\pm\vec{f_1}, \pm\vec{f_2}\}$ belongs to two superbases: $\{\pm\vec{f_1}, \pm\vec{f_2}, \pm(\vec{f_1}+\vec{f_2})\}$ and $\{\pm\vec{f_1}, \pm\vec{f_2}, \pm(\vec{f_1} - \vec{f_2})\}$.
- Any superbasis $\{\pm\vec{f_1}, \pm\vec{f_2}, \pm\vec{f_3}\}$ contains three bases.



Then we can draw a 3-valence graph in $\mathbb{R}^2$, with edges corresponding to bases, and vertices corresponding to superbases. Moreover, one can observe that the vertices and edges that involve a given vector $\pm\vec{f}$ (e.g. $\pm\vec{e}_1$) form a path. Therefore, we can add a face bounded by this path to our topograph and identify it with $\pm\vec{f}$. In the resulting fully labeled topograph:

- each region is labeled with a vector $\pm\vec{f}$,
- two regions separated by an edge form a basis,



- three regions around a vertex form a superbasis.

This graph is known as the *Conway's topograph*.

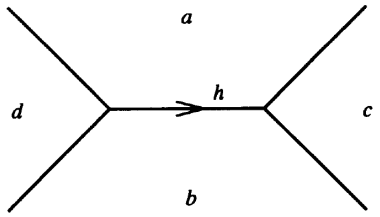
Up to this point, the discussion has nothing to do with quadratic forms. Now, we fix an integral quadratic form Q, and call $Q(\vec{v})$ the *norm* of $\vec{v}$. It turns out that if we know the norms at the three vectors of some superbasis, then the norms of all other vectors are determined! This follows from a simple fact that

$$Q(\vec{v}_1 + \vec{v}_2) + Q(\vec{v}_1 - \vec{v}_2) = 2(Q(\vec{v}_1) + Q(\vec{v}_2)).$$

This formula tells us that if we let

$$a = Q(\vec{v}_1), \quad b = Q(\vec{v}_2), \quad c = Q(\vec{v}_1 + \vec{v}_2), \quad d = Q(\vec{v}_1 - \vec{v}_2)$$

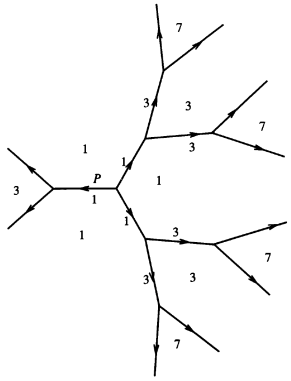
then $d, a + b, c$ forms an arithmetic progression.



Besides marking the norm on each region (a, b, c, d in the figure above), we also mark each edge with a direction and an appropriate number $h > 0$. The above figure means that

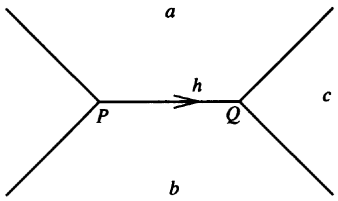
$$c = (a + b) + h \qquad \text{and} \qquad d = (a + b) - h.$$

If $c = d = a + b$, then we omit the arrow and the number h. For instance, below is the marking of $x^2 + xy + y^2$ on part of the topograph.



It is not hard to see that if we know the markings around a vertex, then all the remaining markings can be determined (providing some basic properties of the topograph which we will prove later, e.g. it is connected).

Lemma 10.4 (Climbing lemma)**.** *Suppose a, b, h in the figure below are all positive.*



Then c is also positive, and the edges that emerge from Q both point away from Q.

Proof. This follows from a direct computation using the above arithmetic progression law. □

Proposition 10.5. *The connected component of the topograph containing the vertex $\{\pm\vec{e}_1, \pm\vec{e}_2, \pm(\vec{e}_1 - \vec{e}_2)\}$ has no* cycles*.*

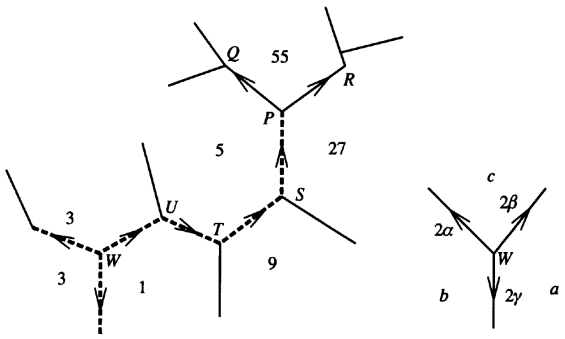
Proof. Consider the quadratic form $Q = x^2 + xy + y^2$; part of its topograph was drawn in previous examples. The vertex $\{\pm\vec{e}_1, \pm\vec{e}_2, \pm(\vec{e}_1 - \vec{e}_2)\}$ is a *well* with respect to Q: all edges are pointed outward from this vertex. Now, by the climbing lemma, all of the numbers involved keep getting larger and larger, therefore the topograph cannot have any cycles. □

This proof is quite interesting, in that we are proving a property of the topograph (which is independent of any quadratic form) by considering the

markings of a particular quadratic form. We will use the similar strategy to prove that the topography is *connected*.

Definition 10.6. A quadratic form Q is called *positive (semi)definite* if $Q(\vec{v}) > 0$ (resp. $Q(\vec{v}) \geq 0$) for all $\vec{v} \neq 0$. The notion of *negative (semi)definite* is similarly defined.

Let us look at the topograph of a positive definite quadratic form whose values at some superbasis are $5, 27, 55$.



The climbing lemma shows that if we walk away from P through either Q or R, the numbers will increase. Instead, we step down to S, at which the values are $5, 9, 27$. Repeating this process, we find ourselves stepping down against the (increasing) flow along the dashed path $STUW$ in the figure.

We stop at W because all three arrows of W are pointed outwards. We call a superbasis W a *well* if its three edges are all pointed outwards. Say the edge marks are $2\alpha, 2\beta, 2\gamma \geq 0$, and the values at the superbasis are $a, b, c > 0$. Then the arithmetic progress law says that

$$2\alpha = b + c - a, \quad 2\beta = c + a - b, \quad 2\gamma = a + b - c,$$

and so

$$a = \beta + \gamma, \quad b = \gamma + \alpha, \quad c = \alpha + \beta.$$

This process shows that there always exists a well for any positive definite Q.

Lemma 10.7 (Well lemma)**.** *Suppose we have a well for a positive definite form. Then the three vectors in this superbasis are the three primitive vectors of smallest norm.*

Proof. Let $\{\pm\vec{e}_1, \pm\vec{e}_2, \pm\vec{e}_3\}$ denote the superbasis at the well. Write a general vector $\vec{v} \in \mathbb{Z}^2$ as

$$\vec{v} = m_1\vec{e}_1 + m_2\vec{e}_2 + m_3\vec{e}_3.$$

One can verify that

$$Q(\vec{v}) = \alpha(m_2 - m_3)^2 + \beta(m_3 - m_1)^2 + \gamma(m_1 - m_2)^2.$$

Also note that since $\vec{e}_1 + \vec{e}_2 + \vec{e}_3 = 0$, simultaneously subtract m_1, m_2, m_3 by a same number would yield the same vector $\vec{v}$.

Suppose $\vec{v}$ is a primitive vector that is not in the superbasis, then all of the differences $m_i - m_j$ are nonzero, so $Q(\vec{v}) \geq \alpha + \beta + \gamma$ which is at least as big as each of $a = \beta + \gamma$, $b = \gamma + \alpha$, and $c = \alpha + \beta$. □

Proposition 10.8. *The topograph is connected.*

Proof. Consider again the positive definite quadratic form $Q = x^2 + xy + y^2$. It has a well with three vectors in the superbasis has norm $1, 1, 1$. Moreover, by the above argument, any other primitive vector has norm $Q(\vec{v}) \geq 1+1+1 = 3$. Therefore the well of this quadratic form is unique. By climbing down, any primitive vector has to be connected to this well. □

Remark 10.9. When the edge marking of a well α, β, γ are all positive, then the well is unique, and we call this a *simple well*.

