2022年春季学期复变函数读书报告汇编

报告人：

严仲谨　刘遥　徐啸宇　郑植　熊茳楠　房奕嘉　薛博文

李昊达　何芊杜　宋典毅　梁思恒　雷语昊　陈航　胡永乐

李骁然　尹顺　虞鹏浩　方星竹　丘瑞岑　梅文九　李乔

陈嘉和　饶睿　陈冠伊　余学一　吴宇阳　王伟涛　温家睿

李振鹏　吴迪　王可为　沙宪宝　王骏澎

组织者：

方汉隆　宋英齐

Number Theory

Physics

Applied Math

Analysis

Topology

Number Theory

The Arithmetic Class Number Formula

Absrtact: In this reading report, we first introduce the functional equation for Dedekind-Zeta function following Hecke's approach, then we combine the function equation for Dedekind-Zeta function and Dirichlet L-function to derive the discriminant-conductor formula, at the end of this report we give the arithmetic class number formula.

1 Inrtoduction(Analytic Class Number Formula)

For a number field K, the Dedekind-Zeta function for K is defined by:

$$\zeta_K(s) = \sum_I \frac{1}{N(I)^s}$$

where I run through all integral ideals of $\mathcal{O}_K$ and $N(I)$ denotes the ideal norm of I.

It can be easily seen that $\zeta_K(s)$ defines a holomorphic function on $\mathrm{Re}(s) > 1$, and $\zeta_K(s)$ admits the following Euler factorization:

$$\zeta_K(s) = \prod_{\mathfrak{p}\in \mathrm{Spec}(\mathcal{O}_K)} \frac{1}{1-N(\mathfrak{p})^{-s}}$$

Let $\mathfrak{N}_K(t)$ denote the number for integral ideals in $\mathcal{O}_K$ with norm $\leq t$, then:

$$\zeta_K(s) = \sum_{n=1}^{\infty} \frac{\mathfrak{N}_K(n)-\mathfrak{N}_K(n-1)}{n^s}$$

However, it is not easy to calculate $\mathfrak{N}_K(t)$. Let C be an ideal class, consider

$$\zeta_K(s,C) = \sum_{I\in C} \frac{1}{N(I)^s}$$

where I run through all integral ideals in C, and let $\mathfrak{N}_K(t,C)$ denote the number of integral ideals in C with norm $\leq t$.

Let C^{-1} be the inverse class of C, and let $J \in C^{-1}$ be an integral class, then we have the following bijection:

$$\{\text{integral ideal in C with norm} \leq t\} \longleftrightarrow$$
$$\{\text{principal ideal divide by } J \text{ with norm} \leq tN(J)\}$$
$$I \mapsto IJ$$

For an integral ideal I, let $\mathfrak{N}_K(t,I)$ be the number of principal ideal divide by I with norm $\leq t$. Clearly, $\mathfrak{N}_K(t,I)$ is closely realated to the number of elements in I. More precisely, $\mathfrak{N}_K(t,I)$ equals the number orbits of elements in I with norm $\leq t$ under the multiplication of $\mathcal{O}_K^{\times}$.

By the geometry of numbers, we have the maps

$$K \xrightarrow{j} \mathbb{R}^{r_1} \times \mathbb{C}^{r_2} \xrightarrow{\ln} \mathbb{R}^{r_1+r_2}$$

$$x \mapsto (\sigma_1(x), ..., \sigma_{r_1}(x), \tau_1(x), ..., \tau_{r_2}(x)) \mapsto$$

$$(\ln|\sigma_1(x)|, ..., \ln|\sigma_{r_1}(x)|, 2\ln|\tau_1(x)|, ..., 2\ln|\tau_{r_2}(x)|)$$

where $\sigma_1, ..., \sigma_{r_1}$ are all real embeddings of K and $\tau_1, \bar{\tau}_1, ..., \tau_{r_2}, \bar{\tau}_{r_2}$ are all complex embeddings of K. It is easily seen that $r_1 + 2r_2 = n = [K : \mathbb{Q}]$.

It is well known that $j(\mathcal{O}_K)$ is a complete lattice in $\mathbb{R}^{r_1} \times \mathbb{C}^{r_2} = \mathbb{R}^n$, and $\tilde{j}(\mathcal{O}_K)(\tilde{j} = \ln \circ j)$ is a lattice of dimension $r_1 + r_2 - 1$ in $\mathbb{R}^{r_1+r_2}$, hence we have:

Dirichlet's unit theorem: As an abelian group, $\mathcal{O}_K^\times \cong \mu(K) \times \mathbb{Z}^{r_1+r_2-1}$

where $\mu(K)$ denotes the finite group of roots of units in K.

Let $\varepsilon_1, ..., \varepsilon_{r_1+r_2-1}$ be generators of the torsion-free part of $\mathcal{O}_K^\times$, we called them the fundamental units of K, and let $\lambda_k(\varepsilon_i)$ denotes the k-th component of $\tilde{j}(\varepsilon_i)$. It can be easily verify that all the (r_1+r_2-1)-th minors of the (r_1+r_2)×(r_1+r_2-1) matrix $(\lambda_i(\varepsilon_j))$ equals, and we denote it by R_K, the regulator of K.

Let $M = [0,1)\,\tilde{j}(\varepsilon_1) \oplus ... \oplus [0,1)\,\tilde{j}(\varepsilon_{r_1+r_2-1}) \oplus \mathbb{R}(1, ..., 1)$, it follows that the number of orbits of elements in I with norm $\leq t$ under the multiplication of the fundamental units equals the number of elements in I with norm $\leq t$ such that its image under j lies in $\ln^{-1}(M)$.

Let $S(t) = \left\{ (x_1, ..., x_{r_1}, z_1, ..., z_{r_2}) \in \mathbb{R}^{r_1} \times \mathbb{C}^{r_2} : \prod_{i=1}^{r_1} |x_i| \prod_{j=1}^{r_2} |z_j|^2 \leq t \right\}$, then $\mathfrak{N}_K(t, I) = \frac{|S(t) \cap \ln^{-1}(M) \cap j(\mathcal{O}_K)|}{|\mu(K)|}$. After some calculations (c.f. Appendix A), $\mathfrak{N}_K(t, I) = \frac{2^{r_1}(2\pi)^{r_2} R_K}{|\mu(K)|\sqrt{|d_K|}N(I)} t + O(t^{1-\frac{1}{n}})$, where d_K denote the discriminant of K.

It follows that $\mathfrak{N}_K(t, C) = \frac{2^{r_1}(2\pi)^{r_2} R_K}{|\mu(K)|\sqrt{|d_K|}} t + O(t^{1-\frac{1}{n}})$, hence $\zeta_K(s, C)$ can be analytic continuation to $\mathrm{Re}(s) > 1 - \frac{1}{n}$ to a meromorphic function with simple pole at $s = 1$ and $\mathrm{Res}_{s=1}\zeta_K(s, C) = \frac{2^{r_1}(2\pi)^{r_2} R_K}{|\mu(K)|\sqrt{|d_K|}}$. Hence we have:

Analytic Class Number Formula:$\zeta_K(s)$ can be analytic continuation to a meromorphic function on $\mathrm{Re} > 1 - \frac{1}{n}$ with simple pole at $s = 1$ and $\mathrm{Res}_{s=1}\zeta_K(s) = \frac{2^{r_1}(2\pi)^{r_2} R_K h_K}{|\mu(K)|\sqrt{|d_K|}}$, where h_K denote the class number of $\mathcal{O}_K$.

2 Functional equation for Dedekind-Zeta Function

In this section, we prove that for $Z_K(s) = A^s \Gamma(\frac{s}{2})^{r_1} \Gamma(s)^{r_2} \zeta_K(s) (A = 2^{-r_2}\sqrt{\frac{|d_K|}{\pi^n}})$, $Z_K(s)$ can be analytic continuation to an meromorphic function on $\mathbb{C}$ with only two simple poles 0 and 1, and satisfies the functional equation $Z_K(s) = Z_K(1-s)$.

The crucial lemma we will use is that:

Lemma: For $f, g : \mathbb{R}^+ \to \mathbb{R}$, suppose:

$$1. \lim_{x\to+\infty} f(x) = a_0, \lim_{x\to+\infty} g(x) = b_0$$

$$2. f(x) - a_0, g(x) - b_0 \text{ decrease rapidly at infty}$$

$$3.\text{there exists } C, k \in \mathbb{R}, k > 0, \text{ such that } f(\frac{1}{t}) = Ct^k g(t)$$

Then $\mathcal{M}(f - a_0)$,$\mathcal{M}(g - b_0)$ admits holomorphic continuation to $\mathbb{C} - \{0, k\}$,with:

$$1.\mathcal{M}(f - a_0), \mathcal{M}(g - b_0) \text{ have simple poles at 0,k}$$

$$2.\text{Res}_{s=0}\mathcal{M}(f - a_0)(s) = -a_0, \text{Res}_{s=k}\mathcal{M}(f - a_0)(s) = Cb_0$$

$$3.\mathcal{M}(f - a_0)(s) = C\mathcal{M}(g - b_0)(k - s)$$

where $\mathcal{M}(f)$ denotes the Mellin transformation, i.e. $\mathcal{M}(f)(s) = \int_0^{+\infty} f(t)t^s \frac{dt}{t}$

Let $j(I)/\mathcal{O}_K^\times$ denote the set of orbits of elements of $j(I)$ under the multiplication of $\mathcal{O}_K^\times$, and we defines:

$$\zeta_K(s, I) = \sum_{a\in j(I)/\mathcal{O}_K^\times} \frac{1}{N(a)^s}$$

Hence for an ideal class C, fixs an $J \in C^{-1}$, we have $\zeta_K(s, C) = N(J)^s \zeta_K(s, J)$.

Observe that $|N(a)| = \prod_{i=1}^{r_1} |\sigma_i(a)| \prod_{j=1}^{r_2} |\tau_j(a)|^2$, let r=$r_1$+$r_2$, we have:

$$(\frac{1}{2^{r_2}\pi^{\frac{n}{2}}})^s \Gamma(\frac{s}{2})^{r_1} \Gamma(s)^{r_2} \zeta_K(s, I) =$$
$$\sum_{a\in jI/\mathcal{O}_K^\times} \int_{\mathbb{R}_+^r} \exp(-\pi(\sum_{i=1}^{r_1} |\sigma_i(a)|^2 t_i + 2\sum_{j=1}^{r_2} |\tau_j(a)|^2 t_{r_1+j})) \prod_{i=1}^{r_1} |t_i|^{\frac{s}{2}} \prod_{j=1}^{r_2} |t_{r_1+j}|^s \prod_{k=1}^{r} \frac{dt_k}{t_k}$$

Consider the change of variables $\mathbb{R}_+^r \to \mathbb{R}^r, (t_1, ..., t_r) \mapsto (\ln t_1, ..., \ln t_{r_1}, 2\ln t_{r_1+1}, ..., 2\ln t_{r_1+r_2})$, the integration becomes:

$$= \sum_{a\in jI/\mathcal{O}_K^\times} 2^{-r_2} \int_{\mathbb{R}^r} \exp(-\pi(\sum_{i=1}^{r_1} |\sigma_i(a)|^2 e^{t_i} + 2\sum_{j=1}^{r_2} |\tau_j(a)|^2 e^{\frac{t_{r_1+j}}{2}})) \exp(\frac{\sum_{k=1}^{r} t_k}{2} s) \prod_{k=1}^{r} dt_k.$$

Since $\varepsilon_1, ..., \varepsilon_{r_1+r_2-1}, \frac{(1,...,1)}{r}$ forms a basis of $\mathbb{R}^r$ with Jacobian R_K, we change the basis again and gets:

$$= \sum_{a\in jI/\mathcal{O}_K^\times} 2^{-r_2} R_K \int_0^\infty t^{\frac{s}{2}} (\int_{\mathbb{R}^{r-1}} \exp(-\pi(\sum_{i=1}^{r_1} |\sigma_i(a)|^2 \prod_{k=1}^{r-1} |\sigma_i(\varepsilon_k)|^{\lambda_k} t^{\frac{1}{r}} +$$
$$2\sum_{j=1}^{r_2} |\tau_j(a)|^2 \prod_{k=1}^{r-1} |\tau_j(\varepsilon_k)|^{\lambda_k} t^{\frac{1}{2r}})) \prod_{k=1}^{r-1} d\lambda_k) \frac{dt}{t}$$

$$= \tfrac{2^{-r_2}R_K}{|\mu(K)|} \int_0^\infty t^{\frac{s}{2}} (\sum_{a\in I-\{0\}} \int_{[0,2]^{r-1}} \exp(-\pi(\sum_{i=1}^{r_1} |\sigma_i(a)|^2 \prod_{k=1}^{r-1} |\sigma_i(\varepsilon_k)|^{\lambda_k} t^{\frac{1}{r}} + 2\sum_{j=1}^{r_2} |\tau_j(a)|^2 \prod_{k=1}^{r-1} |\tau_j(\varepsilon_k)|^{\lambda_k} t^{\frac{1}{2r}})) \prod_{k=1}^{r-1} d\lambda_k) \tfrac{dt}{t}$$

Let $f(a,\lambda,t) = \exp(-\pi(\sum\limits_{i=1}^{r_1} |\sigma_i(a)|^2 \prod\limits_{k=1}^{r-1} |\sigma_i(\varepsilon_k)|^{\lambda_k} t^{\frac{1}{r}} + 2\sum\limits_{j=1}^{r_2} |\tau_j(a)|^2 \prod\limits_{k=1}^{r-1} |\tau_j(\varepsilon_k)|^{\lambda_k} t^{\frac{1}{2r}}))$, and $\theta(\lambda,t,I) = \sum\limits_{a\in I} f(a,\lambda,t)$. Then by Poisson summation formula for lattice, we have $\theta(\lambda,t^{-1},I) = \frac{t^{-\frac{1}{2}}}{N(I)\sqrt{|d_K|}}\theta(-\lambda,t,I^\vee)$ (c.f. Appendix B), where $I^\vee = \{x\in K : \mathrm{Tr}(xI)\subset\mathbb{Z}\}$ is the dual ideal for I.We have $I^\vee = D_K^{-1}I^{-1}$, where D_K is the differential ideal for K, which satisfies $N(D_K) = |d_K|$.

Since the integration of $\theta(\lambda,t,I)$ on every fundamental meshs in $\mathbb{R}^{r-1}$ equals, we have $\int_{[0,2]^{r-1}} \theta(\lambda,t,I)d\lambda = \int_{[0,2]^{r-1}} \theta(-\lambda,t,I)d\lambda$. Let $F(t,I) = \int_{[0,2]^{r-1}} \theta(\lambda,t,I)d\lambda$, then we have $F(t^{-1},I) = \frac{t^{-\frac{1}{2}}}{N(I)\sqrt{|d_K|}}F(\frac{1}{t},I^\vee)$, with $F(+\infty,I) = \int_{[0,2]^{r-1}} d\lambda = 2^{r-1}$.

Let $Z_K(s,I) = (\frac{1}{2^{r_2}\pi^{\frac{n}{2}}})\Gamma(\frac{s}{2})^{r_1}\Gamma(s)^{r_2}\zeta_K(s,I)$, then $Z_K(s,I) = \mathcal{M}(F(t,I)-F(\infty,I))(\frac{s}{2})$, by the lemma, $Z_K(s,I)$ admits analytic continuation to $\mathbb{C}-\{0,1\}$, with simple pole 0,1 and $\mathrm{Res}_{s=1}Z_K(s,I) = \frac{2^{r_1}R_K}{N(I)|\mu(K)|\sqrt{|d_K|}}$. Moreover, we have $Z_K(s,I) = \frac{1}{N(I)\sqrt{|d_K|}}Z_K(1-s,I^\vee)$.

Since $I^\vee = D_K^{-1}I^{-1}$, for an ideal class C, let $C^\vee = D_K^{-1}C^{-1}$, clearly, $\zeta_K(s) = \sum\limits_C \zeta_K(s,C^\vee)$. Then $N(I)^sZ_K(s,I) = \frac{N(I)^{s-1}}{\sqrt{|d_K|}}Z_K(1-s,I^\vee) = |d_K|^{\frac{1}{2}-s}N(I^\vee)^{1-s}Z_K(1-s,I^\vee)$, hence $|d_K|^{\frac{s}{2}}N(I)^sZ_K(s,I) = |d_K|^{\frac{1-s}{2}}N(I^\vee)^{1-s}Z_K(1-s,I^\vee)$.

Let $Z_K(s) = A^s\Gamma(\frac{s}{2})^{r_1}\Gamma(s)^{r_2}\zeta_K(s)$, then $Z_K(s) = \sum\limits_I |d_K|^{\frac{s}{2}}N(I)^sZ_K(s,I)$, hence $Z_K(s) = Z_K(1-s)$.

3 Discriminant-Conductor Formula

Observe that we have:

$$\Gamma(s)\Gamma(s+\tfrac{1}{2}) = 2^{1-2s}\Gamma(2s)$$

Then $\tilde{Z}_K(s) = |d_K|^{\frac{s}{2}}(\pi^{-\frac{s}{2}}\Gamma(\frac{s}{2}))^{r_1+r_2}(\pi^{-\frac{s}{2}}\Gamma(\frac{1+s}{2}))^{r_2}\zeta_K(s)$ satisfies the functional equation $\tilde{Z}_K(s) = \tilde{Z}_K(1-s)$.

Let χ be an primitive Dirichlet character with conductor f_χ, it is well known that $L(s,\chi) = \sum\limits_{n=0}^{\infty} \frac{\chi(n)}{n^2}$ satisfies the following functional equation, and hence adimts an analytic continuation to $\mathbb{C}$ (c.f. Appdendix C):

$$\Lambda(s,\chi) = W(\chi)\Lambda(1-s,\overline{\chi})$$

where $\Lambda(s,\chi) = (\frac{f_\chi}{\pi})^{\frac{s+\delta}{2}}\Gamma(\frac{s+\delta}{2})L(s,\chi)$, $W(\chi) = \frac{\tau(\chi)}{i^\delta\sqrt{f_\chi}}$, $\tau(\chi)$ the gauss sum of χ, and $\delta = \frac{1-\chi(-1)}{2}$

Suppose $K/\mathbb{Q}$ abelian, then by Kronecker-Weber theorem (c.f. Appendix D), $K \subset \mathbb{Q}(\xi_n)$ for some integer $n > 0$, where $\xi_n = \exp(\frac{2\pi i}{n})$. Hence for abelian K, we have:

$$\zeta_K(s) = \prod_\chi L(s,\chi)$$

where χ run through every characters of Gal$(K/\mathbb{Q})$ (c.f. Appendix E).

It follows that $\prod\limits_\chi \Lambda(s,\chi) = \prod\limits_\chi W(\chi) \prod\limits_\chi \Lambda(1-s,\chi)$, which gives another functional equation for ζ_K. Observe that $\chi(-1) = 1$ if and only if Fix$(\chi^\perp) \subset \mathbb{R}$, and since $K/\mathbb{Q}$ is abelian, either $r_1 = n$ and $K \subset \mathbb{R}$ or $r_2 = \frac{n}{2}$ and $K \not\subset \mathbb{R}$. Comparing the two functional equation, we have (**discrinminant-conductor formula**):

$$\begin{cases} \prod\limits_\chi f_\chi = |d_K| \\ \prod\limits_\chi \tau(\chi) = \sqrt{|d_K|} & \text{if } K \subset \mathbb{R} \\ \prod\limits_\chi \tau(\chi) = i^{\frac{n}{2}}\sqrt{|d_K|} & \text{if } K \not\subset \mathbb{R} \end{cases}$$

4 Arithmetic Class Number Formula

In this section, we will gives explict formula of $L(1,\chi)$ for a primitive Dirichlet character χ, and refined the analytic classs formula for an abelian extension $K/\mathbb{Q}$, which is known as the arithmetic class number formula.

For Re$(s) > 1$, we have $L(s,\chi) = \sum\limits_{a=1}^{f_\chi} \chi(a) \sum\limits_{k=0}^{\infty} \frac{1}{(a+kf_\chi)^s}$. Obverse that

$$\sum_{b=1}^{f_\chi} \xi_{f_\chi}^{(a-n)b} = \begin{cases} f_\chi & \text{if } n \equiv a \bmod f_\chi \\ 0 & \text{else} \end{cases}$$

Hence $\frac{1}{(a+kf_\chi)^s} = \frac{1}{f_\chi} \sum\limits_{n=kf_\chi+1}^{(k+1)f_\chi} \frac{1}{n^s} \sum\limits_{b=1}^{f_\chi} \xi_{f_\chi}^{(a-n)b}$, then $L(s,\chi) = \sum\limits_{a=1}^{f_\chi}\sum\limits_{b=1}^{f_\chi} \frac{\chi(a)\xi_{f_\chi}^{ab}}{f_\chi} \sum\limits_{n=1}^{\infty} \frac{\xi_{f_\chi}^{-nb}}{n^s}$. Since for a nontrivial character, both sides converges for Re$(s) > 0$, it follows that $L(1,\chi) = -\sum\limits_{a=1}^{f_\chi}\sum\limits_{b=1}^{f_\chi} \frac{\chi(a)\xi_{f_\chi}^{ab}}{f_\chi} \ln(1-\xi_{f_\chi}^{-b})$.

It is easy to verify that $\sum\limits_{a=1}^{f_\chi} \chi(a)\xi_{f_\chi}^{ab} = \overline{\chi}(b)\tau(\chi)$, hence we have:

$$L(1,\chi) = -\frac{\tau(\chi)}{f_\chi} \sum_{b=1}^{f_\chi} \overline{\chi}(b) \ln(1-\xi_{f_\chi}^{-b})$$

Obverse that:

$$1-\xi_{f_\chi}^{-b} = 1-\exp(-\tfrac{2\pi ib}{f_\chi}) = \exp(-\tfrac{\pi ib}{f_\chi})(\exp(\tfrac{\pi ib}{f_\chi})-\exp(-\tfrac{\pi ib}{f_\chi})) = 2i\sin(\tfrac{\pi b}{f_\chi})\exp(-\tfrac{\pi ib}{f_\chi}) = \sin(\tfrac{\pi b}{f_\chi})\exp(\pi i(\tfrac{1}{2}-\tfrac{b}{f_\chi}))$$

Since $|\frac{1}{2}-\frac{b}{f_\chi}|<1$, we have $\ln(1-\xi_{f_\chi}^{-b}) = \ln|\sin(\frac{\pi b}{f_\chi})|+i\pi(\frac{1}{2}-\frac{b}{f_\chi})$, hence we have:

$$L(1,\chi) = -\tfrac{\tau(\chi)}{f_\chi}\sum_{b=1}^{f_\chi}\overline{\chi}(b)(\ln|\sin(\tfrac{\pi b}{f_\chi})|-\tfrac{i\pi}{f_\chi}b)$$

obverse that $\begin{cases} \displaystyle\sum_{b=1}^{f_\chi}\overline{\chi}(b)\ln|\sin(\frac{\pi b}{f_\chi})| = 0 & \text{if } \chi(-1)=-1 \\ \displaystyle\sum_{b=1}^{f_\chi}\overline{\chi}(b)b = 0 & \text{if } \chi(-1)=1 \end{cases}$ hence we have for every non-trivial χ:

$$L(1,\chi) = \begin{cases} -\dfrac{\tau(\chi)}{f_\chi}\displaystyle\sum_{b=1}^{f_\chi}\overline{\chi}(b)\ln|\sin(\frac{\pi b}{f_\chi})| & \text{if } \chi(-1)=1 \\ \dfrac{\tau(\chi)\pi i}{f_\chi^2}\displaystyle\sum_{b=1}^{f_\chi}\overline{\chi}(b)b & \text{if } \chi(-1)=-1 \end{cases}$$

For trivial character χ_0, it is obviously that $L(s,\chi_0)=\zeta(s)$, since $\mathrm{Res}_{s=1}\zeta(s)=1$, we have $\mathrm{Res}_{s=1}\zeta_K(s) = \prod\limits_{\chi\neq\chi_0}L(1,\chi)$

Let G denote the group of characters of $\mathrm{Gal}(K/\mathbb{Q})$, and G_0 the subgroup of even characters, then $[G:G_0] = \begin{cases} 1 & \text{if } K\subset\mathbb{R} \\ 2 & \text{if } K\not\subset\mathbb{R} \end{cases}$.

It follows that:

$$\begin{cases} \dfrac{2^n h_K R_K}{|\mu(K)|\sqrt{|d_K|}} = (-1)^{n-1}\displaystyle\prod_{\chi\neq\chi_0}\frac{\tau(\chi)}{f_\chi}\prod_{\chi\neq\chi_0}\sum_{b=1}^{f_\chi}\overline{\chi}(b)\ln|\sin(\frac{\pi b}{f_\chi})| & \text{if } K\subset\mathbb{R} \\ \dfrac{(2\pi)^{\frac{n}{2}} h_K R_K}{|\mu(K)|\sqrt{|d_K|}} = (-1)^{\frac{n}{2}-1}(\pi i)^{\frac{n}{2}}\displaystyle\prod_{\chi\neq\chi_0}\frac{\tau(\chi)}{f_\chi}\prod_{\substack{\chi\neq\chi_0\\ \chi\in G_0}}\sum_{b=1}^{f_\chi}\overline{\chi}(b)\ln|\sin(\frac{\pi b}{f_\chi})|\prod_{\chi\notin G_0}\sum_{b=1}^{f_\chi}\frac{\overline{\chi}(b)b}{f_\chi} & \text{if } K\not\subset\mathbb{R} \end{cases}$$

By discriminant-conductor formula, we have:

$$\prod_{\chi\neq\chi_0}\frac{\tau(\chi)}{f_\chi} = \begin{cases} \dfrac{1}{\sqrt{|d_K|}} & \text{if } K\subset\mathbb{R} \\ \dfrac{i^{\frac{n}{2}}}{\sqrt{|d_K|}} & \text{if } K\not\subset\mathbb{R} \end{cases}$$

Combined those results, we get:

$$h_K R_K = \begin{cases} \prod\limits_{\chi \neq \chi_0} \sum\limits_{b=1}^{f_\chi} \frac{-\overline{\chi}(b)}{2} \ln|\sin(\frac{\pi b}{f_\chi})| & \text{if } K \subset \mathbb{R} \\ \frac{\mu(K)}{2} \prod\limits_{\substack{\chi \neq \chi_0 \\ \chi \in G_0}} \sum\limits_{b=1}^{f_\chi} \frac{-\overline{\chi}(b)}{2} \ln|\sin(\frac{\pi b}{f_\chi})| \prod\limits_{\chi \notin G_0} \sum\limits_{b=1}^{f_\chi} \frac{-\overline{\chi}(b) b}{2 f_\chi} & \text{if } K \not\subset \mathbb{R} \end{cases}$$

5 Appendix A: Asymptotic Analysis for $\mathfrak{N}_K(t, I)$

In this section, we will prove that $\mathfrak{N}_K(t, I) = \frac{2^{r_1}(2\pi)^{r_2} R_K}{|\mu(K)|\sqrt{|d_K|}N(I)} t + O(t^{1-\frac{1}{n}})$.

The following lemma is crucial in our proof:

Lemma: Let D be a subset of $\mathbb{R}^n$, and L a complete lattice in $\mathbb{R}^n$. Suppose the boundary of D is (n-1)-Lipschitz parametrizable, then the number of fundamental meshs of L which intersects with $\partial(tD)$ is $O(t^{n-1})$.

Let $n^-(t)$ be the number of element in $j(I)$ whose fundamental mesh was contained in $\ln^{-1}(M) \cap S(t) = t^{\frac{1}{n}}(\ln^{-1}(M) \cap S(1))$, and $n^+(t)$ the number of element in $j(I)$ whoes fundamental mesh intersects $t^{\frac{1}{n}}(\ln^{-1}(M) \cap S(1))$ nonempty, then $n^+(t) - n^-(t)$ is the number of elements whose fundamental mesh intersects with the boundary of $t^{\frac{1}{n}}(\ln^{-1}(M) \cap S(1))$. By the lemma, we have $n^+ - n^- = O(t^{1-\frac{1}{n}})$.

Clearly, we have $n^- \leq |\mu(K)|\mathfrak{N}_K(t, I) \leq n^+$, and $\mathrm{vol}(j(I))n^- \leq \mathrm{vol}(\ln^{-1}(M) \cap S(1))t \leq \mathrm{vol}(j(I))n^+$, hence we have $\mathfrak{N}_K(t, I) = \frac{\mathrm{vol}(\ln^{-1}(M \cap S(1)))}{|\mu(K)|\sqrt{|d_K|}N(I)} t + O(t^{1-\frac{1}{n}})$

Let $M_0 = [0,1)\,\tilde{j}(\varepsilon_1) \oplus ... \oplus [0,1)\,\tilde{j}(\varepsilon_{r_1+r_2-1}) \oplus (-\infty, 0]\,(1, ..., 1)$, then $\ln^{-1}(M \cap S(1)) = \ln^{-1}(M_0)$. And it is not hard to conclude that $\mathrm{vol}(\ln^{-1}(M_0)) = 2^{r_1}(2\pi)^{r_2} R_K$, hence we have proved that:

$$\mathfrak{N}_K(t, I) = \frac{2^{r_1}(2\pi)^{r_2} R_K}{|\mu(K)|\sqrt{|d_K|}N(I)} t + O(t^{1-\frac{1}{n}})$$

6 Appendix B: Functional Equation for $\theta(\lambda, t, I)$

In this section, we will prove that $\theta(\lambda, t^{-1}, I) = \frac{t^{-\frac{1}{2}}}{N(I)\sqrt{|d_K|}} \theta(-\lambda, t, I^\vee)$.

The crucial result we will use is

Poisson summation formula:Let $f \in \mathrm{S}(R^n)$, and L a complete lattice in $\mathbb{R}^n$, then $\sum\limits_{x \in L} f(x) = \frac{1}{\mathrm{vol}(L)} \sum\limits_{x \in L^\vee} \mathcal{F}f(x)$, where $\mathcal{F}f$ denote the Fourier transform of f, $\mathrm{vol}(L)$ denote the volume of fundamental meshes of L, and $L^\vee$ denote the dual lattive of L (the lattice generated by the dual basis for the basis of L, in particular, $j(I)^\vee = j(I^\vee)$).

Consider the gauss function, $g(x) = \exp(-\pi(x,x)) = \exp(-\pi \sum\limits_{k=1}^{n} x_k^2)$, it is easy to verify that $\mathcal{F}g(x) = g(x)$.

For an positive definite matrix A, let $g_A(x) = \exp(-\pi(x, Ax))$. It is well known that there exist an invertible matrix C s.t. $A = CC^T$, hence $g_A(x) = \exp(-\pi(Cx, Cx))$, and it is easy to vearify that $\mathcal{F}g_A(x) = \frac{1}{\sqrt{|\det A|}} g_{(A^{-1})^T}(x)$.

Hence for $f(x, \lambda, t) = \exp(-\pi(\sum\limits_{i=1}^{r_1} |x_i|^2 \prod\limits_{k=1}^{r-1} |\sigma_i(\varepsilon_k)|^{\lambda_k} t^{\frac{1}{r}} + 2 \sum\limits_{j=1}^{r_2} |x_{r_1+j}|^2 \prod\limits_{k=1}^{r-1} |\tau_j(\varepsilon_k)|^{\lambda_k}) t^{\frac{1}{2r}})$, we have $\mathcal{F}f(x, \lambda, t) = t^{-\frac{1}{2}} f(x, -\lambda, \frac{1}{t})$.

It follows that

$$\theta(\lambda, t, I) = \sum_{a \in j(I)} f(a, \lambda, t) = \frac{t^{-\frac{1}{2}}}{N(I)\sqrt{|d_K|}} \sum_{a \in j(I^\vee)} f(a, -\lambda, \tfrac{1}{t}) = \theta(-\lambda, \tfrac{1}{t}, I^\vee).$$

7 Appendix C: Functional Equation for $L(s, \chi)$

In this section, we will prove that $\Lambda(s, \chi) = W(\chi)\Lambda(1 - s, \overline{\chi})$, where $\Lambda(s, \chi) = (\frac{f_\chi}{\pi})^{\frac{s+\delta}{2}} \Gamma(\frac{s+\delta}{2}) L(s, \chi)$, $W(\chi) = \frac{\tau(\chi)}{i^\delta \sqrt{f_\chi}}$, $\delta = \frac{1-\chi(-1)}{2}$. Without loss of generality, we will assume χ is nontrivial.

Let $\theta(t, \chi) = \sum\limits_{n \in \mathbb{Z}} \chi(n) \exp(-\pi(\frac{n}{f_\chi})^2 t)$, then by poisson summation formula, we have:

$$\theta(t, \chi) = \sum_{a=1}^{f_\chi} \chi(a) \sum_{n \in \mathbb{Z}} \exp(-\pi(n + \tfrac{a}{f_\chi})^2 t) = \sum_{a=1}^{f_\chi} \chi(a) \sum_{n \in \mathbb{Z}} t^{-\frac{1}{2}} \exp(-\pi \tfrac{n^2}{t}) \exp(\tfrac{2\pi i n a}{f_\chi}) =$$

$$t^{-\frac{1}{2}} \sum_{n \in \mathbb{Z}} \exp(-\pi \tfrac{n^2}{t}) \sum_{a=1}^{f_\chi} \chi(a) \xi_{f_\chi}^{an} = \tau(\chi) t^{-\frac{1}{2}} \theta(\tfrac{f_\chi^2}{t}, \overline{\chi})$$

We can modify $\theta(t, \chi)$ slightly by let $\theta(t, \chi) = \sum\limits_{n \in \mathbb{Z}} \chi(n) \exp(\frac{-\pi n^2 t}{f_\chi})$, then we have $\theta(t, \chi) = \tau(\chi) t^{-\frac{1}{2}} \theta(\frac{1}{t}, \overline{\chi})$.

Suppose $\chi(-1) = 1$, it is easy to see that $\frac{\theta(t,\chi)}{2} = \sum\limits_{n=1}^{\infty} \chi(n) \exp(\frac{-\pi n^2 t}{f_\chi})$, and hence $\Gamma(\frac{s}{2}) L(s, \chi) = (\frac{\pi}{f_\chi})^{\frac{s}{2}} \mathcal{M}(\frac{\theta(t,\chi)}{2})(\frac{s}{2})$, and we get the functional equation by the lemma in section 3.

The difficulty occurs when $\chi(-1) = -1$, since for these χ, we have $\theta(t, \chi) \equiv 0$. This is because $\chi(n) \exp(-\frac{\pi n^2 t}{f_\chi})$ is odd function for n.

Notice that $\theta(t, \chi) = \sum\limits_{n \in \mathbb{Z}} n\chi(n) \exp(-\frac{\pi n^2 t}{f_\chi})$ is odd, and $\Gamma(\frac{s+1}{2}) L(s, \chi) = (\frac{\pi}{f_\chi})^{\frac{s+1}{2}} \mathcal{M}(\frac{\theta(t,\chi)}{2})(\frac{s+1}{2})$, it remains to find the functional equation for this $\theta(t, \chi)$.

By poisson summation formula, we have:

$$\sum_{n\in\mathbb{Z}}\chi(n)\exp(-\tfrac{\pi(n+x)^2t}{f_\chi})=\tfrac{\tau(\chi)}{\sqrt{f_\chi}}t^{-\frac{1}{2}}\sum_{n\in\mathbb{Z}}\overline{\chi}(n)\exp(-\tfrac{\pi n^2}{f_\chi t})\exp(\tfrac{2\pi inx}{f_\chi})$$

Differential by x on both sides and let $x=0$, we get:

$$\sum_{n\in\mathbb{Z}}n\chi(n)\exp(-\tfrac{\pi n^2t}{f_\chi})=\tfrac{\tau(\chi)}{i\sqrt{f_\chi}}t^{-\frac{3}{2}}\sum_{n\in\mathbb{Z}}n\overline{\chi}(n)\exp(\tfrac{-\pi n^2}{f_\chi t})$$

Hence we get the functional equation of $L(s,\chi)$ for $\chi(-1)=-1$.

8 Appendix D: Kronecker-Weber Theorem

In this section, we will prove the Kronecker-Weber theorem using some basic facts about higher-ramification group.

Suppose $K/\mathbb{Q}$ is an abelian extension, for prime $\mathfrak{p}|p$, we define the j-th ramification group to be:

$$V_{\mathfrak{p}}^j=\{\sigma\in\operatorname{Gal}(K/\mathbb{Q}):\sigma(x)\equiv x \mod \mathfrak{p}^{j+1},\forall x\in\mathcal{O}_K\}$$

Since $K/\mathbb{Q}$ is abelian, it is easy to see that $V_{\mathfrak{p}}^j$ coincidence for all $\mathfrak{p}|p$, hence we will write V_p^j instead of $V_{\mathfrak{p}}^j$

The followings are the facts we will use in the proof of Kronecker-Weber theorem:

1.$\frac{V_p^0}{V_p^1}$ can be embeded into $\mathbb{F}_p^\times$.

2.$\frac{V_p^j}{V_p^{j+1}}$ can be embeded into the additive group of $\mathbb{F}_p$ for $j\geq 1$.

We call a field cyclotomic if it can be embeded into some $\mathbb{Q}(\xi_n)$. It is obviously that the composite field of two cyclotomic fields is also cyclotic, hence by the structure of finite abelian group, we only needs to prove that every cyclic extension of prime power order is cyclotomic.

First, we will prove that for a cyclic extension $K/\mathbb{Q}$, s.t. $[K:\mathbb{Q}]=p^m$, suppose p is the only prime that ramifies in $[K:\mathbb{Q}]$, then K is the unique subfield of order p^m of $\mathbb{Q}(\xi_{p^{m+1}})$.

Let L be the unique subfiled of order p^m of $\mathbb{Q}_{\xi_{p^{m+1}}}$, we only needs to prove that $[KL:\mathbb{Q}]=p^m$.

Since there exists an canonical embedding $\operatorname{Gal}(KL/\mathbb{Q})\to\operatorname{Gal}(K/\mathbb{Q})\times\operatorname{Gal}(L/\mathbb{Q})$,$\sigma\mapsto(\sigma|_K,\sigma|_L)$, hence $KL/\mathbb{Q}$ is abelian. Noticed that $T_{q,KL/\mathbb{Q}}\subset T_{q,K/\mathbb{Q}}\times T_{q,L/\mathbb{Q}}$, where T_q denotes the ramification group for q, it follows that p is the only prime that ramifies in $KL/\mathbb{Q}$. Notice that the order of element in $\operatorname{Gal}(KL/\mathbb{Q})$ is less than p^m, if $KL/\mathbb{Q}$ is cyclic, then we must have $[KL : \mathbb{Q}] \leq p^m$, hence we only needs to prove that $KL/\mathbb{Q}$ is cyclic (this result only holds for $p \neq 2$).

Lemma: Suppose $K/\mathbb{Q}$ is abelian of order p^n ($p \neq 2$), and p is the only prime that ramifies in $K/\mathbb{Q}$, then $K/\mathbb{Q}$ is cyclic.

Proof: Noticed that the fixed field of T_p is an unramified extension of $\mathbb{Q}$, hence $\mathrm{Fix}(T_p) = \mathbb{Q}$, that is, $\mathrm{Gal}(K/\mathbb{Q}) = T_p = V_p^0$ (hence p must ramifies totally) . Since $\left[T_p : V_p^1\right] |p-1$, we must have $T_p = V_p^1$.

By the structure of finite abelian group, $\mathrm{Gal}(K/\mathbb{Q})$ is abelian if and only if it contains a unique subgroup of order p^{m-1}. Since $\cap_{j=1}^{\infty} V_p^j = \{Id\}$, and $\left[V_p^j : V_p^{j+1}\right] = 1$ or p for $j \geq 1$, there exist a subgroup of order p^{m-1} of $\mathrm{Gal}(K/\mathbb{Q})$ (we will show that V_p^2 is that group).

Let H be a subgroup of order p^{m-1} of $\mathrm{Gal}(K/\mathbb{Q})$, we will proves that $V_p^2 \subset H$ (hence $H = V_p^2$). Let $L = \mathrm{Fix}(H)$, then under the canonical map $\mathrm{Gal}(L/\mathbb{Q}) = \frac{\mathrm{Gal}(K/\mathbb{Q})}{H}$, we have $\frac{V^2_{p,K/\mathbb{Q}}H}{H} \subset V^2_{p,L/\mathbb{Q}}$, hence we only needs to prove that $V^2_{p,L/\mathbb{Q}} = \{Id\}$.

Since p ramifies totally in $L/\mathbb{Q}$, we can reduces to deal with the p-adic case. Namely, for $L/\mathbb{Q}_p$ abelian,totally ramifies,$[L : \mathbb{Q}_p] = p$, we have $V_p^2 = \{Id\}$. Since $L/\mathbb{Q}_p$ totally ramifies, we have $L = \mathbb{Q}_p(\pi_L)$. Let $f(X) = a_0 + a_1X + ... + a_pX^p$ be the minimal polynomial of π_l, then $f^{(1)}(\pi_L) = \prod\limits_{\sigma}(\pi_L - \sigma(\pi_L))$. Suppose $\mathrm{Gal}(L/\mathbb{Q}_p) = \frac{V_p^j}{V_p^{j+1}}$, then $v(f^{(1)}(\pi_L)) = v(\pi_L)^{(j+1)(p-1)}$. Also we have $f^(1)(\pi_L) = a_1 + 2a_2\pi_L + ... + p\pi_L^{p-1}$, hence $v(f^{(1)}(\pi_L)) \leq v(\pi_L)^{2p-1}$. Therefore, we must have $j = 1$, i.e. $V_p^2 = \{Id\}$. □

When p=2, we will use induction on m. It is well known that every quadratic fields is cyclotomic. For $m > 1$, let $L = \mathbb{Q}(\xi_{2^{m+2}} + \xi_{2^{m+2}}^{-1})$, then $L/\mathbb{Q}$ is cyclic of order 2^m. Noticed that K and L has $\mathbb{Q}(\sqrt{2})$ as subfield in common (consider the maximal real subfield of K). Hence $[KL : \mathbb{Q}] \leq 2^{2m-1}$. Choose a generator σ of $\mathrm{Gal}(K/\mathbb{Q})$ and a generator τ of $\mathrm{Gal}(L/\mathbb{Q})$ such that $\sigma|_{K\cap L} = \tau|_{K\cap L}$ (this can be done by pigenhole principal). Then (σ, τ) geneartes a subgroup H of $\mathrm{Gal}(KL/\mathbb{Q})$ of order 2^m. Let $F = \mathrm{Fix}(H)$, then $[F : \mathbb{Q}] \leq 2^{m-1}$, hence F is cyclotomic by induction. Since $H\mathrm{Gal}(KL/L) = \mathrm{Gal}(KL/\mathbb{Q})$, we have $FL = KL$, hence FL is cyclotomic and hence K is cyclotomic.

Then we will prove that for $K/\mathbb{Q}$ cyclic of order p^m, K is cyclotomic. Let $q_1, ..., q_r$ be all the primes $q \neq p$ that ramifies in $K/\mathbb{Q}$, and make induction on r. When r=0, we have already prove that K is cyclotomic. For $r \leq 1$, let $q = q_r$, noticed that $\left[V_q^j : V_q^{j+1}\right] = 1$ or q, hence $V_q^1 = \{Id\}$. It follows that $|T_q| = p^k|q-1$ and T_q is cyclic. Let L be the unique subfild of $\mathbb{Q}(\xi_q)$ of order p^k, we will prove that LK is cyclotomic and hence conclude that K is cyclotomic. Noticed that $T_{q,LK} \subset T_{q,L} \times T_{q,K}$. Since $[LK : \mathbb{Q}]$ is a power of p, $T_{q,LK}$ is also a cyclic group, hence we must have $|T_{q,LK}| \leq p^k$. Also we have $|T_{q,LK}| = e(\mathfrak{q}_{LK}|q) \geq e(\mathfrak{q}_K|q) = |T_{q,K}| = p^k$, hence $|T_{q,LK}| = p^k$. Let K_0 be the fixed field of $T_{q,LK}$, then q is unramified in $K_0/\mathbb{Q}$, hence K_0 is cyclotomic by induction. Noticed that $L \cap K_0$ is unramified, we have $L \cap K_0 = \mathbb{Q}$, hence $[LK_0 : \mathbb{Q}] = [L : \mathbb{Q}]\,[K_0 : \mathbb{Q}] = p^k\,[K_0 : \mathbb{Q}]$. Since $[LK : \mathbb{Q}] = [LK : K_0]\,[K_0 : \mathbb{Q}] = |T_{q,LK}|\,[K_0 : \mathbb{Q}] = p^k\,[K_0 : \mathbb{Q}]$, we have $LK = LK_0$, and hence LK is cyclotomic, which completes the proof of Kronecker-Weber theorem.

9 Appendix E: Sketch of Proof for $\zeta_K(s) = \prod_{\chi} L(s, \chi)$

For an abelian number field K, by Kronecker-Weber theorem, there exist some $n \in \mathbb{Z}_{\geq 1}$, such that $K \subset \mathbb{Q}(\xi_n)$. Noticed that $\mathrm{Gal}(\mathbb{Q}(\xi_n)/\mathbb{Q}) = (\frac{\mathbb{Z}}{n\mathbb{Z}})^\times$, let G denotes the group of dirichlet characters module n, then there exists a one-one correspondence between the subgroups of G and the subfields of $\mathbb{Q}(\xi_n)$, given by $X \leftrightarrow \mathrm{Fix}(X^\perp)$

Let X be the associates group of dirichlet characters for K, if we regard $\chi \in X$ not a dirichlet character module n, but a primitive character, then we get the following filteration for X.

$$\begin{array}{ccccccc}
K & \supset & K_D & \supset & K_T & \supset & \mathbb{Q} \\
\updownarrow & & \updownarrow & & \updownarrow & & \updownarrow \\
\{Id\} & \subset & D_p & \subset & T_p & \subset & \mathrm{Gal}(K/\mathbb{Q}) \\
\updownarrow & & \updownarrow & & \updownarrow & & \updownarrow \\
X & \supset & X_D & \supset & X_T & \supset & \{\chi_0\}
\end{array}$$

where D_p and T_p are the inertia group and ramification group for a prime p, $X_D = \{\chi \in X : \chi(p) \neq 0\}$, $X_T = \{\chi \in X : \chi(p) = 1\}$.

It follows that for every prime p, we have $\prod\limits_{\mathfrak{p}|p} \frac{1}{1-(N\mathfrak{p})^{-s}} = (\frac{1}{1-p^{-f(\mathfrak{p}|p)s}})^{g(\mathfrak{p}|p)} = \prod\limits_{\chi \in X} \frac{1}{1-(\frac{\chi(p)}{p})^s}$. Hence we have $\zeta_K(s) = \prod\limits_{\chi \in X} L(s, \chi)$.

10 Appendix F: The Relative Class Number Formula

Using the arithmetic class number formula, we can compute $h_K R_K$ for every abelian number field K easily. However, if we want to compute the class number h_K, we must compute R_K in advanced. Unfortunately, it is generally not easy to compute R_K, e.g. for real quadratic field $\mathbb{Q}(\sqrt{n})$, computing R_K is equivalent solving the Pell equation $X^2 - nY^2 = 1$, which is proved to be a NP-problem (c.f. Manders and Adleman,*NP-complete decision problems for binary quadratics*,J Comput System Sci 16(1978)168-184).

However, for totally imaginary abelian number field K, there exist a method to compute a factor of h_K directly. In the followings, we let K a totally imaginary abelian number field K, $K^+ = K \cap \mathbb{R}$, h the class number of K, h^+ the class number of K^+, R the regulator of K and R^+ the regulator of K^+.

Let H and H^+ be the Hilbert class fields for K and K^+. Noticed that K/K^+ is totally ramified at archimedean primes, hence $K \cap H^+ = k^+$. Therefore, $h^+ = [H^+ : K^+] = [KH^+ : K] \,|\, [H : K] = h$, i.e. $h^+|h$. We write $h^- = \frac{h}{h^+}$, and calls it the relative class number.

By arithmetic class number formula, we have:

$$hR = \frac{|\mu(K)|}{2} \prod_{\substack{\chi\neq\chi_0 \\ \chi\in G_0}} \sum_{b=1}^{f_\chi} \frac{-\overline{\chi}(b)}{2} \ln|\sin(\frac{\pi b}{f_\chi})| \prod_{\chi\notin G_0} \sum_{b=1}^{f_\chi} \frac{-\overline{\chi}(b)b}{2f_\chi}$$

$$h^+R^+ = \prod_{\substack{\chi\neq\chi_0 \\ \chi\in G_0}} \sum_{b=1}^{f_\chi} \frac{-\overline{\chi}(b)}{2} \ln|\sin(\frac{\pi b}{f_\chi})|$$

Hence we have:

$$h^- = \tfrac{h}{h^+} = \tfrac{R^+}{R}\tfrac{|\mu(K)|}{2} \prod_{\chi\notin G_0} \sum_{b=1}^{f_\chi} \tfrac{-\overline{\chi}(b)b}{2f_\chi}$$

Let U, U^+ be the unit groups of K and K^+, let $Q = [U : \mu(K)U^+]$. In the followings, we will prove that $Q = 1$ or 2 and $\frac{R}{R^+} = \frac{2^r}{Q}$ with $r = \frac{1}{2}[K : \mathbb{Q}] - 1$, and conclude that $h^- = Q|\mu(K)| \prod_{\chi\notin G_0} \sum_{b=1}^{f_\chi} \frac{-\overline{\chi}(b)b}{4f_\chi}$.(The relative class number is also important, e.g. it can be proof that $p|h(\mathbb{Q}(\xi_p))$ if and only if $p|h^-(\mathbb{Q}(\xi_p))$, hence p is a regular prime if and only if $p|h^-(\mathbb{Q}(\xi_p))$ c.f. Washington, *Introduction to Cyclotomic Fields*)

Now we prove that $Q = 1$ or 2. Clearly, $\varepsilon/\bar{\varepsilon} \in \mu(K)$ for $\varepsilon \in U$. Consider the map $\phi : E \to \frac{\mu(K)}{\mu(K)^2}, \varepsilon \mapsto \varepsilon/\bar{\varepsilon} + \mu(K)^2$. Noticed that $\varepsilon \in \mathrm{Ker}(\phi)$ if and only if $\varepsilon = -\bar{\varepsilon}\xi^2$ for some $\xi \in \mu(K)$, if and only if $\varepsilon\xi = -\bar{\varepsilon}\bar{\xi}$ for some $\xi \in \mu(K)$(i.e. $\varepsilon\xi \in U^+$) if and only if $\varepsilon \in \mu(K)U^+$. Hence $\mathrm{Ker}(\phi) = \mu(K)U^+$, and we conclude that $Q = [U : \mu(K)U^+] \,|\, [\mu(K) : \mu(K)^2] = 2$, i.e. $Q = 1$ or 2.

Then, we prove that $\frac{R}{R^+} = \frac{2^r}{Q}$. For an independent subset $\varepsilon_1, ..., \varepsilon_r$ of U, let $R_K(\varepsilon_1, ..., \varepsilon_r)$ denotes the regulator with respect to $\varepsilon_1, ..., \varepsilon_r$. Now let $\varepsilon_1, ..., \varepsilon_r$ be the fundamental unit for K^+, then it forms a independent subset of U. It is clearly that $R_K(\varepsilon_1, ..., \varepsilon_r) = 2^r R_{K^+}(\varepsilon_1, ..., \varepsilon_r) = 2^r R^+$, hence we only needs to prove that $\frac{R}{R_K(\varepsilon_1,...,\varepsilon_r)} = Q$.

Let $\eta_1, ..., \eta_r$ be the fundamental unit for K, then $\varepsilon_j = (\prod\limits_{i=1}^{r} \eta_i^{a_{i,j}})\xi_j$ with $\xi_j \in \mu(K)$. Then $\log|\sigma_k(\varepsilon_j)| = \sum\limits_{i=1}^{n} a_{i,j} \log|\sigma_k(\eta_i)|$, hence $\frac{R}{R_K(\varepsilon_1,...,\varepsilon_r)} = |\det(a_{i,j})|$. By the structure of abelian group, we have $Q = [U : \mu(K)U^+] = |\det(a_{i,j})|$, hence $\frac{R}{R_K(\varepsilon_1,...,\varepsilon_r)} = Q$.

11 Reference

[1] Helmut Hasse, *Über die Klassenzahl abelcher Zahlkörper* (translated by Mikihito Hirabayashi), (1952), first part of this book

[2] David Hilbert, *Die Theorie der algebraischen Zahlkörper* (translated by Iain T.Adamson), (1897), §22-§26,§100-§104 of this book.

[3] Erich Hecke, *Ueber die Zetafunktion beliebiger algebraischer Zahlkörper*, (1917), about the functional equation for Dedekind-Zeta function

[4] Erich Hecke, *Über die L-Funktionen und den Dirichletschen Primzahlsatz für einen beliebigen Zahlkörper*,Nachr. Kgl. Gel. d. Wissensch. Göttingen (1917), about the decomposition law of the class field for K described by the character group X, and the discriminant-conductor formula.

[5] M.J. Greenberg, *An elementary proof of the Kronecker-Weber theorem*, American Math. Monthly, 81 (1974), 601–607, essentially a modern version of the proof in Hilbert's book.

[6] Lawrence C. Washington, *Introduction to Cyclotomic Fields*,(1982). Chapter 4 of this book

[7] Serge Lang, *Algebraic Number Theory*,(1986), Chapter VI and Chapter XIII of this book.

[8] Jürgen Neukirch.*Algebraic Number Theory* (translated by Norbert Schappacher),(1992), Chapther VII§1-§5 of this book and some basic facts about algebraic number theory and class field theory.

An Introduction To An Algorithm Factoring Numbers With Elliptic Curves

Song Dianyi; Yu Penghao

June 2022

1 Abstract

In this article we are going to introduce a new algorithm factoring integers proposed by H.W.Lenstra. It is a method depending on the use of elliptic curves, and it is faster than existing methods when the number has smaller prime divisors.

2 Introduction

This article is divided into three major parts. The first part is devoted to show the basic properties of elliptic curves, and the second part will introduce the structure of the algorithm, then the third part will give an analysis of the method, including its success probability as well as its efficiency.

3 Basics of Elliptic Curves

In this article, we denote by F_p a finite field with cardinality of p, and by A^* the group of units of a ring A with 1.

(3.1) An elliptic curve over field K is a pair of elements $a, b \in K$ with $4a^3 + 27b^2 \neq 0$. These elements are thought of as the coefficients in the Weierstrass equation

$$y^2 = x^3 + ax + b$$

We denote such elliptic curve (a, b) by $E_{a,b}$, or simply by E. The set of points $E(K)$ of such an elliptic curve over K is defined by

$$E(K) = \{(x : y : z) \in P^2(K) : y^2z = x^3 + axz^2 + bz^3\}$$

Here $P^2(K)$ denotes the projective plane over K, and $(x : y : z)$ denotes the equivalence class containing (x, y, z).

Let E be an elliptic curve over K, then the zero point of the curve is the point $(0 : 1 : 0)$, denoted by O. The other points, $(x : y : 1)$, where $x, y \in K$ satisfy the Weierstrass equation $y^2 = x^3 + ax + b$.

The set $E(K)$ has the structure of an abelian group with additive group law, which is defined as follows. First, O is the zero element satisfying $O + P = P + O = P$ for all $P \in E(K)$. For two non-zero points $P = (x_1 : y_1 : 1)$ and $Q = (x_2 : y_2 : 1)$, $P + Q = O$ if and only if $x_1 = x_2$ and $y_1 = -y_2$. Otherwise, let $\lambda \in K$ be determined by $\lambda = \frac{(y_1 - y_2)}{x_1 - x_2}$ if $P \neq Q$ and $\lambda = \frac{(3{x_1}^2 + a)}{2y_1}$ if $P = Q$, and then let $\nu = y_1 - \lambda x_1$. Then $P + Q = R = (x_3 : y_3 : 1)$, where $x_3 = \lambda^2 - x_1 - x_2$ and $y_3 = -\lambda x_3 - \nu$. Such operation can be easily proven by Vieta's Theorem to be well defined.

(3.2) For two elliptic curves $E = E_{a,b}$ and $E' = E_{a',b'}$ defined over K, an isomorphism $E \to E'$ is defined to be an element $u \in K^*$ satisfying both $a' = u^4 a$ and $b' = u^6 b$. Any isomorphism $u : E \to E'$ induces an isomorphism $E(K) \to E'(K)$ of the abelian groups that sends $(x : y : z)$ to $(u^2 x : u^3 y : z)$, denoted by u as well.

An automorphism of an elliptic curve E over K is an isomorphism $E \to E$. The set of automorphisms of E is a subgroup of K^*, denoted by $AutE$ or $Aut_K E$. And it can be easily calculated that $\#AutE$ must be 2 or 4 or 6.

(3.3) The number of elliptic curves over F_p, namely the number of pairs $(a, b) \in F_p \times F_p$ with $4a^3 + 27b^2 \neq 0$, can be easily calculated to be $p^2 - p$.

(3.4) For any elliptic curve E over $\mathbf{F_p}$ we have by a theorem of Hasse

$$\#E(\mathbf{F_p}) = p + 1 - t \quad \text{with } -2\sqrt{p} \leq t \leq 2\sqrt{p}$$

Conversely, if p is a prime greater than 3 and t an integer satisfying $|t| < 2\sqrt{p}$. Then the weighted number of elliptic curves E over $\mathbf{F_p}$ with $\#E(\mathbf{F_p}) = p+1-t$ up to isomorphism is given by a formula

$$\#'\{E : E\, elliptic\, curve\, over\, \mathbf{F_p}, E(\mathbf{F_p}) = p + 1 - t\}/ \cong_{F_p} = H(t^2 - 4p)$$

(3.5) Then we use (3.3) to count the set

$$\{E : E \text{ elliptic curve over } F_p\}/{\cong_{F_p}}$$

of isomorphism classes of elliptic curves over F_p. Since the number of elliptic curves isomorphic to a given elliptic curve E is $\#{F_p}^*/\#AutE = (p-1)/\#AutE$, summing over the representatives of the isomorphism classes and dividing by $(p-1)$ we get

$$\sum_E \frac{1}{\#AutE} = p$$

We express this by writing

$$\#'\{E : E \text{ elliptic curve over } F_p\}/{\cong_{F_p}} = p.$$

In similar expressions, the notation $\#'$ denotes the weighted cardinality, the isomorphism class of E being weighted $(\#AutE)^{-1}$.

(3.6) In this part some properties of binary quadratic forms will be introduced.

Let Δ be a negative integer satisfying $\Delta \equiv 0$ or $1 (mod 4)$. A positive definite integral binary quadratic form of discriminant Δ, or briefly a form, is a polynomial $F = aX^2 + bXY + cY^2$ with $a, b, c \in \mathbf{Z}, a > 0, b^2 - 4ac = \Delta$.

An isomorphism from a form $F = aX^2 + bXY + cY^2$ to a form $F' = a'X'^2 + b'X'Y' + c'Y'^2$ can be expressed by a matrix $\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$ with $\alpha, \beta, \gamma, \delta \in \mathbf{Z}$, $\alpha\gamma - \beta\delta = 1$. In fact, for a better understanding ,we may take $X' = \alpha X + \beta Y$ and $Y' = \gamma X + \delta Y$.

Using some knowledge of linear algebra, the set of automorphisms of a form F is a subgroup of the group $\mathbf{SL_2(Z)}$ with integral entries and determinant 1; such subgroup is denoted by Aut F. It can be easily shown that $AutF$ is a cyclic group of order 2 or 4 or 6.

For fixed Δ, the set of equivalence classes of forms of discriminant Δ is finite, and the Kronecker class number H(Δ) of Δ is defined to be the weighted cardinality of the set defined as follows:

$$H(\Delta) = \#'\{F : F \text{ is a form of discriminant } \Delta\}/\sim$$

with $\sim$ denoting equivalence and the meaning of $\#'$ being making sums with weight, where the equivalence class containing F being counted with weight $(\#AutF)^{-1}$, similiar to the definition in (3.5). It is not hard to show that H(Δ) > 0.

A primitive form $F = aX^2 + bXY + cY^2$ is a form with $gcd(a, b, c) = 1$. h(Δ) denotes the weighted cardinality of the set of equivalence classes of primitive forms of discriminant Δ. By sorting forms with $gcd(a, b, c)$, it is easy to see that

$$H(\Delta) = \sum_d h(\Delta/d^2)$$

for all d satisfying $d|\Delta$ and $\Delta/d^2 \equiv 0$ or $1 (\text{mod } 4)$. The largest of such d is called the conductor f of Δ,and $\Delta_0 = \Delta/f^2$ is the fundamental discriminant associated to Δ; it can be shown that the $d's$ in the above summation are exactly the positive divisors of f.

4 The Structure of The Factoring Algorithm

To unitize the notation in this section, We call a divisor d of a positive integer n non-trivial if $1 < d < n$. And in this section we will describe the structure of the factoring algorithm attempting to find a non-trivial divisor of a positive integer n, or affirm it to be a prime number.

(4.1) To describe the algorithm one needs the definition of Elliptic curve modulo n, and in this case n is a positive integer (that is not necessarily a prime number).

Consider the set of all triples $(x, y, z) \in (Z/nZ)^3$ for which x, y, z generates the unit ideal of Z/nZ, with the group of unit $(Z/nZ)^*$ acting on this set by $u(x, y, z) = (ux, uy, uz)$. The orbit of (x, y, z) is denoted by $(x : y : z)$, and the set of all such orbits by $P^2(Z/nZ)$.

For $a, b \in Z/nZ$, the cubic curve $E = E_{a,b}$, defined similarly to the one in (3.1), is defined over Z/nZ by the equation

$$y^2 = x^3 + ax + b$$

The set of points $E(Z/nZ)$ of such curve over Z/nZ is defined by

$$E(Z/nZ) = \{(x : y : z) \in P^2(Z/nZ) : y^2z = x^3 + axz^2 + bz^3\}$$

If $6(4a^3 + 27b^2) \in (Z/nZ)^*$ then E is called an elliptic curve over Z/nZ.

For general n, there is a partially defined "pseudo-addition" operation on a subset of $E(Z/nZ)$ defined in the following part. For notations, we denote the point $(0 : 1 : 0)$ of $P^2(Z/nZ)$ by O, and we denote by V_n the subset of $P^2(Z/nZ)$ defined as follows:

$$V_n = \{(x : y : 1) : x, y \in (Z/nZ)\} \cup \{O\}$$

For $P \in V_n$ and a prime p dividing n we denote by P_p the point of $P^2(F_p)$ obtained by reducing the coordinates x, y of P modulo p. And it is easily observed that $O_p = P_p$ if and only if $P = O$.

(4.2) In this part the algorithm performing "pseudo-addition" will be presented, and this part of the algorithm will be frequently used in the whole algorithm.

Given $n \in Z_{>1}, a \in Z$ and $P, Q \in V_n$, the algorithm will either calculate a non-trivial divisor d of n, or determines a point $R \in V_n$ with the following property: if p is any prime dividing n and satisfies that there exists $b \in F_p$ such that

$$6(4\bar{a}^3 + 27b^2) \neq 0 \quad \text{for } \bar{a} = (a \text{ mod } p),$$
$$P_p \in E_{\bar{a},b}(F_P), \quad Q_p \in E_{\bar{a},b}(F_P),$$

Then $R_p = P_p + Q_p$ in the group $E_{\bar{a},b}(F_P)$, with the addition defined in (3.1).

Note that the application of the algorithm does not require n to be a composite number, nor do we need to know the prime divisor of n beforehand.

When calculating, if $P = O$ put $R = Q$ and stop, or if $P \neq O$ and $Q = O$ put $R = P$ and stop, these are the trivial cases. Then in the remaining case $P \neq O$, $Q \neq O$, let $P = (x_1 : y_1 : 1)$ and $Q = (x_2 : y_2 : 1)$. Then use the Euclidean algorithm to calculate the value of $gcd(x_1 - x_2, n)$. If this gcd is not 1 or n, denote it by d and stop. If $gcd(x_1 - x_2, n) = 1$ then the Euclidean algorithm also gives the value of $(x_1 - x_2)^{-1}$; in this case put

$$\lambda = (y_1 - y_2)(x_1 - x_2)^{-1},$$
$$x_3 = \lambda^2 - x_1 - x_2, \quad y_3 = \lambda(x_1 - x_3) - y_1$$
$$R = (x_3 : y_3 : 1)$$

and stop. Finally in the case that $gcd(x_1 - x_2, n) = n$, so that $x_1 = x_2$ in Z/nZ. Calculate $gcd(y_1 + y_2, n)$. If it is not 1 or n, denote it by d and stop. If it is n (so that $y_1 = -y_2$), put $R = O$ and stop. If $gcd(y_1 + y_2, n) = 1$, put

$$\lambda = (3{x_1}^2 + a)(y_1 + y_2)^{-1},$$
$$x_3 = \lambda^2 - x_1 - x_2, \quad y_3 = \lambda(x_1 - x_3) - y_1$$
$$R = (x_3 : y_3 : 1)$$

and stop. This finishes the description of this part of the algorithm. And its correctness can be checked by the formulae stated in (3.1), as they go through similar process.

(4.3) In this part the algorithm performing multiplication (a number in Z^+ multiplying a point on the curve) will be introduced.

By repeating the algorithm of addition presented in (4.2), an algorithm of multiplication can accomplish the following. Given $k \in Z^+, n \in Z_{>1}, a \in Z/nZ$ and $P \in V_n$, it either calculates a non-trivial divisor d of n, or it determines a point $R \in V_n$ with the following property: if p is any prime dividing n and satisfies that there exists $b \in F_p$ such that

$$6(4\bar{a}^3 + 27b^2) \neq 0 \quad \text{for } \bar{a} = (a \bmod p),$$
$$P_p \in E_{\bar{a},b}(F_P),$$

then $R_p = k \cdot P_p$ in the group $E_{\bar{a},b}(F_P)$. If this algorithm determines a point R with the stated property, then it is denoted by kP.

If k is given as $k = k_1k_2$, then one can calculate kP by $kP = k_1(k_2P)$. Suppose that k is given as a product

$$k = \prod r^{e(r)},$$

where r ranges over a certain set of positive integers with $e(r)$ being positive integers. It can be easily seen that to multiply a point P by k times it suffices to perform $e(r)$ multiplications by r for each r. To make the proof in the appendix stand, we assume that the multiplications by r are performed in increasing order of r.

Remark: To calculate $r \cdot P$, a good way is by using the binary representation of r, which takes the time of $O(log(r)M(n))$, with $M(n)$ being the time of performing one round of addition.

(4.4) In this part we will introduce the algorithm factoring with elliptic curves with operations stated in (4.2) and (4.3).

(4.4.1) When factoring with one curve, let $n, v, w \in Z_{>1}$ and $a, x, y \in Z/nZ$ be given. An algorithm attempting to find a non-trivial divisor d of n is described below.

For each integer $r \geq 2$, denote by $e(r)$ the largest integer m with $r^m \leq v + 2\sqrt{v} + 1$, and put

$$k = \prod_{r=2}^{w} r^{e(r)}.$$

Let $P = (x : y : 1) \in V_n$, calculate kP with the algorithm in (4.3), If this attempt fails then a non-trivial divisor d of n is found, and the algorithm halts. If kP is calculated successfully then the algorithm halts as well, with the message that it fails to find a non-trivial divisor of n.

(4.4.2) The whole structure of the algorithm is based on (4.4.1), which is mainly applying (4.4.1) repeatedly on different curves. Let $n, v, w, h \in Z_{>1}$ be given, a probabilistic algorithm attempting to find a non-trivial divisor d of n will be described in this part.

First we suppose that the random number generator used in this algorithm can draw the triple $(a, x, y) \in (Z/nZ)^3$ with equal probability given to each triple, and that successive calls to the random number generator are independent.

Then for one round, draw three elements $a, x, y \in Z/nZ$ at random, and apply the algorithm (4.4.1) to n, v, w, a, x, y (the notations are the same as in (4.4.1)). If the result is a non-trivial divisor d of n, halt the whole calculation with the result found. Otherwise, repeat the operation in the next round from drawing three elements $a, x, y \in Z/nZ$ again. The algorithm will halt once a non-trivial divisor of n is found, otherwise it will halt when it has already been applied h rounds.

(4.5) For a better understanding of the algorithm, the number v should be thought of as an upper bound of the divisor d one is trying to find, the parameter w is corresponds to the time one is willing to spend on a single curve, and h is the number of curves that one tries.

It should be noted that the success probability of the algorithm is not 1. It is a function of w and h, which increases as either of w or h increases, with the optimal choice discussed in the proofs in the Appendix.

The time efficiency, which corresponds to the success probability, can be stated as follows:

(4.5.1)There is a function $K : R_{>0} \to R_{>0}$ with

$$K(x) = e^{\sqrt{(2+o(1))logxloglogx}} \quad \text{for } x \to \infty$$

such that the following is true. Let $n \in Z_{>1}$ be an integer that is not a prime power nor divisible by 2 or 3, and let g be any positive integer. Then algorithm (4.4.2), when applied with suitable values for v, w, h, can be used to find, with the success probability at least $1 - e^{-g}$, a non-trivial divisor of n within time

$$gK(p)M(n)$$

where p denotes the least prime divisor of n and $M(n) = O((logn)^2)$ being the time needed to perform a single operation of addition on a curve.

It can be easily noted that the excluded cases in the statement are easy to be checked within a far smaller amount of time as $n \to \infty$, so the algorithm can actually be applied to any positive integer n. The proof of such statement, as well as the suitable choice of the parameters, will be left to the appendix.

5 Appendix1: The Preparation of Mathematical Knowledge

(5.1) Recall the formula $H(\Delta) = \Sigma_d h(\Delta/d^2)$ and the notations defined in (3.6).

The quadratic character $\chi : \mathbf{Z} \to \{0, 1, -1\}$ associated to Δ is defined by

$$\chi(l) \equiv \Delta^{(l-1)/2}(mod\ l), \chi(l) \in \{0, 1, -1\}\, if\ l\ is\ an\ odd\ prime$$

$$\chi(2) = 0, 1, -1 \text{ for } \Delta \equiv 0(\text{mod } 4)\ 1(\text{mod } 8)\ 5(\text{mod } 8)$$

$$\chi(mn) = \chi(m)\chi(n)$$

With the analytic class number formula for h(Δ), we have

$$h(\Delta) = \frac{\sqrt{-\Delta}}{2\pi} L(1, \chi),\ where\, L(s, \chi) = \sum_{n=1}^{\infty} \frac{\chi(n)}{n^s} \text{ for s} \in \mathbf{C}, \text{Re(s)} > 0$$

Recall that $\Delta_0 = \Delta/f^2$, then by induction we can obtain a formula

$$L(1, \chi) = L(1, \chi_0) \prod_{l|f} (1 - \frac{\chi_0(l)}{l})$$

with l ranging over the primes dividing f and χ_0 being the character associated to Δ_0. Combining the formulae with $H(\Delta) = \Sigma_d h(\Delta/d^2)$, we could have

$$H(\Delta) = \frac{\sqrt{-\Delta}}{2\pi} L(1, \chi_0)\psi(f)$$

with $\psi : \mathbf{Z}^* \to \mathbf{R}$ defined by

1. $\psi(l^k) = \frac{l-l^{-k}}{l-1}, l, \frac{l+1-2l^{-k}}{l-1}$, if l is prime, $k \geq 1$ and $\chi_0(l) = 0, 1, -1$
2. $\psi(mn) = \psi(m)\psi(n)$, if $\gcd(m, n) = 1$

Let $\phi(f)$ be Euler's function, it can be shown that $1 \leq \psi(\mathrm{f}) \leq (\mathrm{f}/\phi(\mathrm{f}))^2 = \mathrm{O}(\log(\log\mathrm{f}))^2$

Furthermore, it can be shown that $L(1, \chi_0) = O(log \, \|\Delta_0\|)$. The proofs of both of the inequalities above are beyond our reach. And applying the theorem given in the book of K.Prachar, we can find that there exists a positive effectively computable constant c_1 such that for all $z \in \mathbf{Z_{>1}}$,there exists $\Delta^* < -4$ with the property that

$$L(1, \chi_0) \geq \frac{c_1}{log \, z} \text{ if } |\Delta_0| \leq \mathrm{z}, \Delta_0 \neq \Delta*$$

(5.2) Proposition: Directly following from the inequalities above, we have the inequality that there exists positive constants c_2, c_3 such that for each $z \in \mathbf{Z_{>1}}$,there exists $\Delta^* = \Delta^*(z) < -4$ such that

$$c_2 \frac{\sqrt{-\Delta}}{log \; z} \leq H(\Delta) \leq c_3 \sqrt{-\Delta} \cdot \log|\Delta| \cdot (\log\log|\Delta|)^2,$$

which holds for all $\Delta \in \mathbf{Z}$ with $-z \leq \Delta < 0, \Delta \equiv 0$ or $1(mod \; 4)$, with notice that the left inequality may be invalid if $\Delta_0 = \Delta^*$. ■

(5.3) Proposition: There exist effectively computable positive constants c_4, c_5 such that for each prime number $p > 3$, the following two statements are valid:

(a) If S is a set of integers s with $|s - p - 1| \leq 2\sqrt{p}$, then

$$\#'\{E : E \; elliptic \; curve \; over \; \mathbf{F_p}, \#E(\mathbf{F_p}) \in S\}/ \cong_{F_p} \leq c_4 \cdot \#S \cdot \sqrt{p} \cdot \log(p) \cdot (\log(\log(p)))^2$$

(b)If S is a set of integers s with $|s - p - 1| \leq \sqrt{p}$, then

$$\#'\{E : E \; elliptic \; curve \; over \; \mathbf{F_p}, \#E(\mathbf{F_p}) \in S\}/ \cong_{F_p} \geq c_5(\#S - 2) \cdot \sqrt{p}/\log(p)$$

Proof We know from the conclusion of (3.4) that $\#'\{E : E \; elliptic \; curve \; over \; \mathbf{F_p}, \#E(\mathbf{F_p}) \in S\}/ \cong_{F_p} = \sum_{s \in S} H((p + 1 - s)^2 - 4p)$

Proof of (a) take $z = 4p$ and (5.2), we have $\#'\{E : E \; elliptic \; curve \; over \; \mathbf{F_p}, \#E(\mathbf{F_p}) \in S\}/ \cong_{F_p} = \sum_{s \in S} H((p + 1 - s)^2 - 4p) \leq \sum_{s \in S} \cdot c_3 \cdot \sqrt{4p - (p + 1 - s)^2} \cdot \log(4p - (p + 1 - s)^2) \cdot (\log\log(4p - (p + 1 - s)^2))^2 \leq c_4 \cdot \#S \cdot \sqrt{p} \cdot \log(p) \cdot (\log(\log(p)))^2$.

Proof of (b) Also,we take $z = 4p$. It suffices to show that there are at most two integers t, $|t| \leq \sqrt{p}$, for which the fundamental discriminant associated to $t^2 - 4p$ equals Δ^*. In summation, at most 2 "s" is not suitable. Then we have $\#'\{E : E \; elliptic \; curve \; over \; \mathbf{F_p}, \#E(\mathbf{F_p}) \in S\}/ \cong_{F_p} = \sum_{s \in S} H((p + 1 - s)^2 - 4p) \geq \sum_{\text{suitable s} \in \text{S}} H((p + 1 - s)^2 - 4p) \geq \sum c_2 \frac{\sqrt{4p-(p+1-s)^2}}{\log(4p)} \geq c_5(\#S - 2) \cdot \sqrt{p}/\log(p)$. ■

(5.4) Modular curves. (This part is beyond reach of our knowledge as well, so we keep faithful to the statements given by Lenstra and assume them to be valid in order to continue the proof) We wish to estimate the weighted number of elliptic curves E over $\mathbf{F_p}$ for which $\#E(\mathbf{F_p})$ is divisible by a given prime number l. For this purpose some results about the modular curves $\chi(l)$ and $\chi_1(l)$ will be shown.

Let p be a prime number, $p > 3$ and l a prime number different from p.We can consider pairs (E, P) to analogy the isomorphism of E, which consists of an elliptic curve E over F_p and a point P$\in E(F_P)$ of order l. Two such pairs (E, P) and (E', P') are said to be equivalent over F_p if there exists an isomorphism $u : E \rightarrow E'$ over F_p that maps P to P'. We denote the set of equivalence classes by $Z_1(l)(F_p)$. But if u is allowed to be in the algebraic closure $\bar{F_p}$ of F_p rather than in F_p, a map $E(\bar{F_p}) \to E'(\bar{F_p})$ is also defined. So we obtain the definition of equivalence over $\bar{F_p}$. The set of classes of this equivalence relation is denoted by $Y_1(l)(F_p)$. There is an obvious surjective map $Z_1(l)(F_p) \to Y_1(l)(F_p)$

If C is a complete non-singular irreducible curve of genus g over F_p then by Weil's inequality the cardinality of the set $C(F_p)$ of points of C over F_p satisfies

$$|\#C(F_p) - p - 1| \leq 2g\sqrt{p}$$

Applying this to C=$\chi_1(l)$, using the properties of modular curves, it can be obtained that

$$\#Y_1(l)(F_p) = p + O(l^2 p^{1/2}) \tag{1}$$

Applying Weil's inequality to $C = \chi(l)$, we can find by using properties of modular curve that

$$\#Y(l)(F_p) = p + O(lp^{1/3}) \tag{2}$$

Remark: The proof of both of the above statements are beyond reach of our knowledge, so we put them here without giving a proof.

(5.5) Proposition: Let p, l be primes, $p > 3,\ l \neq p$.

(a) Let E be an elliptic curve over F_p and $P \in E(F_p)$ a point of order l. The subgroup of all $u \in Aut_{F_p}E$ that send P to P is denoted by $A_{E,P}$. Then the number of elements of $Z_1(l)(F_p)$ that map to the class of (E, P) in $Y_1(l)(F_p)$ equals $\#A_{E,P}$

(b) If $p \equiv 1(mod\ l)$, with a primitive l^{th} root of unity $\xi \in F_p$ being chosen. Let E be an elliptic curve over F_p ,and $Q, P \in E(F_P)$ are points of order l satisfying $e_l(P, Q) = \xi$, where e_l denotes the Weil Pairing. Denote by $A_{E,P,Q}$ the subgroup $A_{E,P} \cap A_{E,Q}$ of $Aut_{F_p}(E)$.Then the number of elements of $Z(l)(F_p)$ that map to the class of (E, P, Q) in $Y(l)(F_p)$ equals $\#A_{E,P,Q}$.

Proof of (a) Let E be given by a, b, and let $P = (x : y : 1)$. If E', P' is another such pair, given by a', b', x', y', then (E, P) and (E', P') correspond to the same element of $Y_1(l)(F_p) \iff (a', b', x', y') = (u^4a, u^6b, u^2x, u^3y)$ for some $u \in \bar{F_p}^*$, and to the same element of $Z_1(l)(F_p) \iff u$ can be take in F_P^*. It follows that the number of elements of $Z_1(l)(F_p)$ mapping to the class of (E, P) equals index $[B_{E,P} : C_{E,P}]$, where the subgroups $B_{E,P}, C_{E,P}$ of $\bar{F_p^*}$ are defined by:

$$B_{E,P} = \{u \in \bar{F_p^*} : \{u^4a, u^6b, u^2x, u^3y\} \subset F_p\}$$

$$C_{E,P} = \{u \in \bar{F_p^*} : (u^4a, u^6b, u^2x, u^3y) = (v^4a, v^6b, v^2x, v^3y) \text{ for some v} \in \mathrm{F_p^*}\}$$

For $B_{E,P}$, note that for $u \in \bar{F_p^*}$, $u^4a \in F_p \iff (u^4a)^p = u^4a$, and similarly with u^6b, u^2x, u^3y; hence the map that sends u to u^{p-1} maps $B_{E,P}$ onto the group $\bar{A}_{E,P}$ of all $u \in Aut_{\bar{F_p}}(E)$ sending P to P. It is obvious that the kernel is F_p^*, so that $\#E_{E,p} = \#\bar{A}_{E,P} \cdot \#F_p^*$.

It can be shown that $C_{E,P}$ is generated by F_p^* and $\bar{A}_{E,P}$, so that $\#C_{E,P} = \#F_p^* \cdot \#A\bar{}_{E,P}/\#(\bar{A}_{E,P} \cap F_p^*)$, and note that $\bar{A}_{E,P} \cap F_p^*$ is just $A_{E,P}$.

This proves (a). Although the operations with Weil pairing are beyond reach of our knowledge, we will be faithful to the author's understanding of the correctness of (b). ■

Remark: With the results obtained in the propositions above, we can obtain the following results successively, and the final results stated in (5.8) will be directly used in analyzing the algorithm.

(5.6) Proposition: Let p, l be primes, $p > 3, l \neq p$. Then the number

$$\#'\{\mathrm{E : E\ elliptic\ curve\ overF_p, \#E(F_p) \equiv 0(mod\ l)}\}/\cong_{\mathrm{F_p}}$$

equals

$$\frac{p}{l^2-1} + O(lp^{1/2}),\ if\ p \equiv 1(\mathrm{mod\ l})$$

,

$$\frac{p}{l-1} + O(lp^{1/2}),\ otherwise$$

Note: This property gives a "probability" in random selection of elliptic curve. Recall that $\#'\{E : E$ elliptic curve over $F_p\}/\cong_{F_p} = p$. Dividing the corresponding terms on both sides of the equation, then $\#E(F_p) \equiv 0$ (mod l) tends to $l/(l-1)$ and $l/(l^2-1)$ in these condition as above.

Proof: Let Y_1, Z_1 denote $Y_1(l)(F_p), Z_1(l)(F_p)$ defined in (5.4). Similarly we can use Y, Z to denote $Y(l)(F_p), Z(l)(F_p)$. A theorem in reference[2] says that the group $E(F_p)[l] = \{P \in E(F_p) : lP = O\}$ has order l or l^2. Then we can suppose W to be the set of isomorphism classes of elliptic curves E over F_p with $\#E(F_p) \equiv 0$ (mod l). W can be written as $W = W_1 \cup W_2$,with W_i consisting of the classes of those E with $\#E(F_p)[l] = l^i$, so $W_2 = \varnothing$ unless $p \equiv 1$(mod l).

The map $Z_1 \to W$ mapping the class of (E, P) to the class of E is clearly surjective. (E, P) and (E', P') map to the same element $\iff$ P and P' belong to the same orbit of $Aut_{F_p}E$; also, the size of the orbit is exactly the index $[Aut_{F_p}E : A_{E,P}] = \#Aut_{F_p}E/\#A_{E,P}$, with $A_{E,P}$ defined in (5.5). For a fixed E, we use the orbit summation, we have

$$\sum_P \frac{\#Aut_{F_p}E}{\#A_{E,P}} = l^i - 1$$

Then dividing $\#Aut_{F_p}E$ and summing over E in Z_1 we have

$$\sum \frac{1}{\#A_{E,P}} = (l-1) \cdot \#'W_1 + (l^2-1)\#'W_2$$

By(5.5), Z is a "fiber" of Y, and the left-hand sum add $\frac{1}{\#A_{E,P}}$ for $\#A_{E,P}$ times, then the left-hand sum equals $\#Y_1$, using formula (2) in (5.5) it can be shown that

$$(l-1) \cdot \#'W_1 + (l^2-1) \cdot \#'W_2 = p + O(l^2\sqrt{p}) \tag{3}$$

If p $\not\equiv 1$(mod l),then this means that

$$(l-1)\#'W = p + O(l^2\sqrt{p}),$$

For the second equation, similarly use (5.5)(b) with (P, Q) being a Weil pair, we will know

$$\sum_{(P,Q)} \frac{\#Aut_{F_p}E}{\#A_{E.P,Q}} = l(l^2-1)$$

In the same way we can get $\sum_Z \frac{1}{\#A_{E.P,Q}} = l(l^2-1) \cdot \#'W_2$, similarly by using equation (2), we get $l(l^2-1)\#'W_2 = p + O(l^3\sqrt{p})$. Hence, solving a linear equation in two variables, we have

$$\#'W = \#'W_1 + \#'W_2 = \frac{1}{l-1}((l-1)\#'W_1 + (l^2-1)\#'W_2) - \frac{1}{l^2-1}(l(l^2-1) \cdot \#'W_2)$$

$$= (\frac{1}{l-1} - \frac{1}{l^2-1}) + O(l\sqrt{p})$$

■

(5.7) Proposition: Now we give some bound to be used in the analysis of the algorithm. There exists c_6 such that for all pairs of prime p, l with $p > 3$ we have

$$\#'\{E : E\ elliptic\ curve\ over F_p, \#E(F_p) \not\equiv 0(\text{mod l})\}/\cong_{F_p} \geq c_6 p$$

.

Proof: We only need to minus the inappropriate situation to show that they do not exceed the bound. The left hand side is $((l-2)/(l-1))p + O(l\sqrt{p})$ if p$\not\equiv 0, 1$ (mod l), and $((l^2-l-1)/(l^2-1))p + O(l\sqrt{p})$ if p $\equiv$ 1(mod l). Let c_7 be an appropriate coefficient, satisfying that when $l \leq c_7 p$, the proposition is correct.

Using (5.3)(a) on the set $S = \{s \in \mathbf{Z} : |s - p - 1| \leq 2\sqrt{p}, s \equiv 0 \text{mod l}\}$, which has cardinality $O(1 + \sqrt{p}/l)$, Then we only have the cases of p satisfying $p \leq c_8$ or $l \geq c_9(\log p)(\log\log P)^2 > c_7\sqrt{p}$ remaining to be discussed. But in either of these cases, p is bounded, thus showing the suitable constant c_6 exists. ■

(5.8) Proposition: There is a positive effectively computable constant c_{10} such that for every prime number $p > 3$ the following two statements are valid.

(a) If S is a set of integers s with $|s - p - 1| \leq \sqrt{p}$, then the number of triples $(a, x, y) \in F_p^3$ for which

$$4a^3 + 27b^2 \neq 0, \#E_{a,b}(F_p) \in S,$$

where $b = y^2 - x^3 - ax$, is at least $\frac{c_{10}(\#S-2)p^{\frac{5}{2}}}{\log(p)}$

(b) If l is any prime number,then the number of triples $(a, x, y) \in F_p^3$ for which $4a^3 + 27b^2 \neq 0$, $\#E_{a,b}$ $(F_p)$$\neq$ 0 (mod l), where b = y^2 - $x^3 - ax$, is at least c_{10} p^3 .

This proposition is simple application of the proposition above. Consider (a, b, x, y) with (a, b) denoting elliptic curves and (x, y) denoting a point $(x : y : 1)$ on the elliptic curve. There are at most $(p-1)/\#AutE$ pairs of (a, b), and each $E_{a,b}$ corresponds to $\#E_{a,b}(F_p) - 1$ points $(x : y : 1)$, by taking summation we obtain

$$\sum \frac{(p-1)(\#E(F_p) - 1)}{\#AutE},$$

By using Hasse's theorem and (5.3) we find this is at least

$$c_5(p-1)(p - 2\sqrt{p})(\#S - 2)\sqrt{p}/\log p$$

In the same way, by directly using (5.7) and Hasse's theorem we can get the second equation as well. ■

6 Appendix2: The Estimate Of the Algorithm

In this section, we will estimate the success probability as well as the time efficiency of the algorithm.

(6.1) Proposition: Let $n, v, w \in Z_{>1}$ and $a, x, y \in Z/nZ$ be as in (4.4.1), put $b = y^2 - x^3 - ax \in Z/nZ$ and $P = (x : y : 1) \in V_n$. Suppose that n has prime divisor p and q satisfying the following conditions.

(i) $p \leq v$;
(ii) $6(4\bar{a}^3 + 27\bar{b}^3) \neq 0$ for $\bar{a} = (a \bmod p), \bar{b} = (b \bmod p)$;
(iii) each prime number r dividing $\#E_{\bar{a},\bar{b}}(F_p)$ satisfies $r \leq w$;
(iv) $6(4\hat{a}^3 + 27\hat{b}^3) \neq 0$ for $\hat{a} = (a \bmod q), \hat{b} = (b \bmod q)$;
(v) $\#E_{\hat{a},\hat{b}}(F_q)$ is not divisible by the largest prime number dividing the order of P_p.
Then algorithm (4.4.1) can find a non-trivial divisor of n successfully.

Remark: To apply this proposition to the whole proof of the statement in (4.5), one only need the $n's$ not being a prime power nor divisible by 2 or 3, so the proof will also be limited to these $n's$. Also, to complete the proof we do not need to know the actual value of p and q, we only assume the existence of them.

Proof: It follows from Hasse's Inequality that $\#E_{\bar{a},\bar{b}}(F_P) \leq v + 2\sqrt{v} + 1$. So with $e(r)$ defined as in (4.4.1), denote by α the order of P_p in the group $E_{\bar{a},\bar{b}}(F_p)$, and let t be the largest prime number dividing α, and s satisfies that $t^s || \alpha$. And of course s satisfies $1 \leq s \leq e(t)$. Let

$$k_0 = (\prod_{r=2}^{t-1} r^{e(r)}) \cdot t^{s-1} \qquad ;$$

then it is obvious that $k_0 P_p \neq O_p$ and $k_0 t P_p = O_p$ in the group $E_{\bar{a},\bar{b}}(F_p)$ (this is because the limitation of the exponent of each r by $e(r)$ due to the inequality above).

If $k_0 t P \in V_n$ exists, then we have $k_0 t P = O$ in V_n. But with $k_0 t \cdot P_q = O_q$ and (v) we have $k_0 P_q = O_q$, meaning that $k_0 P = O$ in V_n, thus causing a contradiction. So $k_0 t P$ cannot exist, thus meaning the existence of a non-trivial divisor of n. ■

The next proposition attempts to show the probability that a random triple (a, x, y) can be successful, which is represented in the way of $\frac{N}{n^3}$ as it is stated in the proposition.

(6.2) Proposition: There exists a positive and effectively computable constant c with the following property. Let $n, w, v \in Z_{>1}$, with n having at least two distinct prime divisors greater than 3, and v satisfies that $p \leq v$, where p is the smallest prime divisor of n. Let

$$j = \#\{s \in Z : |s - p - 1| < \sqrt{p}, \text{ every prime divisor of } s \text{ is no greater than } w\}$$

Then let N be the number of triples $(a, x, y) \in (Z/nZ)^3$ that lead to algorithm (4.4.1) finding a non-trivial divisor of n successfully, then N satisfies

$$\frac{N}{n^3} > \frac{c}{logp} \cdot \frac{j-2}{2[\sqrt{p}]+1}.$$

Remark: We are actually looking for the triples (a, x, y) satisfying (6.1), whose number is less than N and also satisfies the inequality.

Proof: Let q be the a prime divisor different from p, For each positive number s, denote by T_s the following set:

$$T_s = \{(\alpha, x_1, y_1) : 4\alpha^3 + 27\beta^2 \neq 0, \#E_{\alpha,\beta}(F_p) = s, \text{ where } \beta = {y_1}^2 - {x_1}^3 - \alpha x_1\}.$$

Denote by $t_{(\alpha,x_1,y_1)}$ the largest prime divisor of the order of the point $(x_1 : y_1 : 1)$ in $E_{\alpha,\beta}(F_P)$ for $(\alpha, x_1, y_1) \in T_s$. Then denote by $U_{(\alpha,x_1,y_1)}$ the following set:

$$U_{(\alpha,x_1,y_1)} = \{(\alpha_2, x_2, y_2) : 4{\alpha_2}^3 + 27{\beta_2}^2 \neq 0, \#E_{\alpha_2,\beta_2}(F_q) \text{ not divisible by } t_{(\alpha,x_1,y_1)}, \\ \text{where } \beta_2 = {y_2}^2 - {x_2}^3 - \alpha_2 x_2\}.$$

To achieve the condition stated in (6.1), we define a set $V(\alpha, x_1, y_1, \alpha_2, x_2, y_2)$ in the following way:

$$V(\alpha, x_1, y_1, \alpha_2, x_2, y_2) = \{(a, x, y) \in (Z/nZ)^3 : (a(modp), x(modp), y(modp)) = (\alpha, x_1, y_1) \\ (a(modq), x(modq), y(modq)) = (\alpha_2, x_2, y_2)\}$$

Then by applying (6.1), with i summing over the set of positive integers whose greatest prime divisor is no larger than w, we have:

$$N \geq \sum_{i} \sum_{(\alpha,x_1,y_1)\in T_i} \sum_{(\alpha_2,x_2,y_2)\in U_{(\alpha,x_1,y_1)}} \#V_(\alpha, x_1, y_1, \alpha_2, x_2, y_2)$$

With the obvious result that $card(V_(\alpha, x_1, y_1, \alpha_2, x_2, y_2)) = \frac{n^3}{(pq)^3}$, applying the conclusion of (5.8)(b) we have $\#U_{(\alpha,x_1,y_1)} \geq c_{10}q^3$, reducing the inequality to

$$\tfrac{N}{n^3} \geq c_{10} \sum \tfrac{\#T_s}{p^3}$$

Remembering the fact that $|s-p-1| \leq \sqrt{p}$, applying the conclusion in (5.8)(a), thus finishing the proof. ■

With Proposition (6.2) proven, we are then able to estimate the success probability of the algorithm in (4.4.2). Based on (6.2), with the same parameters n, v, w, h as in (4.4.2), it is easy to know that the failure probability is $(1-\frac{N}{n^3})^h$, with N defined in (6.2). If we use $f(w) = \frac{j}{2[\sqrt{p}]+1}$ to represent the probability of an integer in the interval $(p+1-\sqrt{p}, p+1+\sqrt{p})$ with all its prime divisor no greater than w (j has the same definition as in (6.2)), then it follows from (6.2) that

$$\tfrac{N}{n^3} > \tfrac{c \cdot f(w)}{3logv},$$

Then $(1-\frac{N}{n^3})^h \leq e^{\frac{-hc\cdot f(w)}{3logv}}$, thus showing that the success probability is at least $1-e^{\frac{-hc\cdot f(w)}{3logv}}$.

Then it comes to the last part of estimating the time efficiency, which is an important property, of the algorithm (4.4.2). Applying the already known knowledge of the Euclidean algorithm, it is easy to know that the time needed performing a single operation of addition is $O((logn)^2)$, denoted by $M(n)$.

Remark: In the original article by Lenstra, it is stated that the time it takes in finishing one round of algorithm (4.4.2) is about $O(hw(logv)M(n))$, in the article the author said the reason to be $logk = O(wlogv)$, where k is the same as in (4.4.1). But due to the existence of the difference between addition chain, in order that (6.1) stands, it is needed that

$$(\prod_{r=2}^{t-1} r^{e(r)})t^s \cdot P$$

be calculated (regardless of it being successful or not) in the process for $3 \leq t \leq w$ and $0 \leq s \leq e(t)$. Then this shows that it is not entirely correct to prove in the original way the author gives. This relation requires that the time efficiency should be estimated in a more precise way considering the aspect above. Still, after such calculation, the result of the estimate remains unchanged.

Then with the success probability fixed, h is of the same magnitude as $\frac{logv}{f(w)}$ defined in (6.2), so the problem comes to minimizing $\frac{w}{f(w)}$.

With an unproved conjecture, which is extended beyond the theorem of Canfield, Erdös, and Pomerance, assumed true. We have the probability of a random positive integer $s \in (x+1-\sqrt{x}, x+1+\sqrt{x})$ (the original theorem applies for $x \leq s$) has all its prime divisor no greater than $L(x)^\alpha$ is $L(x)^{\frac{-1}{2\alpha}+o(1)}$. The function $L(x)$ is defined over the interval (e, ∞) by the equation

$$L(x) = e^{\sqrt{logxloglogx}}$$

Putting x to p, we obtain that

$$f(L(p)^\alpha) = L(p)^{\frac{-1}{2\alpha}+o(1)},$$

with f defined in (6.2). With $w = L(p)^\alpha$, it implies that

$$\tfrac{w}{f(w)} = L(p)^{\frac{1}{2\alpha}+\alpha+o(1)},$$

which is suggesting that the optimal choice of w is $w = L(p)^{\frac{\sqrt{2}}{2}+o(1)}$ as $p \to \infty$.

With p unknown beforehand, in the practical sense p can be substituted by v as it is given that $p \leq v$, and the actual running of the algorithm (4.4.2) can be performed by choosing v in an increasing sequence to avoid the situation that the least prime divisor of n being too large.

Hence the statement in (4.5) is proven.

7 Appendix3: Some Remarks

(7.1) With the Riemann Hypothesis assumed, one can obtain a stronger inequality in (6.2), which is $\frac{N}{n^3} > \frac{c}{loglogp} \cdot \frac{u}{2[\sqrt{p}]+1}$. But with the further analysis following the calculation after (6.2), it reaches no stronger result than the original one.

(7.2) The author stated that the algorithm can also be applied for the purpose of recognizing numbers built up from primes below a certain bound, and in this case the unproved conjecture in the estimation can be substituted by analytic results within reach of present techniques, yet it is beyond reach of this article.

(7.3) In comparison with other previous methods, its expected total factoring time in worst cases (the second largest prime divisor of n is not much less than $\sqrt{n}$) is $L(n)^{1+o(1)}$ as $n \to \infty$, which is also reachable by other methods. The main advantage of this elliptic curve method lies in factoring integer n with smaller prime divisors.

8 References:

1. ·H.W.LENSTRA.JR, Factoring integers with elliptic curves. Annals of Mathematics, Second Series, Vol. 126, No. 3 (Nov., 1987), pp. 649-673

2. Joseph H.Silverman, The Arithmetic of Elliptic Curves. New York Springer, second edition, Vol 106, (2008)

The Weil Bound

Yongle Hu

Abstract

This article introduces some properties of function fields, and proves the Weil bound for some character sums.

1 Function Fields

Let k be a field, and K is an extension of k which includes an element t such that t is transcendental over k, then the field K is called a function field over k, while k is called a constant filed of K.

Let $k[x]$ denote the ring of polynomials in one variable over k, then the quotient field of $k[x]$ is called rational function field, which is denoted by $k(x)$. Clearly $k(x)$ is a function field over k, which is the situation we are most concerned about in this passage.

In fact, $k(x)$ and $\mathbb{Q}$ have many structures in common. One aspect of them is valuation. We need to do some preparation first.

Let G be an abelian group. Then G is called an ordered group, if it can be equipped with an operation '$<$' satifying, for all $a,\ b,\ c \in G$:

(1) Exactly one of these three cases holds true: $a < b, b < a, a = b$.

(2) If $a < b$ and $b < c$, then $a < c$.

(3) If $a < b$, then $a + c < b + c$.

Asssume G is an ordered group. Let ∞ be an symbol not in G satisfying $a < \infty,\ a + \infty = \infty + a = \infty + \infty = \infty$ for any $a \in G$. For all $a, b \in G$, we define $a > b$ if and only if $b < a$, $a \leq b$ if and only if $a < b$ or $a = b$, $a \geq b$ if and only if $a > b$ or $a = b$. Note that G must be torsion free since if $a \in G$ and $a \neq 0$, say $a > 0$, then $na = a + \cdots + a > 0 + \cdots + 0 = 0$.

Let K be a function field over k. A map $v : K \to G \cup \{\infty\}$ is called a valuation of K/k, if it satisfies the following conditions:

(1) $v(ab) = v(a) + v(b),\ \forall a, b \in K$, i.e., v is a group homomorphism;

(2) $v(a+b) \geq \min\{v(a), v(b)\},\ \forall a, b \in K$;
(3) $v(a) = \infty$ if and only if $a = 0$;
(4) v is non-trivial, i.e., $\exists a \neq 0$ such that $v(a) \neq 0$;
(5) $v(c) = 0,\ \forall c \in k^*$.

The second condition is usually called strong triangle inequality. Form this condition, it can be known that if $a, b \in K,\ v(a) < v(b)$, then $v(a+b) \geq v(a)$, $v(-b) = v(b) + v(-1) = v(b)$, and $v(a) = v(a+b-b) \geq \min\{v(a+b),\ v(-b)\} = \min\{v(a+b),\ v(b)\}$. But $v(a) < v(b)$, so $v(a) \geq \min\{v(a+b),\ v(b)\} = v(a+b)$ is the only possible case. As a result, $v(a+b) = v(a)$. By induction, it can be easily got that if $a_1, \ldots, a_n \in K,\ v(a_1) < v(a_i),\ i = 2, \ldots, n$, then

$$v(\sum_{k=1}^{n} a_k) = v(a_1) \tag{1}$$

Let $\mathcal{O}_v = \{a \in K |\ v(a) \geq 0\}$, $\mathcal{P}_v = \{a \in K |\ v(a) > 0\}$. Then $\mathcal{O}_v$ is a subring of K, $\mathcal{P}_v$ is an ideal of $\mathcal{O}_v$. For any $a \in \mathcal{O}_v \setminus \mathcal{P}_v$, $v(a) = 0$, so $v(a^{-1}) = -v(a) = 0$, thus $a^{-1} \in \mathcal{O}_v$, a is a unit of $\mathcal{O}_v$. As a result, $\mathcal{O}_v$ is a local ring and $\mathcal{P}_v$ is the maximal ideal. $\mathcal{O}_v$ is usually called valuation ring associated to v, and $\mathcal{O}_v/\mathcal{P}_v$ is called residue field of K with respect to $\mathcal{P}_v$.

A valuation $v : K \to \mathbb{R} \cup \{\infty\}$ of K/k is called discrete, if $v(K^*)$ have no limit in $\mathbb{R}$. Moreover, if $v(K^*) = \mathbb{Z}$, then v is called normalized.

Similar to the valuation the p-adic valuations in $\mathbb{Q}$, there are two kinds of normalized valuations of $k(x)/k$. One of them is v_p, where $p \in k[x]$ is a monic irreducible polynomial. For any $a \in k(x)$, write $a = p(x)^n \dfrac{f(x)}{g(x)},\ n \in \mathbb{Z},\ p(x) \nmid f(x)g(x)$, then define $v_p(a) = n$. The other is v_∞. For $\dfrac{f(x)}{g(x)} \in k(x)$, define $v_\infty(\dfrac{f(x)}{g(x)}) = \deg(g(x)) - \deg(f(x))$. It can be easily checked that v_∞ and v_p are normalized valuations of $k(x)/k$.

Theorem 1. Let $v : k(x) \to G \cup \{\infty\}$ is a valuation of $k(x)/k$, then there exsists $c \in G$ and $c > 0$ such that $v = v_\infty c$ or $v = v_p c$ for some monic irreducible polynomial $p(x)$.

Proof. (i)If $v(x) \geq 0$, then for $f(x) = \sum\limits_{k=0}^{n} c_i x^i$, $v(f(x)) \geq \min\limits_{1 \leq i \leq n} \{v(c_i x^i)\}$. But $v(c_i x^i) = v(c_i) + i \cdot v(x) \geq 0$, so $v(f(x)) \geq 0$, $f(x) \in \mathcal{O}_v$. As a result, $k[x] \subseteq \mathcal{O}_v$. Set $\mathfrak{p} = \mathcal{P}_v \cap k[x]$, then if $a, b \in k[x]$, $ab \in \mathfrak{p}$, then $ab \in \mathcal{P}_v$, so $a \in \mathcal{P}_v$ or $b \in \mathcal{P}_v$. But $a \in k[x]$ and $b \in k[x]$, so $a \in \mathfrak{p}$ or $b \in \mathfrak{p}$, which means $\mathfrak{p}$ is a prime ideal of $k[x]$. Write $\mathfrak{p} = (p(x))$, where $p(x)$ is a monic irreducible polynomial. For $f(x) \in k[x]$, if

$p(x) \nmid f(x)$, then $f(x) \notin (p(x)) = \mathcal{P}_v \cap k[x]$, so $f(x) \notin \mathcal{P}_v$, moreover $v(f(x)) = 0$. As a result, for any $a \in k(x)$, write $a = p(x)^n \dfrac{f(x)}{g(x)}$, $n \in \mathbb{Z}$, $p(x) \nmid f(x)g(x)$, then $v(a) = v(p(x))n + v(f(x)) - v(g(x)) = v(p(x))v_p(a)$. Since v is nontrivial, $v(p(x)) \neq 0$, then $v(p(x)) > 0$. The proof is completed by setting $c = v(p(x))$.

(ii) If $v(x) < 0$, then for $f(x) = \sum\limits_{i=0}^{n} c_{n-i}x^i$, $c_0 \neq 0$, then $v(c_0x^n) < v(c_ix^{n-i})$, $i = 1, \ldots, n$. Using the equation (1), we have $v(f(x)) = v(c_0x^n) = v(x)\deg(f(x))$. As a result, for any $a \in k(x)$, write $a = \dfrac{f(x)}{g(x)}$, then $v(a) = v(f(x)) - v(g(x)) = v(x)(\deg(g(x)) - \deg(f(x))) = -v(x)v_\infty(a)$. Since $-v(x) > 0$, The proof is completed by setting $c = -v(x)$. □

The Theorem 1 shows that v_∞ and v_p are all the types of valuations of $k(x)/k$. As a corollary of Theorem 1, every valuation $v : K \to \mathbb{R} \cup \{\infty\}$ of $k(x)/k$ is discrete.

Let E is a field, and ∞ is a symbol not in E satisfying: $a+\infty = \infty+a = \infty$, $\forall a \in E$; $a \cdot \infty = \infty \cdot a = \infty$, $\forall a \in E^*$; $\infty \cdot \infty = \infty$. Note that $\infty + \infty$, $0 \cdot \infty$ and $\infty \cdot 0$ are invalid formulas.

A function $\varphi : K \to E \cup \{\infty\}$ is called a place of K/k, if it satisfies the following conditions:
(1) $\varphi(a+b) = \varphi(a) + \varphi(b)$, $\varphi(ab) = \varphi(a)\varphi(b)$, for all $a, b \in K$ such that the right sides of the equations are valid;
(2) φ is notrival, i.e., $\varphi(1) = 1$ and $\exists a \in K$ such that $\varphi(a) = \infty$;
(3) $v(a) \neq 0$ or ∞, for all $a \in k^*$.

Let $\mathcal{O}_\varphi = \{a \in K |\ \varphi(a) \neq \infty\}$, then $\mathcal{O}_\varphi$ is a subring of K. Thus $\varphi : \mathcal{O}_\varphi \to E$ is a ring homomorphism. Set $\mathcal{P}_\varphi = \mathrm{Ker}(\varphi) = \{a \in K |\ \varphi(a) = 0\}$, then $\mathcal{P}_\varphi$ is a prime ideal of $\mathcal{O}_\varphi$. For any $a \in \mathcal{O}_\varphi \setminus \mathcal{P}_\varphi$, $\varphi(a)\varphi(a^{-1}) = \varphi(a \cdot a^{-1}) = \varphi(1) = 1$, so $\varphi(a^{-1}) \neq \infty$, $a^{-1} \in \mathcal{O}_\varphi$, therefore a is a unit of $\mathcal{O}_\varphi$. As a result, $\mathcal{O}_\varphi$ is a local ring and $\mathcal{P}_\varphi$ is the maximal ideal.

Assume $a \in K \setminus \mathcal{O}_\varphi$, then $\varphi(a) = \infty$. If $\varphi(a^{-1}) \neq 0$, then $1 = \varphi(1) = \varphi(a)\varphi(a^{-1}) = \infty$, which is a contradiction. Therefore $\varphi(a^{-1}) = 0$, then $a^{-1} \in \mathcal{P}_\varphi$.

Consider K^* and $\mathcal{O}_\varphi^*$, the multiplicative groups of K and $\mathcal{O}_\varphi$. They are both abelian groups, so the factor group $K^*/\mathcal{O}_\varphi^*$ is also an abelian group. For $a \in K^*$, let $\bar{a} = a\mathcal{O}_\varphi^* \in K^*/\mathcal{O}_\varphi^*$. Then the operation '<' can be defined as follows: $\bar{a} < \bar{b}$ if and only if $ba^{-1} \in \mathcal{P}_\varphi$. For all $\bar{a}, \bar{b}, \bar{c} \in K^*/\mathcal{O}_\varphi^*$, first, if $ba^{-1} \in \mathcal{P}_\varphi$, then $\bar{a} < \bar{b}$; if $ba^{-1} \in \mathcal{O}_\varphi^*$, then $\bar{a} = \bar{b}$; if $ba^{-1} \in K^* \setminus \mathcal{O}_\varphi$, then $ab^{-1} \in \mathcal{P}_\varphi$, so $\bar{b} < \bar{a}$. Second, if $\bar{a} < \bar{b}$ and $\bar{b} < \bar{c}$, then

$ca^{-1} = ba^{-1} \cdot cb^{-1} \in \mathcal{P}_\varphi$, thus $\bar{a} < \bar{c}$. Third, if $\bar{a} < \bar{b}$, then $(bc)(ac)^{-1} = ba^{-1} \in \mathcal{P}_\varphi$, so $\overline{ac} < \overline{bc}$. As a result, $K^*/\mathcal{O}_\varphi^*$ is an ordered group.

Let v is the canonical map $K^* \to K^*/\mathcal{O}_\varphi^*$, and define $v(0) = \infty$ additionally. Then for all $a, b \in K^*$, $v(ab) = \overline{ab} = \bar{a}\bar{b} = v(a) \cdot v(b)$. Say $v(a) \leq v(b)$, then $ba^{-1} \in \mathcal{O}_\varphi$, $(a+b)a^{-1} = 1 + ba^{-1} \in \mathcal{O}_\varphi$, thus $v(a+b) \geq v(a) = \min\{v(a), v(b)\}$. Since there exists $a \in K$ such that $\varphi(a) = \infty$, $a \notin \mathcal{O}_\varphi$, $v(a) \neq 0$. At last, for $a \in k^*$, $\varphi(a) \neq 0$ or ∞, hence $a \in \mathcal{O}_\varphi \setminus \mathcal{P}_\varphi = \mathcal{O}_\varphi^*$, so $v(a) = 0$. As a result, v is a valuation of K/k. Obviously $\mathcal{O}_\varphi = \mathcal{O}_v$ and $\mathcal{P}_\varphi = \mathcal{P}_v$.

Conversely, if v is a valuation of K/k, define $\varphi : K \to \mathcal{O}_v/\mathcal{P}_v \cup \{\infty\}$ as follows: if $a \in \mathcal{O}_v$, then $\varphi(a) = a + \mathcal{P}_v$; otherwise $\varphi(a) = \infty$. It is easy to verify that φ is a place of K/k. We will often simply use $\mathcal{P}_v$ to denote this place.

Therefore, valuation and place are essentially the same concept. Let $\mathcal{P}$ be a place of K/k and v be a valuation corresponding to $\mathcal{P}$. The *residue field* associated to $\mathcal{P}$ is $k(\mathcal{P}) = \mathcal{O}_\mathcal{P}/\mathcal{P} = \mathcal{O}_v/\mathcal{P}_v$. Write $f_\mathcal{P} = [k(\mathcal{P}) : k]$, where $f_\mathcal{P}$ is called the *degree* of the place $\mathcal{P}$.

From now on, we will discuss only the case $K = k(x)$. Assume v is a normalized valuation of $k(x)/k$ and $\mathcal{P}$ is the place corresponding to v. If $v = v_p$, in which $p(x) \in k[x]$ is a monic irreducible polynomial. Let $S_p = \{g \in k[x] \mid p \nmid g\}$. Then $\mathcal{O}_{v_p} = S_p^{-1}k[x]$, $\mathcal{P}_{v_p} = S_p^{-1}(p)$. So $k(\mathcal{P}) = \mathcal{O}_{v_p}/\mathcal{P}_{v_p} = S_p^{-1}k[x]/S_p^{-1}(p) \cong k[x]/(p)$ and $f_\mathcal{P} = [k[x]/(p) : k] = \deg(p)$. In this case, we will write $\mathcal{P} = \mathcal{P}_p$.

Otherwise $v = v_\infty$. For $R(x) = \dfrac{b_m x^m + \cdot + b_0}{c_n x^n + \cdot + c_0} \in k(x)$, $v_\infty(R(x)) = n - m = v_x(x^{n-m}\dfrac{b_m + \cdot + b_0 x^m}{c_n + \cdot + c_0 x^n}) = v_x(\dfrac{b_m x^{-m} + \cdot + b_0}{c_n x^{-n} + \cdot + c_0}) = v_x(R(\dfrac{1}{x}))$. Let $y = \dfrac{1}{x}$, then $v_\infty(R(x)) = v_x(R(y))$. Since $k(x) \cong k(y)$, $k(\mathcal{P}) = \mathcal{O}_{v_\infty}/\mathcal{P}_{v_\infty} \cong \mathcal{O}_{v_x}/\mathcal{P}_{v_x} \cong k[x]/(x) \cong k$. As a result, $f_\mathcal{P} = [k : k] = 1$. In this case, we will write $\mathcal{P} = \mathcal{P}_\infty$.

Let $\mathbb{P}_k$ denote the set of all the places of $k(x)/k$, $\mathbb{D}_k$ denote the free abelian group generated by all the elements of $\mathbb{P}_k$. An element of $\mathbb{D}_k$ is called a divisor. Equivalently, a divisor is a formal sum of all the element of $\mathbb{P}_k$ with integral coeffcients such that only a finite number of coefficients are non-zero. Two divisors $\mathfrak{a} = \sum a_\mathcal{P}\mathcal{P}, \mathfrak{b} = \sum b_\mathcal{P}\mathcal{P}$ are called coprime, if for any place $\mathcal{P}$, either $a_\mathcal{P} = 0$ or $b_\mathcal{P} = 0$.

Assume that $\mathfrak{m}$ is a divisor. Define the degree of $\mathfrak{m}$ by

$$\mathcal{D}(\mathfrak{m}) = \sum_{\mathcal{P} \in \mathbb{P}_k} a_\mathcal{P} f_\mathcal{P}, \text{ where } \mathfrak{m} = \sum_{\mathcal{P} \in \mathbb{P}_k} a_\mathcal{P}\mathcal{P}.$$

A divisor is called finite, if the coefficient of $\mathcal{P}_\infty$ is zero. A divisor is called positive, if all the coefficients are non-negative. In fact, finite positive divisors are essentially the same as monic polynomials in $k[x]$. For a finite positive divisor $\mathfrak{a} = \sum a_p \mathcal{P}_p$, since $a_p \geq 0$ and only a finite number of $a_p \neq 0$, $\prod p(x)^{a_p}$ is a monic polynomial of $k[x]$. Conversely, if $f \in k[x]$ is a monic polynomial and its standard prime factorization decomposition is $f(x) = \prod p(x)^{a_p}$, then $\mathfrak{a} = \sum a_p \mathcal{P}_p$ is a finite positive divisor. In this case, we will write $f(x) = R_\mathfrak{a}(x)$. Then $d(\mathfrak{a}) = \sum a_p f_{\mathcal{P}_p} = \sum a_p \deg(p) = \deg(R_\mathfrak{a})$.

For a finite divisor $\mathfrak{m}$, there exists positive divisors $\mathfrak{a}, \mathfrak{b}$ such that $\mathfrak{m} = \mathfrak{a} - \mathfrak{b}$. Then we write $R_\mathfrak{m}(x) = \dfrac{R_\mathfrak{a}(x)}{R_\mathfrak{b}(x)}$. Conversely, every $m \in k(x)^*$ is corresponding to a divisor $\mathfrak{m}$, which remains true when $k(x)$ is replaced by other function fields.

Suppose that all distinct roots of $R_\mathfrak{a}(x)R_\mathfrak{b}(x)$ in $\bar{k}$ are $\xi_1, \xi_2, \ldots, \xi_d$, then define $d_0(\mathfrak{m}) = d$. Write $R_\mathfrak{m}(x) = \prod\limits_{j=1}^{d} (x - \xi_j)^{a_j}$. Then for $R(x) \in k(x)$, define $R(\mathfrak{m}) = \prod\limits_{j=1}^{d} R(\xi_j)^{a_j}$. Due to Vieta's theorem, $R(\mathfrak{m}) \in k$.

2 The Weil bound for some character sums

From now on, we will assume that k is a finite field of q elements, whose characteristic is p.

Let G be a group, then a homomorphism $\chi : G \to \mathbb{C}^*$ is called a character of G. χ is called non-trivial, if there exists $g \in G$ such that $\chi(g) \neq 1$.

Lemma 2. Assume that G is a finite group and χ is a non-trivial character of G, then $\sum\limits_{a \in G} \chi(a) = 0$.

Proof. Since χ is non-trivial, there exists $g \in G$ such that $\chi(g) \neq 1$. Then since G is finite, $a \mapsto ga$ is a bijection from G to itself. Then

$$\sum_{a \in G} \chi(a) = \sum_{a \in G} \chi(ga) = \sum_{a \in G} \chi(g)\chi(a) = \chi(g) \sum_{a \in G} \chi(a).$$

Because $\chi(g) \neq 1$, we have

$$\sum_{a \in G} \chi(a) = 0. \tag{2}$$

□

Set $G_0 = \{1 + xf(x) \mid f \in k[[x]]\}$, then G_0 is a subgroup of $k[[x]]^*$. Let ω be a non-trivial character of G_0. Assume that N is a positive integer such that $\omega(1 + x^N f) = 1$ for all $f \in k[[x]]$. For $a \in k((x))^*$, write $a(x) = cx^n(1 + xf(x))$, where $c \in k, n \in \mathbb{Z}, f(x) \in k[[x]]$. Then define $\omega(a) = \omega(1 + xf)$. It can be easily verified that ω is a character of the multiplicative group $k((x))^*$.

Let s be a positive integer such that $p^s \geq N$. For any $a \in k((x))^*$, write $a(x) = cx^n(1 + a_1 x + \dots)$. Since p is the characteristic of k, $\omega(a)^{p^s} = \omega((1 + a_1 x + \dots)^{p^s}) = \omega(1 + a_1^{p^s} x^{p^s} + \dots) = 1$. Thus the image of ω is a subset of the p^s-roots of unity.

For $f(x) = 1 + a_1 x + a_2 x^2 + \dots \in G_0$, set $g(x) = 1 + a_1 x + \dots + a_{N-1} x^{N-1}$, $h(x) = a_N + a_{N+1} x + a_{N+2} x^2 + \dots$, then $f(x) = g(x) + x^N h(x)$. Let g^{-1} denote inverse for g in G_0, then $\omega(1 + x^N g^{-1} h) = 1$. As a result,

$$\begin{aligned} \omega(1 + a_1 x + a_2 x^2 + \dots) &= \omega(g + x^N h) = \omega(g)\omega(1 + x^N g^{-1} h) = \omega(g) \\ &= \omega(1 + a_1 x + \dots + a_{N-1} x^{N-1}). \end{aligned} \tag{3}$$

For any $f(x) \in k[x]$, we can naturally consider $f(x)$ as an element of $k[[x]]$. Furthermore, for any $R(x) \in k(x)^*$, write $R(x) = \dfrac{f(x)}{g(x)}$, where $f(x), g(x) \in k[x]^*$. Then $R(x)$ can be considered as an element of $k((x))^*$, so $\omega(R(x))$ is well defined.