On the other hand, if a well is not simple, then without loss of generality, say $\gamma = 0$. Then $a = \beta$ and $b = \alpha$, and the norm is

$$Q(\vec{v}) = b(m_2 - m_3)^2 + a(m_1 - m_2)^2.$$

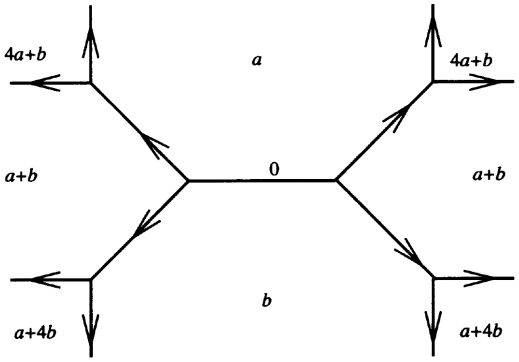
So the value of Q at $m_1\vec{e}_1 + m_2\vec{e}_2$ is $am_1^2 + bm_2^2$, and at the four vectors

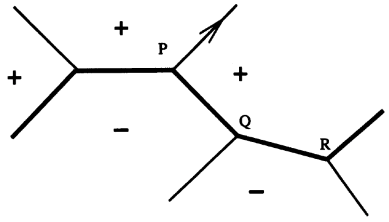
$$\pm\vec{e}_1, \quad \pm\vec{e}_2, \quad \pm(\vec{e}_1 + \vec{e}_2), \quad \pm(\vec{e}_1 - \vec{e}_2)$$

the values are

$$a, \quad b, \quad , a + b, \quad a + b$$

and everywhere else its values are strictly larger. In this case, the positive definite form has two wells on each end of an edge, and every other edge has an arrow pointing away from this edge. We call this a *double well*.

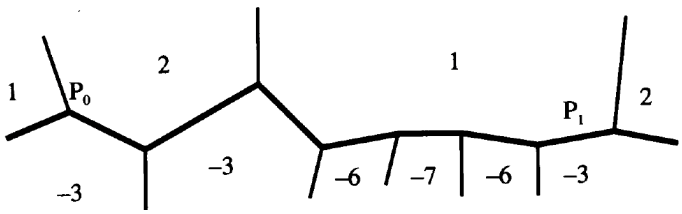




10.2. **Indefinite forms not representing** 0**: The river.** In this case, the topograph must contain an edge lying directly between a positive and a negative value.

The climbing lemma shows that if we climb away from the river on the positive side, the values will continually increase. Similarly, if we move away from the other side, the values get more and more negative. Note that this proves that the river is unique, because the topograph is connected, and if you move away from the river, you will see values of only one sign. So you will not get to another river.

Example 10.10. Consider the indefinite from $Q = x^2 + 4xy - 3y^2$. Its river looks like:



Note that the (values of the) river is *periodic*! One can start with the superbasis P_0 on the river, and finds that it reaches another superbasis P_1 with the same

surrounding values $(1, 2, -3)$. As an application, this *proves* that $x^2 + 4xy - 3y^2 = -2$ has no integral solutions.

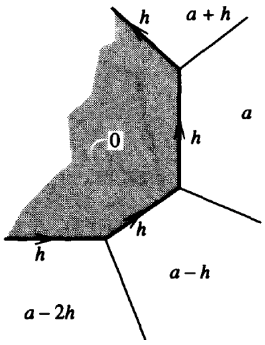
Proposition 10.11. *The river of an integral indefinite quadratic form is periodic.*

Proof. Consider an edge on the river. Write the values on both sides of the edge as a, b (where $ab < 0$), with a marking $h > 0$ on the edge. Up to a change of basis, one can write the quadratic form as $ax^2+hxy+by^2$. The *discriminant* of the quadratic form $d = ab - (\frac{1}{2}h)^2 < 0$ is independent of the choice of a basis. Therefore, for any edge on the river, the corresponding triple (a, b, h) always satisfies

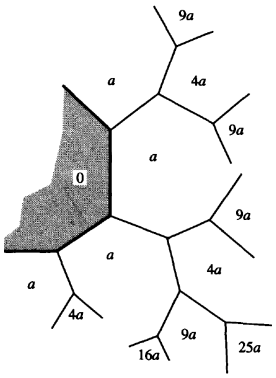
$$ab - \left(\frac{1}{2}h\right)^2 = d.$$

Thus there are only finitely many possible such triples (a, b, h), so such triple must repeat somewhere on the river. □

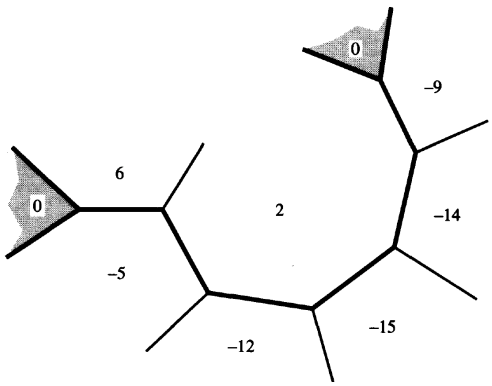
10.3. **Semidefinite forms: The lake.** A *lake* is the region corresponding to a vector where the form represents 0. Then the arithmetic progression law tells us that the values in the regions around a lake form an infinite arithmetic progression, as in the following figure.



However, if a form is semidefinite (either positive or negative), then the h in the figure must be zero; otherwise there would be both positive and negative terms in the sequence $\{\ldots, a - h, a, a + h, \ldots\}$. So it actually looks like the following:



10.4. **Indefinite forms representing** 0**.** In this case, we have a lake, and a non-constant arithmetic progression around the lake, which must change sign somewhere around the lake shore. If the change is directly between positive and negative, then it happens at an edge of some river flowing out from the lake. Since the values on a river is periodic, it must end by flowing into another lake.

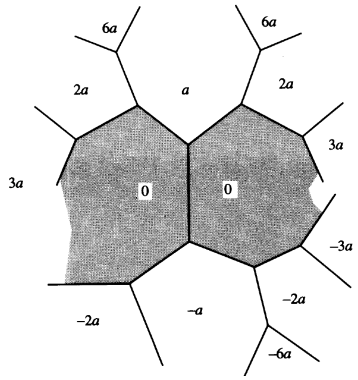


There is a special case in which the river is of zero length, which happens when the arithmetic progression along the lake contains zero. The form is then equivalent to hxy, and the topograph has two lakes abutting along an edge – the *weir* – with positive values on one side and negative ones on the other.

Let us summarize with the following theorem.

Theorem 10.12. *For any integers a, b, h, n, there is an algorithm to decide whether the Diophantine equation*

$$ax^2 + hxy + by^2 = n$$



is solvable for integers $(x, y) \in \mathbb{Z}^2$*, and to find such integers in the case when it is solvable.*

11. Miscellaneous Topics

11.1. **The ambiguous clock.** Suppose that one has a clock with an obstacle that its hour hand and minute hand look exactly the same (which is not suppose to happen!). With a bit of thinking, one realizes that it still is possible to know the exact time at almost every moment. On the other hand, there are moments where, if one exchanges the location of the hour hand and the minute hand, it still is a time position that makes sense. In this case, we cannot tell what time it is, since both positions give different times that make sense.

Question 11.1. *How many times in a day where the clock is ambiguous?*

It turns out that this question has an interesting *topological* answer. Consider the two-dimensional torus $T^2 \cong S^1 \times S^1$. Let α and β be the two circles respectively. One can show that the answer is given by (two times) the *intersection numbers* of the two curves

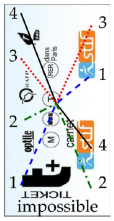
$$\alpha + 12\beta \qquad \text{and} \qquad 12\alpha + \beta$$

and minus 11 (why?). The intersection number can be computed in the *homology group* of T^2:

$$|(\alpha + 12\beta).(12\alpha + \beta)| = 12^2 - 1^2 = 143.$$

Here one has to be careful that the pairing $\alpha.\beta = -\beta.\alpha$ because of certain *orientation* issues.

11.2. **Kontsevich's four polynomial theorem.** In March 2009, two mathematicians, Ghys and Kontsevich, were attending an administrative meeting. Both of them were being bored. Suddenly, Kontsevich passed Ghys a Paris metro ticket containing a scribble and a single word: "impossible".



That was the new theorem he wanted to share with Ghys! Here is the statement.

Theorem 11.2. *Four polynomials P_1, P_2, P_3, P_4 of a real variable x cannot satisfy*

- $P_1(x) < P_2(x) < P_3(x) < P_4(x)$ *for small* $x < 0$.
- $P_1(0) = P_2(0) = P_3(0) = P_4(0)$.
- $P_2(x) > P_4(x) > P_1(x) > P_3(x)$ *for small* $x > 0$.

The *relative position* of the graphs of four real polynomials is subject to some constrains. This was fascinating: a new elementary result on polynomials in 2009! Later on, they put this into a more general context; it turns out to be closely related to singularities in real algebraic geometry, chord diagram, Möbius strip, knots and links, operad theory, Kontsevich's universal invariant of knots, and many others. These can be found in a *very* nice book by Ghys [8].

Proof. Let us define the *valuation* $v(P)$ of a polynomial $P = a_0 + a_1x + a_2x^2 + \cdots$ to be the smallest k such that $a_k \neq 0$.

Replacing P_i by $P_i - P_1$, we can assume that $P_1 = 0$. Since P_2 and P_4 change sign at the origin, their valuations $v(P_2), v(P_4)$ are odd; and $v(P_3)$ is even.

On the other hand, from $0 < P_2(x) < P_3(x) < P_4(x)$ for small $x < 0$, we deduce that $v(P_2) \geq v(P_3) \geq v(P_4)$. Similarly, $P_2(x) > P_4(x) > 0$ for small $x > 0$ implies that $v(P_2) \leq v(P_4)$. This forces $v(P_2) = v(P_3) = v(P_4)$. Contradiction. □

Definition 11.3. Let $n \geq 2$ be an integer and π be a permutation of $\{1, \ldots, n\}$. We say π is a *polynomial interchange* if there exists n polynomials $P_1, \ldots, P_n$ such that

- $P_1(x) < \cdots < P_n(x)$ for small $x < 0$.
- $P_1(0) = \cdots = P_n(0)$.
- $P_{\pi(1)}(x) > \cdots > P_{\pi(n)}(x)$ for small $x > 0$.

Kontsevich's theorem states that $(1243) \in S_4$ is *not* a polynomial exchange. Similarly, one can show that $(1342) \in S_4$ is not a polynomial exchange. It turns out that the remaining 22 permutations in S_4 are polynomial exchange. It is a fun exercise to find four such polynomials for each case.

Definition 11.4. Let $n \geq 2$ be an integer and π be a permutation of $\{1, \ldots, n\}$. We say π is separable if does not "contain" one of the two "forbidden" permutations, i.e. if there do not exist four indices $1 \leq i_1 < i_2 < i_3 < i_4 \leq n$ such that

$$\pi(i_2) < \pi(i_4) < \pi(i_1) < \pi(i_3) \quad \text{or} \quad \pi(i_3) < \pi(i_1) < \pi(i_4) < \pi(i_2).$$

By the definition, a polynomial interchange is necessarily separable. We will show that the converse is also true, i.e. the forbidden permutations found by Kontsevich are the only constrains for such polynomials to exist. Let us begin with a combinatorial lemma.

Lemma 11.5. *Let $n \geq 3$. Suppose π is a separable permutation. Then there is a* proper *interval $I = \{k, k+1, \ldots, k+\ell\} \subseteq \{1, \ldots, n\}$ of length $\ell + 1 \geq 2$ whose image by π is also an interval.*

Proof. We may assume $\pi(1) < \pi(2)$ since otherwise we could replace π by the "reverse" permutation $\overline{\pi}(k) = n + 1 - \pi(k)$. If $\pi(2) = \pi(1) + 1$, then we are done. Hence we assume that $\pi(2) > \pi(1) + 1$.

Consider the smallest integer k such that $\pi(\{2, \ldots, k\})$ contains the interval $J = \{\pi(1) + 1, \ldots, \pi(2)\}$. Observe that $\pi(1) < \pi(k) < \pi(2)$. If the image $\pi(\{2, \ldots, k\})$ is exactly the interval J, then we are done. Otherwise, choose an element $2 < \ell < k$ whose image is not in J.

Observe that $\pi(\ell) < \pi(1)$ is not possible: otherwise we have $1 < 2 < \ell < k$ and $\pi(\ell) < \pi(1) < \pi(k) < \pi(2)$. Therefore we have $\pi(\ell) > \pi(2)$. This also shows that all elements of $\pi(\{2, \ldots, k\})$ is greater than $\pi(1)$. Let $\pi(\ell)$ be the largest among them.

If $\pi(\{2, \ldots, k\})$ is an interval, then we are done. Otherwise, there is at least one "gap" in the image $\pi(\{2, \ldots, k\})$, which must be greater than $\pi(2)$.

Therefore, there exists $m > k$ such that $\pi(2) < \pi(m) < \pi(\ell)$. We thus find $2 < \ell < k < m$ where $\pi(k) < \pi(2) < \pi(m) < \pi(\ell)$. Contradiction. □

Corollary 11.6. *Let $n \geq 3$. Suppose π is a separable permutation. Then there are two consecutive integers whose images are also consecutive.*

Proof. The proof is obvious by induction. □

Theorem 11.7. *A permutation is a polynomial interchange if and only if it is separable.*

Proof. It remains to prove that any separable permutation is a polynomial interchange. By the corollary, there exists $\{i, i+1\}$ such that $\{\pi(i), \pi(i+1)\}$ is also consecutive. Imagine "collapsing" $\{i, i+1\}$ and $\{\pi(i), \pi(i+1)\}$ into single points, we produce a permutation π' on $n-1$ objects which is separable, and therefore a polynomial interchange by induction. It follows that there are $n-1$ polynomials $P_1, \ldots, P_{n-1}$ which interact at the origin according to π'. The only thing remains is to split the i-th polynomial in order to produce n polynomials $P_1, \ldots, P_{i-1}, P_i', P_i'', P_{i+1}, \ldots, P_n$ which interact according to π. This can be done by setting

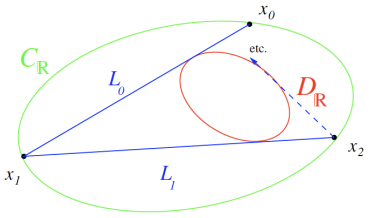
$$P_i'(x) = P_i(x) \quad \text{and} \quad P_i''(x) = P_i(x) + (-x)^N$$

for sufficiently large N, even or odd, according to whether $\pi(i+1) > \pi(i)$ or $\pi(i+1) < \pi(i)$. □

Exercise. Let $a(n)$ be the number of separable permutations of $\{1, \ldots, n\}$. Can you find an explicit formula of $a(n)$ (or equivalently, find the generating function $G(t) = \sum a(n)t^n$)?

Lecture 14

11.3. **The Poncelet problem.** Consider the following problem.



Given any two ellipses, one is contained in the other. Start with any point x_0 on the outer boundary, and draw a line L_0 tangent to the inner one (there are

two choices), continue until it hits the outer ellipse again. Begin again at this new point by drawing the other line through it and tangent to the inner ellipse. Iterating this construction, we may ask whether it ever closes up, i.e. returns to the starting point x_0. A surprising fact is that the answer is *independent* of the choice of the starting point x_0!

Theorem 11.8. *Whether or not the trajectory closed up, is independent of the choice of x_0 and L_0.*

Perhaps even more surprising is that the proof of this theorem uses some non-trivial results of *complex Riemann surfaces*.

Let us denote the boundary of the outer ellipse by C, and the inner one by D. It turns out that the result has not much to do with they are ellipses; the crucial thing here is that they can be defined by the zero set of certain degree two polynomials. One can generalize the situation by considering the complex projective plane $\mathbb{CP}^2$, and two zero sets

$$C = \{F_C(x, y, z) = 0\} \qquad \text{and } D = \{F_D(x, y, z) = 0\}$$

where F_C, F_D are of degree two. Consider the *incidence correspondence*

$$E = \{(x, L) \mid x \in L\} \subseteq C \times \check{D}$$

where $\check{D} \subseteq \check{\mathbb{P}}^2$ is the *dual curve* of D consisting of lines tangent to D. There are two *involutions* of E:

- $\iota_1 \colon E \to E$: each $L \in \check{D}$ meets C in two points (counted with multiplicities), and swapping these two points yield ι_1.
- $\iota_2 \colon E \to E$: each $x \in C$ has two lines tangent to D ("counted with multiplicities"), and swapping these two lines yield ι_2.

Consider the projection

$$\pi \colon E \to C \cong \mathbb{P}^1; \qquad (x, L) \mapsto x.$$

- The projection π has degree 2: there are two lines tangent to D through a general point $x \in C$.
- It has 4 *ramification* points: namely, the points of E fixed by the involution ι_2: a point on C is ramify if and only if it lies in the intersection $C \cap D$, and $|C \cap D| = 4$ by *Bézout's theorem*.

The *Riemann-Hurwitz formula* says that

$$\chi(E) = \deg(\pi) \cdot \chi(C) - r(\pi) = 2 \cdot 2 - 4 = 0.$$

Therefore, E is isomorphic to an *elliptic curve* (i.e. a two dimension torus, topologically) $\mathbb{C}/\Lambda$. Any automorphism is of the form $z \mapsto az + b$ for some $a, b \in \mathbb{C}$, and squaring it gives

$$z \mapsto az + b \mapsto a(az + b) + b = a^2 z + (ab + b).$$

It is an involution if and only if

- $a = 1$ and $b \in \Lambda/2$, or
- $a = -1$.