Now we can state the main theorem of this part.

Theorem 3. Let χ_0 be a character of the multiplicative group k^*, and ω satisfies the above properties. Set $\chi_0(0) = 0$, $\chi_0(\infty)$ and $\omega(0) = 0$ in addition. Let $\lambda(a) = \omega(1 - ax)$ for all $a \in k$. Assume that $\mathfrak{b}$ is a finite divisor and $d = d_0(\mathfrak{b})$, then

$$\left| \sum_{a \in k} \lambda(a) \chi_0[R_{\mathfrak{b}}(a)] \right| \leq (N + d - 2)\sqrt{q}. \tag{4}$$

Proof. For any finite divisor $\mathfrak{m}$ coprime to $\mathfrak{b}$, define that

$$\Delta(\mathfrak{m}) = \omega[R_{\mathfrak{m}}(1/x)]\chi_0[R_{\mathfrak{m}}(\mathfrak{b})].$$

Note that if $\mathfrak{m}$ and $\mathfrak{n}$ are finite divisor coprime to $\mathfrak{b}$, $R_{\mathfrak{m}+\mathfrak{n}}(x) = R_{\mathfrak{m}}(x) R_{\mathfrak{m}}(x)$, thus $\Delta(\mathfrak{m} + \mathfrak{n}) = \Delta(\mathfrak{m})\Delta(\mathfrak{n})$.

Since $|\omega(f)| = 1$ and $|\chi_0(a)| = 1$ for all $f \in k((x))^*, a \in k^*$, $|\Delta(\mathfrak{m})| = 1$ for all finite divisor $\mathfrak{m}$ coprime to $\mathfrak{b}$.

For any divisor $\mathfrak{m}$, the norm of $\mathfrak{m}$ is defined by $|\mathfrak{m}| = q^{\mathcal{D}(\mathfrak{m})}$. The set of all finite positive divisors coprime to $\mathfrak{m}$ is denoted by $C(\mathfrak{m})$.

Assume that $s \in \mathbb{C}$ and $\mathrm{Re}(s) > 1$, then define the $L-$series

$$L(s, \Delta) = \sum_{\mathfrak{a} \in C(\mathfrak{b})} \frac{\Delta(\mathfrak{a})}{|\mathfrak{a}|^s}.$$

We will show that the $L-$series has beautiful structures.

Lemma 4. The $L-$series is well defined when $\mathrm{Re}(s) > 1$, which is a holomorphic function of s. Moreover, it is a polynomial of q^{-s}, and the degree of the polynomial is no more than $N + d - 2$.

Proof of the lemma. Let $\epsilon > 0$, and n be a positive integer. Since any finite positive divisor of degree n corresponds to a monic polynomial of degree n, there are q^n finite positive divisors of degree n. Thus there are at most q^{n-i}a positive divisor $\mathfrak{a}$ of degree n of the form $\mathfrak{a} = i\mathcal{P}_\infty + \mathfrak{a}_1$ where $\mathfrak{a}_1$ is a finite positive divisor of degree $n - i$. As a result, the number of positive divisors of degree n is no more than $\sum\limits_{i=0}^{n} q^{n-i} \leq q^{n+1}$.

Therefore

$$\sum_{\mathfrak{a} \in C(\mathfrak{b})} \left| \frac{\Delta(\mathfrak{a})}{|\mathfrak{a}|^s} \right| \leq \sum_{\mathfrak{a} \text{ is positive}} \frac{1}{|\mathfrak{a}|^{\mathrm{Re}(s)}} \leq \sum_{n=1}^{\infty} \frac{q^{n+1}}{q^{n\mathrm{Re}(s)}} = \frac{q}{1 - q^{1-\mathrm{Re}(s)}}.$$

As a result, $L(s, \Delta)$ is absolute uniform convergence when $\mathrm{Re}(s) > 1 + \epsilon$, so it defines a holomorphic function of s when $\mathrm{Re}(s) > 1 + \epsilon$. Due to the arbitrariness of the choice of ϵ, $L(s, \Delta)$ is a holomorphic function of s when $\mathrm{Re}(s) > 1$.

Note that above proof remains true when Δ is replaced by any character χ of $\mathbb{D}_k$ satisfying $\sup\{|\chi(\mathfrak{a})|\} < \infty$.

Write $R_{\mathfrak{b}}(x) = \prod\limits_{j=1}^{d} (x - \xi_j)^{b_j}$, $\xi_j \in \bar{k}$, $b_j \in \mathbb{Z} \backslash \{0\}$, $j = 1, \ldots, d$. Set $b(x) = \prod\limits_{j=1}^{d} (x - \xi_j)$. Since k is a finite field, every irreducible polynomial in $k[x]$ is separable, thus having no multiple roots. As a result, $b(x)$ is a product of some irreducible polynomials, so $b(x) \in k[x]$.

To prove that $L(s, \Delta)$ is a polynomial of q^{-s} whose degree is no more than $N + d - 2$, we only need to show that for $n \geq N + d - 1$, the coefficient of q^{-sn} is zero, i.e.

$$\sum_{\substack{\mathfrak{a} \in C(\mathfrak{b}) \\ \mathcal{D}(\mathfrak{a}) = n}} \Delta(\mathfrak{a}) = \sum_{\substack{\mathfrak{a} \in C(\mathfrak{b}) \\ \mathcal{D}(\mathfrak{a}) = n}} \omega\left(R_{\mathfrak{a}}\left(\frac{1}{x}\right)\right) \chi_0[R_{\mathfrak{a}}(\mathfrak{b})] = 0. \tag{5}$$

Since every finite positive divisor $\mathfrak{a}$ of degree n is associated to a monic polynomial $R_{\mathfrak{a}}(x)$ of the same degree, the equation (5) can be modified to

$$\sum_{\substack{\deg(f)=n \\ f \text{ is monic, } (f,b)=1}} \omega\left(f\left(\frac{1}{x}\right)\right) \chi_0[f(\mathfrak{b})] = 0. \tag{6}$$

Note that $\chi_0[h(\mathfrak{b})] = \chi_0\left(\prod_{j=0}^{d} h(\xi_j)^{b_j}\right)$, so this value only depends on $h(\xi_j)$. As a result, if $h \equiv g \mod b$, then $h(\xi_j) = g(\xi_j)$, $j = 1, \ldots, d$, so $\chi_0[h(\mathfrak{b})] = \chi_0[g(\mathfrak{b})]$.

Write $f(x) = \sum_{j=0}^{n} a_{n-j}x^j$, $a_0 = 1$, then use the equation (3), $\omega[f(1/x)] = \omega[x^{-n}(1 + a_1x + \cdots + a_nx^n)] = \omega(1 + a_1x + \cdots + a_{N-1}x^{N-1})$, which means the value $\omega[f(1/x)]$ only depends on $a_1, \ldots, a_{N-1}$.

For such a polynomial f, let $f_1(x) = \sum_{j=0}^{N-1} a_jx^{n-j}$, $f_2(x) = \sum_{j=0}^{n-N} a_{n-j}x^j$, then there exists $g(x) \in k[x]$ such that $\deg(g) < d$ and $g \equiv f_2 \mod b$. So $\omega[f(1/x)]\chi_0[f(\mathfrak{b})] = \omega[f_1(1/x)]\chi_0[(f_1+g)(\mathfrak{b})]$. Conversely, if $g(x) \in k[x]$, $\deg(g) < d$, then for any polynomial h such that $\deg(h) \leq n - N - d$, $\deg(g + bh) \leq n - N$, so there exists exactly one f satifying such that $f_2 = g + bh$. Since the number of such h is $q^{n-N-d+1}$, we have

$$\text{LHS of (6)} = q^{n-N-d+1} \sum_{f_1} \omega\left(f_1\left(\frac{1}{x}\right)\right) \sum_{\substack{\deg(g)<d \\ (f_1+g,b)=1}} \chi_0[(f_1 + g)(\mathfrak{b})], \tag{7}$$

where the summation is performed for all f_1 of the form $f_1(x) = \sum_{j=0}^{N-1} a_jx^{n-j}$.

Let $G_b = (k[x]/(b))^*$. For $h \in k[x]$ such that $(h, b) = 1$, write $\bar{h} = h + (b) \in G_b$. Fix f_1, then for any $\bar{h} \in G_b$, there exists exactly one g such that $\deg(g) < d$ and $\overline{f_1 + g} = \bar{h}$.

Note that in this case, $\chi_0[(f_1+g)(\mathfrak{b})] = \chi_0[h(\mathfrak{b})]$, thus

$$\sum_{\substack{\deg(g)<d \\ (f_1+g,b)=1}} \chi_0[(f_1+g)(\mathfrak{b})] = \sum_{\bar{h}\in G_b} \chi_0[h(\mathfrak{b})].$$

Substituting it into equation (7), we have

$$\text{LHS of (6)} = q^{n-N-d+1} \sum_{f_1} \omega\left(f_1\left(\frac{1}{x}\right)\right) \sum_{\bar{h}\in G_b} \chi_0[h(\mathfrak{b})]. \tag{8}$$

Let $G_1 = \{1 + a_1 x + \cdots + a_{N-1}x^{N-1} \mid a_j \in k,\ j = 1,\ldots,N-1\}$, then G_1 can be considered as a subgroup of $(k[x]/(x^N))^*$, and ω can be considered as a non-trivial character of G_1. So by using lemma 2, we have

$$\sum_{f_1} \omega\left(f_1\left(\frac{1}{x}\right)\right) = \sum_{h\in G_1} \omega(h) = 0.$$

Substituting it into equation (8), we gain that the right hand side of equation (8) is zero, therefore the equation (6) is proved. □

From the lemma 4, we know that $L(s,\Delta) = F(q^{-s})$, where F is a polynomial such that $\deg(F) \le N + d - 2$. Since $F(q^{-s})$ can be well defined for all $s \in \mathbb{C}$, we can extend $L(s,\Delta)$ to the whole complex plane.

Write $d_1 = \deg(F)$. Since $\mathfrak{a} = 0$ is the unique element in $C(\mathfrak{b})$ such that $\mathcal{D}(\mathfrak{a}) = 0$ amd $R_0(x) = 1$, the constant term of F is $\Delta(0) = 1$. So we can write $F(x) = \prod_{j=1}^{d_1}(1 - \alpha_j x)$. By comparing the coefficient of the term x of two sides of above equation, we have

$$\sum_{\substack{\mathfrak{a}\in C(\mathfrak{b}) \\ \mathcal{D}(\mathfrak{a})=1}} \Delta(\mathfrak{a}) = -\sum_{j=1}^{d_1} \alpha_j \tag{9}$$

Any finite positive divisor $\mathfrak{a} \in C(\mathfrak{b})$ of degree 1 is associated to a monic polynomial $R_{\mathfrak{a}}(x) = x - a$ where $a \in k$. Then $\omega[R_{\mathfrak{a}}(1/x)] = \omega[x^{-1}(1-ax)] = \omega(1-ax) = \lambda(a)$, $R_{\mathfrak{a}}(\mathfrak{b}) = \prod_{j=1}^{d}(\xi_j - a)^{b_j} = (-1)^{\mathcal{D}(\mathfrak{b})} R_{\mathfrak{b}}(a)$, therefore $\Delta(\mathfrak{a}) = (-1)^{\mathcal{D}(\mathfrak{b})}\lambda(a)R_{\mathfrak{b}}(a)$. Conversely, if $x - a$ is associated to the divisor $\mathfrak{a}$ which is not in $C(\mathfrak{b})$, then $x - a | b(x)$, thus there exists $j_0 \in \{1, \ldots, d\}$ such that $a = \xi_{j_0}$, so $R_{\mathfrak{b}}(a) = \prod_{j=1}^{d}(a - \xi_j)^{b_j} = 0$. Substituting them into equation (9), we gain that

$$\sum_{a \in k} \lambda(a)\chi_0[R_{\mathfrak{b}}(a)] = (-1)^{\mathcal{D}(\mathfrak{b})+1} \sum_{j=1}^{d_1} \alpha_j. \tag{10}$$

According to the original paper [3], it can be proved that this $L-$series divides the zeta-function of an Abelian extension of $k(x)$ by class-field theory. Then by the Riemann hypothesis in function fields[1], any the root s_0 of the zeta-function satisfies that $\mathrm{Re}(s_0) = 1/2$, thus any the root s_0 of $L(s, \Delta)$ satisfies that $\mathrm{Re}(s_0) = 1/2$. Since $L(s, \Delta) = \prod_{j=1}^{d_1}(1 - \alpha_j q^{-s})$, $s_j = \dfrac{\log \alpha_j}{\log q}$ is a root of $L(s, \Delta)$. Therefore $\mathrm{Re}(s_j) = 1/2$, $|\alpha_j| = |q^{s_j}| = q^{\mathrm{Re}(s_j)} = \sqrt{q}$. Substituting them into equation (10), we have

$$\left| \sum_{a \in k} \lambda(a)\chi_0[R_{\mathfrak{b}}(a)] \right| \leq \sum_{j=1}^{d_1} |\alpha_j| \leq d_1\sqrt{q} \leq (N + d - 2)\sqrt{q}. \tag{11}$$

As a result, the theorem 3 is proved. □

In the end, we will dicuss an application of the theorem 3.

Corollary 5. Let $F(x) \in k[x]$ such that $\deg(F) = n$ and $F(0) = 0$. Assume that ψ is a non-trivial character of the additive group of k, and there exists $a_0 \in k$ such that $\psi[F(a_0)] \neq 1$. Then

$$\left| \sum_{a \in k} \psi[F(a)]\chi_0[R_{\mathfrak{b}}(a)] \right| \leq (n + d - 1)\sqrt{q}. \tag{12}$$

Proof. Let $N = n + 1$. Due to the theorem 3, we only need to construct a character ω satifying the above conditions such that $\lambda(a) = \omega(1 - ax) = \psi[F(a)]$.

Let $m \geq n$ be an integer, $k[x_1, \ldots, x_m]$ denote the ring of polynomials in m variables. Consider the elementary symmetric polynomials of $x_1, \ldots, x_m$

$$\sigma_t = (-1)^t \sum_{1 \leq j_1 < \cdots < j_t \leq m} x_{j_1} \ldots x_{j_t},\ t = 1, \ldots, m.$$

[1] The first proof of the Riemann hypothesis in function fields over a finite field owed to A. Weil [2] in 1941. Then a new approach was invented by S. A. Stepanov [6] in 1969, which was simplified by E. Bombieri [7] in 1973.

Write $F(x) = \sum_{i=1}^{n} c_i x^i$, $S_t = \sum_{j=0}^{m} x_j^t$, $t = 1, \dots, n$. Due to Newton's identities, $S_t = \sum_{i=1}^{t-1} \sigma_i S_{t-i} + t\sigma_t$. Thus by induction, there exists $G_t(x_1, \dots, x_t) \in k[x_1, \dots, x_t]$, which is independent of m and has no constant term, such that $S_t = G_t(\sigma_1, \dots, \sigma_t)$, $t = 1, \dots, n$. Then $\sum_{j=1}^{m} F(x_j) = \sum_{i=1}^{n} c_i S_i = G(\sigma_1, \dots, \sigma_n)$, where $G = \sum_{i=1}^{n} c_i G_i \in k[x_1, \dots, x_n]$ independent of m.

For $f(x) = 1 + b_1 x + b_2 x^2 + \dots \in k[[x]]$, set $h(x) = x^n + b_1 x^{n-1} + \dots + b_n \in k[x]$. Let $a_1, \dots, a_n$ be the roots of h in $\bar{k}$, and σ_t be the elementary symmetric polynomials corresponding to $a_1, \dots, a_n$. Then $\sigma_j = b_j$, $j = 1, \dots, n$. Define $\omega(f) = \psi[G(b_1, \dots, b_n)]$. We will show that ω meets all the conditions.

First, let $b_1 = -a \in k$, $b_2 = b_3 = \dots = b_n = 0$, then $h(x) = x^n + ax^{n-1}$, $a_1 = a$, $a_2 = \dots = a_n = 0$, so $\omega(f) = \psi[G(1, 0, \dots, 0)] = \psi[F(a)]$. Thus $\omega(1 - ax) = \psi[F(a)]$.

Second, setting $b_1 = 0$ in addition, we gain that $\omega(1 + x^{n+1} g) = \psi[F(0)] = 1$ for all $g \in k[[x]]$.

Third, for another element $f_1(x) = 1 + b_1' x + b_2' x^2 + \dots \in k[[x]]$, set $h_1(x) = x^n + b_1' x^{n-1} + \dots + b_n' \in k[x]$. Let $a_1', \dots, a_n'$ be the roots of h_1 in $\bar{k}$, and σ_t' be the elementary symmetric polynomials corresponding to $a_1', \dots, a_n'$. Then $\sigma_j' = b_j'$, $j = 1, \dots, n$. Let τ_t be the elementary symmetric polynomials corresponding to $a_1, \dots, a_n$, $a_1', \dots, a_n'$, then $h(x)h_1(x) = x^{2n} + \tau_1 x^{2n-1} + \dots + \tau_{2n}$. Therefore $\omega(f \cdot f_1) = \omega(1 + \tau_1 x + \dots + \tau_n x^n + \dots) = G(\tau_1, \dots, \tau_n) = \psi[\sum_{j=1}^{n} F(a_j) + \sum_{j=1}^{n} F(a_j')] = \psi[\sum_{j=1}^{n} F(a_j)]\psi[\sum_{j=1}^{n} F(a_j')] = \omega(f)\omega(f_1)$. As a result, ω is a character of G_0. Moreover, since there exists $a_0 \in k$ such that $\omega(1 - a_0 x) = \psi[F(a_0)] \neq 1$, ω is non-trivial. □

Note that the condition $\psi[F(a_0)] \neq 1$ is only used to ensure ω is non-trivial, and it can be replaced by other conditions. For instance, assume that $p \nmid n$ .Since ψ is non-trivial, there exists $a \in k$ such that $\psi(a) \neq 1$. Let $\sigma_1 = \dots = \sigma_{n-1} = 0$, $\sigma_n = (nc_n)^{-1} a$, then $G(\sigma_1, \dots, \sigma_n) = c_n S_n + \sum_{i=1}^{n-1} c_i S_i = nc_n \sigma_n + H(\sigma_1, \dots, \sigma_n) = a + H(0, \dots, 0) = a$. Therefore $\omega(1 + \sigma_1 x + \dots + \sigma_n x^n) = \psi[G(\sigma_1, \dots, \sigma_n)] = \psi(a) \neq 1$, thus ω is non-trivial.

Now assumme that $p > 2$, $s \neq 0$, then $d_0(\mathfrak{b}) = 2$ for $R_{\mathfrak{b}}(x) = x^2 - s$. Furthermore, set $F(x) = x$, then from the corollary 5 we gain that

$$\left|\sum_{a\in k}\psi(a)\chi_0(a^2 - s)\right| \leq 2\sqrt{q}. \tag{13}$$

Let $\chi_0(a) = 1$ if a is a square in k^*, otherwise $\chi_0(a) = -1$. For $u, v \in k^*$, we will show that the sum in the left hand side of equation (13) is corresponding to the Kloosterman sum $\sum\limits_{x\in k^*}\psi(ux + vx^{-1})$, which plays a crucial role in the representations of numbers in the form $ax^2 + by^2 + cz^2 + dt^2$ [1].

Let $\mathcal{N}(a)$ denote the number of elements x such that $ux + vx^{-1} = a$, or equivalently, the number of solutions of the equation $(2ux - a)^2 = a^2 - 4uv$. Therefore $\mathcal{N}(a) = \chi_0(a^2 - 4uv) + 1$. As a result,

$$\begin{aligned}\sum_{x\in k^*}\psi(ux + vx^{-1}) &= \sum_{a\in k}\psi(a)\mathcal{N}(a) = \sum_{a\in k}\psi(a)\chi_0(a^2 - 4uv) + \sum_{a\in k}\psi(a)\\ &= \sum_{a\in k}\psi(a)\chi_0(a^2 - 4uv). \end{aligned} \tag{14}$$

The last equation is obtained from the lemma 2.

Setting $s = 4uv$ in the inequality (13) and then substituting it into the equation (14), we get a bound for the Kloosterman sum

$$\left|\sum_{x\in k^*}\psi(ux + vx^{-1})\right| \leq 2\sqrt{q}. \tag{15}$$

The coefficient 2 in the left hand side of (15) can be improved. Malyshev[4] proved that it can be replaced by 1 in many cases. In fact, he showed that for general Kloosterman sum

$$K(u, v; c) = \sum_{1\leq x\leq |c|, (c,x)=1} \exp\left(\frac{2\pi i(ux + vx^{-1})}{c}\right),$$

it can be proved that

$$|K(u, v; c)| \leq \min\left\{\sqrt{(u, c)}\tilde{d}\left(\frac{c}{(u, c)}\right), \sqrt{(v, c)}\tilde{d}\left(\frac{c}{(v, c)}\right)\right\}\sqrt{|c|}, \tag{16}$$

where (m, n) denotes the great commom divisor of integers m and n, $\tilde{d}(m)$ denotes the number of positive divisors of integer m. His proof is also based on the Weil bound which we have proved above.

There are some generalizations for the Weil bound as well. For instance, Bombieri[5] generalized the Weil bound by replacing F with a polynomial in n variables. He also proved that the exponent $1/2$ of q can not be improved and found some best coefficient of $\sqrt{q}$ for some spacial cases.

References

[1] Kloosterman, H. D., *On the representation of numbers in the form* $ax^2+by^2+cz^2+dt^2$. Acta mathematica, 1927, 49(3): 407-464.

[2] Weil, A., *On the Riemann hypothesis in function-fields.* Proc. Natl. Acad. Sci. U.S.A., 1941, 27(7): 345-347.

[3] Weil, A., *On some exponential sums.* Proc. Natl. Acad. Sci. U.S.A., 1948, 34(5): 204–207.

[4] Malyshev, A. V., *On the representation of integers by positive quadratic forms.* Trudy Matematicheskogo Instituta imeni VA Steklova, 1962, 65: 3-212.

[5] Bombieri, E. *On exponential sums in finite fields.* American Journal of Mathematics, 1966, 88(1): 71-105.

[6] Stepanov, S. A., *On the number of points of a hyperelliptic curve over a finite prime field.* Mathematics of the USSR-Izvestiya, 1969, 3(5): 1103.

[7] Bombieri, E., *Counting points on curves over finite fields.* Séminaire Bourbaki vol. 1972/73 Exposés 418–435. Springer, Berlin, Heidelberg, 1974: 234-241.

[8] Salvador, G. D. V., *Topics in the theory of algebraic function fields.* Boston, MA: Birkhäuser, 2006.

[9] Stein, E. M., Shakarchi, R., *Complex analysis.* Princeton University Press, 2010.

读书报告

梅文九　吴宇阳

2022 年 6 月 14 日

1　问题引入

我们主要阅读的文章是 H. D. Kloosterman 的 On the representation of numbers in the form $ax^2+by^2+cz^2+dt^2$. 本文主要旨在介绍这篇论文的主要结果, 勾勒出其大致思路, 补充论文中未出现的一些细节.

如果给定正整数 a,b,c,d, 那么哪些自然数 n 可以表示为 $ax^2+by^2+cz^2+dt^2$ 的形式呢? 我们可以自然地将所有四元正整数组 (a,b,c,d) 分为两类: (1) 所有充分大的正整数都可表示成 $ax^2+by^2+cz^2+dt^2$ 的形式; (2) 存在无穷多个正整数 n, 它们都不能写成 $ax^2+by^2+cz^2+dt^2$ 的形式. 事实上, 这篇论文基本回答了这个问题: 利用论文的结果, 对于绝大多数的 (a,b,c,d), 我们都可以判断它属于以上两种情形中的哪一种.

为了证明这个结果, 我们引入记号 $r(n)$, 表示 n 写成 $ax^2+by^2+cz^2+dt^2$ 的不同表示方法数, 这里仅要求 x,y,z,t 是整数. 论文中证明的主要结果是:

定理: 对任意 $\epsilon>0$, 我们有

$$r(n)=\frac{\pi^2}{\sqrt{abcd}}nS(n)+O\left(n^{\frac{17}{18}+\epsilon}\right).$$

$S(n)$ 将在后面小节中定义. 这样, 如果我们对某组 (a,b,c,d) 能说明, n 充分大时有 $S(n)>Kn^{-\epsilon}$, 那么定理中给出的 $r(n)$ 等于两项之和, 由前一项占主导, 于是在 n 充分大时, 一定会有 $r(n)>0$. 这样, 所有充分大的正整数都至少有一种表示成 $ax^2+by^2+cz^2+dt^2$ 的方法.

为了证明上述定理, 我们将给出五个引理, 它们将一并组成一个主引理. 利用这个主引理, 我们将给出定理的证明. 此后, 我们将对不同的 (a,b,c,d) 讨论 $S(n)$ 的估计, 从而给出分类.

2　重要引理

在本小节中, 我们将先列出论文中证明定理所需的各个引理. 其次, 我们将简述它们的证明思路.

我们取 $N=[\sqrt{n}]$, 并把 N 阶 Farey 数列画到 $\mathbb{C}$ 中圆 $\Gamma:|w|=e^{-\frac{1}{n}}$ 上. 事实上, 我们可以先在实数轴上, 利用 N 阶 Farey 数列对 $[0,1]$ 区间进行划分. 具体来说, 先在 $[0,1]$ 上标出所有分母 q 不超过 N 的不可约分数 $\frac{p}{q}$. 注意到对于两个相邻分数 $\frac{p_1}{q_1}<\frac{p_2}{q_2}$, 我们一定有 $p_2q_1-p_1q_2=1$, 从而我们可以在它们之间添上一个新点 $\frac{p_1+p_2}{q_1+q_2}$, 这样, 对于任意内点 $\frac{p}{q}$, 它左右两个新添上的点组成一个区间, 因此它就会唯一对应于一段区间:

$$j_{p,q}=\left(\frac{p}{q}-\frac{1}{q(q+q_1)},\frac{p}{q}+\frac{1}{q(q+q_2)}\right),$$

这里记 $\dfrac{p_1}{q_1},\dfrac{p_2}{q_2}$ 为 $\dfrac{p}{q}$ 左右两个相邻点. 最后, 记 $j_{0,1}=\left(0,\dfrac{1}{N+1}\right),j_{1,1}=\left(\dfrac{N}{N+1},1\right)$, 这样, 所有 $j_{p,q}$ 就形成了 $[0,1]$ 的一个划分. 我们再把线段 $[0,1]$ 粘到 Γ 上. 现在, 每个分数 $\dfrac{p}{q}$ 就会对应 Γ 上的一段圆弧 $\xi_{p,q}$.

我们在课上已经研究过了 $\vartheta(x)$ 函数, 即 $\vartheta(x)=\sum_{n\in\mathbb{Z}}x^{n^2}$. 那么当 $|w|<1$ 时, 我们有

$$1+\sum_{n=1}^{\infty}r(n)w^n=\vartheta(w^a)\vartheta(w^b)\vartheta(w^c)\vartheta(w^d),$$

而且这个收敛是内闭一致收敛的. 我们将利用

$$r(n)=\frac{1}{2\pi i}\int_{\Gamma}\vartheta(w^a)\vartheta(w^b)\vartheta(w^c)\vartheta(w^d)w^{-n-1}$$

来估计 $r(n)$, 因此我们将先来估计 $\vartheta(w^s)$, 这里 $s=a,b,c,d$.

我们进行一些记号的说明. 下文中 s 将不加说明地指代 $s=a,b,c,d$ 中一者. p,q 总是指代两个互素的正整数. 当 v 是正整数时, 我们记

$$S_{p,q,v}=\sum_{j=0}^{q-1}\exp\left(\frac{2p\pi ij^2}{q}+\frac{2v\pi ij}{q}\right).$$

当 $v\equiv 0 \pmod q$ 时, 此即 Gaussain Sum $S_{p,q}$. 高斯和有已知的公式, 我们将在需要用到时给出.

引理: 在圆弧 $\xi_{p,q}$ 上, 我们有

$$\vartheta(w^s)=\varphi_s+\Phi_s,$$

这里

$$\varphi_s=\sqrt{\frac{\pi}{s}}\frac{S_{sp,q}}{q}\left(\frac{1}{n}-i\theta\right)^{-\frac{1}{2}},$$

$$\Phi_s=\frac{2}{q}\sqrt{\frac{\pi}{s}}\left(\frac{1}{n}-i\theta\right)^{-\frac{1}{2}}\sum_{v=1}^{\infty}S_{sp,q,v}\exp\left(-\frac{\pi^2v^2}{sq^2\left(\dfrac{1}{n}-i\theta\right)}\right).$$

证明: 我们课上讲过了 Poisson 求和公式, 即

$$\sum_{n\in\mathbb{Z}}f(x+n)=\sum_{n\in\mathbb{Z}}e^{2\pi inx}\hat{f}(n).$$

那么我们注意到若记 $f(x)=e^{-\pi x^2\alpha}$, 这里 $\alpha\in\mathbb{C}$ 满足 $Re(\alpha)>0$, 那么我们有 $\hat{f}(x)=\dfrac{e^{-\frac{\pi}{\alpha}x^2}}{\sqrt{\alpha}}$. 从而

$$\sum_{l=-\infty}^{+\infty}e^{-\pi(l+\frac{j}{q})^2\alpha}=\frac{1}{\sqrt{\alpha}}\sum_{l=-\infty}^{+\infty}e^{\frac{2\pi ilj}{q}}e^{-\frac{\pi}{\alpha}l^2},$$

我们让 $\alpha=\dfrac{sq^2}{\pi}\left(\dfrac{1}{n}-i\theta\right)$, 我们可以推出

$$\sum_{l=-\infty}^{+\infty}e^{-(ql+j)^2s(\frac{1}{n}-i\theta)}=\frac{1}{q}\sqrt{\frac{\pi}{s}}\left(\frac{1}{n}-i\theta\right)^{-\frac{1}{2}}\sum_{l=-\infty}^{+\infty}e^{\frac{2\pi ilj}{q}}e^{-\frac{\pi^2l^2}{sq^2(\frac{1}{n}-i\theta)}}.$$

注意到 $w\in\xi_{p,q}$, 故我们写出 $w=\exp\left(\dfrac{2p\pi i}{q}-\dfrac{1}{n}+i\theta\right)$, 有

$$\begin{aligned}\vartheta(w^s) &= \sum_{v=-\infty}^{+\infty} w^{sv^2}\\ &= \sum_{v=-\infty}^{+\infty} \exp\left(\frac{2\pi ipv^2 s}{q} - v^2 s\left(\frac{1}{n} - i\theta\right)\right)\\ &=\sum_{j=0}^{q-1}\sum_{l=-\infty}^{+\infty} \exp\left(\frac{2\pi ip(lq+j)^2 s}{q} - (lq+j)^2 s\left(\frac{1}{n} - i\theta\right)\right)\\ &=\sum_{j=0}^{q-1}\exp\left(\frac{2\pi ipj^2 s}{q}\right)\sum_{l=-\infty}^{+\infty}\exp\left(-(lq+j)^2 s\left(\frac{1}{n} - i\theta\right)\right)\\ &=\sum_{j=0}^{q-1}\exp\left(\frac{2\pi ipj^2 s}{q}\right)\frac{1}{q}\sqrt{\frac{\pi}{s}}\left(\frac{1}{n} - i\theta\right)^{-\frac{1}{2}}\sum_{l=-\infty}^{+\infty}\exp\left(\frac{2\pi ilj}{q}\right)\exp\left(-\frac{\pi^2 l^2}{sq^2\left(\frac{1}{n} - i\theta\right)}\right)\\ &=\frac{1}{q}\sqrt{\frac{\pi}{s}}\left(\frac{1}{n} - i\theta\right)^{-\frac{1}{2}}\sum_{j=0}^{q-1}\exp\left(\frac{2\pi ipj^2 s}{q}\right)\left(1+\sum_{v\neq 0}\exp\left(\frac{2\pi ivj}{q}\right)\exp\left(-\frac{\pi^2 v^2}{sq^2\left(\frac{1}{n} - i\theta\right)}\right)\right).\end{aligned}$$

注意到

$$\sum_{j=0}^{q-1}\exp\left(\frac{2\pi ipj^2 s}{q}\right)\exp\left(\frac{2\pi ivj}{q}\right) = S_{sp,q,v},\quad \sum_{j=0}^{q-1}\exp\left(\frac{2\pi ipj^2 s}{q}\right) = S_{sp,q},$$

同时显然交换 j 的符号可知 $S_{sp,q,v} = S_{sp,q,-v}$, 从而我们完成了引理的证明. □

这就需要我们对 $S_{sp,q,v}$ 进行估计. 论文中通过一系列引理对 $S_{sp,q,v}$ 进行了刻画. 我们先列出这些引理.

引理 2.1: 对于给定的 s, q, v, 我们有以下三种情形之一:

(i) $S_{sp,q,v}$ 对所有 p 恒等于 0;

(ii) 可以找到不依赖于 p 的正整数 v'', 使得

$$S_{sp,q,v} = \exp\left(\frac{2\pi ip'v''}{q}\right)S_{sp,q},$$

这里 p' 满足 $p'p+1\equiv 0 \pmod{p}$;

(ii) 可以找到不依赖于 p 的正整数 v'', 使得

$$S_{sp,q,v} = \frac{(s,2)}{2(s,8)}\exp\left(\frac{2\pi ip'v''}{4q}\right)S_{sp,4q},$$

这里 p' 满足 $p'p+1\equiv 0 \pmod{4p}$.

在介绍**引理 2.2** 之前, 我们再引入一批记号. 在今后出现求和号 $\sum$ 时, 有时会写成 $\sum'$ 的形式, 这指代在给定 q 的情形下对所有 $0<p<q, (p,q)=1$ 的 p 进行求和; 若 $\sum'$ 还有下标, 那么指代在给定 q 的情形下对所有 $0<p<q, (p,q)=1$ 且满足下标中所列条件的 p 进行求和.

对于给定的 p, q, 我们可以这么定义出唯一的 p_1:

$$p(p_1+N)+1\equiv 0 \pmod{q},\quad 0<p_1\leqslant q.$$

我们对于某个正整数 $0\leqslant \mu\leqslant q-1$, 再记

$$\sigma_1 = \sideset{}{'}\sum_{p_1\leqslant\mu} S_{ap,q,v_1}S_{bp,q,v_2}S_{cp,q,v_3}S_{dp,q,v_4}\exp\left(-\frac{2n\pi ip}{q}\right).$$

为了简洁, 记

$$\{S_q^p\} = S_{ap,q}S_{bp,q}S_{cp,q}S_{dp,q}.$$

最后, 在下文中出现 K 时, 总是指一个依赖于 $a, b, c, d, \epsilon, n, q$ 的常数, 对 n, q 有一致上界, 它可能在不同的地方指代不同的常数.

引理 2.2: 若 σ_1 不恒为 0, 我们总能找到整数 v, 它依赖于 v_i, a, b, c, d, q, 但不依赖于 p, P, 使得要么

$$\sigma_1 = \sideset{}{'}\sum_{p_1 \leq \mu} \{S_q^p\} \exp\left(\frac{2\pi i u p}{q} + \frac{2\pi i v p'}{q}\right),$$

这里记 $u = -n, pp' + 1 \equiv 0 \pmod q$; 要么

$$\sigma_1 = K \sum_{P_1 \leq \mu} \{S_{4q}^P\} \exp\left(\frac{2\pi i u P}{4q} + \frac{2\pi i v P'}{4q}\right),$$

这里记 $u = -4n, PP' + 1 \equiv 0 \pmod{4q}$, 此时求和就变成要求对所有 $0 < P < 4q, (P, 4q) = 1$ 且 $P_1 \leqslant \mu$ 的 P 求和, 这里 P_1 为唯一满足 $P(P_1 + N) + 1 \equiv 0 \pmod{4q}$ 且 $0 < P_1 \leqslant 4q$ 的整数.

显然**引理 2.2** 可以立即推出下面的

引理 2.2': 我们总有 $\sigma_1 = K\sigma_2$, 这里记

$$\sigma_2 = \sideset{}{'}\sum_{p_1 \leq \mu} \{S_q^p\} \exp\left(\frac{2\pi i u p}{q} + \frac{2\pi i v p'}{q}\right),$$

其中 q 可能是 σ_1 中 q 的一倍或四倍; 相应地, $u = -n$ 或 $-4n$, 而 $1 + pp' \equiv 0 \pmod q$.

从而问题转化为研究 σ_2 的大小.

对于高斯和 $S_{a,c}$, 我们有以下的结果: 当 $(a, c) = 1$ 时,

$$S_{a,c} = \begin{cases} 0, & c \equiv 2 \pmod 4 \\ e_c\sqrt{c}\left(\dfrac{a}{c}\right), & 2 \nmid c \\ (1+i)e_a^{-1}\sqrt{c}\left(\dfrac{c}{a}\right). & 2 \nmid a, 4 \mid c \end{cases}$$

这里我们记

$$e_m = \begin{cases} 1, & m \equiv 1 \pmod 4 \\ i. & m \equiv 3 \pmod 4 \end{cases}$$

这个结果的证明可以参照, 比如说, Bachmann 的 Die analytische Zahlentheorie 2 (1894), 146-187. 于是通过 (较为冗长的) 计算, 我们可以得到

引理 2.3: 我们有

$$\{S_q^p\} = B\left(\frac{p}{Q_aQ_bQ_cQ_d}\right)\zeta(p, q)q^2,$$

这里 Q_s 代表 q_s 的奇数部分. 我们再定义

$$\eta(p, q, s) = \begin{cases} 1, & 2 \nmid q_s = Q_s \\ 0, & q_s \equiv 2 \pmod 4 \\ \exp\left(\dfrac{1}{4}s_q p Q_s \pi i\right), & q_s = 2^{\mu_s}Q_s, 2 \nmid \mu, \mu > 2 \\ 1 + \exp\left(\dfrac{1}{2}s_q p Q_s \pi i\right). & q_s = 2^{\mu_s}Q_s, 2 \mid \mu, \mu \geqslant 2 \end{cases}$$

最后令 $\zeta(p,q)$ 为

$$\zeta(p,q)=\zeta(p,q,a,b,c,d)=\eta(p,q,a)\eta(p,q,b)\eta(p,q,c)\eta(p,q,d).$$

现在我们取 Q 为 q 的奇数部分, G 是 $(a,Q),(b,Q),(c,Q),(d,Q)$ 的最小公倍数. 我们在 $8\mid q; 4\mid q, 8\nmid q; 2\mid q, 4\nmid q; 2\nmid q$ 时分别定义 Λ 为 $8G; 4G; 2G; G$. 那么显然有 $\Lambda\mid q$, 且 Λ 关于所有 q 有一致上界. 从而由**引理 2.3** 我们可以推得

引理 2.3': 我们有

$$|\sigma_2|\leqslant Kq^2\sum_{\lambda=1}^{\Lambda}\left|\sideset{}{'}\sum_{\substack{p_1\leqslant\mu\\ p\equiv\lambda \pmod{\Lambda}}}\exp\left(\frac{2\pi iup}{q}+\frac{2\pi ivp'}{q}\right)\right|.$$

为了沿着**引理 2.3'** 的方向继续估计 $|\sigma_2|$, 我们就需要研究 $S(u,v;\lambda,\Lambda;q)$, 这里

$$S(u,v;\lambda,\Lambda;q)=\sideset{}{'}\sum_{p\equiv\lambda \pmod{\Lambda}}\exp\left(\frac{2\pi iup}{q}+\frac{2\pi ivp'}{q}\right).$$

论文中此后给出了一列引理 (引理 4a-引理 4e), 最后总结得到以下的

引理 2.4: 若 $\Lambda\mid q$, 则

$$S(u,v;\lambda,\Lambda;q)=O\left(q^{\frac{3}{4}+\epsilon}(u,q)^{\frac{1}{4}}\right),\quad S(u,v;\lambda,\Lambda;q)=O\left(q^{\frac{3}{4}+\epsilon}(v,q)^{\frac{1}{4}}\right).$$

最后, 我们还需证明以下的

引理 2.5: 若 $\Lambda\mid q, \mu<q$, 则对

$$\sigma_4=\sideset{}{'}\sum_{p_1\leqslant\mu, p\equiv\lambda \pmod{\Lambda}}\exp\left(\frac{2\pi iup}{q}+\frac{2\pi ivp'}{q}\right),$$

(回忆这里 p',p_1 满足 $1+pp'\equiv 0 \pmod{q}, p'\equiv p_1+N \pmod{q}$), 我们有

$$|\sigma_4|<Kq^{\frac{7}{8}+\epsilon}(u,q)^{\frac{1}{4}}.$$

结合以上结果, 我们就得到了最终的主引理结果:

主引理: 我们有

$$\sideset{}{'}\sum_{p_1\leqslant\mu}S_{ap,q,v_1}S_{bp,q,v_2}S_{cp,q,v_3}S_{dp,q,v_4}\exp\left(-\frac{2n\pi ip}{q}\right)=O\left(q^{2+\frac{7}{8}+\epsilon}(n,q)^{\frac{1}{4}}\right).$$

我们首先来说明**主引理**. 回忆等式左边就是 σ_1, 而它由**引理 2.2'** 被 σ_2 同数量级控制; 结合**引理 2.3'** 和**引理 2.5** 的结果, 我们得到

$$|\sigma_2|\leqslant Kq^2\sum_{\lambda=1}^{\Lambda}|\sigma_4(u,v)|\leqslant Kq^{2+\frac{7}{8}+\epsilon}\Lambda(u,q)^{\frac{1}{4}}.$$

注意到依据情形的不同, $u=-n$ 或 $u=-4n$, q 也可能变成原来的 q 的四倍; 而 Λ 关于 q 有一致上界, 这就利用前边的 5 个引理给出了**主引理**的证明.

我们再返回, 对每个引理加一点说明 (来证明我确实认认真真看完了所有内容, 而且看得非常辛苦). 下文中将记 $q=q_s(s,q), s=s_q(s,q)$, 那么 $(q_s,s_q)=1$.

引理 2.1 的简略证明: 注意到我们有

$$S_{sp,q,v}=\sum_{j=0}^{q-1}\exp\left(\frac{2\pi ispj^2}{q}+\frac{2\pi ivj}{q}\right)=\sum_{j_1=0}^{q_s-1}\exp\left(\frac{2\pi is_qpj_1^2}{q_s}+\frac{2\pi ivj_1}{q}\right)\sum_{\mu=0}^{(s,q)-1}\exp\left(\frac{2\pi iv\mu}{(s,q)}\right),$$

那么当 $(q,s)\nmid v$ 时对所有 p 恒等于 0. 只需讨论 $(q,s)\mid v$ 的情形, 此时设 $v=(q,s)v'$. 那么

$$S_{sp,q,v}=(s,q)S_{s_qp,q_s,v'}.$$

我们可以分成三种情形: $(1)2\nmid q_s$; $(2)2\mid q_s, 2\mid v'$; $(3)2\mid q_s, 2\nmid v'$. 其中 (3) 的情形最为困难, 我们略去前两种情形的说明 (前两种情形对应于结论中的 (ii)), 直接来讨论 (3) 的情形. 取 p'' 满足

$$s_qp(2p''-1)\equiv -v' \pmod{4q_s},$$

那么我们有

$$\begin{aligned}S_{s_qp,q_s,v'}=&\sum_{j=0}^{q_s-1}\exp\left(\frac{2\pi is_qp(j+p'')^2}{q_s}+\frac{2\pi iv'(j+p'')}{q}\right)\\ =&\exp\left(\frac{2\pi is_qpp''^2}{q_s}+\frac{2\pi iv'p''}{q_s}\right)\sum_{j=0}^{q_s-1}\exp\left(\frac{2\pi is_qpj^2}{q_s}+\frac{2\pi ij(v'+2s_qpp'')}{q_s}\right)\\ =&\exp\left(\frac{2\pi is_qpp''^2}{q_s}+\frac{2\pi iv'p''}{q_s}\right)\sum_{j=0}^{q_s-1}\exp\left(\frac{2\pi is_qp(j^2+j)}{q_s}\right).\end{aligned}$$

这在 $4\mid q_s$ 时恒等于 0, 因为

$$\begin{aligned}\sum_{j=0}^{q_s-1}\exp\left(\frac{2\pi is_qp(j^2+j)}{q_s}\right)=&\sum_{j=0}^{q_s-1}\exp\left(\frac{2\pi is_qp\left(\left(j+\frac{q_s}{2}\right)^2+\left(j+\frac{q_s}{2}\right)\right)}{q_s}\right)\\ =&-\sum_{j=0}^{q_s-1}\exp\left(\frac{2\pi is_qp(j^2+j)}{q_s}\right).\end{aligned}$$

从而我们以下设 $q_s\equiv 2 \pmod 4$. 我们注意到对所有 $(a,b)=1, b\equiv 2 \pmod 4$, 我们有 $S_{a,b}=0$, 这是因为

$$S_{a,b}=\sum_{j=0}^{b-1}\exp\left(\frac{2a\pi i\left(j+\frac{b}{2}\right)^2}{b}\right)=\exp\left(\frac{ab\pi i}{2}\right)\sum_{j=0}^{b-1}\exp\left(\frac{2a\pi ij^2}{b}\right)=-S_{a,b}.$$

从而

$$\begin{aligned}\sum_{j=0}^{q_s-1}\exp\left(\frac{2\pi is_qp(j^2+j)}{q_s}\right)=&\exp\left(-\frac{2\pi is_qp}{4q_s}\right)\sum_{j=0}^{q_s-1}\exp\left(\frac{2\pi is_qp(2j+1)^2}{4q_s}\right)\\ =&\exp\left(-\frac{2\pi is_qp}{4q_s}\right)\left(\frac{1}{2}S_{s_qp,4q_s}-S_{s_qp,q_s}\right)\end{aligned}$$

注意到 $S_{s_qp,q_s}=0$, 故我们推出

$$\begin{aligned}S_{s_qp,q_s,v'}=&\frac{1}{2}\exp\left(\frac{2\pi is_qpp''^2}{q_s}+\frac{2\pi iv'p''}{q_s}-\frac{2\pi is_qp}{4q_s}\right)S_{s_qp,4q_s}\\ =&\frac{1}{2}\exp\left(\frac{\pi i(s_qp-v')p''}{q_s}+\frac{2\pi iv'p''}{q_s}-\frac{2\pi is_qp}{4q_s}\right)S_{s_qp,4q_s}\\ =&\frac{1}{2}\exp\left(\frac{2\pi i(2p''-1)s_qp+4\pi ip''v'}{4q_s}\right)S_{s_qp,4q_s}.\end{aligned}$$

现在取 v'', p' 满足

$$v'^2(s,q)\equiv s_qv'' \pmod{4q_s},\quad 1+pp'\equiv 0 \pmod{4q}.$$

那么

$$(2p''-1)s_qp+2p''v'\equiv(-1+2p'')v'\equiv-\frac{v'^2}{s_qp}\equiv-\frac{v''}{(s,q)p}=\frac{v''p'}{(s,q)},\pmod{4q_s}$$

从而

$$S_{s_qp,q_s,v'}=\frac{1}{2}\exp\left(\frac{2\pi ip'v''}{4q}\right)S_{s_qp,4q_s}.$$

最后我们注意到

$$S_{sp,q,v}=(s,q)S_{s_qp,q_s,v'},\quad S_{sp,4q}=(s,q)S_{s_qp,4q_s},$$

这将完成引理的证明. □

注: 论文中最后一步得出来的结果并不同, 我认为这里作者应该是错写出了

$$S_{sp,4q}=(s,4q)S_{s_qp,4q_s},$$

然而我反复检查觉得作者这里出现了纰漏 (不过也不能完全确定, 因为论文跳步明显). 当然, 这一步并没有在大局上影响定理的证明. 另外, **引理 2.1** 的证明比较冗长也是因为我在看时还没有听说 $S_{a,c}$ 有这样广泛的公式; 相应地, 利用这个公式的话, 上述证明会简洁一些. 论文里这样的纰漏其实是不少的, 以下我就不再专门写**注**来批评一番了.

我们不给出**引理 2.2** 的证明, 先延续我们对引理的讨论.

引理 2.3' 的简略证明: 这是因为记 $Q=(Q,s)Q_s$, 那么

$$\left(\frac{p}{Q_s}\right)=\left(\frac{p}{Q}\right)\left(\frac{p}{(Q,s)}\right),$$

我们推出

$$\left(\frac{p}{Q_aQ_bQ_cQ_d}\right)=\left(\frac{p}{(Q,a)}\right)\left(\frac{p}{(Q,b)}\right)\left(\frac{p}{(Q,c)}\right)\left(\frac{p}{(Q,d)}\right),$$

从而由于 $(Q,s)\mid\Lambda$, 我们有

$$\left(\frac{p+\Lambda}{Q_aQ_bQ_cQ_d}\right)=\left(\frac{p}{Q_aQ_bQ_cQ_d}\right).$$

同时注意到根据定义我们有 $\eta(p,q,s)=\eta(p+\Lambda,q,s)$, 于是

$$\zeta(p+\Lambda,q)=\zeta(p,q).$$

因此由**引理 2.3** 的结果, 我们可以让求和中的 p 依照模 Λ 分类, 从而有

$$\sigma_2=Bq^2\sum_{\lambda=1}^{\Lambda}\left(\frac{\lambda}{Q_aQ_bQ_cQ_d}\right)\zeta(\lambda,q)\sideset{}{'}\sum_{p_1\leqslant\mu,p\equiv\lambda\pmod{\Lambda}}\exp\left(\frac{2\pi iup}{q}+\frac{2\pi ivp'}{q}\right),$$

注意到 $\zeta(\lambda,q)$ 是一致有界的, 从而我们完成了证明. □

现在我们回头看: 事实上, 我们不给出**引理 2.2** 的证明, 是因为**引理 2.2** 本身是不正确的. 事实上, 原论文中的

$$S_{sP,q}=\frac{1}{2}S_{sP,4q}$$

这一步是有问题的, 我们可以给出反例. 为了解决这个问题, 我们进行了大量的探索, 包括查阅文献, 在 Mathematics Stack Exchange 上提问等等, 但都没有进展. 后来, 经过大量思考, 我们发现, 这里我们只需要还是与**引理 2.3'** 中的证明一样, 找到一个有一致上界且整除 q 的 Λ, 然后说明那些所谓的余项以 Λ 为周期.

具体来说, 利用**引理 2.1** 的结果, 我们可以得到

$$\sigma_1 = K \sum_{p_1 \leqslant \mu}{}' S_{ap,q_1} S_{bp,q_2} S_{cp,q_3} S_{dp,q_4} \exp\left(-2\pi i p' \left(\frac{v_1''}{q_1} + \frac{v_2''}{q_2} + \frac{v_3''}{q_3} + \frac{v_4''}{q_4}\right)\right) \exp\left(-\frac{2n\pi i p}{q}\right)$$

这里 q_1, q_2, q_3, q_4 分别对应 $s = a, b, c, d$ 时应用**引理 2.1** 所得的结果, 也就是依据 s, q, v_i 的取值可能取 q 或 $4q$. 当然, 如果全是 q 或 $4q$ 的话, 我们依照原本论文中的思路就不会遇到问题. 因此我们下面假设 σ_1 不取 0; 同时 q_1, q_2, q_3, q_4 中既有取 q 者, 又有取 $4q$ 者. 这样, 我们有某个 q_s 是模 4 余 2 的. 我们记 $q_0 = 4q$, 这样有 $8 \mid q_0$. 我们设 q 的奇数部分, 也就是 q_0 的奇数部分是 Q. 还是取 G 是 $(a, Q), (b, Q), (c, Q), (d, Q)$ 的最小公倍数, 那么我们这时可以取 $\Lambda = 8G$, 这样我们还是有 $\Lambda \mid q_0$. 同时 $\Lambda \leqslant 8abcd$, 故有一致上界. 此时我们只需验证以下性质:

$$\frac{1}{\sqrt{q}} S_{sp,q_i} \left(\frac{p}{Q_s}\right)$$

关于 p 是以 Λ 为周期且有一致上界的. 这样我们还是可以关于 p 模 Λ 进行分类, 从而得到**引理 2.3** 中的估计, 这样就修补了这里的错误.

我们进行分类讨论就可以说明该性质. 我们鉴于篇幅问题只说明一种情形: 即我们得到的是 $S_{sp,q}$, 这里 $4 \mid q_s$. 我们设 $q_s = 2^{\mu_s} Q_s$. 那么我们由求和公式, 得到

$$\begin{aligned}
S_{sq,p} &= (s,q) S_{s_q p, q_s} \\
&= (s,q)(1+i) e_{s_q p}^{-1} \sqrt{q_s} \left(\frac{Q_s}{s_q}\right) \left(\frac{Q_s}{p}\right) \left(\frac{2^{\mu_s}}{s_q p}\right) \\
&= \sqrt{(s,q)} (1+i) e_{s_q p}^{-1} \sqrt{q} \left(\frac{Q_s}{s_q}\right) \left(\frac{p}{Q_s}\right) (-1)^{\frac{(p-1)(Q_s-1)}{2}} \left((-1)^{\frac{(s_q p)^2-1}{8}}\right)^{\mu_s}
\end{aligned}$$

其中用到了二次互反律和 2 是否为二次剩余判定等等. 现在我们发现, 由于 Λ 是 8 的倍数, 而显然 $e_{s_q p}^{-1}, (-1)^{\frac{(p-1)(Q_s-1)}{2}}, \left((-1)^{\frac{(s_q p)^2-1}{8}}\right)^{\mu_s}$ 只与 p 模 8 的余数有关, 这就完成了这种情形下的讨论. 其它情形的讨论是完全类似的.

总之, 我们还是有

$$\sigma_1 = K q_0^2 \sum_{\lambda=1}^{\Lambda} \sum_{\substack{p_1 \leqslant \mu \\ p \equiv \lambda \pmod{\Lambda}}} \exp\left(\frac{2\pi i u p}{q_0} + \frac{2\pi i v p'}{q_0}\right).$$

由此我们还是可以推出**引理 2.3**.

接下来我们尝试给出**引理 2.4** 的证明思路. 这可以分解为以下几步:

引理 2.4.1: 若 $\Lambda_1 \mid q_1, \Lambda_2 \mid q_2, (q_1, q_2) = 1$, 则

$$S(u, v_1; \lambda_1, \Lambda_1; q_1) S(u, v_2; \lambda_2, \Lambda_2; q_2) = S(u, v_1 q_2^2 + v_2 q_1^2; \lambda_1 q_2 + \lambda_2 q_1, \Lambda_1 \Lambda_2; q_1 q_2).$$

引理 2.4.2: 设 $q = w_1^{\xi_1} w_2^{\xi_2} \cdots w_r^{\xi_r}$, 若 $(u, q) = (v, q) = 1, \Lambda = w_1^{\zeta_1} w_2^{\zeta_2} \cdots w_r^{\zeta_r}$, 这里 $\zeta_j \leqslant \xi_j$, ζ_j 可能等于 0. 此时, 可以取一列 v_j, λ_j, 使得 $(v_j, w_j^{\xi_j}) = 1$, 且

$$S(u, v; \lambda, \Lambda; q) = \prod_{j=1}^{r} S(u, v_j; \lambda_j, w_j^{\zeta_j}; w_j^{\xi_j}).$$

这两个引理从直观上就可以看出该怎么证, 利用中国剩余定理即可, 鉴于篇幅不再赘述. 这两个引理告诉我们, u, v 都与 q 互素时, 相当于只需考虑 q 是素数的幂的情形.

引理 2.4.3: 若 $q = w^{\xi}, \Lambda = w^{\zeta}, \zeta \leqslant \xi, (u, w) = (v, w) = 1$, 则

$$|S(u, v; \lambda, \Lambda; q)| < K q^{\frac{3}{4}}.$$

引理 2.4.3 的证明: 我们考虑下式

$$\sigma_3 = \sum_{\lambda}{}' \sum_{u}{}' |S(u, v; \lambda, \Lambda; q)|^4,$$

这里求和号 $\sum'$ 分别指对所有小于 Λ 且与之互素的正整数 λ 求和; 以及对所有小于 q 且与之互素的正整数 u 求和.

断言: σ_3 不依赖于 v 的取值.

断言的证明: 我们写出

$$P \equiv up \pmod q, \quad 1 + PP' \equiv 0 \pmod q,$$

那么 $p' \equiv P'u \pmod q$, 而我们需要对所有 $P \equiv u\lambda \pmod \Lambda$ 进行求和, 即我们有

$$\sum_{\lambda}{}' \sum_{u}{}' \left| \sum_{P \equiv u\lambda \pmod \Lambda}{}' \exp\left(\frac{2\pi i P}{q} + \frac{2\pi i u v P'}{q} \right) \right|^4 = \sum_{\lambda}{}' \sum_{u}{}' |S(1, uv; u\lambda, \Lambda; q)|^4 = \sum_{\lambda}{}' \sum_{u}{}' |S(1, u; \lambda, \Lambda; q)|^4.$$

这是因为既然 λ 取遍所有小于 Λ 且与之互素的正整数, 那么 $u\lambda$ 显然也一样, 于是我们可以在上式中把 $u\lambda$ 改成 λ; 类似地, 我们也可以再把 uv 改成 u, 这就说明了断言.

现在把模的平方写成复数乘自己的共轭, 我们得到

$$\begin{aligned} \sigma =& \sum_{\lambda}{}' \sum_{u}{}' \sum_{p_1, p_2, \pi_1, \pi_2} \exp\left(\frac{2\pi i u (p_1 + p_2 - \pi_1 - \pi_2)}{q} + \frac{2\pi i v (p_1' + p_2' - \pi_1' - \pi_2')}{q} \right) \\ =& \sum_{\lambda}{}' \sum_{p_1, p_2, \pi_1, \pi_2} \exp\left(\frac{2\pi i v H'}{q} \right) c_q(H). \end{aligned}$$

这里第二行中是写出 $H = p_1 + p_2 - \pi_1 - \pi_2, H' = p_1' + p_2' - \pi_1' - \pi_2'$, 并对 u 求和, 并注意到我们有公式

$$c_q(H) = \sum_{p}{}' \exp\left(\frac{2\pi i p H}{q} \right) = \sum_{\delta | (H, q)} \delta \mu \left(\frac{q}{\delta} \right).$$

这里 μ 是 Möbius 函数, 即

$$\mu(n) = \begin{cases} 1 & n = 1; \\ 0 & n\text{被某个质数的平方整除}; \\ (-1)^k & n\text{是}k\text{个互异质数之积}. \end{cases}$$

我们注意到 $q = w^{\xi}$, 那么 $v_w(H) < \xi - 1$ 时 $c_q(H) = 0$; $v_w(H) = \xi - 1$ 时 $c_q(H) = -w^{\xi-1}$; $v_w(H) \geqslant \xi - 1$ 时 $c_q(H) = w^{\xi} - w^{\xi-1} = \varphi(q)$, 因此我们有

$$\sigma_3 = -w^{\xi-1} \sum_{\lambda}{}' \sum_{\substack{p_1, p_2, \pi_1, \pi_2 \\ v_w(H) = \xi - 1}} \exp\left(\frac{2\pi i v H'}{q} \right) + \varphi(q) \sum_{\lambda}{}' \sum_{\substack{p_1, p_2, \pi_1, \pi_2 \\ q | H}} \exp\left(\frac{2\pi i v H'}{q} \right).$$

现在, 注意到 σ_3 与 v 无关, 因此我们在上式中对所有 v 求和, 得到

$$\begin{aligned} \varphi(q)\sigma_3 =& - w^{\xi-1} \sum_{\lambda}{}' \sum_{\substack{p_1, p_2, \pi_1, \pi_2 \\ v_w(H) = \xi - 1}} c_q(H') + \varphi(q) \sum_{\lambda}{}' \sum_{\substack{p_1, p_2, \pi_1, \pi_2 \\ q | H}} c_q(H') \\ =& w^{2\xi-2} N_1 - w^{\xi_1} \varphi(q) N_2 - w^{\xi_1} \varphi(q) N_3 + (\varphi(q))^2 N_4, \end{aligned}$$

这里

$N_1=\sum'_\lambda N_{1,\lambda}$, 这里 $N_{1,\lambda}$ 是满足 $p_1,p_2,\pi_1,\pi_2\equiv\lambda\ (\mathrm{mod}\ \Lambda)$, 且 $H,H'\equiv 0\ (\mathrm{mod}\ w^{\xi-1}), q\nmid H,H'$ 的四元解 (p_1,p_2,π_1,π_2) 的个数 (当然指在模 q 意义下);

$N_2=\sum'_\lambda N_{2,\lambda}$, 这里 $N_{2,\lambda}$ 是满足 $p_1,p_2,\pi_1,\pi_2\equiv\lambda\ (\mathrm{mod}\ \Lambda)$, 且 $H\equiv 0\ (\mathrm{mod}\ w^{\xi-1}), q\nmid H, q\mid H'$ 的四元解 (p_1,p_2,π_1,π_2) 的个数;

$N_3=\sum'_\lambda N_{3,\lambda}$, 这里 $N_{3,\lambda}$ 是满足 $p_1,p_2,\pi_1,\pi_2\equiv\lambda\ (\mathrm{mod}\ \Lambda)$, 且 $H'\equiv 0\ (\mathrm{mod}\ w^{\xi-1}), q\nmid H', q\mid H$ 的四元解 (p_1,p_2,π_1,π_2) 的个数;

$N_4=\sum'_\lambda N_{4,\lambda}$, 这里 $N_{4,\lambda}$ 是满足 $p_1,p_2,\pi_1,\pi_2\equiv\lambda\ (\mathrm{mod}\ \Lambda)$, 且 $q\mid H,H'$ 的四元解 (p_1,p_2,π_1,π_2) 的个数.

于是我们有

$$\varphi(q)\sigma_3\leqslant w^{2\xi-2}N_1+(\varphi(q))^2N_4,$$

我们下面来说明 $N_4=O(q^2), N_1=O(w^{2\xi+2})$. 我们先来估计 N_4, 我们去掉要求 p_1,p_2,π_1,π_2 全部模 Λ 同余的条件, 因此只需考虑

$$p_1+p_2\equiv\pi_1+\pi_2\ (\mathrm{mod}\ q),\quad p_1'+p_2'\equiv\pi_1'+pi_2'\ (\mathrm{mod}\ q).$$

第二个式子等价于 $\pi_1\pi_2(p_1+p_2)\equiv p_1p_2(\pi_1+\pi_2)\ (\mathrm{mod}\ q)$, 从而结合第一个式子, 得到要么

$$p_1+p_2\equiv 0\ (\mathrm{mod}\ q),\quad \pi_1+\pi_2\equiv 0\ (\mathrm{mod}\ q);$$

要么

$$p_1p_2\equiv\pi_1\pi_2\ (\mathrm{mod}\ q),\quad p_1+p_2\equiv\pi_1+\pi_2\ (\mathrm{mod}\ q)$$

第一种情形显然只有 $O(q^2)$ 组解; 第二种情形又可以推出 $(p_1-p_2)^2\equiv(\pi_1-pi_2)^2\ (\mathrm{mod}\ q)$, 于是 $p_1-p_2\equiv\pm(\pi_1-\pi_2)\ (\mathrm{mod}\ q)$, 于是又只有 $O(q^2)$ 组解, 从而 $N_4\leqslant Kq^2$.

对于 N_1, 我们还是一样地减弱要求, 一样地得到在模 $w^{\xi-1}$ 的情形下有 $O(w^{2\xi-2})$; 因此在模 q 意义下解的个数为 $O(w^{2\xi+2})$.

因此我们可以得到估计

$$\varphi(q)\sigma_3\leqslant Kw^{2\xi-2}w^{2\xi+2}+Kq^2+q^2\leqslant Kq^4.$$

注意到 $\varphi(q)=w^{\xi-1}(w-1)$, 因此 $\sigma_3\leqslant Kq^3$. 那么我们推出

$$|S(u,v;\lambda,\Lambda;q)|<Kq^{\frac{3}{4}}.$$

□

引理 2.4.4: 若 $\Lambda\mid q,(u,q)=(v,q)=1$, 则

$$S(u,v;\lambda,\Lambda;q)=O(q^{\frac{3}{4}+\epsilon}).$$

引理 2.4.4 的证明: 从**引理 2.4.2** 及**引理 2.4.3**, 我们得到 $q=w_1^{\xi_1}w_2^{\xi_2}\cdots w_r^{\xi_r}$ 时,

$$|S(u,v;\lambda,\Lambda;q)|\leqslant K^rq^{\frac{3}{4}+\epsilon}.$$

现在注意到

$$K^r<2^{Kr}\leqslant[(1+\xi_1)(1+\xi_2)\cdots(1+\xi_r)]^K,$$

而 $(1+\xi_1)(1+\xi_2)\cdots(1+\xi_r)$ 是 q 的因子个数, 从而很容易说明是 $O(q^\epsilon)$ 的. 因此 $K^r=O(q^\epsilon)$, 于是我们完成了证明. □

引理 2.4.5: 若 $\Lambda \mid q, (u,q)=1$, 则

$$S(u,v;\lambda,\Lambda;q)=O(q^{\frac{3}{4}+\epsilon}).$$

引理 2.4.5 的简略证明: 我们还是略去一些过程, 直接断言可以找到一列 v_j,λ_j,ξ_j', 使得

$$S(u,v;\lambda,\Lambda;q)=\prod_{j=1}^{r}S(u,v_j;\lambda_j,w_j^{\zeta_j};w_j^{\xi_j}),\quad (v,q)=\prod_{j=1}^{r}(v_j,w_j^{\xi_j})=\prod_{j=1}^{r}w_j^{\xi_j'}.$$

这里 $\xi_j'\leqslant\xi_j$, 且 ξ_j 可以取 0. 如对于那些 $\xi_j'=0$ 的 j, 我们已经有 $(v_j,w_j^{\xi_j})=1$, 从而由**引理 2.4.3** 已经有

$$\left|S\left(u,v_j;\lambda_j,w_j^{\zeta_j};w_j^{\xi_j}\right)\right|<Kw_j^{\frac{3}{4}\xi_j}. \tag{1}$$

我们再来考虑那些 $\xi_j'=\xi_j$ 的 j. 此时 $w_j^{\xi_j}\mid v_j$. 因此我们有

$$S\left(u,v_j;\lambda_j,w_j^{\zeta_j};w_j^{\xi_j}\right)=\sum_{p\equiv\lambda \pmod{w_j^{\zeta_j}}}\exp\left(\frac{2\pi iup}{w_j^{\xi_j}}\right).$$

这在 $\zeta_j=0$ 时是 $c_{w_j^{\xi_j}}(u)=\mu(w_j^{\xi_j})$, 从而 (1) 这样的估计还是成立的. 下面设 $\zeta_j\neq0$, 那么我们有

$$S\left(u,v_j;\lambda_j,w_j^{\zeta_j};w_j^{\xi_j}\right)=\sum_{v=0}^{w_j^{\xi_j-\zeta_j}-1}\exp\left(\frac{2\pi iu\lambda_j}{w_j^{\xi_j}}\right)\exp\left(\frac{2\pi iuv}{w_j^{\xi_j-\zeta_j}}\right),$$

这在 $\xi_j\neq\zeta_j$ 时等于 0, 而 $\xi_j=\zeta_j$ 时它只剩一项, 模长为 1, 故 (1) 仍然成立.

最后还剩 $0<\xi_j'<\xi_j$ 的情形. 我们记 $v_j=w_j^{\xi_j'}v_j'$, 那么

$$S\left(u,v_j;\lambda_j,w_j^{\zeta_j};w_j^{\xi_j}\right)=\sum_{p\equiv\lambda_j \pmod{w_j^{\zeta_j}}}\exp\left(\frac{2\pi iup}{w_j^{\xi_j}}+\frac{2\pi iv_j'p'}{w_j^{\xi_j-\xi_j'}}\right).$$

我们分下面三种情形讨论.

(i) 当 $\zeta_j=\xi_j-\xi_j'$ 时, 我们可以取 λ_j' 满足 $1+\lambda_j\lambda_j'\equiv0 \pmod{w_j^{\xi_j-\xi_j'}}$. 那么

$$S=\exp\left(\frac{2\pi iu\lambda_j}{w_j^{\xi_j}}+\frac{2\pi iv_j'\lambda_j'}{w_j^{\xi_j}}\right)\sum_{v=0}^{w_j^{\xi_j'-1}}\exp\left(\frac{2\pi iuv}{w_j^{\xi_j'}}\right)=0.$$

(ii) 当 $\zeta_j<\xi_j-\xi_j'$ 时, 我们可以让

$$p=p_1+vw_j^{\xi_j-\xi_j'},\quad v=0,1,\cdots,w_j^{\xi_j'}-1,$$

这里我们要求 $p_1\equiv\lambda_j \pmod{w_j^{\zeta_j}}$, 并写出 $1+p_1p_1'\equiv0 \pmod{w_j^{\xi_j-\xi_j'}}$, 那么我们有

$$S\left(u,v_j;\lambda_j,w_j^{\zeta_j};w_j^{\xi_j}\right)=\sum_{\substack{0<p_1<w_j^{\xi_j-\xi_j'}\\ p_1\equiv\lambda_j \pmod{w_j^{\zeta_j}}}}\exp\left(\frac{2\pi iup_1}{w_j^{\xi_j}}+\frac{2\pi iv_j'p_1'}{w_j^{\xi_j-\xi_j'}}\right)\sum_{v=0}^{w_j^{\xi_j'-1}}\exp\left(\frac{2\pi iuv}{w_j^{\xi_j'}}\right)=0.$$

(iii) 当 $\zeta_j>\xi_j-\xi_j'$ 时, 我们可以让

$$p=\lambda_j+vw_j^{\zeta_j},\quad v=0,1,\cdots,w_j^{\xi_j-\zeta_j}-1.$$

因此还是让 λ_j' 满足 $1+\lambda_j\lambda_j' \equiv 0 \pmod{w_j^{\xi_j-\xi_j'}}$, 我们有

$$S = \exp\left(\frac{2\pi i u\lambda_j}{w_j^{\xi_j}} + \frac{2\pi i v_j'\lambda_j'}{w_j^{\xi_j-\xi_j'}}\right) \sum_{v=0}^{w_j^{\xi_j-\zeta_j}-1} \exp\left(\frac{2\pi i u v}{w_j^{\xi_j-\zeta_j}}\right) = 0.$$

从而 (1) 的估计对一切情形都是成立的. 因此我们还是类似地, 将不同素数项对应的上界相乘, 我们就完成了**引理 2.4.5** 的证明. □

引理 2.4 的证明: 我们还是先有

$$S = S(u,v;\lambda,\Lambda;q) = \prod_{j=1}^{r} S(u,v_j;\lambda_j,w_j^{\zeta_j};w_j^{\xi_j}),$$

这里

$$q = \prod_{j=1}^{r} w_j^{\xi_j},\ \Lambda = \prod_{j=1}^{r} w_j^{\zeta_j},\ (v,q) = \prod_{j=1}^{r}(v_j, w_j^{\xi_j}),\ (u,q) = \prod_{j=1}^{r}(u, w_j^{\xi_j}).$$

我们注意到, 那些让 $(u,w_j) = (v_j,w_j) = 1$ 的 j 由**引理 2.4.3** 可以给出估计

$$|S| < K w_j^{\frac{3}{4}\xi_j} = K(u,w_j^{\xi_j})^{\frac{1}{4}} w_j^{\frac{3}{4}\xi_j} = K(v_j,w_j^{\xi_j})^{\frac{1}{4}} w_j^{\frac{3}{4}\xi_j}$$

如果只有 $(u,w_j) = 1$, 我们由**引理 2.4.4** 知也成立; 而如果只有 $(v_j,w_j) = 1$, 那么注意到若取 λ' 满足 $1+\lambda\lambda' \equiv 0 \pmod{\Lambda}$, 我们有

$$S(u,v;\lambda,\Lambda;q) = S(v,u;\lambda',\Lambda;q)$$

从而这样的估计还是成立的.

从而我们只剩下退化情形, 即 u,v 都不与 w_j 互素. 我们由对称性, 不妨假设有 $(u,w_j^{\xi_j}) \geqslant (v,w_j^{\xi_j})$. 我们写出

$$(v_j,w_j^{\xi_j}) = w_j^{\xi_j'},\ (u,w_j^{\xi_j}) = w_j^{\xi_j''},\ \xi_j'' \geqslant \xi_j' > 0.$$

我们记 $v_j = v_j' w_j^{\xi_j'}, u = u' w_j^{\xi_j'}$.

如果 $\xi_j' = \xi_j$, 我们直接粗暴地放出

$$|S| < K w_j^{\xi_j} = K w_j^{\frac{1}{4}\xi_j'} w_j^{\xi_j} = K(v_j,w_j^{\xi_j})^{\frac{1}{4}} w_j^{\frac{3}{4}\xi_j} \leqslant K(u,w_j^{\xi_j})^{\frac{1}{4}} w_j^{\frac{3}{4}\xi_j}.$$

而如果 $\xi_j' < \xi_j$, 那么我们分两种情况讨论.

(i) 若 $\zeta_j \leqslant \xi_j - \xi_j'$, 我们有

$$S = \sideset{}{'}\sum_{\substack{p\equiv\lambda_j \pmod{w_j^{\zeta_j}} \\ 0<p<w_j^{\xi_j}}} \exp\left(\frac{2\pi i u' p}{w_j^{\xi_j-\xi_j'}} + \frac{2\pi i v_j' p'}{w_j^{\xi_j-\xi_j'}}\right) = w_j^{\xi_j'} \sideset{}{'}\sum_{\substack{p\equiv\lambda_j \pmod{w_j^{\zeta_j}} \\ 0<p<w_j^{\xi_j-\xi_j'}}} \exp\left(\frac{2\pi i u' p}{w_j^{\xi_j-\xi_j'}} + \frac{2\pi i v_j' p'}{w_j^{\xi_j-\xi_j'}}\right).$$

那么此时注意到 $(v_j', w_j^{\xi_j-\xi_j'}) = 1$, 我们由**引理 2.4.5** 得到

$$|S| < K w_j^{\xi_j'} w_j^{\frac{3}{4}(\xi_j-\xi_j')+\epsilon} = K w_j^{\frac{1}{4}\xi_j'} w_j^{\frac{3}{4}\xi_j+\epsilon} = K(v_j,w_j^{\xi_j})^{\frac{1}{4}} w_j^{\frac{3}{4}\xi_j+\epsilon} \leqslant K(u,w_j^{\xi_j})^{\frac{1}{4}} w_j^{\frac{3}{4}\xi_j+\epsilon}.$$

(ii) 若 $\zeta_j > \xi_j - \xi_j'$, 我们有

$$S = \exp\left(\frac{2\pi i u'\lambda_j}{w_j^{\xi_j-\xi_j'}} + \frac{2\pi i v_j'\lambda_j'}{w_j^{\xi_j-\xi_j'}}\right) w_j^{\xi_j-\zeta_j},$$

这里我们取 λ_j' 使得 $1+\lambda_j\lambda_j' \equiv 0 \pmod{w_j^{\zeta_j}}$. 因此

$$|S| = w_j^{\xi_j-\zeta_j} < w_j^{\xi_j'} \leqslant (v_j, w_j^{\xi_j})^{\frac{1}{4}} w_j^{\frac{3}{4}\xi_j} \leqslant (u, w_j^{\xi_j})^{\frac{1}{4}} w_j^{\frac{3}{4}\xi_j}.$$

因此无论如何, 我们都有

$$|S| < K(u, w_j^{\xi_j})^{\frac{1}{4}} w_j^{\frac{3}{4}\xi_j+\epsilon}, \quad |S| < K(v_j, w_j^{\xi_j})^{\frac{1}{4}} w_j^{\frac{3}{4}\xi_j+\epsilon}.$$

从而将不同素数幂对应的项乘起来即可得到**引理 2.4** 的证明. □

最后我们来给出最为困难的**引理 2.5** 的证明, 这将结束本小节的内容.

引理 2.5 的证明: 我们考虑 $\xi\eta-$ 平面上的正方形区域 $0<\xi\leqslant 1, 0\leqslant \eta < 1$. 在 $\xi-$ 轴上, 我们取出以下 q 个点:

$$\xi = \frac{1}{q}, \frac{2}{q}, \cdots, \frac{q-1}{q}, 1.$$

在那些使得 $(p_1+N, q)=1$ 的点 $\xi = \dfrac{p_1}{q}$ 处, 我们取其纵坐标为

$$\eta = \frac{up+vp'}{q},$$

这里 $p' \equiv p_1 + N \pmod q, 1+pp' \equiv 0 \pmod q$. 这样, 我们在 $\xi\eta-$ 平面上得到了 $\varphi(q)$ 个点, 我们可以取适合的坐标使得它们都位于正方形区域内.

我们待定某个大数 M, 对 $m=0,1,\cdots,M-1$, 记 M_m 为满足下述条件的 p 的个数:

$$0<p_1\leqslant \mu,\ p\equiv \lambda \pmod \Lambda,\ \frac{m}{M} \leqslant \frac{up+vp'}{q} \leqslant \frac{m+1}{M}.$$

那么我们得知

$$\begin{aligned}
&\sideset{}{'}\sum_{\substack{p_1\leqslant\mu, p\equiv\lambda \pmod\Lambda,\\ \frac{m}{M}\leqslant\frac{up+vp'}{q}\leqslant\frac{m+1}{M}}} \exp\left(\frac{2\pi iup}{q}+\frac{2\pi ivp'}{q}\right)\\
=&M_m\exp\left(\frac{2\pi im}{M}\right) + \sideset{}{'}\sum_{\substack{p_1\leqslant\mu, p\equiv\lambda \pmod\Lambda,\\ \frac{m}{M}\leqslant\frac{up+vp'}{q}\leqslant\frac{m+1}{M}}} \left(\exp\left(\frac{2\pi i(up+vp')}{q}\right) - \exp\left(\frac{2\pi im}{M}\right)\right)\\
=&M_m\exp\left(\frac{2\pi im}{M}\right) + O\left(\frac{M_m}{M}\right),
\end{aligned}$$

从而我们有

$$\sigma_4 = \sideset{}{'}\sum_{\substack{p_1\leqslant\mu\\ p\equiv\lambda \pmod\Lambda}} \exp\left(\frac{2\pi iup}{q}+\frac{2\pi ivp'}{q}\right) = \sum_{m=0}^{M-1} M_m \exp\left(\frac{2\pi im}{M}\right) + O\left(\frac{\mu}{M}\right).$$

为了估计 M_m, 我们考察函数 $f(\xi,\eta)$, 定义如下:

(i) 对于 $0<\xi<\dfrac{\mu}{q}, \dfrac{m}{M}<\eta<\dfrac{m+1}{M}$, 定义 $f(\xi,\zeta)=1$;

(ii) 对于在矩形 $0<\xi<\dfrac{\mu}{q}, \dfrac{m}{M}<\eta<\dfrac{m+1}{M}$ 边界上的 (ξ,ζ), 定义 $f(\xi,\zeta)=\dfrac{1}{2}$;

(iii) 对于矩形 $0<\xi\leqslant 1, 0\leqslant\eta<1$ 中的其它点, 定义 $f(\xi,\zeta)=0$. (当 $m=M-1$ 时, 这一条要改成对于矩形 $0<\xi\leqslant 1, 0<\eta\leqslant 1$ 中的其它点, 定义 $f(\xi,\zeta)=0$.)

(iv) 对 ξ,η 都以 1 为周期进行延拓.

现在我们注意到在矩形每条边上至多两个点, 因此我们有

$$M_m = \sideset{}{'}\sum_{p\equiv\lambda \pmod{\Lambda}} f\left(\frac{p_1}{q}, \frac{up+vp'}{q}\right) + O(1).$$

注意到由二维傅立叶变换, 我们有

$$f(\xi,\eta) = \sum_{h=-\infty}^{+\infty}\sum_{h=-\infty}^{+\infty} a_{h,k}e^{2\pi i\xi h}e^{2\pi i\eta k},$$

这里

$$a_{h,k} = \int_0^1\int_0^1 f(\xi,\eta)e^{-2\pi i\xi h}e^{-2\pi i\eta k}d\xi d\eta.$$

这事实上是可以积出来的. 比如说, $h\neq 0$ 时,

$$a_{h,0} = \int_0^1\int_0^1 f(\xi,\eta)e^{-2\pi i\xi h}d\xi d\eta = \int_0^{\frac{\mu}{q}} d\xi e^{-2\pi i\xi h}\int_{\frac{m}{M}}^{\frac{m+1}{M}} d\eta = -\frac{1}{2\pi ihM}\left(e^{-\frac{2\pi ih\mu}{q}}-1\right).$$

类似地, 我们有

$$a_{0,0} = \frac{\mu}{qM};$$
$$a_{0,k} = -\frac{\mu}{2\pi ikq}\left(e^{-\frac{2\pi ik(m+1)}{M}} - e^{-\frac{2\pi ikm}{M}}\right), \quad (k\neq 0);$$
$$a_{h,k} = -\frac{1}{4\pi^2hk}\left(e^{-\frac{2\pi ih\mu}{q}}-1\right)\left(e^{-\frac{2\pi ik(m+1)}{M}} - e^{-\frac{2\pi ikm}{M}}\right), \quad (h,k\neq 0).$$

我们写出

$$\begin{aligned}M_m =& \sideset{}{'}\sum_{p\equiv\lambda \pmod{\Lambda}}\sum_{h=-\infty}^{+\infty}\sum_{k=-\infty}^{+\infty} a_{h,k}\exp\left(\frac{2\pi ihp_1}{q} + \frac{2\pi ik(up+vp')}{q}\right) + O(1)\\ =&\sum_{h=-H}^{+H}\sum_{k=-H}^{+H}\sideset{}{'}\sum_{p\equiv\lambda \pmod{\Lambda}} + \sideset{}{'}\sum_{p\equiv\lambda \pmod{\Lambda}}\sum_{h=-\infty}^{+\infty}\sum_{|k|>H} + \sum_{p\equiv\lambda \pmod{\Lambda}}\sum_{|h|>H}\sum_{k=-\infty}^{+\infty} + O(1)\\ =&\Sigma_1+\Sigma_2+\Sigma_3+O(1).\end{aligned}$$

(i) 我们先来估计 Σ_1. 我们要对 Σ_1 中每一项关于 h,k 是否等于 0 进行讨论.

Σ_1 中 $h=0, k=0$ 的项是

$$a_{0,0}\sideset{}{'}\sum_{p\equiv\lambda \pmod{\Lambda}} 1 = \frac{\mu}{qM}\sideset{}{'}\sum_{p\equiv\lambda \pmod{\Lambda}} 1 = \frac{\mu}{qM}\varphi_\lambda(q).$$

这里 $\varphi_\lambda(q)$ 是某个依赖于 λ,Λ,q 的数, 它不依赖于 m.

Σ_1 中 $h\neq 0, k=0$ 的项是

$$T_1 = -\frac{1}{2\pi iM}\sum_{0<|h|\leqslant H}\frac{1}{h}\left(\exp\left(-\frac{2\pi ih\mu}{q}\right)-1\right)\exp\left(-\frac{2\pi ihN}{q}\right)\sideset{}{'}\sum_{p\equiv\lambda \pmod{\Lambda}}\exp\left(\frac{2\pi ip'h}{q}\right).$$

由**引理 2.4**, 我们可以得到

$$\begin{aligned}|T_1| \leqslant& \frac{K}{M}\sum_{h=1}^{H}\frac{(h,q)^{\frac14}}{h}q^{\frac34+\epsilon} \leqslant Kq^{\frac34+\epsilon}\sum_{\delta|q}\delta^{\frac14}\sum_{\substack{(h,q)=\delta\\ h\leqslant H}}\frac{1}{h} \leqslant Kq^{\frac34+\epsilon}\sum_{\delta|q}\delta^{-\frac34}\sum_{h_1\leqslant\frac{H}{\delta}}\frac{1}{h_1}\\ \leqslant& Kq^{\frac34+\epsilon}\log H\sum_{\delta|q}1 = O(q^{\frac34+\epsilon}\log H).\end{aligned}$$

Σ_1 中 $h=0, k\neq 0$ 的项是

$$T_2=-\frac{1}{2\pi iq}\sum_{0<|k|\leqslant H}\frac{1}{k}\left(\exp\left(-\frac{2\pi ik(m+1)}{M}\right)-\exp\left(-\frac{2\pi ikm}{M}\right)\right)\sideset{}{'}\sum_{p\equiv\lambda \pmod{\Lambda}}\exp\left(\frac{2\pi ikup}{q}+\frac{2\pi ikvp'}{q}\right).$$

还是由**引理 2.4**, 我们可以得到

$$|T_2|\leqslant K\frac{\mu}{q}\sum_{k=1}^{H}\frac{(ku,q)^{\frac14}}{k}q^{\frac34+\epsilon}\leqslant Kq^{\frac34+\epsilon}(u,q)^{\frac14}\sum_{k=1}^{H}\frac{(k,q)^{\frac14}}{k}=O\left(q^{\frac34+\epsilon}(u,q)^{\frac14}\log H\right).$$

Σ_1 中 $h\neq 0, k\neq 0$ 的项记为 T_3 时, 我们类似地推出有

$$|T_3|\leqslant K\sum_{h=1}^{H}\sum_{k=1}^{H}\frac{1}{hk}(hk,q)^{\frac14}q^{\frac34+\epsilon}\leqslant Kq^{\frac34+\epsilon}(u,q)^{\frac14}\sum_{h=1}^{H}\frac{1}{h}\sum_{k=1}^{H}\frac{(k,q)^{\frac14}}{k}=O\left(q^{\frac34+\epsilon}(u,q)^{\frac14}(\log H)^2\right).$$

这样, 把上面这四种项加起来, 我们得到有

$$\Sigma_1=\frac{\mu}{qM}\varphi_\lambda(q)+O\left(q^{\frac34+\epsilon}(u,q)^{\frac14}(\log H)^2\right).$$

(ii) 我们再来估计 Σ_2 及 Σ_3. 我们取一个小常数 ψ, 如下定义区域 $R_1(\psi)$: 我们定义 $R_1(\psi)$ 为正方形 $0<\xi\leqslant 1, 0\leqslant\eta<1$ 中这 7 块条状区域的并:

$$1^\circ\ 0\leqslant\xi\leqslant\psi;\quad 2^\circ\ \frac{\mu}{q}-\psi\leqslant\xi\leqslant\frac{\mu}{q}+\psi;\quad 3^\circ\ 1-\psi\leqslant\xi\leqslant 1;\quad 4^\circ\ 0\leqslant\eta\leqslant\psi;$$

$$5^\circ\ \frac{m}{M}-\psi\leqslant\xi\leqslant\frac{m}{M}+\psi;\quad 6^\circ\ \frac{m+1}{M}-\psi\leqslant\xi\leqslant\frac{m+1}{M}+\psi;\quad 7^\circ\ 1-\psi\leqslant\eta\leqslant 1.$$

这样, 我们可以发现 $R_1(\psi)$ 中含有的标记点数为 $O(\psi q)$. 我们记 $R_2(\psi)$ 为正方形 $0<\xi\leqslant 1, 0\leqslant\eta<1$ 中剩下的区域, 事实上, 可以观察到 $R_2(\psi)$ 由六个矩形组成. 若 $(\xi,\eta)\in R_2(\psi)$, 我们有

$$\xi>\psi;\ \left|\xi-\frac{\mu}{q}\right|>\psi;\ 1-\xi>\psi;\ \eta>\psi;\ \left|\eta-\frac{m}{M}\right|>\psi;\ \left|\eta-\frac{m+1}{M}\right|>\psi;\ 1-\eta>\psi.$$

我们与论文中保持一致, 简记

$$\xi=\frac{p_1}{q},\ \eta=\frac{up+vp'}{q},$$

那么我们有

$$\begin{aligned}|\Sigma_2|=&\left|\sideset{}{'}\sum_{p\equiv\lambda \pmod{\Lambda}}\sum_{h=-\infty}^{+\infty}\sum_{|k|>H}a_{h,k}\exp\left(2\pi ih\xi+2\pi ik\eta\right)\right|\\ \leqslant&K\sideset{}{'}\sum_{p\equiv\lambda \pmod{\Lambda}}\left|\sum_{h\neq 0}\frac{1}{h}\left(e^{2\pi ih(\xi-\frac{\mu}{q})}-e^{2\pi ih\xi}\right)\right|\left|\sum_{|k|>H}\frac{1}{k}\left(e^{2\pi ik(\eta-\frac{m+1}{M})}-e^{2\pi ik(\eta-\frac{m}{M})}\right)\right|\\ &+K\sideset{}{'}\sum_{p\equiv\lambda \pmod{\Lambda}}\left|\sum_{|k|>H}\frac{1}{k}\left(e^{2\pi ik(\eta-\frac{m+1}{M})}-e^{2\pi ik(\eta-\frac{m}{M})}\right)\right|\\ \leqslant&K\sideset{}{'}\sum_{p\equiv\lambda \pmod{\Lambda}}\left|\sum_{|k|>H}\frac{1}{k}\left(e^{2\pi ik(\eta-\frac{m+1}{M})}-e^{2\pi ik(\eta-\frac{m}{M})}\right)\right|\\ \leqslant&K\sideset{}{'}\sum_{p\equiv\lambda \pmod{\Lambda}}\left|\sum_{k>H}\frac{\sin 2\pi k\left(\eta-\frac{m+1}{M}\right)-\sin 2\pi k\left(\eta-\frac{m}{M}\right)}{k}\right|.\end{aligned}$$

这里第二个不等号是因为

$$\begin{aligned}\left|\sum_{h\neq 0}\frac{1}{h}\left(e^{2\pi ih(\xi-\frac{\mu}{q})}-e^{2\pi ih\xi}\right)\right| &= \left|\sum_{h=1}^{\infty}\frac{2}{h}\left(\sin 2\pi h\left(\xi-\frac{\mu}{q}\right)-\sin 2\pi h\xi\right)\right| \\ &\leqslant \left|\sum_{h=1}^{\infty}\frac{2\sin 2\pi h\left(\xi-\frac{\mu}{q}\right)}{h}\right| + \left|\sum_{h=1}^{\infty}\frac{2\sin 2\pi h\xi}{h}\right|.\end{aligned}$$

有一致上界.

由于当 $0<\psi\leqslant\theta\leqslant\pi-\psi<\pi$ 时, 我们有

$$\left|\sum_{k>H}\frac{2\pi k\theta}{k}\right|\leqslant\frac{K}{H\psi},$$

从而我们得到对 $R_2(\psi)$ 内的 (ξ,η), 我们有

$$\left|\sum_{k>H}\frac{\sin 2\pi k\left(\eta-\frac{m+1}{M}\right)-\sin 2\pi k\left(\eta-\frac{m}{M}\right)}{k}\right|<\frac{K}{H\psi}.$$

而对于一般的点, 我们也有上式 $<K$. 从而我们取 $\psi=H^{-\frac{1}{2}}$ 时, 我们有

$$\Sigma_2=O\left(\frac{q}{H\psi}\right)+O(q\psi)=O\left(\frac{q}{\sqrt{H}}\right).$$

其中前者是一般点的贡献, 后者是 $R_2(\psi)$ 内的点的贡献.

完全类似地, 我们可以得到 $\Sigma_3=O\left(\frac{q}{\sqrt{H}}\right)$.

因此我们综合上述讨论, 我们得到

$$M_m=\Sigma_1+\Sigma_2+\Sigma_3+O(1)=\frac{\mu}{qM}\varphi_\lambda(q)+O\left(q^{\frac{3}{4}+\epsilon}(u,q)^{\frac{1}{4}}(\log H)^2\right)+O\left(\frac{q}{\sqrt{H}}\right)$$

从而我们取 $H=q$, 就得到

$$M_m=\frac{\mu}{qM}\varphi_\lambda(q)+O\left(q^{\frac{3}{4}+\epsilon}(u,q)^{\frac{1}{4}}\right).$$

那么回到 σ_4, 我们有

$$\sigma_4=\frac{\mu}{qM}\varphi_\lambda(q)\sum_{m=0}^{M-1}e^{\frac{2\pi im}{M}}+O\left(Mq^{\frac{3}{4}+\epsilon}(u,q)^{\frac{1}{4}}\right)+O\left(\frac{q}{M}\right)=O\left(Mq^{\frac{3}{4}+\epsilon}(u,q)^{\frac{1}{4}}\right)+O\left(\frac{q}{M}\right).$$

我们取 $M=\left[q^{\frac{1}{8}}\right]$, 那么我们就得到

$$\sigma_4=O\left(q^{\frac{7}{8}+\epsilon}(u,q)^{\frac{1}{4}}\right)+O\left(q^{\frac{7}{8}}\right)=O\left(q^{\frac{7}{8}+\epsilon}(u,q)^{\frac{1}{4}}\right).$$

这就完成了**引理 2.5** 的证明. □

3 定理证明

在本小节中, 我们将进行主要定理的证明, 即

$$r(n)=\frac{\pi^2}{\sqrt{abcd}}nS(n)+O(n^{\frac{17}{18}+\epsilon}).$$

之前的文章中, 我们提到了一个证明思路, 即利用

$$r(n) = \frac{1}{2\pi i}\int_{\Gamma} \vartheta(w^a)\vartheta(w^b)\vartheta(w^c)\vartheta(w^d)w^{-n-1}dw$$

来估计 $r(n)$. 我们将圆周 Γ 划分为若干小圆弧 $\xi_{p,q}$, 可得

$$r(n) = \frac{1}{2\pi i}\sum_{q=1}^{N}\sideset{}{'}\sum_{p}\int_{\xi_{p,q}} \vartheta(w^a)\vartheta(w^b)\vartheta(w^c)\vartheta(w^d)w^{-n-1}dw$$

其中

$$w = e^{-\frac{1}{n}+\frac{2\pi ip}{q}+i\theta}, \theta \in \left(-\frac{2\pi}{q(q+q'')}, \frac{2\pi}{q(q+q')}\right)$$

由**引理**,

$$\vartheta(w^s) = \varphi_s + \Phi_s$$

对 $s = a, b, c, d$ 成立, 于是我们可以把 $r(n)$ 表达式中每个 $\vartheta(w^s)$ 拆分, 得到 16 项. 我们记

$$J_1 = \frac{1}{2\pi i}\sum_{q=1}^{N}\sideset{}{'}\sum_{p}\int_{\xi_{p,q}} \varphi_a\varphi_b\varphi_c\varphi_d w^{-n-1}dw$$

J_2 为剩余项的和. 下面我们先来估计 J_1. 将 φ_s 的表达式代入, 得

$$J_1 = \frac{\pi^2}{\sqrt{abcd}}\frac{1}{2\pi i}\sum_{q=1}^{N}\sideset{}{'}\sum_{p} q^{-4}\{S_q^p\}\int_{\xi_{p,q}}\left(\frac{1}{n}-i\theta\right)^{-2} w^{-n-1}dw.$$

再通过

$$\int_{\xi_{p,q}} = \int_{\theta=-\frac{2\pi}{q(q+q'')}}^{-\frac{2\pi}{q(q+N)}} + \int_{\theta=-\frac{2\pi}{q(q+N)}}^{\frac{2\pi}{q(q+N)}} + \int_{\theta=\frac{2\pi}{q(q+N)}}^{-\frac{2\pi}{q(q+q')}}$$

将 J_1 分成三段, 记为

$$J_1 = J_{1,1} + J_{1,2} + J_{1,3}$$

对于 $J_{1,2}$, 通过简单计算我们可知

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

代入 $z = \frac{1}{n} - i\theta$, 得到

$$\left(\frac{1}{n}-i\theta\right)^{-2} = F\left(we^{-\frac{2\pi pi}{q}}\right) + O(1)$$

其中

$$F(z) = \frac{z}{(1-z)^2}.$$

记 η 为圆周 Γ 中去掉 $\left(-\dfrac{2\pi}{q(q+N)}, \dfrac{2\pi}{q(q+N)}\right)$ 的部分, 于是

$$\begin{aligned}
J_{1,2} =& \frac{\pi^2}{\sqrt{abcd}} \frac{1}{2\pi i} \sum_{q=1}^{N} \sum_{p}^{\prime} q^{-4} \{S_q^p\} \int_{\theta=-\frac{2\pi}{q(q+N)}}^{\frac{2\pi}{q(q+N)}} F\left(we^{-\frac{2\pi p i}{q}}\right) w^{-n-1} dw + O\left(\sum_{q=1}^{N} \sum_{p}^{\prime} q^{-4} \left|\{S_p^q\}\right| \int_{\theta=-\frac{2\pi}{q(q+N)}}^{\frac{2\pi}{q(q+N)}} |w|^{-n-1} dw\right) \\
=& \frac{\pi^2}{\sqrt{abcd}} \frac{1}{2\pi i} \sum_{q=1}^{N} \sum_{p}^{\prime} q^{-4} \{S_q^p\} \int_{\theta=-\frac{2\pi}{q(q+N)}}^{\frac{2\pi}{q(q+N)}} F\left(we^{-\frac{2\pi p i}{q}}\right) w^{-n-1} dw + O\left(\sum_{q=1}^{N} \sum_{p}^{\prime} \frac{1}{q^2} \frac{1}{qN}\right) \\
=& \frac{\pi^2}{\sqrt{abcd}} \frac{1}{2\pi i} \sum_{q=1}^{N} \sum_{p}^{\prime} q^{-4} \{S_q^p\} \int_{\Gamma} F\left(we^{-\frac{2\pi i p}{q}}\right) w^{-n-1} dw + K \sum_{q=1}^{N} \sum_{p}^{\prime} q^{-4} \{S_q^p\} \int_{\eta} F\left(we^{-\frac{2\pi p i}{q}}\right) w^{-n-1} dw + O\left(\frac{1}{N}\right) \\
=& \frac{n\pi^2}{\sqrt{abcd}} \sum_{q=1}^{N} \sum_{p}^{\prime} q^{-4} \{S_q^p\} e^{-\frac{2np\pi i}{q}} + K \sum_{q=1}^{N} q^{-4} \int_{\eta} \frac{e^{-\frac{1}{n}-i\theta}}{(1-e^{-\frac{1}{n}-i\theta})^2} e^{1-ni\theta} d\theta \sum_{p}^{\prime} \{S_q^p\} e^{-\frac{2np\pi i}{q}} + O\left(\frac{1}{N}\right) \\
=& \frac{\pi^2}{\sqrt{abcd}} nS(n) - \frac{n\pi^2}{\sqrt{abcd}} \sum_{q=N+1}^{\infty} \sum_{p}^{\prime} q^{-4} \{S_q^p\} e^{-\frac{2np\pi i}{q}} + O\left(\sum_{q=1}^{N} q^{-4} \int_{\eta} \frac{\left|e^{-\frac{1}{n}-i\theta}\right|}{\left|1-e^{-\frac{1}{n}-i\theta}\right|^2} d\theta \left|\sum_{p}^{\prime} \{S_q^p\} e^{-\frac{2np\pi i}{q}}\right|\right) + O\left(\frac{1}{N}\right) \\
=& \frac{\pi^2}{\sqrt{abcd}} nS(n) + O\left(n \sum_{q=N+1}^{\infty} q^{-4} \left|\sum_{p}^{\prime} \{S_q^p\} e^{-\frac{2np\pi i}{q}}\right|\right) + O\left(\sum_{q=1}^{N} q^{-4} \int_{\frac{\pi}{qN}}^{\infty} \frac{d\theta}{\frac{1}{n^2}+\theta^2} \left|\sum_{p}^{\prime} \{S_q^p\} e^{-\frac{2np\pi i}{q}}\right|\right) + O\left(\frac{1}{N}\right)
\end{aligned}$$

其中

$$S(n) = \sum_{q=1}^{\infty} \sum_{p}^{\prime} q^{-4} \{S_q^p\} e^{-\frac{2np\pi i}{q}}$$

最后一个等号是因为

$$\frac{\left|e^{-\frac{1}{n}-i\theta}\right|}{\left|1-e^{-\frac{1}{n}-i\theta}\right|^2} = \frac{1}{e^{\frac{1}{n}} + e^{-\frac{1}{n}} - 2\cos\theta} = \frac{1}{\left(e^{\frac{1}{2n}} - e^{-\frac{1}{2n}}\right)^2 + \left(2\sin\dfrac{\theta}{2}\right)^2} = O\left(\frac{1}{\dfrac{1}{n^2}+\theta^2}\right).$$

当 $\theta \in \left(\dfrac{2\pi}{q(q+N)}, \pi\right)$ 时成立, $\theta \in \left(\pi, 2\pi - \dfrac{2\pi}{q(q+N)}\right)$ 时有对称结果.

由**主引理**,

$$\sum_{p}^{\prime} \{S_q^p\} e^{-\frac{2np\pi i}{q}} = O\left(q^{2+\frac{7}{8}+\epsilon} (n,q)^{\frac{1}{4}}\right)$$

代入得

$$\begin{aligned}
J_{1,2} =& \frac{\pi^2}{\sqrt{abcd}} nS(n) + O\left(n \sum_{q=N+1}^{\infty} q^{-\frac{9}{8}+\epsilon} (n,q)^{\frac{1}{4}}\right) + O\left(\sum_{q=1}^{N} N q^{-\frac{1}{8}+\epsilon} (n,q)^{\frac{1}{4}}\right) + O\left(\frac{1}{N}\right) \\
=& \frac{\pi^2}{\sqrt{abcd}} nS(n) + O\left(n \sum_{\delta|n} \delta^{-\frac{7}{8}+\epsilon} \sum_{q_1 \geq \frac{N+1}{\delta}} q_1^{-\frac{9}{8}+\epsilon}\right) + O\left(N \sum_{\delta|n} \delta^{\frac{1}{8}+\epsilon} \sum_{q_1 \leq \frac{N}{\delta}} q_1^{-\frac{1}{8}+\epsilon}\right) + O\left(\frac{1}{N}\right) \\
=& \frac{\pi^2}{\sqrt{abcd}} nS(n) + O\left(n^{1+\epsilon} \sum_{\delta|n} \delta^{-\frac{7}{8}+\epsilon} \left(\frac{N+1}{\delta}\right)^{-\frac{1}{8}}\right) + O\left(n^{\frac{1}{2}+\epsilon} \sum_{\delta|n} \delta^{\frac{1}{8}+\epsilon} \left(\frac{N}{\delta}\right)^{\frac{7}{8}}\right) + O\left(\frac{1}{N}\right) \\
=& \frac{\pi^2}{\sqrt{abcd}} nS(n) + O\left(n^{\frac{15}{16}+\epsilon} \sum_{\delta|n} \delta^{-\frac{3}{4}+\epsilon}\right) + O\left(n^{\frac{15}{16}+\epsilon} \sum_{\delta|n} \delta^{-\dfrac{3}{4}+\epsilon}\right) + O\left(\frac{1}{N}\right) \\
=& \frac{\pi^2}{\sqrt{abcd}} nS(n) + O\left(n^{\frac{15}{16}+\epsilon'}\right)
\end{aligned}$$

下面计算 $J_{1,1}$ 与 $J_{1,3}$. 由对称性, 我们计算 $J_{1,3}$. 首先将 $J_{1,3}$ 以如下方式拆分为 Σ_1 和 Σ_2:

$$J_{1,3} = \frac{\pi^2}{\sqrt{abcd}} \frac{1}{2\pi i} \sum_{q=1}^{N} \sum_{p}^{\prime} q^{-4} \{S_q^p\} \int_{\frac{2\pi}{q(q+N)}}^{\frac{2\pi}{q(q+q')}} \left(\frac{1}{n} - i\theta\right)^{-2} w^{-n-1} dw = K \sum_{q=1}^{N_1} + K \sum_{q=N_1+1}^{N} = K\Sigma_1 + K\Sigma_2$$

其中 N_1 待定. 下面我们分别计算 Σ_1 和 Σ_2:

$$\begin{aligned}
|\Sigma_1| =&K\left|\sum_{q=1}^{N_1}\sideset{}{'}\sum_{p} q^{-4}\{S_q^p\}\int_{\frac{2\pi}{q(q+N)}}^{\frac{2\pi}{q(q+q')}}\left(\frac{1}{n}-i\theta\right)^{-2}w^{-n-1}dw\right|\\
=&K\left|\sum_{q=1}^{N_1}\sideset{}{'}\sum_{p} q^{-4}\{S_q^p\}\int_{\frac{2\pi}{q(q+N)}}^{\frac{2\pi}{q(q+q')}}\left(\frac{1}{n}-i\theta\right)^{-2}e^{-\frac{2\pi npi}{q}-ni\theta}d\theta\right|\\
\leq&K\sum_{q=1}^{N_1}\sideset{}{'}\sum_{p} q^{-4}q^2\int_{\frac{\pi}{qN}}^{\infty}\frac{d\theta}{\frac{1}{n^2}+\theta^2}\\
\leq&K\sum_{q=1}^{N_1}N\frac{\varphi(q)}{q}\\
=&O\left(NN_1\right),
\end{aligned}$$

$$\begin{aligned}
|\Sigma_2| =&K\left|\sum_{q=N_1+1}^{N}\sideset{}{'}\sum_{p} q^{-4}\{S_q^p\}\int_{\frac{2\pi}{q(q+N)}}^{\frac{2\pi}{q(q+q')}}\left(\frac{1}{n}-i\theta\right)^{-2}w^{-n-1}dw\right|\\
=&K\left|\sum_{q=N_1+1}^{N}\sideset{}{'}\sum_{p} q^{-4}\{S_q^p\}\sum_{\mu=q'+q-N}^{q-1}\int_{\frac{2\pi}{q(N+\mu+1)}}^{\frac{2\pi}{q(N+\mu)}}\left(\frac{1}{n}-i\theta\right)^{-2}e^{-ni\theta-\frac{2\pi npi}{q}}d\theta\right|\\
=&K\left|\sum_{q=N_1+1}^{N} q^{-4}\sum_{\mu=1}^{q-1}\int_{\frac{2\pi}{q(N+\mu+1)}}^{\frac{2\pi}{q(N+\mu)}}\left(\frac{1}{n}-i\theta\right)^{-2}e^{-ni\theta}d\theta\sideset{}{'}\sum_{q'+q-N\leq\mu}\{S_q^p\}e^{-\frac{2\pi npi}{q}}\right|.
\end{aligned}$$

记 $p_1 = q' + q - N$. 由法雷数列的性质, $0 < p_1 \leq q$ 且 $p\left(p_1 + N\right) + 1 \equiv 0 \pmod q$.

于是由**主引理**,

$$\sideset{}{'}\sum_{p1\leq\mu}\{S_p^q\}e^{-\frac{2\pi npi}{q}} = O\left(q^{2+\frac{7}{8}+\epsilon}\left(n,q\right)^{\frac{1}{4}}\right)$$

代入得

$$\begin{aligned}
|\Sigma_2| \leq&K\sum_{q=N_1+1}^{N}q^{-4}\sum_{\mu=1}^{q-1}\int_{\frac{2\pi}{q(N+\mu+1)}}^{\frac{2\pi}{q(N+\mu)}}\frac{d\theta}{\frac{1}{n^2}+\theta^2}q^{2+\frac{7}{8}+\epsilon}\left(n,q\right)^{\frac{1}{4}}\\
\leq&K\sum_{q=N_1+1}^{N}q^{-\frac{9}{8}+\epsilon}\left(n,q\right)^{\frac{1}{4}}\int_{\frac{2\pi}{q(N+q)}}^{\frac{2\pi}{q(N+1)}}\frac{d\theta}{\frac{1}{n^2}+\theta^2}\\
=&O\left(\sum_{q=N_1+1}^{N}\frac{\left(n,q\right)^{\frac{1}{4}}}{q^{\frac{9}{8}-\epsilon}}n^2\left(\frac{2\pi}{q(N+1)}-\frac{2\pi}{q(N+q)}\right)\right)\\
=&O\left(n^{1+\epsilon}\sum_{q=N_1+1}^{\infty}\frac{\left(n,q\right)^{\frac{1}{4}}}{q^{\frac{9}{8}}}\right)\\
=&O\left(n^{1+\epsilon}\sum_{\delta|n}\frac{\delta^{\frac{1}{4}}}{\delta^{\frac{9}{8}}}\sum_{q_1\geq\frac{N_1+1}{\delta}}q_1^{-\frac{9}{8}}\right)\\
=&O\left(n^{1+\epsilon'}N_1^{-\frac{1}{8}}\right)
\end{aligned}$$

取 $N_1 = \lfloor n^{\frac{4}{9}}\rfloor$, 则

$$J_{1,3} = O\left(NN_1\right) + O\left(n^{1+\epsilon'}N_1^{-\frac{1}{8}}\right) = O\left(n^{\frac{17}{18}+\epsilon'}\right),$$

同理,

$$J_{1,1} = O\left(n^{\frac{17}{18}+\epsilon'}\right),$$

于是

$$J_1 = J_{1,1} + J_{1,2} + J_{1,3} = \frac{\pi^2}{\sqrt{abcd}} nS(n) + O\left(n^{\frac{17}{18}+\epsilon}\right).$$

下面计算 J_2. 共有 15 项, 我们取其中一项为例进行计算, 如

$$I = \frac{1}{2\pi i}\sum_{q=1}^{N}\sum_{p}{}' \int_{\xi_{p,q}} \varphi_a\varphi_b\Phi_c\Phi_d w^{-n-1}dw$$

类似上面的操作, 利用

$$\int_{\xi_{p,q}} = \int_{\theta=-\frac{2\pi}{q(q+q'')}}^{-\frac{2\pi}{q(q+N)}} + \int_{\theta=-\frac{2\pi}{q(q+N)}}^{\frac{2\pi}{q(q+N)}} + \int_{\theta=\frac{2\pi}{q(q+N)}}^{-\frac{2\pi}{q(q+q')}}$$

我们可以将 I 分成三段 $I = I_1 + I_2 + I_3$.

先计算 I_2, 将 φ_s 与 Φ_s 的公式代入得

$$I_2 = K\sum_{q=1}^{N} q^{-4}\int_{-\frac{2\pi}{q(q+N)}}^{\frac{2\pi}{q(q+N)}} \left(\frac{1}{n} - i\theta\right)^{-2} e^{ni\theta}\sum_{v_3=1}^{\infty}\sum_{v_4=1}^{\infty}\left(\sum_{p}{}' S_{ap,q}S_{bp,q}S_{cp,q,v_3}S_{dp,q,v_4}\right)\exp\left(-\frac{\pi^2(\frac{v_3^2}{c}+\frac{v_4^2}{d})}{q^2(\frac{1}{n}-i\theta)}\right)d\theta,$$

所以

$$|I_2| \le K\sum_{q=1}^{N} q^{-4}\int_{0}^{\frac{2\pi}{q(q+N)}} \frac{1}{\frac{1}{n^2}+\theta^2}\sum_{v_3=1}^{\infty}\sum_{v_4=1}^{\infty}\left|\sum_{p}{}' S_{ap,q}S_{bp,q}S_{cp,q,v_3}S_{dp,q,v_4}\right|\exp\left(-\frac{\pi^2 n(\frac{v_3^2}{c}+\frac{v_4^2}{d})}{q^2(1+n^2\theta^2)}\right)d\theta.$$

取 $l > 0$ 为一个小量, 通过下面的操作我们将上述和分解为三部分:

$$\sum_{q=1}^{N}\int_{\theta=0}^{\frac{2\pi}{q(q+N)}} = \sum_{0<q\le n^{\frac{1}{2}-l}}\int_{\theta=0}^{\frac{1}{qn^{\frac{1}{2}+l}}} + \sum_{0<q\le n^{\frac{1}{2}-l}}\int_{\theta=\frac{1}{qn^{\frac{1}{2}+l}}}^{\frac{2\pi}{q(q+N)}} + \sum_{n^{\frac{1}{2}-l}<q\le N}\int_{\theta=0}^{\frac{2\pi}{q(q+N)}} = \Sigma_1' + \Sigma_2' + \Sigma_3'$$

我们再分别进行估计:

$$\begin{aligned}
\Sigma_1' \le& K\sum_{0<q\le n^{\frac{1}{2}-l}} q^{-4}\frac{n^2}{q(q+N)}q^{2+\frac{7}{8}+\epsilon}(n,q)^{\frac{1}{4}}\sum_{v_3=1}^{\infty}\sum_{v_4=1}^{\infty}\exp\left(-Ln^{2l}\left(v_3^2+v_4^2\right)\right)\\
=& O\left(n^{\frac{3}{2}+\epsilon}\sum_{q=1}^{N}\frac{(n,q)^{\frac{1}{4}}}{q^{\frac{17}{8}-\epsilon}}exp\left(-Ln^{2l}\right)\right)\\
=& O\left(n^3 exp\left(-Ln^{2l}\right)\right)\\
=& O(n^{\frac{17}{18}+\epsilon}),
\end{aligned}$$

$$\begin{aligned}
\Sigma_2' \le& K\sum_{0<q\le n^{\frac{1}{2}-l}} q^{-4}q^{2+\frac{7}{8}+\epsilon}(n,q)^{\frac{1}{4}}\int_{\theta=\frac{1}{qn^{\frac{1}{2}+l}}}^{\frac{2\pi}{q(q+N)}}\frac{d\theta}{\frac{1}{n^2}+\theta^2}\sum_{v_3=1}^{\infty}\sum_{v_4=1}^{\infty}\exp\left(-L\left(v_3^2+v_4^2\right)\right)\\
\le& Kn^{\epsilon}\sum_{0<q\le n^{\frac{1}{2}-l}}\frac{(n,q)^{\frac{1}{4}}}{q^{\frac{9}{8}}}n\int_{\frac{n^{\frac{1}{2}-l}}{q}}^{\infty}\frac{dt}{t^2+1}\\
\le& Kn^{\frac{1}{2}+l+\epsilon}\sum_{0<q\le n^{\frac{1}{2}-l}}\frac{(n,q)^{\frac{1}{4}}}{q^{\frac{1}{8}}}\\
\le& Kn^{\frac{1}{2}+l+\epsilon}\sum_{\delta|n}\delta^{\frac{1}{4}}\delta^{-\frac{1}{8}}\sum_{0<q_1\le\frac{n^{\frac{1}{2}-l}}{\delta}} q_1^{-\frac{1}{8}}\\
\le& Kn^{\frac{1}{2}+l+\epsilon}n^{\frac{7}{8}(\frac{1}{2}-l)}\\
=& O\left(n^{\frac{15}{16}+\epsilon}\right),
\end{aligned}$$

$$
\begin{aligned}
\Sigma_3' \leq & K \sum_{n^{\frac{1}{2}-l}<q\leq N} q^{-4} q^{2+\frac{7}{8}+\epsilon} (n,q)^{\frac{1}{4}} \int_{\theta=0}^{\frac{2\pi}{q(q+N)}} \frac{d\theta}{\frac{1}{n^2}+\theta^2} \sum_{v_3=1}^{\infty}\sum_{v_4=1}^{\infty} \exp\left(-L\left(v_3^2+v_4^2\right)\right) \\
\leq & K \sum_{q>n^{\frac{1}{2}-l}} (n,q)^{\frac{1}{4}} q^{-\frac{9}{8}+\epsilon} n \int_0^{\frac{2\pi n}{q(q+N)}} \frac{dt}{1+t^2} \\
\leq & K n^{1+\epsilon} \sum_{\delta|n} \delta^{\frac{1}{4}} \delta^{-\frac{9}{8}} \sum_{q_1>\frac{n^{\frac{1}{2}-l}}{\delta}} q_1^{-\frac{9}{8}} \\
= & O\left(n^{\frac{15}{16}+\epsilon}\right).
\end{aligned}
$$

于是, $I_2 = O\left(n^{\frac{17}{18}+\epsilon}\right)$.

对于 I_3, 我们可以采用与之前 $J_{1,3}$ 完全相同的计算方法，可得

$$I_3 = O\left(n^{\frac{17}{18}+\epsilon}\right).$$

同理,

$$I_1 = O\left(n^{\frac{17}{18}+\epsilon}\right).$$

于是,

$$I = I_1 + I_2 + I_3 = O\left(n^{\frac{17}{18}+\epsilon}\right).$$

对于 J_2 剩余的项，我们有同样的结论，于是

$$J_2 = O\left(n^{\frac{17}{18}+\epsilon}\right).$$

综上我们即证明了主要定理:

$$r(n) = \frac{\pi^2}{\sqrt{abcd}} n S(n) + O\left(n^{\frac{17}{18}+\epsilon}\right).$$

□

4 后续讨论

下面我们进行命题的后续讨论, 即对 $S(n)$ 的取值进行讨论, 进而分析出 $ax^2+by^2+cz^2+dt^2$ 属于哪一个类别.

我们先回顾一下 $S(n)$ 的定义

$$S(n) = \sum_{q=1}^{\infty} \sum_{p}' q^{-4} \{S_q^p\} e^{-\frac{2np\pi i}{q}}.$$

我们记

$$A_q = \sum_{p}' q^{-4} \{S_q^p\} e^{-\frac{2np\pi i}{q}},$$

则

$$S(n) = \sum_{q=1}^{\infty} A_q.$$

引理 4.1: 对于 $(q,q')=1, s\equiv p \pmod{q'}, s\equiv r \pmod{q}$, 我们有 $\{S_{qq'}^s\} = \{S_{q'}^p\}\{S_q^p\}$.

证明: 由 $\{S_p^q\}$ 的定义, 我们只需证明 $S_{as,qq'} = S_{ap,q'}S_{ar,q}$.

$$\begin{aligned}
S_{ap,q'}S_{ar,q} =& \sum_{m=0}^{q-1} \exp\frac{2ar\pi im^2}{q} \sum_{n=0}^{q'-1} \exp\frac{2ap\pi in^2}{q'} \\
=& \sum_{m=0}^{q-1}\sum_{n=0}^{q'-1} \exp\frac{2a\pi i(q'rm^2+qpn^2)}{qq'} \\
=& \sum_{\substack{j\equiv m \pmod q \\ j\equiv n \pmod{q'}}} \exp\frac{2a\pi isj^2}{qq'} \\
=& \sum_{j=0}^{qq'-1} \exp\frac{2a\pi isj^2}{qq'} \\
=& S_{as,qq'}.
\end{aligned}$$

于是**引理 4.1** 得证. □

引理 4.2: $(q,q')=1$ 时, $A_qA_{q'} = A_{qq'}$.