Now consider ι_1 and ι_2. Since they both are involutions with fixed points, they belong to the second case:

$$\iota_1(z) = -z + b_1 \quad \text{and} \quad \iota_2(z) = -z + b_2.$$

The process in the Poncelet problem is nothing but the composition $\iota_2 \circ \iota_1$, which therefore is a translation by $\beta \coloneqq b_2 - b_1$ in $\mathbb{C}/\Lambda$. Hence, the trajectory can closed up if and only if there exists $n \in \mathbb{N}$ such that $\beta n \in \Lambda$, which is independent of the starting point (x_0, L_0).

11.4. **Dilogarithm function and its five-term relation.** (Reference: [19])
The dilogarithm function is defined by the power series

$$\mathrm{Li}_2(z) = \sum_{n=1}^{\infty} \frac{z^n}{n^2} \qquad \text{for } |z| < 1.$$

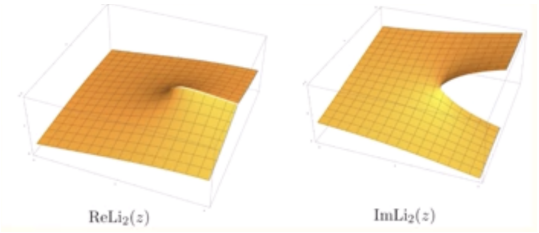
The definition (and the name) come from the analogy with the Taylor series of the ordinary logarithm around 1

$$-\log(1 - z) = \sum_{n=1}^{\infty} \frac{z^n}{n} \qquad \text{for } |z| < 1.$$

On $|z| < 1$ we have

$$\frac{\mathrm{d}}{\mathrm{d}z}\mathrm{Li}_2(z) = -\frac{\log(1 - z)}{z}, \quad \text{hence} \quad \mathrm{Li}_2(z) = -\int_0^z \frac{\log(1 - u)}{u}\,\mathrm{d}u.$$

Using the integral expression, one can holomorphically extend $\mathrm{Li}_2(z)$ to the domain $\mathbb{C}\backslash(1, \infty)$. Below are plots of the real and imaginary parts of $\mathrm{Li}_2(z)$: the real part can be extended continuous (but not analytically) to $(1, \infty)$, but the imaginary part cannot: it jumps by $2\pi i \log|z|$ as z crosses the cut.



Remark 11.9. Most functions have either no exactly computable special values (e.g. Bessel functions), or else a countable, easily describable set of them. For instance, for the Gamma function we have

$$\Gamma(n) = (n-1)!, \qquad \Gamma\left(n+\frac{1}{2}\right) = \frac{(2n)!}{4^n n!}\sqrt{\pi},$$

and for the Riemann zeta function we have

$$\zeta(2) = \frac{\pi^2}{6}, \quad \zeta(4) = \frac{\pi^4}{90}, \quad \zeta(6) = \frac{\pi^6}{945}, \quad \dots,$$

$$\zeta(0) = -\frac{1}{2}, \quad \zeta(-2) = \zeta(-4) = \cdots = 0,$$

$$\zeta(-1) = -\frac{1}{12}, \quad \zeta(-3) = \frac{1}{120}, \quad \zeta(-5) = -\frac{1}{252}, \quad \dots.$$

On the other hand, as far as anyone knows, there are exactly 8 values of z for which z and $\mathrm{Li}_2(z)$ can both be given in closed form:

$$\mathrm{Li}_2(0) = 0, \quad \mathrm{Li}_2(1) = \frac{\pi^2}{6}, \quad \mathrm{Li}_2(-1) = -\frac{\pi^2}{12}, \quad \mathrm{Li}_2\left(\frac{1}{2}\right) = \frac{\pi^2}{12} - \frac{1}{2}\log^2(2),$$

$$\mathrm{Li}_2(-\varphi) = -\frac{\pi^2}{10} - \log^2(\varphi), \quad \mathrm{Li}_2\left(-\frac{1}{\varphi}\right) = -\frac{\pi^2}{15} + \frac{1}{2}\log^2(\varphi),$$

$$\mathrm{Li}_2\left(\frac{1}{\varphi^2}\right) = \frac{\pi^2}{15} - \log^2(\varphi), \quad \mathrm{Li}_2\left(\frac{1}{\varphi}\right) = \frac{\pi^2}{10} - \log^2(\varphi).$$

In contrast to its special values, the dilogarithm function enjoys a lot of functional equations. To begin with, there are two reflection properties

$$\mathrm{Li}_2\left(\frac{1}{z}\right) = -\mathrm{Li}_2(z) - \frac{\pi^2}{6} - \frac{1}{2}\log^2(-z),$$

$$\mathrm{Li}_2(1-z) = -\mathrm{Li}_2(z) + \frac{\pi^2}{6} - \log(z)\log(1-z).$$

Therefore, the six functions

$$\mathrm{Li}_2(z),\ -\mathrm{Li}_2\left(\frac{1}{z}\right),\ -\mathrm{Li}_2(1-z),\ \mathrm{Li}_2\left(\frac{1}{1-z}\right),\ \mathrm{Li}_2\left(\frac{z-1}{z}\right),\ -\mathrm{Li}_2\left(\frac{z}{z-1}\right)$$

are equal modulo elementary functions. Next, there is the two-variable, five-term relation

$$\begin{aligned}&\mathrm{Li}_2(x)+\mathrm{Li}_2(y)+\mathrm{Li}_2\left(\frac{1-x}{1-xy}\right)+\mathrm{Li}_2(1-xy)+\mathrm{Li}_2\left(\frac{1-y}{1-xy}\right)\\ =&\frac{\pi^2}{6}-\log(x)\log(1-x)-\log(y)\log(1-y)+\log\left(\frac{1-x}{1-xy}\right)\log\left(\frac{1-y}{1-xy}\right).\end{aligned}$$

One can simplify the above relation by considering Bloch–Wigner function $D(z)$, which is defined to be

$$D(z)=\mathrm{Im}(\mathrm{Li}_2(z))+\arg(1-z)\log|z|$$

where arg denotes the branch of the argument lying between 0 and 2π. Recall that $\mathrm{Li}_2(z)$ jumps by $2\pi i\log|z|$ as z crosses the cut $(1,\infty)$, so the additional term $\arg(1-z)\log|z|$ makes the Bloch–Wigner function $D(z)$ a continuous function on $\mathbb{C}$.

One can show that all of the functional equations above lose the elementary correction terms when expressed in terms of $D(z)$. In particular,

$$D(z)=D\left(1-\frac{1}{z}\right)=D\left(\frac{1}{1-z}\right)=-D(\frac{1}{z})=-D(1-z)=-D\left(\frac{-z}{1-z}\right)$$

and the five-term relation

$$D(x)+D(y)+D\left(\frac{1-x}{1-xy}\right)+D(1-xy)+D\left(\frac{1-y}{1-xy}\right)=0.$$

Remark 11.10. Starting from $a_1=1-xy$ and $a_2=x$, if we define a sequence $\{a_n\}$ recursively via

$$a_3=\frac{1-a_2}{a_1},$$

then we have

$$a_3=\frac{1-x}{1-xy},\quad a_4=\frac{1-y}{1-xy},\quad a_5=y,\quad a_6=a_1,\quad a_7=a_2.$$

This is called the *pentagon relation*, and is the most basic example of the structure of *cluster algebras*.

The five-term relation of $D(z)$ becomes even cleaner if we think of D as being a function not of a single complex variable but of the *cross-ratio* of four complex numbers:

$$\widetilde{D}(z_0, z_1, z_2, z_3) \coloneqq D\left(\frac{z_0 - z_2}{z_0 - z_3} \cdot \frac{z_1 - z_3}{z_1 - z_2}\right).$$

The six-fold symmetry says that $\widetilde{D}$ is invariant under even permutations of its four variables, and anti-invariant under odd permutations. The five-term relation then becomes

$$\sum_{i=0}^{4} (-1)^i \widetilde{D}(z_0, \ldots, \hat{z_i}, \ldots, z_4) = 0.$$

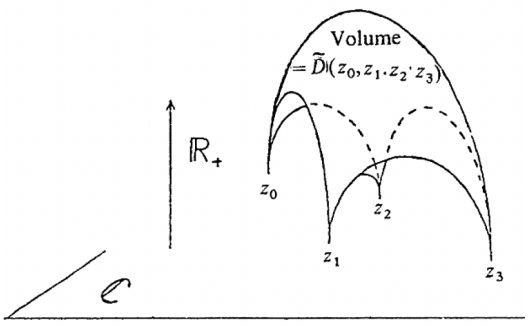
It turns out that the dilogarithm occurs naturally as measurement of volumes in *hyperbolic geometry*. Consider the hyperbolic 3-space

$$\mathbb{H}^3 = \{(z, w) \in \mathbb{C} \times \mathbb{R} \mid w > 0\}$$

equipped with the hyperbolic metric

$$\mathrm{d}s = \frac{\sqrt{\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}w^2}}{w}.$$

The geodesics on $\mathbb{H}^3$ are either vertical lines or semicircles in vertical planes with endpoints in $\mathbb{C} \times \{0\}$; and the geodesic planes are either vertical planes or hemispheres with boundary in $\mathbb{C} \times \{0\}$.



An *ideal tetrahedron* is a tetrahedron whose vertices are all in $\mathbb{C} \times \{0\}$. It is proved by Lobachevsky that the *volume* of the ideal tetrahedron with vertices

z_0, z_1, z_2, z_3 is precisely given by $D(z)$!

$$\text{Volume} = \widetilde{D}(z_0, z_1.z_2, z_3).$$

The (anti)symmetry property of $\widetilde{D}$ under permutations of the z_i is obvious from this geometric interpretation, since renumbering the vertices leaves the tetrahedron unchanged but may reverse the orientation. The five-term relation is also follows immediately from the geometric interpretation: the five tetrahedron spanned by four at a time of $z_0, \ldots, z_4$, counted positively or negatively as in the formula, add up to zero.

11.5. **Quantum dilogarithm, stability conditions, and wall-crossing formula.** (Reference: [11])

Consider the series

$$\mathbb{E}(x) = 1 + \frac{q^{1/2}}{q-1}x + \cdots + \frac{q^{n^2/2}}{(q^n-1)(q^n-q)\cdots(q^n-q^{n-1})}x^n + \cdots$$

which is an element of $\mathbb{Q}(q^{1/2})[\![x]\!]$. The following properties explain why it is considered as a *quantum analogue* of the dilogarithm function.

- As $q \to 1^-$, we have

$$\mathbb{E}(x) \sim \exp\left(-\frac{\mathrm{Li}_2(-x)}{\log q}\right).$$

- For two variables x_1 and x_2 which q-commute in the sense that $x_1x_2 = qx_2x_1$, we have the pentagon relation

$$\mathbb{E}(x_1)\mathbb{E}(x_2) = \mathbb{E}(x_2)\mathbb{E}(q^{-1/2}x_1x_2)\mathbb{E}(x_1).$$

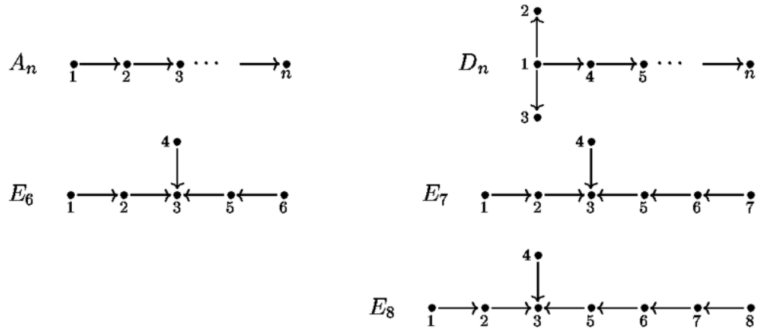
In fact, it can be proved that the classical five-term relation of $\mathrm{Li}_2(x)$ follows from the pentagon relation of the quantum dilogarithm.

The equation is the simplest example of the *wall-crossing formula* for *stability conditions*.

Let Q be one of the following *Dynkin quivers*.

We will associate a whole family of quantum dilogarithm products which are all equal for each Q. The pentagon relation corresponds to the special case where Q is the A_2 quiver.

Definition 11.11. A *representation* of a quiver Q (over a field, say $\mathbb{C}$) consists of the following data:



- a finite dimensional vector space assigned to each vertex $\{V(x)\}_{x:\text{vertex}}$;
- a linear map assigned to each arrow $\{T(a)\colon V(x) \to V(y)\}_{x \xrightarrow{a} y}$.

One can easily define the notion of *morphisms* between representations, and obtains an *abelian category* $\mathrm{Rep}(Q)$ of representations of a quiver Q. A representation is called *simple* if it does not have any nontrivial sub-representations; a representation is called *indecomposable* if it is not isomorphic to the direct sum of two non-trivial representations. It is clear that simple representations are indecomposable, but the converse is not true.

Example 11.12. Let Q be the loop quiver: it has a single vertex, and has a single edge that starts and ends at the same vertex. Then a representation of Q is simply a linear endomorphism $T\colon V \to V$. Two representations T_1, T_2 are isomorphic if they are *similar*, i.e. there exists an invertible linear map $S\colon V \to V$ such that $T_1 = S \circ T_2 \circ S^{-1}$. Therefore,

- the isomorphism classes of representations bijectively correspond to Jordan canonical forms;
- the isomorphism classes of indecomposable representations are the ones that correspond to maps having a single Jordan block

$$\begin{pmatrix} \lambda & 1 & & & 0 \\ & \lambda & 1 & & \\ & & \cdots & \cdots & \\ & & & \cdots & 1 \\ 0 & & & & \lambda \end{pmatrix};$$

- the isomorphism classes of simple representations are all one-dimensional $T\colon \mathbb{C} \to \mathbb{C}$, $z \mapsto \lambda z$, where $\lambda \neq 0$ (why?).

Example 11.13. Let Q be the A_2 quiver. Then

- the isomorphism classes of representations bijectively correspond to matrices of the form
$$\begin{pmatrix} 1 & & & & & 0 \\ & 1 & & & & \\ & & \cdots & & & \\ & & & 1 & & \\ 0 & & & & 0 & 0 \end{pmatrix};$$
- the isomorphism classes of indecomposable representations are
$$\mathbb{C} \to 0, \qquad 0 \to \mathbb{C}, \qquad \mathbb{C} \xrightarrow{\cong} \mathbb{C};$$
- the isomorphism classes of simple representations are
$$\mathbb{C} \to 0, \qquad 0 \to \mathbb{C}.$$

Theorem 11.14 (Gabriel)**.** *The category Rep(Q) has finitely many isomorphism classes of indecomposable representations if and only if Q is one of the Dynkin quivers. Moreover, the isomorphism classes bijectively correspond to the* positive roots *of the* root systems *of the Dynkin diagrams.*

Definition 11.15. A *stability function* on an abelian category $\mathcal{A}$ is a group homomorphism
$$Z \colon K_0(\mathcal{A}) \to \mathbb{C}$$
such that $Z(E) \in \mathbb{H} \cup \mathbb{R}_{<0}$ for all nonzero object $E \in \mathcal{A}$. The *phase* of E is defined to be
$$\phi(E) \coloneqq \frac{1}{\pi} \arg(Z(E)) \in (0, 1].$$

Here, $K_0(\mathcal{A})$ denotes the *Grothendieck group* of $\mathcal{A}$, which is the free abelian group generated by objects of $\mathcal{A}$, modulo the relation that $[A] + [C] = [B]$ for any exact sequence $0 \to A \to B \to C \to 0$. A simpler way to put this is, if there are three objects $E_1, E_2, E_3 \in \mathcal{A}$ satisfies $E_3 \cong E_2/E_1$, then the stability function has to satisfy $Z(E_2) = Z(E_1) + Z(E_3)$.

With a chosen stability function, one says a nonzero object E is *stable* (resp. *semistable*) if for each nonzero proper sub-object F of E, we have $\phi(F) < \phi(E)$ (resp. $\phi(F) \leq \phi(E)$). Note that any simple object is automatically stable as it has no non-trivial sub-objects. Also note that any stable object is indecomposable (why?).

Proposition 11.16 (King)**.** *Let $\mathcal{A} = Rep(Q)$ be the category of representations of a quiver Q, and let $Z\colon K_0(\mathcal{A}) \to \mathbb{C}$ be a stability function. Denote $\mathcal{A}_\phi$ the full subcategory of $\mathcal{A}$ whose objects are the zero object and the semistable objects of phase ϕ.*

- *The subcategory $\mathcal{A}_\phi$ is stable under extensions, kernels, and cokernels in $\mathcal{A}$. In particular, it is an abelian sub-category of $\mathcal{A}$. Its simple objects are precisely the stable objects of phase ϕ.*
- *(Harder–Narasimhan property) Each nonzero object E admits a (unique) filtration*
$$0 = E_0 \subseteq E_1 \subseteq E_2 \subseteq \cdots \subseteq E_n = E$$
where each E_i/E_{i-1} is semistable and $\phi(E_1) > \phi(E_2/E_1) > \cdots \phi(E_n/E_{n-1})$.

In other words, the semistable objects $\{\mathcal{A}_\phi\}$ give a nice "refinement" of the category $\mathcal{A}$, in the sense that every object can be uniquely built from the semistable objects by the Harder–Narasimhan property.

It is important to note that the notion of stability of objects depends on the stability function: as one varies the stability function $Z\colon K_0(\mathcal{A}) \to \mathbb{C}$, a stable object could become unstable.