证明:

$$\begin{aligned}
A_qA_{q'} =& \left(\sum_p{}' \{S_{q'}^p\}e^{-\frac{2np\pi i}{q'}}\right)\left(\sum_r{}' \{S_q^r\}e^{-\frac{2nr\pi i}{q}}\right) \\
=& \sum_p{}'\sum_r{}' \{S_{q'}^p\}\{S_q^r\}e^{-\frac{2n\pi i(pq+rq')}{qq'}} \\
=& \sum_{s\equiv r \pmod q) j\equiv p \pmod{q'}}{}' \{S_{qq'}^s\}e^{-\frac{2n\pi is}{qq'}} \\
=& A_{qq'}.
\end{aligned}$$

□

于是, 由**引理 4.2**,

$$S(n) = \sum_{q=1}^{\infty} A_q = \prod_w \chi_w$$

其中, w 为素数,

$$\chi_w = \sum_{k=0}^{\infty} A_{w^k}$$

记 $\Delta = abcd$, 我们将所有素数分为两类:

(i)$w=2$ 或 $w \mid \Delta$;

(ii)$w \neq 2$ 且 $w \nmid \Delta$.

先对第 (ii) 类的素数求解 χ_w, 对于这些 w, 设 $n = w^l n', w \nmid n'$, 我们有

引理 4.3: 对于第 (ii) 类的 w, 我们有 $\chi_w = \left(1 - \left(\frac{\Delta}{w}\right)\frac{1}{w^2}\right)\left(1 + \sum_{k=1}^{l}\left(\frac{\Delta}{w^k}\right)\frac{1}{w^k}\right)$.

证明: 我们利用高斯和的结论, 可得

$$S_{w^k}^p = w^{2k}\left(\frac{\Delta}{w^k}\right)$$

于是,

$$\begin{aligned} A_{w^k} =& w^{-4k} \sideset{}{'}\sum_{p} \{S_{w^k}^{p}\} e^{\frac{-2np\pi i}{w^k}} \\ =& w^{-2k} \sideset{}{'}\sum_{p} \left(\frac{\Delta}{w^k}\right) e^{-\frac{2np\pi i}{w^k}} \\ =& w^{-2k} \left(\frac{\Delta}{w^k}\right) \sideset{}{'}\sum_{p} e^{-\frac{2n_1 p\pi i}{w^{k-l}}}. \end{aligned}$$

所以

$$A_{w^k} = \begin{cases} \left(\dfrac{\Delta}{w^k}\right) w^{-k-1}(w-1) & k \le l \\ -\left(\dfrac{\Delta}{w^k}\right) w^{-k-1} & k = l+1 \\ 0 & k \ge l+2 \end{cases}$$

故

$$\begin{aligned} \chi_w =& 1 + \sum_{k=1}^{l} \left(\frac{\Delta}{w^k}\right) w^{-k-1}(w-1) - \left(\frac{\Delta}{w^{l+1}}\right) w^{-l-2} \\ =& \left(1 - \left(\frac{\Delta}{w}\right)\frac{1}{w^2}\right) \left(1 + \sum_{k=1}^{l} \left(\frac{\Delta}{w^k}\right) \frac{1}{w^k}\right). \end{aligned}$$

于是**引理 4.3** 得证. □

我们记 $m = \prod_{w|n, w\in(ii)} w^{v_w(n)}$, 同时

$$\chi^{(1)} = \prod_{w\in(i)} \chi_w,$$

$$\chi^{(2)} = \prod_{w\in(ii)} \chi_w.$$

由**引理 4.3**

$$\begin{aligned} \left|\chi^{(2)}\right| =& \left| \prod_{w\in(ii)} \left(1 - (\frac{\Delta}{w})\frac{1}{w^2}\right) \left(1 + \sum_{k=1}^{v_w(m)} (\frac{\Delta}{w^k})\frac{1}{w^k}\right) \right| \\ \ge& \left| \prod_{w\in(ii)} \left(1 - \frac{1}{w^2}\right) \prod_{w\in(ii), w|m} \frac{1 - (\frac{\Delta}{w^{v_w(m)+1}}) w^{-v_w(m)-1}}{1 - (\frac{\Delta}{w}) w^{-1}} \right| \\ \ge& \frac{1}{2} \prod_{w|m} \left(1 - \frac{1}{w}\right) \\ \ge& \frac{\varphi(m)}{2m} \\ >& \frac{K}{\log\log m} \\ \ge& \frac{K}{\log\log n}. \end{aligned}$$

下面计算 $\chi^{(1)}$, 也即计算 χ_2 与 $\chi_w(w \mid \Delta)$.

先考虑 w 为奇素数且 $w \mid \Delta$ 的情况, 设

$$a = w^{\mu_a} a_1, b = w^{\mu_b} b_1, c = w^{\mu_c} c_1, d = w^{\mu_d} d_1$$

不妨设

$$\mu_a \le \mu_b \le \mu_c \le \mu_d.$$

若 $\mu_a \ge 1$, 则对于不是 w 倍数的 n, n 不能表示成 $ax^2+by^2+cz^2+dt^2$. 于是此时的 a,b,c,d 属于第 (2) 类.

若 $\mu_a = 0, \mu_b \ge 1$, 则对于 $n = ax^2+by^2+cz^2+dt^2$, 必有 $n \equiv ax^2 \pmod{w}$, 即 na^{-1} 为 w 的平方剩余. 于是有无穷多个 n 无法被 $ax^2+by^2+cz^2+dt^2$ 表示. 此时 a,b,c,d 也属于第 (2) 类.

对于剩余的 a,b,c,d, 我们有如下结果:

引理 4.4: 对于 $\mu_a = \mu_b = 0, \mu_c \le \mu_d$, 我们有如下结果:

(i)$\mu_c \ge 1, \mu_d \ge 2, \left(\dfrac{ab}{w}\right) = (-1)^{\frac{w+1}{2}}$ 时, 存在一列 n，$\chi_w = 0$;

(ii)$\mu_c = \mu_d = 1, \left(\dfrac{ab}{w}\right) = \left(\dfrac{c_1d_1}{w}\right) = (-1)^{\frac{w+1}{2}}$ 时, 当 $n = w^k$ 时, 存在 $K > 0, \chi_w \sim \dfrac{K}{n}$;

(iii) 对于其它情况, 存在 $K > 0, \chi_w > K, \forall n$. 证明过程略 (利用高斯和结论进行计算，并不困难).

下面对 χ_2 进行类似讨论，设

$$a = 2^{\mu_a}a_1, b = 2^{\mu_b}b_1, c = 2^{\mu_c}c_1, d = 2^{\mu_d}d_1,$$

不妨设

$$\mu_a \le \mu_b \le \mu_c \le \mu_d.$$

若 $\mu_a \ge 1$, 则对于不是 2 倍数的 n, n 不能表示成 $ax^2+by^2+cz^2+dt^2$. 于是此时的 a,b,c,d 属于第 (2) 类.

若 $\mu_a = 0, \mu_b \ge 2$, 则对于 $n = ax^2+by^2+cz^2+dt^2$, 必有 $n \equiv ax^2 \pmod{4}$, 即 na^{-1} 为 4 的平方剩余. 于是有无穷多个 n 无法被 $ax^2+by^2+cz^2+dt^2$ 表示。此时 a,b,c,d 也属于第（2）类.

对于剩余的 a,b,c,d, 我们有如下结果:

引理 4.5: 对于 $\mu_a = 0, \mu_b \ge 1, \mu_c \le \mu_d$, 我们有如下结果:

(i)χ_2 对一列 n 为 0, 如果

$$\begin{aligned}
\mu_a, \mu_b, \mu_c, \mu_d =& 0,1,1,\ge 3;\\
& 0,1,2,\ge 4;\\
& 0,1,\ge 3,\ge 3;\\
& 0,0,\ge 2,\ge 2;\\
& 0,0,0,\ge 3, a \equiv b \equiv c \pmod{4}
\end{aligned}$$

(ii) 对于 $n = 2^k$ 存在 $K > 0, \chi_2 \sim \dfrac{K}{n}$, 如果

$$\begin{aligned}
\mu_a, \mu_b, \mu_c, \mu_d =& 0,1,1,2, \quad a + d_1 \equiv b_1 + c_1 \equiv 4(\ (\mathrm{mod}\ 8)) \quad or \quad b_1 + c_1 + 2a \equiv a + d_1 + 2b_1 \equiv 4 \pmod{8};\\
& 0,1,2,3, \quad b_1 + d_1 \equiv a + c_1 \equiv 4 \pmod{8} \quad or \quad b_1 + d_1 + 2a \equiv a + c_1 + 2b_1 \equiv 4 \pmod{8};\\
& 0,1,1,odd, \quad a + b \equiv c_1 + d_1 \equiv 4 \pmod{8} \quad or \quad a + b + 2c_1 \equiv c_1 + d_1 + 2a \equiv 4 \pmod{8};\\
& 0,0,0,0, \quad a \equiv b \equiv c \equiv d \pmod{4} \quad and \quad a + b + c + d \equiv 4 \pmod{8};\\
& 0,0,0,2, \quad a \equiv b \equiv c \equiv d_1 \pmod{4} \quad and \quad a + b + c + d_1 \equiv 4 \pmod{8}.
\end{aligned}$$

(iii) 对于其它情况，存在 $K > 0, \chi_2 > K, \forall n$.

证明过程略.

综合以上的讨论, 我们可以知道, 如果 a, b, c, d 满足:

1. 不在**引理 4.5** 的 (i)(ii) 中;

2.Δ 的任何素因子不满足**引理 4.4** 的 (i)(ii);

3. 任何奇素数不整除 a, b, c, d 中 ≥ 3 个数;

4.a, b, c, d 中至少一个数为奇数;

5.a, b, c, d 中至少两个数不被 4 整除;

则存在 $K > 0$,

$$S(n) > \frac{K}{\log\log n}.$$

由主要定理, 即可知对于这样的 a, b, c, d, 任意充分大的正整数均可被表示.

如果 a, b, c, d 不满足上面 3/4/5 中任意一个, 则由前面的讨论知有无穷多个 n 不能被表示. 若不满足 1/2, 我们下面进行讨论.

首先我们证明: 若**引理 4.4**(i) 对某个奇素数成立或**引理 4.5**(i) 成立, 则无穷多个 n 不能被表示.

若奇素数 w 满足**引理 4.4** 的 (i)，我们分情况讨论:

①$2 \leq \mu_c \leq \mu_d$, 且 $\left(\dfrac{ab}{w}\right) = (-1)^{\frac{w+1}{2}}$:

我们考虑 $n = wn_1, (n_1, w) = 1$, 下证这样的 n 不能被表示.

对于 $wn_1 = ax^2 + by^2 + cz^2 + dt^2$, 我们可知 $(w, x) = (w, x) = 1$, 否则 $w \mid x, w \mid y$, 推出 $w^2 \mid n$，矛盾! 由 $w \mid ax^2 + by^2$, 我们知 $ab \equiv -b^2 \left(\dfrac{y}{x}\right)^2 \pmod w$ 故 $\left(\dfrac{ab}{w}\right) = \left(\dfrac{-1}{w}\right) = (-1)^{\frac{w-1}{2}}$, 与条件矛盾.

②$\mu_c = 1, \mu_d \geq 2, \left(\dfrac{ab}{w}\right) = (-1)^{\frac{w+1}{2}}$:

此时我们可以完全类似的说明下列 n 无法被表示：

$$n = wn_1, w \nmid n_1, \left(\frac{c_1 n_1}{w}\right) = -1.$$

若**引理 4.5**(i) 成立，则我们可以通过简单讨论说明下列 n 不能被表示：

①$\mu_a, \mu_b, \mu_c, \mu_d = 0, 1, 1, \geq 3$

若 $b_1 + c_1 \equiv 0 \pmod 4$, 则 $n \equiv a + 4 \pmod 8$ 不能被表示;

若 $b_1 + c_1 \equiv 2 \pmod 4$, 则 $n \equiv a + 2b_1 + 4 \pmod 8$ 不能被表示;

②$\mu_a, \mu_b, \mu_c, \mu_d = 0, 1, 2, \geq 4$

若 $a + c_1 \equiv 0 \pmod 4$, 则 $n = 2n_1, n_1 \equiv b_1 + 4 \pmod 8$ 不能被表示;

若 $a + c_1 \equiv 2 \pmod 4$, 则 $n = 2n_1, n_1 \equiv b_1 + 2a + 4 \pmod 8$ 不能被表示;

③$\mu_a, \mu_b, \mu_c, \mu_d = 0, 1, \geq 3, \geq 3$

$n \equiv a + 2b_1 + 4 \pmod 8$ 不能被表示;

④$\mu_a, \mu_b, \mu_c, \mu_d = 0, 0, \geq 2, \geq 2$

若 $a + b \equiv 0 \pmod 4$, 则 $n = 2n_1, 2 \nmid n_1$ 不能被表示;

若 $a + b \equiv 2 \pmod 4$, 则 $n \equiv a + 2 \pmod 4$ 不能被表示;

⑤$\mu_a, \mu_b, \mu_c, \mu_d = 0, 0, 0, \geq 3, a \equiv b \equiv c \pmod 4$

$n \equiv a + b + c + 4 \pmod 8$ 不能被表示.

下面考虑**引理 4.4**(ii) 与**引理 4.5**(ii).

先考虑**引理 4.4**(ii), 即 $\mu_c=\mu_d=1, \left(\dfrac{ab}{w}\right)=\left(\dfrac{c_1d_1}{w}\right)=(-1)^{\frac{w+1}{2}}$ 时, 我们分情况讨论:

①$c_1>1, d_1>1$, 则 $n=w^k, 2\nmid k$ 不能被表示:

证明: 我们先证明一个如下的

引理: $w\nmid A, w\nmid B, \left(\dfrac{AB}{w}\right)=(-1)^{\frac{w+1}{2}}$, 则对于 $\forall X,Y$, 我们有 $2\mid v_w\left(AX^2+BY^2\right)$

引理的证明: 若 $2\nmid v_w\left(AX^2+BY^2\right)$, 我们不妨设 $w\nmid X, w\nmid Y$, 则此时仍有 $w\mid AX^2+BY^2$.

故 $AB\equiv -B^2\left(\dfrac{Y}{x}\right)^2 \pmod{w}$. 于是 $\left(\dfrac{AB}{w}\right)=\left(\dfrac{-1}{w}\right)=(-1)^{\frac{w-1}{2}}$, 与条件矛盾.

回到原题, 若 $w^k=ax^2+by^2+w\left(c_1z^2+d_1t^2\right)$, 由于 k 为奇数, 故由引理知 $c_1z^2+d_1t^2\neq 0$. 又由于 $c_1>1, d_1>1$ 我们可由递降法得知 $ax^2+by^2\neq 0$. 设 $\lambda=v_w\left(ax^2+by^2\right)$, 则由引理, $2\mid\lambda$. 由于 $2\nmid k$, 所以 $k>\lambda$, 且 $w^{k-\lambda}=\dfrac{ax^2+by^2}{w^\lambda}+\dfrac{c_1z^2+d_1t^2}{w^{\lambda-1}}$. 于是 $v_w\left(c_1z^2+d_1t^2\right)=\lambda-1$, 与引理矛盾.

②$c_1=1, d_1\geq 3$, 则 $n=2\cdot w^k, 2\nmid k$ 不能被表示;

③$c_1=1, d_1=2, w\neq 5$, 则 $n=5w^k, 2\nmid k$ 不能被表示;

④$c_1>1, d_1=2, w=5$, 则 $n=7\cdot 5^k, 2\nmid k$ 不能被表示;

⑤$c_1=1, d_1=1, w\neq 3$, 则 $n=3\cdot w^k, 2\nmid k$ 不能被表示;

⑥$c_1=1, d_1=1, w=3, a>1, b>1$, 则 $n=3^{k+1}, 2\nmid k$ 不能被表示;

⑦$c_1=1, d_1=1, w=3, a=1, b>1$, 则 $n=2\cdot 3^{k+1}, 2\nmid k$ 不能被表示.

唯一一种例外情况是 $x^2+y^2+3z^2+3t^2$, 可以表示所有正整数.

上面第②类到第⑦类的证明与第一类完全类似, 就不赘述了. 对于这种例外情况, 证明过程与拉格朗日四平方和定理基本相同, 不是本文的证明重点.

再考虑**引理 4.5**(ii). 我们可以类似上面的过程进行分类讨论, 由于过程太过复杂, 我们在此不给出具体的讨论过程, 只附上最后的结果:

属于第 (1) 类的 $\{a,b,c,d\}$ 有:

$$\{1,2,3,6\},\{1,2,3,22\},\{1,2,3,38\},\{1,2,6,11\},\{1,2,6,19\}$$

$$\{1,1,10,10\},\{1,2,2,9\},\{1,2,2,17\},\{1,2,2,25\},\{1,1,2,18\}$$

$$\{1,2,9,18\},\{1,2,17,18\},\{1,2,18,25\},\{1,1,2,34\},\{1,1,2,50\}$$

$$\{1,2,9,34\},\{1,2,9,50\},\{1,2,17,50\},\{1,1,2,2\},\{1,1,1,1\}$$

$$\{1,1,1,9\},\{1,1,1,17\},\{1,1,1,25\},\{1,1,5,5\},\{1,1,1,36\}$$

$$\{1,1,1,68\},\{1,1,1,100\},\{1,1,4,9\},\{1,1,4,17\},\{1,1,4,25\},\{1,1,5,20\};$$

Kloosterman 未能判定属于哪一类别的 $\{a,b,c,d\}$ 有:

$$\{1,2,17,34\},\{1,2,11,38\},\{1,2,19,38\},\{1,2,19,22\}.$$

这些数组现在也都已经被解决, 都属于第 (1) 类.

其余满足**引理 4.5**(ii) 的 $\{a,b,c,d\}$ 均属于第 (2) 类.

关于这一部分的证明过程，感兴趣的读者可以阅读:

The Completion of a Problem of Kloosterman, by Gordan Pall;

On the expression of a number in the form $ax^2+by^2+cz^2+du^2$, by Ramanujan.

由此我们即完成了整个问题的讨论. 通过解析数论的巧妙计算方法, 我们证明了前面的引理及主要定理, 再经过进一步的讨论, 所有的 $\{a,b,c,d\}$ 都被我们成功分类!

An elementary introduction to Kuznecov's article on modular forms written in 1981

Bowen Xue, Zhenpeng Li

May, 2022

1 Introduction

History has witnessed the fast development of modular forms, which is common in a number of mathematical branches. There is no denying that due to the idea of modular forms, analytical number theory embraces its brand new era. This article is intended to record the results and some proofs roughly, especially those related to analytical number theory. Our main reference is the classic work that belongs to Kuznecov. His estimates, via the use of modular forms and former conclusions, are more precise than his contemporaries'.

In essence, those basic formulae come from the Fourier expansion of some fundamental functions. However, by virtue of the deformation of integrals, we obtain a series of nontrivial results. The reason for our deformation comes from some masters' estimates and rich properties of Bessel function. We will focus on the deformation but skip some inequalities so that we can make our article seem easy.

This article is based on [1] and [2]. They do help us a lot.

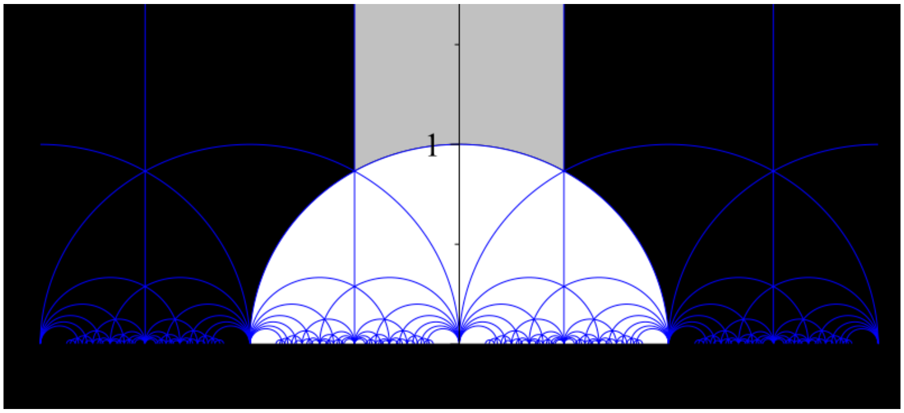
2 Notation

Before we begin our journey, some definitions are important.

Let G be the modular group $PSL(2,\mathbb{Z})$. We equip the upper plane $\mathbb{H}$ with a G-action, that is,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} z = \frac{az+b}{cz+d}.$$

Besides, Laplace operator $\mathcal{L} = -y^2(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2})$ is also frequent. We will consider its eigenfunctions of discrete spectrum, named after cusp forms of weight 0. Here we recognize that they are nontrivial real-analytical automorphic functions, which satisfy the equality $f(gz) = f(z), \forall g \in G$ and finiteness condition $\int_D |f(z)|^2 dz < \infty$, where D is the fundamental field of G and $dz = \frac{dxdy}{y^2}$, the G-invariant measure of $\mathbb{H}$.



The gray domain is the fundamental field.

On the basis of some knowledge on compact operators, we know that the Laplace operator has Lebesgue spectrum of multiplicity one which fills out the semiaxis $\frac{1}{4} \leqslant \lambda < \infty$, and it has a discrete spectrum of finite multiplicity located on the semiaxis $\lambda \geqslant 0$ and having no points of accumulation in every finite interval.

The simplest subgroup of G may be translations $< z \mapsto z + n >$. We call it G_∞.

Hecke defined operators acting on automorphic functions, $T(n), n \in \mathbb{N}_+$. Let us prove some basic propositions.

3 Hecke operators

3.1 Definition and properties

A matrix M is called order n if $\det M = n$. We consider the equivalence relation of M_1 and M_2 if $M_1 = gM_2, g \in G$. It is not difficult to verify that all representatives are of the form

$$\begin{pmatrix} a & b \\ 0 & d \end{pmatrix}, ad = n, d > 0, b = 0, 1, ..., d-1.$$

Hence, the following definition is valid.

Definition 3.1. For every automorphic function f and $n > 0$, we define

$$T(n)(f) = \frac{1}{\sqrt{n}} \sum_{\substack{ad=n \\ d>0}} \sum_{b \bmod d} f\left(\frac{az+b}{d}\right). \tag{3.1}$$

Or equivalently,

$$T(n)(f) = \frac{1}{\sqrt{n}} \sum_{M_i} f(M_i z), \tag{3.2}$$

where M_i runs over all representatives.

It is trivial that the images of Hecke operators are also automorphic. And the following theorem implies that they are commutative.

Theorem 3.2. *n, $m \in \mathbb{N}_+$, then*

$$T(n)T(m) = \sum_{d|(m,n)} T\left(\frac{mn}{d}\right). \tag{3.3}$$

Proof. **Step 1:** If $(m, n) = 1$, then,

$$T(n)T(m)f = \frac{1}{\sqrt{mn}} \sum_{ad=n} \sum_{b \bmod d} \sum_{a'd'=m} \sum_{b' \bmod d'} f\left(\frac{aa'z + a'b + b'd}{dd'}\right). \tag{3.4}$$

Because m and n are coprime, aa' and dd' run over every divisor of mn and $a'b + b'd$ runs over the residue system of dd'. As a result, $T(n)T(m) = T(mn)$.

Step 2: If $m = p$ and $n = p^r$ and p prime, then,

$$T(p)f = \frac{1}{\sqrt{p}} \left(f(pz) + \sum_{b=0}^{p-1} f\left(\frac{z+b}{p}\right)\right). \tag{3.5}$$

Then,

$$p^{\frac{r+1}{2}} T(p^r)T(p)f = \sum_{k=0}^{r} \sum_{t=0}^{p^k-1} \left(f\left(\frac{p^{r-k+1}z + tp}{p^k}\right) + \sum_{b=0}^{p-1} f\left(\frac{p^{r-k} + t + bp^k}{p^{k+1}}\right)\right). \tag{3.6}$$

Similarly, basic number theory tells us that,

$$T(p^r)T(p) = T(p^{r+1})T(p^{r-1}). \tag{3.7}$$

Step 3: If $m = p^s$ and $n = p^r$ and p prime, then we can assume $s \leqslant r$.

If $s < r$, $T(p)T(p^r)T(p^s) = T(p^r)\left(T(p^{s+1}) + T(p^{s-1})\right)$. Hence, we can do induction on $T(p^r)T(p^{s+1})$. If $s = r$, the same as the above case.

All the three steps have told us all. ■

Lemma 3.3. If we let Chebyshev polynomial be

$$U_n(\cos\theta) = \frac{\sin(n+1)\theta}{\sin\theta}, \tag{3.8}$$

then

$$T(p^r) = U_r\left(\frac{1}{2}T(p)\right) = \sum_{0\leqslant k\leqslant r/2} \frac{(-1)^k(r-k)!}{k!(r-2k)!}\left(T(p)\right)^{r-2k}. \tag{3.9}$$

Proof. It is trivial. ■

Corollary 3.4. p prime and $2\cos\theta$ is an eigenvalue of $T(p)$, where $\theta\in\mathbb{C}$. Then $\frac{\sin(r+1)\theta}{\sin\theta}$ is an eigenvalue of $T(p^r)$.

3.2 Inner product of automorphic funtions

Lemma 3.5. $T(n)$ is Hermitian with respect to this inner product of automorphic functions,

$$(f_1, f_2) = \int_D f_1(z)\overline{f_2(z)}dz, \tag{3.10}$$

where D is the fundamental field and dz is the G-invariant measure.

Proof. It suffices to considering the case when p prime. Then,

$$(T(p)f)(z) = \frac{1}{\sqrt{p}}\sum_{j=0}^{p} f(\alpha g_j z) = \frac{1}{\sqrt{p}}\sum_{j=0}^{p} f(\widetilde{\alpha}\widetilde{g_j}z), \tag{3.11}$$

where,

$$\begin{aligned}
\alpha &= \begin{pmatrix}1 & 0\\ 0 & p\end{pmatrix}\\
g_i &= \begin{cases}\begin{pmatrix}1 & j\\ 0 & 1\end{pmatrix} & j=0,...,p-1\\ \begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix} & j=p\end{cases}\\
\widetilde{\alpha} &= \begin{pmatrix}p & 0\\ 0 & 1\end{pmatrix}\\
\widetilde{g_j} &= \begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix} g_j
\end{aligned}$$

As a result, if we change variables by $z' = \alpha g_j z$ and conform the integral domain, we can get,

$$(T(p)f_1, f_2) = \frac{1}{\sqrt{p}} \sum_{j=0}^{p} \int_D f_1(\alpha g_j z)\overline{f_2(z)} dz \tag{3.12}$$

$$= \frac{1}{\sqrt{p}} \int_B f_1(z)\overline{f_2(\alpha^{-1} z)} dz, \tag{3.13}$$

where $B = \bigcup_{j=0}^{p} \alpha g_j D$.

The same as this, we cam get,

$$(f_1, T(p)f_2) = \frac{1}{\sqrt{p}} \int_{\widetilde{B}} f_1(z)\overline{f_2(\widetilde{\alpha} z)} dz, \tag{3.14}$$

where $\widetilde{B} = \bigcup_{j=0}^{p} \widetilde{g}_j D$.

But note that $\alpha^{-1} z = \widetilde{\alpha z} = pz$. So what we only need to do is to compare B and $\widetilde{B}$. We claim that they are both the fundamental field of G_∞ and the proof is reserved for practice. ■

3.3 Relation to eigenfunctions of Laplace operator

We denote ψ as the eigenfunction of the discrete spectrum of Laplace operator equipped with the eigenvalue $\lambda > \frac{1}{4}$. And let $\kappa = \sqrt{\lambda - \frac{1}{4}}$. The Fourier expansion of ψ is clear. And the regular property makes the following formula appropriate,

$$\psi(z) = \sum_{-\infty}^{+\infty} c_n(y) e^{2\pi i n x}. \tag{3.15}$$

So, apply this to characteristic equation, we can get,

$$-y^2 c_n'' + 4\pi^2 n^2 y^2 c_n = \lambda c_n, \tag{3.16}$$

which is the classical Bessel equation. Then,

$$c_n(y) = \rho(n)\sqrt{y} K_{i\kappa}(2\pi|n|y) + \widetilde{\rho}(n)\sqrt{y} I_{i\kappa}(2\pi|n|y), \tag{3.17}$$

with $\rho(n)$ and $\widetilde{\rho}(n)$ to be decided.

But the second component is always unbounded. Finiteness condition request it to be zero.

Simultaneously, when $n = 0$, the concrete calculation suggests that $\rho(0) = 0$. So,

$$\psi(z) = \sum_{n \neq 0} \rho(n) \sqrt{y} K_{i\kappa}(2\pi |n| y) e^{2\pi i n x}. \tag{3.18}$$

Beside, $K_\nu(y)$, considered as a function of ν, is even and entire on the complex plane. When ν purely imaginary and y positive, $K_\nu(y)$ is a real number. As a result, if $\psi(z)$ takes on real value, an extra condition is inevitable,

$$\rho(n) = \overline{\rho(-n)}. \tag{3.19}$$

This part provide a case of Hecke operators acting on special functions.

Lemma 3.6. The same as above and let $n \geqslant 1$, then,

$$(T(n)\psi)(z) = \sum_{m \neq 0} t_n(m) \sqrt{y} K_{i\kappa}(2\pi |n| y) e^{2\pi i n x}, \tag{3.20}$$

where,

$$t_n(m) = \sum_{\substack{d \mid (m,n) \\ d > 0}} \rho\left(\frac{mn}{d^2}\right). \tag{3.21}$$

Proof.

$$T(n)\psi = \frac{1}{\sqrt{n}} \sum_{\substack{ad=n \\ a>0}} \sum_{b \bmod d} \sum_{m \neq 0} \rho(m) \sqrt{\frac{ay}{d}} K_{i\kappa}\left(\frac{2\pi |m| ay}{d}\right) \exp\left(2\pi i m \frac{ax+b}{d}\right), \tag{3.22}$$

and,

$$\sum_{b \bmod d} \exp\left(2\pi i m \frac{b}{d}\right) = \begin{cases} d, & \text{if } d \mid m, \\ 0, & \text{else.} \end{cases}, \tag{3.23}$$

tell us all. ■

Then, using the following proposition, we can choose a special basis in an attempt for simplification.

Proposition 3.7. Hecke operators commute with the Laplace operator.

If V_λ is the λ-characteristic space of $\mathcal{L}$, V_λ is also the invariant space of $T(n)$. We have told readers $\dim V_\lambda < \infty$ and $T(n)$ Hermitian and commutative. Hence, we pose induction on $T(n)$ and make them diagonal under some basis. Note that the dimension is finite, our induction will stop finally. That means, we can choose a basis such that every $T(n)$ act as a stretch on them. That is, if ψ_j is the eigenfunction of λ_j and,

$$T(n)\psi_j = \mu_j(n)\psi_j, \tag{3.24}$$

we can get,

$$\psi_0(n) = \text{const.} \tag{3.25}$$

If we use our Fourier coefficients, we can get,

$$\sum_{\substack{d|(m,n)\\ d>0}} \rho_j\left(\frac{mn}{d^2}\right) = \mu_j(n)\rho_j(m). \tag{3.26}$$

If we take $m = 1$, we get,

$$\mu_j(n) = \frac{\rho_j(n)}{\rho_j(1)}. \tag{3.27}$$

As a result, we get the matrix form of the formula (3.3),

$$\rho_j(n)\rho_j(m) = \rho_j(1) \sum_{d|(m,n)} \rho_j\left(\frac{mn}{d^2}\right). \tag{3.28}$$

Example 3.8.

$$\left(T(n)E\right)(z,s) = \tau_s(n)E(z,s), \tag{3.29}$$

where $E(z,s)$ is Eisenstein series that will be discussed later. Its definition is that,

$$E(z,s) = y^s + \frac{1}{2} \sum_{\substack{(c,d)=1\\ c\neq 0}} \frac{y^s}{|cz+d|^{2s}}. \tag{3.30}$$

And,

$$\tau_s(n) = |n|^{s-1/2} \sum_{\substack{d|n\\ d>0}} d^{1-2s}. \tag{3.31}$$

4 Inner product between real-analytic Poincare series

When $z \in \mathbb{H}$ and $s \in \mathbb{C}$, Poincare defined the series $P_n(z,k) = n^{k-1} \sum\limits_{g \in G_\infty \backslash G} \frac{e^{2\pi i n g z}}{(cz+d)^k}$, where $gz = \frac{az+b}{cz+d}$. Similarly, Selberg defined $U_n(z,s) = \sum\limits_{g \in G_\infty \backslash G} (\mathrm{Im} gz)^s e^{2\pi i n g z}$, which is called real-analytic Poincare series.

This section is the highlight of the entire paper. Due to calculating this inner product from two basis, principal properties of Poincare series are described in two senses, both numerically and analytically. Corollary 5.1 and theorem 5.3 can be understood easily in this way. As for Poincare series, it can be regarded as the expansion of group representation theory, where sums with respect to group elements exist everywhere. Moreover, in the theory of Riemann surfaces, the Riemann θ functions share the sane philosophy.

Remark. When $\mathrm{Re}\, s > 1$, the series absolutely converge. And $U_0(z,s) = E(z,s)$. Besides, we can verify that U_n is automorphic, and if we let $\sigma = \mathrm{Re}\, s$,

$$\mathcal{L}U_n(z,s) = s(1-s)U_n(z,s) + 4\pi n s U_n(z,s+1), \tag{4.1}$$

$$|U_n(z,s)| \leqslant y^\sigma e^{-2\pi n y} + E(z,\sigma) - y^\sigma. \tag{4.2}$$

Theorem 4.1. s_1, $s_2 \in \mathbb{C}$, and $\mathrm{Re}\, s_1$, $\mathrm{Re}\, s_2 > \frac{3}{4}$, $\mathrm{Re}\,(s_1+s_2) < \frac{5}{2}$. Then,

$$\begin{aligned}(U_n(\cdot, s_1), U_m(\cdot, \overline{s_2})) &= \delta_{mn} \frac{\Gamma(s_1+s_2-1)}{(4\pi n)^{s_1+s_2-1}} \\ &+ \left(\sqrt{\frac{n}{m}}\right)^{s_2-s_1} \frac{2^{3-s_1-s_2}}{\sin \pi(s_1-s_2)} \sum_{c=1}^{\infty} \frac{S(n,m;c)}{c^{s_1+s_2}} \Phi(s_1, s_2; \frac{4\pi\sqrt{mn}}{c}),\end{aligned} \tag{4.3}$$

where $S(n,m;c) = \sum\limits_{\substack{1 \leqslant d \leqslant |c| \\ (c,d)=1 \\ dd' \equiv 1 \bmod c}} \exp\left(2\pi i \left(\frac{nd}{c} + \frac{md'}{c}\right)\right)$ is the Klooster's sum and,

$$\Phi(s_1, s_2; x) = \pi \int_1^\infty (u - 1/u)^{s_1+s_2-2} \left(-\sin(\pi s_1) J_{s_1-s_2}(xu) + \sin(\pi s_2) J_{s_2-s_1}(xu)\right) \frac{du}{u}. \tag{4.4}$$

The idea of this proof is in center of Fourier expansion. So, we can first calculate the Fourier coefficients of U_n.

4.1 Fourier expansion of U_n

Lemma 4.2. $\operatorname{Re} s > 1$, $n \in \mathbb{N}$, $z \in \mathbb{H}$, then,

$$U_n(z,s) = \sum_{m=-\infty}^{+\infty} e^{2\pi i m x} B_n(m; y, s), \tag{4.5}$$

where,

$$\begin{aligned} B_n(m; y, s) &= \delta_{mn} y^s e^{-2\pi n y} \\ &+ \frac{1}{2} \sum_{c \neq 0} \frac{S(n,m;c)}{|c|^{2s}} y^{1-s} \int_{-\infty}^{+\infty} \exp\left(-2\pi i m y \xi - \frac{2\pi n}{c^2 y (1 - i\xi)}\right) \frac{d\xi}{(1+\xi^2)^s}. \end{aligned} \tag{4.6}$$

Moreover, when $\operatorname{Re} s > \frac{3}{4}$, the formula is well-defined and holomorphic.

Proof. Dismiss all the strictness, we can get,

$$\begin{aligned} U_n(z,s) &= y^s e^{2\pi i n z} + \frac{1}{2} \sum_{\substack{(c,d)=1 \\ c \neq 0 \\ ad \equiv 1 \bmod c}} \frac{y^s}{|cz+d|^{2s}} \exp\left(\frac{2\pi i n a}{c} - \frac{2\pi i n}{c(cz+d)}\right) \\ &= y^s e^{2\pi i n z} + \frac{1}{2} \sum_{c \neq 0} \frac{y^s}{|c|^{2s}} \sum_{\substack{1 \leqslant d \leqslant |c| \\ (c,d)=1}} e^{2\pi i n \frac{a}{c}} f_n\left(x + \frac{d}{c}; c, y, s\right), \end{aligned} \tag{4.7}$$

where,

$$f_n(x; c, y, s) = \sum_{-\infty}^{+\infty} |iy + m + x|^{-2s} \exp\left(\frac{-2\pi i n}{c^2(iy + m + x)}\right). \tag{4.8}$$

Note that f_n has period one, so,

$$f_n = \sum_{m=-\infty}^{+\infty} e^{2\pi i m x} b_n(m; c.y, s), \tag{4.9}$$

where,

$$\begin{aligned} b_n(m; c, y, s) &= \int_0^1 e^{-2\pi i m \xi} f_n(\xi; c, y, s) d\xi \\ &\xlongequal{\text{Change order}} y^{1-2s} \int_{-\infty}^{+\infty} \exp\left(-2\pi i m y \xi - \frac{2\pi n}{c^2 y (1 - i\xi)}\right) \frac{d\xi}{(1+\xi^2)^s}. \end{aligned} \tag{4.10}$$

We may as well omit $[i, i\infty)$ and $(-i\infty, -i]$ for one-valued branch. And we can change the toy contour to $\operatorname{Im}\xi = \Delta$, $-1 < \Delta < 1$. Then,

$$\operatorname{Re}(-i\xi) = \Delta, \quad \operatorname{Re}\left(\frac{-1}{1-i\xi}\right) < 0, \quad \left|\left(1+\xi^2\right)\right|^{-s} < \left((1-|\Delta|)^2 + (\operatorname{Re}\xi)^2\right)^{-\sigma}. \tag{4.11}$$

Hence, for any s for which $\operatorname{Re} s > 1/2$ and any $\Delta \in (-1, 1)$ we have,

$$|b_n(m; c, y, s)| \leqslant A(y, \sigma) e^{-2\pi|m\Delta|y}, \quad \sigma = \operatorname{Re} s. \tag{4.12}$$

As a result, this series converge absolutely when $y > 0$ and $\sigma > 1/2$ and if the following series converges, we can substitute (4.9) in (4.7).

$$\sum_{c\neq 0} \frac{S(n, m; c)}{|c|^{2\sigma}}. \tag{4.13}$$

It really converges when $\sigma > 3/4$. In fact, Kloosterman sums satisfy Weil's estimate,

$$|S(n, m; c)| \leqslant |c|^{1/2} \min\left\{\sqrt{(n,c)}\, d\left(\frac{c}{(n,c)}\right), \sqrt{(m,c)}\, d\left(\frac{c}{(m,c)}\right)\right\}. \tag{4.14}$$

So, for any fixed n, this series is determined by $\sqrt{n} \sum\limits_{c\neq 0} \frac{d(c)}{|c|^{2\sigma - 1/2}}$, which converges when $\sigma > 3/4$.

Replace f_n by its Fourier expansion, and our target will be reached.

■

4.2 One way of calculating inner product

It is easy to see that, the inner product I above satisfies that,

$$I = \sum_{g \in G_\infty \backslash G} \int_D U_n(z, s_1) (\operatorname{Im} gz)^{s_2} \overline{e^{2\pi i m g z}} dz \tag{4.15}$$

$$\overset{G\text{-invariant}}{=\!=\!=\!=} \sum_g \int_{gD} U_n(z, s_1) y^{s_2} \overline{e^{2\pi i m z}} dz \tag{4.16}$$

$$= \int_B U_n(z, s_1) y^{s_2} \overline{e^{2\pi i m z}} dz, \text{ where B is the strip of } [0,1] \times [0, \infty) \tag{4.17}$$

$$= \delta_{mn} \frac{\Gamma(s_1 + s_2 - 1)}{(4\pi n)^{s_1 + s_2 - 1}} + \frac{1}{2} \int_0^\infty y^{s_2 - 2} e^{-2\pi m y} \sum_{c \neq 0} \frac{S(n, m; c)}{|c|^{2 s_1}} y^{s_1} b_n(m; c, y, s_1). \tag{4.18}$$

The last step uses Fourier expansion of U_n. And if we can change the order, we will get the target form of (4.3). To reach this, we note that,

$$|y^{s_1} b_n(m; c, y, s_1)| \leqslant y^{1-\sigma_1} \int_{-\infty}^{\infty} \frac{d\xi}{(1+\xi^2)^{\sigma_1}}, \quad \sigma_1 = \operatorname{Re} s_1. \tag{4.19}$$

So, when we temporarily assume $\sigma_2 > \sigma_1 > 1$ and $m \geqslant 1$, the integrand in (4.18) is majorized by $y^{\sigma_2-\sigma_1-1} e^{-2\pi m y} \sum_{c\neq 0} \frac{|S(n,m;c)|}{c^{2\sigma_1}}$. As a result, change the order of summation over c and integration over y and use the representation of b_n, we can obtain that the inner integral is equal to,

$$\int_0^\infty y^{s_2-s_1-1} \exp\left(-2\pi m(1+i\xi)y - \frac{2\pi n}{c^2(1-i\xi)y}\right) dy. \tag{4.20}$$

Here, the writer skipped a lot of calculation and claimed that using the well-known integral representation for the Hankel function of the first kind of a purely imaginary argument,

$$K_v(z) = \frac{1}{2}\int_0^\infty \exp\left(-\frac{z}{2}\left(u + \frac{1}{u}\right)\right) \frac{du}{u^{v+1}} \quad \operatorname{Re} z > 0, \tag{4.21}$$

we can get, the integral in (4.20) is equal to,

$$\frac{2}{|c|^{s_2-s_1}} \left(\frac{n}{(1+\xi^2)\, m}\right)^{\frac{s_2-s_1}{2}} K_{s_1-s_2}\left(\frac{4\pi\sqrt{nm}}{|c|}\sqrt{\frac{1+i\xi}{1-i\xi}}\right). \tag{4.22}$$

Substitute it into the above formula, the second term in (4.18) is,

$$2\left(\sqrt{\frac{n}{m}}\right)^{s_2-s_1} \sum_{c=1}^{\infty} \frac{S(n,m;c)}{c^{s_1+s_2}} \int_{-\infty}^{\infty} K_{s_1-s_2}\left(\frac{4\pi\sqrt{nm}}{c}\sqrt{\frac{1+i\xi}{1-i\xi}}\right) \left(1+\xi^2\right)^{-\frac{s_1+s_2}{2}} d\xi. \tag{4.23}$$

And if we let $v = \sqrt{\frac{1+i\xi}{1-i\xi}}$ which changes along right unit semicircle from $-i$ to i, we can obtain,

$$(4.23) = (-i)2^{2-s_1-s_2} \int_{-i}^{i} K_{s_1-s_2}(xv)\left(v + \frac{1}{v}\right)^{s_1+s_2-2} \frac{dv}{v}, \quad x = \frac{4\pi\sqrt{nm}}{c}. \tag{4.24}$$

Cut the complex $v-$plane along the negative real semiaxis, and deform the path of the integral above. We can obtain a path from the imaginary axis from $-i\infty$ to $-i$ and from i to $i\infty$, because for each fixed v and fixed $x > 0$, when $|v| \to \infty$ in the right half-plane, $K_v(xv) \ll |v|^{-1/2}$. So the integral along the bigger semicircle converges to 0 when $\sigma_1 + \sigma_2 \leq 5/2$.

And In the integral from i to $i\infty$, we have,

$$K_v(z) = \frac{\pi}{2\sin \pi v} \left\{ e^{-i\pi v/2} J_{-v}\left(ze^{-i\pi/2}\right) - e^{i\pi v/2} J_v\left(ze^{-i\pi/2}\right) \right\}, \tag{4.25}$$

and in the integral from $-i\infty$ to $-i$, we have,

$$K_v(z) = \frac{\pi}{2\sin \pi v} \left\{ e^{i\pi v/2} J_{-v}\left(ze^{i\pi/2}\right) - e^{-i\pi v/2} J_v\left(ze^{i\pi/2}\right) \right\}. \tag{4.26}$$

Hence, substitute the above expressions and combine same terms, we can get the formula (4.4).

In order to erase the extra assumption of s, it suffices to verifying the series in (4.3) converges absolutely and use the principle of analytic continuation. Since we have estimates,

$$|J_{\pm v}(x)| \ll x^{-|\operatorname{Re} v|}, \tag{4.27}$$

and,

$$\int_1^\infty u^\mu J_v(u) du, \tag{4.28}$$

is finite for any $\operatorname{Re} \mu < 1/2$, we can obtain, when $x \to 0+$

$$|\Phi(s_1, s_2; x)| \ll \begin{cases} x^{2-\sigma_1-\sigma_2}, & \min(\sigma_1, \sigma_2) > 1, \\ x - |\sigma_1 - \sigma_2| \ln \frac{1}{x}, & \min(\sigma_1, \sigma_2) \leqslant 1. \end{cases} \tag{4.29}$$

Thus, the general term in the series in (4.3) can be dominated by $o\left(|c|^{-2}|S(n, m; c)|\right)$ if $\min(\sigma_1, \sigma_2) > 1$, and by $o\left(|c|^{-2\min(\sigma_1,\sigma_2)} \ln|c||S(n, m; c)|\right)$ if $\min(\sigma_1, \sigma_2) \geqslant 1$.

And then, every proposition in this section has been proved.

4.3 The other way

The subsection 3.3 has stepped forward a lot. Here we can use their information to calculate the inner product the second time.

Lemma 4.3. The eigenfunctions above are complete in this Hilbert space.

Theorem 4.4. Let s_1 and s_2 be complex variables. For any fixed value of one of them, the inner product is a meromorphic function of the second variable in the entire plane, and for all s_1 and s_2 with $\operatorname{Re} s_j > 1$ it equals to,

$$I = \pi(4\pi)^{1-s_1-s_2}\left(\sqrt{\frac{n}{m}}\right)^{s_2-s_1}\left(\sum_{j=1}^{\infty}\rho_j(n)\overline{\rho_j(m)}\Lambda(s_1,s_2;\kappa_j)\right.$$
$$\left.+\frac{1}{\pi}\int_{-\infty}^{+\infty}\left(\frac{m}{n}\right)^{ir}\sigma_{2ir}(n)\sigma_{-2ir}(m)\Lambda(s_1,s_2;r)\frac{\cosh\pi r}{|\zeta(1+2ir)|^2}dr\right), \quad (4.30)$$

where $\sigma_s(n) = \sum_{d|n} d^s$ and,

$$\Gamma(s_1,s_2;r) = \frac{\Gamma(s_1-1/2+ir)\Gamma(s_1-1/2-ir)\Gamma(s_2-1/2+ir)\Gamma(s_2-1/2-ir)}{\Gamma(s_1)\Gamma(s_2)}. \quad (4.31)$$

Proof. If we let,

$$\begin{aligned}\mathcal{E}_i(f) &= \textstyle\int_D f(z)\overline{\psi_j(z)}dz\\ \mathcal{E}(r,f) &= \textstyle\int_D f(z)\overline{E(z,1/2+ir)}dz,\end{aligned} \quad (4.32)$$

We have Parseval's equality,

$$\left(U_n\left(\cdot,s_1\right),U_m\left(\cdot,\overline{s_2}\right)\right) = \frac{1}{4\pi}\int_{-\infty}^{\infty}\mathcal{E}\left(r,U_n\left(\cdot,s_1\right)\right)\overline{\mathcal{E}\left(r,U_m\left(\cdot,\overline{s_2}\right)}dr$$
$$+\sum_{j=0}^{\infty}\mathcal{E}_j\left(U_n\left(\cdot,s_1\right)\right)\overline{\mathcal{E}_j\left(U_m\left(\cdot,\overline{s_2}\right)\right)}. \quad (4.33)$$

Similarly, using their Fourier coefficients, we will obtain,

$$(U_n(\cdot,s),\psi_j) = \int_0^{\infty}y^{s-2}\int_0^1 e^{2\pi inz}\overline{\psi_j(z)}dxdy \quad (4.34)$$
$$= (2\pi n)^{1/2-s}\overline{\rho_j(n)}\int_0^{\infty}e^{-y}K_{i\kappa_j}(y)y^{s-3/2}dy. \quad (4.35)$$

And we can get a simple form,

$$(U_n(\cdot,s),\psi_j) = 2\pi\sqrt{n}(4\pi n)^{-s}\overline{\rho_j(n)}\frac{\Gamma\left(s-1/2+i\kappa_j\right)\Gamma(s-1/2-i\kappa_j)}{\Gamma(s)}, \quad (4.36)$$

and,

$$\frac{1}{\pi}\left(U_n(\cdot,s),E(\cdot,1/2+ir)\right)$$
$$= 2^{2-2s}(n\pi)^{1/2-s-ir}\sigma_{2ir}(n)\frac{\Gamma(s-1/2+ir)\Gamma(s-1/2-ir)}{\Gamma(s)\Gamma(1/2-ir)\zeta(1-2ir)} \quad (4.37)$$

Substituting these Fourier coefficients in Parseval's equality, we obtain the main assertion of the lemma. We reserve the others for readers.

■

5 Application

In this section, we will only show the outcome from the analysis above without proofs.

Corollary 5.1. m, $n \in \mathbb{N}_+$, $|\operatorname{Im} t| \leqslant \frac{1}{4}$, then,

$$\sum_{j=1}^{\infty} \frac{\rho_j(n)\overline{\rho_j(m)}}{\cosh \pi\kappa_j} H(\kappa_j, t) + \frac{1}{\pi}\int_{-\infty}^{+\infty} \left(\frac{m}{n}\right)^{ir} \sigma_{2ir}(n)\sigma_{-2ir}(m) \frac{H(r,t)}{|\zeta(1+2ir)|^2} dr$$
$$= \frac{\delta_{mn}}{\pi^2} \frac{t}{\sinh \pi t} + \frac{2t}{\pi \sinh(2\pi t)} \sum_{c=1}^{\infty} \frac{S(n,m;c)}{c} \Phi(\frac{4\pi\sqrt{mn}}{c}, t), \quad (5.1)$$

where

$$H(r,t) = \frac{\cosh \pi r}{\cosh \pi(r+t) \cosh \pi(r-t)}, \quad (5.2)$$
$$\Phi(x,t) = x \int_x^{\infty} \left(J_{2it}(u) + J_{-2it}(u)\right) \frac{du}{u}. \quad (5.3)$$

Proof. Let $s_1 = 1 + it$ and $s_2 = 1 - it$. In this case, compare the two forms of the inner product. Intereted readers can complete the remaining proof. ■

Corollary 5.2. Given $\epsilon > 0$, $X \geqslant 2$ and $n \geqslant 1$,

$$\sum_{\kappa_j \leqslant X} \frac{|\rho_j(n)|^2}{\cosh \pi\kappa_j} = \frac{X^2}{\pi^2} + O\left(X \log(X) + Xn^{\epsilon} + n^{\frac{1}{2}+\epsilon}\right). \quad (5.4)$$

Theorem 5.3. $h(r)$ is an even function of complex variables holomorphic on the strip $\{\operatorname{Im} r \leqslant \Delta\}$ with $\Delta > \frac{1}{2}$ and $h(r) = O\left(|r|^{-2-\delta}\right)$ where $\delta > 0$. m, $n \in \mathbb{N}_+$. Then,

$$\sum_{j=1}^{\infty} \frac{\rho_j(n)\overline{\rho_j(m)}}{\cosh \pi\kappa_j} h(\kappa_j) + \frac{1}{\pi}\int_{-\infty}^{+\infty} \left(\frac{m}{n}\right)^{ir} \sigma_{2ir}(n)\sigma_{-2ir}(m) \frac{h(r)}{|\zeta(1+2ir)|^2} dr$$
$$= \frac{\delta_{mn}}{\pi^2} \int_{-\infty}^{+\infty} r \tanh \pi r h(r) dr + \sum_{c=1}^{\infty} \frac{S(n,m;c)}{c} \phi(\frac{4\pi\sqrt{mn}}{c}), \quad (5.5)$$

where,

$$\phi(x) = \frac{2i}{\pi} \int_{-\infty}^{+\infty} J_{2ir}(x) \frac{r}{\cosh \pi r} h(r) dr. \tag{5.6}$$

Remark. The estimate (5.4) can be obtained when we let $m = n$. Now weight function H plays a role of filtration. Moreover, theorem 5.3 can be obtained if we integrate (5.1) over t and change the order of sum and integration.

Theorem 5.4. When n, m fixed and $T \to \infty$,

$$|\sum_{1 \geqslant c \geqslant T} \frac{S(n, m; c)}{c}| \ll T^{1/6} (\ln T)^{1/6}. \tag{5.7}$$

Remark. This estimate is the first nontrivial conclusion all over the world. Ju. V. Linnik conjectured that the average on the left is much smaller than any T^{ϵ}. And Selberg found a counterexample to show that the analog of Linnik's conjecture for an arbitrary discrete subgroup of $SL(2, \mathbb{R})$ is wrong.

6 Summary

We briefly discuss the basic ideas of this topic and some theorems. In this process, we turn to be familiar with modular forms and contemporary analytic number theory. The deeper our grasp of arithmetic group is, the better we can understood the number theory. I think it is what Kuznecov's article suggests. The introduction is still imperfect, and we apologize for all possible errors and fault sentence in advance.

References

[1] Nikolai Vasil'evitch Kuznetsov. Petersson's conjecture for cusp forms of weight zero and linnik's conjecture. sums of kloosterman sums. *Matematicheskii Sbornik*, 153(3):334–383, 1980.

[2] Elias M Stein and Rami Shakarchi. *Complex analysis*, volume 2. Princeton University Press, 2010.

Physics

A journey to space-time singularity

Yu xueyi

June 2022

1 Introduction

This is a reading report on general relativity and space time singularity. The book or article I read is listed in **Section 6 (Reference)**.

Section 2(Causal structure in space time) will introduce the basic concept and result about the causal structure in general relativity, and introduce the concept of Cauchy surface. **Section 3(The longest causal path, and conjugate point on non-spacelike geodesics)** will investigate when can a non-spacelike geodesic be a longest non-spacelike path.

Section 2,3 can be seen as preparation to Section 4,5. **Section 4 (Hawking's singularity theorem on the cosmology)**, **Section 5(Penrose's singularity theorem on the blackhole)** will introduce two exciting evidence of the existence of singularity, one exists in the beginning of universe, the other exists in the death of a massive star.

Because Hawking's singularity theorem do with time-like geodesics, Penrose's singularity theorem do with null-geodesic. Hawking's singularity theorem is easier to understand for beginner, so I don't follow the route of history and introduce it first.

2 Causal structure in space time

The Minkowski space is the four-dim linear space equipped with the (0,2) metric tensor $\begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ on each point. Call the four axis t, x, y, z; t express time; x,y,z express space. The Minkowski space is the mathematics model of space time in special relativity.

But in general relativity, the things is a little different.The space-time is locally a Minkowski space on every point. That's to say,every point in the space-time manifold $\mathcal{M}$ (a four-dim manifold that is smooth enough to ensure the theorem in this article to be right) is equipped with a (0,2) metric tensor g_{ab}, which is equivalent to the metric tensor in the Minkowski space under coordinate transformation. But this space-time ($\mathcal{M}$,g)can be curve in large scale, the curve of space-time is source of gravitation

Definition 2.1. a four-dim vector x^a is said to be timelike if $g_{ab}x^ax^b < 0$; null if $g_{ab}x^ax^b = 0$; spacelike if $g_{ab}x^ax^b > 0$

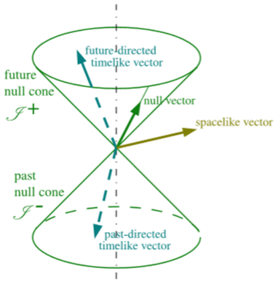


Figure 1: three types of vector

When a particle is moved in space timethe tangent vector of its world line is always timelike (particle with static mass,like electron) or null (particle without static mass,like photon).

Definition 2.2. a C^1 curve $x^\mu(s)$ in space time is said to be non-spacelike path or a casual path if the tangent vector dx^μ/ds is timelike or null on every point.

The causal path $x^\mu(s)$ is a geodesic under proper parameter s if $D^2x^\mu/ds^2 = 0$ along the path .D means covariant derivative. This means that the tangent vector is parallel while moving along the curve. The parameter s is called a affine parameter

Use the metric g, we can define the length of a causal path from p to q is

$$l = \int_{s_p}^{s_q} \sqrt{-g_{ab}\frac{dx^a}{ds}\frac{dx^b}{ds}}\, ds$$

. Attention to the negtive sign. While in Minkowski space , the length is simply

$$l = \int_{s_p}^{s_q} \sqrt{dt^2 - d\vec{x}^2}$$

.

Every substance and information can only travel through non-space like curve.So we have

Definition 2.3. Γ is a subset of $\mathcal{M}$. Then the casual future of Γ, called as $\mathscr{J}^+(\Gamma)$, is the point that can be reached by a future-directed causal path beginning from a point in Γ. The causal past $\mathscr{J}^-(\Gamma)$ can be defined as the same.

. Figure 2 shows the casual future of a point q, Figure 3 shows the casual future of a round ring W in space.

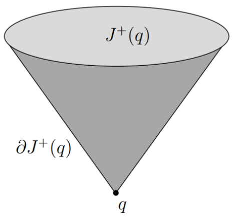


Figure 2: casual future example 1

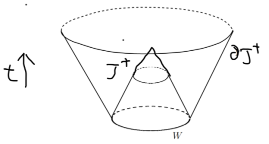


Figure 3: casual future example 2

Definition 2.4. Given two points p,q in $\mathcal{M}$, if p is in the causal future of q, then we can define the causal diamond between p,q , named $\mathcal{D}_q^p$, is the intersect of $\mathcal{J}^+(q)$ and $\mathcal{J}^-(p)$

.

Clearly each causal path from p to q is in $\mathcal{D}_q^p$, most of the time $\mathcal{D}_q^p$ is compact, however there are counter examples. If we moved a point from space time in the causal diamond, then $\mathcal{D}_q^p$ can be non-compact (figure 4)

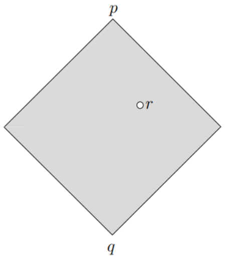


Figure 4: casual diamond $\mathcal{D}_q^p$ is not compact if I moved away a point

If the $\mathcal{D}_q^p$ is compact,then the space of causal path from p to q is compact too when we give $\mathcal{M}$ certain restrictions(Theorem 2.2). This can be seen using the intuition of Arzela-Ascoli theorem, but our space time is not a metric space yet. We can define a metric (Euclid metric) on $\mathcal{M}$ use a local finite atlas and "partition of unity".

Recall that let $\mathcal{M}$ is paracompact, then there exist a local finite atlas $\{U_a, \phi_a\}$, and $\{C^1\}$ functions g_a on each map, such that

(1)$0 \leq g_a \leq 1$;

(2)the support of g_a is contained in U_a ;

(3)$\sum_a g_a(p) = 1$or any p ;

Then the set of g_a is called a partition of unity. g_a is the weights of each map. If there is only one map that cover p. We can define the metric at p simply the metric as in Euclid space. If there are many maps that cover p, we can calculate the weighted average of every metrics use the weights function g_a. Then we can define the length of γ which is a C^1 curve from p to q, then define d(p,q)=inf{length(γ)|γ is a C^1 curve from p to q}. One can proof the topology of $\mathcal{M}$ under the metric d is the same as the natural topology of $\mathcal{M}$ by verifying the set {the open subset of every U_a} is a topological basis of $\mathcal{M}$ in both cases.

Then we can proof the theorem we mentioned before, but before that we should proof a baby version.

Theorem 2.1. Let $(\mathcal{M}, g)$ be the four-dim Minkowsky space. If the $\mathcal{D}_q^p$ is compact,then the space of C^1 causal path from p to q is compact too.

Proof. we can see C^1 causal path $\gamma(s)$ from p to q as a C^1 map from [0,1] to $\mathcal{M}$, such that $\gamma(0) = q, \gamma(1) = p$, we can let the parameter s simply represent the t-axis value of $\gamma(s)$, however we should we should act a linear function on it so that $\gamma(0) = q, \gamma(1) = p$.

$\gamma(s)$ may be not equicontinuous when they are just normal curve.But cause $\gamma(s)$ is causal path, they are equicontinuous: Because $\gamma(s)$ is causal path, then $|dt| > |d\vec{x}|$,(t represent time, $\vec{x}$ represent space) then $ds_E^2 = dt^2 + d\vec{x}^2 < 2 * dt^2$, then $|ds| < \sqrt{2}|dt|$, then $d(\gamma(s_1), \gamma(s_2)) < \sqrt{2}|s_1 - s_2|$.

Recall that γ are all in a compact place $\mathcal{D}_q^p$, use Arzela-Ascoli theorem, function space $\gamma(s)$ is compact, then the curve space γ is compact □

The compactness of the space of causal path means a lot to us, in such space there is a longest causal path from any two compact sets (such as from two points, or from a point to a compact set). But this compactness only holds for some "normal enough" space. There is a restriction that have good physical intuition, that is "globally hyperbolic".

Definition 2.5. A space-time ($\mathcal{M}$,g) is called globally hyperbolic if there is a Cauchy surface $\mathcal{H}$. That is, $\mathcal{H}$ is a spacelike 3-dim submanifold of $\mathcal{M}$, and the point of $\mathcal{M} - \mathcal{H}$ is divided into $\mathcal{J}^+(\Gamma)$ (future) or$\mathcal{J}^-(\Gamma)$ (past). If a point p is in the future of $\mathcal{H}$, then any past-pointed closed casual path without end point will pass though $\mathcal{H}$; if a point q is in the past of $\mathcal{H}$, then any future-pointed closed casual path will pass though $\mathcal{H}$.

The intuition is that if you want to predict what happened in p, you just need to know the data on $\mathcal{H}$, figure 6 shows a counter example of globally hyperbolic space-time. A point r is kicked out from the space time. If you want to predict what will happen at p, you also need the data that comes from the lost point r.

Then we come to the theorem that shows the compactness (theorem2.2, theorem2.3, especially theorem 2.4).

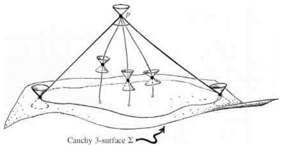


Figure 5: Cauchy surface example

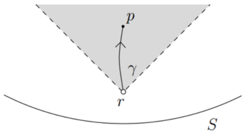


Figure 6: Cauchy surface counter example

Theorem 2.2. if ($\mathcal{M}$,g) is globally hyperbolic, $\mathcal{H}$ is a Cauchy surface. Let q be a point in the past of $\mathcal{H}$, then the future-pointed causal path segment from q that has an end point on $\mathcal{H}$ form a compact space

Proof. Take $\gamma_1, \gamma_2, ..., \gamma_n, ...$ be a infinte sequence of future-pointed C^1 causal path segment from q that has an end point on $\mathcal{H}$.

($\mathcal{M}$,g) is locally a Minkowski space at p.Then there is a open neighborhood of q called U, and a coordinate $(t, \vec{x})$ on U such that the Lorentz metric is $ds^2 = adt^2 + bd\vec{x}^2$, $a \in (-1-\epsilon, -1+\epsilon)$, $b \in (1-\epsilon, 1+\epsilon)$.while ϵ is sufficiently small.

We set the Euclid metric is $ds_E^2 = dt^2 + d\vec{x}^2$, one can verify that U is a metric space under this Euclid metric. And the topology won't change

Use the symbol in the proof of theorem 2.1. Cause γ is timelike, then $ds^2 = a\,dt^2 + b\,d\vec{x}^2 > 0$, then

$$|dt|\frac{1-\epsilon}{1+\epsilon}|d\vec{x}|$$

then

$$ds_E^2 = dt^2 + d\vec{x}^2 < \frac{2}{1-\epsilon}dt^2$$

then

$$|ds| < \sqrt{\frac{2}{1-\epsilon}}|dt|$$

then $d(\gamma(s_1), \gamma(s_2)) < \sqrt{2}|s_1 - s_2|$, then the map $\gamma(s)$ is equicontinuious. Then the image of $\gamma(s)$ is within the closer of U, which is compact. Use Arzela-Ascoli theorem, then γ_n restricted to U will have a subsequence converge to γ_a, which is a causal path segment which has a end point q' on ∂U.Then we can extend the convergence causal path segment from q' use the same method. We can extend it until it pass through $\mathcal{H}$.

To make our proof more rigorous, let's do some subtle work. Each time we extend our convergence causal path segment, we define l_{add} the sup of the Euclid-length of the segment we can add on the original path(we have shown that l can't be zero,because we can always extend).Then each step we will extend until the length add at least $l_{add}/2$, this can be done.

Then our convergence causal path won't have a end point at the past of $\mathcal{H}$. If so, we call the endpoint r. We call the end point of each step of extend r_n, then $r_n \to r$. Use the definition $l_{add}(r_n) \to l_{add}(r)$.

But we have promise that each step we will extend until the length add $l_{add}(r_n)/2$, but the r_n converge, so we have $l_{add}(r_n) \to 0$, then we have $l_{add}(r) = 0$, a contradiction!

Then our convergence causal path won't have a end point at the past of $\mathcal{H}$, we can call it γ. use the global hyperbolic of ($\mathcal{M}$,g), γ will have a end point on $\mathcal{H}$, that's what we want

□

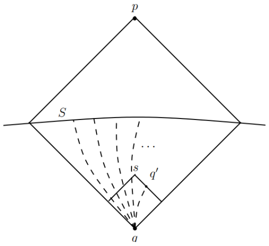


Figure 7: proof of theorem 2.2

Theorem 2.3. The point q is defined in theorem 2.2, then $\mathcal{D}_q^{\mathcal{H}} := \mathcal{J}^+(q) \cap (\mathcal{J}^-(\mathcal{H}) \cup \mathcal{H})$ is compact.

Proof. If $\mathcal{D}_q^{\mathcal{H}}$ is not compact. There will be $q_1, q_2, ..., q_n, ...$ in $\mathcal{D}_p^{\mathcal{H}}$ that don't have a limit point in $\mathcal{D}_q^{\mathcal{H}}$, then the causal path $\gamma_1, \gamma_2, ..., \gamma_n, ...$ (γ_n pass through q and q_n) will have a converge subsequence γ_{nk} that converge to a causal path γ. Then for sufficiently large k, γ_{nk} will be within a neighborhood of γ, which is compact.Then q_{nk} is within the compact neighborhood of γ, but q_{nk} is a subsequence of q_n and won't have a limit point, contradict with the compactness! □

Theorem 2.4. if ($\mathcal{M}$,g) is globally hyperbolic, $\mathcal{H}$ is a Cauchy surface. Let q be a point in the past of a point p, then the C^1 causal path segment from q to p form a compact space.

This theorem is also right if p,q is replaced by compact set. Because the direct product of compact set is compact

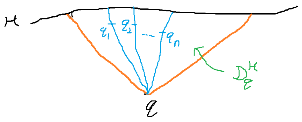


Figure 8: proof of theorem 2.3

Proof. There are two situations: p, q are within two different sides of $\mathcal{H}$, or p, q are within the same side of $\mathcal{H}$

(1)p, q are all in the future of $\mathcal{H}$

$\gamma_1, \gamma_2, ..., \gamma_n, ...$ are causal path from p to q. γ_n can be extend to $\mathcal{H}$ and then call it β_n . From theorem 2.2 we have known that β_n will have limit causal path β. Cause each β_n pass though q, then β will pass though q too. Call the segment on β from q to p γ. Cause β is the limit causal path of β_n, γ is the limit causal path of γ_n.

(2)p is in the future of $\mathcal{H}$, while q is in the past.

$\gamma_1, \gamma_2, ..., \gamma_n, ...$ are causal path from p to q. γ_n can be divided into γ_n^1 and γ_n^1. γ_n^1 are causal path segment from q to $\mathcal{H}$; γ_n^2 are causal path segment from $\mathcal{H}$ to p. From theorem 2.2 we have known that γ_n^1 will have a subsequence $\gamma_{n_i}^1$ converge to γ^1. We also know that $\gamma_{n_i}^2$ will have a subsequence $\gamma_{n_{ij}}^2$ converge to γ^2. Put together γ^1,γ^2 and then form a causal path γ from p to q. Cause $\gamma_{n_{ij}}^1$ converge to γ^1, $\gamma_{n_{ij}}^2$ converge to γ^2, then $\gamma_{n_{ij}}$ converge to γ, then γ is the limit causal path of γ_n. □

From now on we only consider our space-time ($\mathcal{M}$, g) is globally hyperbolic. This have good reason because we always expect the future of our universe is determined by some initial state.

3 The longest causal path, and conjugate point on non-spacelike geodesics

From theorem 2.4 we know in a globally hyperbolic space-time there is a longest causal path from any two compact sets. If the two compact sets is two points p,q (p is in the future of q) we call itthe longest causal path from p to q. Use variation of curve it can be showed that the longest causal path must be a timelike or null geodesics (theorem 3.1).

In the opposite, a timelike or null geodesics may not be the longest causal path from p to q. As an example. Let's see the longitude begin at the north pole N (figure.9), as the path have pass though the south pole S, it's still a geodesic, but no longer a shortest path. We notice that S is a conjugate point of S, that is, some geodesics emitted from N will focus at S again. We will show that a geodesics is a shortest path (in Lorentz metric, the longest causal path) only if it's geodesics and have no conjugate point on it (Theorem 3.1).

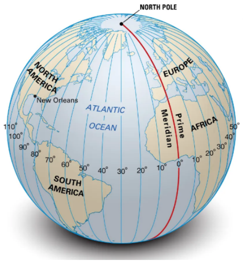


Figure 9: Longitude that have pass though south pole is no longer the shortest path

To define conjugate point we should define variation of curve first.

Definition 3.1. A variation α of a C^1 causal path γ(from q to p) is a C^1 map from $[-\epsilon, +\epsilon] \times [0, s_p]$ (set $s_q = 0$) to $\mathcal{M}$ such that

(1)$\alpha(0, s) = \gamma(s)$

(2)$\alpha(u, s) := \gamma_u(s)$ is also a C^1 causal path from q to p.

$\partial\alpha(u, s)/\partial(u)|_{u=0} := Z(s)$ is called the variation vector. From the definition we have $Z(0) = 0, Z(s_p) = 0$.

Then we define conjugate point

Definition 3.2. (p is conjugate point of point q) p is in the causal future of q, γ is the non-spacelike geodesic from p to q. We call p is a conjugate point of q, if there is a variation of γ, such that

(1)$\gamma_u(0) = q$

(2)$\gamma_u(s_p) = p$

(3)$\gamma_u(s)$ solve the geodesic equation $d^2\gamma(s)/ds^2 = 0$ in first order. That is $d^2\gamma_\epsilon(s)/ds^2 = o(\epsilon)$.

To ensure that γ_u is significantly different from γ. We have to set another restriction. Named the begin vector of $\gamma(s)$ is $\vec{a}(u) = d\,\gamma_u(0)/ds$. Named the begin vector of $\gamma(s)$ is $\vec{a}(u) = d\,\gamma_u(0)/ds$. Named the end vector of $\gamma(s)$ is $\vec{b}(u) = d\,\gamma_u(s_p)/ds$. The restriction is that

(4)

$$\frac{d\,\vec{a}(u)}{du} \neq 0$$

That is, the beginning vector change direction in first order. This is equivalent to

$$\frac{d\,\vec{b}(u)}{du} \neq 0$$

because if

$$\frac{d\,\vec{b}(u)}{du} = 0$$

recall

$$\frac{d^2\gamma_u(s)}{ds^2} = o(u)$$

, then $\vec{a}(u) = o(u)$, then

$$\frac{d\,\vec{a}(u)}{du} = 0$$

a contradiction.

Figure 10: q' is a conjugate point of q on geodesic pq

Then we come to a Necessity about when can a causal path be a longest causal path.

Theorem 3.1. p is in the causal future of p Then a C^1 causal path γ from p to q is a longest causal path only if γ is a non-spacelike geodecsic, and there are no conjugate point of q on γ.

Proof. In variation of curve , named $L(u) := Length(\gamma_u)$ Some calculation (reference(2) Lemma.4.5.4) shows that

$$\frac{\partial L}{\partial u}|_{u=0} = \int_0^{s_p} g(Z(s), \frac{d^2\gamma(s)}{ds^2})ds \tag{1}$$

g(,) is the inner product of two vector use the Lorentz metric g. Z is the variation vector we mentioned above. If γ is not a geodesic . That is, If $d^2\gamma(s)/ds^2)$ isn't zero along γ, we can set Z(s) such that $g(Z(s), d^2\gamma(s)/ds^2) > 0$ when $d^2\gamma(s)/ds^2$ is non-zero. So that $\partial L/\partial u > 0$, then there is some γ_u near γ that is longer than γ. So to be a shortest path, γ must be a geodesic.

If there are conjugate point q' of q on pq, then there is a variation γ_u of qq' such that γ_u solve the geodesic equation in first order, that is $d^2\gamma_\epsilon(s)/ds^2 = o(\epsilon)$. The variaton vector $Z_u(s)$ is continuous in u,z , so it is bounded in a neighborhood of γ . Then formula (1) shows that

$$\frac{\partial L}{\partial u}|_{u=\epsilon} = o(\epsilon)$$

That is, the length change of γ_u don't change in second order.

Recall the restriction that $d\vec{b}(u)/du \neq 0$, $\vec{b}$ is the end vector of qq'. Then γ_u (the variation of qq') together with q'p form a "kink". We will show that by rounding off the kink we can increase the length in second order, then the new curve will be longer than γ.

Named the combination of γ_u and q'p is η_u. We have $\eta_0 = \gamma$. η_u is a curve from q to p whose tangent vector is continuous expect at q'. Named the end vector of γ_u at q' is $\vec{b}_u$. Named the begin vector of $q'p$ at q' is $\vec{a}_u = \vec{b}_0$ (because γ is C^1). A variation of η_u is η_{uw}, the length of η_{uw} is $L_u(w)$

Similar to (1) calculation (reference(2) Lemma.4.5.4) shows that

$$\begin{aligned}\frac{\partial L_u(w)}{\partial w} &= \int_0^{s_{q'}} g(Z_{uw}(s)\frac{d^2\eta_{uw}(s)}{ds^2})ds + \int_{s_{q'}}^{s_p} g(Z_{uw}(s)\frac{d^2\eta_{uw}(s)}{ds^2})ds \\ &\quad + g(Z_{uw}(s_{q'}), b_{uw}) - g(Z_{uw}(s_{q'}), a_{uw})\end{aligned} \tag{2}$$

There is a technical fact that in a neighbourhood of q', For any two vector $\vec{a}$, $\vec{b}$, there is vector $\vec{z}$ such that $g(\vec{z},\vec{b}) - g(\vec{z},\vec{b}) > K||\vec{a}-\vec{b}||\cdot||\vec{z}||$, while $||\vec{z}||$ is the Euclid length of $\vec{z}$. So we can set $Z_{uw}(s'_q)$ is a vector like this, and set $||Z_{uw}(s'_q)|| = 1$. We can also set Z_{uw} on other point such that $d^2Z_{uw}(s)/d^2s$ is zero in $(0, s'_q), (s'_q, s_p)$, and $Z_{uw}(0) = Z_{uw}(s_p) = 0$. This set ensures $d^2\eta_{uw}(s)/ds^2$ is still zeros in $(0, s'_q), (s'_q, s_p)$, which makes things easier.

Notice that from definition of variation vector , we have

$$\frac{\partial(\vec{a}_{uw} - \vec{b}_{uw})}{\partial w} = \frac{\partial^- \vec{Z}_{uw}}{\partial s} + \frac{\partial^+ \vec{Z}_{uw}}{\partial s}$$

The set of Z ensures both one of the right side $\frac{\partial^- \vec{Z}_{uw}}{\partial s}$, $\frac{\partial^+ \vec{Z}_{uw}}{\partial s}$ have Euclid length greater than $C||\vec{Z}_{uw}|| = C$, C is a constant. Then we have

$$||\frac{\partial(\vec{a}_{uw} - \vec{b}_{uw})}{\partial w}|| > 2C$$

So $||\vec{a}_{uw} - \vec{b}_{uw}||$ doesn't reduce to less than $\frac{1}{2}||\vec{a}_u - \vec{b}_u||$ when $w \in [0, ||\vec{a}_u - \vec{b}_u||/2c]$. Then from formula (2), when $w = ||\vec{a}_u - \vec{b}_u||/2c$,we have

$$\begin{aligned}L_u(w) - L_u(0) &= \int_0^w g(Z_{uw}(s'_q), b_{uw}) - g(Z_{uw}(s'_q), a_{uw})dw \\ &> wK||\vec{Z}_{uw}||\frac{||\vec{a}_u - \vec{b}_u||}{2} = wK\frac{||\vec{a}_u - \vec{b}_u||}{2} \\ &= K\frac{||\vec{a}_u - \vec{b}_u||^2}{2c},\end{aligned} \tag{3}$$

Recall $d\vec{b}(u)/du \neq 0$, then $||\vec{a}_u - \vec{b}_u|| = ||\vec{b}_0 - \vec{b}_u|| = ku(1 + o(1))$ so $L_u(w) - L_u(0) = \frac{Kk}{2c}u^2(1 + o(1))(u \to 0)$. This means that η_{uw} do increase the length in second order. So for some sufficiently small u,w; η_{uw} is longer than γ. Then γ is not the longest when there are conjugate point.

□

The following situation is also useful. If S is a spacelike 3-dim surface. q is in the causal past of S. What kind of causal path γ can be the longest from p to S. When can guess that γ should be a geodesic and have no conjugate point on it. There are another necessity that γ must be orthogonal to S, that is the tangent vector of γ at p (p is on S) is orthogonal to the tangent hyperplane of S at p.

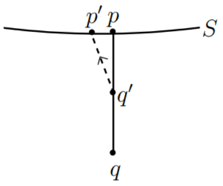


Figure 11: pq has to have no conjugate point q' on S

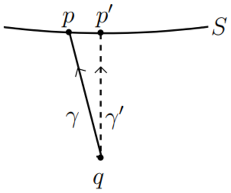


Figure 12: pq has to be orthogonal to S

The definition of conjugate point is much like the situation of two point, but a little different

Definition 3.3. (q is a conjugate point of surface S) S is a spacelike 3-dim surface. q is in the causal past of S. γ is the non-spacelike geodesic from q to S and orthogonal to S at p, p is on γ. We call q is a conjugate point of S at p, if there is a variation of γ, such that

(1)$\gamma_u(0) = q$

(2)$\gamma_u(s_p)$ is still on S

(3)γ_u is still orthogonal to S

(4)And $\gamma_u(s)$ solve the geodesic equation $d^2\gamma(s)/ds^2 = 0$ in first order. That is $d^2\gamma_\epsilon(s)/ds^2 = o(\epsilon)$.

To ensure that γ_u is significantly different from γ. We have to set another restriction. Named the begin vector of $\gamma(s)$ at q is $\vec{a}(u) = d\gamma_u(0)/ds$. The restriction is that

(5)

$$\frac{d\,\gamma_u(s_p)}{du} \neq 0$$

That is, the end point of γ at S change position in first order. This is equivalent to

$$\frac{d\,\vec{a}(u)}{du} \neq 0$$

,

because if

$$\frac{d\,\vec{a}(u)}{du} = 0$$

recall

$$\frac{d^2\gamma_u(s)}{ds^2} = o(u)$$

then

$$\frac{d\,\gamma_u(s_p)}{du} = 0$$

a contradiction.

The necessity of longest causal path in this situation is what we have mentioned above

Theorem 3.2. q is in the causal past of S Then a C^1 causal path γ from q to p(on S) is a longest causal path from q to S, only if

(1)γ is a non-spacelike geodecsic.

(2)There are no conjugate point of q on γ (see figure 11)

(3)And γ is orthogonal to S. (see figure 12)

Proof. From theorem 3.1 , if (1) is not right, Then there is a longer causal path from q to p.

Let's verify (2). If q' is a conjugate point of S at p on qp. Then we can variate q'p to q'p'(p' is still on S) who solve the geodesic equation in first order. Then the length of q'p' don't change in first order. And q'p' with qq' form a "kink" , we can round off the kink as in the proof of thm 3.1. Then the new curve will be longer in second order. Then there will be a longer causal path from q to S

Let's verify (3). γ_u is a variation of γ From calculation(Reference 2, Lemma 4.5.5) similar to formula (1)(2), we get

$$\frac{\partial L}{\partial u}|_{u=0} = \int_0^{s_p} g(Z(s), \frac{d^2\gamma(s)}{ds^2})ds + g(Z(s), \frac{d\gamma(s)}{ds})|s = 0 \tag{4}$$

So we can choose Z such that $Z(0)$ satisfy $g(Z(s), \frac{d\gamma(s)}{ds})|s = 0 > 0$. And Z(s) decay quickly so that $\int_0^{s_p} g(Z(s), \frac{d^2\gamma(s)}{ds^2})ds$ is sufficiently small. So $\frac{\partial L}{\partial u}|_{u=0} > 0$. From variation we can got a longer curve.

□

Theorem 3.2 will be useful in the proof of singularity theorem.

4 Hawking's singularity theorem on the cosmology

Dose the universe have a beginning, or it has existed for infinite time and will still exist? These question is beyond the reach of science until Einstein write his field equation

$$R_{ij} - \frac{1}{2} R g_{ij} = 8\pi T_{ij}$$

in 1915, which can also be written as

$$R_{ij} = 8\pi (T_{ij} - \frac{1}{2} T g_{ij}) = 8\pi \hat{T}_{ij}$$

.

The left hand side is the curvature in space-time, and the left hand side is the matter. It shows how matter curve the space-time. Shortly after this , Alexander Friedmann use the equation to investigate the dynamics of the universe. His work based on the assumption that the universe is uniform isotropic, this is luckly well satisfied by our observable universe according to modern observations. Friedmann's work shows that the universe have a beginning, and may have an end. But whether if uniform isotropic condition isn't satisfied. Some work shows that the deviate of isotropic may prevent singularity from happen.

In this section we will introduce Hawking's singularity theorem on the cosmology. Hawking proofs it in 1970, encouraged by the work of Penrose. This theorem shows that our universe must have a dramatic beginning in some sense. My introduction don't follow the history route. But since the Hawking's singularity theorem only do with time-like geodesic. It's easier to understand.

The proof do a lot with conjugate point. The conjugate point is easy to form in universe. Because the positive energy will cause positive curvature, and positive curvature will cause geodesic come together.

Let's show a convenient way to show when a conjugate point can from

S is a space-like three-dim surface. Let γ be a time-like geodesic orthogonal to S and intersect S at p. Then there is a transfer ϕ_s of a neighborhood U of p on S. That is , from every point q in U, there is a geodesic γ_q begin at q, orthogonal to S and point to the same direction of γ. Define $\phi_s(q) = \gamma_q(s)$. Clearly $\phi_0 = id$. ϕ_s can be seen as a map from a three-dim manifold to another three-dim manifold, $\phi'_s(p)$ is a 3*3 matrix. So We can define

$$A(s) = \sqrt{det(\phi'_s(p)^T \, g_{ij} \, \phi'_s(p))}$$

.

$A(s)$ can be seen as the volume change of U while transfer along γ, clearly $A(0) = 1$.

Theorem 4.1. At some point p' on γ , if $A(s_{p'}) = 0$, then p' is a conjugate point of γ.

Proof. $A(s_{p'}) = 0$ means that for some $\vec{x}$ in linear space, $\phi'_s(p)\vec{x} = 0$. Then choose Q_n in U such that $Q_1, Q_2...$ is on a curve in S that pass P and its tangent vector corresponds to $\vec{x}$ in the tangent space of S at p. Then we have $||\phi_s(Q_n), \phi_s(P)|| = o(||Q_n, P||)$, $||, ||$ is the Euclid length. Use the definition of ϕ_s, then $||\gamma_{Q_n}(s_{p'}) - \gamma_{Q_n}(s_{p'})|| = o(||Q_n, P||)$.We can change γ_{Q_n} a little bit to γ'_{Q_n}, the begin point of γ'_{Q_n} is still Q_n, γ'_{Q_n} is still orthogonal to S. But the end point is moved to $\gamma(s_{p'})$. Cause the distance from $\gamma_{Q_n}(s_{p'})$ to $\gamma_{Q_n}(s_{p'})$ is less then first order. We can still let γ'_{Q_n} solve the geodesic equation in first order. Then from definition P' is conjugate point of S at P. □

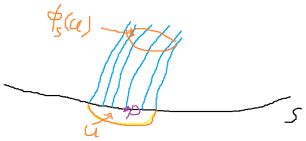


Figure 13: the transfer ϕ_s acts on a neighbor U of p on S

The Raychaudhuri equation (1955) of time-like geodesic shows how $A(s)$ change under the influence of matter. It is just a deformation the 00 component of Einstein equation

$$R_{00} = 8\pi(T_{00} - \frac{1}{2}Tg_{00})$$

to a ordinary differential equation of $A(s)$. The Einstein equation should be used in a proper coordinates.

The coordinate is that: First cause S is a three-dim space like surface, we can set a three-dim vector $\vec{x}$ to every point q on U as a C^1 coordinate, in fact, the coordinate of q is $(0, \vec{x})$. Second, in a neighbor of γ, we set the coordinate of $\gamma_q(s)$ is $(s, \vec{x})$. Since for any $s < s_{p'}$ (p' is the conjugate of S at P). $A(s_{p'}) > 0$, that is $\phi'_s(p)$ is non-singular. So use the inverse function theorem $\phi_s(p)$ is diffeomorphism (so injective) for sufficiently small U as a neighbor of p. Since [0,s] is compact , the Euclid radius of possible U of every s_1 on [0,s] has a positive lower bound. So on [0,s], the coordinate is well defined C^1 coordinate for sufficiently small U. This means that if we got a ODE of A(s), it is hold on any [0,s] for $s < s_p$, so it is hold on $[0, s_p)$.

There is a good property of our coordinate, which will make the calculation much easier.

Theorem 4.2. In this coordinate , the metric tensor on g_{ij} satisfy $g_{i0} = g_{0i} = 0$ (i=1,2,3). This means that $\phi_s(U)$ is still orthogonal to γ while transferring.

This theorem holds because s is a parameter represent the length on each curve. A rigorous proof is this

Proof. Cause γ is orthogonal to S, then $g_{i0}(P) = g_{0i}(P) = 0$ (i=1,2,3) , that is $g_{i0}(\gamma(0)) = 0$. Also

$$\frac{dg_{i0}(\gamma(s))}{ds}|_{s=0} = g(\frac{d\gamma_q(s)}{dx^i}, \frac{d\gamma(s)}{ds})$$

g(,) is the Lorentz inner product, cause

$$\frac{d\gamma_q(s)}{dx^i} = 0$$

, so

$$\frac{dg_{i0}(\gamma(s))}{ds}|_{s=0}$$

is zero. Cause

$$\frac{d^2g_{i0}(\gamma(s))}{d^2s} = g(\frac{d(D^2\gamma_q(s)/d^2s)}{dx^i}, \frac{d\gamma(s)}{ds})$$

.

Cause $\gamma_q(s)$ is geodesic , then $D^2\gamma_q(s)/d^2s = 0$, then

$$\frac{d^2g_{i0}(\gamma(s))}{d^2s} = 0$$

Recall the initial condition

$$g_{i0}(\gamma(0)) = 0$$

,

$$\frac{dg_{i0}(\gamma(s))}{ds}|_{s=0} = 0$$

then we have $g_{i0}(\gamma(s)) = 0$. □

So under the coordinate $(t, \vec{x})$ we defined above, the metric is simply

$$ds^2 = -dt^2 + g_{ij}x_ix_j(i, j = 1, 2, 3) \tag{5}$$

Under this coordinate we have

$$A(s) = \sqrt{det(g_{ij}(s))} \tag{6}$$

Cause the parameter has a intuitive meaning, the time, so we use t instead of s from now on.

From the definition of Ricci curvature we have

$$R_{00} = -\frac{1}{2}\partial_t(g^{ik}\partial_t g_{ik}) - \frac{1}{4}(g^{ik}\partial_t g_{kj})(g^{jm}\partial_t g_{mi}) \tag{7}$$

Then we have

$$R_{00} = -\frac{1}{2}\partial_t Tr(g^{-1}\dot{g}) - \frac{1}{4}Tr(g^{-1}\dot{g})^2 \tag{8}$$

From the equation of A(t) (6) we have

$$\theta = \frac{\dot{A}}{A} = \frac{1}{2}Tr(g^{-1}\dot{g}) \tag{9}$$

θ is called the volume expansion, it describe the volume change of the frontier of a small bunch of geodesics around γ , the "frontier of geodesic bunch" is $\phi_t(U)$ in fact.

Define

$$\sigma^i_j = g^{ik}\dot{g_{kj}} - \frac{1}{3}\delta^i_j Tr(g^{-1}\dot{g}) \tag{10}$$

which describe the shape change of the frontier of a small bunch of geodesics around γ, that is $\phi_t(U)$, is named as "shear".

Then R_{00} can be reduce to

$$R_{00} = -\partial_t\theta - \frac{1}{3}\theta^2 - \frac{1}{4}Tr\,\sigma^2 \tag{11}$$

Recall the Einstein equation

$$R_{00} = 8\pi(T_{00} - \frac{1}{2}Tg_{00}) \tag{12}$$

Define

$$\hat{T}_{00} = T_{00} - \frac{1}{2}Tg_{00} \tag{13}$$

is a kind of matter tensor.

Then (12) is equivalent to

$$R_{00} = 8\pi\hat{T}_{00} \tag{14}$$

Then from (11), (14) we get the the Raychaudhuri equation (1955)

$$\frac{d\theta}{dt} + \frac{1}{3}\theta^2 = -\frac{1}{4}Tr\,\sigma^2 - 8\pi\hat{T}_{00} \tag{15}$$

Recall $\theta = \dot{A}/A$, so (15) is a ODE of A(t)

If $\hat{T}_{00} \geq 0$, which is satisfied by normal matter. Then the right hand side of (15) is non-positive, effect is that $A(t)$ tend to decrease, that means time-like geodesic tends to come together. So we can see in fact the conjugate point is easy to form (See figure.14) .

Theorem 4.3. Assume $\hat{T}_{00} \geq 0$. If $\theta(0) = -\lambda$ is negative, then there is a conjugate point of S on $\gamma(t)$ within $t \in [0, 3/\lambda]$.

Proof. From $\hat{T}_{00} \geq 0$, then the right hand side of (15) is negative.

then we have the left hand side of (15) is

$$\frac{d\theta}{dt} > -\frac{1}{3}\theta^2 \tag{16}$$

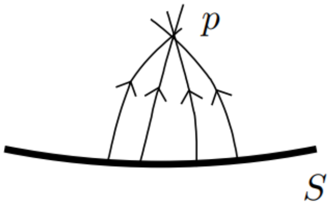


Figure 14: conjugate point p of S is easy to form

So if $\theta(0) = -\lambda$ is negative, Then from the comparision theorem in ODE, we have

$$\theta(t) < \frac{3}{x - 3/\lambda}$$

Then $\theta(t)$ will go to $-\infty$ within $t \in [0, 3/\lambda]$. This meas that A(t) will go to zero within $[0, 3/\lambda]$, then there is a conjugate point of S on $\gamma(t)$ within $[0, 3/\lambda]$. □

$\hat{T}_{00} \geq 0$ is a restriction on matter , which is called the strong energy condition. It is satisfied by non-relativistic matter, radiation . But not satisfied by Dark energy, which is a discovery of modern observation (1998).

Hawking singularity Theorem says that

Theorem 4.4. (Hawking 1970) If the condition below is satisfied

(1)The strong energy condition $\hat{T}_{00} \geq 0$ is satisfied.

(2) If the universe is globally hyperbolic , that is , there is a Cauchy surface S

(3) On S the Hubble constant is everywhere positive. That is , on every point p, the initial expansion $\theta_p(0) = \dot{A}(0)/A(0)$ defined above is positive. And the expansion $\theta_p(0)$ on S has a positive lower bound θ_{min}

Then our universe is time-like geodesic incomplete, that is , there is some time-like geodesic before S that can't be extend to any time parameter to the past. So, the particle traveling though this geodesic may mysteriously have a "beginning" in her time.

(1) is satisfied by normal matter, but sadly, not the dark energy, which dominate in our universe since 9.8 billion years ago (our universe aged 13.7 billion years). (2) is the assumption we always tend to believe. (3) means our universe is expanding, which may be supported by observation.

Proof. If a point q is in the past of S. From theorem 2.3 we know the point on S that can be reached by q by a causal path is compact, From theorem 2.4 we know there is a longest causal path γ from q to S, intersecting with S on p. Then from theorem 3.2 we known that γ must be time-like geodesic orthogonal to S, and there is no conjugate point of S on it.

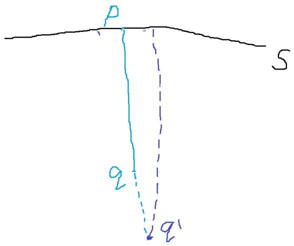


Figure 15: every particle have a finite history...

Notice $\theta_p(0) \geq \theta_{min}$, cause in this case our geodesic is pointing to the past. We have to change the sign of $\theta_p(0)$, then we get $\theta_p(0) \leq -\theta_{min}$ From the strong energy condition $\hat{T}_{00} \geq 0$ (we have discussed the affect of it) , then γ must have a conjugate point q' of S on it when the time parameter is smaller than $\frac{3}{-\theta_p(0)}$, it is equivalent to

$$lenth(pq') \leq \frac{3}{-\theta_p(0)} \leq \frac{3}{\theta_{min}}$$

.

Cause pq have to be longest path, so q' is not on pq', then

$$lenth(pq) \leq lenth(pq') \leq \frac{3}{-\theta_p(0)} \leq \frac{3}{\theta_{min}}$$

.

That means that every point q in the past of S will have a longest-causal path to S that are shorter than $3/\theta_{min}$, that means every causal-path point to the past are shorter than $3/\theta_{min}$ (see figure.15), then our universe is timelike geodesic incompelete

□

The result means that the history of any particle before S is shorter than $3/\theta_{min}$. Where do it begin? We call this place singularity.

In this theorem, there may be many singularity, but from astronomy observation , mysteriously, our universe only have one, we call it the Big-Bang singularity.

5 Penrose singularity theorem on black hole

Stars resist the gravitation by heat pressure of the matter in star. And the heat energy comes from nuclear reaction in the heart of the star. At the end of the life , the star will consume out of its nuclear fuel and lose heat gradually. Then the heat pressure can't resistance the gravitation and then the star will collapse. What will a dead star result to be, it may result as a white dwarf or neutron star . But Chandrasekhar shows that white dwarf have a mass greater than

$1.4M_{sun}$ can't resistance its gravitation by Electron degeneracy pressure and exists stably, Oppenheimer shows that neutron star can't have a mass greater than $3M_{sun}$ can't resistance its gravitation by neutron degeneracy pressure and exists stably. What will a star have a mass greater than $3M_{sun}$ become when it dead, this is a mystery. Oppenheimer and Snyder (1939) shows that if ignore the pressure of matter, and star collapse in a spherical symmetry way. It will collapse to be smaller than its Schwarzschild radius $r_s = 2m$, and will keep collapsing to a singularity. Then all the mass is located at a single point, which seems to be crazy. But the singularity may be seen as a result of perfect spherical symmetry. It's not clear whether singularity will form in real universe.

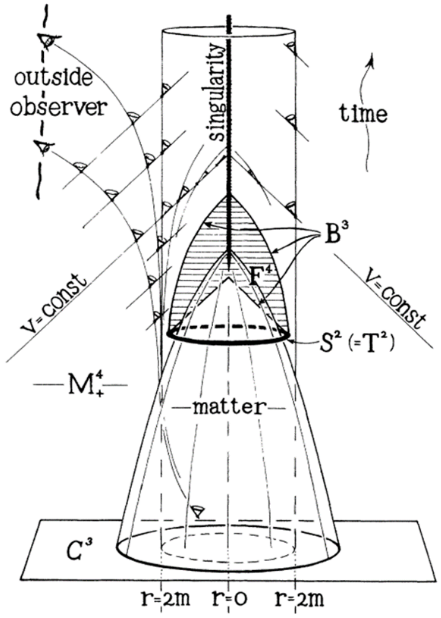


Figure 16: The figure in Penrose's original paper. A star Collapses in a spherical symmetry way will result in a singularity. S^2 is a trapped surface, F^4, B^3 is its future and future boundary.

In 1965 Penrose published his paper "Gravitational collapse and space-time singularities". It shows that the singularity is inevitable if some condition is satisfied

Cause Penrose work do a lot with trapped surface Γ , a compact 2-dim space-like C^1 surface that satisfy some condition , and its causal future $\mathcal{J}^+(\Gamma)$. Let's investigate some property of $\mathcal{J}^+(\Gamma)$, if Γ is a compact 2-dim space-like C^1 surface.

Theorem 5.1. Space-time $(\mathcal{M}, g)$ is globally-hyperbolic. Γ is a compact 2-dim space-like C^1 surface. Then $\partial\mathcal{J}^+(\Gamma)$, the boundary of $\mathcal{J}^+(\Gamma)$, is a consist of the null-geodesic that comes from Γ and orthogonal to Γ

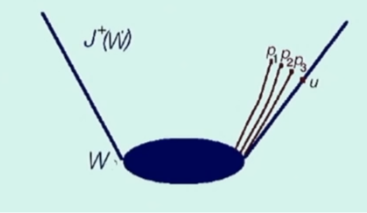


Figure 17: The future boundary of W is made bynull-geodesic that comes from W and orthogonal to W

Proof. If a point u on $\partial\mathcal{J}^+(\Gamma)$ and can be connected to Γ by a causal path γ. Cause Γ is compact , we can choose γ be the longest. Cause u is on the boundary of future, so $length(\gamma) = 0$. If not, $length(\gamma)0$, then in a neighbor U of u, from continuity , every point in U can be connected to Γ by a causal path γ', contradict with the definition of boundary.

So, cause γ is the longest, similarly to theorem 3.2, we have γ is a geodesic orthogonal to Γ. Cause $lenth(\gamma) = 0$, then we have γ is a null-geodesic.

It remain to proof that there do has a causal path γ that connect u with γ . (See fig.17 , in figure.17 W means Γ) . This may not be right if $(\mathcal{M}, g)$ is not globally hyperbolic. Imagine kick off a "ring" R on $\partial\mathcal{J}^+(\Gamma)$ from $(\mathcal{M}, g)$, then the point on $\partial\mathcal{J}^+(\Gamma)$ and behind R it is still on $\partial\mathcal{J}^+(\Gamma)$, but can't connect to Γ by a causal path, because they causal path have to pass the "ring" R.

Now we proof the theorem is right if $(\mathcal{M}, g)$ is globally hyperbolic. Let u be a point on $\mathcal{J}^+(\Gamma)$. Cause u is on the boundary , there will be a sequence of point $p_1, p_2, ...$ in $\mathcal{J}^+(\Gamma)$ such that $p_n \to u$. From the definition of $\mathcal{J}^+(\Gamma)$, there will be a causal path from p_n to γ, name it γ_n. $p_1, p_2, ...$ converge, so they are in a compact set, from globally hyperbolic and theorem 2.4, they form a compact space. So they have a converge sequence γ_{n_k}. It converge to a path γ, Because γ_{n_k} is causal path, , then the tangent vector on every point of γ_{n_k} in non-spacelike, then because γ_{n_k} converge to γ, then the tangent vector on every point of γ in non-spacelike, then γ is a causal path from u to Γ □

Now we can define the trapped surface.

Definition 5.1. (Not rigorous)

Trapped surface:a compact 2-dim space-like C^1 surface Γ that satisfy both of the the two local null geodesic beams(that is, light beams) orthogonal to γ (going outside and inside) at p will decay the area of they wavefront when they leave p.

In figure.16, $S^2 (= T^2)$ is an example of trapped surface. (Figure.19 shows it more clearly). In the case of fig.16, calculation shows that, the gravitation is strong enough that the in-going and out-going light-rays all "going inside"! So the area of the two wave-front all decay. That is a motivational example of trapped surface.

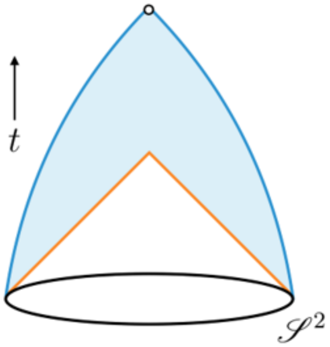


Figure 18: a picture of trapped surface γ, and its causal future. Notice both blue ring and orange ring go smaller when time increase

From continuity, the trapped surface is still a trapped surface if the spherical symmetry is perturbed. Trapped surface can be seen as a result of high matter density and strong gravitation. And Penrose work shows that once the trapped surface is formed, singularity is inevitable.

Let's make the definition of trapped surface more rigorous.

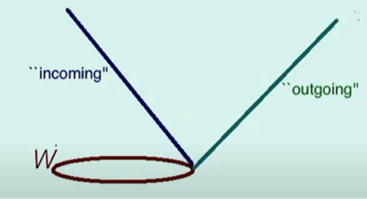


Figure 19: Two light rays orthogonal to W at the same point, one can be seen as "income" direction, and one can be seen as "outcome" direction

First, let's define the two orthogonal direction of light rays. given a point p on Γ, Γ has tangent space W at p, space-time ($\mathcal{M}$,g) also have a tangent space V at p. W is a 2-dim spacelike subspace of 4-dim V. It can be verified that there is only two null-vector $\vec{a}, \vec{b}$ (up to multiplied by a real number) in V that are orthogonal to W. Then set $\vec{a}, \vec{b}$ to be the beginning vector of the two light rays γ_p, η_p . One can be seen as "incoming" , and one as "outgoing" . (see figure 19)

Secondly we can choose a neighbor U of p on Γ. Each point q in U can have a coordinate (x_1, x_2). On every point q of U we can similarly define γ_q, η_q, choose the direction of γ_q, η_q such that γ_q, η_q is continuous in q. $\gamma_q(s)$ has a affine parameter s("affine" means the tangent vector in parallel along the curve , that is $D^2\gamma_q(s)/d^2s$). Set $\gamma_q(0) = q$, then the parameter s of each $\gamma_q(s)$ is determined up to multiply by a constant(the constant of each geodesic γ_q can be distinct from each other).

The same as the case of timelike geodesics, define $\phi_s(q) = \gamma_q(s)$. Clearly

$\phi_0 = id$. ϕ_s can be seen as a map from a 2-dim manifold (U) to another 2-dim manifold, $\phi_s'(p)$ is a 2×2 matrix. So We can define

$$A(s) = \sqrt{det(\phi_s'(p)^T \, g_{ij} \, \phi_s'(p))}$$

A(s) means the wave front of the light ray beams while transferring. (see figure. 20)

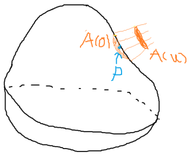


Figure 20: A(s) is a measurement of the area of the wavefront of a local light-ray beam.

Then, choose proper parameter for each $\gamma_q(s)$ so that

$$\frac{\partial X_q(0)}{\partial x_i(q)} = 0 (i = 1, 2)$$

on each q.

$X_q(0)$ means the tangent vector of γ_q at q. The equation means the beginning vector of γ_q is parallel while moving along U.

Then similar to theorem 4.2, we have the wavefront of light beams $\phi_s(U)$ is the orthogonal to γ.

Set $\theta(s) = \dot{A}/A$, θ is called the area expansion, or simply expansion

Then we can define trapped surface rigorously

Definition 5.2. (rigorous)

Trapped surface a compact 2-dim space-like C^1 surface Γ that satisfy both of the the two null geodesic γ_p, η_p orthogonal to γ at p have a positive initial area expansion. That is, $\theta_{1p}(0) > 0, \theta_{2p}(0) > 0$. θ_{1p} means the area expansion of γ_p, θ_{2p} means the area expansion of η_p

use similar calculation in section 4, we got equation similar to (15). This is called the null Raychaudhuri equation, which is a analogy to the original Raychaudhuri equation (15).

$$\frac{d\theta}{ds} + \frac{1}{2}\theta^2 = -\frac{1}{4}Tr\,\sigma^2 - 8\pi\hat{T}_{uu} \tag{17}$$

$\hat{T}_{uu}$ means the component of $\hat{T}$ at the direction of tangent vector of γ at $\gamma(s)$. Notice that (17) is a little different to (15), the coefficient of θ^2 in left hand side change from 1/3 to 1/2, this is because Γ in section 5 is 2-dim , but S in section 4 is 3-dim.

Similar to Section 4, we have

Theorem 5.2. If $\theta(0) = -\lambda$ is negative(this will be satisfied by a trapped surface) . And if the weak energy condition

$$\hat{T}_{uu} \geq 0$$

is satisfied, there will be a conjugate point of Γ at p on $\gamma(s)$ when $s \in [0, 2/\lambda_1]$. (See figure.21)

Proof. The proof is the same as theorem 4.3, you just have to change the coefficient from 3 to 2 . □

But notice in theorem 5.2 we also have to assume $(\mathcal{M}, g)$ is null-like geodesic complete, that is, any null-like geodesic can be extend to sufficiently large affine parameter. If $(\mathcal{M}, g)$ is not null-like geodesic complete, the geodesic may can't extend before it come to a conjugate point. We can say it meet a singularity before it meet a conjugate point.

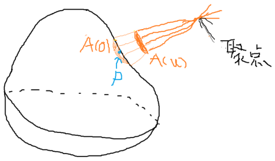


Figure 21: conjugate point will form at some parameter $u \in [0, 2/\lambda_1]$

Finally we come to Penrose singularity theorem

Theorem 5.3. (Penrose 1965)

In space-time $(\mathcal{M}, g)$ If the condition below is satisfied

(1)The weak energy condition $\hat{T}_{uu} \geq 0$ is satisfied.

(2)$(\mathcal{M}, g)$ is globally hyperbolic .And there is a Cauchy surface S that is non-compact

(3)A trapped surface Γ is formed.

Then $(\mathcal{M}, g)$ is null-like geodesic incomplete. In fact this is a evidence of the existence the singularity. We think some null-like geodesic can't extend because it "hit to a singularity". Just as Hawking ones said

"Although we have omitted the singular point from the definition of space-time, we can still recognize the 'holes' left where they have been cut out by the existence of incomplete geodesics."

(1) is satisfied by normal matter, and particularly dark energy. In fact , it's satisfied by any usual relativistic classical matter field. (2) is the assumption we always tend to believe. (3) Happens when matter density is high and gravitation is strong enough.

Proof. **Step 1: If $(\mathcal{M}, g)$ is null-like geodesic complete, then $\partial\mathcal{J}^+(\Gamma)$ is compact**

To get a contradiction , we assume $(\mathcal{M}, g)$ is null-like geodesic complete, that is, any null-like geodesic can be extend to Sufficiently large parameter.

Cause $\partial \mathcal{J}^+(\Gamma)$ is a boundary of a space-time region, then it's closed in $(\mathcal{M}, g)$ and have no boundary as a 3-dim manifold.

Cause Γ is a trapped surface, we will show that, strangely, $\partial \mathcal{J}^+(\Gamma)$ is compact. We say compactness is strange result because you can see the example in figure 1, figure 2, non of the $\partial \mathcal{J}^+$ is compact.

Cause the trapped surface Γ is compact, and $\theta_{1p}(0), \theta_{1p}(0)$ is continuous in P(P is on Γ), Then there will have a positive lower bound of $\theta_{1p}(0), \theta_{1p}(0)$, cal bound λ

Then from theorem 5.1 we got $\partial \mathcal{J}^+(\Gamma)$ is made of null-geodesic segment γ_p that is orthogonal to Γ at some point p on. Cause γ_p should be a shortest path from point on p to Γ. There should be no conjugate point of Γ on segment γ_p. But the weak energy condition tells us that, if keep extend, γ_p will come to a conjugate point when $s \in [0, 2/\lambda]$, so the parameter of point on γ_p won't be large than $2/\lambda$.

We can define map f

$$\Gamma \times [0, 2/\lambda] \times \{1, 2\} \to (\mathcal{M}, g)$$

that is , $f(p, s, 1) = \gamma_p(s)$, $f(p, s, 2) = \eta_p(s)$ (γ_p is null-geodesic going inside, η_p is null-geodesic going outside). We can see f is continuous. Cause the parameter of point on γ_p won't be larger than $2/\lambda$, then

$$\partial \mathcal{J}^+(\Gamma) \subset f(\Gamma \times [0, 2/\lambda] \times \{1, 2\})$$

Cause $\Gamma \times [0, 2/\lambda] \times \{1, 2\}$ is compact, f is continuous, we have $f(\Gamma \times [0, 2/\lambda] \times \{1, 2\})$ is compact. Then $\partial \mathcal{J}^+(\Gamma)$ is a closed subset of compact set, then $\partial \mathcal{J}^+(\Gamma)$ is compact.

Step 2: in fact $\partial \mathcal{J}^+(\Gamma)$ can't be compact

Let's also show $\partial \mathcal{J}^+(\Gamma)$ can't be compact . In fact, we can set a C^1 timelike segment field on $(\mathcal{M}, g)$(such as a direction at each point where $Length_{Lorentz}/Length_{Euclid}$ become the maximal). Then the integral curve of the segment field F will connect a point p at Γ and a point $g(q)$ at Cauchy surface S (See fig.22). Cause segment field F is C^1, then we have map g: $\partial \mathcal{J}^+(\Gamma) \to S$ is continuous.

g is also a injection, if not, two point p, q on $\partial \mathcal{J}^+(\Gamma)$ satisfy $g(p) = g(q)$, then there will be a integral curve of F called α connect q,p(p is in the future) . But α is timelike (See fig.23), so we can start at Γ, walk along a null-geodesic γ to q, and then walk along α to p, then $\gamma \cup \alpha$ is a casual path from Γ to p and its length is positive. Then p won't on $\partial \mathcal{J}^+(\Gamma)$, contradiction. So g is a injection .

Then because segment field F is C^1 , we have g^{-1} is also continuous. Then g is a homeomorphism from $\partial \mathcal{J}^+(\Gamma)$ to a subset S_0 of S. S is not compact, S_0 is homeomorphic to $\partial \mathcal{J}^+(\Gamma)$, so S_0 should be compact. So S_0 should have a boundary , but $\partial \mathcal{J}^+(\Gamma)$ is a three-dim manifold without boundary , a contradiction!

□

Let's see why the assumption that $(\mathcal{M}, g)$ is null-like geodesic complete is needed to get a contradiction. Remember in the first half of our proof, we have shows that if $(\mathcal{M}, g)$ is null-like geodesic complete, $\partial\mathcal{J}^+(\Gamma)$ is compact. If $(\mathcal{M}, g)$ is null-like geodesic incomplete, like the space time in figure.21 (the null geodesic can't extend when they meet the singularity). To make things more easy, we just consider in figure.21 spacetime is 2+1 dim, then the trapped surface S^2 is a 1-dim ring, We can see the future bound of trapped surface $\partial\mathcal{J}^+(S^2)$ is homeomorphism to a 2-dim sphere without a point, because the top of the sphere is a singularity , and it's not in our space-time, we have to remove it. So $\partial\mathcal{J}^+(S^2)$ is homeomorphism to a 2-dim open disk, which is not compact.

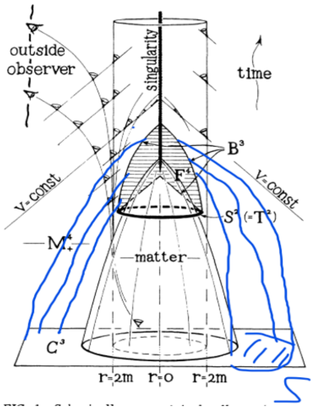


Figure 22: a homeomorphism g from $\partial\mathcal{J}^+(\Gamma)$ to S_0

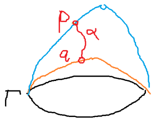


Figure 23: p, q is on $\partial\mathcal{J}^+(\Gamma)$, If there are line segment α from p to q, then α can't be time-like

At the end of massive star. nuclear fuel is consumed, no force can resist gravitation, then star will collapse. we can expect the matter-density will get very high and then a trapped surface will be formed , then things end up in one/some singularity. There are no proper physical theory near the singularity, that's the boundary of our knowledge. What the singularity really is , that's a question that is challenging the wisdom of human being.

The Penrose's work is just a new beginning of the study of black hole. His work give general relativity theorist new tools and perspective. Important work like black hole area non-decreasing theorem (Hawking 1971), No-hair theorem of blackhole(1973 and later) was made after that.

The observation evidence of black holes is also accumulated as time goes by.In 1996 and 1998, two teams lead by Genzel and Ghez, published their observation of the star orbits around Sagittarius A^*, and find it is high massive $(4.1 \times 10^6 \dot{M}_{sun})$, but located in a relatively small area. In 2015 , LIGO first detected the gravitational waves emitted from the merging of two blackholes. In 2019 and 2022, ETH release the "photo" of two massive black holes, one is at the center of galaxy M81, one is at the center of our Milky Way. The second one is exactly Sagittarius A^* which has been studied by teams of Genzel and Ghez.

6 Reference

there are four main reference

(1) **Penrose Roger (1965), "Gravitational collapse and space-time singularities"**

this is the original paper of Penrose singularity theorem on black holes , which only have less than 3 pages.

(2)**S.W.Hawking , and G.F.R.Ellis, "The large scale structure of space time", Cambridge university press,1973**,

this is a professional book on General Relativity and space-time singularity.

(3)**"Light Rays, Singularities, and All That" by Ed Witten**

They are in fact a summer camp lecture notes of Witten in 2018, you can find it on arXiv, you can find the lecture video "Light rays and black holes 1, 2" on internet.

In this lecture, Witten explain in simple language but give reader deep perspective on space-time singularity. This lecture is enjoyable if you are interested in this subject and have basic knowledge on General Relativity.

(4) **"Gravitation " by Charles W. Misner, Kip S. Thorne, and John Archibald Wheeler.**

A Big Mac textbook on General Relativity, written by three master.

An Introduction To Penrose Tiling And Its Algebraic Structures

Qiao Li

June 14th 2022

Contents

1 Introduction

A tiling is a covering of the plane by polygons or other shapes that do not overlap with each other. A tiling which cannot coincide with its original pattern when shifting any finite distance without rotation, is called a aperiodic tiling. The contrast is called an periodic tiling.

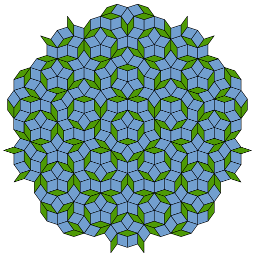


Figure 1: The Penrose tiling

The Penrose tiling, found in the 1970s, is a famous example of an aperiodic tiling. Despite its lack of translational symmetry, it possesses both reflection symmetry and fivefold rotation symmetry, aside with abundant wonderful properties. The tiling was created by Roger Penrose, and was firstly discussed by Penrose and J.H.Conway, who found some of its fantastic features and constructed some of its geometric structure.

In the early 1980s, the algebraic structure of Penrose tiling was discovered, which in my opinion means the beginnning of serious mathematical research of this marvellous phenomenon. In the following years 1982, D.Shechtman found a diffraction pattern in the metallic phase of some Al-Mn alloy turning out to have fivefold rotation symmetry, which contradicts with the existing opinion of the structure of crystals. After fighting against opposition for two years, his work was finally published, which triggered his Nobel prize at 2011 and the new concept of "quasicrystal".

There is the fabulous connection between the Penrose tiling and quasicrystals, just as the connection between periodic tiling and crystals. I will finally show in this paper that the Penrose tiling is a projection of a trivial periodic tiling with a higher dimension, the same as the quasicrystal is the projection of a high dimensional crystal, as it was written in the textbooks of solid state physics.

This paper is organized basically in two parts. At first I will summarize some of the basic properties from tiling and crystal to the fascinating Penrose tiling and quasicrystal, which serves as an exhibition of this topic. And then I will present some of the algebraic structures of the Penrose tiling, in other words do some mathematical approaches, which as an introduction, will show the profounding cannotation of this topic aside from just "Interesting Mathematics".

2 Basic Properties

2.1 From tiling to crystal

A *tiling* is a non-overlapping cover of the plane, where some obvious examples are the square tiling and the regular hexagon tiling. If we do some simple math of a tiling by a regular polygon of n edges, then we have

$$\frac{(n-2)\,\pi}{n} \times m = 2\pi \tag{1}$$

where m is the number of polygon tiles around a perigon, therefore requiring $m \in \mathbb{N}$, and we have

$$\frac{n}{n-2} \in \mathbb{N} \tag{2}$$

which leads to a solution of $n = 3, 4$, or 6.

A *crystal* is a material state where the atoms(or ions) are highly ordered, forming a lattice that extends in all directions. It usually consists of two properties known as the traslational symmetry and the long range orientational order, and the latter are usually characterized by reflection symmetry or rotation symmetry.

The Crystallographic Restriction Theorem states that if a translational symmetric crystal owns a property of n-fold rotation symmetry and reflection symmetry, then n can only take the value of $\{1, 2, 3, 4, 6\}$. The proof are also some simple maths, for if we take one vertice on a horizontal line L in the lattice and suppose the length between two adjacent vertices on the line are d, then we can rotate the line by $l\theta$ and get L_1, here $\theta = \frac{2\pi}{n}$ are the basic angle of n-fold rotation symmetry. After getting L_2 by reflecting L_1 along L, we notice the length between the mth vertice of L_1 and L_2 equals to the length between some two vertices on the original line L, that is, there exists an integer k satisfy

$$2md\cos l\theta = kd \tag{3}$$

by substitution of θ and rearranging the equation we have for every m and l,

$$2m\cos l\frac{2\pi}{n} \in \mathbb{Z} \tag{4}$$

thus we have $n \in \{1, 2, 3, 4, 6\}$.

2.2 Penrose tiling and quasicrystal

Among these tilings, aperiodic tilings seemed to be more interesting than periodic tilings, and many methods to produce an aperiodic tilings were found gradually. An example are the "reptile" method. Its main idea is to create a fractual pattern by putting the tiles together to form a same tile with a bigger size. Thus any finite-distanced shifting can be considered in a large enough pattern same with the tile, and one prooves the shifting cannot coincide with the original tiling.

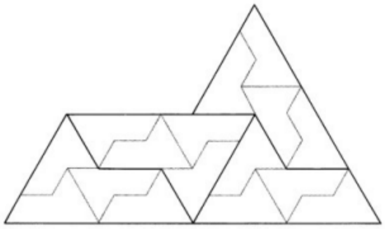


Figure 2: The sphinx reptile figure

But there is the problem that the tiles one can find to create a reptile can always be combined together trivially to form a periodic tiling, like a parallelogram tiling. So mathematicians began searching for an aperiodic tiling S such that for any set of tiles $T \subset S$, T cannot form a periodic tiling.