Remark 11.17. In the *homological mirror symmetry* of Calabi–Yau manifolds, there are certain dualities between *complex geometry* and *symplectic geometry*. For a complex-geometric category $\mathcal{D}$, the input of symplectic geometry precisely gives a stability condition on $\mathcal{D}$, and vice versa.

Example 11.18. Let us consider the example of the A_2 quiver. The Grothendieck group in this case is $K_0(\mathcal{A}) \cong \mathbb{Z}^{\oplus 2}$, which is given by the dimensions of the representations. Therefore, to give a stability function is equivalent to give two complex numbers $z_1, z_2 \in \mathbb{H} \cup \mathbb{R}_{<0}$, where we assign
$$Z_{z_1,z_2}(V \to W) \coloneqq z_1 \dim(V) + z_2 \dim(W).$$
Let us analysis how the set of stable objects varies as z_1, z_2 are changing. First of all, any stable object is indecomposable, therefore is isomorphic to one of the following:
$$\mathbb{C} \xrightarrow{S_1} 0, \qquad 0 \xrightarrow{S_2} \mathbb{C}, \qquad \mathbb{C} \xrightarrow{E=\mathrm{id}} \mathbb{C};$$
moreover, the first two are simple objects, thus are always stable. So our task is to determine for which z_1, z_2 the object E is stable.

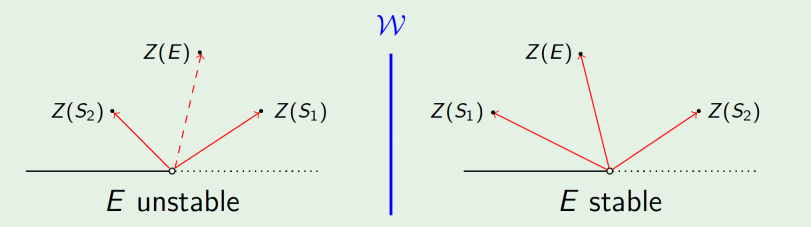
There is a short exact sequence

$$0 \to S_2 \to E \to S_1 \to 0;$$

i.e. S_2 is a subobject of E and $E/S_2 \cong S_1$. (Is S_1 a subobject of E?) The object E is stable with respect to Z_{z_1,z_2} if and only if

$$\arg(Z_{z_1,z_2}(S_2)) < \arg(Z_{z_1,z_2}(E)), \quad \text{or equivalently} \quad \arg(z_1) > \arg(z_2).$$

This is the simplest example of the *wall-crossing phenomenon* in the space of stability conditions.



One can draw a *wall* in the space of stability conditions (in this case, $(\mathbb{H} \cup \mathbb{R}_{<0})^2$), where the set of stable objects remain unchanged in the chambers, but will change as (z_1, z_2) crosses a wall and move to another chamber. For more general quivers (or general abelian categories), there could be infinitely many walls in the space of stability conditions; the walls can even be dense in certain regions.

We will associate a product of quantum dilogarithms with each stability function. First, we need to define the algebra in which these products will be computed. Let Q be a quiver with n vertices. Define an anti-symmetric pairing

$$\lambda \colon \mathbb{Z}^n \times \mathbb{Z}^n \to \mathbb{Z}$$

by

$$\lambda(e_i, e_j) \coloneqq \#\{\text{arrows from } i \text{ to } j\} - \#\{\text{arrows from } j \text{ to } i\}$$

where $\{e_1, \ldots, e_n\}$ is the standard basis of $\mathbb{Z}^n$, and then extend the pairing linearly. The *quantum affine space* $\mathbb{A}_Q$ is the $\mathbb{Q}(q^{1/2})$-algebra generated by the variables x^α, $\alpha \in \mathbb{N}^n$, subject to the relations

$$x^\alpha x^\beta = q^{\lambda(\alpha,\beta)/2} x^{\alpha+\beta}.$$

It is also generated by the variables $x_i = x^{e_i}$, $1 \le i \le n$, subject to the relations

$$x_i x_j = q^{\lambda(e_i,e_j)} x_j x_i.$$

For instance, in the A_2 quiver, we have

$$x_1x_2 = q^{1/2}x^{(1,1)} = qx_2x_1.$$

Theorem 11.19 (wall-crossing formula)**.** *Let Q be a Dynkin quiver, and $Z\colon K_0(Rep(Q)) \to \mathbb{C}$ be a stability function. Then the product*

$$\mathbb{E}_{Q,Z} \coloneqq \prod_{E:stable}^{\curvearrowright} \mathbb{E}(x^{\underline{\dim}(E)})$$

where the factors are in the order of decreasing phases, is independent of the choice of Z.

In particular, the wall-crossing formula for $Q = A_2$ quiver recovers the pentagon identity of quantum dilogarithm.

Lecture 15

Let us sketch the proof of the wall-crossing formula, which is proved by Reineke. First, it is a fact about such formal power series in q that, in order to show $E_{Q,Z}$ is independent of Z, it suffices to prove it under specializing q to prime powers p^m.

Let k be a finite field with q elements, and let $\mathcal{A}$ be the abelian category of k-representations of the quiver Q. Let us introduce the notion of *Hall algebra* $\mathcal{H}(\mathcal{A})$: its elements are formal series with rational coefficients

$$\sum_{[E]\in[\mathcal{A}]} a_E[E]$$

where the sum is taken over the set of isomorphism classes of objects of $\mathcal{A}$. The product of $\mathcal{H}(\mathcal{A})$ is the bilinear map determined by

$$[E_1]\cdot[E_2] = \sum c^F_{E_1,E_2}[F]$$

where $c^F_{E_1,E_2}$ is the number of submodules E_1' of F which are isomorphic to E_1 and such that $F/E_1' \cong E_2$.

Recall that for every stability function Z, one has the corresponding subcategories $\{\mathcal{A}^Z_\phi\}$ with objects given by Z-semistable objects of phase ϕ. The Harder–Narasimhan property can be translated into a Hall algebra identity:

$$\sum_{[E]\in[\mathcal{A}]}[E] = \prod_{\phi:\text{decreasing}}^{\curvearrowright} \sum_{[F]\in[\mathcal{A}^Z_\phi]}[F].$$

In particular, the expression on the right hand side is independent of Z. It remains to translate it in terms of the quantum dilogarithm.

One can define the *evaluation map*

$$\mathrm{ev}\colon \mathcal{H}(\mathcal{A}) \to \mathbb{A}_Q = \mathbb{Q}(q^{1/2})[x^\alpha]$$

by sending

$$[E] \mapsto q^{\langle \underline{\dim}(E), \underline{\dim}(E)\rangle/2} \frac{x^{\underline{\dim}(E)}}{|\operatorname{Aut}(E)|}.$$

An important property of the evaluation map is that it is an *algebra* homomorphism. (Here we regard $q = p^m$ and E as a $\mathbb{F}_q$-representation.) This follows from a formula of Riedtmann that the coefficients $c^F_{E_1,E_2}$ is given by

$$c^F_{E_1,E_2} = q^{-2\dim \operatorname{Hom}(E_1,E_2)} \frac{|\operatorname{Aut}(F)|}{|\operatorname{Aut}(E_1)||\operatorname{Aut}(E_2)|} |\operatorname{Ext}^1(E_1,E_2)_F|.$$

Exercise. Let $k = \mathbb{F}_q$, and let V be an n-dimensional vector space over k. Prove that

$$|\operatorname{Aut}(V)| = (q^n - 1)(q^n - q)\cdots(q^n - q^{n-1}).$$

If E is a stable object, then $\operatorname{Aut}(E) \cong k^\times$, and one can also show that

$$|\operatorname{Aut}(E^{\oplus n})| = (q^n - 1)(q^n - q)\cdots(q^n - q^{n-1}).$$

Thus

$$\mathrm{ev}\left(\sum_n [E^{\oplus n}]\right) = \mathbb{E}(x^{\underline{\dim}(E)})$$

is precisely the quantum dilogarithm! Using the fact that every semistable object is the sum of a unique stable object, we find that

$$\mathrm{ev}\left(\prod_{\phi:\text{decreasing}}^{\curvearrowright} \sum_{[F]\in[\mathcal{A}^Z_\phi]} [F]\right) = \prod_{E:\text{stable}}^{\curvearrowright} \mathbb{E}(x^{\underline{\dim}(E)}).$$

is independent of the choice of Z.