In 1961 Hao Wang guessed that such S do not exist, but in 1964 R. Berger constructed a S satisfy $card\,(S) = 104$. Later D. Knuth decreased the number to 92, and finally in 1974, R. Ponrose gave a construction of S such that $card\,(S) = 4$, which is the original Penrose Tiling.

Suppose there is a crystal lattice $L = \{(x, y) : x, y \in \mathbb{Z}\} \subset \mathbb{R}^2$ on the plane and let $\phi = \frac{1+\sqrt{5}}{2}$ be the golden ratio, we pick out the set of points $T \subset L$ such that every $(x, y) \in T$, the square $S = \{(m, n) : x - 1 < m \leq x, y - 1 < n \leq y\} \subset \mathbb{R}^2$ has an nonempty intersection with the line $y = (\phi - 1)\,x$. Then the projections of all points in T onto the line create a 1-dimensional pattern that is long range orientational ordered, but not translational symmetric. This can be seen as the 1D quasicrystal, also a 1D tiling, if we connect two consecutive projections on the line and take them as 1D tiles. What linked the Penrose tiling with quasicrystals are the generalization to 5D spaces, when an appropriate plane intersect with R^5 lattices and create the Penrose tiling pattern, which will be finally proved as Theorem 2 down below.

2.3 Geometric properties

Apart from the original 6-pieces tiling, two types of the eventual version all consist two pieces, namely "the kite and dart" and "the thick rhombus and the thin rhombus". But the two types are essentially the same, as there exist a transformation between them. As a remark, the Penrose tiles

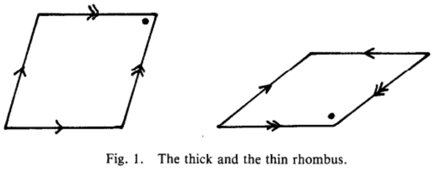


Figure 3: The thick and thin rhombuses with arrows as matching rules

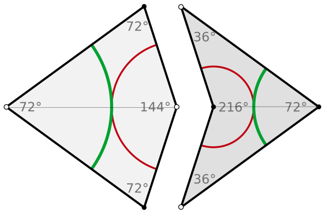


Figure 4: The kite and dart tiles with arcs as matching rules

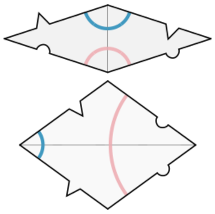


Figure 5: The substitutional tiles serving as matching rules

aren't just two polygons of "kite and dart", or "thick and thin rhombuses", but along with a matching rule to make sure the tiles cannot be combined together to form a trivial parallelogram periodic tiling. The matching rules can be in the form of arrows needing to be matched along along a coincident edge, or arcs inside the polygon needing to be continuous, or simply substitutional tiles as in the example figures.

Using these tiles one can easily create a Penrose tiling, but the numbers of different Penrose tilings are infinite, actually uncountable. Here two tilings are different means that one cannot coincide with the other by shifting a finite distance without rotation. Nevertheless, any finite pattern in any Penrose tiling can be found in every other Penrose tilings, and their numbers in every other Penrose tiling are infinite. Futhermore, Conway proved a so-called *local isomorphism theorem*, stating that a finite pattern of a circle with radius d in any Penrose tiling, can be found in somewhere at most distance $\phi^3 d$ from any point in any other Penrose tilings, here ϕ stands for the golden ratio $\frac{1+\sqrt{5}}{2}$.

3 Algebraic Structures

3.1 Notations and remarks

The main idea of this part is to prove that from five generating parameters $\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4$ along with the requirements $\sum_{j=0}^{4} \gamma_j = 0$ (thus it's four degree of freedom) one can produce a Penrose tiling pattern on R^2, or name it, the complex plane $\mathbb{C}$. I will like to name this Theorem 1 the pentagrid generating theorem. Also by using the pentagrid generating theorem, we can prove the Theorem 2 mentioned above, which I will name it the plane intersection theorem.

A remark is needed, that in this section we will work on the Penrose tiling of thick and thin rhombuses, instead of the kite and dart pattern above. And the matching rules will be in the form of arrows attached to each edge of the thick and thin rhombus, such that the matching of the two rhombus must satisfy that the coincident edge have the arrows of the same direction. So therefore we can give a definition.

Def 3.1. *we call a tiling "rhombus tiling" if the tiles are the thick and thin rhombuses, and a tiling "AR-rhombus tiling" if the two rhombuses are arrowed according to the matching rule.*

We take ζ as the quintic unit root $e^{i\theta}$, and $\theta = \frac{2\pi}{5}$. All the $\sum_j$ means $\sum_{j=0}^{4}$. We take $\lceil x \rceil$ as the roof function of x, means the least integer m satisfy $m \geq x$. So we will start from pentagrids.

3.2 Pentagrids

Def 3.2. *Let $\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4 \in \mathbb{R}$ satisfy $\sum_j \gamma_j = 0$, then define*

$$G_j := \left\{ z \in \mathbb{C} : Re\left(z\zeta^{-j}\right) + \gamma_j \in \mathbb{Z} \right\} \subset \mathbb{C} \tag{5}$$

to be named as *the j-th grid.*

By some deformation we can see that

$$G_j = \left\{ z : z = \zeta^j \left(n - \gamma_j + \alpha i\right), n \in \mathbb{Z}, \alpha \in \mathbb{R} \right\} \tag{6}$$

is a cluster of lines with distance 1 rotated $j\theta$ counterclockwise from the vertical initial state $z = -\gamma_j + \alpha i$.

Then we call

$$G := \bigcup_j G_j \tag{7}$$

the pentagrid (generated by $\left(\gamma_j\right)_j$).

If there does not exist three gridlines intersecting together, we call G a *regular pentagrid*. Otherwise *singular*.

In this paper we always talk about regular pentagrids.

3.3 Theorem 1

We now state the pentagrid generating theorem.

Theorem 1. *By the following operations of a pentagrid G generated by $\left(\gamma_j\right)_j$, one can construct a rhombus tiling and attach arrows to it to create an AR-rhombus tiling, which is a Penrose tiling.*

Proof. We will first show the method to generate a rhombus tiling.
Define

$$K_j\left(z\right) := \left\lceil Re\left(z\zeta^{-j}\right) + \gamma_j \right\rceil \in \mathbb{Z} \tag{8}$$

and $f : \mathbb{C} \to \mathbb{C}$ as

$$f(z) = \sum_j K_j(z)\zeta^j \tag{9}$$

we now consider the behavior of $K_j(z)$. Due to the property of roof functions, we know $K_j(z)$ does not change as long as $Re\left(z\zeta^{-j}\right) + \gamma_j$ is less than or equal to a integer n. Then with reference to expression(6), we can find that $K_j(z)$ holds as a constant between two adjacent lines in G_j.

Thus we can observe the behavior of $f(z)$, and know that $f(z)$ changes(as a complex number, or a vector) by ζ^j iff when z crosses a gridline of G_j. Hence if we consider the intersection point P of a gridline l_i from G_i and l_j from G_j and the four meshes A, B, C, D around P, with A, B on one side of l_i, and A, D on one side of l_j, then we have

$$f(A) - f(B) = f(D) - f(C) = \zeta^j \tag{10}$$

$$f(A) - f(D) = f(B) - f(C) = \zeta^i \tag{11}$$

thus the difference under f between adjacent meshes, be like ζ^j, form an edge of a rhombus, since the module of ζ^j is 1 for all j.
In other words, if we take $f(\mathbb{C}) \subset \mathbb{C}$ as the set of vertices of rhombuses with edge lenth 1, we form a rhombus tiling of the plane $\mathbb{C}$. The proof of this statement only need a check of $|f(z)|$ goes to $+\infty$, which is obvious if we take

$$\lambda_j(z) := K_j(z) - (Re\left(z\zeta^{-j}\right) + \gamma_j) \tag{12}$$

and notice both $0 \leq \lambda_j(z) < 1$ and $f(z) = Re(z) + \sum_j(\gamma_j + \lambda_j(z)\zeta^j)$.

The next thing we do now is to attach arrows to this rhombus tiling. As a remark, the arrowing process is necessary since we have to make sure these rhombuses are put together to form a Penrose tiling instead of some trivial parallelogram periodic tiling.

We will first assign an index to every vertice. Let $\lambda_j(z)$ be defined as above, and we can observe that

$$\sum_j K_j(z) = \sum_j \lambda_j(z) \tag{13}$$

this is because there holds both $\sum_j \zeta^j = 0$ and $\sum_j \gamma_j = 0$ as we supposed. Noticing that the LHS of expression (13) is the sum of five integers while the RHS is the sum of five real numbers each in the interval [0,1), we know that $\sum_j K_j(z) \in \mathbb{Z} \cap [0,5)$. Furthermore $\lambda_j(z) = 0$ means $Re\left(z\zeta^{-j}\right) + \gamma_j \in \mathbb{Z}$, which means z falls on the gridline of G_j. Since we suppose all pentagrids are regular, then for a fixed z, $\lambda_j(z) = 0$ can only hold for at most two j. Thus expression (13) cannot be zero, and we have

$$\sum_j K_j(z) \in \{1,2,3,4\} \tag{14}$$

and we call this value the index of the vertice $f(z)$.

Now we will start attaching arrows according to the index of vertices. First we notice that adjacent vertices in rhombus tiling represents neighbouring meshs in pentagrids, thus transporting along a fixed edge only changes a unique $K_j(z)$ by 1, so the index difference between the starting and ending vertices is also 1. Hence we only have to deal with the situation when a edge connects two vertices of the index {1,2} or {2,3} or {3,4}. We will arrow as follows.

$$2 \twoheadrightarrow 1 \tag{15}$$

$$3 \twoheadrightarrow 4 \tag{16}$$

And between the edge of vertice 2 and 3 we connect with one-head arrow although the directions are yet unknown. Noticing that in the original matching rules of thick and thin rhombuses, the directions of one-head arrows are determined after all the two-head arrows are fixed, then we have finished arrowing the rhombus tiling. Now the only thing we need to prove of theorem 1 is this operation is well-defined, that is, the neighbouring rhombus have the same direction of one-head arrow(if it is a one-head arrow) on their coincident edges.

We will do some simplification. Without loosing generality, we can assume the edge connecting the {2,3} vertice is horizontal, and by the map $\gamma_j \mapsto \gamma_j - Re(d\zeta^{-j})$ for every j, we still have $\sum_j \gamma_j = 0$, while in the expression of G_j, $Re\left(z\zeta^{-j}\right) + \gamma_j$ was replaced by $Re\left((z-d)\zeta^{-j}\right) + \gamma_j$ and thus causing the tiling pattern shifting d rightward for every d. Therefore we can suppose $\gamma_0 = 0$ and the 2-3 edge in the rhombus tiling correspond to the imaginary axis in the pentagrid.

we suppose the neighbouring rhombus correspond to the vertice A, B of the pentagrid, and the two vertice with index {2,3} correspond to the mesh P, Q of the pentagrid. Then by the simplification above, we know that P, Q are adjacent against the gridline G_0. Furthermore we assume A is the intersection of G_0 with G_p, and B is the intersection of G_0 with G_q.

Then by observing the matching rules of the thick and the thin rhombus, we know that the one-head arrow always points from the vertice of the acute angle to the vertice of the obtuse angle. Thus we just need to prove the angel of vertice P in A and in B are both acute or obtuse. From the generating function $f(z)$ we know that the angle of P in A is the angle between the radial ζ^p and the positive real axis $x > 0$, while angle of P in B equal to the angle spanned by radial ζ^j and $x > 0$. Drawing them out we found when i=1,4 the angles of ζ^i and $x > 0$ are acute while when i=2,3 these angle are obtuse. Thus what we need to prove is that p and q have different parity.

imagine z moving upward on the imaginary axis of the pentagrid, we have

$$K_j(iy) = \lceil Re(iy\zeta^{-j}) + \gamma_j \rceil = \lceil \sin(j\theta)y + \gamma_j \rceil, j = 1, 2, 3, 4, y \in \mathbb{R} \tag{17}$$

since the pentagrid is regular and iy is already on G_0, then $\sin(j\theta)y + \gamma_j \in \mathbb{Z}$ can only hold for at most one j. Thus noticing $\sin(1\theta) + \sin(4\theta) = 0$ and $\sin(2\theta) + \sin(3\theta) = 0$, we conclude that $\gamma_1 + \gamma_4 \notin \mathbb{Z}$ and $\gamma_2 + \gamma_3 \notin \mathbb{Z}$.

we then let

$$g_1(y) := K_1(iy) + K_4(iy) - \lceil \gamma_1 + \gamma_4 \rceil \tag{18}$$

$$g_2(y) := K_2(iy) + K_3(iy) - \lceil \gamma_2 + \gamma_3 \rceil \tag{19}$$

and $g_i(y)(i = 1, 2)$ have the form $\lceil a \rceil + \lceil b \rceil - \lceil a + b \rceil$ with $a + b \notin \mathbb{Z}$, and by eumeration we know $g_i(y)(i = 1, 2)$ takes its value in $\{0, 1\}$.

Now we let y goes from $-\infty$ to $+\infty$, and consider the behavior of $g_i(y)$. Since $K_j(z)$ is invariant as long as z do not cross the lines in G_j, and when z do crosses, $K_j(z)$ changes by 1, we know that whenever iy coincide with G_1, $K_1(iy)$ increases by 1 and thus $g_1(y)$ increases by 1. Whenever iy coincide with G_4, $K_4(iy)$ decreases by 1 and thus $g_1(y)$ decreases by 1. The same work with 2,3 as it is with 1,4. So here is our proof of p, q differ in parity.

First we proof $p \neq q$. This is because $g_1(y)$ oscillates between 0 and 1, thus the intersection with lines in G_1 and lines in G_4 alternates. The same works for G_2 and G_3, so we know that two adjacent intersection points A, B cannot be from the same G_i, which means $p \neq q$.

Next if p and q have the same parity, then either $\{p, q\} = \{1, 3\}$ or $\{p, q\} = \{2, 4\}$. We compute the index of vertice P. Note that by simlification above, we have set edge PQ horizontal on the opposite of 0-th grid. Therefore the index of Q is always one more of the index of P. By assumption edge PQ was assigned with a one-head arrow, then vertice P should have index 2 and Q with 3, but since $\gamma_0 = 0$ we have $\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 = 0$, followed by the fact from enumeration that $\lceil \gamma_1 + \gamma_4 \rceil + \lceil \gamma_2 + \gamma_3 \rceil = 1$. Hence

$$Ind(P) = \sum_j K_j(iy) = g_1(y) + g_2(y) + 1 \tag{20}$$

where iy, on the imaginary axis, is between the point A and B. We suppose A is above B.

If $p = 1$ and $q = 3$, then when iy goes from B to A, it have already crossed G_3 but not yet crossed G_1. Thus $g_2(y)$ have already decreased but $g_1(y)$ not yet increased, which indicates $g_1(y) = g_2(y) = 0$. If $(p, q) = (3, 1)$, or $(4, 2)$, or $(2, 4)$, we correspondently have $(g_1(y), g_2(y)) = (1, 1)$, or $(1, 1)$, or $(0, 0)$. Without exception $g_1(y) + g_2(y)$ is even, and thus $Ind(P)$ is odd, which means it cannot equal to 2. So this is a contradiction, and we finished our proof of theorem 1. □

3.4 Theorem 2

We will now state the plane intersection theorem.

Theorem 2. *The vertices of the Penrose tiling generated by the parameters* $\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4$ *in Theorem 1 have the form* $\sum_j k_j\zeta^j$*, where* $(k_0, k_1, k_2, k_3, k_4) \in \mathbb{R}^5$ *satisfy the 5D cube* $\{(x_j)_j \in \mathbb{R}^5 : k_j - 1 < x_j \leq k_j, \forall j\}$ *have an intersection with the plane below:*

$$\prod : \begin{cases} \sum_j x_j = 0 \\ \sum_j (x_j - \gamma_j) Re(\zeta^{2j}) = 0 \\ \sum_j (x_j - \gamma_j) Im(\zeta^{2j}) = 0 \end{cases}$$

Proof. If $(x_j)_j \in \prod$, in other words satisfy the three conditions above, then the vector $(x_j - \gamma_j)_j \in \mathbb{C}^5$ is vertical to the following three vectors in $\mathbb{C}^5$:
$\alpha_0 = (1,1,1,1,1) = (\zeta^{0j})_j$, $\alpha_2 = (\zeta^{2j})_j$, and $\alpha_{-2} = (\zeta^{-2j})_j$. If we expand them to a basis of $\mathbb{C}^5$, the missing two vectors can be $\alpha_1 = (\zeta^j)_j$ and $\alpha_{-1} = (\zeta^{-j})_j$. Thus $(x_j - \gamma_j)_j \in span(\alpha_1, \alpha_{-1})$, and we have $z_1, z_2 \in \mathbb{C}$ such that

$$x_j - \gamma_j = z_1\zeta^j + z_2\zeta^{-j} \tag{21}$$

Noticing the LHS of (21) is real and ζ and ζ^{-j} mutually conjugate, we have z_1 conjugate to z_2, and thus

$$x_j - \gamma_j = 2Re(z_2\zeta^{-j}) = Re(2z_2\zeta^{-j}) \tag{22}$$

Then for any $(k_j)_j$ satisfying the 5D cube attached to it intersects with $\prod$, there exist $(x_j)_j$, or say, $2z_2$, such that

$$k_j = \lceil x_j \rceil \tag{23}$$

or say,

$$k_j = \lceil Re(2z_2\zeta^{-j} + \gamma_j) \rceil \tag{24}$$

hence $\sum_j k_j\zeta^j = f(2z_2)$, and thus is a vertice of the generated Penrose tiling. The other direction is the same. □

4 Summarize

4.1 What we have done

In this paper, we first talked about crystal and tiling of a plane and the correlation between them, then we discussed the similar correlation between quasicrystal and the Penrose tiling. Afterwards we went through some of the geometrical properties of the Penrose tiling, and finally introduced a mathematical approach of its algebraic structures, whose main idea is to generate a pentagrid from five parameters, and present a dual relation between the intersection points of the pentagrid with the rhombus of the Penrose tiling, and the mesh of the pentagrid with the vertice of the Penrose tiling, assigning an index with each vertice and finally generate a Penrose tiling pattern by arrowing the rhombus according to the index of its vertice.

4.2 Forecast

Upon the basis of the algebraic structure, more further properties were found and discussed, not only in algebraic fields, but also analytic fields, such as viewing the space of all Penrose tiling as a metric space, since it has a “natural compact metric topology” and discuss its properties as a strictly ergodic dynamical system, which is done in Robinson, E. A. (1996). The Dynamical Properties of Penrose Tilings. *Transactions of the American Mathematical Society*, 348(11), 4447–4464.

Other triggering thoughts are that mathematics again have gone in front of physics, for the Penrose tiling was found in 1974, and its algebraic structures, which is the main topic of this paper, is found in 1981, just enough to get before the time when it was applied to explain the phenomenon of quasicrystals. Just as before when calculus was found in the 17th centry, and Riemann geometry in the 19th, and many other unlimited examples, mathematics have presented us somewhat magical but end up conclusive methods, tools and models for us to recognize our world.

5 References

1. N.G. de Bruijn, Algebraic theory of Penrose's non-periodic tilings of the plane. I. *Indagationes Mathematicae*

2. N.G. de Bruijn, Algebraic theory of Penrose's non-periodic tilings of the plane. II. *Indagationes Mathematicae*

3. Gardner, M. - *Scientific American*, Jan. 1977, p. 110-121.

4. D. Shechtman, I. Blech, D.Gratias and J.W. Cahn, Metallic Phase with Long-Range Orientational Order and No Translational Symmetry, *Phys. Rev. Lett.*, 53 (1984)1951-1953.

5. https://www.wikipedia.org

An Introduction to Ising Model

Kewei Wang, Xian-bao Sha

June 11, 2022

Abstract

This article is an introduction to Ising model, including the combinatorial solution of 2D Ising model and the thermodynamic properties of a large crystal.

1 Introduction

The Ising model is an important model used in statistical physics to deal with various phase transition problems. A phase transition point of a crystal means a discontinuity point of its particular thermodynamic function (e.g., free energy, internal energy, specific heat, etc., as a function of temperature T) or the derivative of some order of this function. In order to determine whether a phase transition will occur in a crystal, we need to find a way to calculate the thermodynamic functions. The main idea of the Ising model is to simplify the calculation of each physical quantity by arranging the atoms in the crystal in a n-dimensional grid ($n \leqslant 3$) and considering only the mutual energy between adjacent atoms in the grid.

In the following, we will give a concrete construction of Ising model. Take $n = 2$ as an example. Consider a L-row M-column square lattice with $N = L \times M$ lattice points, one atom at each lattice point, and two possible spin states for each atom: $\sigma_i = \pm 1, i = 1, 2, \cdots, N$. Thus, there are 2^N configurations of the spin states of the particles at all lattice points. We assume that the energy of this system consists only of the mutual energy between adjacent atoms. Under this assumption we can obtain the energy (also called the Hamiltonian) of the particular configuration $\sigma = (\sigma_1, \sigma_2, \cdots, \sigma_N)$:

$$E_\sigma := -J \sum_{\{i,j\}} \sigma_i \sigma_j \tag{1.1}$$

where the summation is over all $\{i, j\}$ corresponding to adjacent atoms, and J is the mutual energy constant.

According to statistical physics, all configurations of σ are possible and the probability of each configuration is proportional to $e^{-\beta E_\sigma}$, where $\beta = \frac{1}{k_B T}$, k_B is Boltzmann's constant, and T is the absolute temperature.

What we are concerned with is the normalization constant Z_N of this distribution, that is:

$$Z_N := \sum_{\sigma} e^{-\beta E_\sigma} \tag{1.2}$$

Z_N is also called the partition function in statistical physics. We will give the definition of the relevant thermodynamic functions, and will see that these functions correspond to the partition function and its partial derivatives of different orders, thus the solution of the partition function is the key to solving the Ising model.

Definition 1.1. The free energy ψ_N, internal energy U_N, and specific heat C_N of the system is given by

$$\psi_N := -N^{-1} k_B T \ln Z_N \tag{1.3}$$

$$U_N := \frac{1}{N} \sum_{\sigma} E_\sigma \frac{e^{-\beta E_\sigma}}{Z_N} = k_B T^2 \frac{\partial}{\partial T} \left(-\frac{\psi_N}{k_B T}\right) \tag{1.4}$$

$$C_N := \frac{\partial U_N}{\partial T} \tag{1.5}$$

The results obtained by Onsager are as follows. We will prove this theorem by a combinatorial approach in Section 2.

Theorem 1.2 (Onsager)**.**

$$-\frac{\psi}{k_B T} = \ln 2 + \frac{1}{2\pi^2} \int_0^\pi \int_0^\pi \ln \left((\cosh 2\beta)^2 - \sinh 2\beta (\cos \xi + \cos \eta) \right) \mathrm{d}\xi \mathrm{d}\eta \tag{1.6}$$

when $|\tanh \beta| < \dfrac{1}{4}$*, where* $\psi = \lim\limits_{N \to \infty} \psi_N$*.*

2 Combinatorial Solution of 2D Ising Model

The main idea of solving the 2D Ising model by combinatorial methods is to transform it into a combinatorial counting problem on graphs. The methods contain three main steps, which are shown in the following three subsections respectively.

2.1 Transforming into a Graph Problem

First we will see how the problem is related to graphs. WLOG, we may assume that $J = 1$, or just let $\beta = J/k_B T$.

Consider a 2D square lattice with $L \times L$ sites, and denote all adjacent particle pairs by $\mathcal{N}$, then we can naturally generate a graph $G = (V, \mathcal{N})$ by the lattice, where V contains all sites and $\mathcal{N}$ exactly contains all edges. The partition function of the 2D Ising model on the lattice can be calculated as follow:

$$\begin{aligned} Z_{L^2} &= \sum_{\sigma} e^{-\beta E_\sigma} \\ &= \sum_{\sigma_1, \sigma_2, \cdots, \sigma_{L^2} = \pm 1} \prod_{\{i,j\} \in \mathcal{N}} e^{\beta \sigma_i \sigma_j} \\ &= \sum_{\sigma_1, \sigma_2, \cdots, \sigma_{L^2} = \pm 1} \prod_{\{i,j\} \in \mathcal{N}} (\cosh \beta + \sigma_i \sigma_j \sinh \beta) \\ &= (1 - u^2)^{-L(L-1)} \sum_{\sigma_1, \sigma_2, \cdots, \sigma_{L^2} = \pm 1} \prod_{\{i,j\} \in \mathcal{N}} (1 + \sigma_i \sigma_j u) \end{aligned} \tag{2.1}$$

where $u = \tanh \beta$, and the third equation holds because of $\sigma_i \sigma_j = \pm 1$.

Notice that

$$\prod_{\{i,j\} \in \mathcal{N}} (1 + \sigma_i \sigma_j u) = \sum_{m=0}^{2L(L-1)} u^m \left(\sum_{H \in \mathcal{A}_m} \prod_{\{i,j\} \in \mathcal{N}_H} \sigma_i \sigma_j \right) \tag{2.2}$$

where $\mathcal{A}_m$ is the set of all subgraphs of G that contains m edges, and $\mathcal{N}_H$ is the set of all edges of H.

By observing the degree of each σ_i in (2.2), we can see that since the summation is over $\sigma_i = \pm 1$ for each i, only even subgraphs (that is, the degree of each vertex in the subgraph is even) can make contributions to the summation. Let $\mathcal{A}$ be the set of all even subgraphs of G, and $m(H) = \#_{\mathcal{N}_H}$, then after a simple calculation we can obtain

$$Z_{L^2} = 2^{L^2}(1-u^2)^{-L(L-1)} \left(1 + \sum_{H \in \mathcal{A}} u^{m(H)} \right) \tag{2.3}$$

2.2 An Important Combinatorial Identity

In this subsection, we are going to obtain an important combinatorial identity, which transforms the summation over all even subgraphs into a product of paths, in order to simplify the combinatorial expression (2.3) for the convenience of subsequent solving. But first we need to explain some definitions and notations.

Definition 2.1. A **path** p over G is a sequence of directed edges $(e_1, e_2, \cdots, e_n)$, with each e_k starting at the site where e_{k-1} ended and never goes backwards over e_{k-1}, $k = 2, \cdots, n$. A **closed path** is a path that starts and ends at the same site. The **equivalence class** $[p]$ consists of all equivalent closed paths of p, that is, all circular permutations $(e_k, e_{k+1}, \cdots, e_n, e_1, \cdots, e_{k-1})$ and their inversions $(e_{k-1}, \cdots, e_1, e_n \cdots, e_k)$. A **periodic path** is a periodic sequence which repeats a non-periodic sequence of edges which belongs to a closed path for m times ($m > 1$).

Remark 2.2. In this subsection, all paths we mention are closed path, for simplicity.

Besides, we need to define the sign and weight of a path as follows.

Definition 2.3. The **sign** of a closed path p is given by

$$s(p) := (-1)^{1+t} \tag{2.4}$$

where $2\pi t$ is the angle turned by a tangent vector while traversing p. And the **weight** of p is given by

$$W_p(u) := s(p)u^n \tag{2.5}$$

where n is the number of edges in p.

Example 2.4. A path p is given in Figure 1, whose 4 edges form a square. As we traverse p counterclockwise, the tangent vector will turn by 2π, since at every site it will turn by $\pi/2$. So the sign $s(p) = (-1)^{1+1} = 1$.

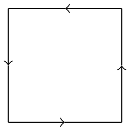


Figure 1: A simple example of sign

With these definitions, we can try to obtain the identity shown in the following theorem.

Theorem 2.5.

$$1+\sum_{H\in\mathcal{A}} u^{m(H)} = \prod_{[p]}(1+W_p(u)) \tag{2.6}$$

where the product is over all equivalence class $[p]$ of non-periodic paths.

This identity was initially proposed as a conjecture in lecture notes by Feynman in 1960, and then proved by Sherman in 1960 and Burgoyne in 1963. The convenience of this transformation lies in the fact that it turns the sum of subgraphs, which is difficult to compute, into a product that is relatively easy to compute.

The proof of the identity contains two steps. First we will expand the product into edge-disjoint paths, that is:

$$\prod_{[p]}(1+W_p(u)) = 1+\sum_{k=1}^{\infty}\sum_{[p_1],[p_2],\cdots,[p_k]} W_{p_1}(u)W_{p_2}(u)\cdots W_{p_k}(u) \tag{2.7}$$

where $[p_1],[p_2],\cdots,[p_k]$ are edge-disjoint. After showing this, what remains to prove is that the terms $W_{p_1}(u)W_{p_2}(u)\cdots W_{p_k}(u)$ where at least one edge e_i is traversed for $r_i>1$ times all cancel out. This part is a bit difficult, therefore is omitted here and only the sketch of the first part will be shown as follows.

Sketch of proof of Theorem 2.5. For each even subgraph H of G, the way to decompose it into different edge-disjoint paths depends on all sites of degree 4 in H. We want to show that

$$\sum_{k=1}^{\infty}\sum_{[p_1]\cup[p_2]\cup\cdots\cup[p_k]=H} W_{p_1}(u)W_{p_2}(u)\cdots W_{p_k}(u) = u^{m(H)} \tag{2.8}$$

WLOG, we may assume that H is connected, otherwise the equality can be obtained by simply multiplying the equalities of connected subgraphs.

For each site of degree 4 in H, the paths will have a crossing at this site, which has three types: selfcrossing, turning left and turning right (see Figure 2). Let t_i be the number of the ith type crossing ($i=1,2,3$), respectively, and let $n_4=t_1+t_2+t_3$. A simple identity is:

$$s(p_1)s(p_2)\cdots s(p_k) = (-1)^{t_1} \tag{2.9}$$

which comes from the fact that a path without any selfcrossings has a sign of 1, and every self-crossing make a -1 contribution to the sign.

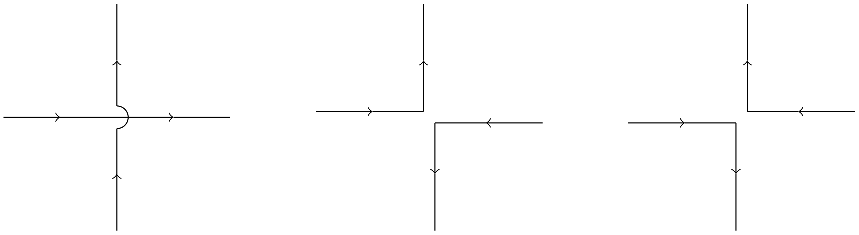


Figure 2: Three types of path crossing

Since every site of degree 4 can have all three types of crossings, the summation in (2.8) can be written as:

$$\begin{aligned}\sum_{k=1}^{\infty}\sum_{[p_1]\cup[p_2]\cup\cdots\cup[p_k]=H} W_{p_1}(u)W_{p_2}(u)\cdots W_{p_k}(u) &= u^{m(H)}\sum_{t_1+t_2+t_3=n_4}\frac{n_4!}{t_1!t_2!t_3!}(-1)^{t_1}1^{t_2}1^{t_3}\\ &= (-1+1+1)^{n_4}u^{m(H)}\\ &= u^{m(H)}\end{aligned} \tag{2.10}$$

Then after making an sum over all even subgraph H, the theorem is proved. □

2.3 Combinatorial Approach to Onsager's Formula

After the effort of the last subsection we have transformed the partition function into

$$Z_{L^2} = 2^{L^2}(1-u^2)^{-L(L-1)}\prod_{[p]}(1+W_p(u)) \tag{2.11}$$

In this subsection we will start from this and finally give a proof of Theorem 1.2.

Let's consider all paths that start at a particular site P_1 and end at the same site $P_{n+1} = (x, y)$ in n steps. Since what we concern is the asymptotic properties of the system as $L \to +\infty$, the boundary conditions can be ignored, and thus we may assume that $P_1 = (0, 0)$. First we need to define the amplitude of this union of paths.

Definition 2.6. The **amplitude** of a path p is given by

$$\overline{W}_p(u) := \alpha^{n_l}\bar{\alpha}^{n_r}u^n \tag{2.12}$$

where $\alpha = e^{i\pi/4}$, and n_l, n_r are the number that p turns left and right, respectively. The **upward amplitude** $U_n(x, y)$ is the summation of amplitude of all paths that start from $(0, 0)$ and moving upward to (x, y) in the n-th step. Similarly is the **downward amplitude** $D_n(x, y)$, **leftward amplitude** $L_n(x, y)$, **rightward amplitude** $R_n(x, y)$ defined. Moreover, if $|x| + |y| > n$ (which means no such path exists), then $U_n(x, y) = D_n(x, y) = L_n(x, y) = R_n(x, y) = 0$.

There is one additional note to be made about this definition: we should give the path a initial direction, so that the turning from it to the first step will also be counted in. For a non-closed path, we can randomly set this direction, for example all upward. But for a closed path p that start and end at $(0, 0)$, we should let this direction be the same as the last step, which makes

$$\overline{W}_p(u) = -W_p(u) \tag{2.13}$$

hold.

With the definitions above, we can easily get the recurrence relation of $U_n(x, y)$ as well as the other three amplitudes as follows.

Proposition 2.7. *For all $n \in \mathbb{N}$, $(x, y) \in \mathbb{Z}^2$,*

$$\begin{aligned}U_n(x,y) &= uU_{n-1}(x,y-1) + 0D_{n-1}(x,y-1) + u\bar{\alpha}L_{n-1}(x,y-1) + u\alpha R_{n-1}(x,y-1)\\ D_n(x,y) &= 0U_{n-1}(x,y+1) + uD_{n-1}(x,y+1) + u\alpha L_{n-1}(x,y+1) + u\bar{\alpha}R_{n-1}(x,y+1)\\ L_n(x,y) &= u\alpha U_{n-1}(x+1,y) + u\bar{\alpha}D_{n-1}(x+1,y) + uL_{n-1}(x+1,y) + 0R_{n-1}(x+1,y)\\ R_n(x,y) &= u\bar{\alpha}U_{n-1}(x-1,y) + u\alpha D_{n-1}(x-1,y) + 0L_{n-1}(x-1,y) + uR_{n-1}(x-1,y)\end{aligned} \tag{2.14}$$

Now we want to calculate these amplitudes at $(0,0)$. The equations above contain translations, but after Fourier transform these translations will be converted to phase shifts, making recurrence relations a lot simpler. Here we use the discrete-time Fourier transform (DTFT), which transforms a sequence into a periodic function.

Lemma 2.8 (DTFT)**.** *Take any $F_n(x,y) \in \mathcal{B}_n(x,y) := \{U_n(x,y), D_n(x,y), L_n(x,y), R_n(x,y)\}$, for all $\xi, \eta \in [0, 2\pi]$, the DTFT of $F_n(x,y)$*

$$\widehat{F}_n(\xi,\eta) := \sum_{x=-\infty}^{\infty} \sum_{y=-\infty}^{\infty} F_n(x,y) e^{-i(\xi x + \eta y)} \tag{2.15}$$

is well defined and the inverse Fourier transform holds:

$$F_n(x,y) = \frac{1}{(2\pi)^2} \int_0^{2\pi} \int_0^{2\pi} \widehat{F}_n(\xi,\eta) e^{i(\xi x + \eta y)} \mathrm{d}\xi \mathrm{d}\eta \tag{2.16}$$

Since $F_n(x,y) = 0$ when $|x| + |y| > n$, the summation in (2.15) is actually a finite sum, so it is well defined. Therefore the sum and the integral in (2.16) can be interchanged, by a simple calculation we can verify that the lemma holds.

Apply the DTFT to (2.14), then the translations will disappear due to the summation and the new recurrence relations are as follows.

Proposition 2.9. *The Fourier transform of $F_n(x,y)$ satisfies:*

$$\phi_n(\xi,\eta) := (\widehat{U}_n(\xi,\eta), \widehat{D}_n(\xi,\eta), \widehat{L}_n(\xi,\eta), \widehat{R}_n(\xi,\eta))^T = uM\phi_{n-1}(\xi,\eta) \tag{2.17}$$

where the entries of M can be denoted by $\alpha, v := e^{-i\eta}, h := e^{i\xi}$:

$$M = \begin{pmatrix} v & 0 & \bar{\alpha}v & \alpha v \\ 0 & \bar{v} & \alpha\bar{v} & \overline{\alpha v} \\ \alpha h & \bar{\alpha}h & h & 0 \\ \overline{\alpha h} & \alpha\bar{h} & 0 & \bar{h} \end{pmatrix} \tag{2.18}$$

Now it's time to finally prove Theorem 1.2.

Proof of Theorem 1.2. Consider all closed paths that start and end at $(0,0)$. Remind that if one such path finally moves upward to $(0,0)$, then its initial direction is also upward. That means when solving $\widehat{U}_n(\xi,\eta)$ by (2.17), $\phi_0(\xi,\eta)$ should be set as $(1,0,0,0)^T$, thus we can obtain

$$\widehat{U}_n(\xi,\eta) = \left((uM)^n(1,0,0,0)^T\right)_1 = (uM)^n(1,1) \tag{2.19}$$

Similar are the other three $\widehat{F}_n(\xi,\eta)$. So the summation over all $\widehat{F}_n(\xi,\eta)$ equals to $\mathrm{tr}(uM)^n$. Combining with the inverse Fourier transform we can obtain

$$\sum_{F_n \in \mathcal{B}_n} F_n(0,0) = \frac{1}{(2\pi)^2} \int_0^{2\pi} \int_0^{2\pi} \mathrm{tr}(uM)^n \mathrm{d}\xi \mathrm{d}\eta \tag{2.20}$$

This summation is actually the sum over all closed paths with length n (periodic and non-periodic), according to (2.13), just differing by a constant. Since there are L^2 sites, we have

$$\sum_{p(n)} W_p(u) = -\frac{L^2}{(2\pi)^2} \int_0^{2\pi} \int_0^{2\pi} \mathrm{tr}(uM)^n \mathrm{d}\xi \mathrm{d}\eta \tag{2.21}$$

For each closed path p that repeats w times, it is counted for $\frac{2n}{w}$ times, and its weight is

$$W_p(u) = (-1)^{w+1} W_{p_0}^w(u) \tag{2.22}$$

where p_0 is the corresponding non-periodic path. Therefore we have

$$\begin{aligned}\sum_{n=1}^{\infty} \frac{1}{2n} \sum_{p(n)} W_p(u) &= \sum_{[p]} \left(\sum_{w=1}^{\infty} \frac{(-1)^{w+1}}{w} W_p^w(u) \right) \\ &= \sum_{[p]} \ln(1 + W_p(u)) \\ &= \ln \prod_{[p]} (1 + W_p(u))\end{aligned} \tag{2.23}$$

When $|u| < \frac{1}{4}$, $\operatorname{tr}(uM) < 1$. By uniformly convergence and by substituting (2.21) we get

$$\begin{aligned}\ln \prod_{[p]} (1 + W_p(u)) &= \frac{L^2}{2(2\pi)^2} \int_0^{2\pi} \int_0^{2\pi} \sum_{n=1}^{\infty} -\frac{\operatorname{tr}(uM)^n}{n} \mathrm{d}\xi \mathrm{d}\eta \\ &= \frac{L^2}{2(2\pi)^2} \int_0^{2\pi} \int_0^{2\pi} \operatorname{tr}(\ln(I - uM)) \mathrm{d}\xi \mathrm{d}\eta \\ &= \frac{L^2}{2(2\pi)^2} \int_0^{2\pi} \int_0^{2\pi} \ln \det(I - uM) \mathrm{d}\xi \mathrm{d}\eta\end{aligned} \tag{2.24}$$

What remains is to calculate is $\det(I - uM)$:

$$\begin{aligned}\det(I - uM) &= (u^2 + 1)^2 + 2u(u^2 - 1)(\cos\xi + \cos\eta) \\ &= \cosh^{-4} 2\beta (\cosh^2 2\beta - \sinh 2\beta (\cos\xi + \cos\eta))\end{aligned} \tag{2.25}$$

Combining (1.3,2.11,2.24), we finally get

$$\begin{aligned}-\frac{\psi}{k_B T} &= \lim_{L \to \infty} L^{-2} \ln Z_{L^2} \\ &= \lim_{L \to \infty} \left(\ln(2(\cosh 2\beta)^{\frac{2(L-1)}{L}}) + \frac{1}{2(2\pi)^2} \int_0^{2\pi} \int_0^{2\pi} \ln \frac{(\cosh^2 2\beta - \sinh 2\beta(\cos\xi + \cos\eta))}{\cosh^4 2\beta} \mathrm{d}\xi \mathrm{d}\eta \right) \\ &= \ln 2 + \frac{1}{2\pi^2} \int_0^{\pi} \int_0^{\pi} \ln(\cosh^2 2\beta - \sinh 2\beta(\cos\xi + \cos\eta)) \mathrm{d}\xi \mathrm{d}\eta\end{aligned} \tag{2.26}$$

and complete the proof. □

3 Thermodynamic properties of a large crystal

In this part,we will discuss the critical value of a large crystal, that is,the temperature(or other thermodynamic quantities) at the singularity,which is exactly the point of phase change.

To simplify the computation,we assume

$$\lambda = -\frac{\psi}{k_B T} = \frac{\text{free energy}}{k_B T} \tag{3.1}$$

We have

$$\begin{aligned} \lambda &= \ln 2 + \frac{1}{2\pi^2}\int_0^\pi \int_0^\pi \ln(\cosh^2 2\beta - \sinh 2\beta(\cos\xi + \cos\eta))\mathrm{d}\xi\mathrm{d}\eta) \\ &= \ln 2\cosh 2\beta + \frac{1}{2\pi^2}\int_0^\pi \int_0^\pi \ln(1 - \frac{\sinh 2\beta}{\cosh^2 2\beta}(\cos\xi + \cos\eta))\mathrm{d}\xi\mathrm{d}\eta) \\ &= \ln 2\cosh 2\beta + \frac{1}{2\pi^2}\int_0^\pi \int_0^\pi \ln(1 - k(\cos\xi + \cos\eta))\mathrm{d}\xi\mathrm{d}\eta) \\ &= \ln 2\cosh 2\beta - \sum_{n=1}^{\infty}\left(\frac{(2n)!}{(n)!}\right)^2 k^{2n} \end{aligned} \tag{3.2}$$

where $2k = \frac{\sinh 2\beta}{\cosh^2 2\beta}$,and the last equation comes from the power series of logarithm.The series converges for $|2k(\cos\xi + \cos\eta)| \leq |4k| < 1$.For $J > 0(k > 0)$, at $k = \frac{1}{4}$, that is, at the critical value $\beta = \beta_C$(or temperature $T_C = 2Jk_B^{-1}\ln^{-1}(\sqrt{2}+1)$) given by

$$\frac{1}{2} = 2k = \frac{\sinh 2\beta}{\cosh^2 2\beta} \tag{3.3}$$

it diverges.Similarly,for $J < 0(k < 0)$,it implies divergence at $k = -\frac{1}{4}, T_C = 2Jk_B^{-1}\ln^{-1}(\sqrt{2}-1)$

The internal energy

$$\begin{aligned} U &= k_B T^2 \frac{\partial}{\partial T}\left(-\frac{\psi}{k_B T}\right) \\ &= -J\coth(2\beta)\left[1 + (2\tanh^2 2\beta - 1)\frac{2}{\pi}F(4k)\right] \end{aligned} \tag{3.4}$$

where $F(x)$ is elliptic integral of the first kind

$$F(x) = \int_0^{\pi/2}(1 - x^2\sin^2\theta)^{-\frac{1}{2}}\mathrm{d}\theta \tag{3.5}$$

F is a elliptic function with a property that

$$F \to \ln[4(1-x^2)]^{-1/2} \text{ as } x \to 1^- \tag{3.6}$$

so it diverges at $4k = 1$.

The specific heat C is given by

$$\begin{aligned} C &= \frac{\partial U}{\partial T} \\ &= \frac{8k}{\pi}(\beta\coth 2\beta)^2[2F(4k) - 2E(4k) + (2\tanh^2 2\beta - 1)(\frac{\pi}{2} + (2\tanh^2 2\beta - 1)F(4k))] \end{aligned} \tag{3.7}$$

where E is the complete elliptic integral of the second kind,defined by

$$E(x) = \int_0^{\pi/2}(1 - x^2\sin^2\theta)^{\frac{1}{2}}\mathrm{d}\theta \tag{3.8}$$

so C divergent at the critical point.

References

[1] Onsager L. Crystal statistics. I. A two-dimensional model with an order-disorder transition[J]. Physical Review, 1944, 65(3-4): 117.

[2] Kac M, Ward J C. A combinatorial solution of the two-dimensional Ising model[J]. Physical Review, 1952, 88(6): 1332.

李政道-杨振宁零点理论简介

王骏澎

2022 年 6 月 14 日

1　引言: 统计物理简介

统计物理 (statistical physics) 是研究由大量微观粒子组成的宏观系统的学科，目前理论较为完善的是平衡态[1]统计物理。像这样一个极其复杂 (粒子数量在 10^{23} 左右的数量级) 的系统，要直接进行具体的力学分析是几乎不可能的。于是，人们开始考虑引入新的假设，从另一个角度对这样的系统进行研究。

统计物理的两条基本假设

遍历假设 (ergodic hypothesis): 对于一个孤立系统[2]，在经过足够长的时间后，系统会经历所有可能的微观状态。

等概率假设 (equal-probability hypothesis): 对处在平衡态的系统，它的每个微观状态的出现概率[3]相同。

在统计物理发展的早期，曾有不少人试图用纯力学手段证明这两条假设，但大部分时候只能在一些极度简化的系统中证明正确性，同时还发现了一些这两条假设不成立的系统。但如果不考虑这两条假设的正确性，直接在它们的基础上进行推导，却可以得到十分丰富的结果；至于这些结果的正确性，则可以直接通过实验来进行验证，而无需纯力学的推导。

系综理论的一些基本概念

系综 (ensemble) 指的是大量具有相同宏观性质的力学系统的集合。在测量一个处于平衡态的宏观系统的某个宏观物理量时，测量通常会持续一段时间，而这段时间对于微观粒子来说，已经是足够之前说的两条基本假设成立的时间了，我们测量得到的物理量，实际上是这个宏观系统所对应的系综的平均值。统计物理中常用的系综有以下几种：

微正则系综 (microcanonical ensemble): 固定粒子数 N，系统总能量 E 和体积 V 的系统所对应的系综。

正则系综 (canonical ensemble): 固定粒子数 N，温度 T 和体积 V 的系统所对应的系综。

巨正则系综 (grand canonical ensemble): 固定化学势 μ，温度 T 和体积 V 的系统所对应的系综。

[1] 指系统不受外力作用，且宏观性质不随时间发生改变的状态。

[2] 与外界既没有物质交换，也没有能量交换的系统。

[3] 按照量子力学的观点，系统只能取一系列分立的状态，因此微观状态的总数会是一个有限的正整数。当然，它在通常的条件下会远大于粒子总数。

在微正则系综中，由于 N, E, V 固定，所以系统的总微观状态数 Ω 也是一个固定值。定义系统的**熵** (entropy)S 为：

$$S(N,E,V) = k_B \ln \Omega(N,E,V)$$

其中 $k_B \approx 1.38 \times 10^{-23} J/K$ 为 Boltzmann 常数。通过一些推导可以证明，这一定义与经典热力学中对熵的定义 $S = \dfrac{\partial Q}{\partial T}$ 是等价的，这就将系统的微观性质与宏观的热力学状态函数联系了起来。从熵出发，可以推导出系统的各种热力学性质，那么自然也可以从微观状态数出发来得到它们。

在正则系综与巨正则系综中，也有类似于微正则系综中的微观状态数这样的，可以导出各种热力学状态函数的微观统计量，即**配分函数** (partition function)$Z(N,T,V)$ 和**巨配分函数** (grand partition function)$\varXi(\mu,T,V)$。正则系综可以想象成要研究的系统与一个大热源接触，而系统与大热源共同构成一个孤立系统并达到热平衡。由于大热源很大，它在向系统传递能量时自身温度几乎不变，因此可以认为温度是固定值。不严格地说，配分函数就是系统和大热源共同构成的孤立系统的微观状态数。类似的，巨正则系综对应的系统可以想象成与一个大热源兼大粒子源接触，二者共同构成平衡态的孤立系统，巨配分函数就是总系统的微观状态数。关于它们的严格定义请读者自行查阅统计物理相关书籍。此外，这两个函数满足关系式

$$F(N,T,V) = -k_B T \ln Z(N,T,V)$$
$$J(\mu,T,V) = -k_B T \ln \varXi(\mu,T,V)$$

其中 F 为 **Helmholtz 自由能** (free energy)，J 为**巨热力势** (grand thermodynamic potential)，它们均为热力学状态函数。

2 相变理论与统计物理

物质中物理、化学性质完全相同，并与其他部分有明显分界线的均匀部分称为一个**相** (phase)。而物质由一个相转变为另一个相的过程称为**相变** (phase transition)，比如常见的水结冰或变为水蒸气的过程就是相变。在相变过程中物质的一些状态函数会发生改变，有些是连续改变的，而有些则不连续。人们根据系统的热力学势函数[4]及其各阶偏导数在相变时的表现对其分了类：热力学势连续而其一阶偏导数不连续的称为一级相变，一阶偏导数连续而二阶偏导数不连续的称为二级相变，依此类推。

现在来考虑一个具体的系统。在一个体积为 V 的容器中，充有某种单原子分子气体，分子数量为 N，此时系统总的分子势能为

$$U = \sum_{1 \leqslant i < j \leqslant N} u(r_{ij})$$

其中 r_{ij} 是第 i 个与第 j 个分子之间的距离，$u(r)$ 表示两个距离为 r 的分子之间的势能。我们假设 u 有以下性质：存在 $b \geqslant a > 0$，使得当 $r \leqslant a$ 时，$u(r) = +\infty$；当 $r > b$ 时，$u(r) = 0$；当 $a < r \leqslant b$ 时，$-\infty < u(r) < +\infty$。在通常的温度和压强下，分子间作用势通常都有与此相近的性质，因此这样的近似是合理的。现在允许这个容器与一个装有同种单原子分子气体的，固定温度 T 和化学势 μ 的大热源兼大粒子源交换能量和物质，那么这是

[4] 比如前面说到的三种系综，它们对应的热力学势函数分别为熵、Helmholtz 自由能和巨热力势

一个对应于巨正则系综的系统。根据统计物理的结论，它的巨配分函数为

$$\Xi(\mu, V, T) = \sum_{N=0}^{M} \frac{Q_N}{N!} y^N$$

其中

$$Q_N = \int \cdots \int_V e^{-\frac{U}{k_B T}} \mathrm{d}\tau_1 \mathrm{d}\tau_2 \ldots \tau_N$$
$$y = \left(\frac{2\pi m k_B T}{h^2} \right)^{\frac{3}{2}} e^{\frac{\mu}{k_B T}}$$

τ_i 是第 i 个分子的坐标，显然 U 是它们的函数，Q_N 实际上就是这 N 个分子所对应的正则系综的配分函数；而 y 被称为系统的**逸度** (fugacity)，其中 $h \approx 6.626 \times 10^{-34} J \cdot s$ 为 Planck 常数，M 是容器 V 能够容纳的最大粒子数。可以看出，Ξ 对于三个自变量都是解析的。由于巨配分函数与巨热力势之间的联系，巨热力势及其各阶偏导数对这三个变量也应该是解析的，那自然不可能有不连续点出现，也就是说，按照统计物理的观点，这种气体似乎是不可能发生相变的。但事实上，现实中单原子分子气体的相变并非不可发生，统计物理的理论在这里与实验产生了冲突。

那么如何用统计物理来解释相变呢？在 20 世纪 50 年代以前，这个问题一直处在争论之中。而 1952 年杨振宁与李政道二人的两篇文章则给出了一个严格的解释。

3 体积趋于无穷时的极限

通常情况下，实际的容器相对于分子来说都是非常巨大的；而即使是解析函数，在取极限的情况下也经常出现奇异性[5]，因此李、杨二人对巨配分函数及其他相关物理量在体积趋于无穷时的行为进行了研究。

对于上一部分给出的系统，气体的压强与密度可以表示为如下形式：

$$\frac{p}{k_B T} = \frac{\ln \Xi}{V}$$
$$\rho = \frac{\partial}{\partial \ln y} \frac{\ln \Xi}{V}$$

要研究极限行为，首先当然要证明极限存在。李、杨二人首先证明了以下结果：

定理 3.1 在固定的温度下，对于正方体容器[6]和取正实值的 y，$\lim\limits_{V \to \infty} V^{-1} \ln \Xi$ 存在，且为 y 的单调递增连续函数。

在证明这条定理前，先来证明两个引理。

引理 3.2 设 V 和 W 是两个正方体容器的体积，它们的边长分别为 L 和 $L + b$(b 是分子间存在相互作用的最大距离)，则有

$$\lim_{L \to \infty} \frac{\ln \Xi(W, y) - \ln \Xi(V, y)}{W} = 0$$

[5] 指连续性、可导性变差

[6] 原文中写的是只要容器表面积是 $O(V^{\frac{2}{3}})$ 的即可，且极限值与容器形状无关，但用他们的方法似乎并不能在这一条件下证明结论。

证明 首先，将容器 V 完全置于 W 内部。考虑巨配分函数的物理意义，可将 $\Xi(W,y)$ 写成 $A_0+A_1+\cdots+A_m$ 的形式，其中 A_i 是 $W\backslash V$ 中有 i 个分子这种情况对于巨配分函数的贡献。

容易看出 $A_0=\Xi(V,y)$，而对于 A_1，由于分子间作用范围有限，且分子有体积 (直径为 a)，所以存在一个正整数 M，每个分子至多只和 M 个分子之间存在相互作用。那么对应的微观状态数也有上限，于是存在正数 β，使得 $A_1\leqslant\beta(W-V)\Xi(V,y)$。

记 $W-V=\Delta$，单个分子的体积为 α，由于分子之间是全同的，应有

$$A_i\leqslant\beta^i\frac{\Delta(\Delta-\alpha)\cdots(\Delta-(i-1)\alpha)}{i!}\Xi(V,y)$$

利用广义二项式定理，求和，可得 $\Xi(W,y)\leqslant(1+\beta\alpha)^{\frac{\Delta}{\alpha}}\Xi(V,y)$，那么就有

$$0<\ln\Xi(W,y)-\ln\Xi(V,y)\leqslant\frac{\Delta}{\alpha}\ln(1+\beta\alpha)$$

由于 $\Delta\sim L^2$，而 $W\sim L^3$，所以引理中的极限为 0，证毕。 □

引理 3.3 固定 $L>0$，则对于边长为 2^iL 的正方体容器 W_i，则极限

$$\lim_{i\to\infty}\frac{\ln\Xi(W_i,y)}{W_i}=K(y)$$

存在。

证明 设 $j>i>0$，将 W_j 分割成 8^{j-i} 个 W_i，显然跨 W_i 边界发生作用的分子数量应该至多正比于边界总面积，即 $3\cdot2^{j-i}\cdot(2^jL)^2$，那么可设这样的分子数不超过 $8^j2^{-i}L^2\gamma$，γ 为常数，则应存在正数 β，使得

$$\Xi(W_j,y)\leqslant\Xi^{8^{j-i}}(W_i,y)\cdot\beta^{8^j2^{-i}L^2\gamma}$$

或

$$\ln\Xi(W_j,y)\leqslant8^{j-i}\ln\Xi(W_i,y)+8^j2^{-i}L^2\gamma\ln\beta$$

现在在每个 W_i 中以合适的方式放入边长为 2^iL-b 的正方体 V_i，可以保证这 8^{j-i} 个 V_i 中的分子只和自己内部的相互作用，则有

$$8^{j-i}\ln\Xi(V_i,y)\leqslant\ln\Xi(W_j,y)$$

那么由 $W_j=8^{j-i}W_i$ 就有

$$\frac{\ln\Xi(V_i,y)}{W_i}\leqslant\frac{\ln\Xi(W_j,y)}{W_j}\leqslant\frac{\ln\Xi(W_i,y)}{W_i}+\frac{\gamma\ln\beta}{2^iL}$$

令 $j\to\infty$，可得

$$\frac{\ln\Xi(V_i,y)}{W_i}\leqslant\underline{\lim}_{j\to\infty}\frac{\ln\Xi(W_i,y)}{W_i}\leqslant\overline{\lim}_{j\to\infty}\frac{\ln\Xi(W_i,y)}{W_i}\leqslant\frac{\ln\Xi(W_i,y)}{W_i}+\frac{\gamma\ln\beta}{2^iL}$$

由于 $\lim\limits_{i\to\infty}\frac{\gamma\ln\beta}{2^iL}=0$，再利用引理 3.2的结论，即可知此引理中的极限存在。 □

定理的证明 沿用引理 3.3中记号，固定 y 并任给 $\varepsilon>0$，存在一个足够大的正方体 W_i 使得

$$\left|\frac{\ln\Xi(W_i,y)}{W_i}-K(y)\right|<\varepsilon$$

任给一个边长不小于 2^iL 的正方体 Ω_0，则应存在正整数 n_0，使得其边长在 n_02^iL 到 $(n_0+1)2^iL$ 之间。记边长为 n_02^iL 的正方体为 Ω_1，边长 $(n_0+1)2^iL$ 的正方体为 Ω_2，边长为 2^iL-b 的正方体为 V。由引理 3.2结论，不妨设 $\left|\dfrac{\ln\Xi(W_i,y)-\ln\Xi(V_i,y)}{W_i}\right|<\varepsilon$。

仿照引理 3.3中的讨论，有

$$\frac{\ln\Xi(V_i,y)}{W_i}\leqslant\frac{\ln\Xi(\Omega_j,y)}{\Omega_j}\leqslant\frac{\ln\Xi(W_i,y)}{W_i}+\frac{\gamma\ln\beta}{2^iL}$$

其中 $j=1,2$。如果 i 取得足够大，满足 $\dfrac{\gamma\ln\beta}{2^iL}<\varepsilon$，则由此可得

$$\left|\frac{\ln\Xi(\Omega_j,y)}{\Omega_j}-K(y)\right|<2\varepsilon$$

最后，不妨设 $\varepsilon<1$，如果 n_0 也取得足够大，满足

$$\left(1+\frac{1}{n_0}\right)^3-1<1-\left(1-\frac{1}{n_0}\right)^3<\varepsilon$$

，那么我们有

$$\begin{aligned}\left|\frac{\ln\Xi(\Omega_1,y)}{\Omega_2}-K(y)\right|&\leqslant\frac{\Omega_1}{\Omega_2}\left|\frac{\ln\Xi(\Omega_1,y)}{\Omega_1}-K(y)\right|+\left(1-\frac{\Omega_1}{\Omega_2}\right)|K(y)|\\&\leqslant(|K(y)|+2)\varepsilon\end{aligned}$$

和

$$\begin{aligned}\left|\frac{\ln\Xi(\Omega_2,y)}{\Omega_1}-K(y)\right|&\leqslant\frac{\Omega_2}{\Omega_1}\left|\frac{\ln\Xi(\Omega_2,y)}{\Omega_2}-K(y)\right|+\left(\frac{\Omega_2}{\Omega_1}-1\right)|K(y)|\\&\leqslant(|K(y)|+2)\varepsilon\end{aligned}$$

注意到

$$\frac{\ln\Xi(\Omega_1,y)}{\Omega_2}\leqslant\frac{\ln\Xi(\Omega_0,y)}{\Omega_0}\leqslant\frac{\ln\Xi(\Omega_1,y)}{\Omega_2}$$

可知

$$\left|\frac{\ln\Xi(\Omega_0,y)}{\Omega_0}-K(y)\right|<(|K(y)|+2)\varepsilon$$

由于 $|K(y)|$ 是一个确定的有限值，再由 Ω_0 和 ε 的任意性即可知，对于正方体容器 V，有

$$\lim_{V\to\infty}\frac{\ln\Xi(V,y)}{V}=K(y)$$

至此，定理 3.1证毕[7]。 □

现在我们只是知道了 $\dfrac{\ln\Xi(V,y)}{V}$ 在 $V\to\infty$ 时的极限存在，并不清楚其性质，因此接下来对其性质进行研究。首先注意到在 V 有限时，Ξ 是 y 的多项式，且系数均为正。设它的 M 个零点分别为 $y_1,y_2,\ldots,y_M$，则有

$$\Xi(V,y)=\prod_{i=1}^{M}\left(1-\frac{y}{y_i(V)}\right)$$

[7] 当然，对于其他形状的容器，如球体等，我们应该也可以通过将其分割为正方体来近似的手段来证明它们收敛到同一极限，但表面积是 $O(V^{\frac{2}{3}})$ 的这一条件是否足够，笔者暂时无法判断。

显然这 M 个零点均不为非负实数，且对称分布在实轴两侧 (因为实系数多项式的复根总是成对出现)。假设存在一个以 $\eta\in\mathbb{C}$ 为圆心，$\sigma>0$ 为半径的圆盘 C，对于充分大的 V，均有 $y_i(V)\notin R$，那么记 $z=y-\eta, z_i=y_i-\eta$，则有

$$\varXi(V,y)=\prod_{i=1}^{M}\left(1-\frac{z}{z_i(V)}\right)\left(\frac{z_i(V)}{y_i(V)}\right)$$

于是

$$\frac{\ln\varXi(V,y)}{V}=\frac{1}{V}\left(\sum_{i=1}^{M}\ln\left(1-\frac{z}{z_i(V)}\right)+\sum_{i=1}^{M}\ln\frac{z_i(V)}{y_i(V)}\right)$$

在圆盘 C 内恒有 $\left|\frac{z}{z_i(V)}\right|<1$，于是可将所有 ln 展开成幂级数，得

$$\frac{\ln\varXi(V,y)}{V}=\sum_{n=0}^{\infty}b_n(V)z^n$$

其中

$$b_0(V)=\frac{1}{V}\sum_{i=1}^{M}\ln\frac{z_i(V)}{y_i(V)}$$

而对于 $n\geqslant 1$，则有

$$b_n(V)=-\frac{1}{nV}\sum_{i=1}^{M}\frac{1}{z_i^n(V)}$$

在圆盘 C 内恒有 $|z_i(V)|\geqslant\sigma$，那么可以得到对 $n\geqslant 1$，均有

$$|b_n(V)|\leqslant\frac{M}{nV\sigma^n}$$

注意到 M/V 是有上界的，它不超过分子体积的倒数，所以我们有

$$\varlimsup_{n\to\infty}|b_n(V)|^{\frac{1}{n}}\leqslant\frac{1}{\sigma}$$

由 Cauchy-Hadamard 公式，对于任意的足够大的 V，$\frac{\ln\varXi(V,y)}{V}$ 对于 y 都是 C 中的解析函数。如果代入 $z=0$，则有

$$\frac{\ln\varXi(V,\eta)}{V}=b_0(V)$$

根据之前的结论，$\lim\limits_{V\to\infty}b_0(V)$ 存在，记此极限为 $b_0(\infty)$。类似的，通过求 $\frac{\ln\varXi(V,y)}{V}$ 对 y 的各阶偏导数，容易证明对所有自然数 n，$b_n(\infty)$ 都是存在的。当然，由极限的性质易知它们满足

$$|b_n(\infty)|\leqslant\frac{1}{n\sigma^n}\sup_{V>0}\frac{M}{V}(n\geqslant 1)$$

所以幂级数 $\sum_{n=0}^{\infty}b_n(\infty)z^n$ 也在 C 内收敛。再由幂级数的内闭绝对一致收敛性知，求极限的顺序可交换，即

$$\sum_{n=0}^{\infty}b_n(\infty)z^n=\lim_{V\to\infty}\sum_{n=0}^{\infty}b_n(V)z^n=\lim_{V\to\infty}\frac{\ln\varXi(V,y)}{V}$$

也就是说，$\lim\limits_{V\to\infty}\frac{\ln\varXi(V,y)}{V}$ 在圆盘 C 内是 y 的解析函数。同时由 ln 的解析性，可知

$$\frac{\partial}{\partial\ln y}\frac{\ln\varXi(V,y)}{V},\frac{\partial^2}{(\partial\ln y)^2}\frac{\ln\varXi(V,y)}{V},\frac{\partial^3}{(\partial\ln y)^3}\frac{\ln\varXi(V,y)}{V},\ldots$$

在 $V\to\infty$ 时的极限也都存在，且也都是 y 的解析函数。此结论很容易推广到任意开集，于是我们得出结论

定理 3.4 设 R 是复平面上一个开区域，且对于足够大的 V，$\Xi(V,y)$ 在 R 中恒不为 0，则当 $V\to\infty$ 时

$$\frac{\ln\Xi(V,y)}{V},\ \frac{\partial}{\partial\ln y}\frac{\ln\Xi(V,y)}{V},\ \frac{\partial^2}{(\partial\ln y)^2}\frac{\ln\Xi(V,y)}{V},\dots$$

的极限均存在且均为 y 的解析函数。

由此定理以及密度和压强的表达式，我们可以知道，在 $V\to\infty$ 时

$$\rho=\frac{\partial}{\partial\ln y}\frac{p}{k_BT}$$

也是成立的。

4 对相变的解释

定理 3.4告诉我们，对于复平面上始终不包含巨配分函数零点的区域，压强、密度等物理量都是逸度的解析函数。实际情况下，逸度只能取正实数，那么我们可以分两种情况来考虑：

(1) 若存在一个包含整个正实轴的区域 R，使得对于足够大的 V 和 $y\in R$，均有 $\Xi(V,y)\neq 0$，则 $V\to\infty$ 时，p,ρ 在实轴上仍是 y 的解析函数，此时不会有相变发生。事实上，此时 p 和 ρ 都是 y 的单调递增函数，p 的单调性是显然的，因为 Ξ 就是 y 的单调递增函数同时二人也给出了一个 ρ 的单调性的简单证明：

要证明 ρ 是 y 的单调递增函数，只要证明对于有限的 V，有

$$\frac{\partial^2\ln\Xi(V,y)}{(\partial\ln y)^2}>0$$

成立即可。记 $t=\ln y$，则此时有

$$\Xi=\sum_{i=0}^{M}a_ie^{it}$$

直接计算得

$$\frac{\partial\ln\Xi}{\partial t}=\sum_{i=0}^{M}\frac{ia_ie^{it}}{\Xi}$$

$$\frac{\partial^2\ln\Xi}{\partial t^2}=\sum_{i=0}^{M}\frac{i^2a_ie^{it}}{\Xi}-\left(\sum_{i=0}^{M}\frac{ia_ie^{it}}{\Xi}\right)^2$$

考虑一个取值为 $0,1,\dots,M$ 的随机变量，它取值为 i 的概率为 $\frac{a_ie^{it}}{\Xi}$，则 $\frac{\partial\ln\Xi}{\partial t}$ 是其数学期望，而 $\frac{\partial^2\ln\Xi}{\partial t^2}$ 是其方差，自然大于 0。所以 ρ 在 $V\to\infty$ 时也是 y 的单调递增函数。

(2) 若不存在 (1) 中所说的区域，即无论 V_0 取多大，对于 $V>V_0$，$\Xi(V,y)$ 对于 y 的零点总是在正实轴上有聚点，则此时，在聚点附近，定理 3.4就可能不成立，那么 $\lim\limits_{V\to\infty}\frac{\ln\Xi}{V}$ 对 $\ln y$ 的各阶导数就可能出现不连续性。如果一阶导数，也就是密度，出现了突变，那么就是一级相变，日常生活中会遇到的相变基本都是这种相变；如果密度连续而 $\frac{\mathrm{d}^2}{(\mathrm{d}\ln y)^2}\lim\limits_{V\to\infty}\frac{\ln\Xi}{V}$ 不连续，则为二级相变，如两种液氦之间的转变。当然，依此类推，还可以有三级、四级相变等，但对这些人们很少研究。

由此可以看出，定理 3.4成功将相变问题与“巨配分函数在复平面上的零点分布”以一种非常简单的方式联系了起来，这便是他们所做的开创性的工作。事实上，在李、杨之前，并非没有人尝试用统计物理的理论去解释相变只是一直没能出现合理且令人信服的理论。比如 Mayer 及其合作者在 1937、1938 年的一系列文章中就在尝试用统计物理解释相变，也对李、杨产生了一些启发。但 Mayer 为了估算配分函数，引入了大量数学近似，其理论准确性难以评估，也在解释液体的状态方程时出现了问题。在李、杨二人的第一篇文章的最后一部分，就将他们的理论与 Mayer 理论进行了对比，并指出了其错误。

5 计算实例

李、杨二人的相变理论的思路很容易理解，但要验证其正确性却并不容易，因为这涉及到巨配分函数的零点分布问题，而实际上巨配分函数的具体形式通常难以写出，计算其零点分布并求出极限更是难上加难。但对于一些特定的较为简化的模型，计算巨配分函数并讨论其零点分布却是有可能的。在李、杨二人随后发表的第二篇文章中，就对于“格气” (lattice gas) 模型进行了研究，给出了其巨配分函数的表达式，并零点分布进行了研究。

首先来简要介绍格气模型。假设有一个共有 M 个格点的正方形点阵，在格点中填充有分子，分子之间存在相互作用力，且填充在同一格点中的分子之间的势能为 $+\infty$，这样的物理模型叫格气模型。李、杨二人为了研究其巨配分函数，将其等价映射到了外加磁场的 Ising 模型上，至于是如何映射的，本文不作介绍，有兴趣的读者请自行查阅原文。总之，根据映射的结果，他们给出了以下巨配分函数的表达式：

$$\varXi = e^{\frac{MH}{k_BT}} \sum_{n=0}^{M} P_n z^n$$

其中 z 是一个与逸度 y 成正比的量，H 则是一个与映射到的 Ising 模型的外加磁场有关的量。记 $\Gamma = \{1, 2, \ldots, M\}$，则

$$P_n = \sum \prod_{i\in A} \prod_{j\in\Gamma\setminus A} a_{ij}$$

其中 A 是 Γ 的一个有 n 个元素的子集，求和取遍 A 的所有取法 (共 $\begin{pmatrix} M \\ n \end{pmatrix}$ 项)。$a_{ij} = e^{\frac{u_{ij}}{k_BT}}$，$u_{ij}$ 是第 i 个格点和第 j 个格点中的分子之间的势能，显然 $u_{ij} = u_{ji}$，$P_0 = P_M = 1$。

李、杨二人给出了以下结论：

定理 5.1 在格气模型中，若对于 $i \neq j$，均有 $u_{ij} \leqslant 0$，则巨配分函数作为 z 的函数，其零点均在复平面中的单位圆上。

这个定理的条件相当宽泛，它对于点阵的大小、形状、维数，以及分子间作用的范围等都没有限制。由于 $e^{\frac{MH}{k_BT}} \neq 0$，只要证明与它相乘的那个 z 的多项式的零点均在单位圆上即可。我们有如下更为一般的定理：

定理 5.2 (李政道-杨振宁单位圆定理) 设 $M \geqslant 2$，$a_{ij} \in [-1, 1]$，$1 \leqslant i, j \leqslant M$ 且 $i \neq j, a_{ij} = a_{ji}$。记 $\Gamma_M = \{1, 2, \ldots, M\}$，$I$ 为 Γ_M 的任意子集，记其中元素个数为 $|I|$，令 $P_n = \sum\limits_{|I|=n} \prod\limits_{i\in I} \prod_{j\in\Gamma_M\setminus I} a_{ij} (n = 0, 1, 2, \ldots, M)$，则多项式

$$\mathcal{P}(z) = \sum_{n=0}^{M} P_n z^n$$

的零点均在复平面中的单位圆上。

对于此定理中的多项式，我们注意到 $P_n = P_{M-n}$，于是当 $z \neq 0$ 时，我们有

$$\mathcal{P}(1/z) = \sum_{n=0}^{M} P_n \frac{1}{z^n} = \frac{1}{z^M} \sum_{n=0}^{M} P_n z^n = \frac{\mathcal{P}(z)}{z^n}$$

那么当 z_0 是 $\mathcal{P}$ 的零点时，$1/z$ 也是其零点，因此我们只需证明 $\mathcal{P}$ 的零点都在单位圆盘内即可 (0 显然不是零点)。

为了证明此结论，李、杨二人引入了如下的 M 元多项式：

$$\mathfrak{P}(z_1, z_2, \ldots, z_M) = \sum_{I \subset \Gamma_M} \left(\prod_{i \in I} z_i \cdot \prod_{\substack{i \in I \\ j \in \Gamma_M \setminus I}} a_{ij} \right)$$

容易看出，此式满足

$$\mathfrak{P}(z, z, \ldots, z) = \mathcal{P}(z)$$
$$z_1 z_2 \ldots z_M \mathfrak{P}(z_1^{-1}, z_2^{-1}, \ldots, z_M^{-1}) = \mathfrak{P}(z_1, z_2, \ldots, z_M)$$

为了证明定理，李、杨二人通过归纳法证明了以下引理：

引理 5.3 假设所有 a_{ij} 均不为 ± 1 或 0，$\mathfrak{P}(z_1, z_2, \ldots, z_M) = 0$ 且 $|z_n| \geqslant 1, n = 1, 2, \ldots, M$，则 $|z_n| = 1$。

证明 记 $M = m$ 时的多项式 $\mathfrak{P}$ 为 $\mathfrak{P}_m$，此结论对 $\mathfrak{P}_1$ 显然成立，因为 $\mathfrak{P}_1 = 1 + z_1$。对于 $M = 2$，有

$$\mathfrak{P}_2(z_1, z_2) = 1 + a_{12}(z_1 + z_2) + z_1 z_2$$

当 $\mathfrak{P}_2 = 0$ 时，有

$$z_1 = -\frac{1 + a_{12} z_2}{z_2 + a_{12}}$$

由于 $|a_{12}| < 1$，由此可得当 $|z_2| > 1$ 时，必有 $|z_1| < 1$，那么要同时满足 $|z_1| \geqslant 1$ 和 $|z_2| \geqslant 1$，就必有 $|z_1| = |z_2| = 1$。

现在进行归纳。假设引理的结论对于 $M-1, M-2 (M \geqslant 3)$ 成立，但对 M 不成立，则存在一组 $(\alpha_1, \alpha_2, \ldots, \alpha_M)$，使得 $|\alpha_n| \geqslant 1 (n = 1, 2, \ldots, M), \mathfrak{P}_M(\alpha_1, \alpha_2, \ldots, \alpha_M) = 0$，且存在一个 n_0 使得 $|\alpha_{n_0}| > 1$。通过交换所有的 $a_{n_0 j}$ 和 a_{Mj}，我们总可以认为这个 n_0 就是 M。

现在固定 $(z_1, z_2, \ldots, z_{M-2})$ 为 $(\alpha_1, \alpha_2, \ldots, \alpha_{M-2})$，令 $\mathfrak{P}_M(\alpha_1, \ldots, \alpha_{M-2}, z_{M-1}, z_M) = 0$，将 z_{M-1} 看作 z_M 的函数 $z_{M-1}(z_M)$，则有 $z_{M-1}(\alpha_M) = \alpha_{M-1}$。

将 $\mathfrak{P}_M(\alpha_1, \ldots, \alpha_{M-2}, z_{M-1}, z_M)$ 写成 $A + Bz_{M-1} + Cz_M + Dz_{M-1}z_M$ 的形式，其中 A, B, C, D 均与 z_{M-1}, z_M 无关，则有

$$z_{M-1} = -\frac{A + Cz_M}{B + Dz_M}$$

若 $D \neq 0$，则存在极限

$$\mathfrak{z}_{M-1} \triangleq \lim_{|z_M| \to \infty} z_{M-1} = -\frac{C}{D}$$

接下来我们要证明 $D \neq 0, |\mathfrak{z}_{M-1}| < 1$。为此，先计算出 $C + Dz_{M-1}$，我们有

$$\begin{aligned}
\mathfrak{P}(z_1,z_2,\dots,z_M) &= \sum_{I\subset\Gamma_M}\left(\prod_{i\in I} z_i\cdot\prod_{\substack{i\in I\\ j\in\Gamma_M\setminus I}} a_{ij}\right)\\
&= \sum_{\substack{I\subset\Gamma_M\\ M\in I}}\left(\prod_{i\in I} z_i\cdot\prod_{\substack{i\in I\\ j\in\Gamma_M\setminus I}} a_{ij}\right)+\sum_{\substack{I\subset\Gamma_M\\ M\notin I}}\left(\prod_{i\in I} z_i\cdot\prod_{\substack{i\in I\\ j\in\Gamma_M\setminus I}} a_{ij}\right)\\
&= \sum_{I\subset\Gamma_{M-1}}\left(z_M\prod_{i\in I} z_i\cdot\prod_{\substack{i\in I\\ j\in\Gamma_{M-1}\setminus I}} a_{ij}\cdot\prod_{j\in\Gamma_{M-1}\setminus I} a_{Mj}+\prod_{i\in I} z_i\cdot\prod_{\substack{i\in I\\ j\in\Gamma_M\setminus I}} a_{ij}\right)\\
&= \sum_{I\subset\Gamma_{M-1}}\left(z_1z_2\dots z_M\prod_{j\in\Gamma_{M-1}\setminus I}\frac{a_{Mj}}{z_j}\cdot\prod_{\substack{i\in I\\ j\in\Gamma_{M-1}\setminus I}} a_{ij}+\prod_{i\in I} z_i a_{i,M}\cdot\prod_{\substack{i\in I\\ j\in\Gamma_M\setminus I}} a_{ij}\right)\\
&= z_1z_2\dots z_M\mathfrak{P}_{M-1}\left(\frac{a_{1M}}{z_1},\frac{a_{2M}}{z_2},\dots,\frac{a_{M-1,M}}{z_{M-1}}\right)\\
&\quad+\mathfrak{P}_{M-1}(z_1a_{1M},z_2a_{2M},\dots,z_{M-1}a_{M-1,M})
\end{aligned}$$

所以有

$$\begin{aligned}
C+Dz_{M-1} &= \alpha_1\alpha_2\dots\alpha_{M-2}z_{M-1}\mathfrak{P}_{M-1}\left(\frac{a_{1M}}{\alpha_1},\dots,\frac{a_{M-2,M}}{\alpha_{M-2}},\frac{a_{M-1,M}}{z_{M-1}}\right)\\
&= a_{1M}a_{2M}\dots a_{M-1,M}\mathfrak{P}_{M-1}\left(\frac{\alpha_1}{a_{1M}},\dots,\frac{\alpha_{M-2}}{a_{M-2,M}},\frac{z_{M-1}}{a_{M-1,M}}\right)\\
&= \alpha_1\alpha_2\dots\alpha_{M-2}z_{M-1}\mathfrak{P}_{M-2}\left(\frac{a_{1,M-1}a_{1M}}{\alpha_1},\frac{a_{2,M-1}a_{2M}}{\alpha_2},\dots,\frac{a_{M-2,M-1}a_{M-2,M}}{\alpha_{M-2}}\right)\\
&\quad+a_{1M}a_{2M}\dots a_{M-1,M}\mathfrak{P}_{M-2}\left(\frac{a_{1,M-1}}{a_{1M}}\alpha_1,\frac{a_{2,M-1}}{a_{2M}}\alpha_2,\dots,\frac{a_{M-2,M-1}}{a_{M-2,M}}\alpha_{M-2}\right)
\end{aligned}$$

由此可以看出 D 的表达式，而 $D=0$ 等价于

$$\mathfrak{P}_{M-2}\left(\frac{\alpha_1}{a_{1,M-1}a_{1M}},\frac{\alpha_2}{a_{2,M-1}a_{2M}},\dots,\frac{\alpha_{M-2}}{a_{M-2,M-1}a_{M-2,M}}\right)=0$$

由于 $|\alpha_i|\geqslant 1, |a_{ij}|<1$，所以 $\left|\dfrac{\alpha_i}{a_{i,M-1}a_{i,M}}\right|>1$，这与归纳假设中，引理的结论对 $M-2$ 成立矛盾，所以必有 $D\neq 0$，那么 $\mathfrak{z}_{M-1}$ 存在且满足 $C+D\mathfrak{z}_{M-1}=0$，这等价于

$$\mathfrak{P}_{M-1}\left(\frac{\alpha_1}{a_{1M}},\dots,\frac{\alpha_{M-2}}{a_{M-2,M}},\frac{\mathfrak{z}_{M-1}}{a_{M-1,M}}\right)=0$$

利用 $|\alpha_i|\geqslant 1, |a_{ij}|<1$，以及归纳假设中结论对 $M-1$ 成立，可知 $\left|\dfrac{\mathfrak{z}_{M-1}}{a_{M-1,M}}\right|<1$，所以 $|\mathfrak{z}_{M-1}|<1$。那么由 $z_{M-1}(z_M)$ 的连续性，应存在一个 α'_M，满足 $|\alpha'_M|>1$，且 $\alpha'_{M-1}\triangleq z_{M-1}(\alpha'_M)$ 满足 $|\alpha'_{M-1}|$。再对 $z_1,z_2,\dots,z_{M-2}$ 重复这样的讨论，可以得到一组 $(\beta_1,\beta_2,\dots,\beta_M)$，满足 $\mathfrak{P}_M(\beta_1,\beta_2,\dots,\beta_M)=0, |\beta_1|=|\beta_2|=\cdots=|\beta_{M-1}|=1, |\beta_M|>1$。

最后,将 $\mathfrak{P}_M(\beta_1,\beta_2,\dots,\beta_M)$ 写成 β_M 的线性函数的形式:$\mathfrak{P}_M(\beta_1,\beta_2,\dots,\beta_M)=A'\beta_M+B'=0$，根据之前的结论，$A'=0$ 等价于

$$\mathfrak{P}_{M-1}\left(\frac{\beta_1}{a_{1M}},\dots,\frac{\beta_{M-2}}{a_{M-2,M}},\frac{\beta_{M-1}}{a_{M-1,M}}\right)=0$$

由归纳假设，这不可能，所以 $A' \neq 0$，则满足 $A'\beta_M + B' = 0$ 的 β_M 应该是唯一的。另外，对于 $z \in \mathbb{C}, |z| = 1$，我们注意到 $z = \bar{z}^{-1}$，以及 $\mathfrak{P}_M$ 的系数均为实数，可得

$$\begin{aligned}\mathfrak{P}_M(\beta_1, \beta_2, \dots, \bar{\beta}_M^{-1}) &= \mathfrak{P}_M(\bar{\beta}_1^{-1}, \bar{\beta}_2^{-1}, \dots, \bar{\beta}_M^{-1}) \\ &= \overline{(\beta_1\beta_2 \dots \beta_M)^{-1}\mathfrak{P}_M(\beta_1, \beta_2, \dots, \beta_M)} = 0\end{aligned}$$

即 $A'\bar{\beta}_M^{-1} + B' = 0$，由于 $|\beta_M| > 1$ 所以 $\beta_M \neq \bar{\beta}_M^{-1}$，这与之前得出的唯一性矛盾。所以引理的结论对 M 也成立，至此，引理证毕。 □

引理 5.3结论结合 $\mathfrak{P}(z, z, \dots, z) = \mathcal{P}(z)$ 即可证得定理 5.2在 $a_{ij} \in (-1, 0) \cup (0, 1)$ 时的情况。要证明一般情况，我们还需要另一个引理：

引理 5.4 设 G 是复平面上一个连通的开集，$\{f_n\}$ 是一列解析函数。若 $\{f_n\}$ 在 G 的紧子集上一致收敛到解析函数 f，且 f_n 均在 G 上没有零点，则 f 恒为 0 或 f 也在 G 上没有零点。

证明 假设 f 不恒为 0，且存在零点，由解析函数的零点孤立性定理，存在 $z_0 \in G$ 和 $r > 0$，使得以 z_0 为圆心，r 为半径的圆盘 D 满足 $\overline{D} \subset G, f(z_0) = 0$，且 $\forall z \in D, z \neq z_0, f(z) \neq 0$。

由 f 的解析性，可以取 $\delta > 0$，使得 $\inf_{z\in\partial D} |f(z)| \geqslant \delta$，那么存在 $N \in \mathbb{N}_+$，使得当 $n \geqslant N$ 时，有 $\inf_{z\in\partial D} |f_n(z)| \geqslant \delta/2$。由 Weierstrass 定理，$f_n'$ 在 ∂D 上是一致收敛到 f' 的，则 f_n'/f_n 也是一致收敛到 f_n 的。设 f 在 z_0 处的零点为 m 重，则由辐角原理

$$2\pi m\mathrm{i} = \int_{\partial D} \frac{f'(z)}{f(z)}\mathrm{d}z = \lim_{n\to\infty} \int_{\partial D} \frac{f_n'(z)}{f_n(z)}\mathrm{d}z = 0$$

出现矛盾。所以若 f 不恒为 0，则在 G 上没有零点。 □

定理的证明 对于 a_{ij} 均不为 $\pm1, 0$ 的情况，由引理 5.3结论，以及 $\mathfrak{P}(z, z, \dots, z) = \mathcal{P}(z)$ 立即可得。若存在等于 ±1 或 0 的 a_{ij}，可选取一列 $\{a_{ij}^{(k)}\}$，使得它们均在 $(-1, 0) \cup (0, 1)$ 上，且 $\lim\limits_{k\to\infty} a_{ij}^{(k)} = a_{ij}$。记由 $a_{ij}^{(k)}, 1 \leqslant i, j \leqslant M$ 所生成的类似定理 5.2中那样的多项式为 $\mathcal{P}^{(k)}$，则容易看出，在紧集上，$\mathcal{P}^{(k)}$ 一致收敛于 $\mathcal{P}$，且它们均在区域 $|z| > 1$ 上没有零点。显然 $\mathcal{P}$ 不恒为零，于是由引理 5.4结论，$\mathcal{P}$ 也在 $|z| > 1$ 上没有零点，再由 $\mathcal{P}(z^{-1}) = z^{-M}\mathcal{P}(z)$ 知其零点均在单位圆上。 □

李、杨给出的这个结果是相当惊人的，巨配分函数的零点数量通常来讲是个非常巨大的数字，而这一结论告诉我们，格气模型中这么多的零点竟然全部分布在单位圆上。同时这个漂亮的结果也告诉我们，格气模型只可能存在单一相变点，因为此时实轴上至多只有一个点可能成为巨配分函数的零点的聚点。

除了单位圆定理，李、杨二人还在他们第二篇论文的第五部分对格气模型的巨配分函数进行了更多的计算，这里就不再进行更多的介绍了，请有兴趣的读者自行查阅。

6 参考文献

[1] C. N. Yang and T. D. Lee, *Statistical Theory of Equations of State and Phase Transitions. I. Theory of Condensation*, Phys. Rev., 87, 404 (1952)

[2] T. D. Lee and C. N. Yang, *Statistical Theory of Equations of State and Phase Transitions. II. Lattice Gas and Ising Model*, Phys. Rev., 87, 410 (1952)

[3] 刘川, 热力学与统计物理, 北京: 北京大学出版社, 2022

Applied Math

An Introduction to Quantum Computing

Yijia Fang; Zhongjin Yan

May 2022

Abstract

In this article, we are going to introduce the basic results of quantum computation and the most famous and important algorithms for this quantum computational model.

Contents

1 Introduction

In recent years, quantum computing has become a widely discussed topic. The significance of studying quantum computing is that it may be more powerful than classical models of computing, such as Turing machines. This may allow quantum computing to be exponentially faster than normal Turing machines. And, it's very important that it is physically possible for quantum computers to be realized. Next, we will show some basic results of quantum computing step-by-step.

2 Quantum Superposition And Qubits

Ordinary computers operate with states built from a finite number of bits. Each bit has two states, 0 or 1. For quantum computers, the analogical objects are called qubits. Like bits, each qubit has two states denoted by 0 or 1 as well. However, the special feature of a qubit is that it can be in both two basic states, denoted by 0 and 1, at the same time, while the normal bit can only be one of these two states. As is customary in physics, we use Dirac notation, $|0\rangle$ and $|1\rangle$, to denote the basic states. And a qubit is allowed to be in any state of a vector on the unit ball of $\mathbb{C}^2$, written as

$$\alpha_0|0\rangle + \alpha_1|1\rangle,\ \text{where } \alpha_0, \alpha_1 \in \mathbb{C},\ |\alpha_0|^2 + |\alpha_1|^2 = 1.$$

A state of this form is called a *superposition* of the basic states, and here α_0, α_1 are called *amplitudes.* We can find that $|0\rangle$, $|1\rangle$ are actually the basis vectors of $\mathbb{C}^2$. According to the principle of quantum mechanics, when isolated from outside, a qubit can stay in the superposition, until it's measured. When measuring a qubit, the amplitude wave collapses, and we will get $|0\rangle$ with probability $|\alpha_0|^2$, or $|1\rangle$ with probability $|\alpha_1|^2$. Similarly, We give the definition of the m-bit quantum register.

Definition 2.1. *A m-bit quantum register is a system composed of m qubits, whose state is a superposition of 2^m basic states, namely a vector on the unit ball of $\mathbb{C}^{2^m}$, written as*

$$\sum_{\substack{x_j \in \{0,1\} \\ j \in \{1,\ldots,n\}}} \alpha_{x_1,\ldots,x_n} |x_1, \ldots, x_n\rangle\,,\ \textit{where} \sum_{\substack{x_j \in \{0,1\} \\ j \in \{1,\ldots,n\}}} |\alpha_{x_1,\ldots,x_n}|^2 = 1.$$

When measuring the register, we will obtain the value $|x_1, \ldots, x_n\rangle$ with probability $|\alpha_{x_1,\ldots,x_n}|^2$, and collapse the state of the register to the vector $|x_1, \ldots, x_n\rangle$. ∎

Remark 2.2. *Here $\mathbb{C}^{2^m}$ is actually $(\mathbb{C}^2)^{\otimes m}$, the tensor product of m spaces of a single qubit $\mathbb{C}^2$.*

Remark 2.3. *For simplicity, we usually omit the normalization factor of states. For example, $|0\rangle - |1\rangle$ means $\frac{1}{\sqrt{2}}|0\rangle - \frac{1}{\sqrt{2}}|1\rangle$.*

3 Quantum Computation And BQP

3.1 Quantum Operations

According to the principle of quantum mechanics, the operations that we can do on a quantum register are as follows:

Definition 3.1. *A quantum operation for an m-qubit register is a unitary transformation $F : \mathbb{C}^{2^m} \to \mathbb{C}^{2^m}$. It maps a quantum register to another register linearly.*

Remark 3.2. *Quantum operations are all reversible, since it's a unitary transformation and we can do the inverse operation.*

Remark 3.3. *Quantum operations can be identified with unitary matrix when choosing a certain basis, such as the basis $\{|x\rangle | x \in \{0,1\}^m\}$.*

Here are some important examples of quantum operations below.

Example 3.4 (Flipping qubits)**.** *Flipping the first qubit of an m-qubit register means applying the* NOT *operation on the first qubit, which can be done as a quantum operation mapping the state $|b, x\rangle$ to the state $|1-b, x\rangle$ for any $b \in \{0,1\}$ and $x \in \{0,1\}^{m-1}$.*

Example 3.5 (Reordering qubits)**.** *We can exchange the values of several qubits by applying a permutation on basic states, which is a quantum operation since it can be expressed by a unitary matrix.*

Example 3.6 (Copying qubits)**.** *Notice that copying one qubit to another as a classical operation is not reversible. In quantum computing, we use the operation $|x, y\rangle \mapsto |x, x \oplus y\rangle$ instead, where the second qubit is often taken to be* 0*.*

Example 3.7 (Rotation on single qubit)**.** *Regarding* 1*-bit qubit as a two-dimensional vector. Rotation on a single qubit can be described by the matrix* $\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$*. Notice that it becomes flipping when* $\theta = \pi$*.*

Example 3.8 (AND of two bits)**.** *Just like copying qubits,* AND *of two bits as a classical operation is not reversible either. We can also use an additional* 0 *qubit on* z*, and take the "reversible* AND*" to be the operation* $|x, y, z\rangle \mapsto |x, y, z \oplus (x \wedge y)\rangle$*. This operation is often known as the Toffoli gate. Similarly, we can define a "reversible* OR*" to be the operation* $|x, y, z\rangle \mapsto |x, y, z \oplus (x \vee y)\rangle$*.*

Example 3.9 (The Hadamard operation)**.** *The Hadamard operation is the map* $|0\rangle$ *to* $|0\rangle + |1\rangle$ *and* $|1\rangle$ *to* $|0\rangle - |1\rangle$*. The corresponding matrix is* $\frac{1}{\sqrt{2}}$ $\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$*. If we apply the Hadamard operation to every qubit of an* m*-qubit register, and denote* $x \odot y$ *to be the inner product of the space* $\mathbb{F}_2^m$*, then* $|x\rangle = |x_1, x_2, \ldots, x_m\rangle$ *is mapped to* $\sum\limits_{y \in \{0,1\}^m} (-1)^{x \odot y} |y\rangle$*. Especially,* $|0^m\rangle$ *is mapped to* $\sum\limits_{x \in \{0,1\}^m} |x\rangle$*.*

However, we still need to give a precise definition of quantum computing. First, only *local* operations are possible to be physically implemented and applied efficiently. This kind of operation is called an *elementary quantum operation* or a *quantum gate.*

Definition 3.10. *An elementary quantum operation, or a quantum gate is a quantum operation that only acts on one or two qubits of the register.*

Theorem 3.11. *We can realize any arbitrary quantum operation with elementary quantum operations.*

Proof. See A. Y. Kitaev, M. Vyalyi, and A. Shen. *Classical and Quantum Computation.* AMS Press, 2002, Page 65. □

Theorem 3.12 (Kitaev)**.** *For every* $D \geq 3$ *and* $\epsilon \geq 0$*, there is an integer* $l \leq 100(D \log 1/\epsilon)^3$ *such that the following is true:*

Every $D \times D$ unitary matrix U can be approximated as a product of unitary matrices $U_1, \ldots, U_l$ such that

$$|U_{i,j} - (U_l \cdots U_1)_{i,j}| < \epsilon \text{ forevery } i, j \in \{1, \ldots, D\}$$

and each U_i correspond to applying either the Hadamard gate, the Toffoli gate, or the phase shift gate $\begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix}$, on at most three qubits.

Proof. See A. Y. Kitaev, M. Vyalyi, and A. Shen. *Classical and Quantum Computation.* AMS Press, 2002, Page 77. □

Remark 3.13. *More generally, for a finitely generated subgroup of $\mathbf{SU}(D)$, we can choose a sequence of generators no more than $O((\log(1/\delta))^{3+\varepsilon})$, whose products approximate a given operator with precision δ, by an algorithm of $O(poly(\log(1/\delta)))$ time. See the Solovay-Kitaev theorem.*

3.2 Classical Computation

Before defining **BQP**, let's give a brief explanation of classical computation.

Definition 3.14. *A Turing machine M is a tuple (Γ, Q, δ) where:*

- *Γ called alphabet of M is a finite set of the symbols on M's tape.*
- *Q is a finite set of M's possible states.*
- *$\delta : Q \times \Gamma \to Q \times \Gamma \times \{L, R, S\}$ is the transition function of M.*

Definition 3.15. *For a nonempty finite set S, we define $S^* := \bigcup_{n=0}^{\infty} S^n$*

Definition 3.16. *$f : \{0,1\}^* \to \{0,1\}^*$ and $T : \mathbb{N} \to \mathbb{N}$. We say a Turing machine M computes f in $T(n)$-time if $\forall x \in \{0,1\}^*$, when initialize M to the start configuration on input x, M halts after at most $T(|x|)$ steps with $f(x)$ on its tape.*

Definition 3.17. *$T : \mathbb{N} \to \mathbb{N}$. A language $L \subseteq \{0,1\}^*$ is in $\mathbf{DTIME}(T(n))$ iff its characteristic function $\chi_L(x) = \begin{cases} 0 & x \notin L \\ 1 & x \in L \end{cases}$ is in $\mathbf{DTIME}(T(n))$.*

Definition 3.18. $\mathbf{P}$:= $\bigcup_{c \geqslant 1} \mathbf{DTIME}(n^c)$.

Definition 3.19. *A Boolean circuit is a DAG(directed acyclic graph) with 0-in-degree nodes denoting the inputs and 0-out-degree nodes denoting the outputs, and other nodes labeled with OR, AND, or NOT called gates. The size of a Boolean circuit C, denoted by $|C|$, is the number of its vertices.*

Definition 3.20. $T : \mathbb{N} \to \mathbb{N}$. *A $T(n)$-size circuit family is $\{C_n\}_{n\in\mathbb{N}}$ such that C_n has n inputs and* 1 *output, and $|C_n| \leqslant T(n)$.*

Definition 3.21. $T : \mathbb{N} \to \mathbb{N}$. *Say a Boolean function $f : \{0,1\}^* \to \{0,1\} \in$ $\mathbf{SIZE}(T(n))$ if there is a $T(n)$-size circuit family $\{C_n\}_{n\in\mathbb{N}}$ such that $\forall x \in \{0,1\}^n$, $f(x) = C_n(x)$.*

Definition 3.22. *A circuit family $\{C_n\}_{n\in\mathbb{N}}$ is $\mathbf{P}$-uniform if there is a polynomial-time Turing machine that on input 1^n outputs the description of circuit C_n.*

Theorem 3.23. *A language L is computable by a $\mathbf{P}$-uniform circuit family iff $L \in\mathbf{P}$.*

Proof. See Arora S, Barak B. *Computational Complexity: A Modern Approach*, Cambridge University Press, 2009, Page 111. □

3.3 Quantum Time Complexity and BQP

Now, let's give the definition of quantum computing and the complexity class **BQP** (**B**ounded error, **Q**uantum, **P**olynomial time).

Definition 3.24 (Quantum computing and time complexity)**.** *Let $T : \mathbb{N} \to \mathbb{N}$. We say a Boolean function $f : \{0,1\}^* \to \{0,1\}$ is* ***computable in quantum*** $T(n)$***-time*** *if there is a polynomial-time classical Turing Machine that $\forall n \in \mathbb{N}$ on input $(1^n, 1^{T(n)})$ outputs $(F_1, \ldots, F_{T(n)})$ such that $\forall x \in \{0,1\}^*$ we can obtain the correct value of $f(x)$ with probability at least $\frac{2}{3}$ by the following process:*

1. *Initialize an m-qubit register to the state $|x0^{m-n}\rangle$, where $m \leqslant T(n)$.*

2. *Apply $F_1, \ldots, F_{T(n)}$ one by one to the register.*

3. *Measure the register and get a value Y.*

4. *Output Y_1.*

Let $f : \{0,1\}^* \to \{0,1\}^l, f(x) = (f_1(x), \ldots, f_l(x))$, we say f is computable in quantum $T(n)$-time if $f_j(x)$ is computable in quantum $T(n)$-time $\forall j = 1, \ldots, l$.

Definition 3.25 (class **BQP**). *A Boolean function $f : \{0,1\}^* \to \{0,1\}^l$ is in* **BQP** *if there is a polynomial $p : \mathbb{N} \to \mathbb{N}$ such that f is computable in quantum $p(n)$-time. We say a language $L \subseteq \{0,1\}^*$ is in* **BQP** *iff its characteristic function $\chi_L(x) = \begin{cases} 0 & x \notin L \\ 1 & x \in L \end{cases}$ is in* **BQP**.

Remark 3.26. *Another definition of quantum computing and* **BQP** *is* ***quantum circuit****. A quantum circuit is a DAG(directed acyclic graph) with 0-in-degree nodes denoting the inputs and 0-out-degree nodes denoting the outputs, and other nodes denoting the quantum gates.*

Definition 3.27. *A language $L \subseteq \{0,1\}^*$ is in* **BQP** *iff there exists a* **P***-uniform polynomial-size quantum circuit family $\{C_n\}_{n \in \mathbb{N}}$ over some finite universal gates and a polynomial q such that $\forall n \in \mathbb{N}, x \in \{0,1\}^n$ have:*

- *$x \in L \Rightarrow C_n(|x\rangle|0\rangle^{\otimes q(n)})$ accepts with probability $\geqslant \frac{2}{3}$*
- *$x \notin L \Rightarrow C_n(|x\rangle|0\rangle^{\otimes q(n)})$ accepts with probability $\leqslant \frac{1}{3}$*

Notice that some examples (flipping, reversible AND, reversible NOT) of quantum operations listed above can take the place of the fundamental classical operations (NOT, AND, OR). In fact, we can efficiently compute any classical operation using quantum operations as long as we have sufficient free qubits:

Lemma 3.28. *If $f : \{0,1\}^n \to \{0,1\}^m$ is computable by a Boolean circuit of size S, then there is a quantum circuit of size $2S + m$ computing the mapping $|x\rangle|0^{2m+S}\rangle \mapsto |x\rangle|f(x)\rangle|0^{m+S}\rangle$.*

Proof. Replacing the Boolean gates by their corresponding quantum operations, we get a map $|x\rangle|0^{2m+S}\rangle \mapsto |x\rangle|0^m\rangle|f(x)\rangle|z\rangle$, denoted φ. Next copy $f(x)$ to $|0^m\rangle$ using copying operation m times. Then if we apply the reversing operation φ^{-1}, it will return to the original case except the copy of $f(x)$, that is, $\varphi^{-1} : |x\rangle|f(x)\rangle|f(x)\rangle|z\rangle \mapsto |x\rangle|f(x)\rangle|0^{m+S}\rangle$.

Corollary 3.29. P⊆BPP⊆BQP.

Proof. By Theorem 3.23, Definition 3.27 and Lemma 3.28, we know that **P⊆BQP**. Since we can simulate a coin toss using a Hadamard gate, we get **P⊆BPP⊆BQP**. □

4 Grover's Search Algorithm

Theorem 4.1. *There is a quantum algorithm such that, for every polynomial-time computable function* $f : \{0,1\}^n \to \{0,1\}$*, we can find a string* $a \in \{0,1\}^n$ *satisfying* $f(a) = 1$ *in* $poly(n)2^{\frac{n}{2}}$ *time (if such string* a *exists).*

Proof. We use an $n{+}1{+}m$-qubit register, where m is large enough to compute the transformation $|x\sigma 0^m\rangle \mapsto |x(\sigma \oplus f(x))0^m\rangle$ by Lemma 3.28.
Initial state: $|0^{n+1+m}\rangle$
Step 1. Apply Hadamard operation on first n qubits.
Do step 2. and step 3. for $2^{n/2}$ times as follows:
Step 2. Compute $|x\sigma 0^m\rangle \mapsto |x(\sigma \oplus f(x))0^m\rangle$;
if $\sigma = 1$ multiply by -1;
compute $|x\sigma 0^m\rangle \mapsto |x(\sigma \oplus f(x))0^m\rangle$ again.
Step 3. Apply Hadamard operation on first n qubits;
if first n-qubits are all zero, then flip $n+1$st qubit;
if the $n+1$st qubit is 1, then multiply by -1;
if the first n-qubits are all zero, then flip $n+1$st qubit;
apply Hadamard operation on first n qubits again.
Step 4. Measure register and let a' be the obtained value of the first n qubits. If $f(a') = 0$, repeat Step 2. and Step 3. for $2^{n/2}$ times until $f(a') = 1$.
Output: a' such that $f(a') = 1$.
Notice that step 2 and step 3 reflects the vector around $\mathbf{u} = \frac{1}{2^{n/2}} \sum_{x \in \{0,1\}^n} |x\rangle$ and $\mathbf{e} = \sum_{f(x)=0} |x\rangle$, hence the vector rotates the angle $2\langle e, u\rangle$ towards $|a\rangle$ after one loop. Hence in $O(1/\theta) = O(2^{n/2})$ steps, it will be so close to $|a\rangle$ that their inner product is larger than $\cos\theta$. If measuring now, we get $|a\rangle$ with constant probability. Thus, this algorithm has time complexity of $O(poly(n)2^{n/2})$. □

5 Simon's Algorithm

Theorem 5.1. *There is a polynimial-time quantum algorithm such that, for a polynomial-size classical circuit of a function $f : \{0,1\}^n \to \{0,1\}^n$, find the string $a \in \{0,1\}^n$ satisfying $f(x) = f(x \oplus a)$ for every $x \in \{0,1\}^n$ (if such a exists).*

Remark 5.2. *The problem to find such a "period" of f is often called Simon's problem.*

Proof. We use an $2n + m$-qubit register, where m is sufficiently large to compute the transformation $|xz0^m\rangle \mapsto |x(z \oplus f(x))0^m\rangle$.
Initial state: $|0^{2n+m}\rangle$.
Step 2i-1. Apply Hadamard operation on first n qubits;
compute $|xy0^m\rangle \mapsto |x(y \oplus f(x))0^m\rangle$;
apply Hadamard operation on first n qubits again.
Step 2i. Measure first n qubits of register to obtain a value y_i such that $y_i \odot a = 0$.
Repeat until we get y_i enough to retrieve a.
In fact, when $k \geq 2n$ there will be $n-1$ linearly independent y_i with high probability by the lemma below. □

Lemma 5.3. *Choose n vectors uniformly at random from $\mathbb{F}_2^n$, then with probability at least $1/5$ the vectors are linearly independent.*

Proof. $p = \prod_{i=0}^{n-1} \frac{2^n - 2^i}{2^n} = \prod_{i=1}^{n}(1 - 2^{-i}) = \prod_{i=1}^{n}(1 + \frac{1}{2^i - 1})^{-1} > \frac{3}{8}e^{-1/2} > \frac{1}{5}$. □

6 Shor's Algorithm

6.1 Quantum Fourier transformation

Definition 6.1. *For a vector $f = (f(0), \dots, f(M-1)) \in \mathbb{C}^M$, the Fourier transformation of f is a vector $\hat{f} = (\hat{f}(0), \dots, \hat{f}(M-1)) \in \mathbb{C}^M$ defined by:*

$$\hat{f}(x) = \frac{1}{\sqrt{M}} \sum_{y=0}^{M-1} f(x)\omega_M^{xy},\ x = 0, \dots, M-1 \text{ and } \omega_M = e^{\frac{2\pi i}{M}}$$

Remark 6.2. *We know that the Fourier basis*

$$\left\{\frac{1}{\sqrt{M}}\left(1,\omega_M^{-x},\dots,\omega_M^{-(M-1)x}\right)\right\}_{x=0,\dots,M-1}$$

is an orthonormal basis. So the Fourier transform $FT_M : f \mapsto \hat{f}$ is a unitary operation.

Based on the divide-and-conquer idea, we have the quantum operation that change the state of a quantum register to its Fourier transform.

Lemma 6.3. *For any positive integer m, we can change the state of a quantum register $f = \sum_{x=0}^{2^m-1} f(x)|x\rangle$ to its Fourier transform $\hat{f} = \sum_{x=0}^{2^m-1} \hat{f}(x)|x\rangle$ using only $O(m^2)$ elementary quantum operations.*

Proof. Let $W_m = \operatorname{diag}\{\omega^0,\dots,\omega^{2^m-1}\}$ be a $2^m \times 2^m$ diagonal matrix, H be the Hadamard gate. Suppose we have run $FT_{2^{m-1}}$ on the first $m-1$ qubits. Apply W on the first $m-1$ qubits on $|x_1,\dots,x_{m-1},1\rangle$, then apply H on the last qubit, and move the last qubit to the first one. States are as follows:

$$\begin{aligned}
&FT_{2^{m-1}}f_0|0\rangle + FT_{2^{m-1}}f_1|1\rangle\\
\to&FT_{2^{m-1}}f_0|0\rangle + W_{m-1}\cdot FT_{2^{m-1}}f_1|1\rangle \quad (m \text{ quantum gates})\\
\to&FT_{2^{m-1}}f_0(|0\rangle+|1\rangle) + W_{m-1}\cdot FT_{2^{m-1}}f_1(|0\rangle-|1\rangle) \quad (1 \text{ quantum gate})\\
&= (FT_{2^{m-1}}f_0 + W_{m-1}\cdot FT_{2^{m-1}}f_1)|0\rangle + (FT_{2^{m-1}}f_0 - W_{m-1}\cdot FT_{2^{m-1}}f_1)|1\rangle\\
\to&|0\rangle(FT_{2^{m-1}}f_0 + W_{m-1}\cdot FT_{2^{m-1}}f_1) + |1\rangle(FT_{2^{m-1}}f_0 - W_{m-1}\cdot FT_{2^{m-1}}f_1)\\
&= \hat{f} \quad (1 \text{ quantum gate})
\end{aligned}$$

Let $T(m)$ be the number of quantum gates used for FT_{2^m}. Then we know that $T(m) = T(m-1) + O(m)$, so $T(m) = O(m^2)$. □

6.2 Reducing Factoring to Order Finding

Lemma 6.4. *If odd nonprime n is not a prime power, then the probability that a uniformly random element $x \in (\mathbb{Z}/n\mathbb{Z})^\times$ has the property that* $\operatorname{ord}(x) = 2r, r \in \mathbb{Z}_+$ *and* $\{\gcd(n, x^r+1), \gcd(n, x^r-1)\} \cap \{1, n\} = \varnothing$ *is at least $\frac{1}{4}$.*

Proof. Only prove when $n = pq$ for primes $p > q \geqslant 3$. Let

$$\begin{aligned}
\varphi : (\mathbb{Z}/n\mathbb{Z})^\times &\to (\mathbb{Z}/p\mathbb{Z})^\times \times (\mathbb{Z}/q\mathbb{Z})^\times\\
x &\mapsto (x_p, x_q)
\end{aligned}$$

then $\operatorname{ord}(x) = \operatorname{lcm}(\operatorname{ord}(x_p), \operatorname{ord}(x_q))$ and $\varphi(-1) = (-1, -1)$. Since all elements with odd order in $(\mathbb{Z}/p\mathbb{Z})^\times$ is a proper subgroup of $(\mathbb{Z}/p\mathbb{Z})^\times$, it contains at most half of the elements of $(\mathbb{Z}/p\mathbb{Z})^\times$. So the probability that x_p has even order is at least $\frac{1}{2}$. Let

$$G_l = \{x \in (\mathbb{Z}/q\mathbb{Z})^\times \mid \operatorname{ord}(x) = 2^j * c,\ c \text{ is odd and } j \leqslant l\},$$

then $G_0 \leqslant \cdots \leqslant G_{l-1} \leqslant G_l \leqslant \cdots \leqslant (\mathbb{Z}/q\mathbb{Z})^\times$. Let

$$\begin{aligned} f : G_l &\to G_{l-1} \\ x &\mapsto x^2 \pmod q \end{aligned}$$

then $\ker(f) \supseteq \{1, -1\}$, so $\#G_{l-1} \geqslant \#G_l/2$. Let $\operatorname{ord}(x_p) = 2^{s_p}c_p$ and $\operatorname{ord}(x_q) = 2^{s_q}c_q$, where c_p and c_q are odd. Then the probability that $s_p = s_q$ is at most $\frac{1}{2}$. So the probability that $s_p \geqslant 1$ and $s_p \neq s_q$ is at least $\frac{1}{4}$.

When $s_p \geqslant 1$ and $s_p \neq s_q$, $\operatorname{ord}(x) = 2^{\max(s_p,s_q)} \operatorname{lcm}(c_p, c_q)$ and $r = \operatorname{ord}(x)/2$. Since $s_p \neq s_q$, $\varphi(x^r) \neq (-1, -1)$, and $x^r \not\equiv -1 (\operatorname{mod} n)$. So $x^r \pm 1 \not\equiv 0 \pmod n$ and $(x^r + 1)(x^r - 1) = x^{2r} - 1 \equiv 0 \pmod n$. Thus we have $\{\gcd(n, x^r + 1), \gcd(n, x^r - 1)\} \cap \{1, n\} = \varnothing$. □

6.3 Shor's order-finding algorithm

Lemma 6.5. *There is a polynomial-time quantum algorithm that on input (a, n) outputs $\operatorname{ord}_n(a)$.*

Proof. Let $m = \lceil 5 \log(n) \rceil$. We use a $m + \operatorname{poly}(\log(n))$-qubit register. Since $x \mapsto a^x (\operatorname{mod} n)$ is computable in $\operatorname{poly}(\log(n))$ time, we can compute $|x\rangle|y\rangle \mapsto |x\rangle|y \oplus (a^x (\operatorname{mod} n))\rangle$.

The algorithm is as follows:

Step 1. Run QFT to the first m qubits.
Step 2. Compute $|x\rangle|y\rangle \mapsto |x\rangle|y \oplus (a^x (\operatorname{mod} n))\rangle$.
Step 3. Measure last n qubits and get y_0.
Step 4. Run QFT to the first m qubits.
Step 5. Measure the first m qubits and get x.
Step 6. Find $\frac{p}{q}$ such that $\gcd(p, q) = 1$ and $\left|\frac{x}{2^m} - \frac{p}{q}\right| \leqslant \frac{1}{10 \cdot 2^m}$.
Step 7. If $a^q \equiv 1 (\operatorname{mod} n)$, output q.

States here are as follows:

$$\frac{1}{\sqrt{M}}\sum_{x=0}^{2^m-1}|x\rangle|0^n\rangle$$

$$\rightarrow\frac{1}{\sqrt{M}}\sum_{x=0}^{2^m-1}|x\rangle|a^x(\operatorname{mod} n)\rangle$$

$$\rightarrow\frac{1}{\sqrt{K}}\sum_{l=0}^{K-1}|x_0+lr\rangle|y_0\rangle,$$

x_0 is the smallest number s.t. $a^{x_0}\equiv y_0(\operatorname{mod} n), K=\left\lfloor\frac{M-x_0-1}{r}\right\rfloor$

$$\rightarrow\frac{1}{\sqrt{M}\sqrt{K}}\left(\sum_{x=0}^{2^m-1}\sum_{l=0}^{K-1}\omega_{2^m}^{(x_0+lr)x}|x\rangle\right)|y_0\rangle$$

Lemma 6.6. *The number of* $x\in\{0,\dots 2^m-1\}$ *such that* $0\leqslant xr(\operatorname{mod} 2^m)<r/10$ *and* $\gcd(\lfloor xr/2^m\rfloor,r)=1$ *is at least* $\Omega(r/\log r)$*.*

Lemma 6.7. *If* $0\leqslant xr(\operatorname{mod} 2^m)<r/10$*, then before Step 5, the coefficient of* $|x\rangle$ *is at least* $\Omega(\frac{1}{\sqrt{r}})$*.*

With these two lemmas, we know that with a probability of at least $\Omega(1/\log r)$, the measured value x has the property mentioned in Lemma 6.6. The property shows that, for $c=\lfloor xr/2^m\rfloor$, we have $|xr-c2^m|<r/10$, and

$$\left|\frac{x}{2^m}-\frac{c}{r}\right|<\frac{1}{10\cdot 2^m}<\frac{1}{4n^4}$$

According to the theory of continued fraction, we know that in this condition the algorithm will output r. So the algorithm have a probability of at least $\Omega(1/\log r)$, and then $\Omega(1/\log n)$ to output r. Then we can repeatedly run it several times and take the smallest output to increase the probability of successfully getting $\operatorname{ord}_n(a)$. □

Finally, we come to the proof of the main theorem.

Theorem 6.8 (Shor)**.** *There is a quantum algorithm that on input* n *outputs the prime factorization of* n *in* $\operatorname{poly}(\log(n))$ *time.*

Proof. We only need to give an algorithm that on input n output a nontrivial factor of n. Because we can run the algorithm recursively to get the prime factorization of n.

The algorithm is as follows:

Step 1. If n is even, return 2; else proceed to Step 2.

Step 2. For $k = 2, \dots \log n + 1$, if $n = m^k$, return m; else proceed to Step 3.

Step 3. Choose an $a \in \{1, \dots, n-1\}$ uniformly randomly. Compute $b = \gcd(a, n)$ with Euclid's algorithm. If $b > 1$, return b; else proceed to Step 4.

Step 4. Compute $r = \mathrm{ord}_n(a)$. If r is odd, return "n is prime"; else proceed to Step 5.

Step 5. Compute $d = \gcd(a^{r/2} - 1, n)$. If $d > 1$, return d; else return "n is prime". □

By lemmas above, we know that this quantum algorithm is polynomial-time. And we can get an factor with a probability of at least $\frac{1}{4}$. Repeat it several times then we can increase the probability of success to at least $\frac{2}{3}$.

7 Conclusion

So far, we have given the definition of classical and quantum computing, and we know **P**$\subseteq$**BPP**$\subseteq$**BQP**. Grover's search algorithm, Simon's algorithm and Shor's algorithm show us that quantum computing may be strictly more powerful than classical computing. Since we haven't found a polynomial-time probabilistic computation for integer factorization, and we believe there isn't. Shor's algorithm makes us believe that **BPP**$\neq$**BQP**, and shows us that many encryption schemes such as RSA may be not safe after the quantum computer is realized physically.

References

[1] Arora S, Barak B. *Computational Complexity: A Modern Approach*, Cambridge University Press, 2009.

[2] A. Y. Kitaev, M. Vyalyi, and A. Shen. *Classical and Quantum Computation.* AMS Press, 2002.

On Symbolic-Numeric Methods of Integrating Rational Functions

Chen Guanyi

May 31th, 2022

1. Introduction

Precise computation of integrals of rational functions turns out important in more advanced integrating algorithms. Typical methods of rational integration are either numeric, which means aiming at an approximate output (of an actual value), or symbolic, which means aiming at a precise formula output that exists theoretically.

Both kinds of methods have advantages and disadvantages: as for numerical ones, the output is usually accurate enough especially for definite integration, yet on ill-conditioned integrals purely numerical methods are over-sensitive to approximate processings, like integrating on intervals that nearly contain a root of denominator. For symbolic ones, although precise algebraic expressions of indefinite integrals help further analysis of results (for example asymptotic analysis), the formula is likely to be inaccessible when roots of denominator are complicated or even impossible to be solved in radicals. However, proper hybrid methods that rely on both numeric and symbolic processes may be able to combine their advantages and avoid their shortcomings.

Based on ideas of Article [1] and Book [2], we will first build basic definitions and theorems, together with some fundamental algorithms; and then introduce two numeric-symbolic methods of rational integrations, called N-PFD and N-LRT, give a posteriori computable bound of errors and compare their behaviours on typical examples to decide which method seems better. Based on typical symbolic algorithms, numerical methods to approximate roots are applied.

2. Basic Definitions, Theorems and Algorithms

2.1 Rational Integrals and Decomposition

Definition 2.1.1 A **rational integral** is the definite (or indefinite) integral of two polynomials with real coefficients. (In wider range complex coefficients are considered, but not here.) Whether definite or indefinite, we will use the notation $\int f(x) =$ for one of the anti-derivatives of $f(x)$.

According to Partial Fraction Theorem we've learned for rational functions, the following theorem is trivial:

Theorem 2.1.2 Suppose $f(x) = \dfrac{P(x)}{Q(x)}$ is the rational function to be integrated, then there exists a unique decomposition (in the sense of differing a constant) of $\int f(x) = \int \dfrac{G(x)}{H(x)} + \dfrac{C(x)}{D(x)}$ such that G, H, C, D are polynomials of real coefficients and H has no repeated roots.

Proof: According to Partial Fraction Theorem, the only concern is uniqueness. Suppose there is another decomposition $\int f(x) = \int \frac{G_1(x)}{H_1(x)} + \frac{C_1(x)}{D_1(x)}$, then $\int (\frac{G_1(x)}{H_1(x)} - \frac{G(x)}{H(x)})$ is a nontrivial rational function, yet the derivative of any nontrivial rational function has repeated factors on denominator, and $\frac{G_1(x)}{H_1(x)} - \frac{G(x)}{H(x)}$ has none since the denominator is the product of some non-repeating factors of $Q(x)$. Contradiction. □

Definition 2.1.3 We call $\frac{C(x)}{D(x)}$ the **rational part** of the integral $\int f(x)$, and $\int \frac{G(x)}{H(x)}$ the **transcendental part**.

A symbolic algorithm for finding out the rational part is necessary, since two parts of integrals have different terms: One is rational and the other consists of log and arctan terms, and the rational part is relatively easier to deal with. To build an algorithm, we first need some basic methods for polynomials which are able to apply on computers:

Algorithm 2.1.4(Extended Euclid Algorithm, EEA) Suppose $f, g \in \mathbb{R}[x](\deg f \geq \deg g)$, then we can find $u = \gcd(f, g)$ and $s, t \in \mathbb{R}[x]$ such that $sf + tg = u$ and $\deg s < \deg g, \deg t < \deg f$ unless the right hand side be zero, by the following steps (regardless of multiplying a real constant in output):

Step 1: Let $u = f, v = g, \mathbf{S} = (1, 0), \mathbf{T} = (0, 1)$

Step 2: If $v = 0$ go to Step 4;

Step 3: Otherwise do Euclid Algorithm $u = qv + r$, let $\mathbf{L} = \mathbf{S} - q\mathbf{T}$, $u = v$, $v = r$, $\mathbf{S} = \mathbf{T}$, $\mathbf{T} = \mathbf{L}$, and go back to Step 2;

Step 4: print $u = \gcd(f, g)$, $\mathbf{S} = (s, t)$.

All the non-trivial u, v's that generated in order, from f to $\gcd(f, g)$, is called the **remainder sequence** of f and g.

Proof: It is obvious that when finishing any step, the formulas $\mathbf{S} \cdot (f, g) = u$, $\mathbf{T} \cdot (f, g) = v$, $\gcd(f, g) = \gcd(u, v)$ remain unchanged, and in the last step $v = 0$, so $u = \gcd(f, g)$. Besides the $\mathbf{S}$ in final output is the previous $\mathbf{T}$, and by induction the maximal degree of the second item of that $\mathbf{T}$ will not exceed

$$\sum_{i=1}^{k} \deg q_i = \sum_{i=1}^{k} (\deg u_i - \deg v_i) < \deg f,$$

where the second equality won't hold because if in the previous step before u, v are changed, $\deg v = 0$ already holds, then v should be 0 afterwards and the $\mathbf{S}$ produced in the same step will be directly printed, which contradicts the assumption. By $\deg t < \deg f$ we find $\deg s < \deg g$ by $sf + tg = u$ unless $\deg g + \deg f \leq \deg u$, which means they are both constants and s, t are therefore constants. □

Algorithm 2.1.5(Square-free Decomposition) Suppose $f \in \mathbb{F}[x]$ where $\mathrm{char}\mathbb{F} = 0$, then it's easily seen that $f = \prod_{i=1}^{m} h_i^i$ is unique (in the sense of multiplying a nonzero constant) where $\deg h_m > 0$ and h_i all in $\mathbb{F}[x]$. We can extract the h_i's by the following steps:

Step 1: Let $u = \gcd(f, f')$, $v_1 = f/u$, $w_1 = f'/u$, $i = 1$;

Step 2: Let $h_i = \gcd(v_i, w_i - v_i')$, $v_{i+1} = v_i/h_i$, $w_{i+1} = (w_i - v_i')/h_i$, $i = i + 1$;

Step 3: If $v_i \neq 1$ do Step 2 again;

Step 4: Print all h_i's.

Proof: To prove the algorithm is valid, we consider if h_1 is the product of distinct square-free prime factors of f, and if $v_2 = u/\gcd(u, u')$ and $w_2 = u'/\gcd(u, u')$ hold. In fact, noticing that u equals f divided by the product of all its distinct prime factors, the algorithm will repeats extracting all prime factors of order i once it can be viewed as replacing the initial f by u in the first iteration and repeating until finish.

By simple computation $h_1 = \gcd(f/u, fu'/u^2)$. For any $p^k||f(k \geq 1)$, $v_p(h_1) = 1$ iff $k = 1$ is easily checked since $v_p(u') \geq 0$ and $v_p(u) = k-1$ for $k \geq 1$, and $v_p(u') = k-2$ for $k \geq 2$; for $v_2 = u/(u^2h_1/f)$ and $w_2 = u'/(u^2h_1/f)$, we only need to verify

$$\gcd(u, u') = u^2h_1/f,$$

where the factors of both sides will belong to f. For $k = 1$ both sides have zero order of p, and for $k \geq 2$ both sides have $k-2$, which proves the validity. □

Algorithm 2.1.6(Partial Fraction Method) Suppose $Q(x) = \prod_{i=1}^{m} q_i(x)$ and $\deg P < \deg Q$ where q_i's are relatively prime. Then we can find f_1, f_2 s.t.

$$\frac{P}{Q} = \frac{f_1}{q_1} + \frac{f_2}{Q/q_1},\ \deg f_1 < \deg q_1,\ \deg f_2 < \deg(Q/q_1).$$

Algorithm and Proof: Use Extended Euclid Algorithm for $q_1 = s$ and $\frac{Q}{q_1} = v$ to find $1 = su + tv$, then $\frac{P}{Q} = \frac{Pt}{u} + \frac{Ps}{v}$, so we can replace Pt and Ps by $Pt(\bmod u)$ and $Ps(\bmod v)$, and the right hand side becomes $\frac{P_0}{Q}$ where $\deg P_0 < \deg Q$, yet $Q|P_0 - P$, which means $P_0 = P$.

Repeating the process above we can find f_i's s.t. $\deg f_i < \deg q_i$, and

$$\frac{P}{Q} = \sum_{i=1}^{m} \frac{f_i}{q_i}. \quad \square$$

Algorithm 2.1.7(Hermite Method) We consider the following steps to find the rational part symbolically (Suppose $\deg P < \deg Q$):

Step 1: Apply Square-free Decomposition to Q and suppose $Q(x) = \prod_{i=1}^{m} q_i(x)^i$. By applying Partial Fraction Method we only need to consider the decomposition of $\int \frac{f_i(x)}{q_i(x)^i}$.

Step 2: For each i, notice that $\gcd(q_i, q_i') = 1$, we find polynomials a, b s.t. $aq_i + bq_i' = 1$, and it is easily checked that

$$\int \frac{f_i(x)}{q_i(x)^i} = \int \frac{f_ia + (f_ib)'/(i-1)}{r_i^{i-1}} - \frac{f_ib}{(i-1)r_i^{i-1}},$$

which means the order of the denominator left has decreased.

Step 3: Repeat Step 2 until all denominators are prime. In this way we extract the rational part successfully, and with accurate form since the whole process is purely symbolic (given the context that coefficients are accurate, such as when coefficients are in $\mathbb{Q}$ or wider structures like $\mathbb{Q}(\sqrt{2})$ in definitions of computer algebra systems).

2.2 Transcendental Part

Compared with the rational part, the transcendental part is more difficult to handle especially when it becomes unavoidable to take approximate roots of denominator.

Algorithm 2.2.1(Partial Fraction Decomposition, PFD) We focus on $\mathbb{C}[x]$ we actually find that the transcendental part $\dfrac{G(x)}{H(x)}$ is able to be written in the sum of $\dfrac{c_i}{x-z_i}$'s, where c_i is the residue of $\dfrac{G}{H}$ at z_i. We have

$$\int \frac{G(x)}{H(x)} = \sum_{i=1}^{k} c_i \mathrm{log}(x - z_i),$$

but in real form we must combine the items of complex roots. As $\dfrac{G(x)}{H(x)}$ has real coefficients, we push $\bar{x}$ into x and get $c_i = \bar{c_j}$ for $z_i = \bar{z_j}$. By taking exponents, the following two formulas are trivial (in the corresponding branch):

$$\log(x - z_i) + \log(x - \bar{z_i}) = \log(x^2 - |z_i|^2),$$
$$i(\log(x - z_i) - \log(x - \bar{z_i})) = 2\mathrm{arctan}(\frac{x - \mathrm{Re}(z_i)}{-\mathrm{Im}(z_i)}),$$

so whenever we reach an expression of the partial fraction form we can express our result in real functions.

By simple complex analysis we know $c_k = \dfrac{G(z_k)}{H'(z_k)}$, and whenever we find roots (such as when we are able to solve them out), we are able to give a symbolic anti-derivative.

The PFD Algorithm requires splitting H into irreducible factors, and in the context of precise calculation, unnecessary algebraic numbers like $\sqrt{2}$ may be introduced for unnecessary problems like $\displaystyle\int \frac{2x}{x^2+2}$. However, if we focus on residues and their different following sums, we may introduce a new algorithm. For the same c as a residue, the sum of all corresponding items is

$$c \sum_{i:c_i=c} \log(x - z_i) = c\log(\prod_{z\in z_i's:G-cH'|_z=0} (x - z)) = c\log(\mathrm{gcd}(H, G - cH')),$$

But we need certain tools to find c_i's and their following sums without computing the z_i's. Luckily, by the concept and properties of the Rothstein-Trager Resultant, we are able to view c_i's as a root of a certain polynomial and represent the logarithms again with respect to c_i's.

Definition 2.2.2 Define the **resultant** of two polynomials f, g by

$$\mathrm{Res}_x(f, g) = \det \phi_{f,g},$$

where $\phi_{f,g} : P_m \times P_n \to P_{m+n}, m = \deg g, n = \deg f$ is the linear transform defined by

$$\phi_{f,g}(s, t) = sf + tg,$$

and P_k represents the k-dimensional linear space of all polynomials with degree below k.

For convenience of computation, we select the standard basis $B_k = \{1, x, \ldots, x^{k-1}\}$ for every P_k, and represent $\phi_{f,g}$ in the standard basis as a matrix $S = S(f, g)$.

Definition 2.2.3 The matrix $S_(f, g)$ is called the **Sylvester matrix** of f and g.

Theorem 2.2.4 We reach the following directly from definition:

(1)$\mathrm{gcd}(f, g) = 1$ iff $\phi_{f,g}$ has a trivial kernel, iff $\phi_{f,g}$ is bijective, iff $\mathrm{Res}_x(f, g) = \det S_(f, g) \neq 0$. (2)$S(f, g)$ is a $(m + n) \times (m + n)$ matrix with coefficients $f_n, \ldots, f_0$ beginning from (i, i) to $(i + n, i)$ $(1 \leq i \leq m)$ in the first m columns, and $g_m, \ldots, g_0$ beginning from $(j - m, j)$ to (j, j) $(m + 1 \leq j \leq m + n)$ in the last n columns, and 0 in all other positions.

Now we will find if we consider $\Re(c) = \mathrm{Res}_x(H, G - cH')$ as a polynomial of c, called the **Rothstein-Trager resultant**, we can find our residues by finding roots of $\Re$.

However, it's not easy to compute resultants. Although we know the definitions properties of $\mathrm{Res}(f, g)$ for polynomial $f, g\,(n = \deg f > \deg g = m)$ it is still hard to be calculated quickly, especially as a determinant. Based on special 'parallelogram' structure of Sylvester matrix, a natural idea is transform and decompose the matrix in smaller ones. Notice that we can link polynomial multiplication with matrix structure, we may do EEA to decompose (f, g) and the Sylvester matrix at the same time. The following lemma is useful:

Lemma 2.2.5 Suppose $f = qg + \rho r$ where $n = \deg f \geq \deg g = m > \deg r = d$, f, g, r monic, ρ a constant. Then

$$\mathrm{Res}_x(f, g) = (-1)^{mn} \rho^m \mathrm{Res}_x(g, r).$$

Proof: We focus on the matrix structure. Denote $f = \sum_{i=0}^{n} f_i x^i$ $(f_n = 1)$, and define g_i, q_i, ρ_i's similarly.

In matrix form, the condition become

$$\begin{bmatrix} f_n \\ f_{n-1} \\ \vdots \\ \vdots \\ \vdots \\ f_0 \end{bmatrix} - \begin{bmatrix} g_m & & \\ g_{m-1} & \ddots & \\ \vdots & & 1 \\ g_0 & & \vdots \\ & \ddots & \vdots \\ & & g_0 \end{bmatrix} \begin{bmatrix} q_{n-m} \\ q_{n-m-1} \\ \vdots \\ \vdots \\ \vdots \\ q_0 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \rho r_d \\ \vdots \\ \rho r_0 \end{bmatrix},$$

which means we can eliminate the first m columns in $S(f, g)$ into coefficients of r by the last n columns, and then switch the left and right parts to get

$$(-1)^{mn} \cdot \det \begin{bmatrix} g_m & & & & & \\ g_{m-1} & \ddots & & \rho r_d & & \\ \vdots & \ddots & g_m & \vdots & \ddots & \\ g_0 & & g_{m-1} & \rho r_0 & & \rho r_d \\ & \ddots & \vdots & & \ddots & \vdots \\ & & g_0 & & & \rho r_0 \end{bmatrix} = (-1)^{mn} \rho^m \det S_* \det S(g, r) = (-1)^{mn} \rho^m \det S(g, r),$$

where S_* is the top left part of this transformed matrix, upper-triangular with diagonals all 1. □

By extracting prime term coefficients, we directly lead to the following using Lemma 2.2.5 and induction:

Theorem 2.2.6 (Computability of Resultants) Suppose $f = \rho_0 r_0$, $g = \rho_1 r_1$ are the initial coprime polynomials with $\rho_0, \rho_1 \in \mathbb{F}$, r_0, r_1 monic, and apply Monic EEA to represent the remainder sequence with form $\rho_i r_i$ where r_i monic and $\rho_i \in \mathbb{F}$, $\deg r_i = n_i$ for $i = 1, \ldots, l$. Then

$$\mathrm{Res}_x(f, g) = (-1)^{\tau} \rho_0^{n_1} (\prod_{j=1}^{l} \rho_j^{n_{j-1}}) \mathrm{Res}_x(r_{l-1}, r_l = 1) = (-1)^{\tau} \rho_0^{n_1} (\prod_{j=1}^{l} \rho_j^{n_{j-1}}),$$

where $\tau = \sum_{j=1}^{l-1} n_{j-1} n_j$.

Another important problem is computing $\gcd(H, G - cH')$ for residue c. We can actually find $\gcd(H, G - yH')$ on field $\mathbb{R}(y)$ first, which can be done by EEA; Suppose the remainder sequence is $R_0(x, y), \ldots, R_l(x, y)$. When y is substituted by any specific value of y, the actual process of EEA won't change in the beginning, since the decomposition into quotient and remainder is unique under restriction of degrees, and substitution gives a valid decomposition, which means substitution into the corresponding $R_i(x, y)$ leads to the same result of the algorithm in $\mathbb{R}[x]$ in every step till finish. So the moment that substitution into the next polynomial equals zero, the previous polynomial in remainder sequence gives $\gcd(H, G - cH')$ after substitution.

Now, If we know c repeats k times as a residue, then $\deg \gcd(H, G - cH') = k$ and there exists a unique R_i denoted by $R_{i(k)}$ such that $\deg_x R_i = k$. In this case

$$\gcd(H, G - cH') = R_{i(k)}(x, c),$$

and computation is finished. So under this algorithm being able to find c_i's are not enough: We need to know multiplicities of them when all roots of H are exhausted, since we need to know which degree to substitute into. Luckily the multiplicity of residues and their corresponding roots are also contained in the resultant, in the following means:

Theorem 2.2.7 $\Re(c) = M \prod\limits_{H(z)=0} (G(z) - cH'(z))$($M$ is a nonzero constant).

Proof: We use a quite special view in this equality: H is fixed and the theorem is a claim for any arbitrary choose of coefficients of G for $\deg G < \deg H$. The non-zero constant M must be totally determined by H, because the coefficient of $c^{\deg H}$ is equal when we only view c as variable in two sides, yet in the left side, any single monomial with degree $\deg H$ has coefficients all chosen on the first $\deg H$ columns in the complete monomial expansion of the corresponding Sylvester matrix; In the right side $c^{\deg H}$ has coefficient $(-1)^{\deg H}$. Hence in our special view M is a constant once H is fixed.

As a statement for real-coefficient multi-variable polynomial identity we find it holds once it holds on a dense subset of the whole space.

Since H has distinct roots $z_1, \ldots, z_k$ where $k = \deg H$, we consider the choose of G that makes $\dfrac{G(z_i)}{H'(z_i)}$'s all distinct. Since M is fixed, both sides is of degree $\le k$ for variable c and same non-repeating roots with a total number of k exist and exhaust on both sides, we find it holds naturally, and meanwhile $M \ne 0$(for any given H in our condition) is proved.

Now we only need to proof the choose of G is dense. The linear map from coefficients of G as $(g_0, \ldots, g_{k-1})$ (Assume freely chosen in degree $0, \ldots, k-1$) to $(G(z_1), \ldots, G(z_k))$ is bijective according to Vandermond Determinant Theorem, so in any open ball in $\mathbb{R}^k$ a choose of $(g_0, \ldots, g_{k-1})$ that makes $\dfrac{G(z_i)}{H'(z_i)}$'s all distinct exists, for $H'(z_i)$'s are non-zero constants here. □

Corollary 2.2.8 The multiplicity of root c in $\Re(c)$ equals $\deg(\gcd(H, G - cH'))$.

By Corollary 2.2.8, we are able to extract all residues and its multiplicities totally into roots and its multiplicities of $\Re(c)$. We first compute $\Re(c) = \prod\limits_{i=1}^{k} U_i(c)^i$ by Square-free Decomposition and for roots of U_i we know their multiplicities are i. Notice that when computing the resultant on field $\mathbb{R}(y)$ by Monic EEA, we solve calculating the log terms simultaneously.

One last problem is left: there may be complicated complex log terms that needs combination, and we may first think the relation $i(\log(\dfrac{A+Bi}{A-Bi})) = 2\arctan(\dfrac{A}{B})$ suffice, as in PFD Method. However, this relationship holds only in the same analytic branch and in this case singularities of B may make this fail. Suppose we want to calculate $\displaystyle\int_1^2 \frac{x^4 - 3x^2 + 6}{x^6 - 5x^4 + 5x^2 + 4}$ for example, the indefinite integral will be printed as $\arctan(\dfrac{x^3 - 3x}{x^2 - 2})$, which is first-type discontinuous at $\sqrt{2}$ and thus impossible. In order to avoid singularities we must introduce new methods for combining log terms.

Lemma 2.2.9 Suppose $A, B, C, D \in \mathbb{R}[x]$, $A, B, C, D \neq 0$, $BD - AC = G = \gcd(A, B)$, then

$$\frac{d}{dx}\log(\frac{A+Bi}{A-Bi}) = \frac{d}{dx}\log(\frac{-B+Ai}{-B-Ai}), \quad i\frac{d}{dx}\log(\frac{A+Bi}{A-Bi}) = 2\frac{d}{dx}\arctan(\frac{AD+BC}{G}) + i\frac{d}{dx}\log(\frac{D+Ci}{D-Ci}).$$

Proof: Denote $\dfrac{AD+BC}{G}$ by P. Notice that

$$-\frac{-B+Ai}{-B-Ai} = \frac{A+Bi}{A-Bi} = (\frac{P+i}{P-i})(\frac{D+Ci}{D-Ci}), \quad i\frac{d}{dx}(\log(\frac{P+i}{P-i})) = 2\frac{d}{dx}\arctan P,$$

the statement is trivial. □

from lemma and EEA we can develop a singularity removal method:

Algorithm 2.2.10 (Rioboo's Method) We start from $i\dfrac{d}{dx}\log(\dfrac{A+Bi}{A-Bi})$, and according to Lemma 2.2.9 we suppose $\deg A \geq \deg B > \deg G$ (if $\deg B = \deg G$, then this term can be directly transformed into a continuous arctan term.)

By EEA, we find C, D s.t. $BD-AC = G$(and all $\neq 0$ according to $\deg A \geq \deg B > \deg G$), and $\max\{\deg C, \deg D\} < \max\{\deg A, \deg B\}$. Again by Lemma 2.2.9 we only need to compute on

$$2\arctan(\frac{AD+BC}{G}) + i\frac{d}{dx}\log(\frac{D+Ci}{D-Ci}),$$

and this process finishes in at most $\max\{\deg A, \deg B\} + 1$ steps.

Therefore we build a new algorithm only making use of Square-free Decomposition and EEA:

Algorithm 2.2.11 (Lazard-Rioboo-Trager, LRT) Compute $\displaystyle\int \frac{G}{H}$ in the following steps:

Step 1: Compute Rothstein-Trager resultant $\Re(c)$ in way of Theorem 2.2.6 and store all $R_i(x, y)$'s;

Step 2: Make Square-free Decomposition of $\Re(c)$ into $\displaystyle\prod_{j=1}^{k} U_j(c)^j$, find roots of U_j's and follow the following formula:

$$\int \frac{G}{H} = \sum_{j=1}^{m} \sum_{c:U_j(c)=0} c\log(R_{i(j)}(x, c));$$

Step 3: Similarly, by combining conjugate residues and log terms, and transforming to arctan terms for $\log(X + iY)$ term by Rioboo's Method, we manage to express the final result.

3. Numeric-Symbolic Methods and Error Estimation

3.1 Introduction: Difficulties in Introducing Numerical Methods

Now that we have got the brief basic knowledge about symbolic algorithms like LRT and PFD Method, we find that the procedure of the whole integration can be completely changed when numeric approximation of coefficients and roots are introduced:

(1) **For coefficients:** In the context of rational coefficients or coefficients in fine algebraic express ($\mathbb{Q}(\sqrt{2})$ for example), Coefficients need no perturbation; while in more complex cases perturbing coefficients into floats may be necessary. In this case the multiplicity of square-free decomposition results, as an 'isolated', 'zero-measure' property in space of polynomials may be destroyed;

(2) **For roots:** Even in situations of integer coefficients roots may be unable to solve in radicals. PFD requires rootfinding of the denominator which is likely to destruct conjugate structures for roots(less likely to happen) and repeated residues. LRT require rootfinding of resultant which may result in destroying multiplicities too.

To solve Difficulty 1, a possible method is introducing **approximated symbolic algorithms**, such as finding an approximate GCD by an approximate EEA, or doing approximate Square-free Decomposition in order to detect nearby problems in increased structure. (The increase in multiplicity means the non-commonness of the problem also increased, together with the efficiency of algorithms.) This also avoids *ill-conditioning* roots (within a distance too small) caused by perturbing singular inputs to a nearby float-coefficient input, and is a kind of **Singular-Problem Detection** in its nature. We may also apply this in the way which we introduce later, yet this means the actual integration solves the singular problem instead of the input itself.

Let's analyze this in detail. Approximate algorithms are only able to apply in *parts possible to be viewed 'singular'*, so it is applied in approximately but not square-free parts, assumed in denominators of the original function, which leads the Hermite part in approximated sense. Suppose the input and the singular problem are

$$f = \frac{C}{D} + \int \frac{G}{H} \text{ and } [f] = [\frac{C}{D}] + [\int \frac{G}{H}],$$

then we computed the rational part as $[\frac{C}{D}]$ and we must respectively replace the denominator of f into approximated result to simplify the computed part left, which actually turns f into $[f]$. So the rest part we operate on is in fact $[\int \frac{G}{H}]$ instead of $\int \frac{G}{H}$. Therefore all the steps left will work on the singular problem; when two numeric-symbolic methods on the following transcendental step has any error, the error is in fact *between the singular one and the computed one*, instead of the input and the computed one!

In this way, the nearby most singular problem is successfully detected and assuming we aim at getting closer to this singular problem, we only need to analyze the error as in accurate-input situations, and suppose the rational part is *exact*. we change the nearly-square parts of the factorization of ill-conditioned denominator into square parts, and the Hermite algorithm actually printed can be seen as the *exact rational part for a nearby singular problem* .Hence the difficulties of inaccurate coefficients (in the sense of inaccuracy to possible singular problems) are solved.

On the contrary, even when no singular problems are near, inaccurate input may be inevitable, yet we can still control its influence. Besides, difficulty 2 seems also inevitable especially in more complicated situations of given denominators. Although the error may be unavoidable in complicated situations, we still allow its appearance since in realistic models *complicated rational integrals are approximations to actual systems as well.* Once we manage to control error in integration sufficiently, it can be made relatively smaller than model approximation error.

On this thesis we immediately find two methods available: First, we can also assume the initial input to be $f(x) \in \mathbb{Q}(x)$ since this is viewed as another approximation of complicated possible coefficients that able to be in any demanded precision. We may always assume this: when no singular problems are near, we assume the initial input in $\mathbb{Q}(x)$ is already *exact* for convenience in error analysis, and in this context purely symbolic algorithms besides numerical rootfinding is available on computers. Difficulty 1 is totally solved in this assumption. Second, we may develop formulas to make the error computable by the results and variables produced in numreic-symbolic algorithms. By controlling error by computing it and decreasing the initial tolerance when it fails to reach our expectation, we solve difficulty 2 as well.

Both the distance to the original problem, and the precision we've given our final result is important enough. Now suppose the initial input is $f(x) \in \mathbb{Q}(x)$, and due to algorithms we actually compute the integral of $\hat{f}$. We will provide two kinds of error analysis(where $\Delta f = f - \hat{f}$):

(1) **Backward Error:** Defined by $BE(f) = ||\Delta f||_\infty = \max\limits_{[a,b]} |\Delta f|$, representing the distance between what we actually compute to the original problem; A computable backward error means we can control it to be smaller than the error of approximation of model.

(2) **Forward Error:** Defined by $FE(f) = |\int_a^b \Delta f|$, representing the difference in actual result of a definite integral, and $[a, b]$ should be chosen sufficiently far away from roots of denominator H. A computable forward error means we can control the result of definite integral as precise as besides distance to the original problem,

In the following subsections we will show both of them are computable and approach 0 by $O(\epsilon)$, which means precisely controlling error by decreasing tolerance ϵ is possible. Then we can add the procedure of **Error Computation**. We have a powerful tool in MAPLE computer algebra system to find roots in any demanded error tolerance ϵ, so the error can be made small enough as well if we decrease our initial tolerance.

Generally, by *Singular-Problem Detection* and *Error Computation*, we are able to maximize our efficiency.

3.2 Procedures of N-PFD and N-LRT

Now we introduce symbolic-numeric methods based on classical symbolic methods (PFD and LRT), denoted by **N-PFD** and **N-LRT**:

Algorithm 3.2.1 An **N-PFD** procedure is based on a PFD procedure, but different in:

(1)Use numerical rootfinding and residue-computation instead of symbolical;

(2)Detect conjugate pairs and identical residues (may be used to combine terms) by a user-supplied tolerance e_c, in order to revive structure.

Algorithm 3.2.2 An **N-LRT** procedure is based on a LRT procedure, but different in:

(1)Use numerical rootfinding towards all U_i's;

(2)Detect identical residues and conjugate pairs (may be used to combine terms) by a user-supplied tolerance e_c, in order to revive structure before using Rioboo's Method.

In both procedures, the input is a rational function $f(x) = \dfrac{P(x)}{Q(x)}$ (over $\mathbb{Q}$), together with necessary tolerance settings, for example a tolerance ϵ for relative error in rootfinding. and in the end we actually output an expression of a nearby integrand:

$$\int \hat{f}(x)dx = \frac{C}{D} + \sum v_i \log(V_i) + \sum w_j \arctan(W_j),$$

along with an (linear) error estimation.

3.3 Error Estimations and Linear Error Computations

We consider the backward and forward errors based on N-LRT and N-PFD.

Theorem 3.3.1(Computable Backward Error) Suppose the input tolerance is ϵ (relatively, which means for any root r, $|\Delta r| \leq \epsilon |r|$). Then

$$BE(f) = max_x |\sum_k Re(M(x, r_k))| + O(\epsilon^2),$$

where the principle term is $O(\epsilon)$ and r_k range in the evaluated roots (of different polynomials in different algorithms), and M is computable. The bound is finite on any closed bounded interval without any root of $Q(x)$.

Proof for N-PFD: Since the rational part is computed exactly we need only consider the transcendental part.(This also happens in the rest error analysis.) Suppose $\dfrac{G(x)}{H(x)} = \sum_{i=1}^{\deg H} \dfrac{c_i}{x - \gamma_i}$, then we actually approximated c_i

and γ_i to $\hat{c_i}$ and $\hat{\gamma_i}$. Notice that

$$\begin{aligned}
\frac{c_i}{x-\gamma_i} - \frac{\hat{c_i}}{x-\hat{\gamma_i}} &= \frac{\Delta c_i(x-\hat{\gamma_i}) + \hat{c_i}\Delta\gamma_i}{(x-\gamma_i)(x-\hat{\gamma_i})} + O(\Delta\gamma_i^2)\\
&= \frac{c'(\gamma_i)\Delta\gamma_i}{x-\hat{\gamma_i}-\Delta\gamma_i} + \frac{\hat{c_i}\Delta\gamma_i}{(x-\hat{\gamma_i})(x-\hat{\gamma_i}-\Delta\gamma_i)} + O(\Delta\gamma_i^2)\\
&= \frac{c'(\gamma_i)\Delta\gamma_i}{x-\hat{\gamma_i}} + \frac{\hat{c_i}\Delta\gamma_i}{(x-\hat{\gamma_i})^2} + O(\Delta\gamma_i^2),\\
0 = H(\gamma_i) &= H(\hat{\gamma_i}) + H'(\hat{\gamma_i})\Delta\gamma_i + O(\Delta\gamma_i^2)\\
\Longrightarrow \Delta\gamma_i &= -\frac{H(\hat{\gamma_i})}{H'(\hat{\gamma_i})},\ \text{and } |\Delta\gamma_i| \le \epsilon|\gamma_i|,\ 1 \le i \le \deg H,
\end{aligned}$$

We have the obvious estimation $M(x,r) = (\frac{c'(r)}{x-r} + \frac{c(r)}{(x-r)^2})\frac{H(r)}{H'(r)}$ as an $O(\epsilon)$ term. □

Proof for N-LRT: Since $\int \frac{G}{H} = \sum_{j=1}^{m} \sum_{c:U_j(c)=0} c\log(R_{i(j)}(x,c)) \Longrightarrow \frac{G}{H} = \sum_{j=1}^{m} \sum_{c:U_j(c)=0} c\frac{\frac{\partial R_{i(j)}}{\partial x}(x,c)}{R_{i(j)}(x,c)}$ and $c_{j,k}(1 \le j \le m, 1 \le k \le \deg U_j)$'s are perturbed, suppose into $c_{\widehat{j,k}}$. Given that by partial derivatives of $c_{j,k}$ we know

$$\begin{aligned}
\Delta f(x, c_{1,1}, \ldots, c_{m,\deg U_m}) &= \sum_{j=1}^{m}(\sum_{k=1}^{\deg U_j}(\frac{\frac{\partial R_{i(j)}}{\partial x}(x,c)}{R_{i(j)}(x,c)} + c(\frac{\frac{\partial^2 R_{i(j)}}{\partial x\partial c}(x,c)}{R_{i(j)}(x,c)} - \frac{\frac{\partial R_{i(j)}}{\partial x}(x,c)\frac{\partial R_{i(j)}}{\partial c}(x,c)}{R_{i(j)}(x,c)^2}))|_{c=c_{\widehat{j,k}}}\Delta c_{j,k}) + O(||\Delta c||^2)\\
&\triangleq \sum_{j=1}^{m}(\sum_{k=1}^{\deg U_j}(\xi_{i(j)}(c_{\widehat{j,k}}, x)\Delta c_{j,k})) + O(||\Delta c||^2),
\end{aligned}$$

and similarly noticing $\Delta c_{j,k} = -\frac{U_j(c_{\widehat{j,k}})}{U_j'(c_{\widehat{j,k}})}$ and $\Delta c_{j,k} \le \epsilon|c_{\widehat{j,k}}|$, we have the obvious estimation $M(x,r) = \xi_{i(j)}(x,r)\frac{U_j(r)}{U_j'(r)}$($j$ determined by $r = c_{j,k}$) as an $O(\epsilon)$ term. (In both algorithms the bounded statement is trivial.)□

Theorem 3.3.2(Computable Forward Error) Suppose the input tolerance is ϵ (relatively, which means for any root r, $|\Delta r| \le \epsilon|r|$). Then

$$FE(f) = max_x|\sum_k M(r_k, s_k, x)| + O(\epsilon^2),$$

where the principle term is $O(\epsilon)$ and $r_k + is_k$ ($r_k, s_k \in \mathbb{R}$) range in the evaluated roots (of different polynomials in different algorithms), and M is computable. The bound is finite on any closed bounded interval without any root of $Q(x)$.

Proof for N-LRT: Notice that when any function g has a primitive g^*, we have $|\int_a^b g(x)dx| \le |g^*(a)| + |g^*(b)|$, so we only need to bound a primitive of the error function in the form of our theorem. Suppose $c_k = a_k + ib_k$ ($a_k, b_k \in \mathbb{R}$, $1 \le k \le m$) is a subsequence of all residues that only contain one of each conjugate pair, and $R_j (j = j(k))$ is the corresponding subresultant of c_k. We are able to rewrite the transcendental part as

$$\int \frac{G}{H} = \sum_{k=1}^{m}(a_k\log(V_k) + 2b_k\arctan\frac{W_{1k}}{W_{2k}}),$$

where $W_{1k} = \mathrm{Re}(R_j(c_k,x))$, $W_{2k} = \mathrm{Im}(R_j(c_k,x))$ and $V_k = W_{1k}^2 + W_{2k}^2$ for complex roots (for real roots we have $V_k = W_{1k}$, since their conjugate are themselves), which means they together with their partial derivatives of

a_k and b_k are computable. In this case we can estimate the error by linear approximation:

$$\begin{aligned}\int \Delta f dx = &\sum_{k=1}^{m}(\log(V_k)\Delta a_k + (\frac{\partial V_k}{\partial a_k}\Delta a_k + \frac{\partial V_k}{\partial b_k}\Delta b_k)\frac{\hat{a_k}+\Delta a_k}{V_k} \\ &+ ((W_{2k}\frac{\partial W_{1k}}{\partial a_k} - W_{1k}\frac{\partial W_{2k}}{\partial a_k})\Delta a_k + (W_{2k}\frac{\partial W_{1k}}{\partial b_k} - W_{1k}\frac{\partial W_{2k}}{\partial b_k})\Delta b_k)\frac{2(\hat{b_k}+\Delta b_k)}{W_{1k}^2+W_{2k}^2}) \\ &+ O(||\Delta c||^2),\end{aligned}$$

and by $-\frac{U_j(\hat{c_k})}{U_j'(\hat{c_k})} = \Delta c_k = \Delta a_k + i\Delta b_k$ we can change the $\frac{\hat{a_k}+\Delta a_k}{V_k}$ and $\frac{2(\hat{b_k}+\Delta b_k)}{W_{1k}^2+W_{2k}^2}$ terms into $\frac{\hat{a_k}}{V_k}$ and $\frac{2\hat{b_k}}{W_{1k}^2+W_{2k}^2}$ terms first to extract the $O(||\Delta c||^2)$ terms, and then replace the Δa_k's and Δb_k's in linear terms. Therefore the functions $M(r_k, s_k, x)$ of size $O(\epsilon)$ are computable. □

Proof for N-PFD: Suppose $\gamma_k = \alpha_k + i\beta_k$ ($\alpha_k, \beta_k \in \mathbb{R}$, $1 \leq k \leq m$) is a subsequence of all roots of H that only contain one of each conjugate pair, and $c(\gamma_k) = c_k = a_k + ib_k$ ($a_k, b_k \in \mathbb{R}$, $1 \leq k \leq m$). Following the notations above we let $W_{1k} = x - \alpha_k$, $W_{2k} = -\beta_k$, and $V_k = W_{1k}^2 + W_{2k}^2$ for complex roots (for real roots we have $V_k = W_{1k}$, since their conjugate are themselves), which means they together with their partial derivatives of α_k and β_k are computable. Again we estimate the error by linear approximation:

$$\begin{aligned}\int \Delta f dx = &\sum_{k=1}^{m}(\log(V_k)(\frac{\partial a_k}{\partial \alpha_k}\Delta\alpha_k + \frac{\partial a_k}{\partial \beta_k}\Delta\beta_k) + (\frac{\partial V_k}{\partial \alpha_k}\Delta\alpha_k + \frac{\partial V_k}{\partial \beta_k}\Delta\beta_k)\frac{a_k}{V_k} \\ &+ \frac{2\beta_k b_k(\Delta\alpha_k + \Delta\beta_k)}{(\alpha_k - x)^2 + \beta_k^2} + 2(\frac{\partial b_k}{\partial \alpha_k}\Delta\alpha_k + \frac{\partial b_k}{\partial \beta_k}\Delta\beta_k)\arctan(\alpha_k - x, \beta_k)) \\ &+ O(||\Delta c||^2),\end{aligned}$$

since linear estimations multiplied by $\Delta\alpha_k$ or $\Delta\beta_k$ we can replace the c_k, α_k, β_k's appeared by their approximation $\hat{c_k}, \hat{\alpha_k}, \hat{\beta_k}$'s, and notice $-\frac{H(\hat{\gamma_k})}{H'(\hat{\gamma_k})} = \Delta\gamma_k = \Delta\alpha_k + i\Delta\beta_k$, we similarly replace $\Delta\alpha_k$'s and $\Delta\beta_k$'s and find the result $M(r_k, s_k, x)$ of size $O(\epsilon)$ computable. □

By proof of Theorem 3.3.2 we directly get the corollary:

Corollary 3.3.3 The error term M in Theorem 3.3.2 can be decomposed into M_1 and M_2 corresponding to log and arctan terms respectively by the form given in proof; and both of them are $O(\epsilon)$.

4. Experiments and Comparison: Which One is Better?

4.1 Analysis and Experiment Settings

To decide which one generally performs better, we have many dimensions in consideration, like efficiency, stability towards singularities or precise ill-conditioned problems and so on.

Intuitively, The runtime and instability of N-PFD is generally decided by the degree of the denominator and heights of the coefficients in polynomial; Yet the runtime and instability of N-LRT can be different exponentially under the same degree and heights due to instablity of determinants. So N-PFD may actually be more likely to perform better.

In order to test the actual performance of two algorithms, we select 5 experiments to compare their behaviour, with the assumption that input is precise enough that approximate GCD don't work. They test two algorithms in different senses.

Experiment 4.1.1 $f_1(x) = \dfrac{1}{x^n - 2}$;

Experiment 4.1.1' $f_1(x) = \dfrac{1}{x^n + 2}$;

Experiment 4.1.2 $f_2(x) = \dfrac{1}{x^n + x - 2}$;

Experiment 4.1.3 $f_3(x) = [n, n]_{Ei(x)}(x)$, where $[m, n]_u(x)$ denotes the Padé approximation of order $[m/n]$ of u, and $Ei(x) = \dfrac{e^x}{x}$;

Experiment 4.1.4 $f_4(x) = \dfrac{2x}{x^2 - (1+t)^2}$ $(t \longrightarrow 0)$;

Experiment 4.1.5 $f_5(x) = \dfrac{2x}{x^2 + t^2}$ $(t \longrightarrow 0)$;

The first two experiments serve as a basic comparison of their runtime and stability in same size of denominator; The third experiment tests their stability of performance towards large coefficients (for moderate n the coefficients are already large); The fourth experiment tests when singularities are just outside $[-1, 1]$, how near can we get the interval bounds near singularities before the error exceeds a fixed tolerance radius(like 0.01). We *assume* the fixed radius as its original meaning because this is not clearly illustrated in the context of the article: once we assume the tolerance is some multiple of ϵ, when ϵ sufficiently small the tolerated interval would be somehow fixed since the error term is $O(\epsilon)$ and the constants multiplied is determined by x (in the sense of ignoring the $o(\epsilon)$ differences); this problems remains even when we assume the error to be relative. The last experiment tests how two algorithms perform when the integrated function has nearly real singularities on the imaginary axis.

We also note that the interval of integration may be set to $\mathbb{R}$, since the error analysis only requires finding the maximum value of an error function on the integration zone, which is accessible.

4.2 Comparison of Runtime

For Experiment 4.1.1 and 4.1.2, we select n from 40 logarithmically spaced values from 8 to 377, and we find the following results:

(1)The runtime (when error analysis is open) of two algorithms are shown in the figures below.

We see that in the same size of denominator, the N-LRT Method behaves considerably poorer in Experiment 4.1.2. In Experiment 4.1.1, the performance in runtime of two algorithms are near, but in Experiment 4.1.2 N-PFD is clearly the winner with a runtime nearly identical to Experiment 4.1.1. This means the coefficients of resultant truly get considerably larger as the problem are perturbed a bit, and the N-LRT can be really slowed down in experiments because of this.

Also we find that N-LRT has a more unpredictable efficiency in error analysis, because when we turn off the error analysis the runtime of N-PFD don't change apparently while for N-LRT it changed apparently in Experiment 4.1.1 (not in 4.1.2). Given that error is usually several magnitudes smaller than tolerance, when we turn off the error analysis N-LRT can win in Experiment 4.1.1, which means there are certain problems for N-LRT to behave better. But this also show that error analysis bring more and unpredictable influence in runtime in N-LRT.

And for Experiment 4.1.3 the N-PFD Algorithm completely win in speed, which is shown in the following table:

which means, considering the robustness of efficiency towards high-coefficient problems, N-PFD performs better as well.

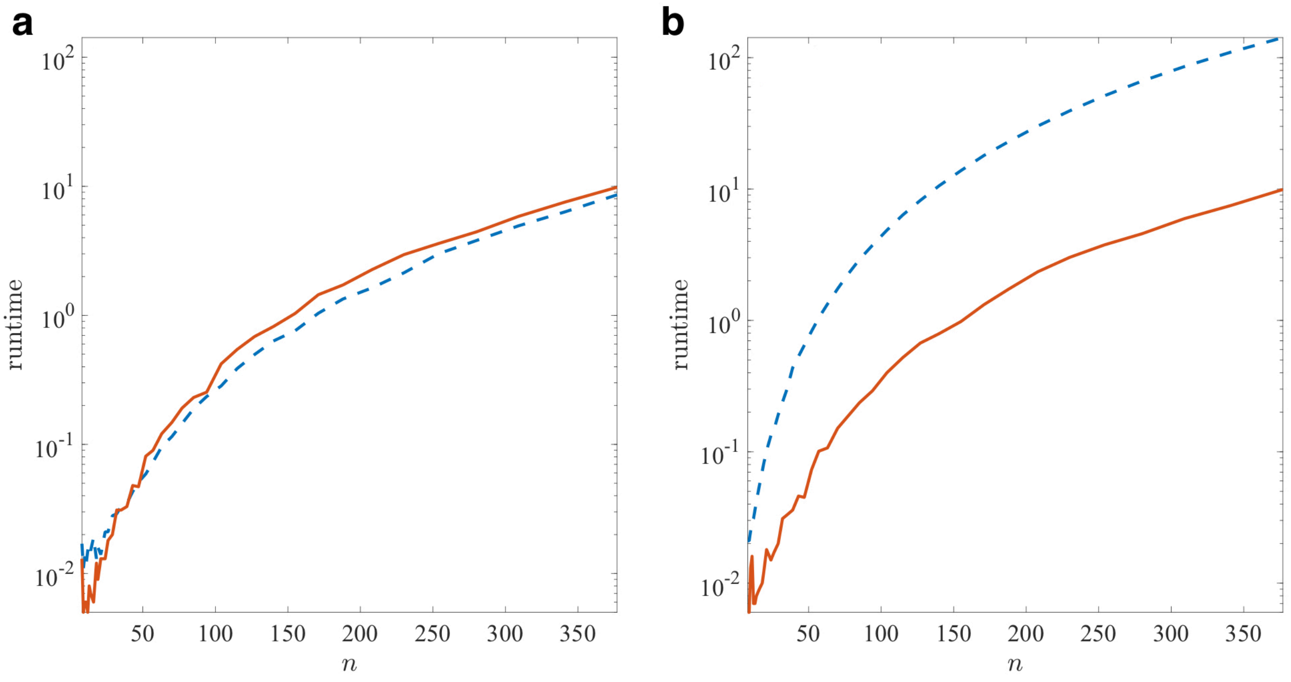


Fig. 1: Different Performances of Different Algorithms (Left: f_1, Right: f_2; Red: N-PFD, Blue: N-LRT)

Runtimes	N-PFD	N-LRT
n=8	0.01s	0.04s
n=13	0.02s	0.18s
n=21	0.04s	2.5s

Tab. 1: Runtime For Two Algorithms in Different Cases

4.3 Comparison of Precision and Singularity-Stability

For problems free from singularities, we set the integrating interval as $\mathbb{R}$ and compute the forward and backward errors. Take $n = 128$ as an example in Experiment 4.1.1', as we reduce the tolerance we get the following results, showing both algorithms perform strongly although N-LRT performs better in several magnitudes.

ϵ	FE of N-LRT	BE of N-LRT	FE of N-PFD	BE of N-PFD
2^{-34}	$8 \cdot 10^{-16}$	$1 \cdot 10^{-15}$	$2 \cdot 10^{-15}$	$2 \cdot 10^{-12}$
2^{-55}	$3 \cdot 10^{-55}$	$2 \cdot 10^{-53}$	$1 \cdot 10^{-39}$	$1 \cdot 10^{-38}$
2^{-89}	$1 \cdot 10^{-75}$	$2 \cdot 10^{-73}$	$9 \cdot 10^{-59}$	$8 \cdot 10^{-58}$
2^{-144}	$6 \cdot 10^{-95}$	$7 \cdot 10^{-93}$	$3 \cdot 10^{-77}$	$2 \cdot 10^{-76}$

Tab. 2: Forward and Backward Errors For Two Algorithms in Experiment 4.1.1', $n = 128$

As for singularity problems like Experiment 4.1.3 (when $n = 8$, a singularity at 10.949), we test the minimum possible width of a symmetric interval around the singularity before the error exceeds the fixed tolerance. The results in article is follows:

this shows that even though N-LRT performs better on ϵ sufficiently small, the decreasing of width for N-LRT is

ϵ	FEW of N-LRT	BEW of N-LRT	FEW of N-PFD	BEW of N-PFD
2^{-34}	$4 \cdot 10^{-3}$	$6 \cdot 10^{-2}$	$1 \cdot 10^{-14}$	$6 \cdot 10^{-7}$
2^{-55}	$7 \cdot 10^{-23}$	$9 \cdot 10^{-12}$	$8 \cdot 10^{-20}$	$3 \cdot 10^{-10}$
2^{-89}	$4 \cdot 10^{-32}$	$2 \cdot 10^{-16}$	$2 \cdot 10^{-28}$	$1 \cdot 10^{-14}$
2^{-144}	$2 \cdot 10^{-34}$	$1 \cdot 10^{-17}$	$3 \cdot 10^{-31}$	$6 \cdot 10^{-16}$

Tab. 3: Widths For Acceptable Forward and Backward Errors For Two Algorithms in Experiment 4.1.3, $n = 8$

not as stable and predictable as N-PFD, and both algorithms perform well with only slight differences. Furthermore in the case with a singularity the error difference is smaller. Also, the singularity-stability for both algorithms are already enough: The tolerated width can be made sufficiently small as ϵ decreases, and when really so close to the singularity as the magnitude shown, the Padé approximation of $Ei(x)$ is no longer a good one, threrfore this width is not a true concern in application.

Therefore, although N-LRT may perform better in numerical precision of the result, no significant advantage is exhibited towards N-PFD.

Two additional tests in Experiment 4.1.4 and 4.1.5 also show the excellent precision and singularity-stability with little difference. With sufficiently small ϵ in Experiment 4.1.4, the tolerated width is even smaller than 2ϵ (ϵ in single side), which means even wih $t = \epsilon$ the definite integral on $[1, 1]$ is able to be computed. Also in Experiment 4.1.5, with the same input tolerances (The fixed radius default, and the ϵ is 2^{-53}(also a default value)), the forward error bound of two algorithms are $1.9 \cdot 10^{-57}$ for $t = 0.1$ and only increased to $1.7 \cdot 10^{-42}$ for $t = 10^{-16}$. And also two algorithms behave in little difference in two examples on numerical precision and singular stability.

4.4 Conclusion

In general N-PFD has **more efficient and more predictable behaviour** than N-LRT and therefore better. In particular problems where exact integrals are needed N-LRT may be suitable, but as for precision there is little advantage for N-LRT compared with N-PFD.

5. Future Works

Although two algorithms are already well enough, there are still particular special problems to be solved. One example is taking better identical residue detection, such as when the detection tolerance is $\epsilon = \epsilon_d$, several choices of coalescing occur in integration of

$$g(x) = \sum_{i=1}^{4} \frac{1 + (i - \frac{5}{2})\epsilon}{x - \alpha_i},$$

where coalescing all residues and two groups of residues give two different expressions that may behave differently near singularities. So a proper method to decide the way of coalescing according to implementation may be developed in future works.

6. References

[1]Symbolic-Numeric Integration of Rational Functions, Robert H.C. Moir , Robert M. Corless , Marc Moreno Maza , Ning Xie, 2019.

[2]Mathematical Theory of Computer Algebra System, Chao Li, Wei Ruan, Long Zhang, Xiang Zhang, 2010.

An Introduction to Fast Fourier Transform

Weitao Wang

Abstract

The report will introduce the concept of Discrete Fourier Transform (DFT) and several applications. Next, we will focus on Fast Fourier Transform (FFT), an effieicent algorithm to compute DFT. We will give detailed description of the algorithm and analysis of its complexity. It will be shown that the complexity to calculate the DFT of an N-array is $N \log N$ with this algorithm, compared to N^2 by direct calculation. Results of numerical experiment will be provided.

1 Introduction to DFT

As FFT is an algorithm to compute DFT, we introduce the concept, properties and applications of DFT at first, so that we can learn the motivation to develop the algorithm.

1.1 Discrete Fourier Transform (DFT)

The Discrete Fourier Transform (DFT)is a discrete version of continuous Fourier Transform. Given the function f defined on finite points: $0, 1, 2, ..., N-1$, the DFT of f is also a function on $\{0, 1, 2, ..., N-1\}$, which is defined as:

$$\hat{f}(j) = \sum_{0}^{N-1} f(k)\omega^{-jk}$$

, whereω is the principal N th root of unity

$$\omega = e^{\frac{2\pi i}{N}}$$

We can observe that DFT is an approximation to continuous Fourier Transform to some extent, actually, let g a complex-valued function defined on $\mathbb{R}$ and has support on $[0, 1]$, and

$$f = g \quad on \quad 0, 1, 2, ..., N-1$$

, then when N is large enough, we have:

$$\begin{aligned} \hat{f}_{continuous}(j) &= \int_0^1 e^{-2\pi i x j} f(x)\ dx \\ &\approx \frac{1}{N} \sum_{k=0} N - 1 e^{\frac{-2\pi i k j}{N}} \\ &= \frac{1}{N} \hat{f}_{DFT}(j) \end{aligned}$$

Therefore it's not hard to believe that the DFT and its calculation is of some significance in many fields across science and engineering. For the discrete nature of computer, we have to use DFT when we need numerical results of Fourier Transform.

1.2 Properties of DFT

1.2.1 Inverse Transform (IDFT)

$$f(n) = \frac{1}{N} \sum_{k=0}^{N-1} \hat{f}(k) e^{i \frac{2\pi k}{N} n}$$

The proof is omitted.

1.2.2 Parseval Theorem

$$\sum_{n=0}^{N-1} f_n g_n^* = \frac{1}{N}\sum_{k=0}^{N-1} \bar{f}_n \bar{g}_n^*$$

The proof is omitted.

1.3 Applications of DFT

1.3.1 Signal Processing

In the field of signal processing, though continuous signals are often studied, in the real world we often deal with digital signals, which are naturally discrete. For Fourier Transform is often applied in signal processing, DFT also plays an import role. Besides, DFT can be considered as a sampling of Discrete Time Fourier Transform (DTFT), which is actually a continuous signal.Oppenheim et al., 1997

1.3.2 In Number Theory

Amazingly, DFT has an application in number theory. The discrete version of Fourier Inversion formula is a key step in the proof of Dirichlet's theorem Stein and Shakarchi, 2011, which states that if q and l are positive integers wit no common factor, then the progression

$$l,\ l+q,\ l+2q,\cdots,\ l+kq,\cdots$$

contains infinite prime numbers.

2 Fast Fourier Transform

2.1 Description of the Algorithm

By taking advantage of the periodicity of $\{1, \omega, \omega^2, ..., \omega^{N-1}$, we can get an algorithm to compute DFT of $f(x)$ defined on $\{0, 1, 2, ..., N-1\}$ with complexity of $N \log N$, compared with N^2 by direct calculation. Below we describe the algorithm.

As we showed before, DFT could be considered as an approximation to continuous Fourier Transform, or a sampling of DTFT in signal processing. It is reasonable to argue that as long as we can get N large enough, we can choose whatever form of N we like. It means, N could be chosen as a prime number, as well as the form of r^m. We will see that there is real gain to choose N as a highly composite number. For simplicity, we suppose

$$N = r_1 r_2$$

to describe the algorithm, where r_1 and r_2 is not required to be prime number. We will see that the result and procedure where

$$N = r_1 r_2 ... r_m$$

is intrinsically the same when we learn the case that $N = r_1 r_2$.

Recalling the formula for DFT:

$$\hat{f}(j) = \sum_{0}^{N-1} f(k)\omega^{-jk}$$

As $N = r_1 r_2$, we can re-index as follows

$$j = j_1 r_1 + j_0, \quad j_0 = 0, 1, \cdots, r_1 - 1, \quad j_1 = 0, 1, \cdots, r_2 - 1$$

$$k = k_1 r_2 + k_0, \quad k_0 = 0, 1, \cdots, r_2 - 1, \quad k_1 = 0, 1, \cdots, r_1 - 1$$

Then, we can write:

$$\hat{f}(j) = \sum_{k_0}\sum_{k_1} f(k_1 r_2 + k_0)\omega^{jk_1 r_2}\omega^{jk_0} \tag{1}$$

since we have by periodicity of ω:

$$\omega^{jk_1 r_2} = \omega^{(j_0 + j_1 r_1)(k_1 r_2)} = \omega^{j_0 k_1 r_2}$$

we can write 1 as:

$$\hat{f}(j) = \sum_{k_0} \omega^{jk_0} \sum_{k_1} f(k_1 r_2 + k_0)\omega^{j_0 k_1 r_2} \tag{2}$$

And in 2 the inner sum $\sum_{k_1} f(k_1 r_2 + k_0)\omega^{j_0 k_1 r_2}$ depends only on j_0 and k_0. Therefore we can define f_1 as:

$$f_1(j_0, k_0) = \sum_{k_1} f(k_1 r_2 + k_0)\omega^{j_0 k_1 r_2} \tag{3}$$

Then we can write:

$$\hat{f}(j_1, j_0) = \sum_{k_0} f_1(j_0, k_0)\omega^{j_1 r_1 + j_0 k_0} \tag{4}$$

We can see from 3 that it needs r_2 operations (An operation means a complex multiplication followed by a complex addition, so as the following.) to get $f_1(j_0, k_0)$. And after that, from 4 we can see that it needs r_1 operations to compute $\hat{f}(j_1, j_0)$. For there are N points in total where we need to compute the value of $\hat{f}$, this two-step algorithm described above needs

$$T = N(r_1 + r_2)$$

operations in total.
After learning the case when $N = r_1 r_2$, we consider the case that

$$N = r_1 r_2 \cdots r_m$$

Expressing the indices as follows:

$$j = j_0 + j_1 r_2 \cdots r_m + \cdots + j_{m-1} r_m$$

$$k = k_0 + k_1 r_{m-1} \cdots r_1 + \cdots + k_{m-1} r_1$$

Thus we have

$$\hat{f}(j_0, j_1, \cdots, j_{m-1}) = \sum_{k_0}\sum_{k_1}\cdots\sum_{k_{m-1}} f(k_0, k_1, \cdots, k_{m-1})\omega^{jk} \tag{5}$$

And we can separate 5 into m inner sums, as in 2, thus giving an m step algorithm, each step needing r_i operations respectively. Therefore the m-step algorithm requires

$$T = N(r_1 + r_2 + \cdots + r_m)$$

operations.

2.2 Analysis of Complexity

As be shown above, we have the theorem about operations needed by FFT algorithm to compute an N-point DFT. That is,

Theorem 2.1. *Suppose $N = r_1 r_2 \cdots r_m$, to compute the DFT of f, which is defined on $\{0, 1, 2, \cdots, N-1\}$, with FFT algorithm, requires*

$$T = N(r_1 + r_2 + \cdots + r_m)$$

operations.

With the theorem, we immediately have the corollary: Suppose $N = r^m$, the operations required $T(r)$ is:

$$T(r) = N \log_r N$$

Proof. By the theorem above, in this case, the total number of operations is

$$T = Nmr$$

And obviously

$$m = \log_r N$$

Thus we have

$$T(r) = rN \log_r N$$

□

r	$\frac{r}{\log_2 r}$
2	2.00
3	1.88
4	2.00
5	2.15
6	2.31
7	2.49
8	2.67
9	2.82
10	3.01

Table 1: r and $E(r)$

Sampling Rate (N)	Time of DFT	Time of FFT
2048	0.28	0.02
4096	1.63	0.05
8192	4.47	0.10
16384	22.23	0.30

Table 2: Comparison between direct computation and FFT

Corollary 2.2. *Suppose $N = r^m$, the complexity of FFT is $O(N \log N)$.*

Proof. By 2.2, operations required is $rN \log_r N$, hence we know the complexity is $O(N \log N)$ □

The case that $N = r^m$ is important because such indices enable a simple realization of the algorithm. Now we consider the efficiency with different choices of r.
To compare the efficiency of different choice of r, we define the quantity E:

Definition 2.3.

$$E(r) = \frac{T(r)}{N \log_2 N}$$

where $T(r)$ is normalized by $N \log_2 N$, thus representing the operations required for the same number of N.
By 2.2, we have:

$$\begin{aligned} T(r) &= \frac{rN \log_r N}{N \log_2 N} \qquad (6) \\ &= \frac{r}{\log_2 r} \qquad (7) \end{aligned}$$

The value $E(r)$ for different r is listed below: (figures from Cooley and Tukey, 1965) It can be seen that among the integers, $r = 3$ is the most efficient. However, the gain is not significant (about 6%) compared to $r = 2$ and $r = 4$, and $r = 2$ or $r = 4$ offers important advantages for computation because of the binary arithmetic of computers. It is more efficient both in locating in the storage and in multiplication. Besides, even choosing $r = 10$ increases the computation no more than 50%, hence it's not absolutely unacceptable.

2.3 Results of Numerical Experiments

With a python program, we compare the time consumption of direct computation of DFT and FFT. To do this, we generate a simple 1-D signal and observe the time consumption to calculate its DFT with both methods while changing the sample rate (which is the N we discussed before). The code can be found in the appendix, whose functions are taken from Kong et al., 2020. And the results are in Table 2.

Appendix

The code for the numerical experiments is:

```
import matplotlib.pyplot as plt
import numpy as np
import timeit


def DFT(x):
    """
    Function to calculate the
    discrete Fourier Transform
    of a 1D real-valued signal x
    """

    N = len(x)
    n = np.arange(N)
    k = n.reshape((N, 1))
    e = np.exp(-2j * np.pi * k * n / N)

    X = np.dot(e, x)

    return X


def FFT(x):
    """
    A recursive implementation of
    the 1D Cooley-Tukey FFT, the
    input should have a length of
    power of 2.
    """
    N = len(x)

    if N == 1:
        return x
    else:
        X_even = FFT(x[::2])
        X_odd = FFT(x[1::2])
        factor = \
            np.exp(-2j * np.pi * np.arange(N) / N)

        X = np.concatenate(
            [X_even + factor[:int(N / 2)] * X_odd,
             X_even + factor[int(N / 2):] * X_odd])
        return X
def gen_sig(sr):
    ’’’
    function to generate
    a simple 1D signal with
    different sampling rate
    ’’’
    ts = 1.0/sr
    t = np.arange(0,1,ts)

    freq = 1.
    x = 3*np.sin(2*np.pi*freq*t)
    return x

# sampling rate
bei=2
sr = 2048*bei
print(sr)
array=gen_sig(sr)
print(timeit.timeit(’DFT(array)’,"from __main__ import DFT, array",number=1))
print(timeit.timeit(’FFT(array)’,"from __main__ import FFT, array",number=1))
```

Listing 1: Code for Numerical Experiments

References

Cooley, J. W., & Tukey, J. W. (1965). An algorithm for the machine calculation of complex fourier series. Mathematics of computation, 19(90), 297–301.

Kong, Q., Siauw, T., & Bayen, A. (2020). Python programming and numerical methods: A guide for engineers and scientis Academic Press.

Oppenheim, A. V., Willsky, A. S., Nawab, S. H., Hernández, G. M., et al. (1997). Signals & systems. Pearson Educación.

Stein, E. M., & Shakarchi, R. (2011). Fourier analysis: An introduction (Vol. 1). Princeton University Press.

Integral Transformation and its Applications in Signal Processing

Siheng.Liang

2022.6.12

13pt

Abstract

This article is an introduction to the integral transformation in signal processing.Including Discrete Fourier Transform(DFT),Fast Fourier Transform(FFT) and Hilbert Transform(HT).

1 Introduction

Begin with a real-valued continuous-time signal $x(t)$,in the mathematical sense we can use $continuous-time\ Fourier\ transform$ and convert it from the time domain to the frequency domain.And the signal defined in the frequency domain is complex symmetric. Thus, the negative frequency half of the signal spectrum contains redundant information with respect to the positive frequency half.So the analytic signal was created to remove this spectral redundacy by deleting the negative frequency half of the signal transform.The analytic signal has been demonstrated have lots of advantages compared with the original real-valued signal.So how to form analytic signal is very important.

1.1 Basic Definition

Definition 1.1 (continous-time signal)**.** We said $x(t)$ a continous-time signal while $x(t)$ is a continuous function from $\mathbb{R} \to \mathbb{R}$

Definition 1.2 (Sample rate)**.** Suppose we want to collect samples from a continuous signal,we choose an interval T and sampling at time interval T.Then we have a discrete-time signal $x(nT) = x(n)$ and we called $\frac{1}{T}$ the sample rate.

Definition 1.3 (Phase Factor)**.** We defined the phase factor as $W_N = e^{-i\frac{2\pi}{N}}$ and $W_N^k = e^{-i\frac{2\pi}{N}k}$

1.2 Integral Transformation

Definition 1.4 $(continuous - timeFouriertransform)$**.** Suppose $x(t)$ is a continuous-time signal and f is frequency,we define Fourier transform as following

$$\mathrm{X}(f) = \int_{-\infty}^{+\infty} x(t)e^{-ift}\mathrm{d}t \tag{1.1}$$

Definition 1.5 $(inverseFouriertransform)$**.** The inverse Fourier transform is defined as following

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \mathrm{X}(f)e^{ift}\mathrm{d}f \tag{1.2}$$

Definition 1.6 $(discrete - timeFouriertransform)$**.** Suppose we have a N-point discrete-time signal $\{x(n)\}$,its DFT is defined as following

$$\mathrm{X}(k) = \sum_{n=0}^{N-1} x(n)W_N^{kn}, \qquad k = 0, 1, \cdot\cdot\cdot, N-1 \tag{1.3}$$

Definition 1.7 $(inversediscrete - timeFouriertransform)$**.** The inverse transform of DFT is defined as following

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} \mathrm{X}(k)W_N^{kn}, \qquad n = 0, 1, \cdot\cdot, N-1 \tag{1.4}$$

Definition 1.8 $(continuous - timeHilberttransform)$**.** Suppose $x(t)$ is a continuous-time signal,its Hilbert transform is defined as following

$$x\hat{(t)} = \frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{x(\tau)}{t-\tau}\mathrm{d}\tau \tag{1.5}$$

2 Algorithm Optimization

Let's review the **DFT** algorithm,if we have a real-valued N-point discrete-time signal,We assume that $N = 2^v$ without losing generality.So we have **DFT** algorithm as following.

$$\mathrm{X}(k) = \sum_{n=0}^{N-1} x(n)W_N^{kn}, \qquad k = 0, 1, \cdot\cdot\cdot, N-1 \tag{2.1}$$

From (2.1) We can see that if we want to get $\mathrm{X}(k)$,we need

$$\mathrm{X}(k) = \sum_{n=0}^{N-1} \bullet \tag{2.2}$$

$N-1$ times addition,and

$$x(0)e^{-i\frac{2\pi}{N}0k}+x(1)e^{-i\frac{2\pi}{N}1k}+\cdots+x(N-1)e^{-i\frac{2\pi}{N}(N-1)k} \tag{2.3}$$

N times multiplication,that's mean if we want to get $\{\mathrm{X}(k)\}(k=0\sim N-1)$,we need N^2 times multiplication and $N(N-1)$ times addition,so if we compute the **DFT** directly, the time complexity of the algorithm will be $O(N^2)$
The algorithm with time complexity of $O(N^2)$ are not suitable for larger data sizes,so we need an effective algorithm.We begin with the property of W_N^{kn}.

Property 2.1 (Periodicity)**.** $W_N^{n(N-K)}=W_N^{-nk}, W_N^{k(N-n)}=W_N^{-nk}$

Proof.

$$W_N^{n(N-k)}=e^{-i\frac{2\pi}{N}(N-k)n}=e^{-i\frac{2\pi}{N}nN}e^{i\frac{2\pi}{N}nk}=e^{-i2\pi n}W_N^{-nk}=W_N^{-nk} \tag{2.4}$$

$$W_N^{k(N-n)}=e^{-i\frac{2\pi}{N}(N-n)k}=e^{-i\frac{2\pi}{N}kN}e^{i\frac{2\pi}{N}nk}=e^{-i2\pi k}W_N^{-nk}=W_N^{-nk} \tag{2.5}$$

□

Corollary 2.1. $W_N^{nk}=W_N^r$ *where r satisfies* $r\equiv(nk) \mod N$

Proof. Assume that $nk=Nq+r(q\in\mathbb{Z})$,we have

$$W_N^r=W_N^{nk-Nq}=W_N^{nk}W_N^{-Nq}=W_N^{nk}e^{2\pi iq}=W_N^{nk} \tag{2.6}$$

□

Property 2.2 (Symmetry)**.** $W_N^{nk+\frac{N}{2}}=-W_N^{nk}$

Proof. $W_N^{nk+\frac{N}{2}}=e^{-i\frac{2\pi}{N}nk}e^{-i\frac{2\pi}{N}\frac{N}{2}}=-e^{-i\frac{2\pi}{N}nk}=-W_N^{nk}$ □

Use these two properties,decompose the expression of DFT into two parts.

$$\begin{aligned}
\mathrm{X}(k)&=\sum_{n=0}^{N-1}x(n)W_N^{kn}\quad k=0,1,\cdots,N-1\\
&=\sum_{m=0}^{\frac{N}{2}-1}x(2m)W_N^{2mk}+\sum_{m=0}^{\frac{N}{2}-1}x(2m+1)W_N^{k(2m+1)}\\
&=\sum_{m=0}^{\frac{N}{2}-1}x(2m)W_{N/2}^{mk}+W_N^k\sum_{m=0}^{\frac{N}{2}-1}x(2m+1)W_{N/2}^{km}
\end{aligned}$$

Define $G(k)=\sum_{m=0}^{\frac{N}{2}-1}x(2m)W_{N/2}^{mk}$ and $H(k)=\sum_{m=0}^{\frac{N}{2}-1}x(2m+1)W_{N/2}^{km}$

Property 2.3. *$G(k)$ and $H(k)$ satisfies $G(k+\frac{N}{2}) = G(k)$ and $H(k+\frac{N}{2}) = H(k)$*

Proof. $G(k+\frac{N}{2}) = \sum_{m=0}^{\frac{N}{2}-1} x(2m)W_{N/2}^{r(k+N/2)} = \sum_{m=0}^{\frac{N}{2}-1} x(2m)W_{N/2}^{rk} = G(k)$ □

Similarly, it can be shown that $H(k)$ is periodic,so we have

$$\mathrm{X}(k) = \begin{cases} G(k) + W_N^k H(k), & k \leq \dfrac{N}{2} - 1; \\ G(k-\dfrac{N}{2}) - W_N^{k-\frac{N}{2}} H(k-\dfrac{N}{2}), & k \geq \dfrac{N}{2} \end{cases} \tag{2.7}$$

So if we obtained $G(k)$ and $H(k)$ for $k \in \{0,1,\cdots \frac{N}{2}-1\}$,to get $\mathrm{X}(k)$,we need to compute $W_N^k H(k)$,$G(k) + W_N^k H(k)$ and $G(k) - W_N^k H(k)$ for $k \in \{0,1,\cdots \frac{N}{2}-1\}$.So we need $\frac{N}{2}$ times multiplication and N times addition.

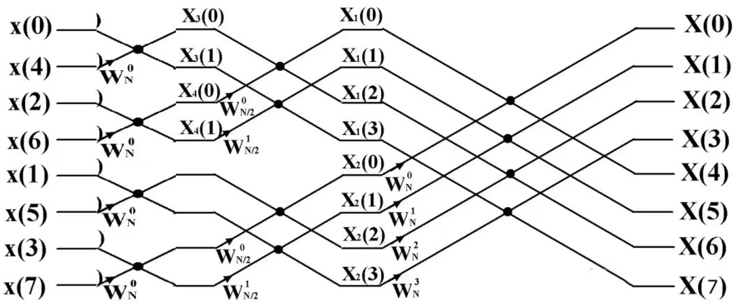
We defined $\frac{N}{2}$-points discrete-time signal $f_1(m) = x(2m)$ and $f_2(m) = x(2m+1)$ for $m \in \{0,1,\cdot\cdot,\frac{N}{2}-1\}$,we have

$$G(k) = \sum_{m=0}^{\frac{N}{2}-1} f_1(m)W_{N/2}km \tag{2.8}$$

$$H(k) = \sum_{m=0}^{\frac{N}{2}-1} f_2(m)W_{N/2}km \tag{2.9}$$

So $G(k)$ and $H(k)$ for $k \in \{0,1,\cdots \frac{N}{2}-1\}$ is the DFT of $\{f_1(m)\}$ and $\{f_2(m)\}$,so we can use same method on $\{f_1(m)\}$ and $\{f_2(m)\}$ for $m \in \{0,1,\cdots \frac{N}{2}-1\}$.As we assumed that $N = 2^v$,this method can be used for v times.

For $v = 3$,the process of the fast Fourier algorithm can be represented by the following butterfly diagram.



Consider the time complexity of FFT,dividing the algorithm into v steps,at each step we initially have a sequence of length N,and we need N times additions and $\frac{N}{2}$ times multiplications to obtain a new sequence as the initial sequence for the next step.So the time complexity of FFT is $O(N\log_2 N)$.

3 Applications

3.1 Analytic Signal

Definition 3.1. Suppose we have a continuous-time real-valued signal $x(t)$ and if a complex-valued signal $z(t)$ satisfy $\mathrm{Re}\{z(t)\} = x(t)$ and $Z(f) = 0$ when the frequency $f < 0$,we call $z(t)$ is continuous-time analytic signal corresponding to $x(t)$

Let $z_r(t) = \mathrm{Re}\{z(t)\}$ and $z_i(t) = \mathrm{Im}\{z(t)\}$,we introduce the orthogonality between the real and imaginary components of the analytic signal.

Property 3.1. *Suppose* $z_r(t) = \mathrm{Re}\{z(t)\}$ *and* $z_i(t) = \mathrm{Im}\{z(t)\}$*,we have*

$$\int_{-\infty}^{+\infty} z_r(t) z_i(t) \mathrm{d}t = 0 \tag{3.1}$$

To prove this property,we first introduce two approaches to creat analytic signal $z(t)$.

3.1.1 Creat Analytic Signal from Time domain

Suppose $x(t)$ is a continuous-time real-valued signal,we will prove that $z(t) = x(t) + i\hat{x(t)}$ is the analytic signal corresponding to $x(t)$.

Proof.

Lemma 3.1 (Dirichlet integral)**.** $\int_{-\infty}^{+\infty} \frac{\sin(ft)}{t} \mathrm{d}t = \pi sgn(f)$

We just need to show for $f < 0$,$\int_{-\infty} +\infty x(t)e^{-ift} + i\hat{x(t)}e^{-ift}\mathrm{d}t = 0$

$$\begin{aligned}
\mathcal{F}(z(t)) &= \int_{-\infty}^{+\infty} x(t)e^{-ift} + i\hat{x(t)}e^{-ift}\mathrm{d}t \\
&= \int_{-\infty}^{+\infty} x(t)\cos(ft) - ix(t)\sin(ft)\mathrm{d}t + \frac{i}{\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{x(\tau)}{t-\tau} e^{-ift}\mathrm{d}\tau\mathrm{d}t \\
&= \int_{-\infty}^{+\infty} x(t)\cos(ft) - ix(t)\sin(ft)\mathrm{d}t + \frac{i}{\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{x(\tau)}{t-\tau} e^{-ift}\mathrm{d}t\mathrm{d}\tau \\
&= \int_{-\infty}^{+\infty} x(t)\cos(ft) - ix(t)\sin(ft)\mathrm{d}t + \frac{i}{\pi}\int_{-\infty}^{+\infty} \pi x(\tau) e^{i(\frac{\pi}{2} - f\tau)}\mathrm{d}\tau \\
&= 0
\end{aligned}$$

□

By this method we can also obtain that the Fourier transform of the imaginary part is a conjugate odd function.

Property 3.2. *Suppose* $\mathrm{Z}_i(f) = \mathcal{F}(iz_i(t))$,*we have*

$$\mathrm{Z}_i(f) = \begin{cases} \mathrm{X}(f), & f > 0 \\ 0, & f = 0 \\ -\bar{\mathrm{X}}(-f), & f < 0 \end{cases} \tag{3.2}$$

3.1.2 Creat Analytic Signal from Frequency domain

Suppose $\mathrm{X}(f)$ defined over the frequency interval $-\infty < f < +\infty$, then we defined

$$\mathrm{Z}(f) = \begin{cases} 2\mathrm{X}(f), & f > 0 \\ X(0), & f = 0 \\ 0, & f < 0 \end{cases} \tag{3.3}$$

Use inverse Fourier transform to obtained $z(t)$,we prove $z(t)$ is analytic signal.

Proof. We just need to show $\mathrm{Re}\{z(t)\} = x(t)$

$$\mathrm{Re}\{z(t)\} = \frac{1}{2}(z(t) + z\bar{(t)}) = \frac{1}{2\pi}\int_0^{+\infty}(X(f) + X\bar{(f)})e^{ift}\mathrm{d}f$$

Notice that $X\bar{(f)} = X(-f)$ and $\frac{1}{2\pi}\int_{-\infty}^{+\infty} X(f)e^{ift}\mathrm{d}f = x(t)$,we completed the proof □

Now,we prove the orthogonality between the real and imaginary components of the analytic signal.

Proof. Notice that $z_r(t) = x(t)$ and $z_i(t) = x\hat{(t)} = \frac{1}{\pi}\int_{-\infty}^{+\infty}\frac{x(\tau)}{t-\tau}\mathrm{d}\tau$

$$\begin{aligned}\int_{-\infty}^{+\infty} z_r(t)z_i(t)\mathrm{d}t &= \frac{1}{\pi}\int_{-\infty}^{+\infty} x(t)\int_{-\infty}^{+\infty}\frac{x(\tau)}{t-\tau}\mathrm{d}\tau\mathrm{d}t \\ &= \frac{1}{\pi}\int\int\frac{x(t)x(\tau)}{t-\tau}\mathrm{d}\tau\mathrm{d}t\end{aligned}$$

By symmetry,we obtained that $\int\int\frac{x(t)x(\tau)}{t-\tau}\mathrm{d}\tau\mathrm{d}t = 0$. □

3.2 Discrete-Time "Analytic" Signal

Now,let's consider the discrete-time situation,suppose we have a N-points discrete-time signal $\{x(n)\}$ obtained by sampling a bandlimited real-valued continuous-time signal $x(nT) = x(n)$ at periodic time intervals of T seconds.There are two properties we wish to satisfy in order for $z(n) = z_r(n) + iz_i(n)$ to be an analytic-like discrete-time signal.

Property 3.3. *The real part of $z(n)$ must exactly yield the original discrete-time sequence.*

$$z_r(n) = x(n) \quad \forall n \in \{0, 1, \cdot\cdot, N-1\} \tag{3.4}$$

Property 3.4. *The real and imaginary components must be orthogonal over the finite interval.*

$$\sum_{n=0}^{N-1} z_r(n) z_i(n) = 0 \tag{3.5}$$

Consider three cases of the analytic-like discrete-time signal that differ in their sample rates.Suppose X(k) is the DFT of $x(n)$

3.2.1 Computing Standard Discrete-Time ”Analytic” Signal

We use same method as the continous-time situation.We defined

$$\mathrm{Z}(k) = \begin{cases} \mathrm{X}(0), & k = 0 \\ 2\mathrm{X}(k), & 1 \le k \le \dfrac{N}{2} \\ 0, & \dfrac{N}{2} + 1 \le k \le N-1 \end{cases} \tag{3.6}$$

And we use inverse transform to obtain $z(n)$.However,this method is not suitable for the discrete case.Consider the data vector as following.

$$x(n) = [4, 2, -2, -1, 3, 1, -3, 1]$$

Use (3.6) we get $z(n)$ as following.

$$z(n) = [3.875-0.396i, 2.125+3i, -2.125+1.811i, -0.875-2.293i, 2.875-1.104i, 1.125+3i, -3.125-0.311i, 1.125-3.707i]$$

We notice that the real part is not equal to the original data.The problem arises at the boundary point.Processing the boundary points, we obtain the following correction formula.

$$\mathrm{Z}(k) = \begin{cases} \mathrm{X}(0), & k = 0 \\ 2\mathrm{X}(k), & 1 \le k \le \dfrac{N}{2} - 1 \\ \mathrm{X}(\dfrac{N}{2}), & k = \dfrac{N}{2} \\ 0, & \dfrac{N}{2} + 1 \le k \le N-1 \end{cases} \tag{3.7}$$

Now,we prove that (3.7) satisfy Property3.3 and 3.4

Proof. We prove Property3.3 at first.Notice that

$$z(n) = \frac{1}{N}\sum_{k=0}^{\frac{N}{2}} \mathrm{Z}(k)W_N^{-nk} \tag{3.8}$$

$$= \frac{1}{N}(\mathrm{X}(0) + \sum_{k=1}^{\frac{N}{2}-1} 2\mathrm{X}(k)W_N^{nk} + \mathrm{X}(\frac{N}{2})\cos(n\pi)) \tag{3.9}$$

And

$$\mathrm{X}(k) = \sum_{n=0}^{N-1} x(n)W_N^{kn} \tag{3.10}$$

We calculate the factor of each $x(m)$,while $m \in \{0, 1, \cdot\cdot, N-1\}$,let it be A_n^m

$$\mathrm{A}_n^m = \frac{1}{N}(1 + 2\sum_{k=1}^{\frac{N}{2}-1} W_N^{k(m-n)} + \cos(n\pi)\cos(m\pi)) \tag{3.11}$$

So if $m = n$,apparently that $\mathrm{A}_n^n = 1$,and if $m \neq n$,let $q = W_N^{m-n}$,we have

$$\mathrm{A}_n^m = \frac{(1 - \cos(m\pi)\cos(n\pi))(1+q)}{(1-q)} \tag{3.12}$$

Notice that $\bar{q} = \frac{1}{q}$,we get that $\mathrm{A}_n^m + \bar{\mathrm{A}}_n^m = 0$. So we get that $\mathrm{Re}(z(n)) = x(n)$

Next we prove Property3.4.Notice that

$$z_i(n) = \frac{1}{2i}(z(n) - \bar{z}(n)) = \frac{1}{i}(\sum_{m\neq n} x(m)\mathrm{A}_n^m) \tag{3.13}$$

So we just need to prove that.

$$\sum_{n=0}^{N-1} x(n)\sum_{m\neq n} x(m)\mathrm{A}_n^m = 0 \tag{3.14}$$

Notice that $\mathrm{A}_n^m + \mathrm{A}_m^n = 0$.We completed the proof. □

3.2.2 Computing Decimated Discrete-Time "Analytic" Signal

It is usually desirable in digital hardware implementations of digital signal processing operations to use the lowest sample rate consistent with preservation of the signal information without aliasing.So sometimes we need reduce sampling rate.The following equation shows that the $\frac{N}{2}$-points analytic signal can be generated directly from the N-points analytic signal.

$$\mathrm{Z}(k) = \begin{cases} \mathrm{X}(0) + \mathrm{X}(\frac{N}{2}), & k = 0 \\ 2\mathrm{X}(k), & 1 \leq k \leq \frac{N}{2} - 1 \end{cases} \tag{3.15}$$

Use inverse DFT and scale by factor $\frac{1}{2}$ to obtain decimated discrete-time "analytic" signal $z(n)$ of half the original sample rate.

Proof. We proof that $\mathrm{Re}\{z(n)\} = x(2n)$

$$\begin{aligned} z(n) &= \frac{1}{2}\frac{2}{N}\sum_{k=0}^{\frac{N}{2}-1} \mathrm{Z}(k)W_{N/2}^{-nk} \\ &= \frac{1}{N}(\sum_{k=1}^{\frac{N}{2}-1} 2\mathrm{X}(k)W_{N/2}^{-nk} + (X(0) + X(\frac{N}{2}))) \end{aligned}$$

Notice that

$$\mathrm{X}(k) = \sum_{n=0}^{N-1} x(n)W_N^{nk} \tag{3.16}$$

Use the same notation as in the previous section.

$$\mathrm{A}_n^m = \frac{1}{N}(1 + \cos(m\pi) + 2\sum_{k=1}^{\frac{N}{2}-1} W_N^{k(m-2n)}) \tag{3.17}$$

So if $m = 2n$,we have $\mathrm{A}_n^{2n} = 1$ and if $m \neq 2n$,we have $\mathrm{A}_n^m + \bar{\mathrm{A}}_n^m = 0$.The orthogonality can be proved in a similar way to the previous section.So we complete the proof. □

3.2.3 Computing Interpolated Sample Rate Discrete-Time "Analytic" Signal

In some cases we need a more accurate reconstruction of the signal, so encrypted sampling is performed.

$$\mathrm{Y}(k) = \begin{cases} \mathrm{X}(k), & 0 \leq k \leq \frac{N}{2} - 1 \\ \frac{1}{2}\mathrm{X}(\frac{N}{2}), & k = \frac{N}{2} \\ 0, & \frac{N}{2} + 1 \leq k \leq NM - \frac{N}{2} - 1 \\ \frac{1}{2}\mathrm{X}(\frac{N}{2}), & k = NM - \frac{N}{2} \\ \mathrm{X}(k), & NM - \frac{N}{2} + 1 \leq k \leq NM - 1 \end{cases} \tag{3.18}$$

Compute the NM-points inverse DFT and scale by M,we obtained NM-points discrete-time signal $y(n)$.We prove that.

$$y(n) = \begin{cases} x(\frac{n}{M}), & n = pM \quad p \in \{0, 1, \cdot\cdot, N-1\} \\ 0, & else \end{cases} \tag{3.19}$$

Proof. We assume that $n = pM$,then we have

$$y(n) = M\frac{1}{NM}(\sum_{k=0}^{\frac{N}{2}-1} \mathrm{X}(k)W_{NM}^{-nk} + \frac{1}{2}\mathrm{X}(\frac{N}{2})W_{NM}^{-n\frac{N}{2}} + \frac{1}{2}\mathrm{X}(\frac{N}{2})W_{NM}^{-n(NM-\frac{N}{2})} + \sum_{k=\frac{N}{2}+1}^{N-1} \mathrm{X}(k)W_{NM}^{-n(k+NM-N)})$$

$$= \frac{1}{N}(\sum_{m=0}^{N-1} x(m)(\sum_{k=0}^{\frac{N}{2}-1} W_N^{k(m-p)} + \cos(m\pi)\cos(p\pi) + \sum_{k=\frac{N}{2}+1}^{N-1} W_N^{k(m-p)})$$

Use same method as (3.11) we have $y(pM) = x(p)$,and for $n \neq pM$ we have $y(n) = 0$ □

We called $y(n)$ a trigonometrically interpolated discrete-time signal from $x(n)$. The following algorithm tell us how to form the analytic-like signal of $y(n)$ by N-points discrete-time signal $x(n)$.

$$\mathrm{Z}(k) = \begin{cases} \mathrm{X}(0), & k = 0 \\ 2\mathrm{X}(k), & 1 \leq k \leq \frac{N}{2} - 1 \\ \mathrm{X}(\frac{N}{2}), & k = \frac{N}{2} \\ 0, & \frac{N}{2} + 1 \leq k \leq NM - 1 \end{cases} \tag{3.20}$$

Use same method as 3.2.1 and 3.2.2,we can easily verify that it satisfies the properties 3.3 and 3.4.

References

[1] S. Lawrence and Marple, Jr,Computing the Discrete-Time “Analytic” Signal via FFT

[2] John G.Proakis and Dimitris G.Manolakis,Digital Signal Processing Principles,Algorithms,and Applications

Analysis

Interpolation and Corona Problem

Liu Yao, Li Xiaoran

June 2022

1 Introduction

Let B denote the Banach algebra of bounded analytic functions in $|z| < 1$ under the maximal norm. (We'll prove B is a Banach algebra in section 2)

For $f_1, ..., f_n \in B$, consider $I = I(f_1, ..., f_n)$ the ideal generated by $f_1, ..., f_n$.

If $I = B$, then there exist $g_1, ..., g_n \in B$ such that

$$f_1g_1 + f_2g_2 + ... + f_ng_n = 1$$

$$\Rightarrow |f_1(z)| + ... + |f_n(z)| \geqslant \frac{1}{\max_{1\leqslant i\leqslant n} ||g_i||} > 0$$

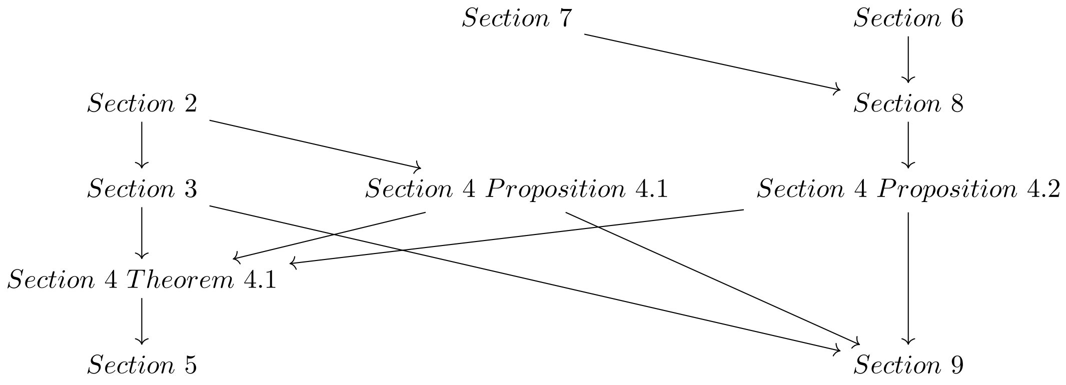
Corona Problem says if $f_1, ..., f_n$ satisfies

$$|f_1| + ... + |f_n| \geqslant \delta > 0$$

then $I = B$.

This is the main motivation of this paper. We'll prove it in section 9.

The relationships between section 2 to section 9 are as follows:



Section 4 Theorem 4.1 and section 5 study the secondary motivation: interpolation.

Interpolation studies what conditions on $a_1, a_2, ... \in \{z : \ |z| < 1\}$ and $w_1, w_2, ... \in \mathbb{C}$ does there exist $f \in B$ such that

$$f(a_\nu) = w_\nu, \qquad \nu = 1, 2, ...$$

We will use the notation $A_1, A_2, ...$ for numerical constants.

2 Hardy Space

Definition 2.1 (Hardy Space)**.** *Denote H^p $(1 \leqslant p \leqslant +\infty)$ the norm space of functions G analytic in $|z| < 1$ under the norm*

$$\|G\|_p = \lim_{r\to 1}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}|G(re^{i\theta})|^p d\theta\right)^{\frac{1}{p}} \tag{2.1}$$

The lemma below shows that the limit in (2.1) exists, maybe infinity.

Lemma 2.1. *For every analytic function f on $|z| < 1$, $(\frac{1}{2\pi}\int_0^{2\pi}|f(re^{i\theta})|^p)^{\frac{1}{p}}$ grows when r grows $(0 < r < 1)$.*

Proof. For $0 < r < R < 1$,

$$\begin{aligned}
f(re^{i\theta}) &= \frac{1}{2\pi}\int_0^{2\pi} f(Re^{i\theta})Re(\frac{Re^{i\phi}+re^{i\theta}}{Re^{i\phi}+re^{i\theta}})d\phi \\
&= \frac{1}{2\pi}\int_0^{2\pi} f(Re^{i\theta})\frac{1-(\frac{r}{R})^2}{1-2\frac{r}{R}\cos(\phi-\theta)+(\frac{r}{R})^2}d\phi \\
|f(re^{i\theta})| &\leq \int_0^{2\pi}\frac{1}{2\pi}\frac{1-(\frac{r}{R})^2}{1-2\frac{r}{R}\cos(\phi-\theta)+(\frac{r}{R})^2}|f(Re^{i\phi}|d\phi \\
&\leq (\int_0^{2\pi}\frac{1}{2\pi}\frac{1-(\frac{r}{R})^2}{1-2\frac{r}{R}\cos(\phi-\theta)+(\frac{r}{R})^2}|f(Re^{i\phi}|^p d\phi)^{\frac{1}{p}} \\
(\frac{1}{2\pi}\int_0^{2\pi}|f(re^{i\theta})|^p d\theta)^{\frac{1}{p}} &\leq (\frac{1}{2\pi}\int_0^{2\pi}d\theta\int_0^{2\pi}\frac{1}{2\pi}\frac{1-(\frac{r}{R})^2}{1-2\frac{r}{R}\cos(\phi-\theta)+(\frac{r}{R})^2}|f(Re^{i\phi}|^p d\phi)^{\frac{1}{p}} \\
&= (\frac{1}{2\pi}\int_0^{2\pi}|f(Re^{i\phi}|^p d\phi)^{\frac{1}{p}}
\end{aligned}$$

□

Theorem 2.1. *H^p is a Banach Space*

Proof. We only prove the completeness.

Let $G_1, G_2, \dots, G_n, \dots$ be a Cauchy sequence in H^p.

$\forall 0 < r < 1$, choose R such that $r < R < 1$.

$\forall \epsilon \in \mathbb{R}$, choose N such that $\forall n > m > N$, $\|G_n - G_m\|_p < \epsilon$. Then $\forall z \in D_r = \{a \in \mathbb{C} : |z| < r\}$,

$$\begin{aligned}
|G_n(z)-G_m(z)| &= \left|\frac{1}{2\pi i}\int_{|\xi|=R}\frac{G_n(\xi)-G_m(\xi)}{\xi - z}\mathrm{d}\xi\right| \\
&\leqslant \frac{R}{2\pi}\int_0^{2\pi}\frac{|G_n(Re^{i\theta})-G_m(Re^{i\theta})|}{|Re^{i\theta}-z|}\mathrm{d}\theta \leqslant \frac{R}{R-r}\int_0^{2\pi}\frac{1}{2\pi}|G_n(Re^{i\theta})-G_m(Re^{i\theta})|\mathrm{d}\theta \\
&\leqslant \frac{R}{R-r}(\int_0^{2\pi}\frac{1}{2\pi}|G_n(Re^{i\theta})-G_m(Re^{i\theta})|^p\mathrm{d}\theta)^{\frac{1}{p}} \leqslant \frac{R}{R-r}||G_n-G_m||_p \\
&\leqslant \frac{R}{R-r}\epsilon
\end{aligned}$$

So G_n uniformly converges on D_r as $n \to \infty$. Since r is arbitrary, we conclude that G_n uniformly converges to G on any compact subset of $|z| < 1$ where G analytic on $|z| < 1$.

For $0 < r < 1$, since G_n converges uniformly to G in D_r as $n \to \infty$,

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} |G_n(re^{i\theta})|^p \mathrm{d}\theta \to \frac{1}{2\pi}\int_{-\pi}^{\pi} |G(re^{i\theta})|^p \mathrm{d}\theta \qquad (n \to \infty)$$

$$\begin{aligned}\Rightarrow \left(\frac{1}{2\pi}\int_{-\pi}^{\pi} |G(re^{i\theta})|^p \mathrm{d}\theta\right)^{\frac{1}{p}} &= \lim_{n\to+\infty}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} |G_n(re^{i\theta})|^p \mathrm{d}\theta\right)^{\frac{1}{p}} \\ &\leqslant \lim_{n\to+\infty} \|G_n\|_p \qquad (\textit{by Lemma } 2.1) \\ &\Rightarrow G \in H^p\end{aligned}$$

And it's easy to prove

$$||G_n - G||_p \to 0,\ n \to +\infty$$

□

In the case $p = +\infty$, H^∞ consists of all bounded analytic functions in $|z| < 1$, with norm $||f||_\infty = sup\{|f(z)|,\ |z| < 1\}$. Denote $B = H^\infty$. In the remaining contexts, $||\cdot||$ refers to $||\cdot||_\infty$

Proposition 2.1. *$1 \leqslant p \leqslant q \leqslant +\infty$, if $f \in H^q$, then $f \in H^p$ and $||f||_p \leqslant ||f||_q$*

Proof. For any $0 < r < 1$,

$$(\frac{1}{2\pi}\int_0^{2\pi} |f(re^{i\theta})|^p d\theta)^{\frac{1}{p}} \leqslant (\frac{1}{2\pi}\int_0^{2\pi} |f(re^{i\theta})|^q d\theta)^{\frac{1}{q}}$$

$$\Rightarrow ||f||_p \leqslant ||f||_q$$

□

Now enumerate some properties of Hardy space in [2] p.96-p.100 without proofs:

Proposition 2.2. *Let $f \in H^p$, $1 \leqslant p \leqslant +\infty$ and denote its zeros in $|z| < 1$ by $\zeta_1, \zeta_2, \ldots$ with multiplicity. Then $\sum(1 - |\zeta_j|) < +\infty$*

Proposition 2.3. *Let $f \in H^p$, $1 \leqslant p \leqslant +\infty$, then $\exists \widetilde{f} \in L^p[0, 2\pi]$ such that*

$$f(re^{i\theta}) \to \widetilde{f}(\theta), \quad a.e.\ and\ L^1, \quad as\ r \to 1^-$$
$$and \quad ||f||_p = (\tfrac{1}{2\pi}\textstyle\int_0^{2\pi} |\widetilde{f}(\theta)|^p \mathrm{d}\theta)^{1/p}$$

Denote $f(e^{i\theta}) = \widetilde{f}(\theta)$. Moreover,

$$f(re^{i\theta}) = \frac{1}{2\pi}\int_0^{2\pi} \frac{1 - r^2}{1 + r^2 - 2rcos(\theta - \phi)} f(e^{i\phi})\mathrm{d}\phi, \quad \forall r \in [0, 1),\ \theta \in \mathbb{R}$$

Proposition 2.4. *$p \geqslant 1$, f is a L^p function on the unit circle. Then $\exists g \in H^p$ such that $g(e^{i\theta}) = f(e^{i\theta})$ if and only if $\int_0^{2\pi} f(e^{i\theta})e^{in\theta}\mathrm{d}\theta = 0$ for all $n > 0$.*

3 A Useful Theorem

Theorem 3.1. *Let σ_ν be a sequence of open subintervals of $(0,1)$. Denote by σ'_ν the interval obtained from σ_ν by adjoining to σ_ν equally long intervals on both sides, and let the length of σ_ν be l_ν. Then the conditions*

$$\sum_{\sigma_j \subset \sigma'_i} l_j \leq Cl_i \tag{3.1}$$

for all i and a fixed constant $C > 2$, imply

$$\sum_{\nu=1}^{\infty}(\frac{1}{l_\nu})(\int_{\sigma_\nu} f(x)dx)^2 \leq AC\int_0^1 f(x)^2dx \tag{3.2}$$

for all square integrable function f.

Proof. We first reduce the theorem to the case that all σ_ν have the form $(\frac{k}{2^n}, \frac{k+1}{2^n})$, where $k, n \in \mathbb{Z}^+$.

As we know, every σ_ν is included in the union of at most four equally long intervals of the form $(\frac{k}{2^n}, \frac{k+1}{2^n})$, where $\frac{1}{2^n} \leq \frac{l_\nu}{2}$. We can get it by this. First, $\sigma_\nu \subset (0,1)$, then divide $[0,1]$ through the half to $[0, \frac{1}{2}], [\frac{1}{2}, 1]$. Second, divide the interval which intersect σ_ν. Continue until one interval $[\frac{k}{2^n}, \frac{k+1}{2^n}] \subset \sigma_\nu$. Let ω_i be the intervals obtained as above and let m_i be the length of ω_i. (3.1) imply that

$$\sum_{\omega_j \subset \omega_i} m_j = \sum_{\sigma_j \subset \sigma'_i} m_j \leq \sum_{\sigma_j \subset \sigma'_i} \frac{l_j}{2} \leq \frac{Cl_i}{2} \leq 2Cm_i = cm_i \tag{3.3}$$

so it is sufficient to prove it by assume $\sigma_\nu = [\frac{k}{2^n}, \frac{k+1}{2^n}]$.

Let $X_\nu(x)$ be the characteristic function of ω_ν and introduce the kernel

$$K(x,y) = \sum_{\nu=1}^{\infty} X_\nu(x)X_\nu(y)/m_\nu$$

$$\begin{aligned}\sum_{\nu=0}^{\infty}\frac{1}{m_\nu}(\int_{\omega_\nu} f(x)dx)^2 &= \int_0^1 f(x)dx\int_0^1 K(x,y)f(y)dy \\ &\leq \left\{\int_0^1 f(x)^2dx\right\}^{\frac{1}{2}}\left\{\int_0^1 dx(\int_0^1 K(x,y)f(y)dy)^2\right\}\end{aligned}$$

We write $m_{ij} = m(\omega_i \cap \omega_j)$, and

$$F = \sum_{i,j}\frac{m_{ij}}{\sqrt{m_i}\sqrt{m_j}}x_ix_j \tag{3.4}$$

where $x_i = \frac{1}{\sqrt{m_i}}\int_{\omega_i} f(x)dx$. If we can prove $F \leq AC\sum_{\nu=1}^{\infty} x_i^2$, then

$$\sum_{\nu=0}^{\infty}\frac{1}{m_\nu}(\int_{\omega_\nu} f(x)dx)^2 \leq \left\{\int_0^1 f(x)^2dx\right\}^{\frac{1}{2}} AC\sum_{\nu=1}^{\infty} x_i^2 = AC\left\{\int_0^1 f(x)^2dx\right\}^{\frac{1}{2}}\left\{\sum_{\nu=0}^{\infty}\frac{1}{m_\nu}(\int_{\omega_\nu} f(x)dx)^2\right\}^{\frac{1}{2}}$$

To discussion F. First, we pay attention to a fact that $\omega_i \subset \omega_j$ or $\omega_j \subset \omega_i$ or $\omega_i \cap \omega_j = \emptyset$. Denote $G = \{i : \omega_i\}$. Let G_1 be set of i where ω_i is not contained in any larger interval ω_j. Let G_2 be the set of all $i \in G - G_1$ where ω_i is not contained in any larger interval ω_j. Denote G_3 the same, so as G_k. Denote $i \in G_k$,let $G_{i\nu}$ be the set of j where $\omega_j \subset \omega_i$ and $j \in G_{\nu+k}, \nu = 0, 1, \ldots.$

For $i \in G$,denote $a_\nu = \sum_{j \in G_{i\nu}} m_j$,and we have

$$\sum_{\nu=n}^{\infty} a_\nu \leq c a_n \qquad n = 0, 1, \ldots \tag{3.5}$$

$$a_0 \geq a_1 \geq a_2 \geq \cdots \geq 0 \tag{3.6}$$

We claim that under the condition (3.5) and (3.6), we have

$$a_n \leq 4(1 - \frac{1}{c})^n a_0$$

We can assume $a_0 = 1$.Let $N > c$ Take $\{b_\nu\}, b_0 = 1; b_n = 0$ when $n > N$, and satisfy

$$\sum_{\nu=n}^{N} b_\nu = c b_n, \qquad n \leq N - c$$

we have $b_n = (1 - \frac{1}{c})^n$. Hence,

$$a_{N-c} \leq b_{N-c} = (1 - \frac{1}{c})^{N-c}$$

which is our assertion.

$$\begin{aligned} F^2 \leq & \left\{ \sum_{i=1}^{\infty} \frac{x_i}{m_i} \sum_{\nu=0}^{\infty} \sum_{j \in G_{i\nu}} \sqrt{m_j} x_j + \sum_{j=1}^{\infty} \frac{x_j}{m_j} \sum_{\nu=0}^{\infty} \sum_{i \in G_{j\nu}} \sqrt{m_i} x_i \right\}^{\frac{1}{2}} \\ & \leq 4 \sum_{i=1}^{\infty} x_i^2 \cdot \sum_{i=1}^{\infty} \frac{1}{m_i} (\sum_{\nu=0}^{\infty} \sum_{j \in G_{i\nu}} \sqrt{m_j} x_j)^2 \end{aligned}$$

We define $k > 0$ by $k^2 = 1 - \frac{1}{c}$

$$\begin{aligned} & \sum_{i=1}^{\infty} \frac{1}{m_i} (\sum_{\nu=0}^{\infty} \sum_{j \in G_{i\nu}} \sqrt{m_j} x_j)^2 \\ \leq & \sum_{i=1}^{\infty} \frac{1}{m_i} \sum_{\nu=0}^{\infty} k^{-\nu} (\sum_{j \in G_{i\nu}} \sqrt{m_j} x_j)^2 \cdot \sum_{\nu=0}^{\infty} k^{\nu} \\ \leq & \frac{1}{1-k} \sum_{i=1}^{\infty} \frac{1}{m_i} \sum_{\nu=0}^{\infty} k^{-\nu} \sum_{j \in G_{i\nu}} m_j \cdot \sum_{j \in G_{i\nu}} x_j^2 \\ \leq & \frac{4}{1-k} \sum_{\mu=1}^{\infty} \sum_{i \in G_\mu} \sum_{\nu=0}^{\infty} k^{\nu} \sum_{j \in G_{i\nu}} x_j^2 = \frac{4}{1-k} \sum_{\nu=0}^{\infty} \sum_{\nu=0}^{\infty} k^{\nu} \sum_{j \in G_{\mu+\nu}} x_j^2 \\ \leq & \frac{4}{1-k} \sum_{j=1}^{\infty} x_j^2 \sum_{\nu=0}^{\infty} k^{\nu} = \frac{4}{(1-k)^2} \sum_{i=1}^{\infty} x_i^2 \end{aligned}$$

□

Theorem 3.2. *Let $\mu(z)$ be a non-negative measure in $|z| < 1$ and assume that*

$$\mu(S) \leq Cl \tag{3.7}$$

holds for all sets S of the form

$$S = \left\{ re^{i\theta} : r \geqslant 1 - l, \theta_0 \leq \theta \leq \theta_0 + l \right\}, \qquad l \leq 1 \tag{3.8}$$

Then there is an absolute constant A_{16} so that

$$\int_D |G(z)|^p d\mu(z) \leq A_{16} C \|G\|_p^p \tag{3.9}$$

for all $G \in H^p, p \geqslant 1$.Conversely, if (3.9) holds for a certain constant C , $\mu(S)$ satisfies (3.7) with C independent of S.

Proof. Let us first assume (3.9) holds for some constant C and consider S as (3.8).

Let

$$G(z) = \left(\frac{1 - |a|^2}{(1 - z\bar{a})^2} \right)^{\frac{1}{p}}$$

where $a = (1 - l)e^{i\theta_0 + (\frac{1}{2})l}$, then

$$\|G\|_p^p = \frac{1}{2\pi} \int_0^{2\pi} \frac{1 - (1 - l)^2}{|1 - e^{i\theta(1-l)}|^2} = \frac{1}{2\pi} \int_0^{2\pi} \frac{1 - (1 - l)^2}{1 - 2(1 - l)\cos\theta + (1 - l)^2} = 1$$

$$|G(z)|^p = \frac{1 - (1 - l)^2}{|1 - z\bar{a}|^2} = \frac{2l - l^2}{|\frac{a}{1-l} - z(1 - l)|^2} \geqslant \frac{2l - l^2}{l^2} \geqslant \frac{1}{2l}$$

By (3.9), $\mu(S) \leq 2Cl$.

Now we prove the converse. Denote

$$r_{\nu n} = \left\{ z : 2^{-1-n} < 1 - |z| \leq 2^{-n}, \frac{\nu \cdot 2\pi}{2^{n+1}} \leq \arg z < \frac{(\nu + 1) \cdot 2\pi}{2^{n+1}} \right\} \tag{3.10}$$

where $n = 0, 1, \cdots ;\ v = 0, 1, \cdots, 2^{n+1} - 1$.

Let $z_{\nu n} = (1 - 2^{-n})e^{\frac{\nu + \frac{1}{2}}{2^{n+1}} 2\pi i}$,and $\omega_{\nu n}$ be the range of $\arg z$ for z in $r_{\nu n}$,where $n = 0, 1, \cdots ; v = 0, 1, \cdots, 2^{n+1} - 1$. We first assume $G \neq 0$, since otherwise by Proposition 2.2 and [1] Chapter 5 Problem 2, assume A is the Blaschke product constructed by the zeros of G, and replace G by $\frac{G}{A}$, the left side of (3.9) increases because $A(z) < 1$, but the norm of G doesn't change.

If we replace G by $G^{\frac{2}{p}}$, we change the situation to $p = 2$. It's sufficient to prove there exists a absolute constant A_1 such that for any harmonic function u with boundary value $f \in L^2(0, 2\pi)$, we have

$$\int_D u(z)^2 d\mu(z) \leq A_1 C \int_0^{2\pi} f(\theta)^2 d\theta \tag{3.11}$$

because, assume (3.11),for $\forall G \in H^p, G = u + iv$, where u, v are hormonic functions, and

$$\int_{D_r} |G(z)|^2 d\mu(z) = \int_{D_r} |u(z)|^2 + |v(z)|^2 d\mu(z) \leq A_1 C \int_{\partial D_r} |u(z)|^2 + |v(z)|^2 d\mu(z) \to \|G\|_2^2$$

Back to the proof, we have

$$\int_D u(z)^2 d\mu(z) \le \sum_{n\in\mathbb{Z},0\le\nu\le 2^{n+1}-1} \mu(r_{\nu n})u(z_{\nu n})^2 \tag{3.12}$$

where $z_{\nu n}$ is the max point of u in $r_{\nu n}$, and $1-2^{-n} \le |z| \le 1-2^{-n-1}$. For $z_{\nu n}$, we denote ω_j^0 be the arc ω_{ij}, $j \le n$ which $arg(z_{\nu n})$ belongs,and denote $\omega_j^1 =$ arc $\omega_{i+1,j}$, $\omega_j^{-1} =$ arc $\omega_{i-1,j}$. By Proposition 2.3

$$u(z_{\nu n})^2 = (\frac{1}{2\pi}\int_0^{2\pi} \frac{1-r^2}{1+r^2-2rcos(\theta-\phi)} f(\phi)\mathrm{d}\phi)^2 \le (\sum_{j=0}^{n} \frac{1}{2\pi}\int_{\omega_j^{k_j}} \frac{1-r^2}{1+r^2-2r\cos(\theta-\phi)} f(\phi)d\phi)^2$$

where $z = re^{i\theta}$. The $k_j \in \{-1,0,1\}$ is chosen so that $\omega_j^{k_j}$ is disjoint with $\omega_{j+1}^{k_{j+1}}$. When on $\omega_j^{k_j}, |\phi-\theta| \in [\frac{\pi}{2^{(j+1)}}, \frac{\pi}{2}], j < n$, when $j = n, |\phi-\theta|$ can be zero. We have

$$\begin{aligned}
u(z_{\nu n})^2 &\le (\sum_{j=0}^{n} \frac{1}{2\pi}\int_{\omega_j^{k_j}} \frac{1-r^2}{(r-1)^2+2r(1-\cos(\theta-\phi)} f(\phi)d\phi)^2 \\
&= (\sum_{j=0}^{n} \frac{1}{2\pi}\int_{\omega_j^{k_j}} \frac{1-r^2}{(r-1)^2+4r(\sin(\frac{\theta-\phi}{2}))^2} f(\phi)d\phi)^2 \\
&\le (\sum_{j=0}^{n} \frac{1}{2\pi}\int_{\omega_j^{k_j}} \frac{1-r^2}{(r-1)^2+\frac{4}{\pi}r(\theta-\phi)^2} f(\phi)d\phi)^2 \qquad (sin\theta \le \frac{2}{\pi}\theta) \\
&\le (\sum_{j=0}^{n} \frac{1}{2\pi}\int_{\omega_j^{k_j}} \frac{1-r^2}{(r-1)^2+\frac{4}{\pi^2}r(\frac{\pi}{2^{j+1}})^2} f(\phi)d\phi)^2 \\
&\le (\sum_{j=0}^{n} \frac{1}{2\pi}\int_{\omega_j^{k_j}} \frac{1-(1-2^{-n-1})^2}{\frac{1}{2^{2n+2}}+4(1-2^{-n-1})\frac{1}{2^{2j+2}}} f(\phi)d\phi)^2 \\
&\le (\sum_{j=0}^{n} \frac{1}{2\pi}\int_{\omega_j^{k_j}} 2^{2j-n-1} f(\phi)d\phi)^2 \le \sum_{j=0}^{n} \frac{1}{16\pi^2} 2^{3j-2n} (\int_{\omega_j^{k_j}} f(\phi)d\phi)^2 \sum_{j=0}^{n} 2^j \\
&= A_3 \sum_{j=0}^{n} 2^{3j-n} (\int_{\omega_j^{k_j}} f(\phi)d\phi)^2 \le A_3 \sum_{k=-1}^{1}\sum_{j=0}^{n} 2^{3j-n} (\int_{\omega_j^{k_j}} f(\phi)d\phi)^2
\end{aligned}$$

with (3.12), we get

$$\begin{aligned}
\int_D u(z)^2 d\mu(z) &\le \sum_{n\in\mathbb{Z},0\le\nu\le 2^{n+1}-1} \mu(r_{\nu n})u(z_{\nu n})^2 \qquad (3.13)\\
&\le A_3 \sum_{n\in\mathbb{Z},0\le\nu\le 2^{n+1}-1} \mu(r_{\nu n})2^{2n} \sum_{k=-1}^{1}\sum_{j=0}^{n} 2^{3(j-n)}(\int_{\omega_j^k} f(\phi)d\phi)^2 \qquad (3.14)\\
&= A_3\pi^2 \sum_{j=0}^{\infty}\sum_{i=0}^{2^{j+1}-1} (\frac{2^{j+1}}{2\pi}\int_{\omega_{ij}} f(\phi)d\phi)^2 \sum_{n\geqslant j} 2^{j-n}\mu(r_{\nu n}) \qquad (3.15)\\
&= A_3\pi^2 \sum_{j=0}^{\infty}\sum_{i=0}^{2^{j+1}-1} \lambda_{ij}(\frac{2^{j+1}}{2\pi}\int_{\omega_{ij}} f(\phi)d\phi)^2 \qquad (3.16)
\end{aligned}$$

where ν is that $\arg z_{\nu n} \in \omega_{i-1,j}\bigcup\omega_{i,j}\bigcup\omega_{i+1,j}$. We denote $\lambda_{ij} = \sum_{n\geqslant j} 2^{j-n}\mu(r_{\nu n})$.
For every k, m, let $S_{\nu n} = \left\{re^{i\theta} : r \geqslant 1-\frac{1}{2^n}, \frac{\nu.2\pi}{2^{n+1}} \le \theta < \frac{(\nu+1).2\pi}{2^{n+1}}\right\}$

$$\begin{aligned}
\sum_{\omega_{ij}\subset\omega_{km}} \lambda_{ij} &= \sum_{r_{\nu n}\subset\omega_{km}} \mu(r_{\nu n}) \sum_{j\le n} 2^{-j} = \sum_{S_{\nu n}\subset\omega_{km}} 2^{m-n}\mu(S_{\nu n}) \qquad (3.17)\\
&\le \sum_{S_{\nu n}\subset\omega_{km}} 2^{m-n}C2^{-n} \le C\sum_{n\geqslant m} 2^{n-m}2^{m-n}2^{-n} \le A_5C2^{-m} \qquad (3.18)
\end{aligned}$$

Now given f, we modify λ_{ij}. If $\lambda_{i1} \geqslant \frac{1}{2}$ we do not change λ_{i1}. If $\lambda_{i1} < \frac{1}{2}$, we add λ_{i1} to $\lambda_{2i,2}$ or $\lambda_{2i+1,2}$ and replace λ_{i1} by zero. This can be done so the right hand side of (3.16) increase, since the corresponding two integrals for $j = 2$ have the integral for $j = 1$ as their mean-value. We do this for all λ_{i1} and obtain coefficients $\lambda_{ij}^{(1)}$. We now treat $\lambda_{i2}^{(1)}$ in the same manner adding $\lambda_{i2}^{(1)}$ to a suitable $\lambda_{k3}^{(1)}$ if $\lambda_{i2}^{(1)} < 2^{-2}$. Continuing in this way obtain coefficients λ_{ij}^* which satisfy (3.17) with A_5C replace by $(A_5 + 2)$. (3.17) imply that

$$2^{-m} \le \frac{\lambda_{im}^*}{2\pi} < \frac{1}{2\pi}(A_5+2)C2^{-m} \qquad (3.19)$$

(3.17) imply that

$$\sum_{\omega_{ij}^*\subset\omega_{km}} 2^{-j} \le \sum_{\omega_{ij}^*\subset\omega_{km}} \frac{\lambda_{ij}^*}{2\pi} \le (A_5+2)C2^{-m}$$

where ω_{ij}^* denote the ω_{ij} where $\lambda_{ij}^* \neq 0$. By Theorem3.1.

$$\sum_{j=1}^{\infty} \sum_{\substack{0\le i\le 2^{j+1}-1;\\ \lambda_{ij}^*\neq 0}} 2^j(\int_{\omega_{ij}^*} f(\phi)d\phi)^2 \le \frac{1}{2\pi}A(A_5+2)C\int_0^{2\pi} f(x)^2dx$$

(3.19) imply that

$$
\begin{aligned}
&A_3\pi^2\sum_{j=0}^{\infty}\sum_{i=0}^{2^{j+1}-1}\lambda_{ij}(\frac{2^{j+1}}{2\pi}\int_{\omega_{ij}}f(\phi)d\phi)^2\\
\leq &A_3\pi^2\sum_{j=0}^{\infty}\sum_{i=0}^{2^{j+1}-1}\lambda^*_{ij}(\frac{2^{j+1}}{2\pi}\int_{\omega^*_{ij}}f(\phi)d\phi)^2\\
\leq &A_3\pi^2\sum_{j=0}^{\infty}\sum_{\substack{0\leq i\leq 2^{j+1}-1;\\ \lambda^*_{ij}\neq 0}}C(A_5+2)2^{-j}(\frac{2^{j+1}}{2\pi}\int_{\omega^*_{ij}}f(\phi)d\phi)^2\\
= &A^*_5\sum_{j=0}^{\infty}\sum_{\substack{0\leq i\leq 2^{j+1}-1;\\ \lambda^*_{ij}\neq 0}}2^j(\int_{\omega^*_{ij}}f(\phi)d\phi)^2\\
\leq &\frac{1}{2\pi}A^*_5A(A_5+2)C\int_0^{2\pi}f(x)^2dx
\end{aligned}
$$

□

4 0-1 interpolations

Theorem 4.1. *$b_\nu, c_\nu \in \mathbb{C}$ mutually different, $|b_\nu| < 1, |c_\nu| < 1,\ \nu = 1, 2, ...$ $\sum\limits_{\nu=1}^{\infty}(1-b_\nu) < +\infty$, $\sum\limits_{\nu=1}^{\infty}(1-c_\nu) < +\infty$, $B(z) = \prod\limits_{\nu=1}^{\infty}\frac{b_\nu - z}{1-\overline{b_\nu}z}\frac{|b_\nu|}{b_\nu}$, $C(z) = \prod\limits_{\nu=1}^{\infty}\frac{c_\nu - z}{1-\overline{c_\nu}z}\frac{|c_\nu|}{c_\nu}$ ([1] p.157 Problem2 shows that $B(z)$ is holomorphic in $|z| < 1$ with zeros exactly at b_ν , and similar for $C(z)$)*

Then $\exists f \in B$ such that

$$f(b_\nu) = 0, \quad f(c_\nu) = 1, \quad \nu = 1, 2, ... \tag{4.1}$$

if and only if $\exists\delta > 0$ such that

$$|B(z)| + |C(z)| \geqslant \delta, \quad \forall|z| < 1 \tag{4.2}$$

If (4.2) holds, (4.1) can be solved with $||f|| \leqslant \delta^{-A_{11}}, \delta < 1/2$

Proof. Assume $f \in B$ satisfies (4.1), then

$$f = Bg, \quad f-1 = hC, \quad ||g|| = ||f||, \quad ||h|| = ||f-1|| \leqslant ||f|| + 1$$

$$\Rightarrow Bg - Ch = 1$$

$$\Rightarrow 1 \leqslant ||g||\,|B(z)| + ||h||\,|C(z)| \leqslant (||f|| + 1)(|B(z)| + |C(z)|)$$

$$\Rightarrow |B(z)| + |C(z)| \geqslant (||f|| + 1)^{-1}$$

For the converse, we need two propositions:

Proposition 4.1. *$a_1, ..., a_s, w_1, ..., w_s \in \mathbb{C},\ |a_\nu| < 1, \nu = 1, 2, ..., s,\ a_1, ..., a_s$ mutually different, $A(z) = \prod\limits_{\nu=1}^{s} \frac{a_\nu - z}{1-\overline{a_\nu} z} \frac{|a_\nu|}{a_\nu}$, then*

1.
$$inf\{||f||,\ f \in B,\ f(a_\nu) = w_\nu,\ \nu = 1, 2, ..., s\} = sup\{|\sum_{\nu=1}^{s} \frac{G(a_\nu)w_\nu}{A'(a_\nu)}|,\ G \in H^1, ||G||_1 = 1\} \tag{4.3}$$

2. *$\exists f_0 \in B$ such that $f_0(a_\nu) = w_\nu,\ \nu = 1, 2, ..., s$, and*
$$||f_0|| = inf\{||f||,\ f \in B,\ f(a_\nu) = w_\nu,\ \nu = 1, 2, ..., s\}$$

Proposition 4.2 (Carleson's lemma)**.** *$a_1, ..., a_s \in \mathbb{C}$ mutually different, $|a_\nu| < 1,\ \nu = 1, ..., s,\ A(z) = \prod\limits_{\nu=1}^{s} \frac{a_\nu - z}{1-\overline{a_\nu} z} \frac{|a_\nu|}{a_\nu},\ 0 < \epsilon \leqslant 1/4, 0 < \kappa < A_{14}$ (A_{14} is an absolute constant less than $1/8$ we'll determine later)*

Then there exist a finite number of disjoint regions $\Omega_1, ..., \Omega_p$ in $|z| < 1$ with rectifiable boundaries $\Gamma = \cup_{j=1}^{p} \partial\Omega_j$ satisfies

1. *$a_1, ..., a_s \in \bigcup_{j=1}^{p} \Omega_j$*
2. *$\forall z \in \Gamma,\ \epsilon \leqslant |A(z)| \leqslant \epsilon^\kappa$*
3. *Let μ be the measure on $|z| < 1$ defined by $\mu(E) =$ arc length of $E \cap \Gamma$ (E is a Borel subset of $\{z, |z| < 1\}$) ,then $\forall l \in (0, 1],\ \theta_0 \in \mathbb{R}$,*
$$\mu(\{re^{i\theta} \mid 1 - l \leqslant r \leqslant 1,\, \theta_0 \leqslant \theta \leqslant \theta_0 + l\}) \leqslant A_{15}\epsilon^{-2} l \tag{4.4}$$

Now assume (4.2) holds with $\delta < 1/2$
for $s \in \mathbb{N}$, define

$$B_s(z) = \prod_{\nu=1}^{s} \frac{b_\nu - z}{1 - \overline{b_\nu} z} \frac{|b_\nu|}{b_\nu},\ C_s(z) = \prod_{\nu=1}^{s} \frac{c_\nu - z}{1 - \overline{c_\nu} z} \frac{|c_\nu|}{c_\nu}$$

Choose $\epsilon \in (0, 1/4]$, $\kappa \in (0, A_{14})$ so that $\epsilon^\kappa = \delta/2$ and let Γ be the curves in Proposition4.2 with respect to $c_1, ..., c_s,\ \epsilon, \kappa$, then for $z \in \cup\overline{\Omega_j}$,

$$\epsilon < |C_s(z)| \leqslant sup\{|C_s(z)|, z \in \Gamma\} \leqslant \epsilon^\kappa \leqslant \frac{\delta}{2} \tag{4.5}$$

$$\Rightarrow |B_s(z)| \geqslant \delta - |C_s(z)| \geqslant \frac{\delta}{2} > 0 \tag{4.6}$$

For $G \in H^1$

$$\frac{1}{2\pi i} \int_\Gamma \frac{G(z)}{B_s(z)C_s(z)} \mathrm{d}z = \sum_{\nu=1}^{s} \frac{G(c_\nu)}{B_s(c_\nu)C_s'(c_\nu)} = \sum_{\nu=1}^{s} \frac{G(c_\nu)}{\frac{\mathrm{d}}{\mathrm{d}z}(B_s(z)C_s(z))|_{z=c_\nu}}$$

By Proposition4.1, suppose f_s in B such that $f_s(b_\nu) = 0, f_s(c_\nu) = 1,\ \nu = 1, 2, ..., s$, and

$$\begin{aligned}||f_s|| &= inf\{||f||,\ f(b_\nu) = 0, f(c_\nu) = 1,\ \nu = 1, 2, ..., s\}\\ &= sup\{|\sum_{\nu=1}^{s} \frac{G(c_\nu)}{\frac{\mathrm{d}}{\mathrm{d}z}(B_s(z)C_s(z))|_{z=c_\nu}}|,\ G \in H_1, ||G||_1 = 1\}\\ &= sup\{|\frac{1}{2\pi i}\int_\Gamma \frac{G(z)}{B_s(z)C_s(z)}\mathrm{d}z|,\ G \in H_1, ||G||_1 = 1\}\\ &\leqslant (\frac{\delta}{2})^{-1}\epsilon^{-1}\frac{1}{2\pi}sup\{|\int_\Gamma |G(z)|\,|\mathrm{d}z|,\ G \in H_1, ||G||_1 = 1\} \quad (by\ (4.5), (4.6))\\ &\leqslant (\frac{\delta}{2})^{-1}\epsilon^{-1}\frac{1}{2\pi}A_{16}A_{15}\epsilon^{-2}||G||_1 \quad (by\ Proposition 4.2\ 3,\ Theorem 3.2)\\ &= \frac{A_{15}A_{16}}{2\pi}(\frac{\delta}{2})^{-1-3/\kappa}\end{aligned}$$

Notice that f_s is independent of the choice of ϵ and κ, so let $\kappa = A_{14}/2,\ \epsilon = (\delta/2)^{\kappa^{-1}}$

$$||f_s|| \leqslant \frac{A_{15}A_{16}}{2\pi}(\frac{\delta}{2})^{-1-6/A_{14}} \leqslant \delta^{-A_{11}}$$

By [1]p.225 Theorem 3.3 and Arzela-Ascoli theorem, a subsequence of $\{f_s\}_{s=1}^{+\infty}$ convergent to $f \in B$, then $||f|| \leqslant \delta^{-A_{11}},\ f(b_\nu) = \lim_{s\to+\infty} f_s(b_\nu) = 0,\ f(c_\nu) = \lim_{s\to+\infty} f_s(c_\nu) = 1$ □

Proof of Proposition 4.1. For $f \in B$ satisfying $f(a_\nu) = w_\nu,\ \nu = 1, 2, ..., s$ and $G \in H^1, ||G||_1 = 1$,

$$\begin{aligned}||f|| = ||\frac{f}{A}|| &= \frac{1}{2\pi}\int_0^{2\pi} |G(e^{i\theta})|\,||\frac{f}{A}||\mathrm{d}\theta\\ &\geqslant \frac{1}{2\pi}|\int_0^{2\pi} \frac{G(e^{i\theta})f(e^{i\theta})}{A(e^{i\theta})}ie^{i\theta}\mathrm{d}\theta|\\ &= |\frac{1}{2\pi i}\int_{|z|=1} \frac{G(z)f(z)}{A(z)}\mathrm{d}z|\\ &= |\sum_{\nu=1}^{s} \frac{G(a_\nu)f(a_\nu)}{A'(a_\nu)}| = |\sum_{\nu=1}^{s} \frac{G(a_\nu)w_\nu}{A'(a_\nu)}|\end{aligned}$$

$\therefore inf\{||f||,\ f \in B,\ f(a_\nu) = w_\nu,\ \nu = 1, 2, ..., s\} \geqslant sup\{|\sum_{\nu=1}^{s} \frac{G(a_\nu)w_\nu}{A'(a_\nu)}|,\ G \in H^1, ||G||_1 = 1\}$

For the remaining proof, we need two lemmas:

Lemma 4.1. $1 \leqslant q \leqslant +\infty,\ g_n \in H^q,\ n = 1, 2, ...$. *If* g_n *converges uniformly to* f *on any compact subset of* $|z| < 1$, *then* $||f||_q \leqslant \liminf_{n\to\infty} ||g_n||_q$

Proof. For any $0 < r < 1$, because g_n converges uniformly to f on $|z| = r$,

$$(\frac{1}{2\pi}\int_0^{2\pi} |f_q(re^{i\theta})|^q\mathrm{d}\theta)^{1/q} = \liminf_{n\to\infty}(\frac{1}{2\pi}\int_0^{2\pi} |g_n(re^{i\theta})|^q\mathrm{d}\theta)^{1/q} \leqslant \liminf_{n\to\infty} ||g_n||_q$$

$$\Rightarrow ||f||_q \leqslant \liminf_{n\to\infty} ||g_n||_q$$

□

Lemma 4.2. $1 \leqslant q \leqslant +\infty,\ g_n \in H^q,\ ||g_n||_q \leqslant M,\ n = 1, 2, \ldots$. *Then there exists a sequence of positive integers* $\{n_k\}$ *such that* g_{n_k} *converges uniformly to a holomorphic function* f *on any compact subset of* $|z| < 1$

Proof. For any $0 < r < 1$, by Proposition 2.3

$$g_n(re^{i\theta}) = \frac{1}{2\pi}\int_0^{2\pi} \frac{1 - r^2}{1 + r^2 - 2rcos(\theta - \phi)} g_n(e^{i\phi})\mathrm{d}\phi$$

$$\begin{aligned}\Rightarrow |g_n(re^{i\theta})| &\leqslant \frac{1}{2\pi}\int_0^{2\pi} \frac{1 - r^2}{1 + r^2 - 2r}|g_n(e^{i\phi})|\mathrm{d}\phi \\ &\leqslant \frac{1+r}{1-r}(\frac{1}{2\pi}\int_0^{2\pi}|g_n(e^{i\phi})|^q)^{1/q} = \frac{1+r}{1-r}||g_n||_q \leqslant \frac{1+r}{1-r}M \\ \Rightarrow sup\{|g_n(z)|,\ |z| \leqslant r\} &= sup\{|g_n(re^{i\theta})|,\ \theta \in \mathbb{R}\} \leqslant \frac{1+r}{1-r}M\end{aligned}$$

Therefore $\{g_n\}$ uniformly bounded on $|z| \leqslant r$.