Another formulation of the wall-crossing formula is by studying certain cluster-like transformations on a *twisted torus*. Define a ring $\oplus_{\alpha\in\mathbb{Z}^n}\mathbb{C}\cdot x^\alpha$, with multiplication given by

$$x^\alpha \cdot x^\beta := (-1)^{-\lambda(\alpha,\beta)} x^{\alpha+\beta}.$$

For instance, for A_2 quiver we have

$$x_1x_2 = x^{e_1}x^{e_2} = -x^{(1,1)} = -x_2x_1.$$

(This ring can be considered as the ring of functions of certain twisted torus.) For each stability function, we can assign a *change of variables* $S_Z(\ell)$ for each *ray* $\ell \subseteq \mathbb{H} \cup \mathbb{R}_{<0}$ as follows:

$$x^\alpha \mapsto x^\alpha \prod_{Z(\beta)\in\ell} (1-x^\beta)^{-\Omega_Z(\beta)\lambda(\alpha,\beta)}$$

where $\Omega_Z(\beta)$ counts the number of isomorphism classes of stable objects (with respect to Z) of class $\beta \in \mathbb{Z}^n$. Observe that $S_Z(\ell)$ is a non-trivial change of variables only if ℓ supports some stable objects.

Let us consider again the A_2 example. Let Z be a stability function, and let ℓ_1, ℓ_2, ℓ_E be the rays that contain $Z(S_1), Z(S_2), Z(E)$, respectively. Then

$$S(\ell_1)\colon x_1 \mapsto x_1, \quad x_2 \mapsto x_2(1-x_1),$$

$$S(\ell_2)\colon x_1 \mapsto x_1(1-x_2)^{-1}, \quad x_2 \mapsto x_2,$$

and when E is stable, we have

$$S(\ell_E)\colon x_1 \mapsto x_1(1+x_1x_2)^{-1}, \quad x_2 \mapsto x_2(1+x_1x_2).$$

The *wall-crossing formula* states that the clockwise product

$$\prod_{\ell}^{\curvearrowright} S(\ell)$$

is independent of the choice of Z. In the case of A_2 quiver, we have the pentagon relation

$$S(\ell_1) \circ S(\ell_2) = S(\ell_2) \circ S(\ell_E) \circ S(\ell_1).$$

By formulating the wall-crossing formula in this way, we will see that it is actually an infinite-dimensional analogue of a more classical phenomenon in the study of ordinary differential equations with *irregular singularities*, called the *Stokes phenomenon*.

11.6. **Borel summation and resurgence.** Let us consider the following *Euler's equation*

$$x^2 \frac{\mathrm{d}f(x)}{\mathrm{d}x} = x - f(x).$$

Our aim is to find all possible solutions of this equation on the real line $x \in \mathbb{R}$. We can begin with searching for a solution in the form of a power series

$$f = a_0 + a_1 x + a_2 x^2 + \cdots .$$

By comparing the coefficients of both sides of the equation, we find

$$a_0 = 0, \quad a_1 = 1, \quad a_{k+1} = (-1)^k k! \quad \text{for } k \geq 1,$$

and therefore

$$f = \sum_{k=0}^{\infty} (-1)^k k! x^{k+1}.$$

The answer is nice and simple, but it has a couple of major problems: First, applying the ratio test, we find

$$\lim_{k\to\infty} \frac{(k+1)! x^{k+2}}{k! x^{k+1}} = \infty$$

for any nonzero x. Hence the series only converges at $x = 0$, which is rather unsettling. Second, it is a basic fact that the solutions of a first-order differential equation like ours are not unique, but should instead depend on a parameter, corresponding to the choice of an initial condition, say at $x = 0$. But here we found only a single solution.

To find other solutions, one can consider the associated *homogeneous* equation

$$x^2 \frac{\mathrm{d}g}{\mathrm{d}x} = -g,$$

which has solutions

$$g = a e^{\frac{1}{x}}, \quad \text{where } a \text{ is an arbitrary constant.}$$

Hence the general solution should be of the form

$$f = \sum_{k=0}^{\infty} (-1)^k k! x^{k+1} + a e^{\frac{1}{x}}.$$

These two major problems arose because at the point $x = 0$, where we have centered our expansion, the equation has a *singularity*. This becomes clearer if we rewrite it as

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \frac{1}{x} - \frac{f}{x^2}.$$

In this form, it is evident that the equation has a pole at $x = 0$, so the usual existence and uniqueness theorem for analytic solutions of the equation does not hold there. Indeed, the correction term $e^{1/x}$ that we missed has an *essential* singularity at $x = 0$, so there is no way we could have found it by considering power series.

How to resolve these issues? One possible way is to consider expansion about a different point $x_0 \neq 0$, where the equation is analytic. Then the problems above would not appear, but the series would unfortunately not have a simple expression as above. More importantly, such expansions would not help us understand the most interesting aspect of the equation, which is the interesting behavior of the solutions near the singular point $x = 0$.

We thus wish to find a way to somehow "resum" the divergent series and obtain an actual solution of the equation. Remarkably, just by "resumming" this single divergent series, we will in fact recover the other term $e^{1/x}$ that we thought were missing.

There are many approaches to resumming divergent series, such as Abel summation, Césaro summation, ζ-function regularization, etc. The method that is appropriate for our setting is the *Borel summation*. One can show that

$$x^{k+1} = \frac{1}{k!} \int_0^\infty t^k e^{-t/x} \, \mathrm{d}t.$$

Let us substitute this into our series solution.

$$\begin{aligned} f(x) &= \sum_{k=0}^{\infty} (-1)^k k! x^{k+1} \\ &= \sum_{k=0}^{\infty} (-1)^k \int_0^{\infty} t^k e^{-t/x} \, \mathrm{d}t \\ \text{“}&= \text{”} \int_0^{\infty} \left(\sum_{k=0}^{\infty} (-1)^k t^k \right) e^{-t/x} \, \mathrm{d}t \\ \text{“}&= \text{”} \int_0^{\infty} \frac{1}{1+t} e^{-t/x} \, \mathrm{d}t. \end{aligned}$$

This calculation is obviously illegal: one cannot interchange the order of summation and integration since our series diverges; and the formula

$$\sum_{k=0}^{\infty} (-1)^k t^k = \frac{1}{1+t}$$

holds only when $|t| < 1$. The second issue is somehow less serious, because the theory of analytic continuation implies that there is only one natural way to extend the function to the region $|t| \geq 1$.

On the other hand, observe that when $x > 0$, the integrand

$$\frac{e^{-t/x}}{1+t}$$

has exponential decay as $t \to \infty$. Hence the expression

$$f(x) = \int_0^{\infty} \frac{e^{-t/x}}{1+t} \, \mathrm{d}t.$$

actually defines a nice smooth function on $x > 0$. This function is the *Borel sum* of our divergent series. We can verify that this function does satisfy the

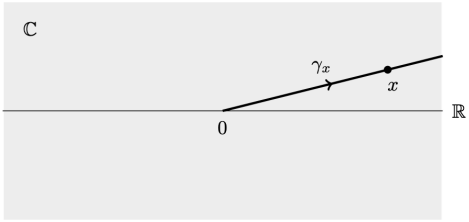
differential equation:

$$\begin{aligned}\frac{\mathrm{d}f}{\mathrm{d}x} &= \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{e^{-t/x}}{1+t}\right)\mathrm{d}t \\ &= \int_0^\infty \frac{e^{-t/x}}{1+t}\cdot\frac{t}{x^2}\,\mathrm{d}t \\ &= \frac{1}{x^2}\int_0^\infty e^{-t/x}\left(1-\frac{1}{1+t}\right)\mathrm{d}t \\ &= \frac{1}{x}-\frac{f}{x^2}.\end{aligned}$$

How about the solutions for $x < 0$? In this case, the integral expression above diverges, so the solution is no longer valid. At this point, it is helpful to expand our viewpoint and allow x to be of *complex-valued*. Observe that for the integral

$$f(x) = \int_0^\infty \frac{e^{-t/x}}{1+t}\,\mathrm{d}t$$

to converge, all that is necessary is that $\mathrm{Re}(x) > 0$, so this expression makes sense for any value of x in the right half-plane, giving a holomorphic function. So we should try to use analytic continuation to define the function in the region $\mathrm{Re}(x) < 0$. To do so, it is useful to modify our contour of integration.

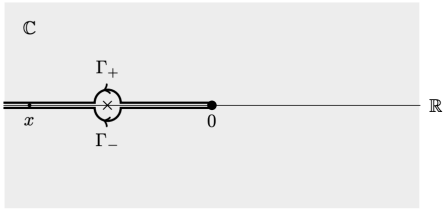


Let γ_x denotes the ray from 0 to $\infty \in \mathbb{CP}^1$ in the direction of x, then

$$f(x) \coloneqq \int_{\gamma_x} \frac{e^{-t/x}}{1+t}\,\mathrm{d}t$$

would converge, so long as $x \notin \mathbb{R}_{\leq 0}$. Due to the path-independence of contour integrals, this new definition of $f(x)$ is the same as the previous one on $\mathrm{Re}(x) > 0$, and so we obtain a holomorphic extension of the solution $f(x)$ to all values of $x \in \mathbb{C}\backslash\mathbb{R}_{<0}$.