By [1] p.225 Theorem 3.3 and Arzela-Ascoli Theorem, there exists a subsequence of $\{g_n\}$ converging uniformly on any compact subset of $|z| < 1$ to a holomorphic function f. □

For $1 \leqslant q \leqslant +\infty$, denote $m_p = inf\{||f||_q,\ f \in H^q,\ f(a_\nu) = w_\nu,\ \nu = 1, 2, ..., s\}$

Choose $g_1, g_2, ... \in H^q$ such that $g_n(a_\nu) = w_\nu,\ \nu = 1, 2, ..., s,\ n = 1, 2, ...$ and $||g_n||_q$ decreases and tends to m_q as n tends to infinity.

By Lemma 4.2, we can assume $g_n \rightrightarrows f_q$ on any compact subset of $|z| < 1$. Then $f_q(a_\nu) = w_\nu,\ \nu = 1, 2, ..., s$, and by Lemma 4.1,

$$||f_q||_q \leqslant m_q = inf\{||f||_q,\ f \in H^q,\ f(a_\nu) = w_\nu,\ \nu = 1, 2, ..., s\}$$

$$\Rightarrow ||f_q|| = m_q$$

Now let $3 \leqslant q < +\infty$.

For $\eta, \phi \in \mathbb{R},\ g \in H^q,\ \widetilde{f}(z) = f_q(z) + \eta e^{i\phi}A(z)g(z)$ satisfies $\widetilde{f} \in H^q$ and $\widetilde{f}(a_\nu) = w_\nu,\ \nu = 1, 2, ..., s$. Therefore

$$||f_q||_q \leqslant ||f_q + \eta e^{i\phi}Ag||_q \tag{4.7}$$

$$\begin{aligned}|f_q(e^{i\theta}) + \eta e^{i\phi}A(e^{i\theta})g(e^{i\theta})|^q &= ((f_q + \eta e^{i\phi}Ag)(\overline{f}_q + \eta e^{-i\phi}\overline{A}\overline{g}))^{q/2} \\ &= (|f_q|^2 + \eta(f_q\overline{A}\overline{g}e^{-i\phi} + \overline{f}_qAge^{i\phi}) + \eta^2|g|^2)^{q/2} \\ \Rightarrow \frac{\mathrm{d}|f_q + \eta e^{i\phi}Ag|^q}{\mathrm{d}\eta} &= \frac{q}{2}|f_q + \eta e^{i\phi}Ag|^{q-2}(f_q\overline{A}\overline{g}e^{-i\phi} + \overline{f}_qAge^{i\phi} + 2\eta|g|^2)\end{aligned}$$

By mean value theorem,

$$\begin{aligned}\left.\frac{\mathrm{d}||f_q + \eta e^{i\phi}Ag||_q^q}{\mathrm{d}\eta}\right|_{\eta=0} &= \lim_{\eta\to 0}\frac{\int_{|z|=1}(|f_q + \eta e^{i\phi}Ag|^q - |f_q|^q)\mathrm{d}\theta}{2\pi\eta} \\ &= \lim_{\eta\to 0}\int_{|z|=1}\frac{q}{4\pi}|f_q + \epsilon\eta e^{i\phi}Ag|^{q-2}(f_q\overline{A}\overline{g}e^{-i\phi} + \overline{f}_qAge^{i\phi} + 2\epsilon\eta|g|^2)\mathrm{d}\theta\end{aligned}$$

where $\epsilon \in (0,1)$ relevant to η and θ.

$$\begin{aligned}
&|f_q + \epsilon\eta e^{i\phi}Ag|^{q-2}|f_q\overline{A}\overline{g}e^{-i\phi} + \overline{f}_qAge^{i\phi} + 2\epsilon\eta|g|^2| \\
&\leqslant (|f_q| + \eta|g|)^{q-2}(2|f_q|\,|g| + 2\eta|g|^2) \\
&\leqslant 2^{q-2}(|f_q|^{q-2} + \eta^{q-2}|g|^{q-2})(|f_q|\,|g| + \eta|g|^2) \\
&= 2^{q-2}(|f_q|^{q-1}|g| + \eta^{q-1}|g|^q + \eta|f_q|^{q-2}|g|^2 + \eta^{q-2}|f_q|\,|g|^{q-1}) \\
&\leqslant 2^{q-2}(\frac{(q-1)|f_q|^q + |g|^q}{q} + \eta\frac{(q-2)|f_q|^q + 2|g|^q}{q} + \eta^{q-2}\frac{|f_q|^q + (q-1)|g|^q}{q} + \eta^{q-1}|g|^q) \\
&\leqslant C(|f_q|^q + |g|^q)
\end{aligned}$$

where C is independent of η.

The boundary function of f_q and g is L^q, so by bounded convergence theorem,

$$\begin{aligned}
\left.\frac{\mathrm{d}||f_q + \eta e^{i\phi}Ag||_q^q}{\mathrm{d}\eta}\right|_{\eta=0} &= \lim_{\eta\to 0}\int_{|z|=1}\frac{q}{4\pi}|f_q + \epsilon\eta e^{i\phi}Ag|^{q-2}(f_q\overline{A}\overline{g}e^{-i\phi} + \overline{f}_qAge^{i\phi} + 2\epsilon\eta|g|^2)\mathrm{d}\theta \\
&= \int_{|z|=1}\frac{q}{4\pi}|f_q|^{q-2}(f_q\overline{A}\overline{g}e^{-i\phi} + \overline{f}_qAge^{i\phi})\mathrm{d}\theta
\end{aligned}$$

By (4.7),

$$\int_{|z|=1}|f_q|^{q-2}(f_q\overline{A}\overline{g}e^{-i\phi} + \overline{f}_qAge^{i\phi})\mathrm{d}\theta = 0, \quad \forall\phi\in\mathbb{R}$$

Let $\phi = 0, \frac{\pi}{2}$, we get

$$\int_{|z|=1}|f_q|^{q-2}(f_q\overline{A}\overline{g} + \overline{f}_qAg)\mathrm{d}\theta = 0 \tag{4.8}$$

$$\int_{|z|=1}|f_q|^{q-2}(-if_q\overline{A}\overline{g} + i\overline{f}_qAg)\mathrm{d}\theta = 0$$

$$\Rightarrow \int_{|z|=1}|f_q|^{q-2}(-f_q\overline{A}\overline{g} + \overline{f}_qAg)\mathrm{d}\theta = 0 \tag{4.9}$$

Add up (4.8) and (4.9),

$$\int_{|z|=1}|f_q|^{q-2}\overline{f}_qAg\mathrm{d}\theta = 0, \qquad \forall g\in H^q$$

Let $g(z) = z^n,\ n = 0, 1, ...,$

$$\int_0^{2\pi}|f_q(e^{i\theta})|^{q-2}\overline{f}_q(e^{i\theta})A(e^{i\theta})e^{in\theta}\mathrm{d}\theta = 0, \quad n = 0, 1, ...$$

Notice that $|f_q(e^{i\theta})|^{q-2}\overline{f}_q(e^{i\theta})A(e^{i\theta})e^{-i\theta}$ is L^1 on $\theta\in[0,2\pi]$. By Proposition 2.4, There is a function $F_q\in H^1$ such that $F_q = m_q^{1-q}z^{-1}|f_q|^{q-2}\overline{f}_qA$ on $|z| = 1$. We have

$$||F_q||_1 = \frac{1}{2\pi}\int_0^{2\pi}m_q^{1-q}|f_q(e^{i\theta})|^{q-1}\mathrm{d}\theta \leqslant m_q^{1-q}(\frac{1}{2\pi}\int_0^{2\pi}|f_q(e^{i\theta})|^q\mathrm{d}\theta)^{\frac{q-1}{q}} = m_q^{1-q}m_q^{q-1} = 1$$

$$
\begin{aligned}
m_q = m_q^{1-q}\frac{1}{2\pi}\int_0^{2\pi}|f_q(e^{i\theta})|^q\mathrm{d}\theta &= \frac{1}{2\pi}\int_0^{2\pi}\frac{e^{i\theta}F_q(e^{i\theta})f_q(e^{i\theta})}{A(e^{i\theta})}\mathrm{d}\theta \\
= \frac{1}{2\pi i}\int_{|z|=1}\frac{F_q(z)f_q(z)}{A(z)}\mathrm{d}z &= \sum_{\nu=1}^{s}\frac{F_q(a_\nu)w_\nu}{A'(a_\nu)}
\end{aligned}
\tag{4.10}
$$

By Proposition 2.1, for any $1 \leqslant q_1 \leqslant q_2 \leqslant +\infty$,

$$m_{q_2} = ||f_{q_2}||_{q_2} \geqslant ||f_{q_2}||_{q_1} \geqslant m_{q_1}$$

$\Rightarrow \lim\limits_{q\to+\infty} m_q$ exists and $m_\infty \geqslant \lim\limits_{q\to+\infty} m_q$

By Proposition 2.1,

$$||f_q||_1 \leqslant ||f_q||_q = m_q \leqslant m_\infty, \quad \forall\, 3 \leqslant q < +\infty$$

By Lemma 4.2, there exists an increasing sequence $\{q_k\}$ of positive real numbers greater than 3 such that q_k tends to infinity and f_{q_k} converges uniformly to f_0 on any compact subset of $|z| < 1$. Then $f_0(a_\nu) = w_\nu,\ \nu = 1, 2, ..., s$

By Lemma 4.1,

$$||f_0||_p \leqslant \lim_{k\to+\infty}||f_{q_k}||_p \leqslant \lim_{k\to+\infty}||f_{q_k}||_{q_k} = \lim_{k\to+\infty} m_{q_k} = \lim_{q\to+\infty} m_q \quad \forall 1 \leqslant p < +\infty$$

$$\Rightarrow m_\infty \leqslant ||f_0||_\infty = \lim_{p\to+\infty}||f_0||_p \leqslant \lim_{q\to+\infty} m_q \leqslant m_\infty$$

$$\Rightarrow ||f_0|| = m_\infty = \lim_{q\to+\infty} m_q \tag{4.11}$$

Notice that $||F_{q_k}||_1 \leqslant 1, k = 1, 2, \dots$, by Lemma 4.2, we can assume F_{q_k} converges uniformly to F_0 on any compact subset of $|z| < 1$. By Lemma 4.1, $||F_0||_1 \leqslant 1$

Replace q in (4.10) with q_k and let k tends to infinity, and by (4.11), we get

$$m_\infty = \sum_{\nu=1}^{s}\frac{F_0(a_\nu)w_\nu}{A'(a_\nu)}$$

$$
\begin{aligned}
&\Rightarrow inf\{||f||,\ f \in B,\ f(a_\nu) = w_\nu,\ \nu = 1, 2, ..., s\} \\
&= m_\infty = \sum_{\nu=1}^{s}\frac{F_0(a_\nu)w_\nu}{A'(a_\nu)} \leqslant sup\{|\sum_{\nu=1}^{s}\frac{G(a_\nu)w_\nu}{A'(a_\nu)}|,\ G \in H^1, ||G||_1 = 1\}
\end{aligned}
$$

$\therefore inf\{||f||,\ f \in B,\ f(a_\nu) = w_\nu,\ \nu = 1, 2, ..., s\} = sup\{|\sum_{\nu=1}^{s}\frac{G(a_\nu)w_\nu}{A'(a_\nu)}|,\ G \in H^1, ||G||_1 = 1\}$

And f_0 satisfies

$$f_0(a_\nu) = w_\nu,\ \nu = 1, 2, ..., s$$

$$||f_0|| = m_\infty = inf\{||f||,\ f \in B,\ f(a_\nu) = w_\nu,\ \nu = 1, 2, ..., s\}$$

□

5 A General Interpolation Theorem

Theorem 5.1. *Let $\{z_\nu\}_1^\infty$ be a sequence of mutually different complex numbers in $|z| < 1$ such that*

$$\prod_{\nu\in\mathbb{N},\, \nu\neq\mu} \left|\frac{z_\nu - z_\mu}{1 - z_\nu \overline{z_\mu}}\right| \geqslant \delta_1 > 0, \quad \mu = 1, 2, ... \tag{5.1}$$

Then for any sequence $\{w_\nu\}$ of complex numbers in $|z| \leqslant 1$, there is a function $f \in B$ such that $f(z_\nu) = w_\nu$

Remark 5.1. *(5.1) implies* $\sum\limits_{\nu=0}^{\infty}(1 - |z_\nu|) < +\infty$

Remark 5.2. *[3] proved that for any bounded complex number sequence $\{w_\nu\}$, there is a function $f \in B$ such that $f(z_\nu) = w_\nu$ if and only if (5.1) holds.*

Denote $\rho(z, w) = |\frac{z-w}{1-z\overline{w}}| \quad (|z| < 1,\, |w| < 1)$

Lemma 5.1. *([1] p.251 Exercise 13) If f is an automorphism of $\{z,\, |z| < 1\}$, then $\rho(f(z), f(w)) = \rho(z, w)$ for all $|z| < 1,\, |w| < 1$*

Lemma 5.2. *For $|z_1|, z_2|, |w| < 1$,*

$$\frac{\Big|\rho(z_1, w) - \rho(z_2, w)\Big|}{1 - \rho(z_1, w)\,\rho(z_2, w)} \leqslant \rho(z_1, z_2) \leqslant \frac{\rho(z_1, w) + \rho(z_2, w)}{1 - \rho(z_1, w)\,\rho(z_2, w)}$$

Proof. By Lemma 5.1, we can assume $w = 0$. It's sufficient to prove that

$$\frac{\Big||z| - |w|\Big|}{1 - |z|\,|w|} \leqslant \rho(z, w) \leqslant \frac{|z| + |w|}{1 - |z|\,|w|}, \qquad \forall\, |z| < 1,\, |w| < 1$$

Denote $C(z_0, r) = \{z \in \mathbb{C},\, |z - z_0| = r\}, \quad z_0 \in \mathbb{C},\ r \in \mathbb{R}_+$

Fix w and $|z| = r < 1$. Let z range on $C(z_0, r)$. Write $\rho(z, w)$ in a form of Mobiüs transformation:

$$\rho(z, w) = \left|\frac{z - w}{1 - z\overline{w}}\right| = \left|\frac{\frac{1}{\overline{w}} - w}{1 - z\overline{w}} - \frac{1}{\overline{w}}\right|$$

$$\begin{aligned}
z \in C(z_0, r) &\Rightarrow 1 - z\overline{w} \in C(1, r|w|) \\
&\Rightarrow \frac{1}{1 - z\overline{w}} \in C\left(\frac{1}{2}(\frac{1}{1 - r|w|} + \frac{1}{1 + r|w|}), \frac{1}{2}(\frac{1}{1 - r|w|} - \frac{1}{1 + r|w|})\right) \\
&\qquad = C\left(\frac{1}{1 - r^2|w|^2}, \frac{r|w|}{1 - r^2|w|^2}\right) \\
&\Rightarrow \frac{\frac{1}{\overline{w}} - w}{1 - z\overline{w}} - \frac{1}{\overline{w}} \in C\left(\frac{1 - |w|^2}{\overline{w}}\frac{1}{1 - r^2|w|^2} - \frac{1}{\overline{w}}, \frac{1 - |w|^2}{|w|}\frac{r|w|}{1 - r^2|w|^2}\right) \\
&\qquad = C\left(\frac{(r^2 - 1)|w|^2}{\overline{w}(1 - r^2|w|^2)}, \frac{r(1 - |w|^2)}{1 - r^2|w|^2}\right)
\end{aligned}$$

$$\Rightarrow |\rho(z,w)| \geqslant \left| \left| \frac{(r^2-1)|w|^2}{\overline{w}(1-r^2|w|^2)} \right| - \frac{r(1-|w|^2)}{1-r^2|w|^2} \right|$$

$$= \left| \frac{(1+r|w|)(r-|w|)}{1-r^2|w|^2} \right| = \frac{\big||z|-|w|\big|}{1-|z|\,|w|}$$

$$|\rho(z,w)| \leqslant \left| \frac{(r^2-1)|w|^2}{\overline{w}(1-r^2|w|^2)} \right| + \frac{r(1-|w|^2)}{1-r^2|w|^2} = \frac{|z|+|w|}{1-|z|\,|w|}$$

□

Proof of Theorem 5.1. Consider an arbitrary decomposition of $\{z_\nu,\ \nu = 1,2,...,s\}$ into disjoint sets I_B and I_C. Let

$$B(z) = \prod_{z_\nu \in I_B} \frac{z_\nu - z}{1-\overline{z_\nu}z}, \quad C(z) = \prod_{z_\nu \in I_C} \frac{z_\nu - z}{1-\overline{z_\nu}z}$$

Denote

$$D_\nu = \{z \in \mathbb{C},\ |z|<1,\ \rho(z,z_\nu) < \frac{\delta_1}{3}\}, \quad \nu = 1,2,...,s$$

$$V_B = \{z \in \mathbb{C},\ |z| \leqslant 1\} \backslash \left(\bigcup_{z_\nu \in I_B} D_\nu \right)$$

$$V_C = \{z \in \mathbb{C},\ |z| \leqslant 1\} \backslash \left(\bigcup_{z_\nu \in I_C} D_\nu \right)$$

(5.1) implies that

$$\delta_1 < 1, \qquad \rho(z_\nu, z_\mu) \geqslant \delta_1, \quad \forall\, \nu \neq \mu \tag{5.2}$$

Suppose $\exists\, \nu \neq \mu$ such that $D_\nu \cap D_\mu \neq \emptyset$, choose $z \in D_\nu \cap D_\mu$, then

$$\rho(z,z_\nu) < \frac{\delta_1}{3}, \quad \rho(z,z_\mu) < \frac{\delta_1}{3}$$

By Lemma 5.2,

$$\rho(z_\mu, z_\nu) \leqslant \frac{\rho(z,z_\nu) + \rho(z_\mu,z)}{1-\rho(z,z_\nu)\,\rho(z_\mu,z)} \leqslant \frac{\frac{\delta_1}{3}+\frac{\delta_1}{3}}{1-\frac{1}{3}\times\frac{1}{3}} = \frac{3}{4}\delta_1 < \delta_1$$

contradiction to (5.2).

Therefore D_ν are mutually disjoint, hence

$$V_B \cup V_C = \{z \in \mathbb{C},\ |z| \leqslant 1\} \tag{5.3}$$

By definition of V_B, $B(z)$ has no zeros in V_B. So

$$\min_{z \in V_B} |B(z)| = \min_{z \in \partial V_B} |B(z)| \tag{5.4}$$

Since

$$\partial V_B = \{z \in \mathbb{C},\ |z| = 1\} \cup \left(\bigcup_{z_\nu \in I_B} \{z \in \mathbb{C},\ |z|<1,\ \rho(z,z_\nu) = \frac{\delta_1}{3}\} \right)$$

Now calculate $|B(z)|$ on each item above. First, $|B(z)| = 1$ if $|z| = 1$. Now suppose $\rho(z, z_\nu) = \frac{\delta_1}{3},\ z_\nu \in I_B$. By Lemma 5.2 and (5.2), for any $\mu \neq \nu$,

$$\begin{aligned}\rho(z, z_\mu) &\geqslant \frac{\rho(z_\nu, z_\mu) - \rho(z_\nu, z)}{1 - \rho(z_\nu, z_\mu)\,\rho(z_\nu, z)} \\ &= \frac{\big(1 - \rho(z_\nu, z_\mu)\big)\big(\rho(z_\nu, z_\mu) - \rho(z_\nu, z)(1 + \rho(z_\nu, z_\mu) + \rho(z_\nu, z_\mu)^2)\big)}{1 - \rho(z_\nu, z_\mu)\,\rho(z_\nu, z)} + \rho(z_\nu, z_\mu)^2 \\ &\geqslant \frac{\big(1 - \rho(z_\nu, z_\mu)\big)\,\big(\rho(z_\nu, z_\mu) - \frac{\delta_1}{3} \cdot 3\big)}{1 - \rho(z_\nu, z_\mu)\,\rho(z_\nu, z)} + \rho(z_\nu, z_\mu)^2 \\ &\geqslant \rho(z_\nu, z_\mu)^2\end{aligned}$$

$$\begin{aligned}\Rightarrow |B(z)| &= \rho(z, z_\nu) \prod_{z_\mu \in I_B,\, \mu \neq \nu} \rho(z, z_\mu) \\ &\geqslant \frac{\delta_1}{3} \prod_{\mu \neq \nu} \rho(z, z_\mu) \geqslant \frac{\delta_1}{3} \prod_{\mu \neq \nu} \rho(z_\nu, z_\mu)^2 \geqslant \frac{\delta_1^3}{3}\end{aligned}$$

By (5.4),

$$|B(z)| \geqslant \frac{\delta_1^3}{3}, \qquad \forall\, z \in V_B \tag{5.5}$$

Same for $C(z)$, we have

$$|C(z)| \geqslant \frac{\delta_1^3}{3}, \qquad \forall\, z \in V_C \tag{5.6}$$

By (5.3), (5.5), (5.6),

$$|B(z)| + |C(z)| \geqslant \frac{\delta_1^3}{3}, \qquad \forall\, |z| \leqslant 1 \tag{5.7}$$

Now, let $w_\nu = u_\nu + iv_\nu$. We arrange $\{z_\nu\}_1^s$ so that $u_1 \leqslant u_2 \leqslant ... \leqslant u_s$ and define $u_0 = 0$. By Theorem 4.1 and (5.7), choose $f_\nu \in B,\ \nu = 1, ..., s$ such that $f_\nu(z_i) = 1,\ 1 \leqslant i \leqslant \nu,\quad f_\nu(z_i) = 0, \nu < i \leqslant s$ and $||f_\nu|| \leqslant (\frac{\delta_1^3}{3})^{-A_{11}}$. Then

$$g(z) = \sum_{\nu=1}^{s} (u_\nu - u_{\nu-1}) f_\nu(z)$$

satisfies $g(z_\nu) = u_\nu,\ \nu = 1, 2, ..., s$, and

$$||g|| \leqslant \sum_{\nu=1}^{s} |u_\nu - u_{\nu-1}|\, ||f_\nu|| \leqslant (\frac{\delta_1^3}{3})^{-A_{11}} \left(|u_1| + \sum_{\nu=2}^{s} (u_\nu - u_{\nu-1})\right) \leqslant 3(\frac{\delta_1^3}{3})^{-A_{11}}$$

Similarly, there exists $h \in B$ such that $h(z_\nu) = v_\nu,\ \nu = 1, 2, ..., s$ and $||h|| \leqslant 3(\frac{\delta_1^3}{3})^{-A_{11}}$

Then $g + ih$ satisfies $g(z_\nu) + ih(z_\nu) = w_\nu,\ \nu = 1, 2, ..., s$ and $||g + ih|| \leqslant 6(\frac{\delta_1^3}{3})^{-A_{11}}$. Let $s \to \infty$ and choose a convergent sequence. □

6 Construction of Γ

In section 6,7 and 8, our aim is to prove Proposition 4.2 rewritten below:

Proposition 4.2 $a_1, ..., a_s \in \mathbb{C}$ mutually different, $|a_\nu| < 1$, $\nu = 1, ..., s$, $A(z) = \prod_{\nu=1}^{s} \frac{a_\nu - z}{1-\overline{a_\nu}z}\frac{|a_\nu|}{a_\nu}$, $0 < \epsilon \leqslant 1/4$, $0 < \kappa < A_{14}$ (A_{14} is an absolute constant less than 1/8 we'll determine later)

Then there exist a finite number of disjoint regions $\Omega_1, ..., \Omega_p$ in $|z| < 1$ with rectifiable boundaries $\Gamma = \cup_{j=1}^{p} \partial\Omega_j$ satisfies

1. $a_1, ..., a_s \in \bigcup_{j=1}^{p} \Omega_j$
2. $\forall z \in \Gamma$, $\epsilon \leqslant |A(z)| \leqslant \epsilon^\kappa$
3. Let μ be the measure on $|z| < 1$ defined by $\mu(E) =$ arc length of $E \cap \Gamma$ (E is a Borel subset of $\{z, |z| < 1\}$) ,then $\forall l \in (0, 1]$, $\theta_0 \in \mathbb{R}$,

$$\mu(\{re^{i\theta} \mid 1 - l \leqslant r \leqslant 1, \theta_0 \leqslant \theta \leqslant \theta_0 + l\}) \leqslant A_{15}\epsilon^{-2}l \tag{6.1}$$

We will follow the notation from Proposition 4.2 in section 6 and 8.

Recall

$$r_{\nu n} = \left\{z : \frac{1}{2^{n+1}} < 1 - |z| \leqslant \frac{1}{2^n}, \frac{\nu \cdot 2\pi}{2^{n+1}} \leqslant \arg z < \frac{(\nu+1) \cdot 2\pi}{2^{n+1}}\right\}$$

Choose $N \in \mathbb{Z}_+$ such that

$$\frac{\epsilon}{2(2\pi + 1)} \leqslant 2^{-N} < \frac{\epsilon}{2\pi + 1} \tag{6.2}$$

Divide $r_{\nu n}$ into 2^{2N} regions:

$$\left\{z : \frac{1 + \frac{k}{2^N}}{2^{n+1}} < 1 - |z| \leqslant \frac{1 + \frac{k+1}{2^N}}{2^{n+1}}, \frac{2\pi(\nu + \frac{l}{2^N})}{2^{n+1}} \leqslant \arg z < \frac{2\pi(\nu + \frac{l+1}{2^N})}{2^{n+1}}\right\}, \quad k, l = 0, 1, ..., 2^N - 1$$

Denote the 2^{2N} regions above by $r_{\nu n}(i)$, $i = 1, 2, ..., 2^{2N}$

For $0 < \delta < 1$, define

$$a(\delta) = \{z : \ |z| < 1, |A(z)| < \epsilon\}$$

$$b(\delta) = \{z : \ |z| \leqslant 1, |A(z)| > \epsilon\}$$

By [1] p.251 Exercise 13 (b), $A'(z) \leqslant \frac{1}{1-|z|^2} \leqslant \frac{1}{1-|z|}$ for $|z| < 1$. Therefore for $z_1, z_2 \in r_{\nu n}(i)$, by mean value inequality and (6.2)

$$\begin{aligned} |A(z_1) - A(z_2)| &\leqslant (\frac{1}{2^{n+1+N}} + \frac{2\pi}{2^{n+1+N}}) \sup_{z \in r_{\nu n}(i)} |A'(z)| \\ &\leqslant \frac{1 + 2\pi}{2^{n+1+N}} \frac{1}{2^{-n-1}} < (1 + 2\pi)\frac{\epsilon}{2\pi + 1} = \epsilon \end{aligned} \tag{6.3}$$

Define

$$\alpha = \bigcup_{r_{\nu n}(i) \cap a(\epsilon) \neq \emptyset} r_{\nu n}(i)$$

By (6.3),

$$a(\epsilon) \subset \alpha \subset a(2\epsilon) \tag{6.4}$$

Denote $\beta = b(\epsilon^\kappa)$, then $\alpha \cap \beta = \emptyset$ because $2\epsilon < \epsilon^\kappa$

We first construct a subset of $|z| < 1$, called P, consisting of some boundaries of $r_{\nu n}(i)$ and separating α and β. Then choose a subset of P, called Γ such that Γ is the boundary of $\bigcup \partial\Omega_j$ and $\alpha \subset \bigcup \partial\Omega_j$ and $(\bigcup \partial\Omega_j) \cap \beta = \emptyset$

Assume $P = \emptyset$ first and add lines into P by the laws below:

1. Add all boundaries of $r_{\nu 0}(i),\ v = 0, 1,\ i = 1, 2, ..., 2^{2N}$ and the circle $|z| = 1$ into P.

2. (a) If r_{01} intersects β

   Add the arc $|z| = 1/2$, $0 \leqslant arg(z) \leqslant \pi/2$ and the segments $1/2 \leqslant |z| \leqslant 1$, $arg(z) = 0, \pi/2$ into P. Say $\{z :\ 1/2 \leqslant |z| \leqslant 1,\ 0 \leqslant arg(z) \leqslant \pi/2\}$ is a t-set.

   For every $r_1 = r_{01}(i) \subset \alpha$, suppose $r_1 = \{z : c_1 \leqslant |z| \leqslant d_1,\ u_1 \leqslant arg(z) \leqslant v_1\}$. Add the arc $|z| = c_1$, $u_1 \leqslant arg(z) \leqslant v_1$ and the segments $c_1 \leqslant |z| \leqslant 1$, $arg(z) = u_1, v_1$ into P. These lines together with $|z| = 1$, $u_1 \leqslant arg(z) \leqslant v_1$ enclose a domain called $h(r_1)$. Add into P all boundary lines those $r_{\nu n}(i) \subset h(r_1)$ for which $n \leqslant N + 1$.

   Next we consider those $r_2 = r_{\nu 2}(i) \subset \alpha \cap \{z : 0 \leqslant arg(z) \leqslant \pi/2\}$ which aren't contained in any $h(r_1)$ appeared above. We do a construction similar to r_1 and add into P the corresponding arcs, segments and boundary lines of those $r_{\nu n}(i) \subset h(r_2)$ for which $n \leqslant N + 2$

   Next we consider those $r_3 = r_{\nu 3}(i) \subset \alpha \cap \{z : 0 \leqslant arg(z) \leqslant \pi/2\}$ which aren't contained in any $h(r_1)$ and $h(r_2)$ appeared above and do the same things as r_1 and r_2. This process ends when no $r_{\nu n}(i)$ with the required properties remains.

   The process will surely come to an end because $\alpha \subset a(2\epsilon)$ and $|A(z)| = 1$ on $|z| = 1$

   (b) If r_{01} doesn't intersects β

   Proceed r_{20} and r_{21} separately according to the same rules in (a) and (b). If r_{20} intersects β, then say $\{z :\ 3/4 \leqslant |z| \leqslant 1,\ 0 \leqslant arg(z) \leqslant \pi/4\}$ is a t-set. Similarly, if r_{21} intersects β, then say $\{z :\ 3/4 \leqslant |z| \leqslant 1,\ \pi/4 \leqslant arg(z) \leqslant \pi/2\}$ is a t-set.

   This process is continued until n large enough such that $r_{\nu n} \subset \beta$ for all ν.

3. r_{11}, r_{21}, r_{31} are proceeded in the same way as r_{01}. Call $r_{01} \cup r_{11} = \{z :\ 1/2 \leqslant |z| \leqslant 1,\ 0 \leqslant arg(z) \leqslant \pi\}$ and $r_{21} \cup r_{31} = \{z :\ 1/2 \leqslant |z| \leqslant 1,\ \pi \leqslant arg(z) \leqslant 2\pi\}$ the s-sets of the zeroth generation. Call all t-sets appeared above belongs to the zeroth generation.

4. In the construction 1. 2. 3. described above, we have obtained a number of disjoint sets $h(r_l)$. If the range of arguments of numbers in r_l is $[2\pi t 2^{-k}, 2\pi(t + 1)2^{-k}]$, we have included in P all boundary lines of $r_{\nu n}(i)$ inside $h(r_l)$ except those in sets $s \subset h(r_l)$:

$$s = \left\{ z :\ 1 - 2^{-k} \leqslant |z| \leqslant 1,\ \frac{2\pi t}{2^k} \leqslant arg(z) \leqslant \frac{2\pi(t + 1)}{2^k} \right\} \tag{6.5}$$

   For different s's, the ranges of $arg(z)$ have no interior points in common. We say that these s-sets belong to the first generation. We now proceeds 2 for $r_{2t,k}$ and $r_{2t+1,k} \subset s$ and add lines into P according to the roles given. This construction gives new t-sets, of the first generation, and s-sets, of the second generation. The process ends when no new s-sets arises (i.e. until s-sets don't intersect α).

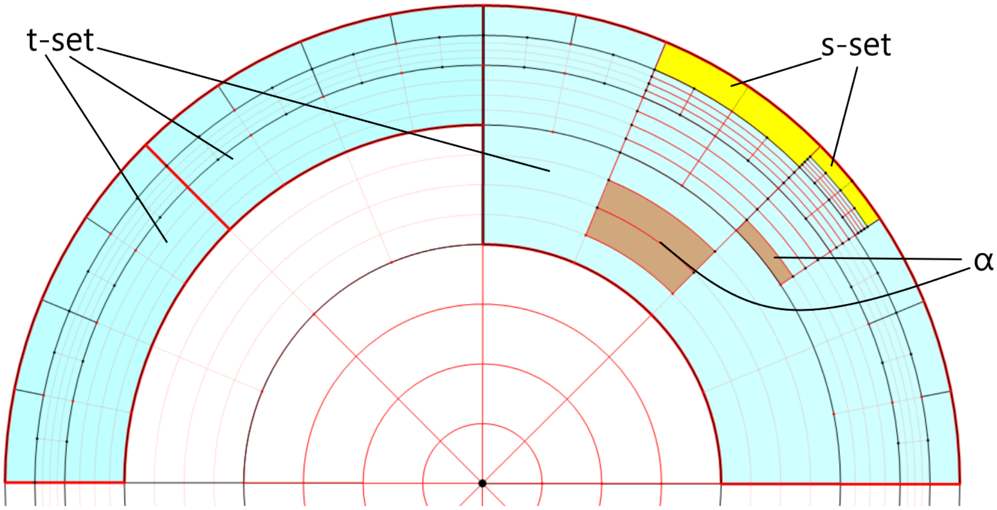


Figure 1: s-set and t-set

It is easy to see that P separates α and β in the sense: if γ is a continuous curve joining α and β, then γ has to intersect P.

P divides the unit disk into finitely many regions, and each region couldn't intersect α and β simultaneously. If a region intersects α, call it α-region, otherwise call it β-region. Then β is contained in the union of all closure of β-regions. Let Ω be the closure the union of all α-regions. and $\Omega_1, ..., \Omega_p$ are all connected components of Ω. Let $\Gamma = \partial\Omega \subset P$, then

$$\alpha \subset \Omega \Rightarrow a_1, ..., a_s \in \Omega$$

$$\alpha \cap \partial\Omega = \emptyset,\ \beta \cap \partial\Omega = \emptyset\ \Rightarrow \forall z \in \Gamma,\ \epsilon \leqslant |A(z)| \leqslant \epsilon^\kappa$$

Now we have seen that Γ satisfies 1,2 in Proposition 4.2. To prove Γ satisfies (6.1) in Proposition 4.2, notice that $\Gamma \subset P$, so it's enough to prove

$$\mu'(\{re^{i\theta} \mid 1-l \leqslant r \leqslant 1,\ \theta_0 \leqslant \theta \leqslant \theta_0 + l\}) \leqslant A_{15}\epsilon^{-2}l \qquad \forall l \in (0,1],\ \theta_0 \in \mathbb{R} \tag{6.6}$$

where $\mu'(E) =$ arc length of $E \cap P$ for $E \subset \{z : \ |z| < 1\}$

We'll prove (6.6) in section 8.

7 Harmonic Measure

Definition 7.1 (Harmonic measure)**.** *[5] Let D be an connected open subset of $\mathbb{C} \cup \{\infty\}$ whose boundary is disjoint union of some simple rectifiable curves.*

Any continuous function $f : \partial D \to \mathbb{R}$ determines a unique continuous function $H_f : \overline{D} \to \mathbb{R}$ such that H_f harmonic in D and $H_f|_{\partial D} = f$.

For each $x \in D$, there is a unique probability measure $\omega(x, D)$ on ∂D such that for any continuous function $f : \partial D \to \mathbb{R}$, we have

$$H_f(x) = \int_{\partial D} f(y) \mathrm{d}\omega(x, D)(y)$$

The measure $\omega(x, D)$ is called the harmonic measure with respect to D and $x \in D$.

Lemma 7.1 (Maximal principle). *$\omega(x, y)$ be the harmonic measure with respect to D, then any harmonic function get its maximal value in ∂D.*

Lemma 7.2 (Hall's Lemma). *$H = \{z : Re(z) > 0\}$, E is a closed subset of H, $D = H - E$, $E^* = \{i|z| : z \in E\}$.*

Denote $\omega(z) = \omega(z, H \backslash E)(\partial E)$, $z \in D$. Then $\omega(z)$ is the bounded harmonic function on D for which $\omega(iy) = 0 (-\infty < y < +\infty)$ and $\omega(z) = 1$ for $z \in \partial E$. Let $\omega^*(z) = \omega(z, H)(E^*)$, then

$$\omega^*(z) = \frac{1}{\pi} \int_{E^*} \frac{x dt}{x^2 + (y - t)^2} \tag{7.1}$$

For $x + iy \in D$,

$$\omega(x + iy) \geqslant \frac{2}{3} \omega^*(x - i|y|) \tag{7.2}$$

Proof. Suppose first that E consists of finite number of radical segments

$$\left\{ re^{i\theta_k} : a_k < r < b_k \right\}, \qquad k = 1, 2 \ldots, n; \quad |\theta_k| < \frac{\pi}{2}$$

with the intervals (a_k, b_k) disjoint. Let

$$G(z, \zeta) = \ln |\frac{z + \bar{\zeta}}{z - \zeta}|$$

denote the Green function of H, and consider the funtion

$$U(z) = \frac{1}{2\pi} \int_E \frac{1}{\xi} G(z, \zeta) ds, \qquad \zeta = \xi + i\eta,$$

where ds is the element of arclength on E. We claim that

$$\omega(x) \leq U(x), \qquad x > 0 \tag{7.3}$$

and

$$U(z) < \frac{3}{2}, \qquad Re(z) > 0. \tag{7.4}$$

For $|\zeta| = \rho, Re(\zeta) \geqslant 0$, the function $\frac{1}{\xi} G(x, \zeta)$ attains its minimum for $\xi = 0$; hence

$$\frac{1}{\xi} G(x, \zeta) \geqslant \frac{2x}{x^2 + \rho^2}, \qquad |\zeta| = \rho$$

which gives

$$U(x) \geqslant \frac{1}{\pi} \int_{E^*} \frac{x}{x^2 + \rho^2} d\rho = \omega^*(x)$$

which prove (7.3). Fix $z = x + iy$, and let $M(\rho)$ be the maximum of $\frac{1}{\zeta}G(z, \zeta)$ over the part of the circle $|\zeta - z| = \rho$ where $Re(\zeta) \geqslant 0$. Since pn this circle

$$\frac{1}{\zeta}G(z, \zeta) = \frac{1}{2\zeta} \ln(1 + \frac{4x\xi}{\rho^2})$$

is decreasing function of ξ,

$$M(\rho) = \begin{cases} \frac{1}{x-\rho} \ln(\frac{2x}{\rho} - 1), & x > \rho \\ \frac{2x}{\rho^2} & x \leq \rho \end{cases}$$

Now let $\phi(\rho)$ denote the total length of the part of E which lies in the disk $|\zeta - z| < \rho$. Since $\phi(\rho) \leq 2\rho$ and $M(\rho)$ is a decreasing function, we have

$$\begin{aligned} U(z) &\leq \frac{1}{2\pi} \int_0^\infty M(\rho) d\phi(\rho) = -\frac{1}{2\pi} \int_0^\infty \phi(\rho) dM(\rho) \\ &\leq -\frac{1}{\pi} \int_0^\infty \rho dM(\rho) = \frac{1}{\pi} \int_0^\infty M(\rho) d\rho = \frac{\pi}{2} + \frac{2}{\pi} < \frac{3}{2} \end{aligned}$$

Thus (7.4) is proved. Then we have $\frac{3}{2}\omega(z) - U(z) \geqslant 0$ on ∂D, so by the maximum principle, the same is true in D. Thus the function

$$\varphi(z) = \frac{3}{2}\omega(z) - \omega^*(z)$$

is non-negative on positive real axis; while $\varphi(iy) = 0 for y \in \mathbb{R}$, and $\varphi(z) \geqslant \frac{1}{2}$ for $z \in E$. By maximum principle, then ,

$$\omega(x + iy) \geqslant \frac{2}{3}\omega^*(x + iy), \qquad x > 0, y < 0.$$

By symmetry,

$$\omega(x + iy) \geqslant \frac{2}{3}\omega^*(x - iy), \qquad x > 0, y > 0.$$

For general compact set E, choose $\epsilon > 0$ and consider

$$S_\epsilon = \{z : \omega(z) > 1 - \epsilon\}$$

Clearly, $\partial E \subset S_\epsilon$. Choose a set $\widetilde{E}$ which consists of a finite number of radical segments with nonoverlapping projections, for which $\widetilde{E}^* = E^*$, let $\widetilde{\omega}(z)$ be the harmonic measure of $\widetilde{E}$. By what just proved,

$$\widetilde{\omega}(x + iy) \geqslant \frac{2}{3}\omega^*(x - i|y|)$$

since $\widetilde{E}^* = E^*$. The function $\omega(z) - \widetilde{\omega}(z)$ vanishes on the imaginary axis, is $\geqslant 0$ on ∂E, and is $\geqslant -\epsilon$ whenever it is defined on $\widetilde{E}$. Thus by the maximum principle,

$$\omega(x + iy) + \epsilon \geqslant \widetilde{\omega}(x + iy) \geqslant \frac{2}{3}\omega^*(x - i|y|)$$

for $(x + iy) \in D$. Now let $\epsilon \to 0$, and the lemma is proved for campact sets E.

Finally, suppose E is closed but unbounded. Let E_r be the intersection of E with disk $|z| \leq r$, let E_r^* be its circular projection, and let $\omega_r(z)$ and $\omega_r^*(z)$ denote the respective harmonic measures. Then

$$\omega(x+iy) \geqslant \omega_r(x+iy) \geqslant \frac{2}{3}\omega_r^*(x-i|y|) \tag{7.5}$$

for each point $x+iy \in D$, and we have the $\omega_r^*(z) \to \omega^*(z)$ pointwise as $r \to \infty$. This completes the proof. □

Corollary 7.1. *$H = \{z : Im(z) > 0\}$, E is a closed subset of H, $D = H - E$, $E^* = \{|z| : z \in E\}$.*

Denote $\omega(z) = \omega(z, H\backslash E)(\partial E)$ is the bounded harmonic function in D for which $\omega(y) = 0(-\infty < y < +\infty)$ and $\omega(z) = 1$ for $z \in \partial E$ ($\omega(z)$ is also called harmonic measure of E respect to $H - E$). Finally, let

$$\omega^*(z) = \frac{1}{\pi}\int_{E^*} \frac{ydt}{y^2+(x-t)^2} \tag{7.6}$$

be the harmonic measure of E^* with respect to H.

For $(x+iy) \in D$,

$$\omega(x+iy) \geqslant \frac{2}{3}\omega^*(-|x|+iy) \tag{7.7}$$

Let R be the annulus $\rho < |z| < 1$, and let E_1 be a closed subset of R which does not divide the plane. Let $\omega_1(z)$ be the harmonic measure of E_1 with respect to $R - E_1$ and let

$$E_1^* = \left\{e^{i\theta} : re^{i\theta} \in E_1\right\} \tag{7.8}$$

be the radial projection of E_1 onto the outer boundary of R. For fixed $\beta < \pi/|\ln\rho|$, let F_1^* be the part of E_1^* such that $|\theta| \leq \beta|\ln\rho|$.

Lemma 7.3. *If $\rho^{\frac{1}{3}} \notin E_1$,*

$$|F_1^*| \leq \sqrt{3}|\ln\rho|(e^{\pi\beta} + e^{-\pi\beta} + 1)\omega_1(\rho^{\frac{1}{3}}) \tag{7.9}$$

Proof. The mapping ψ

$$\zeta = \xi + i\eta = z^{i\pi/\ln\rho} = r^{i\pi/\ln\rho}e^{-\pi\theta/\ln\rho} \tag{7.10}$$

maps R to H, and denote φ be the inverse of ψ. Define set $E = \{\zeta \in H : \varphi(\zeta) \in E_1\}, E^* = \{\xi \in H : \varphi(\xi) \in E_1^*\}, \omega(\zeta)$ be the harmonic measure of E respect to $H - E$, ω^* be the harmonic measure of E^* respect to H. Since

$$u(x) = \int_{\partial E} u(y)\omega(x,y)dy$$
$$u_1(w) = \int_{\partial E_1} u_1(z)\omega_1(w,z)dz$$

Let $z = \varphi(y), w = \varphi(x), u = u_1 \circ \varphi$, then

$$u(x) = \int_{\partial E} u(y)\omega_1(\varphi(x),\varphi(y))d\varphi(y)$$

so

$$\omega(x,y)dy = \omega_1(\varphi(x),\varphi(y))d\varphi(y)$$

then

$$\omega(x) = \int_{\partial E}\omega(x,y)dy = \int_{\partial E_1}\omega_1(\varphi(x),\varphi(y))d\varphi(y) = \omega_1(\varphi(x))$$

the same,

$$\omega^*(x) = \omega_1^*(\varphi(x))$$

when $|z| \leq \sqrt{\rho}$, $\zeta = \xi + i\eta = \psi(z)$, and $\xi < 0$, so by Corollary7.1

$$\omega_1(\rho^{\frac{1}{3}}) = \omega(e^{\frac{i\pi}{3}}) \geq \frac{2}{3}\omega^*(e^{\frac{2\pi i}{3}}) = \frac{2}{3}\omega_1^*(\rho^{\frac{2}{3}})$$

Let F^* denote the image of F_1^* under ψ.

$$\omega_1^*(\rho^{\frac{2}{3}}) = \omega^*(e^{i2\pi/3}) \geq \frac{1}{\pi}\int_{F^*}\frac{\frac{\sqrt{3}}{2}d\xi}{\frac{3}{4}+(-\frac{1}{2}-\xi)^2}$$

$$= \frac{\sqrt{3}}{2|\ln\rho|}\int_{F_1^*}\frac{e^{-\frac{\pi\theta}{\ln\rho}}}{e^{-2\frac{\pi\theta}{\ln\rho}}+e^{-\frac{\pi\theta}{\ln\rho}}+1}d\theta \geq \frac{\sqrt{3}}{2|\ln\rho|}\frac{|F_1^*|}{e^{\pi\beta}+e^{-\pi\beta}+1}$$

In the end,

$$|F_1^*| \leq \sqrt{3}|\ln\rho|(e^{\pi\beta}+e^{-\pi\beta}+1)\omega_1(\rho^{\frac{1}{3}})$$

□

8 Discussion of P

This section is to prove (6.6), which is rewritten below:

$$\mu'(\{re^{i\theta} \mid 1-l \leqslant r \leqslant 1,\ \theta_0 \leqslant \theta \leqslant \theta_0 + l\}) \leqslant A_{15}\epsilon^{-2}l \qquad \forall l \in (0,1],\ \theta_0 \in \mathbb{R} \tag{8.1}$$

where $\mu'(E) =$ arc length of $E \cap P$ for $E \subset \{z :\ |z| < 1\}$.

We'll simplify it.

First, it's enough to prove (8.1) for $l \leqslant 1/2$, because for $1/2 < l \leqslant 1$, from the first construction law of P,

$$\begin{aligned}
&\mu'(\{re^{i\theta} \mid 1-l \leqslant r < 1-\frac{1}{2},\ \theta_0 \leqslant \theta \leqslant \theta_0 + l\})\\
\leqslant\ &\mu'(\{re^{i\theta} \mid 1-l \leqslant r \leqslant 1-\frac{1}{2})\\
\leqslant\ &2^{N+1}(l-\frac{1}{2}) + \left[\frac{l-\frac{1}{2}}{2^{-N-1}}\right]\cdot 2\pi \leqslant 2^{N+1}(2\pi+1)(l-\frac{1}{2})\\
\leqslant\ &2\times 2(2\pi+1)\epsilon^{-1}(2\pi+1)(l-\frac{1}{2}) \qquad (by\ (6.2))\\
\leqslant\ &4(2\pi+1)^2\epsilon^{-2}(l-\frac{1}{2})
\end{aligned}$$

Denote

$$s_{\nu n} = \left\{ z : 0 \leqslant 1 - |z| \leqslant \frac{1}{2^n}, \frac{\nu \cdot 2\pi}{2^n} \leqslant \arg z < \frac{(\nu+1)\cdot 2\pi}{2^n} \right\}, \qquad n \in \mathbb{N}_+, \nu = 0, 1, ..., 2^n - 1$$

$$t_{\nu n} = \left\{ z : 0 \leqslant 1 - |z| \leqslant \frac{1}{2^n}, \frac{\nu \cdot 2\pi}{2^{n+1}} \leqslant \arg z < \frac{(\nu+1)\cdot 2\pi}{2^{n+1}} \right\}, \qquad n \in \mathbb{N}_+, \nu = 0, 1, ..., 2^{n+1} - 1$$

It's easy to see that any set of the form $\{re^{i\theta} \mid 1-l \leqslant r \leqslant 1,\, \theta_0 \leqslant \theta \leqslant \theta_0 + l\}$, $l \in (0, 1/2]$, $\theta_0 \in \mathbb{R}$ is contained in $s_{\nu n} \cup s_{\nu+1,n}$ for some ν and n such that $l \leqslant 1/2^n < 2l$.

If we have proved

$$\mu'(s_{\nu n}) \leqslant A_{18}\epsilon^{-2}\frac{1}{2^n}, \qquad \forall\, n, \nu \tag{8.2}$$

Then

$$\mu'(\{re^{i\theta} \mid 1-l \leqslant r \leqslant 1,\, \theta_0 \leqslant \theta \leqslant \theta_0 + l\}) \leqslant \mu'(s_{\nu n}) + \mu'(s_{\nu+1,n}) \leqslant 2A_{18}\epsilon^{-2}\frac{1}{2^n} \leqslant 4A_{18}\epsilon^{-2} l$$

which implies (8.1). So it's enough to prove (8.2).

Notice that all t-sets (we have defined t-sets in section 6 the second law of construction of P) are of the form $t_{\nu n}$. Denote $T = \{t_{\nu n} : t_{\nu n} \text{ is t-set }\}$. The equation below is the final simplification of (8.1):

$$\mu'(t_{\nu n}) \leqslant A_{19}\epsilon^{-2}\frac{1}{2^n}, \qquad \forall t_{\nu n} \in T \tag{8.3}$$

Lemma 8.1. *(8.3) ⇒ (8.2).*

Proof. Suppose (8.3) holds. For any $s_{\nu n}$, choose the smallest s-set S (maybe of zeroth generation) containing it.

Notice that a t-set either disjoint with $s_{\nu n}$, or containing $s_{\nu n}$, or contained in $s_{\nu n}$. The t-sets in S of the same generation as S are disjoint. So $s_{\nu n}$ either contains some t-sets in S of the same generation as S, or is contained in a t-set in S of the same generation as S.

For the former case, $P \cap s_{\nu n} \subset \partial s_{\nu n} \cup (\bigcup P \cap t_{\nu' n'})$ where $t_{\nu' n'}$ takes all t-sets in $s_{\nu n}$ of the same generation as S. These $t_{\nu' n'}$'s argument range (i.e. $[\frac{\nu'\pi}{2^{n'}}, \frac{(\nu'+1)\pi}{2^{n'}})$) are disjoint, therefore $\frac{2\pi}{2^n} \geqslant \sum \frac{\pi}{2^{n'}}$, hence (8.3) ⇒ (8.2).

For the latter case, $s_{\nu n}$ is contained in a t-set t_0. Let $h_1, h_2, ..., h_p$ be all $h(r_l)$ contained in t_0 in the law 2(a) of construction of P, and the s-sets at the bottom of these h_k are denoted $s_1, s_2, ..., s_p$.

Each s_k is of the form $s_{\nu' n'}$, either disjoint with $s_{\nu n}$, or containing $s_{\nu n}$, or contained in $s_{\nu n}$. s_k containing $s_{\nu n}$ is impossible because S is the smallest s-set containing $s_{\nu n}$. So s_k either disjoint with or contained in $s_{\nu n}$.

By construction of P,

$$P \cap t_0 = \partial t_0 \cup \left(\bigcup_{k=1}^{p} \left(P \cap (h_k \backslash s_k) \right) \right) \cup \left(\bigcup_{k=1}^{p} \left(P \cap s_k \right) \right)$$

$$\Rightarrow P \cap s_{\nu n} \subset \partial s_{\nu n} \cup \left(\bigcup_{s_k \subset s_{\nu n}} \left(P \cap (h_k \backslash s_k) \right) \right) \cup \left(\bigcup_{s_k \subset s_{\nu n}} \left(P \cap s_k \right) \right) \tag{8.4}$$

Denote l_k the length of argument range of h_k, then $\sum\limits_{s_k \subset s_{\nu n}} l_k \leqslant \frac{2\pi}{2^n}$. Now estimate the length of three parts on the right side of (8.4):

- Length of $\partial s_{\nu n} \leqslant \frac{4\pi+2}{2^n}$.
- By construction of P, it's easy to show that

$$\begin{aligned} \textit{length } P \cap (h_k \backslash s_k) &\leqslant 2(N+1)2^N l_k \leqslant 2^{2N+1} l_k \\ &\leqslant 8(2\pi+1)^2 \epsilon^{-2} l_k \quad (\textit{by } (6.2)) \end{aligned}$$

$$\begin{aligned} \Rightarrow \textit{length} \bigcup_{s_k \subset s_{\nu n}} \big(P \cap (h_k \backslash s_k)\big) &\leqslant \sum_{s_k \subset s_{\nu n}} \textit{length } P \cap (h_k \backslash s_k) \\ &\leqslant \sum_{s_k \subset s_{\nu n}} 8(2\pi+1)^2 \epsilon^{-2} l_k \leqslant 16\pi(2\pi+1)^2 \epsilon^{-2} \frac{1}{2^n} \end{aligned}$$

- Same as the former case, it's easy to show that (8.2) holds s-sets. So

$$\textit{length} \bigcup_{s_k \subset s_{\nu n}} \big(P \cap s_k\big) \leqslant \sum_{s_k \subset s_{\nu n}} \textit{length } (P \cap s_k) \leqslant \sum_{s_k \subset s_{\nu n}} \frac{A_{18}}{2\pi} \epsilon^{-2} l_k \leqslant A_{18} \epsilon^{-2} \frac{1}{2^n}$$

Add the three parts, we get (8.2). □

So it's enough to prove (8.3).

Proof of (8.3). Choose an arbitrary $t_{\nu n} \in T,\ n \in \mathbb{N}$. Recall

$$t_{\nu n} = \left\{ z : 0 \leqslant 1 - |z| \leqslant \frac{1}{2^n}, \frac{\nu \cdot 2\pi}{2^{n+1}} \leqslant \arg z < \frac{(\nu+1) \cdot 2\pi}{2^{n+1}} \right\}$$

$$r_{\nu n} = \left\{ z : \frac{1}{2^{n+1}} < 1 - |z| \leqslant \frac{1}{2^n}, \frac{\nu \cdot 2\pi}{2^{n+1}} \leqslant \arg z < \frac{(\nu+1) \cdot 2\pi}{2^{n+1}} \right\}$$

By definition of t-sets, $r_{\nu n} \cap \beta \neq \emptyset$.
Choose $z_0 \in r_{\nu n} \cap \beta$, let $\rho = |z_0|^3,\ z_0 = \rho^{1/3} e^{i\theta_0},\ E_1 = \alpha \cap r_{\nu n}$. Then $z_0 \notin E_1$.

$$1 - \frac{1}{2^n} \leqslant |z_0| \leqslant 1 - \frac{1}{2^{n+1}} \quad \Rightarrow \quad (1 - \frac{1}{2^n})^3 \leqslant \rho \leqslant (1 - \frac{1}{2^{n+1}})^3 < 1 - \frac{1}{2^n}$$

$$\Rightarrow E_1 \subset r_{\nu n} \subset R = \{z : \rho < |z| < 1\}$$

Let $E_1^* = \left\{ e^{i\theta} : re^{i\theta} \in E_1 \right\}$, F_1^* be part of E_1^* such that $|\theta - \theta_0| \leq \pi |\ln \rho|$
Notice $\forall\, re^{i\theta} \in E_1 \subset r_{\nu n}$,

$$|\theta - \theta_0| \leqslant \frac{\pi}{2^n} \leqslant (-3\pi) \ln(1 - \frac{1}{2^{n+1}}) = -\pi \ln(1 - \frac{1}{2^{n+1}})^3 \leqslant -\pi \ln \rho = \pi |\ln \rho|$$

which implies $E_1^* = F_1^*$.
By Lemma 7.3,

$$|E_1^*| = |F_1^*| \leqslant \sqrt{3} |\ln \rho| (e^{\pi^2} + e^{-\pi^2} + 1) \omega_1(\rho^{\frac{1}{3}} e^{i\theta_0}) \leqslant A_{10} 2^{-n} \omega_1(\rho^{\frac{1}{3}} e^{i\theta_0}) \tag{8.5}$$

here $\omega_1(z) = \omega(z, R \backslash E_1)(\partial E_1)$.

Denote $D=\{z:\ |z|<1\}$. $\omega(z,D\backslash E_1)(\partial E_1)$ is a harmonic function on $D\backslash E_1$ with boundary value 0 on ∂D and 1 on ∂E_1.By maximal principal, $0\leqslant\omega(z,D\backslash E_1)(\partial E_1)\leqslant 1$. Let

$$f_1:\partial(R\backslash E_1)\to\mathbb{R}:\ f_1(z)=\begin{cases}0 & ,z\in\partial D\\ 1 & ,z\in\partial E_1\\ \omega(z,D\backslash E_1)(\partial E_1) & ,|z|=\rho\end{cases}$$

Then $\omega(z,D\backslash E_1)(\partial E_1)|_{R\backslash E_1}$ is the harmonic function with boundary value f_1.

$$\begin{aligned}\Rightarrow\omega(z,D\backslash E_1)(\partial E_1)&=\int_{\partial(R\backslash E_1)}f_1(y)\mathrm{d}\omega(z,R\backslash E_1)(y)\\&=\int_{\partial E_1}\mathrm{d}\omega(z,R\backslash E_1)(y)+\int_{|z|=\rho}\omega(z,D\backslash E_1)(\partial E_1)\mathrm{d}\omega(z,R\backslash E_1)(y)\\&\geqslant\int_{\partial E_1}\mathrm{d}\omega(z,R\backslash E_1)(y)=\omega(z,R\backslash E_1)(\partial E_1)=\omega_1(z),\quad z\in R\backslash E_1\end{aligned}\tag{8.6}$$

By (6.4), $E_1\subset\alpha\subset a(2\epsilon)\subset D$. Let

$$f_2:\partial(D\backslash a(2\epsilon))\to\mathbb{R}:\ f_2(z)=\begin{cases}0 & ,z\in\partial D\\ \omega(z,D\backslash E_1)(\partial E_1) & ,z\in\partial a(2\epsilon)\end{cases}$$

Then $\omega(z,D\backslash E_1)(\partial E_1)|_{D\backslash a(2\epsilon)}$ is the harmonic function with boundary value f_2.

$$\begin{aligned}\Rightarrow\omega(z,D\backslash E_1)(\partial E_1)&=\int_{\partial(D\backslash a(2\epsilon))}f_2(y)\mathrm{d}\omega(z,D\backslash a(2\epsilon))(y)\\&=\int_{\partial a(2\epsilon)}\omega(z,D\backslash E_1)(\partial E_1)\mathrm{d}\omega(z,D\backslash a(2\epsilon))(y)\\&\leqslant\int_{\partial a(2\epsilon)}\mathrm{d}\omega(z,D\backslash a(2\epsilon))(y)=\omega(z,D\backslash a(2\epsilon))(\partial a(2\epsilon)),\quad z\in D\backslash a(2\epsilon)\end{aligned}\tag{8.7}$$

Notice that $(log|A(z)|)\Big|_{D\backslash a(2\epsilon)}$ is a harmonic function on $D\backslash a(2\epsilon)$ with boundary value 0 on ∂D and $log(2\epsilon)$ on $\partial a(2\epsilon)$.

$$\Rightarrow log|A(z)|=\int_{\partial a(2\epsilon)}log(2\epsilon)\mathrm{d}\omega(z,D\backslash a(2\epsilon))(y)=log(2\epsilon)\cdot\omega(z,D\backslash a(2\epsilon))(\partial a(2\epsilon)),\quad z\in D\backslash a(2\epsilon)$$

Let $z=z_0\in\beta\subset D\backslash a(2\epsilon)$, by $|A(z_0)|\geqslant\epsilon^\kappa$ and $0<\epsilon\leqslant\frac{1}{4}$,

$$\omega(z_0,D\backslash a(2\epsilon))(\partial a(2\epsilon))=\frac{log|A(z_0)|}{log(2\epsilon)}\leqslant\frac{log(\epsilon^\kappa)}{log(2\epsilon)}=\kappa\frac{|log(\epsilon)|}{|log(\epsilon)|-log2}\leqslant 2\kappa\tag{8.8}$$

By (8.6), (8.7) and (8.8),

$$\omega_1(z_0)\leqslant 2\kappa\tag{8.9}$$

By (8.5) and (8.9),

$$|E_1^*|\leqslant A_{10}2^{-n+1}\kappa\tag{8.10}$$

Suppose $t_{\nu n}$ belongs to the j-th generation. Let L_g be the total length of argument range of all t-sets contained in $t_{\nu n}$ and of generation g, $g = j, j+1, \ldots$. Suppose $l_1, \ldots, l_\sigma$ are lengths of argument range of each t-set contained in $t_{\nu n}$ and of generation $j+1$ respectively. Then by the construction of P,

$$L_{j+1} = \sum_{i=1}^{\sigma} l_i \leqslant |E_1^*| \tag{8.11}$$

By (8.10) and (8.11),

$$L_{j+1} \leqslant A_{10} 2\kappa \cdot 2^{-n} = \frac{A_{10} 2\kappa}{\pi} L_j$$

Recall that in Proposition 4.2, we required $0 < \kappa < A_{14}$ and said A_{14} would be determined later. Now it's time: let $0 < A_{14} < 1/8$ and $\frac{2A_{10}A_{14}}{\pi} \leqslant 1/2$. Then

$$L_{j+1} \leqslant \frac{1}{2} L_j \tag{8.12}$$

If we replace $t_{\nu n}$ by any t-set contained in $t_{\nu n}$ and of generation $j+1$, a correspondent (8.12) still holds. Add these correspondent (8.12)s, we get

$$L_{j+2} \leqslant \frac{1}{2} L_{j+1}$$

Similarly,

$$L_{g+1} \leqslant \frac{1}{2} L_g, \qquad g = j, j+1, \ldots$$

$$\Rightarrow \sum_{g=j}^{\infty} L_g \leqslant \sum_{g=0}^{\infty} \frac{1}{2^g} L_j = 2L_j \tag{8.13}$$

For a t-set t_0 with length of argument range l_0, denote $s_1, \ldots, s_p$ all s-sets of the next generation contained in t_0. Suppose lengths of argument range of $s_1, \ldots, s_p$ are $l_1, \ldots, l_p$ respectively. Denote $t_1, \ldots, t_q$ all t-sets of the next generation contained in t_0. Then

$$P \cap \left(t_0 \backslash (\bigcup_{i=1}^{q} t_i) \right) = \bigcup_{1 \leqslant i \leqslant p,\, h(r_k)\widetilde{*}s_i} \Big(\partial s_i \cup \Big(P \cap \big(h(r_k) \backslash s_i \big) \Big) \Big)$$

where $h(r_k)\widetilde{*}s_i$ means s_i lies at the bottom of $h(r_k)$.

Recall $T = \{ t_{\nu n} : t_{\nu n} \text{ is t-set } \}$.

For each $1 \leqslant i \leqslant p$ and $h(r_k)\widetilde{*}s_i$,

$$\mathit{length}\ \partial s_i \leqslant 4 l_i$$

$$\mathit{length}\ P \cap \big(h(r_k) \backslash s_i \big) \leqslant 2(N+1)2^N l_i$$

$$\Rightarrow \mathit{length}\ P \cap \left(t_0 \backslash (\bigcup_{i=1}^{q} t_i) \right) \leqslant \sum_{i=1}^{p} (4 + 2(N+1)2^N) l_i \leqslant 2^{2N+2} \sum_{i=1}^{p} l_i \leqslant 16(2\pi+1)^2 \epsilon^{-2} l_0$$

$$
\begin{aligned}
\Rightarrow \mu'(t_{\nu n}) = length\ P \cap t_{\nu n} &= \sum_{t_0 \subset t_{\nu n},\, t_0 \in T} length\ P \cap \left(t_0 \backslash (\bigcup_{i=1}^{q} t_i) \right) \\
&\leqslant \sum_{t_0 \subset t_{\nu n},\, t_0 \in T} 16(2\pi+1)^2 \epsilon^{-2} l_0 \\
&= 16(2\pi+1)^2 \epsilon^{-2} \sum_{g=j}^{\infty} L_g \leqslant 32(2\pi+1)^2 \epsilon^{-2} L_j \qquad (by\ (8.13)) \\
&= 32(2\pi+1)^2 \epsilon^{-2} \frac{\pi}{2^n} = A_{19} \epsilon^{-2} \frac{1}{2^n}
\end{aligned}
$$

□

9 Corona Problem

Under Theorem 4.1, we want to get a more powerful Theorem, and that's the Corona Problem.

Theorem 9.1. *Let $f_1(z), f_2(z), \cdots, f_n(z)$ be given functions in B such that*

$$|f_1(z)| + |f_2(z)| + \cdots + |f_n(z)| \geqslant \delta > 0 \tag{9.1}$$

for some δ. Then $I(f_1, f_2, \cdots, f_n) = B$. Furthermore, if $\|f_\nu\| \leq 1, \nu = 1, 2, \cdots, n$, and $\delta \leq \frac{1}{2}$, there exists $p_\nu(z) \in B$,and absolute value $A_{13} > 1 + A_{12}$ so that

$$\sum_{\nu=1}^{n} p_\nu f_\nu = 1 \qquad \|p_\nu\| \leq n! 2^{n A_{13}} \delta^{-A_{13} n} \tag{9.2}$$

Before prove the theorem, we need the following theorem.

Theorem 9.2. *Let $A(z)$ be the finite Blaschke product and assume that the set $z = \{z : |A(z)| < \delta\}, \delta < \frac{1}{2}$, has the (simply connected) components $D_1, D_2, \ldots, D_q$. Let $F_i(z)$ be holomorphic in D_i and assume $|F_i(z)| < 1$ there. Then the interpolation problem*

$$f(a_\nu) = F_i(a_\nu) \qquad a_\nu \in D_i, f \in B \tag{9.3}$$

has a solution f with $\|f\| < \delta^{-A_{12}}$.

Proof. Choose a ϵ so that $\epsilon^\kappa = \epsilon$ and construct Γ relatively $A(z)$ and ϵ, and define $F(z) = F_i(z), z \in D_i$. By Proposition 4.1 there exist f_0 satisfied (9.3), such that

$$
\begin{aligned}
\|f_0\| &= \sup_{\|G\|_1 = 1} |\sum_{\nu=1}^{s} \frac{G(a_\nu) F_i(a_\nu)}{A'(a_\nu)}| = \sup_{\|G\|_1 = 1} |\frac{1}{2\pi i} \int_\Gamma \frac{F(z)G(z)}{A(z)} dz| \\
&\leq \epsilon^{-1} \frac{1}{2\pi} \int_\Gamma |G(z)| dz \leq \delta_{-A_{12}}
\end{aligned}
$$

□

Lemma 9.1. *Let f(z) be an analytic in the open unit disk D and continuous in $\bar{D}$. Suppose $0 < |f(z)| \leq 1$ on $|z| = 1$ and let E is nonempty. Then there exists a sequence $\{B_n(z)\}$ of finite Blaschke products with simple zeros, such that $|B_n(z)| \to |f(z)|$ uniformly in each closed subset of $(\bar{D} - \bar{E})$, and $B_n(z) \to f(z)$ uniformly in each closed subset D.*

Proof. Let S be an arbitrary closed subset of $(\bar{D} - \bar{E})$. Because $f(z) \neq 0, |z| = 1$,therefore has at most a finite number of zeros in D. Then, since it's clear that a finite Blaschke product can be approximated by one with simple zeros, uniformly in $\bar{D}$, it's enough to suppose f does not vanish in $\bar{D}$. We can assume $f(0) > 0$, and we have,

$$f(z) = \exp\left\{\frac{1}{2\pi}\int_0^{2\pi}\frac{e^{it}+z}{e^{it}-z}\ln|f(e^{it})|dt\right\}$$

Now let $\omega_k = e^{2\pi ik/n}$, and let

$$f_n(z) = \exp\left\{\frac{1}{n}\sum_{k=1}^{n}\frac{\omega_k+z}{\omega_k-z}\ln|f(\omega_k)|\right\}$$

Then $f_n(z) \to f(z)$ uniformly in S. Let

$$\epsilon_k = -\frac{1}{n}\ln|f(\omega_k)|$$

so that

$$0 \leq \epsilon_k \leq -\frac{1}{n}\ln\mu = \delta_n \tag{9.4}$$

where μ is the minimum of $|f(z)|$ on $|z| = 1$. Choosing n so large that $\delta_n < \frac{1}{2}$, let

$$1 - \rho_k^2 = 2\epsilon_k \qquad a_k = \rho_k\omega_k$$

and define

$$B_n(z) = \prod_{k=1}^{n}\frac{\bar{a}_k}{|a_k|}\frac{a_k - z}{1 - \bar{a}_k z}$$

Note that $|a_k| = 1$ if $\epsilon_k = 0$, so that the corresponding factor in $B_n(z)$ is trivial. A calculation gives

$$\begin{aligned}2\ln|B_n(z)| &= \sum_{k=1}^{n}\ln\frac{|a_k-z|^2}{|1-\bar{a}_kz|^2} = \sum_{k=1}^{n}(\frac{|a_k-z|^2}{|1-\bar{a}_kz|^2}-1) + \sum_{k=1}^{n}O(\frac{|a_k-z|^2}{|1-\bar{a}_kz|^2}-1)^2\\
&= \sum_{k=1}^{n}(\frac{|a_k-z|^2-|1-\bar{a}_kz|^2}{|1-\bar{a}_kz|^2}) + \sum_{k=1}^{n}O(\frac{|a_k-z|^2-|1-\bar{a}_kz|^2}{|1-\bar{a}_kz|^2})^2\\
&= \sum_{k=1}^{n}(\frac{(|z|^2-1)(1-|a_k|^2)}{|1-\bar{a}_kz|^2}) + \sum_{k=1}^{n}O(\frac{(|z|^2-1)(1-|a_k|^2)}{|1-\bar{a}_kz|^2})^2\\
&= -2(1-|z|^2)\sum_{k=1}^{n}\epsilon_k|1-\bar{a}_kz|^{-2} + \sum_{k=1}^{n}O(2\epsilon_k)^2\\
&= -2(1-|z|^2)\sum_{k=1}^{n}\epsilon_k|1-\bar{a}_kz|^{-2} + nO(\delta_n)^2 = -2(1-|z|^2)\sum_{k=1}^{n}\epsilon_k|1-\bar{a}_kz|^{-2} + O(\delta_n)\end{aligned}$$

uniformly in $\bar{D}$.

$$\ln|f_n(z)| = \frac{1}{n}\sum_{k=1}^{n}Re(\frac{\omega_k+z}{\omega_k-z})\ln|f(\omega_k)| = -\sum_{k=1}^{n}\epsilon_k\frac{1-|z|^2}{|1-\bar{\omega}_kz|}$$

From this we can deduce,

$$\ln|B_n(z)| - \ln|f_n(z)| = -(1-|z|^2)\sum_{k=1}^{n}\epsilon_k(\frac{1}{|1-\rho_k\bar{\omega}_k z|^2} - \frac{1}{|1-\bar{\omega}_k z|^2}) + O(\delta_n)$$

$$= -(1-|z|^2)\sum_{k=1}^{n}\epsilon_k(\frac{|z|^2(1-\rho_k^2) - 2|z|(1-\rho_k)\cos(\theta - \frac{2k\pi}{n})}{|1-\rho_k\bar{\omega}_k z|^2|1-\bar{\omega}_k z|^2}) + O(\delta_n)$$

$$= -(1-|z|^2)\sum_{k=1}^{n}\delta_n(\frac{|z|^2(2\delta_n) - 2|z|(\sqrt{\delta_n})\cos(\theta - \frac{2k\pi}{n})}{|1-\rho_k\bar{\omega}_k z|^2|1-\bar{\omega}_k z|^2}) + O(\delta_n)$$

$$= O(\sqrt{\delta_n})$$

Hence $\ln|B_n(z)| \to \ln|f(z)|$, which implies $|B_n(z)| \to |f(z)|$, uniformly in S. Since $B_n(0) > 0$, it also follows that $B_n(z) \to f(z)$ uniformly in each disk D_r, by

$$\frac{B_n(z)}{f(z)} = \exp\left\{\frac{1}{2\pi}\int_0^{2\pi}\frac{e^{it}+z}{e^{it}-z}\ln|\frac{B_n(e^{it})}{f(e^{it})}|dt\right\} \to 1$$

□

Proof of Corona Problem. We first assume that (9.1) holds and prove (9.2) by induction on n. $n = 1$ is clear. Let us assume (9.2) holds for $n-1$, since Theorem 9.1 is invariant under conformal map, (9.2) satisfied for all simple connected domain.

Consider first the case when $f_n(z)$ is a finite Blaschke product $B(z)$ with simple zeros $b_1, b_2, \ldots, b_s$. The set $|B(z)| < \frac{\delta}{2}$ has the components $D_1, D_2, \ldots, D_q$. In each D_i there exist functions $P_{i\nu}$ such that

$$\sum_{\nu=1}^{n-1} P_{i\nu}(z) f_\nu(z) = 1 \qquad \|P_{i\nu}\| \leq (n-1)!(\frac{2}{\delta})^{(n-1)A_{13}}$$

by theorem 9.2, there exist functions $p_\nu \in B, \nu = 1, 2, \ldots, n-1$, such that

$$p_\nu(b_j) = P_{i\nu}(b_j), \quad b_j \in D_i; \qquad \|p_\nu\| \leq (n-1)!(\frac{2}{\delta})^{(n-1)A_{13}+A_{13}}$$

the function $p_n(z)$ defined by

$$p_n(z) = (1 - \sum_{1}^{n-1} p_\nu(z) f_\nu(z)) B(z)^{-1}$$

belongs to B, and now $\{p_\nu(z)\}_1^n$ satisfied (9.2) with exponent nA_{13}.

For the general case, we choose $\rho < 1$ and replace f_ν by $g_\nu(z) = f_\nu(\rho z)$. If we can prove (9.2) for an infinite sequence $\rho_\mu \to 1$, we have proved it generally. Since if

$$\sum p_{\rho_\mu \nu} g_{\rho_\mu \nu} = 1 \qquad \|p_{\rho_\mu \nu}\| \leq n!(\frac{2}{\delta})^{nA_{13}} \tag{9.5}$$

by Ascoli theorem we can take subseqeuce $\{\rho_\mu\}$ such that for every $\nu = 1, 2, \ldots, n$. $p_{\rho_\mu \nu}$ converges uniformly on every compact subset of D, thus converge to a holomorphic function p_ν with.

$$\sum_{\nu=1}^{n} p_\nu f_\nu = 1 \qquad \|p_\nu\| \leq n!(\frac{2}{\delta})^{nA_{13}}$$

We choose ρ so that $g_n(z) \neq 0$ on $|z| = 1$ and choose $G_n(z)$ analytic and $\neq 0$ in $|z| < 1$ such that

$$|G_n(e^{i\theta})| = \min(|g_n(e^{i\theta})|^{-1}, 2\delta^{-1}). \tag{9.6}$$

since G analytic and $\neq 0$, we conclude that $log\frac{1}{|G|}$ is harmonic function with bounded value $\max(log|g_n(e^{i\theta})|, \, log\frac{\delta}{2})$. Since harmonic function get its maximal value on the boundary, we deduce $|G| \geqslant \min(|g_n(e^{i\theta})|^{-1}, 2\delta^{-1}) \geqslant \min(\frac{2}{\delta}, 1) = 1$

We will prove that the functions $g_1, g_2, \ldots, G_n g_n$ satisfied (9.1) for $\frac{1}{2}$. Since if $|g_n(z)| \leq \frac{\delta}{2}$, Then $|g_1| + |g_2| + \cdots + |g_{n-1}| \geqslant \frac{\delta}{2}$. Otherwise, $|g_n| \geqslant \frac{\delta}{2}$, $|G_n g_n| \geqslant \frac{\delta}{2}$.

By lemma 9.1, there exists a sequence $\{B_k(z)\}$ of finite Blaschke products with simple zeros converging uniformly to $G_n(z)g_n(z)$ outside any neighborhood of the set on $|z| = 1$ where $|g_n(z)| \leq \frac{\delta}{2}$. We apply the above result to $g_1, \ldots, g_{n-1}, B_k$ and let $k \to \infty$ and observe that

$$\lim_{k\to\infty} [\inf_{|z|<1} (|g_1| + \cdots + |g_{n-1}| + |B_k|))] \geqslant \frac{\delta}{2} \tag{9.7}$$

This is because we can choose a sequence of neighborhood U_n converging to $H = \left\{|z| = 1 : |g_n(z)| \leq \frac{\delta}{2}\right\}$. To each U_n we choose B_n such that

$$|B_n(z) - G_n(z)g_n(z)| < \frac{1}{n^2}, \forall z \in U_n$$

We can choose N, such that $\forall n > N, |g_1(z)| + \cdots + |g_{n-1}(z)| > \frac{\delta}{2}, \forall z \in D - U_n$, and (9.7) satisfied $\forall z \in U_n$.

Thus we can obtain a selection of convergent subsequences, coefficients $p_1, \ldots, p_n \in B$ such that

$$p_1 g_1 + \cdots + (p_n G_n) g_n = 1$$

By above special case,

$$\|p_n G_n\| \leq n!(\frac{2}{\delta})^{nA_{13}}$$

Finally,if (9.2) not holds, there is a sequence z_j, so that $\lim_{j\to\infty} f_\nu(z_j) = 0$. This then imply 1 is not in I. □

References

[1] Complex Analysis, Elias M. Stein, Rami Shakarchi

[2] An Introduction to harmonic analysis, Yitzhak Katznelson

[3] An Interpolation problem for bounded analytic functions, Amer.J.Math. 80(1958), 921-930

[4] A Representation formula for the Dirichlet integral, Math.Z,73(1960),190-196.

[5] https://en.m.wikipedia.org/wiki/Harmonic_measure#cite_ref-1

[6] Theory of H^p Spaces, Peter L. Duren

[7] Interpolations by Bounded Analytic Functions and the Corona Problem, Lennart Carleson, 1962

Sumsets and Arithmetical Progression

Rui Rao, Di Wu

1 Introduction

We are interested in the structure of the sets and the sumsets. For example, if we consider two sets A and B in $\mathbb{R}$, both having n elements and these elements being 'general', their sumset

$$A + B := \{m \,|\, m = a + b, a \in A, b \in B\}$$

will consist of n^2 elements.

But interesting things happen if they do have some relations such that the sumset have only elements of order $O(n)$. This means some of the sums are the same, and we can guess easily that they have some artihmetical relations. For generality, we will work in a torsion-free abelian group. That will cover many of our applications such as integers, Euclidean spaces and so on. The relation and the result will be described in what follows.

Definition 1.1 (Arithmetical progression)**.** *Assume G is an abelian group, $a, q_1, q_2 \dots q_d \in G$, $l_1, l_2 \dots l_n \in \mathbb{N}^*$. Then we call*

$$P(q_1, q_2 \dots q_d; l_1, l_2 \dots l_d) = \{n : n = a + x_1q_1 + x_2q_2 \cdots + x_dq_d, 0 \leq x_i \leq l_i\}$$

a d-dimensional arithmetical progression. We call the number of its elements the size of the arithmetical progression.

Now come to the main theorem following Z. Ruzsa [1].

Theorem 1.2 (Z. Ruzsa)**.** *Assume G is a torsion-free abelian group. For any $\alpha > 1$,we can find $d = d(\alpha), C = C(\alpha)$ such that, for any $A, B \subseteq G$ satisfies $|A| = |B| = n$ and $|A+B| \leq \alpha n$, there is an arithmetical progression P,dimension at most d and size at most Cn,with $A \subseteq P$.*

Remark 1.3. *The $d(\alpha)$ and $C(\alpha)$ can be given explicit formulas.*

2 Finite abelian groups

In this and the following two sections, we will show the tools and lemmas for the proof of the main theorem. Most of these proofs are taken from [1].

First we introduce some notations about sumsets. We have defined $A+B$, similarly we define

$$A - B := \{n \,|\, n = a - b, a \in A, b \in B\},$$

and

$$2A := A + A.$$

By the same way we can define sumsets like $3A$ and $4A - 5B$, etc.

This section we deal with finite abelian groups to get some useful estimates. We will focus on characters first.

Definition 2.1 (Bohr set)**.** *Assume G is a finite abelian group and let $\gamma_1, \gamma_2 \dots \gamma_k$ be some characters of G, $0 < \varepsilon \leq \frac{1}{2}$, then the set*

$$B(\gamma_1, \gamma_2 \dots \gamma_k; \varepsilon) := \{g \in G : |arg\ \gamma_j(g)| \leq 2\pi\varepsilon\ for\ all\ j\}$$

will be called a Bohr (k, ε)-set.

Lemma 2.2 (Bogolyubov)**.** *Assume G is an abelian group of order m, $A \subset G$,and $|A| = n = \beta m$,then $D = 2A - 2A$ contains a Bohr $(k, \frac{1}{4})$-set, with $k \leq \beta^{-2}$.*

Proof. Step 1: Denote Γ the set of characters of G. For $\gamma \in \Gamma$, write $f(\gamma) = \Sigma_{a\in A}\gamma(a)$. We have the following orthogonality

$$\sum_{\gamma} \gamma(a)\overline{\gamma(b)} = \begin{cases} 0, a \neq b, \\ 1, a = b, \end{cases}$$

By direct calculation, we have

$$\sum_{\gamma} |f(\gamma)|^2 = mn,$$

and

$$\sum_{\gamma} \gamma(g) = \begin{cases} m, g = e, \\ 0, else. \end{cases}$$

Now we consider

$$h(x) = \sum_{\gamma} |f(\gamma)|^4 \gamma(x).$$

We claim that if $h(x) \neq 0$, then $x \in D$. If $x \notin D = 2A - 2A$, then we calculate

$$\begin{aligned} h(x) &= \sum_{\gamma} f(\gamma)^2 \overline{f(\gamma)^2} \gamma(x) \\ &= \sum_{\gamma \in \Gamma, a,b,c,d \in A} \gamma(a)\gamma(b)\gamma(-c)\gamma(-d)\gamma(x) \\ &= \sum_{a,b,c,d \in A} \sum_{\gamma} \gamma(x + a + b - c - d) \\ &= 0. \end{aligned}$$

And this proves our claim.

Step 2: Let γ_0 be the principal character (i.e $\gamma_0(g) = 1$ for all $g \in G$). We split $\Gamma - \{\gamma_0\}$ into two parts:

$$\Gamma_1 = \{\gamma : f(\gamma) \geq \sqrt{\beta} n\},$$
$$\Gamma_2 = \{\gamma : f(\gamma) < \sqrt{\beta} n\}.$$

Then we construct the Bohr $(k, \frac{1}{4})$-set B where these γ_i take all $\gamma \in \Gamma_1$ and $k = |\Gamma_1|$. We want to show $B \subseteq D$. By step 1, it suffices to show that $h(x) \neq 0$ when $x \in B$. If $x \in B$, $Re(\gamma(x)) > 0$ for all $\gamma \in \Gamma_1$, so we have the inequality

$$\begin{aligned} Re(h(x)) &> n^4 + Re\Big(\sum_{\gamma \in \Gamma_2} |f(\gamma)|^4 \gamma(x) \Big) \\ &\geq n^4 - |\beta n^2 \big(\sum_{\gamma} |f(\gamma)|^2 \big)| \\ &\geq n^4 - \beta n^3 m = 0 \end{aligned}$$

so we have $B \subset D$.

Step 3: In this part we prove that $k = |\Gamma_1| < \beta^{-2}$. By the definition of Γ_1, we have

$$km^2\beta^3 = k\beta n^2 \leq \sum_{\gamma \in \Gamma_1} |f(\gamma)|^2 < \sum_{\gamma \in \Gamma} |f(\gamma)|^2 = mn = m^2\beta,$$

so we get $k < \beta^{-2}$.

$\square$

Next we consider cyclic groups only. We prove Bohr set must contain arithmetical progressions. By Freiman homomorphism (introduced later), this is enough for our purposes. Let $G = \mathbb{Z}/m\mathbb{Z}$, and we use residues to express the element of G. Its characters can be expressed with an residue u:

$$\gamma_u(x) = e^{2\pi iux/m}.$$

Lemma 2.3. *Let m be a positive integer, $u_1, u_2, \cdots, u_k$ be residues with $(u_1, u_2, \cdots, u_k, m) = 1$, $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_k$ real numbers satisfying $0 < \varepsilon_i < \frac{1}{2}$, then there are residues $q_1, q_2, \cdots, q_k$, and $l_1, l_2, \cdots, l_k \in \mathbb{N}$ such that the set*

$$P = \{x_1q_1 + x_2q_2 + \cdots + x_kq_k \,|\, |x_i| \leq l_i\},$$

is contained in $B(u_1, u_2, \cdots, u_k; \varepsilon_1, \varepsilon_2, \cdots, \varepsilon_k)$. And the sums in P are all distinct with

$$|P| > \delta m,$$

here

$$\delta = \frac{\varepsilon_1 \cdots \varepsilon_k}{k^k}.$$

Proof. Let L be a k-dimensional lattice (seen in a $\mathbb{R}$-vector space) of $(x_1, x_2, \cdots, x_k)$ satisfying

$$x_1 \equiv xu_1, \cdots, x_k \equiv xu_k \mod m,$$

with some integer x. Since we have the coprimality condition, every space of m^k have exactly m points in L, hence its determinate is m^{k-1}.

Let Q be a rectangle defined by

$$Q = \{(x_1, x_2, \cdots, x_k) \,|\, |x_i| \leq \varepsilon_i, \forall i\}.$$

Let $\lambda_1, \cdots, \lambda_k$ denote the successive minima of Q with respect to L. A classical result of Minkowski show

$$\lambda_1\lambda_2 \cdots \lambda_k \leq 2^k \frac{\det L}{\text{vol} Q} = \frac{m^{k-1}}{\varepsilon_1 \cdots \varepsilon_k}.$$

By definition of λ_i, we can find $\{a_i\} \subseteq L$ linearly independent with $a_i \in Q\lambda_i$, and if we write

$$a_i = (a_{i1}, a_{i2}, \cdots, a_{ik}),$$

then $|a_{ij}| \le \lambda_i \varepsilon_j$ and $a_{ij} \equiv q_i u_j \mod m$ for some $q_i(< m)$. For

$$l_i = \left[\frac{m}{k\lambda_i} \right],$$

we set $P = \{x_1 q_1 + x_2 q_2 + \cdots + x_k q_k \,|\, |x_i| \le l_i\}$. Direct calculation shows

$$P \subseteq B = B(u_1, u_2, \cdots, u_k; \varepsilon_1, \varepsilon_2, \cdots, \varepsilon_k).$$

Next we show sums in P are all distinct. If $x_1, \cdots, x_k$ and $y_1, \cdots, y_k$ correspond the same value in P, then $z_i = x_i - y_i$ satisfy

$$\sum_i z_i q_i = 0, \ |z_i| \le 2l_i,$$

Multiplying u_j we get

$$\sum_i z_i a_{ij} \equiv 0 \mod m,$$

But

$$|\sum_i z_i a_{ij}| \le \sum_i 2 l_i \lambda_i \varepsilon_i \le \sum_i \frac{2m\varepsilon_i}{k} < m,$$

then$\sum_i z_i a_{ij} = 0$, so $\sum_i z_i a_i = 0$. But $\{a_i\}$ is linearly independent, $z_i = 0$, which is what we want. Then immediately we get

$$|P| > \frac{m^k}{k^k \lambda_1 \lambda_2 \cdots \lambda_k} \ge \delta m.$$

□

Combining lemma 2.2 and lemma 2.3, we get an important result.

Lemma 2.4. *Let m be a prime number, and A be a nonempty subset of $\mathbb{Z}/m\mathbb{Z}$ with $|A| = \beta m$. There are residues $q_1, q_2, \cdots, q_k$, and $l_1, l_2, \cdots, l_k \in \mathbb{N}$ such that the set*

$$P = \{x_1 q_1 + x_2 q_2 + \cdots + x_k q_k \,|\, |x_i| \le l_i\},$$

is contained in $2A - 2A$. And the sums in P are all distinct with

$$|P| \ge \delta m,$$

here

$$k \le \beta^{-2},$$
$$\delta = (4k)^{-k}.$$

3 Freiman homomorphism

In this section we introduce homomorphism in the sense of Freiman, and study its properties. This concept can be defined in general abelian groups.

Definition 3.1. *Let G_1, G_2 be abelian groups, $A_1 \subseteq G_1, A_2 \subseteq G_2$ and $r > 1$ a fixed positive integer. If a mapping $\phi : A_1 \to A_2$ satisfies that for any $x_1, \cdots, x_r, y_1, \cdots, y_r \in A_1$(not necessarily distinct), the equation*

$$x_1 + \cdots + x_r = y_1 + \cdots + y_r,$$

will imply

$$\phi(x_1) + \cdots + \phi(x_r) = \phi(y_1) + \cdots + \phi(y_r),$$

we call ϕ a Freiman homomorphism of order r, or F_r-homomorphism. We call it a isomorphism if it is one-to-one. When we do not specify r, we mean $r = 2$.

Lemma 3.2. *Let G, G' be abelian groups. If $P' \subseteq G'$ is a Freiman image of a arithmetical progression $P(q_1, \cdots, q_d; l_1, \cdots, l_d; a)$, then there are q_i', a' such that*

$$P' = P(q_1', \cdots, q_d'; l_1, \cdots, l_d; a'),$$

and the homomorphism is given by

$$\phi(a + x_1q_1 + \cdots + x_dq_d) = a' + x_1q_1' + \cdots + x_dq_d'.$$

Proof. we define

$$a' = \phi(a), \quad q_i' = \phi(a + q_i) - \phi(a),$$

and we can prove by induction that

$$\phi(a + x_1q_1 + \cdots + x_dq_d) = a' + x_1q_1' + \cdots + x_dq_d'.$$

We use induction on

$$r = x_1 + x_2 \cdots + x_n,$$

and by the definition of q_i' the statement is correct when $r \leq 1$.

Now assume that $r \geq 2$ and the statement holds for all smaller r, and let's just further assume $x_1 \geq 1$, then by the definition, we have

$$\phi(a+x_1q_1+x_2q_2\cdots+x_dq_d)+\phi(a) = \phi(a+(x_1-1)q_1+x_2q_2\cdots+x_dq_d))+\phi(a+x_1),$$

$$\begin{aligned}\phi(a+x_1q_1+x_2q_2\cdots+x_dq_d) &= a'+(x_1-1)q_1'+x_2q2'\cdots+x_dq_d'+a'+x_1'-a'\\ &= a'+x_1q_1'+x_2q_2'\cdots+x_dq_d'.\end{aligned}$$

Then by induction we complete the proof. □

Lemma 3.3. *Let G, G' be abelian groups. $A \subseteq G, A' \subseteq G'$ are F_r-isomorphic sets. Write $r = r'(k+l)$ with positive integers r', k, l. Then $kA - lA$ and $kA' - lA'$ is $F_{r'}$-isomorphic.*

Proof. For

$$x = a_1 + \cdots + a_k - b_1 - \cdots - b_l \in kA - lA,$$

we define

$$\psi(x) = \phi(a_1) + \cdots + \phi(a_k) - \phi(b_1) - \cdots - \phi(b_l),$$

where ϕ is the F_r isomorphism between G, G'. We can check easily that it is well defined and is a $F_{r'}$-isomorphism. □

4 Estimate of sumsets

In [1] there are two lemmas the author didn't give proof (lemma 5.1 and lemma 5.2). These two lemmas is useful in the final theorem.

We need some preparations for the proof of these lemmas. In what follows we are always dealing with finite sets in a fixed torsion-free abelian group.

Lemma 4.1. *Set $f_B(X) = \dfrac{|X+B|}{|X|}$. If X satisfies $f_B(X) \le f_B(Z)$ for every $Z \subset X$, then $f_B(X) \ge f_B(X+C)$.*

Proof. Assume $k = \dfrac{|X+B|}{|X|}$, then $\dfrac{|Z+B|}{|Z|} \ge k$ for all $Z \subset X$, and we will use induction on |C| to prove this lemma.

If $|C| = 0$, it's trivial.

If for any C with $|C| < s$, this lemma is correct. For $|C| = s$, we take $c_0 \in C$ and let $C = c_0 \cup C'$, where $|C'| = s-1$, let $T = (X+c_0) \cap (X+C'-c_0)$

$$\begin{aligned}|X+B+(C'\cup c_0)| &= |X+B+C'| + |(X+B+c_0)\setminus(X+B+C')|\\ &\le |X+B+C'| + |X+B| - |T+B|\\ &\le k|X+C'| + k|X| - k|T| = k(|X+C'|+|X|-|T|)\end{aligned}$$

because

$$|X| = |X + c_0| \geq |T|.$$

If we divide both sides $|X + C|$, then we finish the proof. □

Lemma 4.2.

$$|A||B - C| \leq |A - B||A - C|.$$

Proof. Consider the following map

$$\begin{aligned} \Phi : \qquad & A \times (B - C) \longrightarrow (A - B) \times (A - C), \\ & a \times x \longrightarrow (a - b_0) \times (a - c_0), \end{aligned}$$

here we takes $b_0 = b(x)$ as fixed for each x such that there exists $c \in C$ satisfying $b_0 - c = x$ and c_0 is $b_0 - x$.

The following chain show Φ is reversible (hence injective):

$$\begin{aligned} (a - b_0) \times (a - c_0) \to & (a - b_0, b_0 - c_0) \\ & \to (a - b_0, b_0 - c_0, b_0) \to (a, b_0, c_0). \end{aligned}$$

□

Lemma 4.3 (lemma 5.2 of [1], see also [2])**.** *If $|B| = n, |B + A| = \alpha n$, then for any positive integer k, l, we have*

$$|kA - lA| \leq \alpha^{k+l} n$$

Proof. Take $X \subset B$ such that $\dfrac{|X + A|}{|X|}$ takes the minimum.

By lemma 4.1 we have $\dfrac{|X + A|}{|X|} \leq \alpha$ and

$$|X + mA||X| \leq |X + A||X + (m-1)A| \leq \alpha|X||X + (m-1)A| \cdots \leq \alpha^m|X|.$$

So using lemma 4.2 we get

$$|kA - lA| \leq \frac{|X + kA||X + lA|}{|X|} \leq \alpha^{k+l}|X| \leq \alpha^{k+l} n.$$

□

Lemma 4.4 (lemma 5.1 of [1], see also [2]). *If $A \subseteq \mathbb{Z}, r \geq 2, N \in \mathbb{Z}$ s.t $|rA - rA| = N$, then for any $m > 2r(N-1)$ there is a subset $A' \subseteq A$ satisfy the following statement:*

i) $|A'| \geq \dfrac{|A|}{r}$,

ii) A' is F_r-isomorphic to a subset of $\mathbb{Z}/m\mathbb{Z}$.

Proof. Choose a large prime number q s.t. $q > \max\{rA - rA\}$, now consider the map

$$\mathbb{Z} \overset{mod\ q}{\longrightarrow} \mathbb{Z}/q\mathbb{Z} \overset{\lambda_1 x+\lambda_2}{\longrightarrow} \mathbb{Z}/q\mathbb{Z} \longrightarrow \{1, 2, \ldots, q\} \quad \overset{\text{mod}}{\longrightarrow}{}^{N} \mathbb{Z}/N\mathbb{Z},$$

which bring A's subset A' to the subset B' of $\mathbb{Z}/N\mathbb{Z}$.

Next we choose λ_1, λ_2 and A' to make this map a F_r-isomorphism and $|A'| \geq \dfrac{|A|}{r}$. It's obvious that the first two map is F_r-isomorphism, so we need to ensure that the map

$$\mathbb{Z}/q\mathbb{Z} \to \{1, 2, \ldots, q\}$$

and

$$\{1, 2, \ldots, q\} \to \mathbb{Z}/N\mathbb{Z}$$

are F_r-isomorphism.

First, for a certain λ_1 we choose λ_2 that the map

$$\mathbb{Z}/q\mathbb{Z} \to \{1, 2 \ldots q\}$$

is a F_r-isomorphic on A':

Since there is always a integer k s.t the set $\{k + 1 + q\mathbb{Z}, k + 2 + q\mathbb{Z} \ldots k + [\dfrac{q}{r}] + q\mathbb{Z}\}$ has $\dfrac{|A|}{r}$ elements in $|A''|$, here $A'' = \{\lambda_1 x + q\mathbb{Z}, x \in A\}$, under this condition we can choose $\lambda_2 = -k$ and $|A'| = \{x | k < \lambda x \leq k + [\dfrac{q}{r}], x \in A\}$, then map

$$\mathbb{Z}/q\mathbb{Z} \to \{1, 2, \ldots, q\}$$

is a F_r-isomorphism on A'.

Then we choose λ_1 to make

$$\phi : \mathbb{Z} \to \mathbb{Z}/q\mathbb{Z} \overset{\lambda_1 x+\lambda_2}{\longrightarrow} \mathbb{Z}/q\mathbb{Z} \longrightarrow \{1, 2 \ldots q\} \overset{modN}{\to} \mathbb{Z}/N\mathbb{Z}$$

a F_r-isomorphism, if

$$a_1 + a_2 + \cdots + a_r \neq b_1 + b_2 + \cdots + b_r,$$

but

$$\phi(a_1) + \phi(a_2) + \cdots + \phi(a_r) = \phi(b_1) + \phi(b_2) + \cdots + \phi(b_r),$$

take

$$d = a_1 + a_2 + \cdots + a_r - b_1 - b_2 - \cdots - b_r > 0,$$

then by the definition of ϕ, we know $N|t$ (where $t \equiv \lambda d \mod q$ and $0 \leq t < q$), because q is a prime, for each d there are only $[\frac{q}{N}]$ bad λ_1. And there are at most N different d, so there is always λ_1 and λ_2 satisfy ϕ is a F_r-isomorphism on A', where $|A'| \leq \frac{|A|}{r}$.

□

5 Proof of the main theorem

We finish the final proof.

Proof of the main theorem. Since A is finite, we can assume that $A \subseteq \mathbb{Z}^v$ for some v.

For any r, we can find a set $A_2 \subseteq \mathbb{Z}$ that is F_r-isomorphic to A. In fact, the map

$$(a_1, \cdots, a_v) \rightarrow a_1 + ta_2 + \cdots + t^{v-1}a_v,$$

will work for sufficiently large t. We will use the case $r = 8$.

Now we use lemma 4.3 for $2A_2 - 2A_2$, getting

$$|2A_2 - 2A_2| = |2A - 2A| \leq \alpha^4 n.$$

And we apply lemma 4.4 with $r = 8$ and any prime $m > 2r|2A - 2A|$. So we can always choose m prime with

$$m < 4r|2A - 2A| < 32\alpha^4 n,$$

and get $A' \subseteq A_2$ that is F_8-isomorphic to a $T \subseteq \mathbb{Z}/m\mathbb{Z}$ with $|A'| \geq n/8$. Then we can use lemma 2.4 to get a k dimensional arithmetical progression $P \subseteq 2T - 2T$ with $|P| > \delta n$, and δ and k depend on α only.

Now we come back to A. T is F_8 isomorphic to $A^* \subseteq A$, which can be extended to a F_2-isomorphism between $2T - 2T$ and $2A^* - 2A^*$ (lemma 3.3).

The image of P, denoted by P^*, is still a arithmetical progression in $2A-2A$ (lemma 3.2).

Select maximum subset $a_1, \cdots, a_s \in A$ such that $P^* + a_i$ are pairwise disjoint. This is always possible, and since any of them belong to $A + P^* \subseteq 3A - 2A$, we have (by lemma 4.3 again)

$$s \leq \frac{|3A-2A|}{|P^*|} \leq \frac{\alpha^5 n}{\delta n} = \frac{\alpha^n}{\delta}.$$

Now A can be covered by

$$A \subseteq \bigcup_i (a_i + P^* - P^*) \subseteq \{a_1, \cdots, a_s\} + P^* - P^*.$$

since for any $a \in A$,$\exists a_i$ such that $a + P^* = a_i + P^*$. Easy to see the set $\{a_1, \cdots, a_s\}$ is covered by

$$P(a_1, \cdots, a_s; 1, \cdots, 1; 0),$$

and $P^* - P^*$ is still a k dimensional arithmetical progression with

$$|P^* - P^*| \leq 2^k |P^*| \leq 2^k |2A - 2A| \leq 2^k \alpha^4 n.$$

Combining these two arithmetical progressions, we finally get that A is contained in an arithmetical progression of $s + k$ dimension with its size bounded by $2^{s+k}\alpha^4 n$. The proof is complete. □

6 Review

This theorem can be applied to many situations, such as $B = -A$ or $B = A$, and we can easily get many similar results. But there can be certain improvements.

One thing is that this theorem doesn't state what happens when there is torsion part. We guess the whole theorem is correct in any abelian groups and even these constant can be as the same. The problem is that lemmas in Section 4 which offer certain estimates are central in the final proof, while it seems hard to be stated in torsion cases. We may need some subtle changes in these proofs.

Another thing is that although d, C is independent of n, they seem to be too large (d is in fact exponential on α). The first theorem of [4] says if we restrict ourselves to the case $B = A$, we can have

$$d < \alpha + 1, \qquad \text{for sufficently large}\, n$$

which is just linear on α. Although we are dealing general cases, we think polynomial bound should be enough. The key point is lemma 2.4. It is a general result while we use it in a much more specific case.

References

[1] Ruzsa, Imre Z. ”Generalized arithmetical progressions and sumsets.” Acta Mathematica Hungarica 65.4 (1994): 379-388.

[2] Ruzsa, Imre Z. ”Arithmetic progressions in sumsets.” Acta Arithmetica 60.2 (1991): 191-202.

[3] Freiman, Gregory A. ”Structure theory of set addition.” ASTERISQUE-SOCIETE MATHEMATIQUE DE FRANCE 258 (1999): 1-20.

[4] Freiman, Gregory A. ”What is the structure of K if K+ K is small?.” Number Theory. Springer, Berlin, Heidelberg, 1987. 109-134.

Newlander-Nirenberg Theorem

Xiong Jiangnan

June 10, 2022

1 Introduction

Definition 1.1. A complex manifold is a manifold M, together with:
(1) Atlas $\{(U_\alpha, \phi_\alpha)\}$, where $U_\alpha \subset M$ open, $\phi_\alpha : U_\alpha \to \mathbb{D}^n$ homeomorphism.
where $\mathbb{D}^n$ is the unit open disk in $\mathbb{C}^n$.
(2) Transition map $\phi_\alpha \circ \phi_\beta^{-1} : \phi_\beta(U_\beta) \to \phi_\alpha(U_\alpha)$ is holomorphic.
Note: (2) implies that the transition maps are biholomorphic.
These data are called the complex structure on M.

Definition-Lemma 1.2. Let V be an $\mathbb{R}$-vector space.
(1) A complex structure on V is a linear map $J : V \to V$ s.t. $J^2 = -I$
(2) Complexification of V is the tensor product $V_\mathbb{C} = V \otimes_\mathbb{R} \mathbb{C}$
(3) Given complex structure J, we can regard V as a $\mathbb{C}$-vector space by setting $i \cdot v = Jv$, we denote this $\mathbb{C}$-vector space by V_J.
(4) J can extend to a $\mathbb{C}$-linear map $J : V_\mathbb{C} \to V_\mathbb{C}$ which commutes with the complex conjugation, inducing a decomposition $V_\mathbb{C} = V_\mathbb{C}^{1,0} \oplus V_\mathbb{C}^{0,1}$.
where $V_\mathbb{C}^{1,0}$ is i-eigenspace, $V_\mathbb{C}^{0,1}$ is $(-i)$-eigenspace.
(5) We have isomorphisms: $V_\mathbb{C}^i \cong V_\mathbb{C}^{-i}$ $\overline{\mathbb{C}}$-linear, $V_\mathbb{C}^i \cong V_J$ $\mathbb{C}$-linear.

Similar constructions can be made on manifolds.

Definition 1.3. An almost complex manifold is a manifold M^{2n}, together with (1,1)-tensor J, which is a complex structure on (co)tangent spaces pointwise. More precisely:
(1) $J \in \Gamma(M, T^*M \otimes TM)$, called the almost complex structure.
(2) $\forall p \in M$, $J_p : T_pM \to T_pM$ is a complex structure on T_pM.
(2') $\forall p \in M$, $J_p : T_p^*M \to T_p^*M$ is a complex structure on T_p^*M.

Proposition 1.4. *A complex manifold admits a natural almost complex structure.*

Proof. Let M be complex manifold of complex dimension n.
Then its underlying real manifold M_0 is of real dimension $2n$.
We have canonical $\mathbb{R}$-isomorphism $TM \cong TM_0$.
Multiplied by i on TM gives an almost complex structure on M_0. □

There are various examples of almost complex structures which do not arise from complex structures, for example S^6. This inverse problem is answered by Newlander-Nirenberg Theorem.

Theorem 1.5 (Newlander-Nirenberg)**.** *Let (M, J) be an almost complex manifold. Then there is a complex structure on M which induces the almost complex structure J if and only if J is integrable.*

Remark 1.6. For smooth manifolds and smooth J, this theorem is a corollary of Frobenius' theorem. We treat with weaker smoothness in this report.

2 Local Representation

Let M be a manifold.
Lowercase letters $j, k, l, \cdots$ denote indices from 1 to n.
Greek letters $\mu, \nu, \lambda, \cdots$ denote indices from 1 to $2n$.
The Einstein summation convention is employed.
Recall the complexification $TM_{\mathbb{C}} = TM \otimes_{\mathbb{R}} \mathbb{C}$, when M is $\mathbb{R}$-manifold

If M is a complex manifold, (z^j) be a complex coordinate.
The induced almost complex structure can be described as follows:
$(x^j = \operatorname{Re} z^j, y^j = \operatorname{Im} z^j)$ is real coordinate for the underlying manifold M_0.
$\mathrm{d}z^j \mapsto i\mathrm{d}z^j$ gives the almost complex structure $J : \mathrm{d}x^j \mapsto -\mathrm{d}y^j,\ \mathrm{d}y^j \mapsto \mathrm{d}x^j$
Moreover, $\mathrm{d}z^j, \mathrm{d}\overline{z}^j$ form a $\mathbb{R}$-frame of $T(M_0)_{\mathbb{C}}$.
The extension $J : T^*_{\mathbb{C}}M_0 \to T^*_{\mathbb{C}}M_0$ is $\mathrm{d}z^j \mapsto i\mathrm{d}z^j,\ \mathrm{d}\overline{z}^j \mapsto -i\mathrm{d}\overline{z}^j$

Now assume (M, J) is an almost complex manifold.
Let $(x^1, \cdots, x^{2n})$ be a $\mathbb{R}$ local coordinate system.
We can introduce complex coordinate (not necessarily analytic) by

$$z_j = x_j + i \cdot x_{j+n},\ z_{j+n} = \overline{z_j} = x_j - i \cdot x_{j+n}$$

All complex coordinates (z^μ) in this report satisfy the convention $z^{j+n} = \overline{z^j}$, and we often refer to it as (z^j), without explicitly mentioning $(\overline{z}^j)$.

Let $J \in \Gamma(M, T^*M_{\mathbb{C}} \otimes TM_{\mathbb{C}})$ be the natural extension.
Then J has local representation

$$J = J^\lambda_\mu \mathrm{d}z^\mu \otimes \frac{\partial}{\partial z^\lambda} \tag{2.1}$$

$$J\mathrm{d}z^\lambda = J^\lambda_\mu \mathrm{d}z^\mu \tag{2.2}$$

$$J\frac{\partial}{\partial z^\mu} = J^\lambda_\mu \frac{\partial}{\partial z^\lambda} \tag{2.3}$$

Now J is almost complex structure means

$$J_\mu^\lambda J_\lambda^\nu = -\delta_\mu^\nu \tag{2.4}$$

Definition 2.1. A function $w = w(z)$ of (z^j) is called analytic with respect to the almost complex structure J, if

$$J\mathrm{d}w = i\mathrm{d}w,\ \ J\mathrm{d}\overline{w} = -i\mathrm{d}\overline{w} \tag{2.5}$$

A chart (ζ^μ) is called analytic w.r.t. J if each ζ^j is.

Lemma 2.2. *J is induced by a complex manifold if and only if M can be covered by analytic complex coordinate charts.*

Proof. Let $(\zeta^\mu), (\eta^\mu)$ be two coordinate system analytic w.r.t. J
We shall prove that the transition map $(\zeta^j) \mapsto (\eta^j)$ is holomorphic.
It suffices to prove that $\dfrac{\partial \eta^j}{\partial \overline{\zeta}^k} = 0$. Calculate as follows:

$$\mathrm{d}\eta^j = \frac{\partial \eta^j}{\partial \zeta^k}\mathrm{d}\zeta^k + \frac{\partial \eta^j}{\partial \overline{\zeta}^k}\mathrm{d}\overline{\zeta}^k \tag{2.6}$$

$$J\mathrm{d}\eta^j = \frac{\partial \eta^j}{\partial \zeta^k}J\mathrm{d}\zeta^k + \frac{\partial \eta^j}{\partial \overline{\zeta}^k}J\mathrm{d}\overline{\zeta}^k \tag{2.7}$$

$$i\mathrm{d}\eta^j = i\frac{\partial \eta^j}{\partial \zeta^k}\mathrm{d}\zeta^k - i\frac{\partial \eta^j}{\partial \overline{\zeta}^k}\mathrm{d}\overline{\zeta}^k \tag{2.8}$$

$$(2.6) + i(2.8)\colon\ \frac{\partial \eta^j}{\partial \overline{\zeta}^k}\mathrm{d}\overline{\zeta}^k = 0 \tag{2.9}$$

□

This lemma shows that the question whether the almost complex structure comes from a complex manifold is purely local.

We now discuss this local problem.

By diagonalization, we can apply a suitable linear transformation, and then assume that (2.5) holds for $Z = (z^j)$ at $Z = 0$, i.e.

$$J\mathrm{d}z^j(0) = i\mathrm{d}z^j(0),\ \ J\mathrm{d}\overline{z}^j(0) = -i\mathrm{d}\overline{z}^j(0) \tag{2.10}$$

Under this assumption, we now give some equivalent description of (2.5).

Lemma 2.3. *For $w = w(Z)$, the following are equivalent near $Z = 0$.*
(1) w is holomorphic w.r.t. J
(2) $J\mathrm{d}w = i\mathrm{d}w$

(3) $i\dfrac{\partial w}{\partial z^\lambda} = \dfrac{\partial w}{\partial z^\mu} J^\mu_\lambda$

(4) $i\dfrac{\partial w}{\partial z^j} = \dfrac{\partial w}{\partial z^k} J^k_j + \dfrac{\partial w}{\partial \overline{z}^k} J^{k+n}_j$, $i\dfrac{\partial w}{\partial \overline{z}^j} = \dfrac{\partial w}{\partial z^k} J^k_{j+n} + \dfrac{\partial w}{\partial \overline{z}^k} J^{k+n}_{j+n}$

(5) $i\dfrac{\partial w}{\partial \overline{z}^j} = \dfrac{\partial w}{\partial z^k} J^k_{j+n} + \dfrac{\partial w}{\partial \overline{z}^k} J^{k+n}_{j+n}$

Proof. (1) $\Longrightarrow$ (2): trivial.
(2) $\Longrightarrow$ (1): J preserves complex conjugate.
(2) $\Longleftrightarrow$ (3): $\mathrm{d}w = \dfrac{\partial w}{\partial z^\mu}\mathrm{d}z^\mu$, $J\mathrm{d}w = \dfrac{\partial w}{\partial z^\mu} J^\mu_\lambda \mathrm{d}z^\lambda$, compare the coefficients.
(3) $\Longleftrightarrow$ (4): trivial.
(4) $\Longleftrightarrow$ (5): We restate the problem in the language of matrices:

$$J = [J^\mu_\lambda] = \begin{bmatrix} J^k_j & J^{k+n}_j \\ J^k_{j+n} & J^{k+n}_{j+n} \end{bmatrix}, X = \begin{bmatrix} \dfrac{\partial w}{\partial z^\mu} \end{bmatrix} = \begin{bmatrix} \frac{\partial w}{\partial z^k} \\ \frac{\partial w}{\partial \overline{z}^k} \end{bmatrix}$$

To prove that $(iI - J)X = 0$ is equivalent to the last n rows.

$$iI - J = \begin{bmatrix} i\delta^k_j - J^k_j & -J^{k+n}_j \\ -J^k_{j+n} & i\delta^k_j - J^{k+n}_{j+n} \end{bmatrix} \tag{2.11}$$

$$(iI - J)(0) = \begin{bmatrix} i\delta^k_j - J^k_j(0) & -J^{k+n}_j(0) \\ -J^k_{j+n}(0) & i\delta^k_j - J^{k+n}_{j+n}(0) \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 2iI \end{bmatrix} \tag{2.12}$$

$$(iI + J)(0) = \begin{bmatrix} i\delta^k_j + J^k_j(0) & J^{k+n}_j(0) \\ J^k_{j+n}(0) & i\delta^k_j + J^{k+n}_{j+n}(0) \end{bmatrix} = \begin{bmatrix} 2iI & 0 \\ 0 & 0 \end{bmatrix} \tag{2.13}$$

Claim: $A \in \mathbb{C}^{N\times N}$, $A^2 = I$, then $\mathrm{rank}(iI - A) + \mathrm{rank}(iI + A) = N$

The claim is an easy linear algebraic problem, we omit the proof.
By (2.12), the last n rows of $iI - J$ is linearly independent near $Z = 0$.
By (2.13), $\mathrm{rank}(iI + J) \geq n$ near $Z = 0$.
Now by Claim, $\mathrm{rank}(iI - J) \leq n$ near $Z = 0$.
So $\mathrm{rank}(iI - J) = n$ near $Z = 0$, and its row space is generated by the last n rows.
The proof is completed. □

Now we can solve from Lemma 2.3.(5) to obtain that

$$\frac{\partial w}{\partial \overline{z}^j} = a^k_j \frac{\partial w}{\partial z^k} \tag{2.14}$$

where a^k_j is defined near $Z = 0$ by $[a^k_j] = [i\delta^k_j - J^{k+n}_{j+n}]^{-1}[J^k_{j+n}]$
More precisely: $(i\delta^k_j - J^{k+n}_{j+n})a^l_k = J^l_{j+n}$
Then (2.14) is equivalent to that w is analytic w.r.t. J.
Note that $a^k_j(0) = 0$.

3 Integrability Condition

Assume that we are given analytic coordinate (ζ^j) w.r.t. J, then:

$$J\mathrm{d}\zeta^j = i\mathrm{d}\zeta^j,\ \frac{\partial \zeta^j}{\partial \overline{z}^k} = a^l_k \frac{\partial \zeta^j}{\partial z^l} \tag{3.1}$$

$$\mathrm{d}\zeta^j = \frac{\partial \zeta^j}{\partial z^l}\mathrm{d}z^l + \frac{\partial \zeta^j}{\partial \overline{z}^k}\mathrm{d}\overline{z}^k = \frac{\partial \zeta^j}{\partial z^l}\mathrm{d}z^l + a^l_k\frac{\partial \zeta^j}{\partial z^l}\mathrm{d}\overline{z}^k = \frac{\partial \zeta^j}{\partial z^l}(\mathrm{d}z^l + a^l_k\mathrm{d}\overline{z}^k) \tag{3.2}$$

Let $u^l = \mathrm{d}z^l + a^l_k\mathrm{d}\overline{z}^k$, then (3.2) is $\mathrm{d}\zeta^j = \dfrac{\partial \zeta^j}{\partial z^l}u^l$

Let $[b^l_j] = \left[\dfrac{\partial \zeta^j}{\partial z^l}\right]^{-1}$, then $u^l = b^l_j\mathrm{d}\zeta^j$

$$\mathrm{d}u^l = \mathrm{d}b^l_j \wedge \mathrm{d}\zeta^j = \frac{\partial \zeta^j}{\partial z^k}\mathrm{d}b^l_j \wedge u^k \tag{3.3}$$

This leads to our first formulation of integrability condition.

The integrability condition: the exterior differential of u^l is a sum of exterior products of 1-forms with u^k.

This formulation is close to the modern description of integrability condition. (We'll return to it in the end of this report)

However, it's not easy to use. We shall reformulate it into a more computable form.

Define operators $L_j = \dfrac{\partial}{\partial \overline{z}^j} - a^k_j\dfrac{\partial}{\partial z^k}$

$$\begin{aligned} L_kL_j &= \left(\frac{\partial}{\partial \overline{z}^k} - a^l_k\frac{\partial}{\partial z^l}\right)\left(\frac{\partial}{\partial \overline{z}^j} - a^m_j\frac{\partial}{\partial z^m}\right) \\ &= \frac{\partial^2}{\partial \overline{z}^k\partial \overline{z}^j} - a^l_k\frac{\partial^2}{\partial z^l\partial \overline{z}^j} - \frac{\partial a^m_j}{\partial \overline{z}^k}\frac{\partial}{\partial z^m} - a^m_j\frac{\partial^2}{\partial \overline{z}^k\partial z^m} + a^l_k\frac{\partial a^m_j}{\partial z^l}\frac{\partial}{\partial z^m} + a^l_ka^m_j\frac{\partial^2}{\partial z^l\partial z^m}\end{aligned}$$

Note that $\dfrac{\partial^2}{\partial z^\mu\partial z^\lambda} = \dfrac{\partial^2}{\partial z^\lambda\partial z^\mu}$, we get:

$$L_jL_k - L_kL_j = \left(\frac{\partial a^m_j}{\partial \overline{z}^k} - a^l_k\frac{\partial a^m_j}{\partial z_l} - \frac{\partial a^m_k}{\partial \overline{z}^j} + a^l_j\frac{\partial a^m_k}{\partial z_l}\right)\frac{\partial}{\partial z^m} \tag{3.4}$$

Let $L^m_{j,k} = \dfrac{\partial a^m_j}{\partial \overline{z}^k} - a^l_k\dfrac{\partial a^m_j}{\partial z^l}$, then (3.4) is

$$L_jL_k - L_kL_j = (L^m_{j,k} - L^m_{k,j})\frac{\partial}{\partial z^m}$$

Proposition 3.1. *The integrability condition is equivalent to that L_j commutes with each other.*

Proof. By (3.4), L_j commutes with each other if and only if $L^m_{j,k} = L^m_{k,j}$

$$\begin{aligned}
\mathrm{d}u^m &= \mathrm{d}(\mathrm{d}z^m + a^m_j \mathrm{d}\overline{z}^j) = \mathrm{d}a^m_j \wedge \mathrm{d}\overline{z}^j \\
&= \frac{\partial a^m_j}{\partial z^l}\mathrm{d}z^l \wedge \mathrm{d}\overline{z}^j + \frac{\partial a^m_j}{\partial \overline{z}^k}\mathrm{d}\overline{z}^k \wedge \mathrm{d}\overline{z}^j \\
&= \frac{\partial a^m_j}{\partial z^l}\mathrm{d}z^l \wedge \mathrm{d}\overline{z}^j + a^l_k \frac{\partial a^m_j}{\partial z^l}\mathrm{d}\overline{z}^k \wedge \mathrm{d}\overline{z}^j + L^m_{j,k}\mathrm{d}\overline{z}^k \wedge \mathrm{d}\overline{z}^j \\
&= \frac{\partial a^m_j}{\partial z^l} u^l \wedge \mathrm{d}\overline{z}^j + L^m_{j,k}\mathrm{d}\overline{z}^k \wedge \mathrm{d}\overline{z}^j \\
&= -(\frac{\partial a^m_j}{\partial z_l}\mathrm{d}\overline{z}^j) \wedge u^l + L^m_{j,k}\mathrm{d}\overline{z}^k \wedge \mathrm{d}\overline{z}^j
\end{aligned}$$

If $L^m_{j,k} = L^m_{k,j}$, then $L^m_{j,k}\mathrm{d}\overline{z}^k \wedge \mathrm{d}\overline{z}^j = 0$, the integrability condition holds.
On the other hand, assume that for 1-forms $v_l = v_{l,k}\mathrm{d}z^k + v_{l,k+n}\mathrm{d}\overline{z}^k$

$$L^m_{j,k}\mathrm{d}\overline{z}^k \wedge \overline{z}^j = v_l \wedge u^l$$

Then we have:

$$\begin{aligned}
&L^m_{j,k}\mathrm{d}\overline{z}^k \wedge \overline{z}^j = v_l \wedge u^l = (v_{l,k}\mathrm{d}z^k + v_{l,k+n}\mathrm{d}\overline{z}^k) \wedge (\mathrm{d}z^l + a^l_j \mathrm{d}\overline{z}^j) \\
&= v_{l,k}\mathrm{d}z^k \mathrm{d}z^l + v_{l,k+n}\mathrm{d}\overline{z}^k \mathrm{d}z^l + v_{l,k}a^l_j \mathrm{d}z^k \mathrm{d}\overline{z}^j + v_{l,k+n}a^l_j \mathrm{d}\overline{z}^k \mathrm{d}\overline{z}^j \\
&= v_{l,k}\mathrm{d}z^k \mathrm{d}z^l + (v_{k,j+n} - v_{l,k}a^l_j)\mathrm{d}\overline{z}^j \mathrm{d}z^k + v_{l,k+n}a^l_j \mathrm{d}\overline{z}^k \mathrm{d}\overline{z}^j
\end{aligned}$$

By comparing coefficients, we get

$$\begin{aligned}
v_{l,k} &= v_{k,l} \\
v_{k,j+n} &= v_{l,k}a^l_j
\end{aligned}$$

Therefore

$$\begin{aligned}
&v_{l,k+n}a^l_j = v_{m,l}a^m_k a^l_j = v_{l,m}a^m_k a^l_j = v_{m,j+n}a^m_k = v_{l,j+n}a^l_k \\
&L^m_{j,k}\mathrm{d}\overline{z}^k \wedge \overline{z}^j = v_{l,k+n}a^l_j \mathrm{d}\overline{z}^k \wedge \mathrm{d}\overline{z}^j = 0
\end{aligned}$$

We get $L^m_{j,k} = L^m_{k,j}$. The proof is completed. □

Conclusion: If J is induced by a complex manifold, then the integrability condition holds: L_j commutes with each other.

We now need to do the converse. Assume that the integrability condition holds, to find a complex analytic coordinate system (ζ^j).

Lemma 3.2. *Assume that the following equations hold:*

$$\frac{\partial z^k}{\partial \overline{\zeta}^j} = -a^k_m \frac{\partial \overline{z}^m}{\partial \overline{\zeta}^j} \tag{3.5}$$

Then (ζ^j) *is analytic w.r.t.* J

Proof. Evaluation at $Z = 0$, then $\dfrac{\partial z^k}{\partial \overline{\zeta}^j}(0) = 0$.

Since $(z^\mu), (\zeta^\lambda)$ are two coordinate system, $\det\left(\dfrac{\partial z^\mu}{\partial \zeta^\lambda}\right) \neq 0$

Then $\left(\dfrac{\partial \overline{z}^k}{\partial \overline{\zeta}^j}\right)$ is non-singular near $Z = 0$

$$0 = \frac{\partial \zeta^j}{\partial \overline{\zeta}^k} = \frac{\partial \zeta^j}{\partial z^l}\frac{\partial z^l}{\partial \overline{\zeta}^k} + \frac{\partial \zeta^j}{\partial \overline{z}^m}\frac{\partial \overline{z}^m}{\partial \overline{\zeta}^k} = -a^l_m \frac{\partial \overline{z}^m}{\partial \overline{\zeta}^k}\frac{\partial \zeta^j}{\partial z^l} + \frac{\partial \zeta^j}{\partial \overline{z}^m}\frac{\partial \overline{z}^m}{\partial \overline{\zeta}^k} = \frac{\partial \overline{z}^m}{\partial \overline{\zeta}^k} L_m \zeta^j$$

Therefore $L_m\zeta^j = 0$, (ζ^j) is analytic w.r.t. J □

4 Integral operator

Through this section, we use $z, \zeta, \cdots$ to denote single complex variable in $\mathbb{C}$

We want to solve equation of the following type:

$$\frac{\partial w}{\partial \overline{z}} = f(z) \tag{4.1}$$

Definition 4.1. Define an operator T as follows

$$Tf(\zeta) = \frac{1}{2\pi i}\int_D \frac{f(z)}{z - \zeta}\mathrm{d}z\mathrm{d}\overline{z} \tag{4.2}$$

where $D = B(0, R) = \{z : |z| < R\}$ for some fixed $R > 0$

Example 4.2. For all $\zeta \in D = B(0, R)$

$$T1(\zeta) = \frac{1}{2\pi i}\int_D \frac{\mathrm{d}z\mathrm{d}\overline{z}}{z - \zeta} = \overline{\zeta} \tag{4.3}$$

Proof. Let $z = x + iy = re^{i\theta}$. Let $S(0, r) = \{z \in \mathbb{C} : |z| = r\}$

$$\frac{1}{2\pi i}\int_D \frac{\mathrm{d}z\mathrm{d}\overline{z}}{z - \zeta} = \frac{-1}{\pi}\int_D \frac{\mathrm{d}x\mathrm{d}y}{z - \zeta} = \frac{-1}{\pi}\int_0^R r\mathrm{d}r \int_{S(0,r)} \frac{\mathrm{d}\theta}{z - \zeta}$$

$$\phi(r) := \int_{S(0,r)} \frac{\mathrm{d}\theta}{z - \zeta} = \int_{S(0,r)} \frac{ire^{i\theta}\mathrm{d}\theta}{iz(z - \zeta)} = \int_{S(0,r)} \frac{\mathrm{d}z}{iz(z - \zeta)}$$

If $\zeta = 0$, then $\forall 0 < r < R$

$$\phi(r) = \int_{S(0,r)} \frac{\mathrm{d}z}{iz(z-\zeta)} = 2\pi i \cdot \operatorname*{Res}_{z=0} \frac{1}{iz^2} = 0$$

If $\zeta \neq 0$,

$$\phi(r) = \begin{cases} 2\pi i \cdot \operatorname*{Res}_{z=0} \dfrac{1}{iz(z-\zeta)} = -\dfrac{2\pi}{\zeta}, \text{ if } 0 < r < |\zeta| \\ 2\pi i \cdot \left(\operatorname*{Res}_{z=0} \dfrac{1}{iz(z-\zeta)} + \operatorname*{Res}_{z=\zeta} \dfrac{1}{iz(z-\zeta)} \right) = 0, \text{ if } |\zeta| < r < R \end{cases}$$

In both cases,

$$\frac{1}{2\pi i} \int_D \frac{\mathrm{d}z\mathrm{d}\overline{z}}{z-\zeta} = \frac{-1}{\pi} \int_0^R \phi(r) r\mathrm{d}r = \int_0^{|\zeta|} \frac{2}{\zeta} r\mathrm{d}r = \frac{|\zeta|^2}{\zeta} = \overline{\zeta}$$

□

Theorem 4.3. *Let $\alpha > 0$, $f \in C^\alpha(D)$, $D = B(0,R)$, i.e.*

$$\mathrm{Lip}_\alpha(f) = \sup_{z_1, z_2 \in D} \frac{|f(z_1) - f(z_2)|}{|z_1 - z_2|^\alpha} < \infty$$

Then $Tf(\zeta) \in C^1(D)$, with derivatives:

$$\frac{\partial}{\partial \overline{\zeta}} Tf(\zeta) = f(\zeta) \tag{4.4}$$

$$\frac{\partial}{\partial \zeta} Tf(\zeta) = \frac{1}{2\pi i} \int_D \frac{f(z) - f(\zeta)}{(z-\zeta)^2} \mathrm{d}z\mathrm{d}\overline{z} \tag{4.5}$$

Proof. Recall: $\frac{\partial}{\partial \zeta} = \frac{1}{2}(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y})$, $\frac{\partial}{\partial \overline{\zeta}} = \frac{1}{2}(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y})$

$$\begin{aligned} \frac{\partial}{\partial \overline{\zeta}} Tf &= \frac{1}{2}(\frac{\partial}{\partial x} Tf + i\frac{\partial}{\partial y} Tf) \\ &= \lim_{h \to 0} \left(\frac{Tf(\zeta + h) - Tf(\zeta)}{2h} + i\frac{Tf(\zeta + ih) - Tf(\zeta)}{2h} \right) \\ &= \lim_{h \to 0} \int_D \frac{f(z)}{2h} \left(\frac{1}{z - \zeta - h} + \frac{i}{z - \zeta - ih} - \frac{i+1}{z-\zeta} \right) \mathrm{d}z\mathrm{d}\overline{z} \\ &= \lim_{h \to 0} \int_D \frac{\frac{1-i}{2} h f(z)}{(z-\zeta)(z-\zeta-h)(z-\zeta-ih)} \mathrm{d}z\mathrm{d}\overline{z} \\ &= \lim_{h \to 0} \int_D \frac{\frac{1-i}{2} h (f(z) - f(\zeta))}{(z-\zeta)(z-\zeta-h)(z-\zeta-ih)} \mathrm{d}z\mathrm{d}\overline{z} + f(\zeta) \frac{\partial}{\partial \overline{\zeta}} T1 \\ &= \lim_{h \to 0} \frac{1-i}{2} h \int_D \frac{f(z) - f(\zeta)}{(z-\zeta)(z-\zeta-h)(z-\zeta-ih)} \mathrm{d}z\mathrm{d}\overline{z} + f(\zeta) \end{aligned}$$

Now estimate

$$\begin{aligned}&\left|\int_D \frac{f(z)-f(\zeta)}{(z-\zeta)(z-\zeta-h)(z-\zeta-ih)}\mathrm{d}z\mathrm{d}\overline{z}\right|\\ &\leq C\int_D \frac{|f(z)-f(\zeta)|}{|z-\zeta||z-\zeta-h||z-\zeta-ih|}\mathrm{d}x\mathrm{d}y\\ &\leq C\mathrm{Lip}_\alpha(f)\int_D \frac{1}{|z-\zeta|^{1-\alpha}|z-\zeta-h||z-\zeta-ih|}\mathrm{d}x\mathrm{d}y\end{aligned}$$

Let $D_1 = B(\zeta, \frac{h}{2})$, $D_2 = B(\zeta + h, \frac{h}{2})$, $D_3 = B(\zeta + ih, \frac{h}{2})$

$$\begin{aligned}&\int_{D_1} \frac{1}{|z-\zeta|^{1-\alpha}|z-\zeta-h||z-\zeta-ih|}\mathrm{d}x\mathrm{d}y\\ &\leq \frac{C}{h^2}\int_0^{\frac{h}{2}} \frac{1}{r^{1-\alpha}} r\mathrm{d}r = C\cdot h^{\alpha-1}\end{aligned}$$

$$\begin{aligned}&\int_{D_2} \frac{1}{|z-\zeta|^{1-\alpha}|z-\zeta-h||z-\zeta-ih|}\mathrm{d}x\mathrm{d}y\\ &\leq \frac{C}{h^{2-\alpha}}\int_0^{\frac{h}{2}} \frac{1}{r} r\mathrm{d}r = C\cdot h^{\alpha-1}\end{aligned}$$

$$\begin{aligned}&\int_{D_3} \frac{1}{|z-\zeta|^{1-\alpha}|z-\zeta-h||z-\zeta-ih|}\mathrm{d}x\mathrm{d}y\\ &\leq \frac{C}{h^{2-\alpha}}\int_0^{\frac{h}{2}} \frac{1}{r} r\mathrm{d}r = C\cdot h^{\alpha-1}\end{aligned}$$

Let $D_4 = B(\zeta, \frac{3h}{2}) - (D_1 \cup D_2 \cup D_3)$, $D_5 = D - B(\zeta, \frac{3h}{2})$

$$\begin{aligned}&\int_{D_4} \frac{1}{|z-\zeta|^{1-\alpha}|z-\zeta-h||z-\zeta-ih|}\mathrm{d}x\mathrm{d}y\\ &\leq C\cdot h^2\cdot\frac{1}{h^{3-\alpha}} = C\cdot h^{\alpha-1}\\ &\int_{D_5} \frac{1}{|z-\zeta|^{1-\alpha}|z-\zeta-h||z-\zeta-ih|}\mathrm{d}x\mathrm{d}y\\ &\leq C\int_{\frac{3h}{2}}^{R} \frac{1}{r^{3-\alpha}} r\mathrm{d}r \leq C\int_{\frac{3h}{2}}^{\infty} \frac{1}{r^{2-\alpha}}\mathrm{d}r = C\cdot h^{\alpha-1}\end{aligned}$$

Therefore we have

$$\left|\int_D \frac{f(z)-f(\zeta)}{(z-\zeta)(z-\zeta-h)(z-\zeta-ih)}\mathrm{d}z\mathrm{d}\overline{z}\right| \leq C\mathrm{Lip}_\alpha(f)\cdot h^{\alpha-1} \tag{4.6}$$

Similar argument holds for $\frac{\partial}{\partial\zeta}$:

$$
\begin{aligned}
&\frac{\partial}{\partial\zeta}Tf = \frac{1}{2}(\frac{\partial}{\partial x}Tf - i\frac{\partial}{\partial y}Tf)\\
=&\lim_{h\to 0}\int_D \frac{f(z)}{2h}\left(\frac{1}{z-\zeta-h} - \frac{i}{z-\zeta-ih} - \frac{1-i}{z-\zeta}\right)\mathrm{d}z\mathrm{d}\overline{z}\\
=&\lim_{h\to 0}\int_D \frac{f(z)\cdot(z-\zeta-\frac{i+1}{2}h)}{(z-\zeta)(z-\zeta-h)(z-\zeta-ih)}\mathrm{d}z\mathrm{d}\overline{z}\\
=&\lim_{h\to 0}\int_D \frac{(f(z)-f(\zeta))\cdot(z-\zeta-\frac{i+1}{2}h)}{(z-\zeta)(z-\zeta-h)(z-\zeta-ih)}\mathrm{d}z\mathrm{d}\overline{z}\\
=&\int_D \frac{f(z)-f(\zeta)}{(z-\zeta)^2}\mathrm{d}z\mathrm{d}\overline{z} + \lim_{h\to 0}\int_D \frac{\frac{i+1}{2}h(f(z)-f(\zeta))}{(z-\zeta)(z-\zeta-h)(z-\zeta-ih)}\mathrm{d}z\mathrm{d}\overline{z}\\
-&\lim_{h\to 0}\int_D \frac{ih^2(f(z)-f(\zeta))}{(z-\zeta)^2(z-\zeta-h)(z-\zeta-ih)}\mathrm{d}z\mathrm{d}\overline{z} = \int_D \frac{f(z)-f(\zeta)}{(z-\zeta)^2}\mathrm{d}z\mathrm{d}\overline{z}
\end{aligned}
$$

The proof is completed. □

Lemma 4.4 (Mean Value Theorem)**.** *Let ω be a convex domain, $f(z) \in C^1(\omega)$*

$$\mathrm{Lip}_1(f) \le \sup\left|\frac{\partial f}{\partial z}\right| + \sup\left|\frac{\partial f}{\partial \overline{z}}\right| \tag{4.7}$$

$$f(z_1) - f(z_0) = \frac{\partial f}{\partial z}\cdot(z_1 - z_0) + \frac{\partial f}{\partial \overline{z}}\cdot(\overline{z}_1 - \overline{z}_0) \tag{4.8}$$

Proof. Let $z_t = z_0 + t(z_1 - z_0)$, $g(t) = f(z_t) = f(z_0 + t(z_1 - z_0))$

$$
\begin{aligned}
|f(z_1) - f(z_0)| &= |\int_0^1 g'(t)\mathrm{d}t| \le \int_0^1 |g'(t)|\mathrm{d}t\\
g'(t) &= \frac{\partial f}{\partial z}(z_t)\cdot(z_1 - z_0) + \frac{\partial f}{\partial \overline{z}}(z_t)\cdot(\overline{z_1} - \overline{z_0})
\end{aligned}
$$

This proves (4.7). For (4.8), use Mean Value Theorem in $\mathbb{R}^2$ twice:

$$
\begin{aligned}
&f(z_1) - f(z_0) = \frac{\partial f}{\partial x}(z')\cdot(x_1 - x_0) + \frac{\partial f}{\partial y}(z'')\cdot(y_1 - y_0)\\
=&\left(\frac{\partial f}{\partial z} + \frac{\partial f}{\partial \overline{z}}\right)\frac{z_1 - z_0 + \overline{z}_1 - \overline{z}_0}{2} + \left(i\frac{\partial f}{\partial z} - i\frac{\partial f}{\partial \overline{z}}\right)\frac{z_1 - z_0 - \overline{z}_1 + \overline{z}_0}{2i}\\
=&\frac{\partial f}{\partial z}\cdot(z_1 - z_0) + \frac{\partial f}{\partial \overline{z}}\cdot(\overline{z}_1 - \overline{z}_0)
\end{aligned}
$$

Be careful what this formula means. □

Theorem 4.5. *We have the following estimates:*

$$|Tf| \leq C \cdot R \cdot \sup|f| \tag{4.9}$$

$$\left|\frac{\partial}{\partial\overline{\zeta}}Tf\right| \leq \sup|f| \tag{4.10}$$

$$\left|\frac{\partial}{\partial\zeta}Tf\right| \leq C \cdot R^{\alpha} \cdot \mathrm{Lip}_{\alpha}(f) \tag{4.11}$$

$$\mathrm{Lip}_1(Tf) \leq \sup|f| + C \cdot R^{\alpha} \cdot \mathrm{Lip}_{\alpha}(f) \tag{4.12}$$

$$\mathrm{Lip}_{\alpha}\left(\frac{\partial}{\partial\overline{\zeta}}Tf\right) = \mathrm{Lip}_{\alpha}(f) \tag{4.13}$$

$$\mathrm{Lip}_{\alpha}\left(\frac{\partial}{\partial\zeta}Tf\right) \leq C \cdot \mathrm{Lip}_{\alpha}(f) \tag{4.14}$$

Proof. By (4.4), it's trivial to get (4.10),(4.13)

$$|Tf(\zeta)| \leq C\int_D \frac{|f(z)|}{|z-\zeta|}\mathrm{d}x\mathrm{d}y \leq C\sup|f|\int_{B(0,2R)}\frac{1}{|z|}\mathrm{d}x\mathrm{d}y = CR\sup|f|$$

$$\left|\frac{\partial}{\partial\zeta}Tf(\zeta)\right| \leq C\int_D \frac{|f(z)-f(\zeta)|}{|z-\zeta|^2}\mathrm{d}x\mathrm{d}y \leq C\int_{B(0,2R)}\frac{\mathrm{Lip}_{\alpha}(f)}{|z|^{2-\alpha}}\mathrm{d}x\mathrm{d}y = CR^{\alpha}\mathrm{Lip}_{\alpha}(f)$$

$$\mathrm{Lip}_1(Tf) \leq \sup\left|\frac{\partial}{\partial\zeta}Tf(\zeta)\right| + \left|\frac{\partial}{\partial\overline{\zeta}}Tf(\zeta)\right| \leq \sup|f| + CR^{\alpha}\mathrm{Lip}_{\alpha}(f)$$

$\forall\zeta_1,\zeta_2 \in D$, let $D_1 = B(\zeta_2, 2|\zeta_1-\zeta_2|)\cap D$, $D_2 = D - D_1$

$$\begin{aligned}\frac{\partial}{\partial\zeta}Tf(\zeta_1) - \frac{\partial}{\partial\zeta}Tf(\zeta_2) &= \frac{1}{2\pi i}\int_{D_1}\frac{f(z)-f(\zeta_1)}{(z-\zeta_1)^2}\mathrm{d}z\mathrm{d}\overline{z} - \frac{1}{2\pi i}\int_{D_1}\frac{f(z)-f(\zeta_2)}{(z-\zeta)^2}\mathrm{d}z\mathrm{d}\overline{z}\\ &+\frac{1}{2\pi i}\int_{D_2}(f(z)-f(\zeta_1))\left(\frac{1}{(z-\zeta_1)^2}-\frac{1}{(z-\zeta_2)^2}\right)\mathrm{d}z\mathrm{d}\overline{z} - \frac{1}{2\pi i}\int_{D_2}\frac{f(\zeta_1)-f(\zeta_2)}{(z-\zeta_2)^2}\mathrm{d}z\mathrm{d}\overline{z}\end{aligned}$$

$$\left|\int_{D_1}\frac{f(z)-f(\zeta_1)}{(z-\zeta_1)^2}\mathrm{d}z\mathrm{d}\overline{z}\right| \leq C\int_{B(\zeta_1,3|\zeta_1-\zeta_2|)}\frac{\mathrm{Lip}_{\alpha}(f)}{|z-\zeta_1|^{2-\alpha}}\mathrm{d}x\mathrm{d}y \leq C\mathrm{Lip}_{\alpha}(f)|\zeta_1-\zeta_2|^{\alpha}$$

$$\left|\int_{D_1}\frac{f(z)-f(\zeta_2)}{(z-\zeta_2)^2}\mathrm{d}z\mathrm{d}\overline{z}\right| \leq C\int_{B(\zeta_2,2|\zeta_1-\zeta_2|)}\frac{\mathrm{Lip}_{\alpha}(f)}{|z-\zeta_2|^{2-\alpha}}\mathrm{d}x\mathrm{d}y \leq C\mathrm{Lip}_{\alpha}(f)|\zeta_1-\zeta_2|^{\alpha}$$

$$
\begin{aligned}
&\left|\frac{1}{2\pi i}\int_{D_2}(f(z)-f(\zeta_1))\left(\frac{1}{(z-\zeta_1)^2}-\frac{1}{(z-\zeta_2)^2}\right)\mathrm{d}z\mathrm{d}\overline{z}\right|\\
=&\left|\frac{1}{2\pi i}\int_{D_2}(f(z)-f(\zeta_1))\frac{(\zeta_1-\zeta_2)(2z-\zeta_1-\zeta_2)}{(z-\zeta_1)^2(z-\zeta_2)^2}\mathrm{d}z\mathrm{d}\overline{z}\right|\\
\leq& C\int_{\mathbb{C}-B(\zeta_2,2|\zeta_1-\zeta_2|)}\frac{\mathrm{Lip}_\alpha(f)|\zeta_1-\zeta_2|}{|z-\zeta_2|^{3-\alpha}}\mathrm{d}x\mathrm{d}y\\
=&C\int_{2|\zeta_1-\zeta_2|}^{\infty}\frac{\mathrm{Lip}_\alpha(f)|\zeta_1-\zeta_2|}{r^{3-\alpha}}\mathrm{d}r=C\mathrm{Lip}_\alpha(f)|\zeta_1-\zeta_2|^\alpha
\end{aligned}
$$

$$
\begin{aligned}
&\frac{1}{2\pi i}\int_{D_2}\frac{\mathrm{d}z\mathrm{d}\overline{z}}{(z-\zeta_2)^2}=\frac{1}{2\pi i}\int_{D_2}\frac{\partial}{\partial\zeta_2}\left(\frac{-1}{z-\zeta_2}\right)\mathrm{d}z\mathrm{d}\overline{z}=\frac{\partial}{\partial\zeta_2}\left(\frac{-1}{2\pi i}\int_{D_2}\frac{1}{z-\zeta_2}\mathrm{d}z\mathrm{d}\overline{z}\right)\\
&=\frac{\partial}{\partial\zeta_2}\left(\frac{1}{2\pi i}\int_{D_1}\frac{1}{z-\zeta_2}\mathrm{d}z\mathrm{d}\overline{z}-\frac{1}{2\pi i}\int_{D}\frac{1}{z-\zeta_2}\mathrm{d}z\mathrm{d}\overline{z}\right)=\frac{\partial}{\partial\zeta_2}(0-\overline{\zeta_2})=0
\end{aligned}
$$

(4.14) is proved by these together. □

5 Higher Dimension Case

We now deduce corresponding conclusions in higher dimension.
We work in $\mathbb{C}^n$, let $0<r<\frac{1}{4}, 0<\alpha<1$ fixed, $D=B(0,r)\subset\mathbb{C}$
Consider functions defined on $\Omega=\{(\zeta^1,\cdots,\zeta^n):|\zeta^j|\leq r,\forall 1\leq j\leq n\}$
Define the integral operators:

$$
T^jf(\zeta^1,\cdots,\zeta^n)=\frac{1}{2\pi i}\int_D f(\zeta^1,\cdots,\zeta^{j-1},\tau,\zeta^{j+1},\cdots,\zeta^n)\frac{\mathrm{d}\tau\mathrm{d}\overline{\tau}}{\tau-\zeta^j}\tag{5.1}
$$

Let ∂_j denote either $\dfrac{\partial}{\partial\zeta^j}$ or $\dfrac{\partial}{\partial\overline{\zeta}^j}$
Let δ_j denote a difference quotient operator of the form:

$$
\delta_jf=|\delta\zeta^j|^{-\alpha}(f(\zeta^1,\cdots,\zeta^j+\delta\zeta^j,\cdots,\zeta^n)-f(\zeta^1,\cdots,\zeta^j,\cdots,\zeta^n))\tag{5.2}
$$

Theorem 5.1. *Under suitable differentiability conditions, we have:*
(1) $\dfrac{\partial}{\partial\overline{\zeta}^j}T^jf=f$

$$
\frac{\partial}{\partial\zeta^j}T^jf=\frac{1}{2\pi i}\int_D\frac{f(\zeta^1,\cdots,\zeta^{j-1},\tau,\zeta^{j+1},\cdots,\zeta^n)-f(\zeta^1,\cdots,\zeta^n)}{(\tau-\zeta^j)^2}\mathrm{d}\tau\mathrm{d}\overline{\tau}
$$

(2) $\forall k \neq j$, $\partial_k T^j = T^j \partial_k$, $\delta_k T^j = T^j \delta_k$, $\delta_k \partial_j = \partial_j \delta_k$
(3) We have the following estimates:

$$\left|T^j f\right| \leq Cr \sup |f| \tag{5.3}$$

$$\left.\begin{aligned} &\left|\frac{\partial}{\partial \bar{\zeta}^j} T^j f\right| \leq \sup |f| \\ &\left|\frac{\partial}{\partial \zeta^j} T^j f\right| \leq Cr^\alpha \sup |\delta_j f| \end{aligned}\right\} \left|\partial_j T^j f\right| \leq \sup |f| + Cr^\alpha \sup |\delta_j f| \tag{5.4}$$

$$\left|\delta_j T^j f\right| \leq Cr^{1-\alpha} \sup |f| + Cr \sup |\delta_j f| \tag{5.5}$$

$$\left.\begin{aligned} &\left|\delta_j \frac{\partial}{\partial \bar{\zeta}^j} T^j f\right| \leq \sup |\delta_j f| \\ &\left|\delta_j \frac{\partial}{\partial \zeta^j} T^j f\right| \leq C \sup |\delta_j f| \end{aligned}\right\} \left|\delta_j \partial_j T^j f\right| \leq C \sup |\delta_j f| \tag{5.6}$$

Proof. (1) follows from Thm 4.3, (2) is obvious, (3) follows from Thm 4.5 □

Corollary 5.2 (potential theoretic lemma).

$$\begin{aligned} &\sup |T^j f| + r^\alpha \sup |\delta_j T^j f| + r \sup |\partial_j T^j f| + r^{1+\alpha} \sup |\delta_j \partial_j T^j f| \\ &\leq Cr \sup |f| + Cr^{1+\alpha} \sup |\delta_j f| \end{aligned} \tag{5.7}$$

Theorem 5.3. *Consider the following equations under suitable differentiability conditions*

$$\frac{\partial \omega}{\partial \bar{\zeta}^j} = f_j \tag{5.8}$$

Write $F = (f_1, \cdots, f_n)$, define the combined integral operator:

$$TF = \sum_{s=0}^{n-1} \frac{(-1)^s}{(s+1)!} \sum_{j_1, \cdots, j_s, k \text{ distinct}} T^{j_1} \frac{\partial}{\partial \bar{\zeta}^{j_1}} \cdots T^{j_s} \frac{\partial}{\partial \bar{\zeta}^{j_s}} T^k f_k \tag{5.9}$$

Then $TF \in C^1$, with derivatives

$$\frac{\partial}{\partial \bar{\zeta}^j} TF = f_j + \sum_{s=0}^{n-2} \frac{(-1)^s}{(s+2)!} \sum_{j_1, \cdots, j_s, k, j \text{ distinct}} T^{j_1} \frac{\partial}{\partial \bar{\zeta}^{j_1}} \cdots T^{j_s} \frac{\partial}{\partial \bar{\zeta}^{j_s}} T^k \left(\frac{\partial f_k}{\partial \bar{\zeta}^j} - \frac{\partial f_j}{\partial \bar{\zeta}^k} \right) \tag{5.10}$$

In particular, if F satisfies the compatibility relations

$$\frac{\partial}{\partial \bar{\zeta}^k} f_j = \frac{\partial}{\partial \bar{\zeta}^j} f_k \tag{5.11}$$

then TF is a solution to (5.8)

Proof.

$$
\begin{aligned}
&\frac{\partial}{\partial\overline{\zeta}^j}TF=\sum_{s=0}^{n-1}\frac{(-1)^s}{(s+1)!}\sum_{j_1,\cdots,j_s,k\text{ distinct}}\frac{\partial}{\partial\overline{\zeta}^j}T^{j_1}\frac{\partial}{\partial\overline{\zeta}^{j_1}}\cdots T^{j_s}\frac{\partial}{\partial\overline{\zeta}^{j_s}}T^kf_k\\
&=\sum_{s=0}^{n-1}\frac{(-1)^s}{(s+1)!}\left(\sum_{j_1,\cdots,j_s,k,j\text{ distinct}}T^{j_1}\frac{\partial}{\partial\overline{\zeta}^{j_1}}\cdots T^{j_s}\frac{\partial}{\partial\overline{\zeta}^{j_s}}T^k\frac{\partial}{\partial\overline{\zeta}^j}f_k\right.\\
&+\sum_{j_1,\cdots,j_s,j\text{ distinct}}T^{j_1}\frac{\partial}{\partial\overline{\zeta}^{j_1}}\cdots T^{j_s}\frac{\partial}{\partial\overline{\zeta}^{j_s}}f_j\\
&+\left.\sum_{j_1,\cdots,\hat{j_l},\cdots,j_s,k\text{ distinct}}T^{j_1}\frac{\partial}{\partial\overline{\zeta}^{j_1}}\cdots\widehat{T^{j_l}\frac{\partial}{\partial\overline{\zeta}^{j_l}}}\cdots T^{j_s}\frac{\partial}{\partial\overline{\zeta}^{j_s}}T^k\frac{\partial}{\partial\overline{\zeta}^j}f_k\right)\\
&=\sum_{s=0}^{n-2}\left(\sum_{j_1,\cdots,j_s,k,j\text{ distinct}}T^{j_1}\frac{\partial}{\partial\overline{\zeta}^{j_1}}\cdots T^{j_s}\frac{\partial}{\partial\overline{\zeta}^{j_s}}T^k\frac{\partial}{\partial\overline{\zeta}^j}f_k\cdot\left(\frac{(-1)^s}{(s+1)!}+(s+1)\frac{(-1)^{s+1}}{(s+2)!}\right)\right.\\
&+\left.\sum_{j_1,\cdots,j_s,k,j\text{ distinct}}T^{j_1}\frac{\partial}{\partial\overline{\zeta}^{j_1}}\cdots T^{j_s}\frac{\partial}{\partial\overline{\zeta}^{j_s}}T^k\frac{\partial}{\partial\overline{\zeta}^k}f_j\cdot\frac{(-1)^{s+1}}{(s+2)!}\right)+f_j\\
&=\sum_{s=0}^{n-2}\frac{(-1)^s}{(s+2)!}\sum_{j_1,\cdots,j_s,k,j\text{ distinct}}T^{j_1}\frac{\partial}{\partial\overline{\zeta}^{j_1}}\cdots T^{j_s}\frac{\partial}{\partial\overline{\zeta}^{j_s}}T^k\left(\frac{\partial}{\partial\overline{\zeta}^j}f_k-\frac{\partial}{\partial\overline{\zeta}^k}f_j\right)+f_j
\end{aligned}
$$

□

6 Normed Function Spaces

We introduce some norms on the space of functions on $\Omega=\{(\zeta^1,\cdots,\zeta^n):|\zeta^j|<r,\forall j\}$
Recall that:
∂_j denotes either $\dfrac{\partial}{\partial\zeta^j}$ or $\dfrac{\partial}{\partial\overline{\zeta}^j}$
δ_j denotes a difference quotient operator of the form:

$$\delta_jf=|\delta\zeta^j|^{-\alpha}(f(\zeta^1,\cdots,\zeta^j+\delta\zeta^j,\cdots,\zeta^n)-f(\zeta^1,\cdots,\zeta^j,\cdots,\zeta^n))\tag{6.1}$$

Let ∂^m denote an operator of the form $\partial_{j_1}\cdots\partial_{j_m}$ with $j_1,\cdots,j_m$ distinct.
Let $\partial^{m;j}$ denote an operator of the form $\partial_{j_1}\cdots\partial_{j_m}$ with $j,j_1,\cdots,j_m$ distinct.
Let δ^m denote an operator of the form $\delta_{j_1}\cdots\delta_{j_m}$ with $j_1,\cdots,j_m$ distinct.
Let $\delta^{m;j}$ denote an operator of the form $\delta_{j_1}\cdots\delta_{j_m}$ with $j,j_1,\cdots,j_m$ distinct.

Definition 6.1 (Norms)**.** For functions $z, f : \Omega \to \mathbb{C}$, define

$$H_\alpha[z] = \sum_{k=0}^{n} \frac{r^{k\alpha}}{k!} \sup |\delta^k z| \tag{6.2}$$

$$|z|_n = \sum_{k=0}^{n} \frac{r^k}{k!} \sup |\partial^k z| \tag{6.3}$$

$$|z|_{n+\alpha} = \sum_{k=0}^{n} \frac{r^k}{k!} \sup H_\alpha[\partial^k z] \tag{6.4}$$

$$|f|^j_{n-1+\alpha} = \sum_{k=0}^{n-1} \frac{r^k}{k!} \sup H_\alpha[\partial^{k;j} f] \tag{6.5}$$

where the supremum runs over all suitable operators.

Lemma 6.2. *$H_\alpha[z]$, $|z|_n$, $|z|_{n+\alpha}$, $|f|^j_{n-1+\alpha}$ are indeed norms. And the normed spaces they induced are Banach algebras.*

Proof. It's obvious that they are indeed norms.
The completeness follows from that

$$\partial_j \sum z = \sum \partial_j z, \ \delta_j \sum z = \sum \delta_j z$$

which follows from the dominated convergence theorem.
The rest to prove that they are multiplicative.
By Leibniz rule,

$$\partial_j(fg) = \partial_j f \cdot g + f \cdot \partial_j g, \ \delta_j(fg) = \delta_j f \cdot g + f \cdot \delta_j g$$

$$\frac{\partial^k(fg)}{k!} = \sum_{l=0}^{k} \frac{\partial^l f}{l!} \cdot \frac{\partial^{k-l} g}{(k-l)!}, \ \frac{\delta^k(fg)}{k!} = \sum_{l=0}^{k} \frac{\delta^l f}{l!} \cdot \frac{\delta^{k-l} g}{(k-l)!}$$

It follows that

$$H_\alpha[fg] \le \sum_{m=0}^{\infty} r^{m\alpha} \sum_{k=0}^{m} \sup \frac{|\delta^k f|}{k!} \sup \frac{|\delta^{m-k} g|}{(m-k)!} = H_\alpha[f] H_\alpha[g]$$

$$|fg|_n \le \sum_{m=0}^{\infty} r^{m\alpha} \sum_{k=0}^{m} \sup \frac{|\partial^k f|}{k!} \sup \frac{|\partial^{m-k} g|}{(m-k)!} = |f|_n |g|_n$$

$$|fg|_{n+\alpha} \le \sum_{m=0}^{\infty} r^{m} \sum_{k=0}^{m} \sup H_\alpha \left[\frac{\partial^k f}{k!}\right] \sup H_\alpha \left[\frac{\partial^{m-k} f}{(m-k)!}\right] = |f|_{n+\alpha} |g|_{n+\alpha}$$

$$|fg|^j_{n-1+\alpha} \le \sum_{m=0}^{\infty} r^{m} \sum_{k=0}^{m} \sup H_\alpha \left[\frac{\partial^{k;j} f}{k!}\right] \sup H_\alpha \left[\frac{\partial^{m-k;j} f}{(m-k)!}\right] = |f|^j_{n-1+\alpha} |g|^j_{n-1+\alpha}$$

Therefore the normed space they induced are Banach algebras. □

Definition 6.3. For $Z = (z^1, \cdots, z^n)$, $F = (f_1, \cdots, f_n)$, define

$$|Z|_{n+\alpha} = \sup_{1\le j\le n} |z^j|_{n+\alpha},\ |F|_{n-1+\alpha} = \sup_{1\le j\le n} |f_j|^j_{n-1+\alpha} \tag{6.6}$$

They are indeed norms and induced Banach spaces.

Lemma 6.4.

$$|z|_{n+\alpha} \le \sum_{k,m} \frac{r^{k+m\alpha}}{k!m!} \sup|\delta^m\partial^k z| \tag{6.7}$$

$$r^{k+m\alpha}|\delta^m\partial^k z| \le C|z|_{n+\alpha} \tag{6.8}$$

$$|f|^j_{n-1+\alpha} \le \sum_{k,m} \frac{r^{k+m\alpha}}{k!m!} \sup|\delta^m\partial^{k;j} f| \tag{6.9}$$

$$r^{k+m\alpha} \sup|\delta^m\partial^{k;j} f| \le C|f|^j_{n-1+\alpha} \tag{6.10}$$

Proof. By definition,

$$|z|_{n+\alpha} = \sum_k \frac{r^k}{k!} \sup H_\alpha[\partial^k z] \le \sum_{k,m} \frac{r^{k+m\alpha}}{k!m!} \sup|\delta^m\partial^k z|$$

$$r^{k+m\alpha}|\delta^m\partial^k z| \le Cr^k H_\alpha[\partial^k z] \le C|z|_{n+\alpha}$$

The other two can be proved similarly □

Example 6.5.

$$|\zeta^j|_{n+\alpha} \le (2 + 2^{1-\alpha})r \tag{6.11}$$

Proof.

$$|\zeta^j| \le r,\ \left|\frac{\partial\zeta^j}{\partial\zeta^j}\right| = 1$$

$$|\delta_j\zeta^j| = |\zeta^j_1 - \zeta^j_2|^{1-\alpha} \le (2r)^{1-\alpha}$$

$$|\zeta^j|_{n+\alpha} \le \sum_{k,m} \frac{r^{k+m\alpha}}{k!m!} \sup|\delta^m\partial^k\zeta^j|$$

$$= \sup|\zeta^j| + r\sup\left|\frac{\partial\zeta^j}{\partial\zeta^j}\right| + r^\alpha \sup|\delta_j\zeta^j|$$

$$= (2 + 2^{1-\alpha})r$$

□

Lemma 6.6. $\forall j, l$,

$$|T^j(fg)|^l_{n-1+\alpha} \leq Cr|f|^j_{n-1+\alpha}|g|^l_{n-1+\alpha} \tag{6.12}$$

$$|T^j(fg)|_{n+\alpha} \leq Cr|f|^j_{n-1+\alpha}|g|_{n+\alpha} \tag{6.13}$$

Proof.

$$|T^j(fg)|^l_{n-1+\alpha} \leq \sum_{k,m} \frac{r^{k+m\alpha}}{k!m!} \sup |\delta^m \partial^k T^j(fg)|$$

Let Φ be a term of the following form

$$\Phi = r^{k+m\alpha}\delta^{m;j}\partial^{k;j,l}(fg)$$

We shall estimate terms of the following types:

$$T^j\Phi,\ r^\alpha\delta_j T^j\Phi,\ r\partial_j T^j\Phi,\ r^{1+\alpha}\delta_j\partial_j T^j\Phi$$

which are bounded in absolute value by

$$Cr\sup|\Phi| + Cr^{1+\alpha}\sup|\delta_j\Phi|$$

$$\begin{aligned} r|\Phi| &= r^{k+1+m\alpha}|\delta^{m;j}\partial^{k;j,l}(fg)| \leq r^{k+1+m\alpha}\sum_{s,t} C|\delta^{s;j}\partial^{t;j,l}f| \cdot |\delta^{m-s;j}\partial^{k-t;j,l}g| \\ &= r\sum_{s,t} Cr^{t+s\alpha}|\delta^s\partial^{t;j}f| \cdot r^{k-t+(m-s)\alpha}|\delta^{m-s}\partial^{k-t;l}g| \leq Cr|f|^j_{n-1+\alpha}|g|^l_{n-1+\alpha} \end{aligned}$$

$$\begin{aligned} r^{1+\alpha}|\delta_j\Phi| &= r^{k+1+(m+1)\alpha}|\delta_j\delta^{m;j}\partial^{k;j,l}(fg)| \leq r^{k+1+(m+1)\alpha}\sum_{s,t} C|\delta^s\partial^{t;j}f| \cdot |\delta^{m+1-s}\partial^{k-t;l}g| \\ &= r\sum_{s,t} Cr^{s+t\alpha}|\delta^s\partial^{t;j}f| \cdot |r^{k-t+(m+1-s)\alpha}\delta^{m+1-s}\partial^{k-t;l}g| \leq Cr|f|^j_{n-1+\alpha}|g|^l_{n-1+\alpha} \end{aligned}$$

The other one is proved similarly. □

Corollary 6.7. $\forall j, l$

$$|T^j f|^l_{n-1+\alpha} \leq Cr|f|^l_{n-1+\alpha} \tag{6.14}$$

$$|T^j f|_{n+\alpha} \leq Cr|f|^j_{n-1+\alpha} \tag{6.15}$$

Lemma 6.8.

$$|\partial_j f|^j_{n-1+\alpha} \leq \frac{C}{r}|f|_{n+\alpha} \tag{6.16}$$

Proof.

$$|\partial_j f|^j_{n-1+\alpha} \leq \sum_{k,m} \frac{r^{k+m\alpha}}{k!m!} \sup |\delta^m \partial^{k;j} \partial_j f| \leq \frac{C}{r} |f|_{n+\alpha}$$

□

Corollary 6.9.

$$|T^j \partial_j f|_{n+\alpha} \leq C|f|_{n+\alpha} \tag{6.17}$$

Theorem 6.10.

$$|TF|_{n+\alpha} \leq Cr|F|_{n-1+\alpha} \tag{6.18}$$

Proof. Recall

$$TF = \sum_{s=0}^{n-1} \frac{(-1)^s}{(s+1)!} \sum_{j_1,\cdots,j_s,k \text{ distinct}} T^{j_1} \frac{\partial}{\partial \overline{\zeta}^{j_1}} \cdots T^{j_s} \frac{\partial}{\partial \overline{\zeta}^{j_s}} T^k f_k$$

$$|T^{j_1}\partial_{j_1} \cdots T^{j_s}\partial_{j_s} T^k f_k|_{n+\alpha} \leq C|T^k f_k|_{n+\alpha} \leq Cr|f_k|^k_{n-1+\alpha} \leq Cr|F|_{n-1+\alpha}$$

□

Lemma 6.11.

$$H_\alpha[z] \leq C|z|_n \tag{6.19}$$

Proof. By Mean Value Theorem

$$|\delta_j z| \leq Cr^{1-\alpha} \sup |\partial_j z|$$

$$H_\alpha[z] = \sum_k \frac{r^{k\alpha}}{k!} \sup |\delta^k z| \leq C \sum_k \frac{r^k}{k!} \sup |\partial^k z| = C|z|_n$$

□

We now consider a function $a(Z) = a(z^1, \cdots, z^n)$, and its norm as a function of (ζ^j). Let $\partial^k_Z a$ denote an operator of the form: $\frac{\partial^k}{\partial z^{\mu_1} \cdots \partial z^{\mu_k}}$, ($\mu_1, \cdots, \mu_k$ may not be disjoint) Note that $|Z|_{n+\alpha} = \sup |z^j|_{n+\alpha} = \sup |z^\mu|_{n+\alpha}$

Lemma 6.12. *Assume that $a \in C^n$, $|z|_{n+\alpha} \leq 1$, $|\partial^k_Z a| \leq K$, $\forall 0 \leq k \leq n$, then*

$$H_\alpha[a] \leq C|a|_n \leq \begin{cases} CK \\ CK|Z|_{n+\alpha}, & \textit{if } a(0) = 0 \end{cases} \tag{6.20}$$

Proof.

$$|a|_n = \sum_j \frac{r^j}{j!} \sup |\partial^j a|$$

$\forall j > 0$, by Bruno's Formula, $\partial^j a$ is linear combination of the following terms:

$$\Phi = (\partial_Z^k a)(\partial^{j_1} z^{\mu_1}) \cdots (\partial^{j_k} z^{\mu_k}), \text{ where } \partial^{j_1} \cdots \partial^{j_k} = \partial^j$$

$$r^j |\Phi| \le K |z^{\mu_1}|_n \cdots |z^{\mu_k}|_n \le K |Z|_{n+\alpha}$$

For $j = 0$, $|a| \le K$. If $a(0) = 0$, by Mean Value Theorem,

$$|a(Z)| = |a(Z) - a(0)| = \left| \frac{\partial a}{\partial z^\mu} \cdot (z^\mu - 0) \right| \le CK |Z|_{n+\alpha}$$

□

Theorem 6.13. *Assume $|Z|_{n+\alpha} \le 1$, $a \in C^{2n}$, $|\partial_Z^k a| \le K$, $\forall 0 \le k \le 2n$, then*

$$|a(Z)|_{n+\alpha} \le \begin{cases} CK \\ CK|Z|_{n+\alpha}, \text{ if } a(0) = 0 \end{cases} \tag{6.21}$$

Proof.

$$|a(Z)|_{n+\alpha} = \sum_j \frac{r^j}{j!} H_\alpha[\partial^j a]$$

$\forall j > 0$, $\partial^j a$ is linear combination of

$$\Phi = (\partial_Z^k a)(\partial^{j_1} z^{\mu_1}) \cdots (\partial^{j_k} z^{\mu_k}), \text{ where } \partial^{j_1} \cdots \partial^{j_k} = \partial^j$$

$$r^j H_\alpha[\Phi] \le H_\alpha[\partial_Z^k a] r^{j_1} H_\alpha[\partial^{j_1} z^{\mu_1}] \cdots r^{j_k} H_\alpha[\partial^{j_k} z^{\mu_k}] \le K |Z|_{n+\alpha}$$

For $j = 0$,

$$H_\alpha[a] \le C|a|_n \le \begin{cases} CK \\ CK|Z|_{n+\alpha}, \text{ if } a(0) = 0 \end{cases}$$

□

The same method shows that

Theorem 6.14. *Assume $|Z|_{n+\alpha} \le 1$, $a \in C^{2n-1}$, $|\partial_Z^k a| \le K$, $\forall 0 \le k \le 2n-1$, then $\forall m$*

$$|a(Z)|_{n-1+\alpha}^m \le \begin{cases} CK \\ CK|Z|_{n-1+\alpha}^m, \text{ if } a(0) = 0 \end{cases} \tag{6.22}$$

Theorem 6.15. *Let $Z_1 = (z_1^j), Z_2 = (z_2^j)$, $|Z_1|_{n+\alpha}, |Z_2|_{n+\alpha} \leq 1$, $a \in C^{2n-1}$, $|\partial_Z^k a| \leq K$, $\forall 0 \leq k \leq 2n-1$, then:*

$$|a(Z_1) - a(Z_2)|_{n-1+\alpha}^m \leq CK|Z_1 - Z_2|_{n+\alpha} \tag{6.23}$$

Proof. By Mean Value Theorem

$$a(Z_1) - a(Z_2) = \frac{\partial a}{\partial z^j} \cdot (z_1^j - z_2^j) + \frac{\partial a}{\partial \overline{z}^j} \cdot (\overline{z}_1^j - \overline{z}_2^j)$$

$$|a(Z_1) - a(Z_2)|_{n-1+\alpha}^m \leq C|\partial_Z^1 a|_{n-1+\alpha}^m |Z_1 - Z_2|_{n+\alpha} \leq CK|Z_1 - Z_2|_{n+\alpha}$$

□

7 Proof of Main Theorem

We now solve (3.5)

$$\frac{\partial z^k}{\partial \overline{\zeta}^j} = -a_m^k \frac{\partial \overline{z}^m}{\partial \overline{\zeta}^j}$$

under integrability condition and suitable differentiability conditions.
Let $f_j^k(Z) = -a_m^k \frac{\partial \overline{z}^m}{\partial \overline{\zeta}^j}$, then (3.5) is $\frac{\partial z^k}{\partial \overline{\zeta}^j} = f_j^k(Z)$
Let $F^k = (f_1^k, \cdots, f_n^k)$, let $z_0^k(Z)$ be the value of $TF^k(Z)$ at $\zeta^1 = \cdots = \zeta^n = 0$
Consider the integral equation $Z = \Theta(Z)$:

$$z^k = \zeta^k + TF^k(Z) - z_0^k(Z) =: \Theta^k(Z) \tag{7.1}$$

Lemma 7.1. *A solution of (7.1) satisfies (3.5), for sufficiently small r*

Proof. Assume that (7.1) holds. Let $g_j^k = \frac{\partial z^k}{\partial \overline{\zeta}^j} + a_m^k \frac{\partial \overline{z}^m}{\partial \overline{\zeta}^j}$. To prove: $g_j^k = 0$

Apply $\frac{\partial}{\partial \overline{\zeta}^j}$ to (7.1), $\frac{\partial z^k}{\partial \overline{\zeta}^j} = \frac{\partial}{\partial \overline{\zeta}^j} TF^k(Z)$

$$\begin{aligned} g_j^k &= \frac{\partial z^k}{\partial \overline{\zeta}^j} + a_m^k \frac{\partial \overline{z}^m}{\partial \overline{\zeta}^j} = \frac{\partial}{\partial \overline{\zeta}^j} TF^k(Z) - f_j^k(Z) \\ &= \sum_{s=0}^{n-2} \frac{(-1)^s}{(s+2)!} \sum_{j_1, \cdots, j_s, l, j \text{ distinct}} T^{j_1} \frac{\partial}{\partial \overline{\zeta}^{j_1}} \cdots T^{j_s} \frac{\partial}{\partial \overline{\zeta}^{j_s}} T^l \left(\frac{\partial f_l^k}{\partial \overline{\zeta}^j} - \frac{\partial f_j^k}{\partial \overline{\zeta}^l} \right) \end{aligned}$$

$$\frac{\partial f_l^k}{\partial\overline{\zeta}^j}-\frac{\partial f_j^k}{\partial\overline{\zeta}^l}=\frac{\partial}{\partial\overline{\zeta}^j}\left(-a_m^k\frac{\partial\overline{z}^m}{\partial\overline{\zeta}^l}\right)-\frac{\partial}{\partial\overline{\zeta}^l}\left(-a_m^k\frac{\partial\overline{z}^m}{\partial\overline{\zeta}^j}\right)=\frac{\partial a_m^k}{\partial\overline{\zeta}^l}\cdot\frac{\partial\overline{z}^m}{\partial\overline{\zeta}^j}-\frac{\partial a_m^k}{\partial\overline{\zeta}^j}\cdot\frac{\partial\overline{z}^m}{\partial\overline{\zeta}^l}$$
$$=\frac{\partial a_m^k}{\partial z^p}\frac{\partial z^p}{\partial\overline{\zeta}^l}\frac{\partial\overline{z}^m}{\partial\overline{\zeta}^j}+\frac{\partial a_m^k}{\partial\overline{z}^s}\frac{\partial\overline{z}^s}{\partial\overline{\zeta}^l}\frac{\partial\overline{z}^m}{\partial\overline{\zeta}^j}-\frac{\partial a_m^k}{\partial z^p}\frac{\partial z^p}{\partial\overline{\zeta}^j}\frac{\partial\overline{z}^m}{\partial\overline{\zeta}^l}-\frac{\partial a_m^k}{\partial\overline{z}^s}\frac{\partial\overline{z}^s}{\partial\overline{\zeta}^j}\frac{\partial\overline{z}^m}{\partial\overline{\zeta}^l}$$
$$=\frac{\partial a_m^k}{\partial z^p}\left(g_l^p\frac{\partial\overline{z}^m}{\partial\overline{\zeta}^j}-g_j^p\frac{\partial\overline{z}^m}{\partial\overline{\zeta}^l}\right)+\frac{\partial\overline{z}^s}{\partial\overline{\zeta}^l}\frac{\partial\overline{z}^m}{\partial\overline{\zeta}^j}\left(\frac{\partial a_m^k}{\partial\overline{z}^s}-a_s^p\frac{\partial a_m^k}{\partial z^p}\right)-\frac{\partial\overline{z}^s}{\partial\overline{\zeta}^j}\frac{\partial\overline{z}^m}{\partial\overline{\zeta}^l}\left(\frac{\partial a_m^k}{\partial\overline{z}^s}-a_s^p\frac{\partial a_m^k}{\partial z^p}\right)$$
$$=\frac{\partial a_m^k}{\partial z^p}\left(g_l^p\frac{\partial\overline{z}^m}{\partial\overline{\zeta}^j}-g_j^p\frac{\partial\overline{z}^m}{\partial\overline{\zeta}^l}\right)+\frac{\partial\overline{z}^s}{\partial\overline{\zeta}^l}\frac{\partial\overline{z}^m}{\partial\overline{\zeta}^j}L_{m,s}^k-\frac{\partial\overline{z}^s}{\partial\overline{\zeta}^j}\frac{\partial\overline{z}^m}{\partial\overline{\zeta}^l}L_{m,s}^k=\frac{\partial a_m^k}{\partial z^p}\left(g_l^p\frac{\partial\overline{z}^m}{\partial\overline{\zeta}^j}-g_j^p\frac{\partial\overline{z}^m}{\partial\overline{\zeta}^l}\right)$$

Here used the integrability condition $L_{m,s}^k=L_{s,m}^k$ and commutes the index s,m.
We get a system of linear integral equations of g_j^k

$$g_j^k=\sum_{s=0}^{n-2}\frac{(-1)^s}{(s+2)!}\sum_{j_1,\cdots,j_s,l,j\text{ distinct}}T^{j_1}\frac{\partial}{\partial\overline{\zeta}^{j_1}}\cdots T^{j_s}\frac{\partial}{\partial\overline{\zeta}^{j_s}}T^l\left(\frac{\partial a_m^k}{\partial z^p}\left(g_l^p\frac{\partial\overline{z}^m}{\partial\overline{\zeta}^j}-g_j^p\frac{\partial\overline{z}^m}{\partial\overline{\zeta}^l}\right)\right)\tag{7.2}$$

It suffices to prove that (7.2) admits only the null solution for r sufficiently small.
Now estimate the norm:

$$\sum_{j,k}|g_j^k|_{n-1+\alpha}^j\leq\sum_{j,k,s}C\sum_{j_1,\cdots,j_s,l,j\text{ distinct}}\left|T^{j_1}\partial_{j_1}\cdots T^{j_s}\partial_{j_s}T^l\left(\frac{\partial a_m^k}{\partial z^p}\left(g_l^p\partial_j\overline{z}^m-g_j^p\partial_l\overline{z}^m\right)\right)\right|_{n-1+\alpha}^j$$
$$\leq C\sum_{j,k}\sum_{l\neq j}\left|T^l\left(\frac{\partial a_m^k}{\partial z^p}\left(g_l^p\partial_j\overline{z}^m-g_j^p\partial_l\overline{z}^m\right)\right)\right|_{n-1+\alpha}^j$$
$$\leq Cr\sum_{j,k}\sum_{l\neq j}\left(|g_l^p|_{n-1+\alpha}^l\left|\frac{\partial a_m^k}{\partial z^p}\partial_j\overline{z}^m\right|_{n-1+\alpha}^j+\left|\frac{\partial a_m^k}{\partial z^p}\partial_l\overline{z}^m\right|_{n-1+\alpha}^l|g_j^p|_{n-1+\alpha}^j\right)$$
$$\leq Cr\sum_{j,k}\sup\left|\frac{\partial a_m^k}{\partial z^p}\right|_{n-1+\alpha}^j\sup|\partial_j\overline{z}^m|_{n-1+\alpha}^j\sum_{p,l}|g_l^p|_{n-1+\alpha}^l\leq CKr\sum_{p,l}|g_l^p|_{n-1+\alpha}^l$$

Now for r sufficiently small, $CKr<1$, $\sum_{j,k}|g_j^k|_{n-1+\alpha}^j=0$, $g_j^k=0$ □

Lemma 7.2. *Let* $Z_1=(z_1^j)$, $Z_2=(z_2^j)$. *If* $|Z_1|_{n+\alpha},|Z_2|_{n+\alpha}\leq 4r<1$, *then*

$$|f_j^k(Z)|_{n-1+\alpha}^j\leq CKr\tag{7.3}$$
$$|f_j^k(Z_1)-f_j^k(Z_2)|_{n-1+\alpha}^j\leq CK|Z_1-Z_2|_{n+\alpha}\tag{7.4}$$

Proof.

$$|f_j^k|_{n-1+\alpha}^j=|a_m^k\partial_j\overline{z}^m|_{n-1+\alpha}^j\leq|a_m^k|_{n-1+\alpha}^j|\partial_j\overline{z}^m|_{n-1+\alpha}^j$$
$$\leq CK|Z|_{n+\alpha}\cdot\frac{C}{r}|\overline{z}^m|_{n+\alpha}\leq\frac{CK}{r}|Z|_{n+\alpha}^2\leq CKr$$

$$\begin{aligned}&|f_j^k(Z_1)-f_j^k(Z_2)|_{n-1+\alpha}^j=|a_m^k(Z_1)\partial_j\overline{z}_1^m-a_m^k(Z_2)\partial_j\overline{z}_2^m|_{n-1+\alpha}^j\\ \leq&|a_m^k(Z_1)-a_m^k(Z_2)|_{n-1+\alpha}^j|\partial_j\overline{z}_1^m|_{n-1+\alpha}^j+|a_m^k(Z_2)|_{n-1+\alpha}^j|\partial_j(\overline{z}_1^m-\overline{z}_2^m)|_{n-1+\alpha}^j\\ \leq&CK|Z_1-Z_2|_{n+\alpha}\frac{C}{r}|Z_1|_{n+\alpha}+CK|Z_2|_{n+\alpha}\frac{C}{r}|Z_1-Z_2|_{n+\alpha}\leq CK|Z_1-Z_2|_{n+\alpha}\end{aligned}$$

□

Theorem 7.3. *Equation (7.1) has a solution for r sufficiently small.*

Proof. $\forall|Z|_{n+\alpha}\leq 4r$

$$\begin{aligned}&|\Theta^k(Z)|_{n+\alpha}=|\zeta^k+TF^k(Z)-z_0^k(Z)|_{n+\alpha}\leq|\zeta^k|_{n+\alpha}+2|TF^k(Z)|_{n+\alpha}\\ \leq&(2+2^{1-\alpha})r+Cr|F^k(Z)|_{n-1+\alpha}=(2+2^{1-\alpha})r+Cr\sup|f_j^k(Z)|_{n-1+\alpha}^j\\ \leq&(2+2^{1-\alpha})r+CKr^2<4r\end{aligned}$$

For r sufficiently small.
$\forall|Z_1|_{n+\alpha},|Z_2|_{n+\alpha}\leq 4r$

$$\begin{aligned}&|\Theta^k(Z_1)-\Theta^k(Z_2)|_{n+\alpha}=|TF^k(Z_1)-TF^k(Z_2)+z_0^k(Z_1)-z_0^k(Z_2)|_{n+\alpha}\\ &\leq 2|TF^k(Z_1)-TF^k(Z_2)|_{n+\alpha}=2|T(F^k(Z_1)-F^k(Z_2))|_{n+\alpha}\\ &\leq Cr|F^k(Z_1)-F^k(Z_2)|_{n-1+\alpha}=Cr\sup|f_j^k(Z_1)-f_j^k(C_2)|_{n-1+\alpha}^j\\ &\leq CKr|Z_1-Z_2|_{n+\alpha}\leq\frac{1}{2}|Z_1-Z_2|_{n+\alpha}\end{aligned}$$

Now apply Banach Fixed-Point Theorem. □

Remark 7.4. The proof above only shows that there is a solution (ζ^j) which has mixed derivatives. We need some standard results about elliptic equations to ensure that ζ^j has better smoothness. We omit it here.

8 Modern Formulation

Let (M,J) be an almost complex manifold.
Recall: $TM_{\mathbb{C}}=TM\otimes_{\mathbb{R}}\mathbb{C}$.
The almost complex structure induced a decomposition

$$TM_{\mathbb{C}}=TM^{1,0}\oplus TM^{0,1},\ T^*M_{\mathbb{C}}=T^*M^{1,0}\oplus T^*M^{0,1}\tag{8.1}$$

where $TM^{1,0}$ is the bundle of i-eigenspaces, $TM^{0,1}$ is that of $(-i)$-eigenspaces.
This leads to a decomposition of the bundle of exterior algebras

$$\wedge^nT^*M=\bigoplus_{p+q=n}\wedge^{p,q}T^*M,\ \wedge T^*M_{\mathbb{C}}=\bigoplus_n\wedge^nT^*M=\bigoplus_{p,q}\wedge^{p,q}T^*M\tag{8.2}$$

where $\wedge^{p,q}T^*M = (\wedge^p T^*M^{1,0}) \wedge (\wedge^q T^*M^{0,1})$
Let $\Omega^{p,q}M = \Gamma(M, \wedge^{p,q}T^*M)$ be the collection of (p,q)-forms. Then

$$\Omega^n M = \Gamma(M, \wedge^n T^*M) = \bigoplus_{p+q=n} \Omega^{p,q}M \tag{8.3}$$

Recall the exterior differential operator

$$d : \Omega^n M = \bigoplus_{p+q=n} \Omega^{p,q}M \to \Omega^{n+1}M = \bigoplus_{p'+q'=n+1} \Omega^{p',q'}M \tag{8.4}$$

$\Omega^{p,q}M$ is sent to $\bigoplus_{p'+q'=n+1} \Omega^{p',q'}M$

Let's first look at the case when J is induced by a complex structure.
Let (z^j) be a complex analytic coordinate system.
Then $\mathrm{d}z^j$ forms a local base of $T^*M^{1,0}$, $\mathrm{d}\overline{z}^j$ forms that of $T^*M^{0,1}$
Now a form $\omega \in \Omega^{p,q}$ has the following local representation

$$\omega = \sum_{I,J:|I|=p,|J|=q} \omega_{I,J}\mathrm{d}z^I \wedge \mathrm{d}\overline{z}^J \tag{8.5}$$

we used the multi-index notation: $\mathrm{d}z^I = \mathrm{d}z^{i_1} \wedge \cdots \mathrm{d}z^{i^k}$ when $I = \{i_1, \cdots, i_k\}$

$$\mathrm{d}\omega = \sum_{I,J:|I|=p,|J|=q} \mathrm{d}\omega_{I,J} \wedge \mathrm{d}z^I \wedge \mathrm{d}\overline{z}^J \tag{8.6}$$

$$= \sum_{I,J,j} \frac{\partial \omega_{I,J}}{\partial z^j}\mathrm{d}z^j \wedge \mathrm{d}z^I \wedge \mathrm{d}\overline{z}^J + \sum_{I,J,j}(-1)^p \frac{\partial \omega_{I,J}}{\partial \overline{z}^j}\mathrm{d}z^I \wedge \mathrm{d}\overline{z}^j \wedge \mathrm{d}\overline{z}^J \tag{8.7}$$

Thus $\mathrm{d}\omega \in \Omega^{p+1,q}M \oplus \Omega^{p,q+1}M$
This leads to the modern formulation of integrability condition.

Theorem 8.1. *Let (M, J) be an almost complex manifold. TFAE:*
(1) J is induced by a complex manifold.
(2) For $u^l = \mathrm{d}z^l + a^l_k \mathrm{d}\overline{z}^k$
the exterior differential of u^l is a sum of exterior products of 1-forms with u^k.
(3) the exterior differential d sends $\Omega^{p,q}M$ to $\Omega^{p+1,q}M \oplus \Omega^{p,q+1}M$

Proof. We have showed that (2) $\iff$ (1) $\implies$ (3).
Now assume that $d : \Omega^{1,0}M \to \Omega^{2,0}M \oplus \Omega^{1,1}M = \Omega^{1,0}M \wedge \Omega^1 M$
Claim: $\Omega^{1,0}M = \mathrm{span}\{u^l\}$
Firstly, we prove that $u^l \in \Omega^{1,0}M$:

$$Ju^l = J\mathrm{d}z^l + a^l_k J\mathrm{d}\overline{z}^k = J^l_j \mathrm{d}z^j + J^l_{j+n}\mathrm{d}\overline{z}^j + a^l_k J^{k+n}_j \mathrm{d}z^j + a^l_k J^{k+n}_{j+n}\mathrm{d}\overline{z}^j$$
$$iu^l = i\mathrm{d}z^l + ia^l_k \mathrm{d}\overline{z}^k = i\delta^l_j \mathrm{d}z^j + ia^l_k \delta^k_j \mathrm{d}\overline{z}^j$$

By comparing coefficients, it suffices to prove that:

$$a_k^l(i\delta_j^k - J_{j+n}^{k+n}) = -(-J_{j+n}^l) \tag{8.8}$$

$$a_k^l(-J_j^{k+n}) = -(i\delta_j^l - J_j^l) \tag{8.9}$$

Note that (8.8) is the definition formula of a_k^l,
(8.9) is obtained from (8.8) by a linear transformation.
Now $u^l \in \Omega^{1,0}M$, and obviously u^l are linearly independent,
note that $\dim \Omega^{1,0}M = n$, therefore $\Omega^{1,0}M = \mathrm{span}\{u^l\}$
Thus $\mathrm{d} : \mathrm{span}\{u^l\} \to \Omega^1 M \wedge \mathrm{span}\{u^l\}$, which is exactly what we need. □

References

[1] Shing-Shen Chern. An elementary proof of the existence of isothermal parameters on a surface. *Proceedings of The American Mathematical Society - PROC AMER MATH SOC*, 6, 05 1955.

[2] A. Newlander and L. Nirenberg. Complex analytic coordinates in almost complex manifolds. *Annals of Mathematics. Second Series*, 65, 05 1957.

[3] Raymond Wells. Differential analysis on complex manifolds. 65, 01 1973.

the Unique Conformal Mating between the Ideal Triangle Group and the Anti-polynomial

Qiandu He

June 2022

1 Introduction

In this final project, we will going to introduce how to produce a conformal mating between a map (denoted by ρ) constructed from the ideal triangle group and the anti-polynomials (a shorthand of anti-holomorphic polynomials) $\overline{z}^2$ by using Schwarz reflection map (denoted by σ) of the deltoid. The most of following results and proofs are taken from [1]. In particular, it can be expressed as the following theorem (theorem 1.1 in [1]):

Theorem 1.1. (Dynamics of deltoid reflection).*1) The dynamical plane of Schwarz reflection σ of the deltoid can be partitioned as*

$$\hat{C} = T^{\infty} \sqcup \Gamma \sqcup A(\infty),$$

where T^{∞} is the tiling set, $A(\infty)$ is the basin of infinity, and Γ is their common boundary. Moreover, Γ is a conformally removable Jordan curve.

2) σ is the unique conformal mating of the reflection map $\rho : \overline{\mathbb{D}} \setminus \text{int}\Pi \to \overline{\mathbb{D}}$ and the anti-polynomial $f_0 : \hat{\mathbb{C}} \setminus \mathbb{D} \to \hat{\mathbb{C}} \setminus \mathbb{D}, z \mapsto \overline{z}^2$.

It's hard to understand the mating in above theorem at first glance. Therefore we will describe the basic dynamical objects associated with iteration of Schwarz reflection maps and the meaning of mating. Given a disjoint collection of quadrature domains, we call the complement of their union a droplet. Removing the finitely many singular points from the boundary of a droplet yields the fundamental tile. One can then look at a partially defined anti-holomorphic dynamical system σ that acts on the closure of each quadrature domain as its Schwarz reflection map. Consider the reflection dynamics σ defined on the closure of each quadrature domain, $\hat{C}$ admits a dynamically invariant partition. The first one is an open set called the escaping/tiling set, it is the set of all points that eventually escape to the fundamental tile (for z_0 here, the dynamics of σ can't be iterated forever). The second invariant set is the non-escaping set, namely, the set of all points on which σ can be iterated forever. The last one is their common boundary, where the most chaotic and complex phenomenon happens. On mating, note that the non-escaping set is analogous to the filled Julia set in polynomial dynamics; i.e., the set of points with bounded forward orbits under a polynomial when the tiling set contains no critical points of σ. While the $\sigma-$action on the tiling set exhibits features of reflection groups. As for their common boundary, which is simultaneously analogous to the Julia set of an anti-polynomial (i.e., the boundary of the filled Julia set) and to the limit set of a group.

At the end of these section, let us now detail the organization of this paper. In Section 2, we give a description of the ideal triangle group Π, the associated tessellation of the unit disk, and the reflection map ρ. Here we also define the topological conjugacy $\mathcal{E}$ between ρ and the anti-doubling $\theta \mapsto -2\theta$ on the circle $\mathbb{T}$. In Section 3, we briefly review some general notions and properties of quadrature domains and Schwarz reflection maps. Also, in this section, there are some useful tools used in the proof below . Section 4 is devoted to the study of the dynamics of Schwarz reflection with respect to the deltoid. The principal goal of this section is to present the processing of the proof in [1], but for some difficult parts, we will omit the concrete processing and give some introduction.

2 Ideal Triangle Group

Consider the open unit disk $\mathbb{D}$ in $\mathbb{C}$. Let $\widetilde{C}_1, \widetilde{C}_2, \widetilde{C}_3$ be the hyperbolic geodesics in $\mathbb{D}$ connecting 1 and ω, ω and ω^2, 1 and ω^2 respectively. ($\omega = e^{\frac{2\pi i}{3}}$). The closed ideal triangle bound by these geodesics is called Π below. For each $i \in \{1,2,3\}$, we know that $\widetilde{C}_i$ is an arc of a circle, the reflection of the circle restricted in $\mathbb{D}$ is called ρ_i, thus these three maps generated a subgroup $\mathcal{G} \leqslant Aut(\mathbb{D})$, which we call the ideal triangle group. As for the rest of the disc, we will denote the connected component of $D \setminus \Pi$ containing $\mathrm{int}\rho_i(\Pi)$ by $\mathbb{D}_i$, i.e., $\mathbb{D}_1 \cup \mathbb{D}_2 \cup \mathbb{D}_3 = \mathbb{D} \setminus \Pi$.

We can define $\rho : \mathbb{D} \setminus \mathrm{int}\Pi \to \mathbb{D} : \mathrm{z} \mapsto \rho_{\mathrm{i}}(\mathrm{z})$ if $z \in \mathbb{D}_i \cup \widetilde{C}_i$, for $i = 1,2,3$. Moreover, we can induct a symbolic dynamics through the map ρ as follow: Let $W := \{1,2,3\}$, an element $(i_1, i_2, ...) \in W^{\mathbb{N}}$ is called M-admissible if $M_{i_k, i_{k+1}} = 1$, for all $k \in \mathbb{N}$, and we will denote the set of all M-admissible words in $W^{\mathbb{N}}$ by M^{∞} (M is the 3×3 matrix whose diagonal elements are all zero and other elements are all one). Similarly, we can define the M-admissibility of finite words. Since $\rho(\mathbb{D}_i) \subseteq \mathbb{D}_{i+1} \cup \mathbb{D}_{i+2}, \rho(\mathbb{D}_i) \cap \mathbb{D}_i = \varnothing$, for $i = 1,2,3$ (all subscripts is defined in module 3), the dynamics of ρ is similar to the symbolic dynamics of M^{∞}.

2.1 Tessellation of the Disc

Note that Π is a fundamental domain of $\mathcal{G}$. We can give the tessellation of $\mathbb{D}$ arising from $\mathcal{G}$ to describe the dynamics of ρ.

Definition 2.1. (Tiles). *The images of the fundamental domain* Π *under the elements of* $\mathcal{G}$ *are called tiles. More precisely, for any* $M-$*admissible word*

$(i_1, ..., i_k)$*, we define the tile*

$$T^{i_1,...,i_k} := \rho_{i_1} \circ ... \circ \rho_{i_k}(\Pi).$$

We can write $T^{i_1,...,i_k}$ by ρ in another way. In fact, we have $\rho_{i_k} \circ ... \circ \rho_{i_1}(T^{i_1,...,i_k}) = \Pi$ because $\mathcal{G}$ is a reflection group. Thus $T^{i_1,...,i_k}$ consists of all those $z \in \mathbb{D}$ such that $\rho^{\circ(n-1)}(z) \in \mathbb{D}_{i_n}, \forall 1 \leq n \leq k$ (to make sure the correctly definition of iteration of ρ) and $\rho^{\circ k}(z) \in \Pi$. In other words,

$$T^{i_1,...,i_k} = \bigcap_{n=1}^{k} \rho^{-(n-1)}(\mathbb{D}_{i_n}) \cap \rho^{-k}(\Pi).$$

For a $M-$admissible word $(i_1, i_2, ...)$, let us consider the sequence $\{0, \rho_{i_1}(0), \rho_{i_1} \circ \rho_{i_2}(0), ...\}$. Since $d_{\mathbb{D}}(0, \rho_1(0)) = d_{\mathbb{D}}(0, \rho_2(0)) = d_{\mathbb{D}}(0, \rho_3(0))$ and the element in $\mathcal{G}$ keep the hyperbolic distance, thus any two consecutive points in this points in this sequence is constant. Connecting consecutive points of this sequence by hyperbolic geodesics of $\mathbb{D}$, we obtain a curve corresponding to $(i_1, i_2, ...)$. And all these curves form a dual tree to the $\mathcal{G}-$tessellation of $\mathbb{D}$.

2.2 $\rho-$action on the circle $\mathbb{T}$

First, we extend ρ as an orientation-reversing C^1 double covering. As the division of $\mathbb{D}_1, \mathbb{D}_2, \mathbb{D}_3$, we have $\mathbb{T} = (\partial\mathbb{D}_1 \cap \mathbb{T}) \cup (\partial\mathbb{D}_2 \cap \mathbb{T}) \cup (\partial\mathbb{D}_3 \cap \mathbb{T})$ with the transform matrix M defined above. Also, one can see $\rho|_{\mathbb{T}}$ is an expansive map, therefore, for any element of M^∞, the corresponding infinite sequence of tiles shrinks to a single point of $\mathbb{T}$ (i.e., the curve corresponding to $(i_1, i_2, ...)$ lands at $Q(i_1, i_2, ...)$), which allows us to define a continuous surjection

$$Q : M^\infty \to \mathbb{T},\ (i_1, i_2, ...) \mapsto \bigcap_{n \in \mathbb{N}} \rho^{-(n-1)}(\partial\mathbb{D}_{i_n} \cap \mathbb{T})$$

which semi-conjugates the (left-)shift map on M^∞ (denoted as L) to the map ρ on $\mathbb{T}$.

Definition 2.2. (Semi-conjugation). *Let M, N be topological spaces and f, g be self-homeomorphisms of M, N respectively. f is called semi-conjugated to g, if there is a continuous surjection $h : M \to N$ such that $h \circ f = g \circ h$.*

$$\begin{tikzcd}
M^\infty \arrow[r, "Q"] \arrow[d, "L"] & \mathbb{T} \arrow[d, "\rho|_{\mathbb{T}}"] \\
M^\infty \arrow[r, "Q"] & \mathbb{T}
\end{tikzcd}$$

We would add some explanation for the semi-conjugation and some property of Q here because there are some similar conclusion of the mapping constructed below with some similar proofs.

Actually, one can see the continuity of Q from the image of the landing of curve constructed above. And for any point $x \in \mathbb{T}$, we can select $(i_1, i_2, ...)$

properly such that $\rho^{\circ(n-1)}(x) \in \partial\mathbb{D}_{i_n}, \forall n \in \mathbb{N}$ (if there is one more i_n such that $\rho^{\circ(n-1)}(x)$ landed on $\partial\mathbb{D}_{i_n}$, then we select the smallest one.)

To prove the semiconjugation, it suffices to prove

$$\rho(\bigcap_{n\in\mathbb{N}} \rho^{-(n-1)}(\partial\mathbb{D}_{i_n} \cap \mathbb{T})) = \bigcap_{n\in\mathbb{N}} \rho^{-(n-1)}(\partial\mathbb{D}_{i_{n+1}} \cap \mathbb{T})).$$

Note that $\rho^{\circ(n-1)}(x) \in \partial\mathbb{D}_{i_n} \cap \mathbb{T} \Rightarrow \rho^{\circ(n-2)}(\rho(x)) \in \partial\mathbb{D}_{i_n} \cap \mathbb{T}, \forall n \geq 2$, thus the conclusion is corrected.

2.3 The conjugacy $\mathcal{E}$

We would induct the other expanding double covering of the circle, which is $\overline{z}^2|_{\mathbb{T}}$ (the action of $\overline{z}^2$ on angles of external dynamical rays) as follow:

$$m_{-2} : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z},\ \theta \mapsto -2\theta$$

As the analysis above, The map m_{-2} admits the same parition $\mathbb{R}/\mathbb{Z} = [0, \frac{1}{3}] \cup [\frac{1}{3}, \frac{2}{3}] \cup [\frac{2}{3}, 1]$ with the same transform matrix M. Similarly, we can define a continuous surjection

$$P : M^{\infty} \to \mathbb{R}/\mathbb{Z},\ (i_1, i_2, ...) \mapsto \bigcap_{n\in\mathbb{N}} m_{-2}^{-(n-1)}(I_{i_n})$$

which semi-conjugates the (left-)shift map on M^{∞} to the map m_{-2} on $\mathbb{R}/\mathbb{Z}$ by the similar proof. Thus we get the commutative diagram below:

$$\begin{tikzcd}
\mathbb{R}/\mathbb{Z} \arrow[d, "m_{-2}"] & M^{\infty} \arrow[l, "P"'] \arrow[r, "Q"] \arrow[d, "L"] & \mathbb{T} \arrow[d, "\rho|_{\mathbb{T}}"] \\
\mathbb{R}/\mathbb{Z} & M^{\infty} \arrow[l, "P"'] \arrow[r, "Q"] & \mathbb{T}
\end{tikzcd}$$

and we can obtain the conjugation between $\rho_{\mathbb{T}}$ and m_{-2} by $\mathcal{E} := P \circ Q^{-1}$. In fact, to ensure the mapping is well-defined, we need to pay attention to, for $x \in \mathbb{T}$, the time when the preimage of Q is branched. By the definition of Q, one more selection of i_n implies $\rho^{\circ(n-1)}(x) \in \{1, w, w^2\}$. Because $1, w, w^2$ are all fixed points of ρ, the sequence $\{\rho^{\circ(m-1)}(x) : m \geq n\}$ will be the same point. One can see that the same property happens in the iteration of m_{-2}, so we can well-define the mapping $\mathcal{E}$ on the such point $x \in \mathbb{T}$. For the same reason, the mapping $\mathcal{E}' := Q \circ P^{-1}$ is well-defined and is the inverse of $\mathcal{E}$, therefore, we find the conjugacy between ρ and $\overline{z}^2$ on the circle $\mathbb{T}$.

3 Quadrature Domain, Schwarz Reflection and Some Useful Tools

3.1 Quadrature Domain and Schwarz Reflection

First, we denote the complex conjugation map by ι. We will give the definition of the quadrature domain below. All these notions, properties and proofs is cited from the Section 1 of [2].

For bounded case, a bounded connected open set $\Omega \subseteq \hat{\mathbb{C}}$ is a bounded quadrature domain if it carries a finite node quadrature identity, which means there exists a finite collection of triples (a_k, m_k, c_k), where a_k is points (not necessarily distinct) in Ω, m_k is non-negative integers, and c_k is some complex numbers, such that

$$\forall f \in C(\overline{\Omega}) \cap H(\Omega), \int_\Omega f dA = \sum_k c_k f^{(m_k)}(a_k). \tag{1}$$

Here, $H(\Omega)$ denotes the space of analytic functions in Ω, $C(\overline{\Omega})$ denotes the space of continuous functions in $\overline{\Omega}$. dA is the area measure. Therefore we will always assume $\Omega = \text{int}\overline{\Omega}$. Look back to (1), we can rewrite it by using the contour integral in the right hand side of the quadrature identity,

$$\forall f \in C(\overline{\Omega}) \cap H(\Omega), \int_\Omega f dA = \frac{1}{2i} \oint_{\partial\Omega} f(z) r(z) dz, \tag{2}$$

where,

$$r(z) \equiv r_\Omega(z) = \frac{1}{\pi} \sum_k c_k \frac{m_k!}{(z - a_k)^{m_k+1}}$$

We will call r_Ω the quadrature function and $deg(r_\Omega)$ the order of the quadrature domain Ω. What's more, we see that quadrature function is uniquely determined by the quadrature domain as long as we require that all poles of r be inside Ω and $r(\infty) = 0$.

For unbounded case, we will modify slightly the statement above. Let $\Omega \subseteq \hat{\mathbb{C}}$ such that $\infty \in \Omega$ and int $\overline{\Omega} = \Omega$ (also, we don't want to discuss the case of $\Omega = \hat{\mathbb{C}}$). The unbounded open set Ω is an unbounded quadrature domain if there exists a rational function $r = r_\Omega$ with no poles outside Ω such that

$$f \in C(\overline{\Omega}) \cap H(\Omega), f(\infty) = 0 \Rightarrow \int_\Omega f dA = \frac{1}{2i} \oint_{\partial\Omega} f(z) r(z) dz. \tag{3}$$

The integrals over unbounded domains are always understood in the sense of principal value,

$$\int_\Omega \equiv v.p. \int_\Omega := \lim_{R\to\infty} \int_{\Omega \cap B(0,R)}.$$

and the notions are all same as the bounded case.

To sum up, we have the definition for quadrature function below. In the rest of this subsection, let $\Omega \subseteq \hat{\mathbb{C}}$ (not Riemann sphere itself) be a domain such that $\infty \notin \partial\Omega$ and int $\overline{\Omega} = \Omega$

Definition 3.1. (Quadrature functions). *If there is a rational map r_Ω whose all poles are inside Ω (with $r_\Omega(\infty) = 0$ if Ω is bounded) such that the identity*

$$\int_\Omega f dA = \frac{1}{2i} \oint_{\partial\Omega} f(z) r_\Omega(z) dz$$

holds for every functions $f \in H(\Omega) \cap C(\overline{\Omega})$ (if Ω is unbounded, we further require $f(\infty) = 0$), then we call Ω a quadrature domain and r_Ω the quadrature function of Ω.

Also, we can define the quadrature domain by Schwarz funtions.

Definition 3.2. (Schwarz functions and quadrature domains). *For a domain Ω A Schwarz funtion of Ω is a meromorphic extension of $\iota|_{\partial\Omega}$ to all of Ω. More precisely, a continuous function $S : \overline{\Omega} \to \hat{\mathbb{C}}$ of Ω is called a Schwarz function of Ω if it satisfies the following two properties:*

(1) *S is meromorphic on Ω,*

(2) *$S = \iota$ on $\partial\Omega$.*

The domain Ω is called a quadrature domain if it admits a Schwarz function.

To be further, the map $\sigma := \iota \circ S : \overline{\Omega} \to \hat{\mathbb{C}}$ is the unique anti-meromorphic extension of the Schwarz reflection map with respect to $\partial\Omega$. We will call σ the Schwarz reflection map of Ω.

The following theorem implies these two definition is equivalent. The proof here is cited from the Lemma 3.1. of [2].

Theorem 3.1. (Characterization of quadrature domains). *The following are equivalent.*

(1) *Ω admits a Schwarz funtion,*

(2) *Ω admits a quadrature function r_Ω.*

Proof. Assume Ω has a Schwarz function S. Since S has finitely many poles in Ω, is continuous up to boundary, and of course finite on $\partial\Omega$. We can construct a rational function r which has exactly the same poles and the same principal parts at the poles as S by add the principal part at all poles up (if S has a pole at ∞, then we add the $p(\frac{1}{z})$ where $p(z)$ is the principal part of $f(\frac{1}{z})$ at 0.), thus when Ω is bounded, we select $r(\infty)$ such that $r(\infty) = 0$; when $\infty \in \Omega$, we select $r(\infty)$ such that $\lim_{z\to\infty}(S(z) - r(z)) = 0$.

Here we will induct the Cauchy transform as a tool in this proof. Actually, for a Borel set $E \subseteq \mathbb{C}$ with a compact boundary, we denote by C^E the Cauchy transform of the area measure of E,

$$C^E(z) = \frac{1}{\pi}\int_E k_z(w)dA(w), \quad k_z(w) := \frac{1}{z-w}.$$

We will calculate $C^{\hat{\mathbb{C}}}$ as an example, which is useful in the following proof as well.

Indeed, if $R > |z|$, let $\epsilon > 0$ be small sufficiently such that the disc $|w - z| \leq \epsilon$ is contained in the disc $|w| < R$. We will divide the integral into two parts: $\int_{|w|\leq R} = \int_{|w|\leq R, |w-z|\geq\epsilon} + \int_{|w-z|\leq\epsilon}$.

Note that $dw \wedge d\overline{w} = (dx + idy) \wedge (dx - idy) = -2idx \wedge dy = -2idA$, for the first part, we have

$$\frac{1}{2\pi i}\int_{|w|\leq R, |w-z|\geq\epsilon} \frac{d\overline{w}dw}{z-w} = \frac{1}{2\pi i}(\oint_{|w|=R} - \oint_{|w-z|=\epsilon})\frac{\overline{w}}{z-w}dw,$$

While

$$\oint_{|w|=R} \frac{\overline{w}}{z-w} dw = \oint_{|w|=R} \frac{R^2}{w(z-w)} = 0,$$

$$\begin{aligned}\oint_{|w-z|=\epsilon} \frac{\overline{w}}{z-w} dw &= \oint_{|w-z|=\epsilon} (\frac{\overline{z}}{z-w} - \frac{\overline{z}-\overline{w}}{z-w}) dw \\ &= \oint_{|w-z|=\epsilon} (-\frac{\epsilon^2}{(z-w)^2} + \frac{\overline{z}}{z-w}) dw \\ &= -2\pi i \overline{z}\end{aligned}$$

Therefore, the first integral is equal to $\overline{z}$.

Note that $dA = dxdy = rdrd\theta$, for the second part, we have

$$\frac{1}{\pi} \int_{|w-z|\leq\epsilon} \frac{dxdy}{z-w} = -\frac{1}{\pi} \int_0^{\epsilon} \int_0^{2\pi} \frac{rdrd\theta}{re^{i\theta}} = -\frac{\epsilon}{\pi} \int_0^{2\pi} e^{-i\theta} d\theta = 0.$$

To sum up, we get $C^{\hat{\mathbb{C}}}(z) = \overline{z}$.

We will discuss the unbounded case first. For each $z \in \Omega$ we have

$$\begin{aligned}C^{\Omega^c}(z) &= \frac{1}{\pi} \int_{\Omega^c} \frac{dA(w)}{z-w} \\ &= \frac{1}{2\pi i} \int_{\Omega^c} \frac{dwd\overline{w}}{w-z} \\ &= \frac{1}{2\pi i} \oint_{\partial\Omega} \frac{\overline{w}}{w-z} dw \ (follow\ from\ the\ Green's\ formula) \\ &= \frac{1}{2\pi i} \oint_{\partial\Omega} \frac{S(w)-r(w)}{w-z} dw + \frac{1}{2\pi i} \oint_{\partial\Omega} \frac{r(w)}{w-z} dw.\end{aligned}$$

Note that we used the fact that the boundary of Ω is rectifiable and the assumption that $S(w) - r(w)|_{w=\infty} = 0$ to make sure the first integral above is well-defined. By residue formula, the first integral is equal to $S(z) - r(z)$. Since all the poles of $r(w)$ are inside Ω and $z \in \Omega$, by applying Cauchy's theorem in each component of the interior of Ω^c. Thus we have $S = r + C^{\Omega^c}, \forall z \in \Omega$. Moreover, these three maps are continuous in $\partial\Omega$, so we can expand this identity to the close of Ω.

We can now examine r is the quadrature function of Ω. For all $f \in H(\Omega) \cap C(\overline{\Omega})$ satisfying $f(\infty) = 0$, we have

$$\begin{aligned}\int_{\Omega} f dA &= \frac{1}{2i} \oint_{\partial\Omega} \overline{z} f(z) dz (the\ same\ reason\ as\ above) \\ &= \frac{1}{2i} \oint_{\partial\Omega} S(z) f(z) dz \\ &= \frac{1}{2i} \oint_{\partial\Omega} r(z) f(z) dz + \frac{1}{2i} \oint_{\partial\Omega} C^{\Omega^c}(z) f(z) dz.\end{aligned}$$

Since f and C^{Ω^c} is holomorphic in Ω and all achieve zero at ∞, by using Cauchy's theorem in Ω, the first integral is zero. It follows that Ω admits r as its quadrature function.

Suppose Ω admits quadrature funtion r satisfying the property introduced in the Definition 3.1., especially for the case of $f = \frac{1}{z-w}$ with $z \in \text{int}\Omega^c$. Then

$$C^{\Omega}(z) = \frac{1}{\pi}\int_{\Omega}\frac{dA}{w-z} = \frac{1}{2\pi i}\oint_{\partial\Omega}\frac{r(w)}{z-w} = r(z)$$

Thus for $z \in \partial\Omega$, we have $r(z) + C^{\Omega^c}(z) = C^{\hat{\mathbb{C}}}(z) = \bar{z}$ which means that $S := r + C^{\Omega^c}$ is the Schwarz funtion of Ω.

The bounded case have the same result by similar proof. Note that for bounded case, we can omit the examination of the property of $C^{\Omega}(z) = r(z)$ for $z = \infty$. Since $\infty \notin \partial\Omega$, the property of $S(z)$ in $\partial\Omega$ is still right. □

Moreover, quadrature domains have some other interesting properties, the real analyticity of their boundary, a sufficient and necessary condition for simply connected quadrature domain for instance. (i.e. Theorem 3.4. and Theorem 3.5. in [1]) Here, we won't add more proof of these theorem, because they won't be used in the following proof.

3.2 John Domain

In Section 4.3. of [1], this tool is used to prove conformal removability of the limit set implying uniqueness in the mating theory for theorem 3.2.

Definition 3.3. (Conformal Removability). A compact set $E \subseteq \hat{\mathbb{C}}$ is conformally removable if for all homeomorphisms $f : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$, if f is conformal on $\hat{\mathbb{C}} \setminus E$, then f is conformal on $\hat{\mathbb{C}}$, i.e. is a *Möbius* transformation.

Definition 3.4. (John domain). *A domain $D \subseteq \mathbb{C}$ is called a John domain if there exists $c > 0$ and reference points w_0 such that for any $z_0 \in D$, there exists an arc $\gamma \in D$ joining z_0 to w_0 satisfying*

$$\delta(z) \geq c|z - z_0|, z \in \gamma, \tag{4}$$

where $\delta(z)$ stands for the Euclidean distance between z and ∂D.

Theorem 3.2. *Suppose D is a John domain, then ∂D is conformally removable.*

Here, we can change (5) to another property because of the simple connectedness of T^{∞} and the theorem below.

Theorem 3.3. *For $z \in D$, we denote by γ^z the part of the hyperbolic geodesic of D passing through z and a fixed base-point w_0 that runs from z and lands on ∂D. A simply connected domain D is John domain if and only if there exists $M > 0$ such that for all $z \in D$,*

$$w \in \gamma^z, d_D(z, w) \geq M \Rightarrow \delta(w) \leq \frac{1}{2}\delta(z), \tag{5}$$

where d_D is the hyperbolic distance in D.

The main work in 4.3. is the proof that T^∞ satisfies (5) for some fixed base-point w_0.

3.3 Riemann-Hurwitz Formula

We will use this formula when these two complex manifold are all Riemann surface $\hat{\mathbb{C}}$ to prove the simply-connectedness of A and give a proof of the formula in this section. The following proof is cited from [3].

Theorem 3.4. (Riemann-Hurwitz formula). *Let V and W be domains on $\hat{\mathbb{C}}$ of finite connectivity m and n, respectively, and let $f : V \to W$ be a k-sheeted (ramified) analytic proper map having r critical points (counted by multiplicity). Then*

$$m - 2 = k(n - 2) + r. \tag{6}$$

Here the proper map means that preimages of compact subsets of W are compact. Then f assumes every value exactly k-times for the k-sheets of f.

Proof. The proof of (6) is based on the following lemma.

Lemma 3.5. *Let V be a domain of connectivity m, which is divided by k cross-cuts $c_1, ..., c_k$ (disjoint in V) into l domains $V_1, ..., V_l$, of connectivity $m_1, ..., m_l$, respectively. Then*

$$\sum_{j=1}^{l}(m_j - 2) = m - 2 - k.$$

Here, a cross-cut is a Jordan curve lying in V except for its end points, (the points) which belong to ∂V. We can divide the domain into two domains with lower connectivity.

Proof of lemma 3.5.: We proceed by induction. In the case of $k = 1$, we have either $V \setminus c$ is a domain of connectivity $m - 1$ or consists of two domains V^* and V^{**} of connectivity m^* and m^{**}, respectively, such that $m^* + m^{**} = m + 1$.

For $k > 1$, we first assume that $V \setminus c_1$ is not a domain, which consists of domains V^* and V^{**} of connectivity m^* and m^{**}, respectively. There are k^* and k^{**} cross-cuts in these two domains, such that $k^* + k^{**} = k - 1$. $V*$ is divided into $V_1, ..., V_{L^*}$ by these cross-cuts. Thus inductive assumption, we have

$$\sum_{j=1}^{l^*}(m_j - 2) = m^* - 2 - k^*,$$

similarly

$$\sum_{j=l^*+1}^{l}(m_j - 2) = m^{**} - 2 - k^{**}.$$

Adding up gives the desired result.

Otherwise, however, $V \setminus c_1$ is a domain with multiplicity $m-1$, and is divided by $c_2, ..., c_k$ into domains $V_1, ..., V_m$, thus

$$\sum_{j=1}^{l}(m_j - 2) = (m-1) - 2 - (k-1) = m - 2 - k.$$

Now we come to the proof of the formula itself. First, by applying the Riemann-mapping theorem, we may assume that V and W are all bounded by analytic Jordan curves and singletons.

We will first discuss the case where f is unramified (so that $r = 0$). Then any local branch of f^{-1} may be continued along any curve in W. If W is simply connected, then, by monodromy theorem, f^{-1} is single-valued in W and thus is a conformal mapping. This means that n=1 implies implies $m = k = 1$, thus (6) holds. We proceed by induction. In case $m > 1$, we take a cross-cut c in W, which diminishes the connectivity number: $W^* = W \setminus c$ is $(m-1)-$connected. Suppose $f^{-1}(c)$ is k cross-cuts $c_1, ..., c_k$ of V, and $f^{-1}(W)$ consists of l domains $V_1, ..., V_l$ of connectivity $m_1, ..., m_l$, respectively, and f is a k_j-sheeted proper map $V_j \to W^*$, where $k_1 + ... + k_l = k$. By inductive assumption and lemma 3.5., we get the desired result.

If f is ramified, it has finitely many critical values $w_1, ..., w_s \in W$. Then $W^* = W \setminus \{w_1, ..., w_s\}$ has connectivity $n + s$. Since any w_j has p_j preimages with multiplicities q_j^t such that $\sum_{t=1}^{p_j} q_j^t = k$, and $\sum_{j=1}^{s}\sum_{t=1}^{p_j}(q_j^t - 1) = r$, thus $\sum_{j=1}^{s} = sk - r$, $V^* = f^{-1}(W^*)$ has connectivity $m + sk - r$. Since $f : V^* \to W^*$ is k-sheeted and unramified, we have

$$m + ks - r = k(m + s - 2)$$

and (6) holds true also in this case. □

Remark 3.1. *We would introduce Monodromy theorem as follows. Let f be a function which is analytic in the domain D and let G be a simply-connected region which contains D. For any path γ in G with initial point in D there is an analytic continuation of f on D along γ. Then there is an analytic function $F : G \to \mathbb{C}$ such that $F(z) = f(z)$ for all z in D.*

4 Dynamics of the Deltoid Reflection: Proof for Main Result

4.1 introduction

Suppose $\varphi(z) = z + \frac{1}{2z^2}$ and $\tilde{\iota}(z) = \frac{1}{\bar{z}}$, where $\tilde{\iota}$ is reflection in the unit circle, one can see that this map is univalent in $\hat{\mathbb{C}} \setminus \mathbb{D}$. We define

$$\Omega := \varphi(\hat{\mathbb{C}} \setminus \overline{\mathbb{D}}),\ T := \Omega^c,\ \sigma := \varphi \circ \tilde{\iota} \circ (\varphi|_{\hat{\mathbb{C}} \setminus \mathbb{D}})^{-1} : \overline{\Omega} \to \hat{\mathbb{C}}.$$

$$\begin{tikzcd}
\hat{\mathbb{C}} \setminus \mathbb{D} \arrow[r, "\varphi"] \arrow[d, "\tilde{\imath}"] & \overline{\Omega} \arrow[d, "\sigma"] \\
\overline{\mathbb{D}} \arrow[r, "\varphi"] & \hat{\mathbb{C}}
\end{tikzcd}$$

Note that $\sigma(z) = z$, $\forall z \in \partial\Omega$, thus σ is the associated Schwarz reflection map of Ω. Note that $\varphi(wx) = w\varphi(x)$, it follows that T is symmetric rotation by $\frac{2\pi}{3}$. Moreover, as φ has simple critical points at $1, w, w^2$, ∂T has three $\frac{3}{2}-$cusp points (one can see this by Taylor expansions of $x = cos\theta + \frac{cos2\theta}{2},\ y = sin\theta - \frac{sin2\theta}{2}$). We denote that T^0 is $T - \{\frac{3}{2}, \frac{3w}{2}, \frac{3w^2}{2}\}$.

4.1.1 Schwarz reflection σ

In this section, we will give some properties of σ as a function or a covering mapping.

Proposition 4.1. *The Schwarz reflection map σ of Ω has a double pole at ∞, but no other critical points in Ω.*

Proof. Since the only critical point of φ in $\mathbb{D}$ is at the origin, it follows that σ has a pole at ∞. After analysis of the dominant term of the formula at $w = \infty$, one can see the order of the pole is two.

By definition, one can see $\sigma(\varphi(w)) = \varphi(\frac{1}{\overline{w}})$, thus $\frac{\partial\sigma}{\partial\overline{w}}(\varphi(w)) \cdot \frac{\partial\overline{\varphi(w)}}{\partial\overline{w}} = \frac{\partial\varphi(\frac{1}{\overline{w}})}{\partial\overline{w}}$. Substitute $\varphi(w) = w + \frac{1}{2w^2}$ in it, we have $\overline{\partial}(\varphi(w)) = \overline{w}$ (we denote $\frac{\partial f}{\partial\overline{z}}$ by $\overline{\partial}f$). Since the univalence of $\varphi(w)$ in $\hat{\mathbb{C}} \setminus \mathbb{D}$ and $(\varphi|_{\mathbb{C}\hat{\setminus}\mathbb{D}})^{-1}$ and $\frac{1}{\overline{z}}$ are all well-defined in the corresponding domain, there is no other critical points of σ in Ω. □

Proposition 4.2. *$\sigma : \sigma^{-1}(\Omega) \to \Omega$ is a proper branched $2-$cover branched only at ∞. On the other hand, $\sigma : \sigma^{-1}(T^0) \to T^0$ is a $3-$cover.*

Proof. Suppose that $X_1 = \varphi^{-1}(T^0) \cap \overline{\mathbb{D}}$ and $X_2 = \varphi^{-1}(\Omega) \cap \overline{\mathbb{D}}$, by definition, $\sigma^{-1}(T^0) = (\varphi \circ \tilde{\imath})(X_1)$, $\sigma^{-1}(\Omega) = (\varphi \circ \tilde{\imath})(X_2)$. In fact, one can see that there are three connected components of X_1 and the connectedness of X_2 by considering the preimage of ∂T. What's more, because φ is continuous and the degree of φ is three, $\varphi : X_1 \to T^0$ is a proper covering of degree three and $\varphi : X_2 \to \Omega$ is proper branched covering of degree 2 branched only at 0.

We can construct the univalent components of X_1 and X_2 by φ which implies that $\sigma : (\varphi \circ \tilde{\imath})(X_1) \to T^0$ is a proper covering of degree three and $\sigma : (\varphi \circ \tilde{\imath})(X_2) \to \Omega$ is proper branched covering of degree 2 branched only at ∞. □

Note that $\sigma^{-1}(\Omega) = (\varphi\circ\tilde{\imath})(X_2) \subseteq \varphi(\hat{\mathbb{C}}\setminus\overline{\mathbb{D}}) = \Omega$, we can induce the conclusion below.

Corollary 4.3. *The maps $\sigma^{\circ n} : \sigma^{-n}(\Omega) \to \Omega$ are proper branched covers of degree 2^n branched only at ∞.*

In the dynamical system of ρ in Π, we have $\mathbb{D} = \bigcup_{n\in\mathbb{N}} \rho^{-(n-1)}(\Pi)$. Similarly, we denote that $T^\infty = \bigcup_{n\in\mathbb{N}} \sigma^{-(n-1)}(T^0)$ and call it tiling set. Moreover we will call the components of $\sigma^{-n}T^0$ tiles of rank n. Also, the image of these two dynamical systems is similar. In fact, there are $3 \cdot 2^{n-1}$ tiles of rank n and the union of tiles of rank$\leq n$ is a "polygon" with $3 \cdot 2^n$ iterated preimage of cusps of T as its vertices. Let $\triangle_n$ be a tile of rank $n \geq 1$. It's a "triangle"; one of its sides is a side of a tile of rank $n-1$; we will call it (the side) the base of $\triangle_n$.

4.1.2 The tiling set T^∞

First, we will induct a lemma to prove the geometric properties of T^∞.

Lemma 4.4. *Let K be $\lim_{n\to\infty} K_n$, where $K_n \subseteq K_{n+1}$, $\forall n \in \mathbb{N}$, if K_n are all connected domain, then K is connected. Further, if K_n are all simply connected, then K is simply connected.*

Proof. Consider $x, y \in K$, without loss of generality, suppose $x, y \in K_n$ for some integer numbers n. Because of the connectedness of K_n, we have x, y are connected in K. Thus K is connected domain. Consider a Jordan curve $\gamma \in K$, for any point $z \in \gamma$, it will belong to K_n eventually for some integer numbers n. Because of openness of K_n, there is a neighborhood of z (denoted by B_z)is contained in K_n. The compactness of curve γ induces that there are finite points $z_k : k = 1, 2, ..., m$ and their corresponding neighborhood B_{z_k} such that $\gamma \subseteq \bigcup_{k=1}^m B_{z_k}$, therefore it is contained in K_n for some integer numbers n, and can be shrunk into a single point in K_n because K_n is simple-connected. To sum up, K is simply-connected. □

Proposition 4.5. *T^∞ is a simply connected domain.*

Proof. For $z \in T^\infty$, if z belongs to the interior of some tiles, then it belongs to intT^∞. If z belongs to the boundary of some tiles, assume the order is n, then $\sigma^n(z) \in \partial T$ and not the three cusps, therefore, $\sigma^n(z) \in \text{int}\Pi$*, since the continuity of* σ, we have $z \in T^\infty$. Hence, T^∞ is open.

Note that T^∞ is a increasing union of the connected, simply connected domains $\{\text{int}(\bigcup_{k=0}^n \sigma^{-k}T^0)\}_{n\in\mathbb{N}}$, by lemma, we have T^∞ itself is connected and simply connected. □

Corollary 4.6. *$\hat{\mathbb{C}} \setminus T^\infty$ is a closed, completely invariant set.*

Now we pay attention to the inverse branches of the iterates of σ on T^∞. For a subset X, we denote by $N_\epsilon(X)$ the $\epsilon-$neighborhood of X. Let us fix some small $\epsilon > 0$ and $K > 1$ such that the set $T^{hyp} := N_\epsilon(\overline{T^\infty}) \setminus \overline{N_{K\epsilon(T)}}$ has three simply connected components.

Proposition 4.7. *All inverse branches of $\sigma^{\circ n}, n \geq 1$ are well-defined locally on T^{hyp}. Moreover, σ is hyperbolic on T^{hyp}.*

Proof. Note that $\overline{T^{hyp}} \subseteq \Omega$, and $\sigma : \sigma^{-1}(\Omega) \to \Omega$ is a branched covering and T^∞ is bounded, therefore T^{hyp} doesn't intersect ∞, the critical point of iteraction of σ.

For the second part of the proposition, recall that $|\overline{\partial}\sigma(\varphi(w))| = |w| > 1, \forall w \in \hat{\mathbb{C}} \setminus \overline{\mathbb{D}}$. Since $\overline{T^{hyp}} \subseteq \Omega$ is compact, it follows that $\overline{\partial}\sigma$ have a fixed low bound $\lambda_0 > 1$. □

Remark 4.1. *We omit the proof that T^∞ is bounded. This conclusion will be clearly as a corollary of Proposition 4.9.*

Since points in $T^\infty \setminus T^0$ escape to T^0 under some iteration of σ, we say that T^∞ is the escaping set of σ and $\hat{\mathbb{C}} \setminus T^\infty$ is the non-escaping set of σ.

4.1.3 the basin of ∞

By the double pole at ∞ of σ, ∞ is a super-attracting fixed point of σ. We denote the basin of attraction of ∞ by $A = \{z \in \mathbb{C} \mid \sigma^{\circ n}(z) \to \infty \textit{ as } n \to \infty\}$. Naturally, $A \subseteq \hat{\mathbb{C}} \setminus \overline{T^\infty}$.

Proposition 4.8. *A is a simply connected, completely invariant domain.*

Proof. By definition, for any point $x \in A$, we have $\sigma(x) \in A$ and $\sigma^{-1}(x) \subseteq A$, therefore, A is a completely invariant set.

By the property of ∞, $\forall \epsilon > 0$, $\exists M > 0$, such that $|\overline{\partial}\sigma(z)| < \epsilon$, $\forall |z| > M$. Thus for $|x|, |y| > 2M$, $|\sigma(x) - \sigma(y)| = |\int_\gamma \overline{\partial}\sigma(w)d\overline{w}| < |\int_\gamma \epsilon|d\overline{w}|| < \epsilon|x - y|$. The above fact induces that there is a neighborhood of ∞ (denoted by U) is contained in A and A is the increasing union of the domain $\{\sigma^{-(k-1)}(U)\}_{k\in\mathbb{N}}$. It follows that A is open. What's more, $\forall k \in \mathbb{N}$, every connected component of $\sigma^{-(k-1)}(U)$ has a preimage of ∞. While $\sigma^{-1}(\infty) = \{\infty\}$, $\sigma^{-(k-1)}(U)$ is connected, by lemma 4.4., A is connected.