One can attempt to extend $f(x)$ to points $x \in \mathbb{R}_{<0}$ by choosing contours avoiding $t = -1$ as follows.



Depending on whether we pass above or below the pole, one gets two possible values

$$f(x)^+ = \int_{\Gamma_+} \frac{e^{-t/x}}{1+t}\,\mathrm{d}t \quad \text{and} \quad f(x)^- = \int_{\Gamma_-} \frac{e^{-t/x}}{1+t}\,\mathrm{d}t.$$

These two values do not coincide:

$$f(x)^+ - f(x)^- = 2\pi i \cdot \mathrm{Res}_{t=-1}\left(\frac{e^{-t/x}}{1+t}\right) = 2\pi i \cdot e^{1/x}.$$

This behavior, in which the function defined by a divergent series jumps along a ray, is an example of the *Stokes phenomenon*. The crucial missing term $e^{1/x}$ has "resurged" from crossing the branch cut.

11.7. **Stokes phenomenon of irregular singularities.** Let us consider a slightly more general system of differential equations with an irregular singularity. Let $G = \mathrm{GL}(n, \mathbb{C})$ and $\mathfrak{g} = \mathfrak{gl}(n, \mathbb{C}) = \mathfrak{h} \oplus (\oplus_{\alpha \in \Phi} \mathfrak{g}_\alpha)$, where $\mathfrak{h}$ consists of diagonal matrices, and

$$\Phi = \{\alpha_{ij} = e_i^* - e_j^* | 1 \leq i \neq j \leq n\} \subseteq \mathfrak{h}^*$$

so $\oplus_{\alpha \in \Phi} \mathfrak{g}_\alpha$ corresponds to off-diagonal matrices. Consider the following meromorphic connection of the trivial vector bundle of rank n on $\mathbb{CP}^1$

$$\nabla = \mathrm{d} - \left(\frac{U}{t^2} + \frac{V}{t}\right)\mathrm{d}t$$

where $U = \mathrm{diag}(u_1, \ldots, u_n) \in \mathfrak{h}$ with $u_i \neq u_j$, and $V \in (\oplus_{\alpha \in \Phi} \mathfrak{g}_\alpha)$. The connection has a pole of order 2 at $t = 0$ (an irregular singularity). To find the flat sections $\nabla X = 0$ is the same as to find the solution to the associated system of n ordinary differential equations. For instance, when $n = 1$ the

equation is simply $f'(t) = \frac{U \cdot f(t)}{t^2}$, which is essentially the homogeneous version of the Euler's equation we considered earlier.

The fundamental solution of $\nabla X = 0$ is of the form

$$X = \phi(x) \cdot \begin{pmatrix} e^{-t_1/x} x^{\lambda_1} & & & \\ & e^{-t_2/x} x^{\lambda_2} & & \\ & & \cdots & \\ & & & e^{-t_n/x} x^{\lambda_n} \end{pmatrix}$$

where $\phi(x) \in \mathrm{GL}(\mathbb{C}[\![x]\!])$. The solutions are Borel summable except along the *Stokes rays*

$$\ell_{ij} = \mathbb{R}_{>0}(u_i - u_j) = \mathbb{R}_{>0} \cdot U(\alpha_{ij}).$$

Theorem 11.20. *If r is* not *a Stokes ray, and let $\mathbb{H}_r$ be the half plane with center given by r. There exists a unique holomorphic function*

$$X_r \colon \mathbb{H}_r \to \mathrm{GL}(n, \mathbb{C})$$

such that

$$\frac{\mathrm{d}X_r}{\mathrm{d}t} = \left(\frac{U}{t^2} + \frac{V}{t} \right) X_r$$

and

$$X_r(t) \cdot e^{U/t} \to I \quad \textit{as} \quad t \to 0.$$

Theorem 11.21. *Let Σ be a convex sector bounded by* non-Stokes *rays r_- and r_+. The previous theorem gives two holomorphic functions X_{r_-} and X_{r_+}. Then there exists*

$$g \in \exp\left(\oplus_{U(\alpha) \in \Sigma} \mathfrak{g}_\alpha \right) \subseteq G$$

such that

$$X_{r_+}(t) = X_{r_-}(t) \cdot g \quad \textit{for} \quad t \in \mathbb{H}_{r_-} \cap \mathbb{H}_{r_+}.$$

In particular, when the sector contains exactly one Stokes ray ℓ, the factor g describing how the solutions jump when one crosses the ray is called the *Stokes factor* and is denoted by

$$S_\ell \in \exp\left(\oplus_{U(\alpha) \in \ell} \mathfrak{g}_\alpha \right) \subseteq G$$

Note that it can be uniquely written as

$$S_\ell = \exp\left(\sum_{U(\alpha) \in \ell} \epsilon_\alpha \right) \quad \text{where} \quad \epsilon_\alpha \in \mathfrak{g}_\alpha.$$

A family of such connections is called *isomonodromic* if for all convex sectors Σ the product

$$\prod_{\ell \subseteq \Sigma}^{\curvearrowright} S_\ell$$

is constant as the connections vary.

Remark 11.22. If we know the Stokes factor at a single point in an isomonodromic family of equations, then the Stokes factors at any other points in the family can be uniquely determined.

The parallel between Stokes data and stability conditions on abelian categories can be summarized as follows.

Stokes data	Stability conditions
$U \in \mathfrak{h}$	$Z\colon K_0(\mathcal{A}) \to \mathbb{C}$
Stokes ray $\ell_{ij} = \mathbb{R}_{>0} \cdot U(\alpha_{ij})$	$\mathbb{R}_{>0} \cdot Z(E)$ for E stable
Stokes factor $\epsilon_\alpha \in \mathfrak{g}_\alpha$	Counting stables $\Omega(\gamma)$

In fact, the wall-crossing formula of stability conditions can be regarded as an infinite dimensional categorical generalization of isomonodromic families of connections. See [3] for more details.

Bibliography

[1] M. Artin. *Algebra* (Second Edition). Pearson Education, 2011.

[2] M. Atiyah and I. G. Macdonald. *Introduction to commutative algebra*. Westview Press, Boulder, CO, 2016.

[3] T. Bridgeland and V. Toledano Laredo. *Stability conditions and Stokes factors*. Invent. Math. 187 (2012), no. 1, 61–98.

[4] P. L. Clark. *2010 Summer Course on Model Theory*. Available at: http://alpha.math.uga.edu/ pete/modeltheory2010FULL.pdf

[5] J. H. Conway. *The sensual (quadratic) form*. Carus Mathematical Monographs, 26. Mathematical Association of America, Washington, DC, 1997.

[6] F. Diamond and J. Shurman. *A first course in modular forms*. Graduate Texts in Mathematics, 228. Springer-Verlag, New York, 2005.

[7] M. Einsiedler and T. Ward. *Ergodic theory with a view towards number theory*. Graduate Texts in Mathematics, 259. Springer-Verlag London, Ltd., London, 2011.

[8] É. Ghys. *A singular mathematical promenade*. ENS Éditions, Lyon, 2017.

[9] J. E. Greene and A. Lobb. *The rectangular peg problem*. Ann. of Math. (2) 194 (2021), no. 2, 509–517.

[10] A. Hatcher. *Algebraic topology*. Cambridge University Press, Cambridge, 2002.

[11] B. Keller. *On cluster theory and quantum dilogarithm identities*. Representations of algebras and related topics, 85–116, EMS Ser. Congr. Rep., Eur. Math. Soc., Zürich, 2011.

[12] M. Khovanov. *A categorification of the Jones polynomial*. Duke Math. J. 101 (2000), no. 3, 359–426.

[13] W. B. R. Lickorish. *An introduction to knot theory*. Graduate Texts in Mathematics, 175. Springer-Verlag, New York, 1997.

[14] J. Matousek. *Using the Borsuk-–Ulam Theorem*. Lectures on topological methods in combinatorics and geometry. Universitext. Springer-Verlag, Berlin, 2003.

[15] J.-P. Serre. *A course in arithmetic*. Graduate Texts in Mathematics, No. 7. Springer-Verlag, New York-Heidelberg, 1973.

[16] E. M. Stein and R. Shakarchi. *Complex analysis*. Princeton Lectures in Analysis, 2. Princeton University Press, Princeton, NJ, 2003.

[17] P. Walter. *An introduction to ergodic theory*. Graduate Texts in Mathematics, 79. Springer-Verlag, New York-Berlin, 1982.

[18] D. Zagier. *Newman's Short Proof of the Prime Number Theorem*. The American Mathematical Monthly, 104 (8), 705–708, 1997.

[19] D. Zagier. *The dilogarithm function.* Frontiers in number theory, physics, and geometry. II, 3–65, Springer, Berlin, 2007.