$\sigma : \bigcup_{i=1}^n \sigma^{-(i-1)}(U) \to \bigcup_{i=1}^{n-1} \sigma^{-(i-1)}(U)$ is a proper branched $2-$cover branched only at ∞ (the order is 2). Because of simply-connectedness of U, by Riemann-Hurwitz formula, we have $\bigcup_{i=1}^n \sigma^{-(i-1)}(U)$ is simply-connected by induction on n. Thus A is simply-connected domains by lemma 4.4. □

4.1.4 Singular points

We define $S := \bigcup_{n\in\mathbb{N}} \sigma^{-(n-1)}(T \setminus T^0)$. It's clear that $S \subseteq \partial T^\infty$. Also we have

Proposition 4.9. *$S \subseteq \partial A$.*

Proof. Since A is completely invariant, so is ∂A. For a real $x > 1$, we have

$$\sigma(x + \frac{1}{2x^2}) = \frac{1}{x} + \frac{x^2}{2} > x + \frac{1}{2x^2},$$

thus the forward $\sigma-$orbit of any real $x > \frac{3}{2}$ must converge to ∞. Otherwise, assume the orbit would converges to a fixed positive number $s > \frac{3}{2}$, while this is contradiction to the inequality above, thus $(\frac{3}{2}, +\infty) \subseteq A$, which induced that $\frac{3}{2} \in \partial A$. By definition, we have $S \subseteq \partial A$. □

4.2 Proof for Theorem 4.12. in [1]

As a summary to above analysis, we give the following results about the dynamical plane of σ.

Theorem 4.10. *We have $\partial T^\infty = \partial A = \overline{S}$. Moreover, this set, which we denote by Γ, is a Jordan curve. Moreover, $\hat{\mathbb{C}} = T^\infty \sqcup \Gamma \sqcup A$.*

Proof. Let $\psi^{in} : \overline{\Pi} \to T$ be the homeomorphic extension of a conformal isomorphism such that $\psi^{in}(0) = 0, \psi^{in}(1) = \frac{3}{2}$. Since σ has no critical point in T^∞, the tiles of all rank of σ map diffeomorphically onto T^0 under iterates of σ. Similarly, the tiles of the tessellation of $\mathbb{D}$ arising from the ideal triangle group $\mathcal{G}$ map diffeomorphically onto Π under iterates of ρ. Furthermore, σ and ρ act as identity maps on ∂T^0 and $\partial \Pi$ respectively. This allows us to lift ψ^{in} to a conformal isomorphism $\mathbb{D} \to T^\infty$. Note that the trivial actions of σ and ρ on ∂T^0 and $\partial\Pi$ ensure that this map match on the boundaries of tiles. Thus, by construction, the conformal map ψ^{in} conjugates $\rho : \mathbb{D} \setminus \text{int}\Pi \to \mathbb{D}$ to $\sigma : T^\infty \setminus \text{intT} \to \text{T}^\infty$ by the uniqueness of the Schwarz reflection map. By the local connectedness of Γ (cf.Lemma 4.11.) and *Carathéodory's theorem*, we can extend ψ^{in} to $\mathbb{T}$ continuously. Since the preimage of cusps by $\mathcal{G}$ are dense in $\mathbb{T}$, we have $\overline{S} = \psi^{in}(\mathbb{T})$. In fact, for any point $x \in \partial T^\infty$, there is a sequence of distinct points $\{x_1, x_2, ...\} \subseteq T^\infty$ converging to it, thus $x = \psi^{in}(y)$ for some points $y \in \overline{\mathbb{D}}$. If $y \in \mathbb{D}$, then $x = \psi(y) \in T^\infty$, which is contradiction to the selection of x and openess of T^∞. Therefore $\partial T^\infty \subseteq \psi^{in}(\mathbb{T})$. Similarly, by Open-mapping theorem, we have $\psi^{in}(\mathbb{T}) \subseteq \partial T^\infty$, i.e. $\overline{S} = \partial T^\infty$.

By proposition 4.9.,$S \subseteq \partial A$, therefore, $\partial T^\infty = \overline{S} \subseteq \partial A$. It remains to prove the opposite inclusion. Let $a \in \partial A \setminus \partial T^\infty$. Then $a \notin \overline{T^\infty}$, and there is a open set U such that $a \in U \subseteq \mathbb{C} \setminus T^\infty$. Note that all iterates of σ can be defined in U properly, because they avoid T. Note that $\sigma(\varphi(x)) = \varphi(\frac{1}{\overline{x}})$, and $x + \frac{1}{2x^2} = \frac{1}{\overline{x}} + \frac{\overline{x}^2}{2}$ implies that $2|x - \frac{1}{\overline{x}}| = |\overline{x}^2 - \frac{1}{x^2}| = |x - \frac{1}{\overline{x}}||x + \frac{1}{\overline{x}}|$, thus $x \in \mathbb{T}$, i.e., $\varphi(x) \in \partial T$. For any point x in U, the iteraction series doesn't converge to ∞ (we assume the cluster point is $\tilde{x}$, thus $\tilde{x}$ is a fixed point of σ). By the analysis above, we have $\tilde{x} \in \partial T$, thus $\tilde{x}$ is a cusp point of T. Therefore $\{\sigma, \sigma^{\circ 2}, ..., \sigma^{\circ n}, ...\}$ is a normal family and converge uniformly to a analytic function g or ∞ by Montel theorem, which is contradiction to the fact that $\tilde{x}$ is a cusp point. It follow that $a \in A$, a contradiction.

It is the last part of theorem that Γ (called the limit set of σ) is a Jordan curve. In this article, authors think this fact is based on the locally connectedness of ∂T^∞. I don't know why. □

The rest of this section is the proof of lemma 4.13. in [1].

Lemma 4.11. *Γ is locally connected.*

4.2.1 Local dynamics near cusp points

We consider the dynamics image near $\frac{3}{2}$, the case of other two cusps of T are similar. For $\epsilon > 0$ small enough, let us denote

$$B = B(\frac{3}{2}, \epsilon),\ B^- = B \cap \{z \mid Re(z) < \frac{3}{2}\},\ B^+ = B \setminus B^-.$$

On domain $B \cap \Omega$, we have the following series expansion of $\sigma(\frac{3}{2} + \overline{\delta}^2) = \frac{3}{2} + \delta^2 + k\delta^3 + O(\delta^4)$, thus

$$\sigma(\frac{3}{2} + \overline{\delta}) = \frac{3}{2} + \delta + k\delta^{\frac{3}{2}} + O(\delta^2),$$

where $k > 0$ and the chosen branch of square root is located in $Re(z) \geq 0$. Moreover, we can obtain an asymptotic expansion of the form $\zeta \mapsto \overline{\zeta} + \frac{1}{2} + O(\frac{1}{\zeta})$ on $\kappa(B \cap \Omega)$ by coordinate exchange $\kappa : w \mapsto \frac{k_1}{\sqrt{w - \frac{3}{2}}}$, where k_1 is a proper negative number to make sure the constant $\frac{1}{2}$ in asymptotic expansion.

Comparing with the proof of proposition 4.9., $(\frac{3}{2}, \frac{3}{2} + \epsilon)$ is a repelling direction of σ at $\frac{3}{2}$. And after the coordinate exchange, it was sent to the real axis near ∞. $\kappa(B \cap \Omega)$ is contained in an angle π at ∞. For any $\alpha \in (0, \frac{\pi}{2})$, points with sufficiently large absolute value and lying between the boundary curves $\kappa(B \cap \partial\Omega)$ and the infinite rays $\kappa(\frac{3}{2} + [0, \epsilon)e^{\pm i\alpha})$ eventually escape to $\kappa(B \setminus \Omega)$. In the original coordinates, this means that points sufficiently close to $\frac{3}{2}$ which not located in the real axis will escape to T_0 after iterates of σ. We will record these observations as following.

Proposition 4.12. *If ϵ is sufficiently small, $B^- \subseteq T^\infty$, and $\sigma^{-n}B^+ \to \{\frac{3}{2}\}$. Here σ^{-n} is the branch in B^+ which fixes $\frac{3}{2}$.*

Proof. We would add the proof of some above analysis. First, we need to give precise choice of the branch of square root, for $z \in \mathbb{C}^+$, we define $\sqrt{z}$ as the point that located in the upper plane; for $z \in \mathbb{C}^-$, we define $\sqrt{z}$ as the point that located in the lower plane. What's more, it send the positive real number to the positive one. Secondly, we draw the image of $\kappa(B \cap \Omega)$, one can see the point will move along positive real axis after iteractions. Thus for the point with the sufficiently large absolute value, it will escape to $\kappa(B \setminus \Omega)$. □

What's more, note that locally near the cusps, ∂T^∞ is contained in the "repelling petals" of the cusp points. By definition, $\exists \epsilon > 0$ such that if an orbit in ∂T^∞ stays $\epsilon-$close to a cusp of T, then the orbit lands on this cusp.

Let us denote the Green function of A with pole at ∞, which can be explicitly written as in A

$$G(z) = \lim_{n \to +\infty} \frac{log|\sigma^{\circ n}(z)|}{2^n}.$$

In A^c, we define $G = 0$. By definition, for any $\rho > 0$, σ maps the level curve $\{G = \rho\}$ to level curve $\{G = 2\rho\}$.

Recalling the notions given in section 4.1.1., we have the vertices of the base are iterated preimages of the cusps of T. Note that $\frac{3}{2}$ is connected to ∞ through a external ray, the vertices of the base of $\triangle_n$ are landing points of two external rays in A. We denote by $P(\triangle_n)$ the closed region bounded by the base of $\triangle_n$, the two rays, and the level curve $\{G = \frac{1}{2^n}\}$ (so $\triangle_n \subseteq P(\triangle_n)$), and call it a puzzle piece of rank n.

Proposition 4.13. *For each $n \geq 1$, the sets $\partial T^\infty \cap P(\triangle_n)$ are connected.*

Note that these sets are all $\sigma-$invariant, thus it suffices to prove that the case of $n = 1$. By definition, $\Gamma \setminus \{\frac{3}{2}, \frac{3w}{2}, \frac{3w^2}{2}\}$ can be divided into three parts as the form $\Gamma \cap P(\triangle_1)$. Because T^∞ is a bound simply-connected domain, Γ is a connected set, $\Gamma \cap P(\triangle_1)$ is connected as well by the symmetry of Γ. Combining the $\mathcal{G}-$ray, we have the following proposition naturally by considering the dynamical plane $\mathbb{D}$ of ρ.

Proposition 4.14. *The puzzle pieces separate the impressions of the internal rays of T^∞ (images of $\mathcal{G}-$ rays under ψ^{in}) landing at points of S from each other.*

4.2.2 Local connectivity at "radial" points

Proposition 4.15. *If $x \in \partial T^\infty \setminus S$, then ∂T^∞ is locally connected at x.*

Proof. Note that ∂T^∞ is completely $\sigma-$invariant, then the set $\{x \in \partial T^\infty \mid \partial T^\infty$ is locally connected at x$\}$ is completely $\sigma-$invariant.

Consider the orbit $x_n = \sigma^{\circ n}(x)$ of some $x \in \partial T^\infty \setminus S$. If $d(x_n, T) \to 0$, as $n \to \infty$, then $d(x_n, T \setminus T^0) \to 0$, as $n \to \infty$, or it will landing into T^0 eventually by the fact about "repelling petals ". This would imply that $x \in S$, which contradicts our choice of x.

Thus there is a subsequence of $\{x_n\}$ at a positive distance from T with a cluster point ζ. By the compactness of ∂T^∞, ζ is in ∂T^∞ and not a cusp of T. By replacing ζ by one of its iterated preimages, we can assume that ζ doesn't lie in the impression of the rays at angles $0, \frac{1}{3}, \frac{2}{3}$.

The above assumption on ζ, we can assume $\zeta \in \text{int}P \subseteq P \subseteq \text{int}P_1$, where P, P_1 are puzzle pieces with rank m and 1, respectively. Thus we have $x_n \in \text{int}P$ for infinitely many terms. To prove that ∂T^∞ is locally connected at x, it suffices to show that suitably chosen iterated preimages of $\text{int}P$ produce a basis of open, connected neighborhoods of x in ∂T^∞.

In fact, by proposition 4.7., for each n with $x_n \in \text{int}P$, we can define the inverse branches $\sigma^{-n} : \text{int}P_1 \to \mathbb{C}$, $x_n \mapsto x$. These inverse branches form a normal family on $\text{int}P_1$.

We claim that there is a subsequence such that $(\sigma^{-n})' \to 0$ locally uniformly on $\text{int}P_1$, so on the compact set $P \cap \partial T^\infty$. Indeed, we have $\sigma^{-n_k} \to g$ on $\text{int}P_1$, and we need to prove that $g' = 0$ on some open set $V \subseteq P_1$. Consider that V is a small disc inside the rank one tile contained in P_1, so the preimages $V_k := \sigma^{-n_k} V$ are disjoint open sets. By Koebe distortion theorem, we have

$$area(V_k) = O[diam(V_k)]^2,$$

so $diam(V_k) \to 0$, which implies the claim.

Thus, $diam[\sigma^{-n_k}(\mathrm{int}P \cap \partial T^{\infty})] \to 0$, as $k \to \infty$. These sets are open connected neighborhoods of x in ∂T^{∞} which implies ∂T^{∞} is local connected at x. □

4.2.3 Local connectivity at S

To finish the proof of the local connectivity of ∂T^{∞}, it suffices to check local connectivity at the point $\frac{3}{2}$. Let $\triangle_n^{\pm}$ be the two tiles of rank n which have $\frac{3}{2}$ as a common vertex. Let

$$\tilde{P}_n := int[P(\triangle_n^{+}) \cup P(\triangle_n^{-})]$$

, and for any $n \geq 2, \sigma : \tilde{P}_n \to \tilde{P}_{n-1}$ is a bijection.

The sets $(\partial T^{\infty} \cap \tilde{P}_n) \cup \{\frac{3}{2}\}$ are open and connected. Moreover, their diameters go to zero. Hence, they form a basis of open connected neighborhoods of $\frac{3}{2}$ in the relative topology in ∂T^{∞}.

4.3 Proof for Theorem 4.25. in [1]

In this section, we will study the dynamics near cusp points, the following wedges will be useful. In the dynamical plane of σ, for $\theta \in [0, \frac{\pi}{2})$, we define

$$W_{\theta} := U_{\theta} \cup wU_{\theta} \cup w^2 U_{\theta} \subseteq T^{\infty},$$

where

$$U_{\theta} := \{\frac{3}{2} + re^{i\theta'} : 0 < r < \frac{1}{100}, \frac{\pi}{2} - \theta < \theta' < \frac{3\pi}{2} + \theta\}.$$

Similarly, in the dynamical plane of ρ, for $\theta \in [0, \frac{\pi}{4})$, we define

$$\mathbf{W}_{\theta} := \mathbf{U}_{\theta} \cup w\mathbf{U}_{\theta} \cup w^2\mathbf{U}_{\theta} \subseteq \mathbb{D},$$

where

$$\mathbf{U}_{\theta} := \{1 + re^{i\theta' : 0 < r < \frac{1}{100}, \frac{3\pi}{4} < \theta' < \frac{5\pi}{4}} + \theta\}.$$

First, for two paths $\Gamma_1(\tau)$ and $\Gamma_2(\tau)$ in $\mathbb{D}$, where $0 \geq \tau < \infty$, and let $C > 0$, we denote that the paths shadow each other with a constant C if

$$\forall \tau > 0,\ d_{\mathbb{D}}(\Gamma_1(\tau), \Gamma_2(\tau)) \geq C;$$

and we denote it by:$\Gamma_1 \approx \Gamma_2$.

Secondly, let us fix a positive fixed number $\eta > 0$ and state a "shadowing" lemma for ρ. For a fixed $\theta_0 \in [0, \frac{\pi}{4})$ and $\mathbf{z}_0 \in \mathbb{D} \setminus \mathbf{W}_{\theta_0}$ with $d(\mathbf{z}_0, \mathbf{T}) < \eta$, and $\Gamma^{\mathbf{z}_0}$ be the segment $[\mathbf{z}_0, \zeta_0)$, where $\zeta_0 = \frac{\mathbf{z}_0}{|\mathbf{z}_0|}$. We parametrize $\Gamma^{\mathbf{z}_0}$ such that $d_{\mathbb{D}}(\mathbf{z}_0, \Gamma^{\mathbf{z}_0}(\tau)) = \tau$. If $\rho^{\circ}(\mathbf{z}_0)$ can be defined, we denote it by $\mathbf{z}_n$, $\zeta_n := \rho^{\circ n}(\zeta_0)$, and $\mathbf{z}'_n = |\mathbf{z}_n|\zeta_n$. Let N be a positive integer such that $d(\mathbf{z}_k, \mathbb{T}) < \eta$ and $\mathbf{z}_k \notin \mathbf{W} := \mathbf{W}_{\theta_0}$, for $k = 0, 1, ..., N$.

Lemma 4.16. *For a fixed $\theta_0 \in [0, \frac{\pi}{4})$, we have that $\rho^{\circ N}(\Gamma^{\mathbf{z}_0}) \approx \Gamma^{\mathbf{z}'_N}$, where the shadowing constant $C = C(\theta)$ is independent of $\mathbf{z}_0$ and N.*

Proof. First, since ρ fixed $\mathbb{T}$ as a set, we can extend ρ to the open set $\mathcal{V} := \hat{\mathbb{C}} \setminus \overline{\Pi \cup \tilde{\iota}(\Pi)}$ by Schwarz reflection principle and denote the three connected components of $\mathcal{V}$ by $\mathcal{V}_i, i \in \{1, 2, 3\}$.

For any $k = 0, 1, ..., N$, since $\mathbf{z}_k \notin \mathbf{W}$ and $d(\mathbf{z}_k, \mathbb{T}) < \eta$ for η small enough, we have $\mathbf{z}_k \in \mathcal{V}_i$ for some i. Note that $\mathcal{V}$ is the union of three discs which is vertical to $\mathbb{D}$, $\rho^{-N}(\mathcal{V}_i)$ is a disjoint union of finitely many discs that are invariant under $\tilde{\iota}$, we have $\mathbf{z}_0$ and ζ_0 lies in the same disk $\mathbb{T}_0$ of above ones. By construction, $\mathbf{z}_N$ and ζ_N are in the same component $\mathcal{V}_i$. It follows that there is a mapping $\rho^{-N} : \mathcal{V}_i \to \mathbf{T}_0 : \mathbf{z}_n \mapsto \mathbf{z}_0, \zeta_N \mapsto \zeta_0$. The fact that $\mathbf{z}_N \notin \mathbf{W}$ and $d(\mathbf{z}_N, \mathbb{T}) < \eta$ imply that

$$\mathcal{A} := \mathcal{V}_i \setminus \overline{(\rho^{\circ N}(\Gamma^{\mathbf{z}_0}) \cup \tilde{\iota}(\rho^{\circ N}(\Gamma^{\mathbf{z}_0})))}$$

satisfying that $mod(\mathcal{A})$ has a positive low bounded that denpends only on θ_0 and is independent of $\mathbf{z}_0$ and N. Note that ρ^{-N} on $\mathcal{V}_i$ is an isometry with two corresponding hyperbolic metrics in these two domain. Hence, $\rho^{-N}(\mathcal{A})$ is a annulus of definite modulus surrounding $\overline{\Gamma^{\mathbf{z}_0} \cup \tilde{\iota}(\Gamma^{\mathbf{z}_0})}$ in $\rho^{-N}(\mathcal{V}_i)$.

The lower bound of $mod[\rho^{-N}(\mathcal{A})]$ implies that $\overline{\Gamma^{\mathbf{z}_0} \cup \tilde{\iota}(\Gamma^{\mathbf{z}_0})}$ is uniformly bounded away from the boundary of $\rho^{-N}(\mathcal{V}_i)$, thus the hyperbolic metric of $\mathbb{D}$ and that of $\rho^{-N}(\mathcal{V}_i)$ are both uniformly comparable (here, comparability means equivalence) to the reciprocal of the distance to $\mathbb{T}$. Thus they are comparable to each other on $\Gamma^{\mathbf{z}_0}$, by isometry $\rho^{\circ N} : \rho^{-N}(\mathcal{V}_i) \to \mathcal{V}_i$ with hyperbolic metrics, we have uniformly comparability between hyperbolic metrics of $\mathbb{D}$ and of $\mathcal{V}_i$ on $\rho^{\circ N}(\Gamma^{\mathbf{z}_0})$.

Therefore, we have

$$d_{\mathbb{D}}(x, y) \propto d_{\rho^{-N}(\mathcal{V}_i)}(x, y),$$

$$d_{\mathbb{D}}(\rho^{\circ N}(x), \rho^{\circ N}(y)) \propto d_{\mathcal{V}_i}(\rho^{\circ N}(x), \rho^{\circ N}(y)),$$

and

$$d_{\rho^{-N}(\mathcal{V}_i)}(x, y) = d_{\mathcal{V}_i}(\rho^{\circ N}(x), \rho^{\circ N}(y)),$$

which implies that$\rho^{\circ N} : \Gamma^{\mathbf{z}_0} \to \rho^{\circ N}(\Gamma^{\mathbf{z}_0})$ is a quasi-isometry with constants independent of $\mathbf{z}_0$ and N.

(why?) Therefore, $\rho^{\circ N}(\Gamma^{\mathbf{z}_0})$ is shadowed by a hyperbolic geodistic of $\mathbb{D}$ with one end-point at ζ_N. Since ρ is expanding away from the third roots of unity, we conclude that the other end-point of this shadowing geodesic is bounded away from ζ_N. It follows that $\rho^{\circ N}(\Gamma^{\mathbf{z}_0})$ is shadowed by the geodesic arc $\Gamma^{\mathbf{z}'_N}$, with a shadowing constant independent of $\mathbf{z}_0$ and N.

□

Remark 4.2. *We add a definition of quasi-isometry between metric spaces (X, d_X) and (Y, d_Y), we say a mapping $f : X \to Y$ is isometry with constants $\lambda > 0$ and $k \geq 0$ if for all x and x' in X, the following inequality holds:*

$$\lambda^{-1} d_X(x, x') - k \leq d_Y(f(x), f(y)) \leq \lambda d_X(x, x') + k.$$

Especially here we have $k = 0$.

Next, we would state two uniform estimates for the model map ρ. Here, under ψ^{in}, $0 \in T^{\infty}$ corresponds to $0 \in \mathbb{D}$ and curves γ^{z} in T^{∞} are now geodesic rays through the origin. We set $W := W_{\theta_0}$ for fixed $\theta_0 \in [0, \frac{\pi}{2})$.

Since Γ is a Jordan curve, we can prove the following lemma by contradiction.

Lemma 4.17. $\forall \epsilon > 0, \exists M > 0$ *such that if* $\delta(z') \geq \eta, w' \in \gamma^{z'}$, *and* $d_{T^{\infty}}(w', z') \geq M$, *then* $\delta(w') \leq \epsilon\delta(z')$.

Lemma 4.18. $\forall \epsilon > 0, \exists M > 0$ *such that if* $\sigma(z') \in N_{hyp}(W, C)$, *where* $N_{hyp}(W, C) := \{z \in T^{\infty} : d_{T^{\infty}}(z, W) \leq C\}$, *$C$ is the shadowing constant from lemma 4.16.,* $w' \in \gamma^{z'}$, *and* $d_{T^{\infty}}(w', z') \geq M$, *then* $\delta(w') \leq \epsilon\delta(z')$.

By the following estimation, we will obtain the theorem 4.25. in [1].

4.4 Proof for the Subsection 4.4.3. in [1]

We now show that the Schwarz reflection of the deltoid arises as the unique conformal mating of the anti-polynomial $\overline{z}^2$ and the reflection map ρ.

First, since A is simply connected in $\hat{\mathbb{C}}$, we have a Riemann uniformization $\psi^{out} : \hat{C} \setminus \mathbb{D} \to \overline{A}$ such that $\infty \mapsto$ and $1 \mapsto \frac{3}{2}$. For $\sigma-$action on non-escaping set $\hat{A}$, we can conjugate it to $f_0 : \mathbb{C} \hat{\setminus} \mathbb{D} \to \mathbb{C} \hat{\setminus} \mathbb{D}, f_0(z) = \overline{z}^2$ by ψ^{out}.

It is based on the fact that σ has only critical at ∞ as a pole with order 2. Thus in the coordinate under ψ^{out}, the corresponding function f_0 with the only critical at ∞ satisfies that $f_0 : \hat{\mathbb{C}} \setminus \mathbb{D} \to \hat{\mathbb{C}} \setminus \mathbb{D}$. Consider $\frac{1}{f(\frac{1}{\overline{z}})}$, we can write it as a degree two anti-Blaschke product. Thus we have $\frac{1}{f_0(\frac{1}{\overline{z}})} = \overline{z}^2$, i.e., $f_0(z) = \overline{z}^2$.

Secondly, we will now show the deltoid group $\mathcal{G}_{\triangle} \subseteq Aut(T^{\infty})$ is conformally equivalent to the ideal triangle group $\mathcal{G}$. Precisely, we have the following proposition.

Proposition 4.19. *Let T_j^1 (j=1,2,3) be the three tiles of rank* 1, *so we have*

$$\sigma^{-1}T^0 = T_1^1 \sqcup T_2^1 \sqcup T_3^1.$$

Then each map $\sigma : T_j^1 \to T^0$ extends to a conformal automorphism $\sigma_j : T^{\infty} \to T^{\infty}$. The deltoid group $\mathcal{G}_{\triangle} := \langle\sigma_1, \sigma_2, \sigma_3\rangle \subseteq Aut(T^{\infty})$ is conformally conjugate to the ideal triangle group $\mathcal{G}$.

Proof. Recall that in proof of theorem 4.10., we have proved that $\psi : \mathbb{D} \to T^{\infty}$ conjugates ρ to σ, and hence in particular, conjugates $\rho_j|_{\rho_j(\Pi)}$ to $\sigma|_{T_j^1}$. Also, the desired between $\mathcal{G}$ and $\mathcal{G}_{\triangle}$ is given by ψ^{in}. □

The final part is related to introduce $\sigma-$action on the limit set. For this common boundary, on the one hand, $\sigma : \Gamma \to \Gamma$ is topologically equivalent to $f_0 : \mathcal{J} \to \mathcal{J}$, where $\mathcal{J} = \mathbb{T}$ is Julia set of f_0. On the other hand, $\sigma : \Gamma \to \Gamma$ is topologically equivalent to the Markov map $\rho : \Lambda \to \Lambda$ where $\Lambda = \mathbb{T}$ is the limit set of the ideal triangle group $\mathcal{G}$.

Recall in section two we have given a homeomorphism $\mathcal{E} : \mathbb{T} \to \mathbb{T}, 1 \mapsto 1$. Here we will use it to glue two dynamical systems together.

Proposition 4.20. *There is a unique orientation-preserving homeomorphism $\mathcal{E} : \mathbb{T} \to \mathbb{T}$ that conjugates ρ and f_0 on $\mathbb{T}$.*

Proof. It suffices to prove the uniqueness here. We can consider the orientation-preserving automorphism of $\mathbb{T}$ commuting with $f_0 : \mathbb{T} \to \mathbb{T}$ and fixing 1. One can see it must be identity map by consider the images of π, $\frac{\pi}{2}$, ..., and check they are all fixed point one by one. □

We can summarize this discussion as follows.

(1) We have two conformal dynamical systems

$$\rho : \overline{\mathbb{D}} \setminus \text{int}\Pi \to \overline{\mathbb{D}}, \text{ and } f_0 : \hat{\mathbb{C}} \setminus \mathbb{D} \to \hat{\mathbb{C}} \setminus \mathbb{D}.$$

We also have a mating tool, the homeomorphism $\mathcal{E} : \mathbb{T} \to \mathbb{T}$ which conjugates ρ on the limit set and f_0 on the Julia set.

(2)($Topological\ mating$). Define $X = \overline{\mathbb{D}} \vee_{\mathcal{E}} (\hat{\mathbb{C}} \setminus \mathbb{D})$, $Y = X \setminus \text{int}\Pi$, so X is a topological sphere, and Y is a closed Jordan disc in X. The well-defined topological map $\eta := \rho \vee_{\mathcal{E}} f_0 : Y \to X$ is the topological mating between ρ and f_0

(3)($Conformalmating$). The two Riemann uniformization, ψ^{in} and ψ^{out}, glue together into a homeomorphism

$$H : (X, Y) \to (\hat{\mathbb{C}}, \overline{\Omega})$$

which is conformal outside $H^{-1}(\Gamma)$ and which conjugates η to σ.

(4)($Uniqueness\ of\ conformal\ mating$). There is only one conformal structure on X compatible with the standard structure on $X \setminus H^{-1}(\Gamma)$. Indeed, one can see this by the comformal removability of Γ.

Now we come back to the proof of Theorem 1.1. The first part follows from theorem 4.10. and section 4.3. The second statement is the content of section 4.4.

References

[1] Seung-Yeop Lee, Mikhail Lyubich, Nikolai G. Makarov, and Sabyasachi Mukherjee. Dynamics of Schwarz reflections: the mating phenomena. *arXiv e-prints*, pages 1–24, November 2018.

[2] Seung-Yeop Lee and Nikolai Makarov. Topology of quadrature domains. *Journal of the American Mathematical Society*, 29:334,335,347,348, 07 2013.

[3] Norbert Steinmetz. The formula of riemann-hurwitz and iteration of rational functions. *Complex Variables and Elliptic Equations*, 22:203–206, 09 1993.

Algebraic-Analytic Correspondence

Xingzhu Fang and Haoda Li

May, 2022

Abstract

We introduce the classical GAGA theorems of Serre along with some applications and generalizations of it.

Contents

1 Introduction and Statement

GAGA is a general principle in geometry connecting the seemingly unrelated algebraic geometry and analytic geometry. We usually use *GAGA-type result* to address results involving a comparison between objects in analytic geometry and their analogs in algebraic geometry. On one hand, this type of results allow us to introduce analytic methods into algebraic geometry over $\mathbb{C}$ (or a general char $= 0$ field via Lefschetz principle), e.g. Kodaira's original proof of Kodaira vanishing; on the other hand, algebraicity of analytic objects can sometimes be of great use to address problems of analytic nature.

To begin with, we need to give $X(\mathbb{C})$ the structure of complex analytic variety for any scheme X locally of finite type over $\mathbb{C}$.

Definition 1.1 (Analytification)**.** The functor sending any complex analytic space to the set of morphisms to X as ringed spaces over $\mathbb{C}$ is representable and represented by X^{an}, the analytification of X, with universal morphism $h : X^{\text{an}} \to X$.

Remark 1.2. By evaluating the functors at $\mathbb{C}$, the points of X^{an} is identified with $X(\mathbb{C})$. Also note that analytification functor is fully faithful by Yoneda lemma.

Remark 1.3. After showing the flexibility of representability under taking closed subschemes, open subsets and product space, one is reduced to analytify $\mathbb{A}^1_{\mathbb{C}}$ as $\mathbb{C}$, while both sides assigns a complex analytic space to the globally defined holomorphic functions on it.

Remark 1.4. The morphism h is flat: we may simply observe that the stalks of $\mathcal{O}_X$ and $\mathcal{O}_{X^{\text{an}}}$ at the same point are local Noetherian with the same completion.

Then we have to analytify sheaves.

Definition 1.5. For any sheaf of $\mathcal{O}_X$ modules $\mathcal{F}$, define its analytification to be

$$\mathcal{F}^{\text{an}} := h^{-1}\mathcal{F} \otimes_{h^{-1}\mathcal{O}_X} \mathcal{O}_{X^{\text{an}}}$$

Remark 1.6. Recalling the famous theorem of Oka that $\mathcal{O}_{X^{\text{an}}}$ is coherent, analytification sends coherent sheaves to coherent sheaves.

Now we can state the main theorem.

Theorem 1.7 (GAGA)**.** *X is a **proper** scheme over $\mathbb{C}$. Analytification functor gives an equivalence between the categories of coherent sheaves over X and X^{an}, inducing isomorphisms on cohomology groups.*

2 The Proof

The theorem is reduced to projective case by Chow's lemma and induction on dimension of support.The projective case follows from the case $X = \mathbb{P}^r_{\mathbb{C}}$ by simply considering ideal sheaves.So we assume X is the projective space in the following paragraphs.The proof is divided into three parts.

Theorem 2.1 (GAGA, part 1)**.** *For every coherent sheaf $\mathcal{F}$ on X, and every integer $q \geq 0$, the pullback along h*

$$h^* : H^q(X, \mathcal{F}) \to H^q(X^{\text{an}}, \mathcal{F}^{\text{an}})$$

is bijective.

Proof. Using Hilbert's syzygy and the exact sequence

$$0 \to \mathcal{O}_X(n-1) \to \mathcal{O}_X(n) \to \mathcal{O}_H(n) \to 0$$

for and hyperplane H in X to do induction on dimension and n, we are reduced to the case $\mathcal{F} = \mathcal{O}_X$. In this case, H^0 consists of constant functions by Liouville's theorem (on analytic side) and properness (on algebraic side). All higher cohomologies vanishes by Hodge theory (on analytic side) and standard computation (on algebraic side). □

Theorem 2.2 (GAGA, part 2)**.** *If $\mathcal{F}$ and $\mathcal{G}$ are two coherent sheaves on X, then every analytic homomorphism from $\mathcal{F}^{\mathrm{an}}$ to $\mathcal{G}^{\mathrm{an}}$ comes from a unique algebraic homomorphism from $\mathcal{F}$ to $\mathcal{G}$. In other words, the natural map*

$$\mathrm{Hom}_{\mathcal{O}_X}(\mathcal{F},\mathcal{G}) \to \mathrm{Hom}_{\mathcal{O}_{X^{\mathrm{an}}}}(\mathcal{F}^{\mathrm{an}},\mathcal{G}^{\mathrm{an}})$$

is bijective.

Proof. It suffices to show $\mathcal{H}om(\mathcal{F},\mathcal{G})^{\mathrm{an}} \cong \mathcal{H}om(\mathcal{F}^{\mathrm{an}},\mathcal{G}^{\mathrm{an}})$. By taking stalks, this reduces to a general fact for flat ring map. □

Theorem 2.3 (GAGA, part 3)**.** *For every coherent analytic sheaf $\mathcal{F}$ on X^{an}, there exists a coherent algebraic sheaf $\mathcal{G}$ on X such that $\mathcal{G}^{\mathrm{an}}$ is isomorphic to $\mathcal{F}$. Moreover, such $\mathcal{G}$ is unique, up to unique isomorphism.*

Proof. Induction on the dimension of X. It suffices to show $\mathcal{F}(n)$ is globally generated for n large, as $\mathcal{O}_X(-n)$ is algebraic and $\mathcal{F}$ is exhibited as an algebraic kernel. Take arbitrary hyperplane H and n large enough, forming corresponding exact sequence

$$0 \to \mathcal{K}(n) \to \mathcal{F}(n-1) \to \mathcal{H}(n) \to 0$$

$$0 \to \mathcal{H}(n) \to \mathcal{F}(n) \to \mathcal{F}_H(n) \to 0.$$

By induction hypothesis, $\mathcal{K}$ and $\mathcal{F}_H$ are algebraic since they supports on H. Hence for n large enough, $H^2(X^{an},\mathcal{K}(n)) = H^1(X^{an},\mathcal{F}_H(n) = 0$, so $H^1(X,\mathcal{F}(n-1)) \twoheadrightarrow H^1(X^{an},\mathcal{H}(n)) \twoheadrightarrow H^1(X^{an},\mathcal{F}(n))$. By a theorem of Cartan, coherent sheaves have finite dimensional cohomologies on compact complex manifolds, so the dimension of $H^1(X^{an},\mathcal{F}(n))$ is eventually constant and all the surjective arrows become isomorphisms. In conclusion, we have $H^0(X^{an},\mathcal{F}(n)) \twoheadrightarrow H^0(X^{an},\mathcal{F}_H(n))$. Take n large enough such that $\mathcal{F}_H$ is globally generated at a point x, then so is $\mathcal{F}$ at x by Nakayama lemma. Varying $x \in H$ and use compactness now completes the proof. □

3 Applications and Generalizations

3.1 Chow's theorem

Theorem 3.1 (Chow's theorem)**.** *Every closed analytic subset of projective space is algebraic.*

Proof. This is a direct corollary of Theorem 1.7. Let X be a projective space, and Y be a closed analytic subset of X. Recall that the sheaf $\mathcal{H}_Y = \mathcal{H}_X/\mathcal{I}(Y)$ is a coherent analytic sheaf on X^{an}, thus there is a coherent algebraic sheaf $\mathcal{F}$ on X with $\mathcal{H}_Y = \mathcal{F}^{\mathrm{an}}$ by Theorem 1.7. Therefore

$$Y = \mathrm{Supp}\,\mathcal{H}_Y = \mathrm{Supp}\,\mathcal{F}^{\mathrm{an}} = \mathrm{Supp}\,\mathcal{F}$$

is Zariski-closed. □

Chow's theorem 3.1 leads to some fundamental comparison results:

Corollary 3.2. *If X is an algebraic variety, every compact analytic subset X' of X is algebraic.*

Proof. Let Y be a projective variety, U a Zariski-open dense subset of Y, and $f : U \to X$ a surjective regular map whose graph Γ is Zariski-closed in $X \times Y$. Now $\Gamma' := \Gamma \cap (X' \times Y)$ is compact, since X' and Y are compact and Γ is closed. Therefore the image Y' of the projection from Γ' to Y is also compact. But $Y' = f^{-1}(X')$, thus Y' is an analytic subset of U hence of Y, therefore Y' is a Zariski-closed subset of Y by Chow's theorem 3.1. From this we see that $X' = f(Y')$ is Zariski-closed in X. □

Corollary 3.3. *Every holomorphic map f from a compact algebraic variety X to an algebraic variety Y is regular.*

Proof. Let Γ be the graph of f, which is a compact analytic subset of $X \times Y$ since f is holomorphic. Applying Corollary 3.2 now completes the proof. □

Remark 3.4. Combining Corollary 3.3 with Riemann existence theorem, we see that the category of compact Riemann surfaces is equivalent with the category of projective complex algebraic curves.

3.2 Comparison of coverings

Theorem 3.5 (Grothendieck's Riemann existence theorem)**.** *Let X be a $\mathbb{C}$-scheme locally of finite type. The functor*

$$\Phi : \mathbf{F\acute{E}t}_X \to \mathbf{F\acute{E}t}_{X^{\mathrm{an}}}, \ (f : X' \to X) \mapsto (f^{\mathrm{an}} : X'^{\mathrm{an}} \to X^{\mathrm{an}})$$

induces an equivalence between the categories of finite étale coverings of X and X^{an}.

proof (sketch). The proper case follows (more or less) directly from the main theorem 1.7, via the sheaf-theoretic description of (finite) coverings. And since morphisms are algebraizable, the functor Φ is fully faithful. It then suffices to prove Φ is essentially surjective. For this, through a long march of (step-wise) straight and easy reductions, we may reduce to the case where X is regular affine.

We now assume that $X = \operatorname{Spec} A$ is connected, affine and regular. Note that by Chow's lemma, there exists a compactification $j : X \to P$ of X such that P is proper and j is a dominant open immersion. Then resolve the singularities of P by taking blow ups of points in $P \setminus X$ to obtain a proper regular scheme R over P. By resolution of singularities over $\mathbb{C}$, we have a dominant open immersion $k : X \to R$ such that $j = r \circ k$.

Now suppose the finite étale covering $\mathcal{X}' \to X^{\mathrm{an}}$ can be extended to a finite covering $\mathcal{R}' \to R^{\mathrm{an}}$. Then by proper case, there exists a finite covering $R' \to R$ such that $R'^{\mathrm{an}} \simeq \mathcal{R}'$. Let $X' = R'|_X$, then

$$X'^{\mathrm{an}} = R'|_X^{\mathrm{an}} \simeq R'^{\mathrm{an}}|_{X^{\mathrm{an}}} \simeq \mathcal{R}'|_{X^{\mathrm{an}}} = \mathcal{X}'.$$

It thus remains to show that $\mathcal{X}'$ can be extended to R^{an}. This problem is local on $R^{\text{an}} \setminus X^{\text{an}}$.

For an $x \in R^{\text{an}} \setminus X^{\text{an}}$, since X^{an} and R^{an} are regular, there exists an open neighbourhood $V \subset R^{\text{an}}$ of x with a biholomorphic map

$$\phi : V \overset{\simeq}{\to} \mathbb{C}^n, x \mapsto 0, \phi(V \cap (R^{\text{an}} \setminus X^{\text{an}})) = Z(x_1, \ldots, x_p) \subset \mathbb{C}^n$$

where $p = \text{codim}_{R^{\text{an}}} R^{\text{an}} \setminus X^{\text{an}}$. Let $U = \mathbb{C}^n$ and $U_0 = \mathbb{C}^n - Z(x_1, \ldots, x_p) = (\mathbb{C} \setminus \{0\})^p \times \mathbb{C}^{n-p}$. Now recall that there is an equivalence between the category $\mathbf{F\acute{E}t}_U$ (resp. $\mathbf{F\acute{E}t}_{U_0}$) of finite étale coverings of U (resp. U_0) and the category $\mathbf{FTopCov}_U$ (resp. $\mathbf{FTopCov}_{U_0}$) of finite topological covers of U (resp. U_0). A careful topological check then completes the proof. □

As a direct consequence, we have:

Corollary 3.6. *Let K be a number field and X be a smooth proper scheme over K. Then the profinite completion of the fundamental group of $(X \times_K \mathbb{C})^{\text{an}}$ does not depend on the choice of the embedding $K \hookrightarrow \mathbb{C}$.*

Proof. From Grothendieck's Riemann existence theorem 3.5 and Grothendieck's Galois theory formalism we see that, the profinite completion of the fundamental group of $(X \times_K \mathbb{C})^{\text{an}}$ is isomorphic to the étale fundamental group of $X_{\overline{K}}$, thus does not depend on the choice of the embedding $K \hookrightarrow \mathbb{C}$ in particular. □

3.3 Riemann-Hilbert Correspondence

3.3.1 A Baby Version

Let X be a complex manifold or smooth algebraic variety over $\mathbb{C}$. Riemann-Hilbert correspondence is concerning the relation between the following two categories.

• **Loc**(X) A local system L is a locally free $\mathbb{C}_X$ sheaf of finite rank, where $\mathbb{C}_X$ is the constant sheaf with value in $\mathbb{C}$.

• **Conn**(X) An integrable connection M consists of a vector bundle M on X and a flat connection on M.

We have the following famous theorem.

Theorem 3.7. *X is a complex manifold, then the category of local systems and integrable connections on X are naturally equivalent.*

Proof. We will construct the equivalence in both directions.

Let L be a local system on X, then we simply assign $M = L \otimes_{\mathbb{C}_X} \mathcal{O}_X$ and the connection is induced by the exterior derivative $d : \mathcal{O}_X \to \Omega^1_X$.Obviously M is an integrable connection. On the other direction, for any integrable connection M with structure map $\nabla : M \to M \otimes_{\mathcal{O}_X} \Omega^1_X$, we assign the sheaf of horizontal sections $L = M^{\nabla} = \{s \in M : \nabla s = 0\}$ .It remains to show L is locally free over $\mathbb{C}_X$ with the same rank as M.A section of L is the same as a integrable submanifold of the total space of M, hence the result follows from Frobenius' theorem, see Chapter 19 of [10]. □

Next we look at the relation between integrable connections on smooth X and its analytification X^{an}.In the projective case, GAGA gives an equivalence between the two categories, after identifying connections in a categorical way in both categories(notice that ∇ is not a morphism of coherent sheaves).For general X, we have to carefully specify so called regular connections.

When X is a smooth curve, we consider it's smooth compactifiction $\overline{X}$ and view points in $\overline{X}-X$ as "singularities". We assume $\{p\}=\overline{X}-X$ for simplicity. A meromorphic connection M is informally an integrable connection with possible pole at p.Write M in coordinates near p, solving horizontal section of the connection is the same as solving an ordinary differential equation in complex variables with possible pole at p. ODE tells us the solution is at least meromorphic (i.e. not essential singularity at p) iff the equation has pole at p of at most order one. In this case, the meromorphic connection is called regular. Any connection on X naturally extends to a meromorphic connection on $\overline{X}$, hence it's called regular if the extended one is. For higher dimensional smooth varieties/manifolds, regularity is defined to be tested by curves.

The upshot is the following theorem.

Theorem 3.8 (Deligne [4]). *X smooth algebraic variety over $\mathbb{C}$, then analytification induces an equivalence* $\mathbf{Conn}^{\mathrm{reg}}(X)\cong\mathbf{Conn}(X^{\mathrm{an}})$.

Proof. Fix a smooth compactification $\overline{X}$ of X with complementary a divisor, whose existence is guaranteed by Hironaka's resolution of singularity.

The proof divides into three steps by the diagram below.

$$\begin{tikzcd}
\mathbf{Conn}^{\mathrm{reg}}(\overline{X},D) \arrow[r] \arrow[d] & \mathbf{Conn}^{\mathrm{reg}}(\overline{X}^{\mathrm{an}},D^{\mathrm{an}}) \arrow[d] \\
\mathbf{Conn}^{\mathrm{reg}}(X) \arrow[r] & \mathbf{Conn}(X^{\mathrm{an}})
\end{tikzcd}$$

The left vertical arrow is an equivalence by definition. Observing that being regular is equivalent in algebraic and analytic setting, the upper horizontal arrow is ensured by categorical characterization of connections and GAGA principle (note here we uses properness of $\overline{X}$). The right vertical arrow is called Deligne's Riemann Hilbert correspondence and is of analytic nature.The crucial essentially surjective part is actually local (after using a technical assumption to make the choice unique) and use the interpretation of local systems as representation of fundamental group to reduce to a explicit calculation of matrices. For a complete proof, see [4] or [9]. □

3.3.2 The Complete Statement

We have to explain lots of definitions.The reader should notice they are just glueing the baby version along some stratification.

$D_c^b(X)$ is the full subcategory of the (bounded) derived category of $\mathbb{C}_{X^{an}}$ modules, with cohomologies being constructible sheaves in the following sense.

Definition 3.9 (Constructible Sheaf)**.** A sheaf $\mathcal{F}$ of $\mathbb{C}_{X^{an}}$ modules is called constructible if there exists a smooth stratification of X, such that $\mathcal{F}$ is locally constant of finite rank over $\mathbb{C}_{X^{an}}$ on each stratum.

Remark 3.10. Note any constructible sheaf is generically a local system.

Definition 3.11. (Perverse sheaves) We define the perverse t-structure on $D^b_c(X)$ by:

$$D^{\leq 0} := \{K \in \mathrm{Ob}(D^b_c(X)) | \, \mathrm{Supp}(\mathscr{H}^i K) \leq -i, \forall i \in \mathbb{Z}\}, D^{\geq 0} := \mathbb{D}_X D^{\leq 0}.$$

Define the category of perverse sheaves to be the heart of perverse t-structure:

$$\mathbf{Perv}(X) := (D^b_c(X))^{\heartsuit} = D^{\leq 0} \cap D^{\geq 0}.$$

We turn to the differential side. Let X be an smooth algebraic variety/complex manifold.

Definition 3.12. Define filtered ring (D_X, F) by

$$F_l D_X = \{P \in \mathcal{E}nd_{\mathbb{C}} \mathcal{O}_X : [P, f] \in F_{l-1} D_X, \forall f \in \mathcal{O}_X\}, F_{-1} D_X = 0, D_X = \bigcup F_l D_X.$$

Remark 3.13. Local calculation gives $grD_X = \pi_* \mathcal{O}_{T^*X}$.

Remark 3.14. D_X is a coherent sheaf of (noncommutative!) rings, with local rings Noetherian of global dimension $\dim X$. So it makes sense to talk about coherent D_X-modules.

Remark 3.15. An integrable connection is the same as a D_X-module which is coherent over $\mathcal{O}_X$.

Definition 3.16 (Holonomicity)**.** Let $\pi : T^*X \to X$ be the cotangent bundle. For any coherent D_X-module M, take a good filtration F (choice doesn't matter). Define the characteristic variety $Ch(M) = supp(\mathcal{O}_{T^*X} \otimes_{\pi^{-1}grD_X} \pi^{-1}(gr^F M))$. Then we have $dimCh(M) \geq dimX$ and say M is holonomic when the equality holds.

Remark 3.17. A holonomic D_X-module is generically an integrable connection.

$D^b_{rh}(D_X)$ is the full subcategory of the (bounded) derived category of (left) D_X-modules, with regular holonomic cohomologies.

Let's move to the de Rham functor.The motivation comes from solving differential equations.Let $P \in D_X$, $M = D_X/D_X P$, then $\mathcal{H}om_{D_X}(M, \mathcal{O}_X) = \{f \in \mathcal{O}_X : Pf = 0\}$ is the solutions of P. Let X be a smooth algebraic variety and ω_X be the canonical sheaf of X.

Definition 3.18 (De Rham and solution functors)**.**

$$DR_X M := \omega_{X^{an}} \otimes^L_{D_{X^an}} M^{an};$$

$$Sol_X M := R\mathcal{H}om_{D_{X^{an}}}(M^{an}, \mathcal{O}_{X^{an}}).$$

Proposition 3.19. $DR_X \cong Sol_X(\mathbb{D}_X M)[dimX]$.

Remark 3.20. Taking horizontal sections is the same as $DR_X[dimX] \cong R\mathcal{H}om_{D_X}(\mathcal{O}_X, -)$.

Finally, the main theorem is

Theorem 3.21 (Riemann-Hilbert Correspondence)**.** *For a smooth algebraic variety X, the de Rham functor*

$$DR_X : D^b_{rh}(D_X) \to D^b_c(X)$$

is an equivalence of categories with six functor and Verdier dual, transferring standard t-structure on the left to perverse t-structure on the right.

In particular, it induces an equivalence between the category of regular holonomic D-modules and the category of perverse sheaves.

3.4 Comparison of K-Theories

There is also a comparison result between the algebraic K-theory $K(\mathbb{C})$ of $\mathbb{C}$ and the K-theory ku of complex vector bundles, namely the following theorem of Suslin.

Theorem 3.22 (Suslin)**.** *There is a natural map from $K(\mathbb{C})$ to* ku*, inducing an isomorphism $K(\mathbb{C})/n \cong \mathrm{ku}/n$ for each positive integer n.*

Remark 3.23. Recall that Bott periodicity tells us the structure of ku is rather simple, on the other hand, algebraic K-theories are often very complicated. Thus, this theorem of Suslin connects the relatively complicated $K(\mathbb{C})$ to the relatively simple ku, allowing us to get more hold of $K(\mathbb{C})$.

We assume the following result of Suslin. For a proof, see [6].

Theorem 3.24 (Suslin rigidity)**.** *Let (A, I) be a Henselian pair, $n \in A^\times$ be an integer. Then $K(A, I)/n = 0$. i.e., the natural map $K(A) \to K(A/I)$ is an isomorphism modulo n.*

We first recall the construction of the mentioned K-theories in the language of condensed mathematics.

Definition 3.25 (Algebraic K-theory)**.** Let R be a ring. Then the *algebraic K-theory of R* is the groupification of the $\mathbb{E}_\infty$-monoidal anima $(\mathbf{cProj}(R), \oplus)$, where $\mathbf{cProj}(R)$ denotes the groupoid of finitely generated projective R-module (viewed as an anima). We denote the algebraic K-theory of R by $K(R) \in \mathbf{Grp}_{\mathbb{E}_\infty}(\mathbf{Ani}) = \mathbf{Sp}_{\geq 0}$.

Definition 3.26 (Complex topological K-theory)**.** The *topological K-theory of complex vector bundles* is the groupification of the $\mathbb{E}_\infty$-monoidal anima $(\mathbf{Vect}_\mathbb{C}, \oplus)$, where $\mathbf{Vect}_\mathbb{C}$ denotes the topological groupoid of finite dimensional complex vector spaces. We denote the topological K-theory of complex vector bundles by ku $\in \mathbf{Grp}_{\mathbb{E}_\infty}(\mathbf{Ani}) = \mathbf{Sp}_{\geq 0}$.

It is easy to check that

$$\pi_0 \operatorname{Hom}(X, \mathrm{ku}) = (\mathbf{Bun}_\mathbb{C}(X), \oplus)^{\mathrm{gp}},$$

thus Definition 3.26 coincides with the classical definition.

We now make the following generalization of Definition 3.25 for condensed rings.

Definition 3.27. Let $R \in \mathbf{CondRing}$ be a condensed ring. Then the *algebraic K-theory of R* is given by

$$K(R) = (\mathbf{ExDisc} \ni S \mapsto K(R(S))) \in \mathbf{CondSp}_{\geq 0}.$$

Write

$$\mathbf{cProj}(R) = (\mathbf{ExDisc} \ni S \mapsto \mathbf{cProj}(R(S))) \in \mathbf{Mon}_{\mathbb{E}_\infty}(\mathbf{CondAni}),$$

then $K(R) = \mathbf{cProj}(R)^{\mathrm{gp}}$ is its groupification.

From now on, we use $\mathbb{C}$ to denote its corresponding condensed ring. Then the classical algebraic K-theory of complex numbers is now just $K(\mathbb{C})(*)$. The following theorem connects $K(\mathbb{C})$ and ku, indicating that ku is the "non-archimedean" part of $K(\mathbb{C})$.

Theorem 3.28. *We have*

$$\mathbf{cProj}(\mathbb{C}) = \coprod_{n=0}^{\infty} B\mathrm{GL}_n(\mathbb{C}),$$

where $B\mathrm{GL}_n(\mathbb{C})$ denotes the classifying space of $\mathrm{GL}_n(\mathbb{C})$ as a condensed group. The natural map $\mathrm{GL}_n(\mathbb{C}) \to \mathrm{h}(\mathrm{GL}_n(\mathbb{C}))$ induces a map $\mathbf{cProj}(\mathbb{C}) \to \coprod_{n=0}^{\infty} B(\mathrm{h}(\mathrm{GL}_n(\mathbb{C}))) = \mathbf{Vect}_{\mathbb{C}}$, giving a map $K(\mathbb{C}) \to \mathrm{ku}$ after taking groupification, which is an isomorphism after taking solidification. In particular, $K(\mathbb{C})^{\blacksquare} = \mathrm{ku}$.

3.5 Other GAGA Theorems

GAGA results also appears in other types of geometries.

Theorem 3.29 (Formal GAGA)**.** *Let A be an adic Noetherian ring, with defining ideal I. Denote $Y = \operatorname{Spec} A, Y_n = \operatorname{Spec} A/I^n, \hat{Y} = \operatorname{Spf} A$. Let X be a Noetherian scheme, separated and of finite type over Y, with I-adic completion $\hat{X}$. Then completion gives an equivalence between the categories of coherent sheaves over X and $\hat{X}$ with proper support over Y and $\hat{Y}$ respectively, inducing isomorphism on cohomologies (i.e. completion commutes with higher pushforward).*

Proof. The cohomology part is the formal function theorem, see Chapter 29 of [15] or Chapter 3, Section 11 of [8]. The existence part also firstly deal with projective case, proving a coherent sheaf becomes globally generated after twisting enough times in a rather direct way, and then deal with general case by Noetherian induction and Chow's lemma. For the whole account of the proof (and more applications), see [5]. □

Theorem 3.30 (Rigid GAGA)**.** *Let X be a proper scheme over a non-archimedean field K. Then rigid analytification (similarly defined) induces an equivalence between categories of coherent sheaves over X and X^{rig}, inducing isomorphism on cohomologies.*

Proof. The proof carries *mutatis mutandis*, assuming Kiehl's proper mapping theorem (rigid version of Grothendieck coherence or Cartan's theorem). For more details, see [2]. □

We record three more developments of GAGA theorems.

Noticing the difference of the essential techniques in proving different GAGA theorems, Peter Scholze develops condensed mathematics to treat archimedean geometry in a non-archimedean way.The following theorem of Ostrowski is a first glimpse of how the analytification works.

Theorem 3.31 (Ostrowski). *A is any $\mathbb{C}$-algebra, then the standard norm on $\mathbb{C}$ induces a homeomorphism $M^{\mathrm{Berk}}(A) = \mathrm{Hom}(A, \mathbb{C})$, where M^{Berk} is the Berkovich spectrum.*

For any cocomplete closed symmetric monoidal stable ∞-category $\mathcal{C}$, we can assign a natural locale (i.e. pointless topological space) $S(\mathcal{C})$ whose closed subsets correspond to idempotent algebras of $\mathcal{C}$. For any closed subset A, we assign $\mathcal{C}(A) = \mathbf{Mod}_A(\mathcal{C})$ and $\mathcal{C}(U) = \mathcal{C}/\mathcal{C}(S(\mathcal{C}) - U)$, giving rise to a sheaf of ∞-category on the locale $S(\mathcal{C})$, called the structure sheaf. Define a categorified locale to be $(X, \mathcal{C})$, where X is a locale and $\mathcal{C}$ a cocomplete closed symmetric monoidal stable ∞-category, together with a morphism $f : S(\mathcal{C}) \to X$. The structure sheaf of X is the pushforward of that of $S(\mathcal{C})$.

For finite type algebra A over $\mathbb{C}$, define $S(A) = S(D(\mathrm{Liq}_p(A)))$ and a open subset $S(A, A)$ by certain convergence conditions. We have natural maps $S(A, A) \to M^{\mathrm{Berk}}(A) \to \mathrm{Spec}\, A(\mathbb{C})$ and $S(A) \to \mathrm{Spec}\, A^{\mathrm{op}}$.Descend by Zariski topology, any X separated of finite type over $\mathbb{C}$ give rise to $(X(\mathbb{C}), C^{\mathrm{an}}(X))$ and $(X, C^{\mathrm{alg}}(X))$.

Theorem 3.32. *For X proper, there is a natural equivalence of cocomplete closed symmetric monoidal $D(\mathrm{Liq}_p(\mathbb{C}))$-linear stable ∞-categories*

$$C^{\mathrm{an}}(X) \cong C^{\mathrm{alg}}(X).$$

There is also a formal GAGA for good moduli spaces.

Definition 3.33. A quasi-compact and quasi-separated morphism of locally Noetherian algebraic stacks $\phi : X \to Y$ is a good moduli space morphism if

- (ϕ is Stein) the morphism $\mathcal{O}_Y \to \phi_* \mathcal{O}_X$ is an isomorphism, and
- (ϕ is cohomologically affine) the functor $\phi_* : \mathbf{QCoh}(\mathcal{O}_X) \to \mathbf{QCoh}(\mathcal{O}_Y)$ is exact.

A stack is said to have the resolution property if every coherent sheaf has a surjection from a vector bundle.

Theorem 3.34 ([7]). *Suppose $\mathfrak{X} \to \mathrm{Spec}\, A$ is a good moduli space, where A is a complete Noetherian local ring with maximal ideal $\mathfrak{m}$ and $\mathfrak{X}$ is of finite type over $\mathrm{Spec}\, A$. Let $\hat{\mathfrak{X}}$ denote the formal completion of $\mathfrak{X}$ with respect to $\mathfrak{m}$.*

(1) The completion functor $\mathbf{Coh}(\mathfrak{X}) \to \mathbf{Coh}(\hat{\mathfrak{X}})$ is fully faithful.

(2) Suppose $\mathfrak{X}_0 = \mathfrak{X} \times_{\mathrm{Spec}\, A} \mathrm{Spec}\, A/\mathfrak{m}$ has the resolution property. TFAE:

(quot) $\mathfrak{X}$ is the quotient of an affine scheme by GL_n for some n.

(quot') $\mathfrak{X}$ is the quotient of an algebraic space by an affine algebraic group.

The above conditions imply the following equivalent conditions:

(res) $\mathfrak{X}$ has the resolution property.

(res') Every coherent sheaf on $\mathfrak{X}_0$ has a surjection from a vector bundle on $\mathfrak{X}$.

The above conditions imply

(GAGA) The completion functor $\mathbf{Coh}(\mathfrak{X}) \to \mathbf{Coh}(\hat{\mathfrak{X}})$ is an equivalence.

If the unique closed point of $\mathfrak{X}$ has affine stabilizer group then (res) implies (quot'), and if $\mathfrak{X}$ has affine diagonal then (GAGA) implies (res').

GAGA theorem also appears in derived geometry, see [11].

Theorem 3.35. *Let $f : X \to Y$ be a proper morphism of derived complex Artin stacks locally almost of finite presentation. Then the diagram commutes.*

$$\begin{tikzcd}
\mathbf{Coh}^{-}(X) \arrow[r, "(-)^{\mathrm{an}}"] \arrow[d, "f_*"] & \mathbf{Coh}^{-}(X^{\mathrm{an}}) \arrow[d, "f_*^{an}"] \\
\mathbf{Coh}^{-}(Y) \arrow[r, "(-)^{\mathrm{an}}"] & \mathbf{Coh}^{-}(Y^{\mathrm{an}})
\end{tikzcd}$$

Theorem 3.36. *Let X be a proper derived Artin stack locally almost of finite presentation over $\mathbb{C}$. Then the analytification functor induces an equivalence $\mathbf{Coh}(X) \cong \mathbf{Coh}(X^{\mathrm{an}})$.*

References

[1] M. Artin, Algebraization of formal moduli: II. Existence of modifications, *Ann. of Math.* **91** (1970), 88–135.

[2] S. Bosch, *Lectures on formal and rigid geometry*, SFB 478 (preprint; June, 2005), http://www.math1.uni-muenster.de/sfb/about/publ/heft378.ps.

[3] D. Clausen, P. Scholze, *Condensed Mathematics and Complex Geometry*, lecture notes, https://people.mpim-bonn.mpg.de/scholze/Complex.pdf.

[4] P. Deligne, Equations Différentielles à Points Singuliers Réguliers, Springer-Verlag Berlin Heidelberg, 1970.

[5] B. Fantechi; L. Göttsche; L. Illusie; S. L. Kleiman; N. Nitsure; A. Vistoli, *Fundamental algebraic geometry: Grothendieck's FGA explained.* Math. Surv. and Mono., **123**. American Math. Soc., Providence, RI, 2005. x+339 pp.

[6] O. Gabber, K-Theory of Henselian Local Rings and Henselian Pairs, *Algebraic K-Theory, Commutative Algebra, and Algebraic Geometry*, vol. **126** of *Contemp. Math.*, 59–70.

[7] A. Geraschenko, D. Z-Brown, Formal GAGA for good moduli spaces, arXiv: 1208.2882.

[8] R. Hartshorne, *Algebraic Geometry*, Springer-Verlag, New York, 1977.

[9] R. Hotta, K. Takeuchi, T. Tanisaki, D-Modules, Perverse Sheaves, and Representation Theory. English Edition: Birkhäuser Boston, 2008. Progress in Mathematics Vol. 236.

[10] J. M. Lee, Introduction to Smooth Manifolds, Graduate Texts in Mathematics volume 218, Springer Science+Business Media New York 2012.

[11] M. Porta, Derived complex analytic geometry I: GAGA theorems, arXiv: 1506.09042.

[12] M. Raynaud, Géométrie Algébrique et Géométrie Analytique, in *Revêtements Étales et Groupe Fondamental (SGA 1)*, Springer, 1971.

[13] J-P. Serre, Géométrie Algébrique et Géométrie Analytique, *Annales de l'Institut Fourier* **6** (1956), 1–42.

[14] L. Tang, Y. Li and others, Condensed mathematics, unpublished lecture notes, https://www.bananaspace.org/wiki/.

[15] R. Vakil, *The Rising Sea: Foundations of Algebraic Geometry*, unpublished lecture notes, http://math.stanford.edu/∼vakil/216blog/FOAGnov1817public.pdf.

Topology

An introduction to dessins d'enfants

Hang Chen and Haoda Li

May, 2022

Abstract

The ideas about dessins d'enfants are originally outlined by Alexander Grothendieck in his famous program *Esquisse d'un programme*, we introduce some developments of it in this paper. Dessins d'enfants are topological and combinatorial objects that reflects rich geometric and arithmetic information. The first five sections are devoted to give an overview of the theory of dessins d'enfants. They cover topics like Belyi's theorem and the Grothendieck correspondence, along with Galois action on dessins; which shed light on the geometric and arithmetic feature of dessins d'enfants. Examples and applications are given in Section 6 and Section 7, respectively. The prerequisites needed to read the applications in Section 7 are partly covered in the appendices.

Contents

1 Introduction

The theory of dessins d'enfants is considered by Grothendieck as one of the most important discoveries he made in his mathematical career, according to Grothendieck himself in the *Esquisse d'un programme* [3]. As easily seen in the Definition 2.0.1 given in Section 2, such objects are purely of topological and combinatorial data. However, these objects contain rich geometric and arithmetic information. For example, as we shall see in Section 5 and Section 7.3, they yield interesting descriptions of the absolute Galois group $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ and its representations (see Corollary 5.1.4 and Theorem 7.3.2, for example). Heuristically speaking, they lead to comparisons between arithmetic Galois groups and geometric fundamental groups, which are somewhat highly nontrivial since naturally the arithmetic fundamental group is an extension of the arithmetic Galois group and the geometric fundamental group. Moreover, the theory of dessins d'enfants is considered as the starting point of the study of anabelian geometry.

An important feature of dessins d'enfants is that they correspond to étale coverings of $\mathbb{P}^1_{\overline{\mathbb{Q}}}\backslash\{0,1,\infty\}$, in particular the absolute Galois group $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ acts on them. More precisely, the category of dessins d'enfants is equivalent to the category $\mathbf{F\acute{E}t}_{\mathbb{P}^1_{\overline{\mathbb{Q}}}\backslash\{0,1,\infty\}}$ of finite étale coverings of $\mathbb{P}^1_{\overline{\mathbb{Q}}} \setminus \{0,1,\infty\}$, which is equivalent to the category $\mathbf{FTopCov}^{\mathrm{Gal}(\mathbb{C}/\overline{\mathbb{Q}})}_{\mathbb{P}^1_{\mathbb{C}}\backslash\{0,1,\infty\}}$ of finite topological ramified coverings of $\mathbb{P}^1_{\mathbb{C}}$ defined over $\overline{\mathbb{Q}}$ and branching only at some points in $\{0,1,\infty\}$ (by Grothendieck's Riemann existence theorem and that $\mathbb{P}^1_{\mathbb{C}}\backslash\{0,1,\infty\}$ is integral as a scheme, see [14] for details). This is called the Grothendieck correspondence, a topological constructive proof is given in Section 3.2, another proof via the cartographical group is given in Section 3.3, see Section 3 for details.

Using the topological and combinatorial datum, we may define the automorphism group $\mathrm{Aut}(\mathcal{D})$ and the monodromy group $\mathrm{Mon}(\mathcal{D})$ of a dessin $\mathcal{D}$, they coincide with the automorphism group $\mathrm{Aut}(f_{\mathcal{D}})$ and the monodromy group $\mathrm{Mon}(f_{\mathcal{D}})$ of the ramified cover $f_{\mathcal{D}}$ obtained via the Grothendieck correspondence, see Section 4 for details. They are both interesting invariants of dessins, as they are direct to compute and invariant under the action of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, in particular they can be used to check whether two dessins are in the same Galois orbit. It is also worth mentioning that $\mathrm{Aut}(\mathcal{D})$ is in fact isomorphic to the centralizer $Z(\mathrm{Mon}(\mathcal{D}))$ of $\mathrm{Mon}(\mathcal{D})$ in S_n.

There are many interesting applications of the theory of dessins d'enfants. For example, it yields a new perspective of Davenport's bound on $f^3 - g^2$, leading to a clean and short new proof of Zannier's main result in [13]. Also, we may use Belyi's theorem to show that the *abc*-conjecture implies Faltings theorem. For details, see Section 7.

2 The definition of Dessin d'enfants

Definition 2.0.1. A *dessin d'enfant*, or simply a dessin, is a pair $(X, \mathcal{D})$, where X is an oriented compact topological surface, and $\mathcal{D} \subset X$ is a finite graph such that:

(1) $\mathcal{D}$ is connected.

(2) $\mathcal{D}$ can be put a bipartite structure, namely the vertices can be marked with two distinct marks in such a way that the direct neighbors of any given vertex are all of the opposite mark.

(3) $X \setminus \mathcal{D}$ is the union of finitely many topological discs, which is called the faces of $\mathcal{D}$.

Definition 2.0.2. A clean dessin is a dessin where all the vertices with a particular mark have degree 2. And it's to be understood that in the definition of the dessin the condition (2) is removed. For a graph satisfying the conditions (1) and (3), a dessin is associated by giving all the vertices the same mark and placing a new vertex with a different mark in the middle of each edge.

Note that a dessin is more than a mere abstract graph, since it's equipped with a certain embedding in a given topological surface. The genus of a dessin $(X, \mathcal{D})$ is simply the genus of the topological surface X. When the topological surface X is clear from the context, we will denote the dessin simply by $\mathcal{D}$.

3 Grothendieck correspondence

3.1 Belyi's theorem

There's a one-one correspondence between dessin d'enfants and Belyi functions, which is the main reason why we consider the theory of dessin d'enfants. Belyi's celebrated theorem states that a Riemann surface with a Belyi's function on it if and only if it's defined over $\overline{\mathbb{Q}}$, and Belyi's functions are defined over $\overline{\mathbb{Q}}$.

Definition 3.1.1. A morphism $f : X \to \mathbb{P}^1_{\mathbb{C}}$ all of whose critical values lie in $\{0, 1, \infty\}$ is called a *Belyi morphism*. We call f clean if all the ramification orders over 1 are equal to 2.

Definition 3.1.2. If X is an algebraic curve defined over $\overline{\mathbb{Q}}$ and f is a Belyi function on it, we call the couple (X, f) a *Belyi pair*. Two Belyi pairs are said to be *equivalent* if they are equivalent as ramified coverings.

Theorem 3.1.3 (Belyi's theorem)**.** *Let S be a compact Riemann surface, then the following statements are equivalent:*

(a) X is defined over $\overline{\mathbb{Q}}$.

(b) There exists a non-constant holomorphic function $f : X \to \mathbb{P}^1_{\mathbb{C}}$ all of whose branch values lie in $\{0, 1, \infty\}$, i.e., a Belyi function.

Full proof is given in appendix, here we just prove $(a) \Rightarrow (b)$ as the construction is useful in creating some dessins. And a partly proof of $(b) \Rightarrow (a)$(different from the approach taken in appendix) is given based on a main criterion just to show maybe some insights of the theorem.

Lemma 3.1.4. *Let f be a morphism from Riemann surface S to $\mathbb{P}^1_{\mathbb{C}}$ and all critical values of f lie in $\overline{\mathbb{Q}} \cup \{\infty\}$. Then there exists a function $P : \mathbb{P}^1_{\mathbb{C}} \mapsto \mathbb{P}^1_{\mathbb{C}}$ such that $g = P \circ f$ is a Belyi function. Moreover, P can be chosen to be a polynomial.*

Proof. A general observation is given first that the following equation holds.

$$\text{Branch}(g \circ f) = \text{Branch}(g) \cup g(\text{Branch}(f)) \tag{3.1.4.1}$$

Step 1: Constructing a function only ramifies in rational numbers.

Let S be the set of all critical values of f and all their conjugates under $\text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. If $S \in \mathbb{Q}$, go to the next step. If not, set $m_1(z) = \prod_{s\in S}(z - s) \in \mathbb{Q}[z]$. The branch values of $m_1 \circ f$ are contained in $S_1 = m_1(\{\text{roots of } m_1' \cup \{0, \infty\})$ by equation 3.1.4.1. Since $m_1' \in \mathbb{Q}[z]$, S_1 contains all the conjugates of S_1 under $\text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. Thus, set $m_2(z) = \prod_{s\in S_1}(z-s) \in \mathbb{Q}[z]$, then we have $\deg m_2 \leq \deg m_1' < \deg m_1$. So we can construct $m_i \in \mathbb{Q}[z]$ recursively, and the degrees of m_i decrease successively until for some $n, \deg(m_n) = 0$. Let $h = m_{n-1} \circ m_{n-2} \circ \cdots \circ m_1 \circ f$, then all the branch values are contained in $\mathbb{Q}$.

Step 2: If there's a branch value $m/(m + n) \in \mathbb{Q} \cap (0, 1)$

Consider the *Belyi polynomial*

$$P_{m,n}(z) = \frac{(m + n)^{m+n}}{m^m n^n} z^m(1 - z)^n \tag{3.1.4.2}$$

which transforms both 0 and 1 to 0, and $m/(m + n)$ to 1. It ramifies only at the points $x = 0, 1, m/(m + n), \infty$. Hence, $\text{Branch}(P_{m,n} \circ h) = \text{Branch}(h) - \{m/(m + n)\}$. We can reduce branch values by one in $\mathbb{Q} \cap (0, 1)$ by composing h with a Belyi polynomial.

Step 3: If there are branch values in $\mathbb{Q} - [0, 1]$

Composing Möbius transformation $M(z) = 1/z$ or $M(z) = 1 - z$ can help to transform some branch values to $[0, 1]$.

After finishing step 1, we can use step 2 and step 3 alternately to obtain the wanted Belyi function. □

Proof of (a)⇒(b). Let $S = S_F$ be a compact Riemann surface where $F(X,Y) = p_0(X)Y^n + p_1(X)Y^{n-1} + \cdots + p_n(X)$. Considering the function $\mathbf{x}$: $(x,y) \mapsto x$, then $\mathrm{Branch}(\mathbf{x}) \subset \{\text{roots of } p_0\} \cup \{\text{first ordinates of common roots of } F_Y \text{ and } F\}$. By Bézout's theorem, the branch values of $\mathbf{x}$ are all in $\overline{\mathbb{Q}}$. Then by Lemma 3.1.4, the proof of this half is done. □

This approach can be applied to construct some Belyi morphisms, and then by Grothendieck correspondence some dessins.

For (b)⇒(a), a criterion for definability over $\overline{\mathbb{Q}}$ is given without proof.

Theorem 3.1.5. *Let S be a compact Riemann surface, the following conditions are equivalent:*

- *S is defined over $\overline{\mathbb{Q}}$.*
- *The family $\{S^\sigma\}_{\sigma \in \mathrm{Gal}(\mathbb{C}/\mathbb{Q})}$ contains only finitely many isomorphism classes of Riemann surfaces.*

Proof of (a)⇒(b) based on the criterion. From this criterion the proof of Belyi theorem is quickly accessed. If $f : S \to \mathbb{P}^1_{\mathbb{C}}$ is a Belyi function, then for arbitrary $\sigma \in \mathrm{Gal}(\mathbb{C})$, the degree of $f^\sigma : S^\sigma \to \mathbb{P}^1_{\mathbb{C}}$ unchanged. σ acts trivially on set $\{0, 1, \infty\}$, then f and f^σ have the same branch values. So the monodromies of f^σ have only finite possibilities when σ goes through $\mathrm{Gal}(\mathbb{C})$, which shows f^σ are in finite equivalent classes of coverings. Then apply theorem 3.1.5 to finish the proof. □

3.2 A proof via a topological construction

In the original work of Grothendieck, he only considered the clean dessins and proves the correspondence of the clean dessins and clean belyi pairs. But the proof of construction also works for the general situations. In the next section, we will introduce another proof of Grothendieck correspondence of clean dessins d'enfants by considering action of cartographical group on the set of flags of a dessin, which can also be modified to suit non-clean case.

We first construct a dessin from a Belyi pair. For a given Belyi pair (S, f), let $\mathcal{D}_f = f^{-1}[0,1]$, where $[0,1]$ is the segment of the real line on $\mathbb{P}^1_{\mathbb{C}}$, the vertices of the graph are $f^{-1}(0)$ and $f^{-1}(1)$ with different marks which equip the graph a bipartite structure. Then $(S, \mathcal{D}_f)$ is a dessin.

Proposition 3.2.1. *The dessin $(S, \mathcal{D}_f)$ satisfies the following properties:*

(1) $\mathcal{D}_f$ is a dessin d'enfant.

(2) Each face of the dessin has exactly one point in $f^{-1}(\infty)$.

(3) Each of the sets $f^{-1}[0,1]$, $f^{-1}[0,\infty]$ and $f^{-1}[1,\infty]$ is a union of topological segments. All of them together are the complete set of edges of a triangle decomposition $\mathcal{T}(\mathcal{D}_f)$ of S. This shows $\mathcal{D}_f$ satisfies the condition (1) and (3) of the definition of dessins.

(4) $\deg(f)$ *agrees with the number of edges of* $\mathcal{D}_f$.

(5) The multiplicity of f *at a vertex* v *of* $\mathcal{D}_f$ *coincides with the degree of the vertex. The multiplicity of a point in the set* $f^{-1}(\infty)$ *agrees with half the valency of the face where the point is.*

(6) If f *is clean,* $\mathcal{D}_f$ *is clean.*

Next, we are going to construct a Belyi pair from a given dessin $(X, \mathcal{D})$, including equipping the topological surface X with a Riemann surface structure. Suppose the two marks on the dessin are $\circ$ and $\bullet$, and the edges of the dessin are labelled with numbers from 1 to n.
Step 1: We shall first construct a triangle decomposition $\mathcal{T} = \mathcal{T}(\mathcal{D})$ of X associated to $\mathcal{D}$. Choose a *centre* in each of the faces of $\mathcal{D}$ and mark them with $\star$. For each pair (v, A) where v is a vertex in the boundary of a face A, draw a segment γ_A^v that starts at v and ends at the centre of A, and it's not allowed to meet any edge except at v it self. Then the γ_A^v divide X into some triangles, satisfying following properties:

- Every triangle contains one vertex of each type $\circ, \bullet$ and $\star$.
- For an edge numbered j, there are two triangles T_j^+ and T_j^- of which j is a common edge. A triangle of which j is an edge is denoted by T_j^+ if the circuit $\circ \to \bullet \to \star \to \circ$ follows the positive orientation of δT_j^+, and by T_j^- otherwise.

Step 2: For a triangle T_j^+, we can construct a homeomorphism f_j^+ from the triangle T_j^+ to $\bar{\mathbb{H}}^+ := \mathbb{H} \cup \mathbb{R} \cup \infty$ satisfying the following condition:

$$f_j^+ : \begin{cases} \partial T_j^+ \longrightarrow & \mathbb{R} \cup \infty \\ \circ \longmapsto & 0 \\ \bullet \longmapsto & 1 \\ \star \longmapsto & \infty \end{cases} \tag{3.2.1.1}$$

And similarly construct f_j^- from the adjacent triangle T_j^- to $\bar{\mathbb{H}}^- := \hat{\mathbb{C}} \setminus \mathbb{H}$ that coincides with f_j^+ in the intersection $T_j^+ \cap T_j^-$ and verifies also equation 3.2.1.1. To ensure the compatibility condition on the boundary, we can first construct the homomorphisms on ∂T_j^δ $(\delta \in \{+, -\})$ and extend them to T_j^δ. Since these two homeomorphisms agree on $T_j^+ \cap T_j^-$, we say they can be *glued together*.

Glue together the collection of homeomorphisms $f_j^\pm : T_j^\pm \to \bar{\mathbb{H}}^\pm$ to construct a continuous function $f_{\mathcal{T}(\mathcal{D})} : X \to \hat{\mathbb{C}}$ whose restriction on $X^* = X \setminus f_{\mathcal{T}(\mathcal{D})}^{-1}\{0, 1, \infty\} \to \hat{\mathbb{C}} \setminus \{0, 1, \infty\}$ is a topological covering. So X^* inherits from $\hat{\mathbb{C}}$ a unique Riemann surface structure such that $f_j^\pm$ is holomorphic. Furthermore X can be converted into a compact Riemann surface denoted $S_{\mathcal{T}(\mathcal{D})}$ such that $f_{\mathcal{T}(\mathcal{D})}$ becomes a morphism from $S_{\mathcal{T}(\mathcal{D})}$ to $\hat{\mathbb{C}}$.

Moreover, it can be checked that for a different triangle decomposition $\mathcal{L}(\mathcal{D})$, $(S_{\mathcal{L}(\mathcal{D})}, f_{\mathcal{L}(\mathcal{D})})$ and $(S_{\mathcal{T}(\mathcal{D})}, f_{\mathcal{T}(\mathcal{D})})$ are in the same equivalent class of covering, which means that modulo the equivalence of coverings, the pair $(S_{\mathcal{L}(\mathcal{D})}, f_{\mathcal{L}(\mathcal{D})})$ depends only on the dessin $(X, \mathcal{D})$. Therefore, we shall write $(S_{\mathcal{D}}, f_{\mathcal{D}})$ instead of $(S_{\mathcal{T}(\mathcal{D})}, f_{\mathcal{T}(\mathcal{D})})$.

The following proposition is direct from the above construction procedure.

Proposition 3.2.2. *The pair $(S_{\mathcal{D}}, f_{\mathcal{D}})$ satisfies the following properties:*

(1) $(S_{\mathcal{D}}, f_{\mathcal{D}})$ is a Belyi pair.

(2) The same number equations as in Proposition 6.

(3) $f_{\mathcal{D}}^{-1}([0,1]) = \mathcal{D}$.

It's easy to check the two correspondences given above are mutually inverse.

Theorem 3.2.3. *The two correspondences*

$$\begin{array}{rcl} \{\textit{Equiv. classes of dessins}\} & \longrightarrow & \{\textit{Equiv. classes of Belyi pairs}\} \\ (X, \mathcal{D}) & \longmapsto & (S_{\mathcal{D}}, f_{\mathcal{D}}) \\ (S, \mathcal{D}_f) & \longleftarrow\!\!\dashv & (S, f) \end{array}$$

are mutually inverse.

Definition 3.2.4. A function $R : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$ is called Belyi extending if it satisfies the following conditions:

(1) R is a Belyi function.

(2) R is defined over the rationals.

(3) $R(\{0, 1, \infty\}) \subset \{0, 1, \infty\}$.

If (S, f) is a Belyi pair and R is a Belyi extending, then $R \circ f$ is still a Belyi function. It will be shown in the section 6 that Belyi-extendings can be used to create new invariants of Galois action.

Take $R = 1/4z(1-z)$. If f^c denote $R \circ f$, $\mathcal{D}^c$ denote the dessin corresponding to f^c, then f^c is a clean Beiyi function and therefore $\mathcal{D}^c$ is a clean dessin. Actually, $\mathcal{D}^c$ can be obtained by changing all the vertices of $\mathcal{D}$ white and adding a black vertex in the middle of each edge.

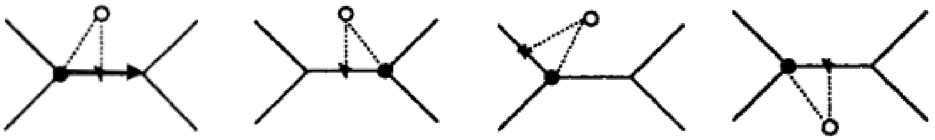
Take $R = 1/z$. Let $\mathcal{D}$ be a clean dessin and f is the corresponding clean Belyi function. The dual dessin formed from the preimages of the line segment $[1, \infty]$ corresponds to the $R \circ f = 1/f$.

3.3 Cartographical group and clean dessin

Definition 3.3.1. The cartographical group C_2 is generated by σ_0, σ_1 and σ_2 with the relations $\sigma_0^2 = \sigma_1^2 = \sigma_2^2 = 1$ and $(\sigma_0\sigma_2)^2 = 1$. The oriented cartographical group C_2^+ is the subgroup of index 2 of C_2 containing all even words of C_2, generated by $\rho_0 = \sigma_1\sigma_0$, $\rho_1 = \sigma_0\sigma_2$ and $\rho_2 = \sigma_2\sigma_1$, with the relations $\rho_1^2 = 1$ and $\rho_0\rho_1\rho_2 = 1$.

Definition 3.3.2. Let $\mathcal{D}$ be a clean dessin with marking as above, and the vertices marked $\bullet$ are of degree 2. Then the *flag set* $F(\mathcal{D})$ is the set of triangles in the triangle decomposition $\mathcal{T}(\mathcal{D})$, the elements of which are called flags, and the *oriented flag set* is the collection of all T_j^+.

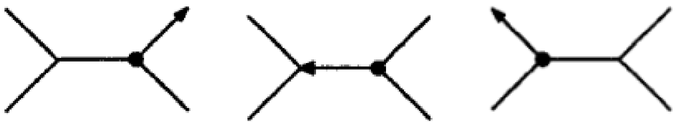
The action of the group C_2 on $F(\mathcal{D})$ is defined as in Figure 1.



(a) F (b) $\sigma_0(F)$ (c) $\sigma_1(F)$ (d) $\sigma_2(F)$

Figure 1: The action of $\sigma_0, \sigma_1, \sigma_2$ on a flag F

Only the group element in C_2^+ preserves the orientation. And the oriented flags can be uniquely determined by a vertex $\circ$ and an edge coming out of the vertex. In this view, the actions given by $\rho_0(F)$, $\rho_1(F)$ and $\rho_2(F)$ are shown by Figure 2.



(a) $\rho_0(F)$ (b) $\rho_1(F)$ (c) $\rho_2(F)$

Figure 2: The action of ρ_0, ρ_1, ρ_2 on a flag F

Definition 3.3.3. Let $B_{F,D}$ be the set of elements of C_2^+ fixing a flag F, which is a subgroup of finite index in C_2^+. Since C_2^+ acts transitively on $F^+(\mathcal{D})$, for any other flag F', $B_{F',D}$ is conjugate to $B_{F,D}$ in C_2^+.

Theorem 3.3.4. *There is a bijection between the isomorphism classes of clean dessins and the conjugacy classes of subgroups of C_2^+ of finite index.*

Proof. By Lemma 3.3.3, a dessin can be associated with a conjugacy class of subgroups of C_2^+ of finite index by the stabilizer of a flag. Then our only task is to construct a dessin from a subgroup of C_2^+ of finite index fitting this correspondence.

If B be the set of elements of C_2^+ fixing a flag F, then the action of C_2 on the flag set is isomorphic to its action on the coset space $H = C_2/B$. So imagine the elements in the coset space $H = C_2/B$ be flags with three vertices marked with $\circ, \bullet, \star$ respectively. A flag is positively oriented if and only if it's contained in C_2^+/B. By the definition of the action of σ_0, two flags have the common $\circ - \star$ edge if and only if they are in the same σ_0-obit. The $\bullet - \star$ and $\circ - \star$ edges can be identified in the similar way. Thus, we can glue the flags together in a unique way such that the dessin obtained corresponds to conjugacy class of B. $\square$

The fundamental group of $\mathbb{P}^1_{\mathbb{C}} - \{0, 1, \infty\}$ is denoted by π_1, generated freely by loops γ_0, γ_1 around 0 and 1. Recall the classical results:

Lemma 3.3.5. *There is a bijection between the conjugacy classes of subgroups of finite index of π_1 and isomorphism classes of finite coverings X of $\mathbb{P}^1_{\mathbb{C}}$ ramified only over $0, 1, \infty$.*

For the clean type, let $\pi_1' = \pi_1/\langle\gamma_1^2\rangle$, then we have the following:

Corollary 3.3.6. *There is a bijection between the conjugacy classes of subgroups of finite index of π_1' and isomorphism classes of finite coverings X of $\mathbb{P}^1_{\mathbb{C}}$ ramified only over $0, 1, \infty$ such that the ramification over 1 is of degree at most 2.*

Theorem 3.3.7. *There is a bijection between the set of equivalent classes of clean dessins and the set of equivalent classes of clean Belyi pairs.*

This theorem is an immediate consequence of Lemma 3.3.6 and Theorem 3.3.4.

Remark 3.3.8. Actually, the actions ρ_1, ρ_2 defined here are consistent with σ_0, σ_1 in the next section, and the argument above can be adopted in the non-clean general case, which is related with the Fuchsian group description of Belyi pairs.

Let p, q be the order of the actions of ρ_1, ρ_2 on flag set. $\Pi_{p,q}$ denotes the hyperbolic regular q-gon with angles $2\pi/p$ in $\mathbb{H}$, and $\Gamma_{p,q} \subset PSL_2(\mathbb{R})$ is the group generated by two hyperbolic rotation β, γ. β is the rotation of $\Pi_{p,q}$ about the centre of angle $2\pi/p$, and γ is the rotation of $\mathbb{H}$ about a vertex of $\Pi_{p,q}$ of angle $2\pi/p$. β, γ satisfies the relations $\beta^q = \gamma^p$.

Then the homomorphism $\phi : C_2^+ \to \Gamma_{p,q} : \rho_1 \mapsto \beta, \rho_2 \mapsto \gamma$ is well defined. Set $\Gamma_{\mathcal{D},F} = \phi(B_{\mathcal{D},F})$. Since the conjugate class of $\Gamma_{\mathcal{D},F}$ is independent of the choice of F, so we may denote it by $\Gamma_{\mathcal{D}}$.

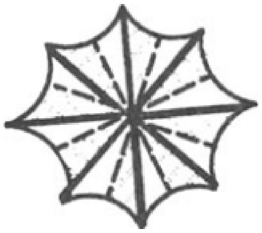


Figure 3: $\Pi_{p,q}$

The actions of β, γ on $\Gamma_{p,q}/\Gamma_{\mathcal{D},F}$ are isomorphic to the actions of ρ_1, ρ_2 on $C_2^+/B_{\mathcal{D},F}$. Thus, the Belyi pair $(S_{\mathcal{D}}, f)$ isomorphic to $(\mathbb{H}/\Gamma_{\mathcal{D}}, \pi)$ where π is the natural projection $\mathbb{H}/\Gamma_{\mathcal{D}} \to \mathbb{H}/\Gamma_{p,q}$. And the triangles in Figure 3 maps to flags by the projection $\mathbb{H} \to \mathbb{H}/\Gamma_{\mathcal{D}}$.

And obviously this also can be done in non-clean case with some adjustment.

4 Permutation representation pair

From now on, we assume a dessin is marked with white and black color such that the corresponding Belyi function sends white vertices to 0 and black vertices to 1.

Definition 4.0.1. Let $(X, \mathcal{D})$ be a dessin. If the dessin has n edges, label the edges of the dessin with integer $\{1, 2, \ldots, n\}$. Consider two element σ_0, σ_1 in the permutation group S_n. Since the edges of the dessin is labelled, we can define the action of σ_0, σ_1 on the set of edges. σ_0 permutes an edge i to the following edge under positive rotation around the white vertex of i, and σ_1 permutes an edge i to the next edge by positive rotation around the black vertex of i. Then (σ_0, σ_1) is called the *permutation representation pair* of the dessin.

Proposition 4.0.2. *The permutation representation pair satisfies the following properties:*

(1) The circles of the σ_0 is in one-one correspondence with the white vertices of the dessin. And the degree of the corresponding vertex agrees with the length of the circle.

(2) The circles of the σ_0 is in one-one correspondence with the black vertices. And the degree of the corresponding vertex agrees with the length of the circle.

(3) The circles of the $\sigma_1\sigma_0$ is in one-one correspondence with the faces.

$$\begin{aligned} (4)\ \#\{\textit{cycles of}\ \sigma_0\} &= \#\{f_{\mathcal{D}}^{-1}(0)\}, \\ \#\{\textit{cycles of}\ \sigma_1\} &= \#\{f_{\mathcal{D}}^{-1}(1)\}, \\ \#\{\textit{cycles of}\ \sigma_1\sigma_0\} &= \#\{f_{\mathcal{D}}^{-1}(\infty)\}. \end{aligned}$$

(5) The genus of the dessin can be calculated by the formula

$$2 - 2g = \#\{\textit{cycles of}\ \sigma_0\} + \#\{\textit{cycles of}\ \sigma_1\} - n + \#\{\textit{cycles of}\ \sigma_1\sigma_0\}.$$

(The Euler-Poincare characteristic of X corresponding to the polygonal decomposition)

(6) the group $\langle\sigma_0, \sigma_1\rangle$ acts transitively on the edges.

Proof. The first three properties can be easily viewed by the definition of (σ_0, σ_1), for the obit of an edge i under the action of σ_0 is just all the edges that have the common white vertex with i, similar for σ_1. The statement of the faces needs slightly more thoughts to consider the rotation of the triangle introduced in section 3.2.

Based on the first three propositions, the fourth one comes from the Grothendieck correspondence, and the fifth is an application of the Riemann-Hurwitz formula. The last one is because a dessin is a connected graph. □

If we label the edges of the dessin in a different way, and get a new permutation representation pair (σ_0', σ_1'), then there's a permutation in S_n that conjugate σ_0 to σ_0' and σ_1 to σ_1'. Actually, dessins are in one-one correspondence with permutation representation pairs up to conjugation. The way to construct a dessin from a given pair is to be discussed next, and the uniqueness is due to the relationship between the permutation representation pair and monodromy of corresponding Belyi function.

Proposition 4.0.3. *Let $\sigma_0, \sigma_1 \in S_n$, and $\langle\sigma_0, \sigma_1\rangle$ is a transitive subgroup. There exists a dessin of n edges whose permutation representation pair is precisely (σ_0, σ_1).*

One possible idea is to construct the faces first. Starting from an arbitrary edge i, the sequence $i, \sigma_0 i, \sigma_1\sigma_0 i, \sigma_0\sigma_1\sigma_0 i, \cdots$ iterates through the edges of a face containing i in counterclockwise direction. The common vertex of i and $\sigma_0(i)$ is marked white, and the common vertex of i and $\sigma_1(i)$ is marked black. When we have all faces constructed and all vertex marked, we can glue the same edge together such that the white vertices are glued to white ones and black to black. Then a dessin has been constructed now and obviously (σ_0, σ_1) is the dessin's permutation representation pair.

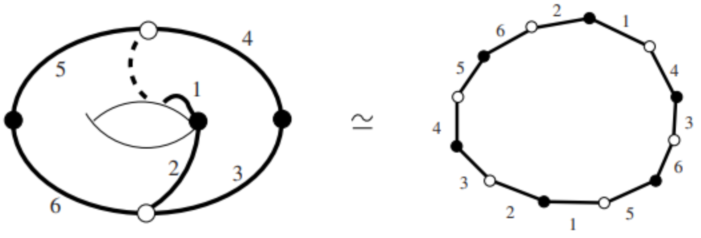


Figure 4: $\sigma_0 = (1, 5, 4)(2, 6, 3), \sigma_1 = (1, 2)(3, 4)(5, 6)$

4.1 Monodromy of dessins

The fundamental group $\pi_1(\mathbb{P}^1_{\mathbb{C}} \setminus \{0, 1, \infty\}, y)$ is a rank 2 free group. Let $y = 1/2$, then $\pi_1(\mathbb{P}^1_{\mathbb{C}} \setminus \{0, 1, \infty\}, y)$ is generated by γ_0, γ_1 that are the loops based at $y = 1/2$ and turning counterclockwise once around the points 0 and 1 respectively. $\sigma_\gamma \in \mathrm{Bij}(f^{-1}(1/2))$ is defined as follows. If $x \in f^{-1}(1/2)$, then we can lift γ to a path $\tilde{\gamma}$ with the initial point x and the point $x' \in f^{-1}(1/2)$. Set $\sigma_\gamma(x) = x'$.

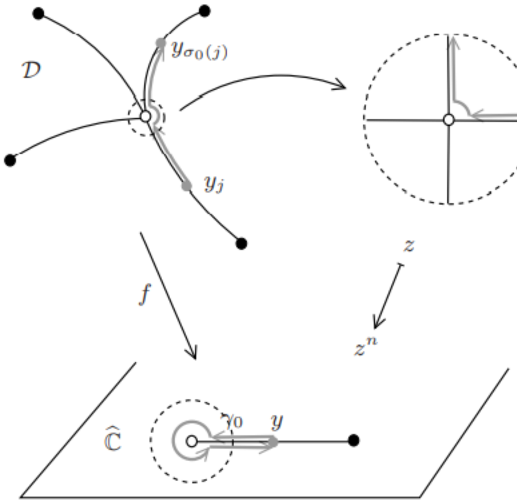


Figure 5: Lift of γ_0

As every edge of the dessin has exactly one point in the inverse image of $1/2$, $\sigma_{\gamma_0}, \sigma_{\gamma_1}$ can be thought to act on the set of edges of the dessin. Then it can be known easily from the construction of the Grothendieck correspondence that $\sigma_{\gamma_0} = \sigma_0$, $\sigma_{\gamma_1} = \sigma_1$.

Proposition 4.1.1. *The permutation representation pair of a dessin and the monodromy of the corresponding Belyi pair are determined by each other.*

Definition 4.1.2. The permutation group generated by σ_0 and σ_1 is called the monodromy group of the dessin, denoted by $\mathrm{Mon}(\mathcal{D})$.

4.2 Automorphisms of dessins

Definition 4.2.1. Let $(X, \mathcal{D})$ be a dessin. $\mathrm{Homeo}^+(X, \mathcal{D})$ is defined to be the set of orientation-preserving homeomorphisms of X that preserve $\mathcal{D}$ as a bicoloured graph. We set

$$\mathrm{Aut}(\mathcal{D}) = \mathrm{Homeo}^+(X, \mathcal{D}) / \sim,$$

where $H_1 \sim H_2$ if $H_1 \circ H_2^{-1}$ preserves setwise every edge of $\mathcal{D}$, called the *automorphisms of the dessin.*

It's clear that an element of $\mathrm{Aut}(\mathcal{D})$ is determined by how it permutes the edges of the dessin, then an element of $\mathrm{Aut}(\mathcal{D})$ can be represented by a permutation in S_n (If we label the edges of a dessin with $\{1, 2, \cdots, n\}$).

As the Grothendieck correspondence, the automorphisms of dessins correspond to automorphisms of the associated Belyi covers. Obviously, an automorphism of Belyi cover can be thought as an element of $\mathrm{Aut}(\mathcal{D})$. The following gives a brief description of why all element in $\mathrm{Aut}(\mathcal{D})$ can be constructed in this way.

Proposition 4.2.2. *The map* $\mathrm{Aut}(f_{\mathcal{D}}) \to \mathrm{Aut}(\mathcal{D}) : H \mapsto H$ *is an isomorphism of group.*

The inverse map can be obtained by the following lemma.

Lemma 4.2.3. *Let $\mathcal{T}(\mathcal{D})$ be a triangle decomposition for the dessin $(X, \mathcal{D})$, and $(S_{\mathcal{T}}, f_{\mathcal{T}})$ is the Belyi pair given by $\mathcal{T}(\mathcal{D})$. $H \in \mathrm{Homeo}^+(X, \mathcal{D})$, then there exits $H_1 \in \mathrm{Aut}(f_{\mathcal{T}})$ such that $H \sim H_1$.*

Proof. The triangle decomposition $H(\mathcal{T}(\mathcal{D}))$ provides x another Riemann surface structure. Due to the uniqueness of corresponding Belyi pair as coverings, there exists an isomorphism of Riemann surface F such that $f_{\mathcal{D}} \circ F = f_{H(\mathcal{D})}$. Then $H_1 = F \circ H$ satisfies $f_{\mathcal{T}} \circ H_1 = f_{\mathcal{T}}$ and equivalent to H. □

$\mathrm{Aut}(\mathcal{D})$ can be easily described by $\mathrm{Mon}(\mathcal{D})$ as shown next. An element of $\mathrm{Aut}(\mathcal{D})$ can be described by an permutation in S_n. And since it's a orientation-preserving homeomorphism preserving the bicolored structure, the corresponding permutation commutes with σ_0 and σ_1 (σ_0 is the rotation positively around white vertex and σ_1 around black vertex). Then we're going to explain that all the permutations with σ_0 and σ_1 are in $\mathrm{Aut}(\mathcal{D})$.

Proposition 4.2.4. *Assume $\sigma \in S_n$ commutes with σ_0 and σ_1, then there exists $H \in \mathrm{Aut}(\mathcal{D})$ making σ the corresponding permutation.*

Proof. Let $\mathcal{T}$ be a triangle decomposition associated with the dessin. As before, we first construct H_i^δ on the triangle $T_i^\delta(\delta \in \{+,-\})$, and then glue them together to get H. Take $H_i^\delta = (f_{\sigma(i)}^\delta)^{-1} \circ f_i^\delta$. To check they can glue, it's enough to check the boundary compatibility conditions. Take H_i^+ for instance:

$$H_i^+ = H_i^- \text{ on } T_i^+ \cap T_i^-$$

$$H_i^+ = H_{\sigma_0(i)}^- \text{ on } T_i^+ \cap T_{\sigma_0(i)}^-$$
$$H_i^+ = H_{\sigma_1(i)}^- \text{ on } T_i^+ \cap T_{\sigma_1(i)}^-$$

The first one comes from the compatibility of f_j^+ and f_j^-. The equation

$$\begin{aligned} H_{\sigma_0(i)}^- = (f_{\sigma(\sigma_0(i))}^-)^{-1} \circ f_{\sigma_0(i)}^- &= (f_{\sigma_0(\sigma(i))}^-)^{-1} \circ f_{\sigma_0(i)}^- \\ &\quad (\text{ on } T_i^+ \cap T_{\sigma_0(i)}^-) \\ &= (f_{\sigma(i)}^+)^{-1} \circ f_i^+ = H_i^+ \end{aligned}$$

shows the remained ones follow from that σ commutes with σ_0 and σ_1. Also, this construction ensured the glued morphism H having the corresponding permutation σ. □

Theorem 4.2.5. $\mathrm{Aut}(\mathcal{D})$ *is isomorphic to $Z(\mathrm{Mon}(\mathcal{D}))$, the centralizer of the monodromy group of $\mathcal{D}$ in S_n.*

The definition of regular dessin parallels the concept of Galois covering.

Definition 4.2.6. A dessin $(X, \mathcal{D})$ is called regular if $\mathrm{Aut}(\mathcal{D})$ acts transitively on the edges of $\mathcal{D}$.

Theorem 4.2.7. *Let $(X, \mathcal{D})$ be a dessin, and (S, f) be the corresponding Belyi pair. The following statements are equivalent.*

(1) $\mathcal{D}$ is regular.

(2) $f : S \to \mathbb{P}^1_{\mathbb{C}}$ is a Galois covering.

(3) $\#\mathrm{Mon}(\mathcal{D}) = \#\{$edges of $\mathcal{D}\} = \deg(f)$

From the Proposition 4.2.4, that the dessin is regular is totally the same with that the corresponding Belyi morphism is a Galois covering. And the last one comes from the results of Galois covering and is used to identify whether a dessin is regular.

If a dessin is regular, $\mathrm{Aut}(\mathcal{D}) \cong Z(\langle \sigma_0, \sigma_1 \rangle)$ acts transitively on the edges of $\mathcal{D}$, then the circles of σ_0 have the same length, the same for σ_0, $\sigma_1\sigma_0$. So all white vertices of the dessin have the same degree, and the same is true for black vertices and faces, and with these propertied a dessin is called uniform. But notice that a uniform dessin is not necessarily regular.

5 Galois action on dessins

We examine the action of the absolute Galois group $G_{\mathbb{Q}} := \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ on the set of dessins of genus 0, 1, and ≥ 2, respectively.

We first describe the natural action of $G_{\mathbb{Q}}$ on the set of dessins.

Definition 5.0.1 (Transform of a dessin)**.** Let $\mathcal{D}$ be a dessin, σ be an element in $G_{\mathbb{Q}}$. Then the *transform (by σ) $\mathcal{D}^\sigma$* of a dessin $\mathcal{D}$ is defined by the composition

$$\begin{tikzcd}
\mathcal{D} \arrow[r, dashed] \arrow[d] & \mathcal{D}^\sigma \\
(S_{\mathcal{D}}, f_{\mathcal{D}}) \arrow[r, "\sigma"] & (S_{\mathcal{D}}^\sigma, f_{\mathcal{D}}^\sigma) \arrow[u]
\end{tikzcd}$$

where the vertical arrows are given as in Theorem 3.2.3.

The following basic properties of the action are by direct verification.

Proposition 5.0.2. *Let $\mathcal{D}$ be a dessin. The following quantities of $\mathcal{D}$ are invariant under the action of $G_{\mathbb{Q}}$:*

(a) The number of edges.

(b) The number of white vertices, black vertices and faces.

(c) The degree of the white vertices, black vertices and faces.

(d) The genus.

(e) The monodromy group.

(f) The automorphism group.

The main result of this section is the faithfulness of the action of $G_{\mathbb{Q}}$ on the set of dessins of any fixed genus:

Theorem 5.0.3 (Faithfulness of Galois action)**.** *For any $g \in \mathbb{Z}_{\geq 0}$, the action of $G_{\mathbb{Q}}$ on the set of dessins of genus g via $\sigma \mapsto (\mathcal{D} \mapsto \mathcal{D}^\sigma)$ is faithful.*

We will examine the three cases $g = 0, 1$, $g \geq 2$ separately.

5.1 Galois action on genus 1 dessins

Recall that the j-invariant classifies Riemann surfaces of genus 1 up to isomorphism, we have

Proposition 5.1.1. *The action of* $\mathrm{Gal}(\mathbb{C}/\mathbb{Q})$ *on the isomorphism classes of compact Riemann surfaces of genus* 1 *is faithful.*

Proof. Let $\sigma \in \mathrm{Gal}(\mathbb{C}/\mathbb{Q})$ be an element not fixing a $z \in \mathbb{C}$. Choose a λ with $j(\lambda) = z$. Then $C_\lambda^\sigma = C_{\lambda^\sigma}$ has j-invariant

$$j(\lambda^\sigma) = j(\lambda)^\sigma = \sigma(z) \neq z,$$

therefore it cannot be isomorphic to C_λ. □

Corollary 5.1.2. *The action of $G_{\mathbb{Q}}$ on the set of dessins of genus 1 is faithful.*

From this we easily deduce:

Theorem 5.1.3. *The action of $G_{\mathbb{Q}}$ on the fundamental group $\pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{\mathbb{Q}}} \backslash \{0, 1, \infty\}, x)$ is faithful.*

Corollary 5.1.4. *There is an injective homomorphism*

$$\rho : G_{\mathbb{Q}} \to \mathrm{Out}(\hat{F}_2),$$

from the absolute Galois group $G_{\mathbb{Q}}$ to the outer automorphism group of the profinite completion of the free group $F_2 = \langle a, b\rangle = \langle a, b, c \mid abc = 1\rangle$.

Proof. By a corollary of Grothendieck's Riemann existence theorem (see [11]), the étale fundamental group $\pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{\mathbb{Q}}} \setminus \{0, 1, \infty\}, x)$ is isomorphic to the profinite completion of the topological fundamental group $\pi_1(\mathbb{P}^1_{\mathbb{C}} \setminus \{0, 1, \infty\}, x) \simeq F_2$, now by an easy check on the action of $G_{\mathbb{Q}}$ on the fundamental group $\pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{\mathbb{Q}}} \setminus \{0, 1, \infty\}, x)$, we see that the preimage of $\mathrm{Inn}(\hat{F}_2)$ of the embedding

$$\tilde{\rho} : G_{\mathbb{Q}} \to \mathrm{Aut}(\hat{F}_2)$$

is trivial, giving rise to an embedding

$$\rho : G_{\mathbb{Q}} \to \mathrm{Out}(\hat{F}_2)$$

as desired. □

5.2 Galois action on genus ≥ 2 dessins

We first recall a classical result of Riemann surfaces.

Proposition 5.2.1. *Two hyperelliptic Riemann surfaces are isomorphic to each other if and only if there is a Möbius transformation relating the branch set of the respective hyperelliptic involutions.*

Proof. This follows directly from the fact that the hyperelliptic involution J of a hyperelliptic Riemann surface S is the only automorphism of order 2 satisfying $S/\langle J\rangle \cong \mathbb{P}^1$. □

From this we may even deduce the faithfulness of the action of $G_{\mathbb{Q}}$ on the set of hyperelliptic curves of genus g.

Theorem 5.2.2. *Let $\sigma \in G_{\mathbb{Q}}$ be a nontrivial element in $G_{\mathbb{Q}}$, and $\alpha \in \overline{\mathbb{Q}}$ be an algebraic number not fixed by σ. For an integer $n \in \mathbb{Z}_{\geq 0}$, let C_n be the curve given by*

$$y^2 = (x-1)(x-2)\cdots(x-(2g+1))(x-(\alpha+n)).$$

Then, there is an n such that C_n^σ is not isomorphic to C_n.

Sketch of proof. Assume otherwise, then invoking Proposition 5.2.1 gives a family $\{f_n\}_{n\in\mathbb{Z}_{\geq 0}}$ of Möbius transformations in which only finitely many are distinct. We may choose three distinct natural numbers n_i $(i = 1, 2, 3)$ such that the three f_{n_i} are identical (denote by f from now on) and we have

$$f_{n_i}(\alpha + n_i) = \sigma(\alpha) + n_i$$

for each i. Then the Möbius transformation $g(z) := f(\alpha + z) - \sigma(\alpha)$ fixes all the n_i hence is identity, giving

$$f(z) = z + \sigma(\alpha) - \alpha.$$

Now

$$k + (\sigma(\alpha) - \alpha) = f(k) = l_k \in \{1, 2, \dots, 2g+1\}$$

for each $k \in \{1, 2, \dots, 2g+1\}$, but then

$$\sigma(\alpha) - \alpha = l_1 - 1 = l_2 - 2 = \cdots = l_{2g+1} - (2g+1)$$

must be 0, contradicting $\sigma(\alpha) \neq \alpha$. □

Corollary 5.2.3. *For any $g \geq 2$, the action of $G_{\mathbb{Q}}$ on the set of dessins of genus g is faithful.*

5.3 Galois action on genus 0 dessins

Notation 5.3.1. We call polynomial Belyi morphisms $f : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$ *Shabat polynomials*, as often used in the literature. Two Shabat polynomials f_1, f_2 are called *linearly equivalent* if they are the same up to a linear change of variables, i.e. $f_1(x) = f_2(ax+b)$ for some $a, b \in \mathbb{C}$. It is direct to check that two Shabat polynomials are linearly equivalent if and only if they are in the same $\mathrm{PSL}_2(\mathbb{C})$-orbit.

We will use the following elementary technical lemma:

Lemma 5.3.2. *(1) Let H_1, H_2 be two monic polynomials of the same degree whose constant terms are 0. Assume that there are polynomials G_1, G_2 with $G_1 \circ H_1 = G_2 \circ H_2$. Then $H_1 = H_2$.*

(2) Let H_1, H_2 be arbitrary polynomials of the same degree such that $G_1 \circ H_1 = G_2 \circ H_2$ for some polynomials G_1, G_2. Then there are constants c, d such that $H_2 = cH_1 + d$.

Proof. See Leila's paper [8]. □

Theorem 5.3.3. *The action of $G_{\mathbb{Q}}$ on the set of linearly equivalent classes of Shabat polynomial is faithful.*

Proof (Lenstra). Let $\sigma \in G_{\mathbb{Q}}$ be a nontrivial element in $G_{\mathbb{Q}}$, and $\alpha \in \overline{\mathbb{Q}}$ not fixed by σ. Set

$$f_\alpha(x) = \int x(x-1)^2(x-\alpha)^3 \in \mathbb{Q}(\alpha)[x],$$

then f_α is a polynomial ramified exactly at $\{0, 1, \alpha\}$ and the ramification indices are distinct. Applying Belyi's algorithm now yields a polynomial $g \in \mathbb{Q}[x]$ such that $h_\alpha = g \circ f_\alpha$ is a Shabat polynomial.

Suppose h_α and h_α^σ are linearly equivalent, then

$$g(f_\alpha(az+b)) = f_\alpha(az+b) = f_\alpha^\sigma(z) = g(f_{\sigma(\alpha)}(z))$$

for some $a, b \in \mathbb{C}$, therefore by Lemma 5.3.2

$$f_\alpha(az+b) = cf_{\sigma(\alpha)}(z) + d$$

for some $c, d \in \mathbb{C}$. Now by our choice of f_α, the map $(z-b)/a$ maps $0, 1, \alpha$ to $0, 1, \sigma(\alpha)$ respectively, forcing $b = 0$ and $a = 1$ therefore $\sigma(\alpha) = \alpha$, leading to a contradiction. □

Corollary 5.3.4. *The action of $G_{\mathbb{Q}}$ on the set of dessins of genus 0 is faithful.*

6 Examples

Some dessins of genus 0:

All genus 0 Riemann surfaces are isomorphic to $\hat{\mathbb{C}}$, then in genus 0 case, we may assume $S = \hat{\mathbb{C}}$.

Example 6.0.1. The two dessins given by $f(z) = z^n$ and Chebyshev polynomial $f(\cos t) = \cos(nt)$ are star-like and chain-like trees as shown in Figure 6.

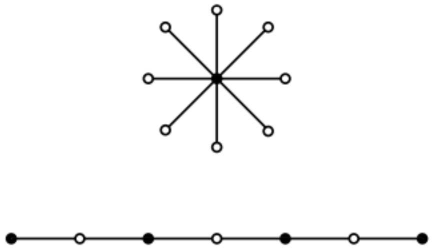


Figure 6: Star-like and chain-like trees

Example 6.0.2. The two dessins shown in Figure 7 are both with six edges and white vertices of degrees 2,2,1,1 and black vertices of degrees 4,1,1. The permutation representation pair of $\mathcal{D}_1$ is $\sigma_0 = (1,5)(6,3)$ and $\sigma_1 = (1,2,3,4)$. The permutation representation pair of

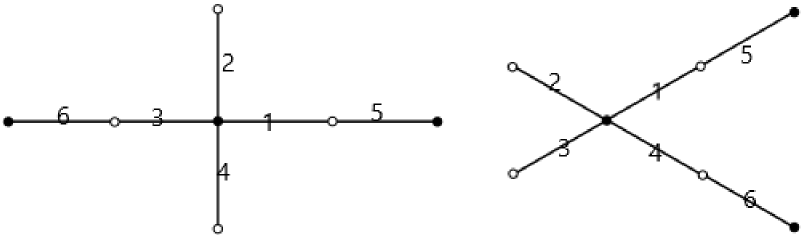


Figure 7: $\mathcal{D}_1$ and $\mathcal{D}_2$

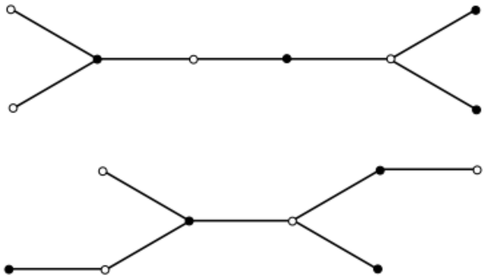


Figure 8: $\mathcal{D}_3$ and $\mathcal{D}_4$

$\mathcal{D}_2$ is $\sigma_0 = (1,5)(6,4)$ and $\sigma_1 = (1,2,3,4)$. Then $\#\,\mathrm{Mon}(\mathcal{D}_1) = 48$, $\#\,\mathrm{Mon}(\mathcal{D}_2) = 120$, and $\mathrm{Aut}(\mathcal{D}_1) = \langle(1,3)(2,4)(5,6)\rangle$, $\mathrm{Aut}(\mathcal{D}_2) = \langle(1)\rangle$. Since the monodromy group and automorphism group are invariants under the action of the absolute Galois group, therefore $\mathcal{D}_1$ and $\mathcal{D}_2$ can not be Galois conjugated.

But for dessins $\mathcal{D}_3$ and $\mathcal{D}_4$ shown in Figure 8, they have the same degree sets and monodromy group and automorphism group. But the monodromy group of $\mathcal{D}_3^c$ and $\mathcal{D}_4^c$ are different, then they are not Galois conjugated.

If R is a Belyi extending, the monodromy group and automorphism group of $R \circ f$ are also invariant under Galois group action. This example shows this method can construct some new invariants. Finding invariants of dessins which completely identify their $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$-orbits is a problem concerned in the theory of dessins d'enfants.

Example 6.0.3. This is an example of calculating the corresponding Belyi function from the dessin considered in the paper [15]. The dessin $\mathcal{D}$ shown in Figure 9 is a clean dessin(all vertices are the preimages of 0, and there's a middle point in each edge being a preimage of 1). Since it has three vertices of degree 2 and two vertices of degree 3, the numerator be f should be of the form:

$$(x^2 + ax + b)^3(x^3 + cx^2 + dx + e)^2 \tag{6.0.3.1}$$

The dessin have three faces of valencies 3,4,5, then the denominator of f should be $(x - f)^3(x - g)^4(x - h)^5$. As $\mathcal{D}$ is clean, $1 - f$ has 6 roots and the multiplicity of each root is 2.

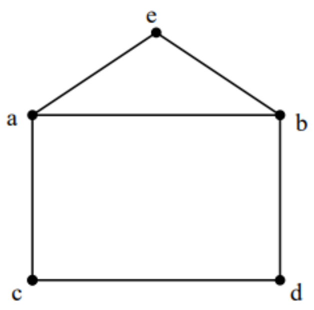


Figure 9: A dessin with 6 edges

So we have:

$$f(x) = k\frac{(x^2+ax+b)^3(x^3+cx^2+dx+e)^2}{(x-f)^3(x-g)^4(x-h)^5} \tag{6.0.3.2}$$

$$f(x) - 1 = k\frac{(x^6+mx^5+nx^4+px^3+qx^2+ux+v)^2}{(x-f)^3(x-g)^4(x-h)^5} \tag{6.0.3.3}$$

The above two equations reaches a system of equations including 15 unknowns and 12 equations. The remaining 3 degrees is because of the automorphism group of $\mathbb{P}^1_{\mathbb{C}}$. More the related results of finding the equation of Belyi function from a clean dessin in genus 0 case can be found in the article [8]

A dessin of genus 1:

Example 6.0.4. Let S be the Riemann surface $\{y^2 = x(x-1)(x-\sqrt{2})\} \cup \{\infty\}$. By the method proposed in the proof of $(a) \Rightarrow (b)$ of Belyi's theorem, we can construct a Belyi function on S: $f = -4(x^2-1)/(x^2-2)^2$:

$$x \xrightarrow{t \mapsto t^2-2} x^2-2 \xrightarrow{t \mapsto -1/t} -1/(x^2-2) \xrightarrow{t \mapsto 4t(1-t)} -4(x^2-1)/(x^2-2)^2 \tag{6.0.4.1}$$

By computing the preimage of $[0,1]$ step by step, the dessin corresponding to f can be obtained as Figure 10.

The only $\sigma \in \mathrm{Gal}(\mathbb{C})$ may act non-trivially on the dessin is $\sigma(\sqrt{2}) = -\sqrt{2}$. $S^\sigma = \{y^2 = x(x-1)(x+\sqrt{2})\} \cup \{\infty\}$, $f^\sigma = -4(x^2-1)/(x^2-2)^2$. And we can obtain the corresponding dessin $\mathcal{D}^\sigma$ by the same means as shown in Figure 11.

The permutation representation pair of $\mathcal{D}$ is $\sigma_0 = (1,7,5,3)(4,8), \sigma_1 = (1,2,3,4,5,6,7,8)$. The permutation representation pair of $\mathcal{D}^\sigma$ is $\sigma_0 = (2,4,6,8)(3,7), \sigma_1 = (1,2,3,4,5,6,7,8)$. They are not conjugate, so $\mathcal{D}$ and $\mathcal{D}^\sigma$ are not isomorphic dessin. Actually, $j(\sqrt{2}) \neq j(-\sqrt{2})$ (j denotes the classical j-invariant) shows that S and S^σ are not isomorphic Riemann surface. Hence, there are two dessins in the orbit of $\mathcal{D}$.

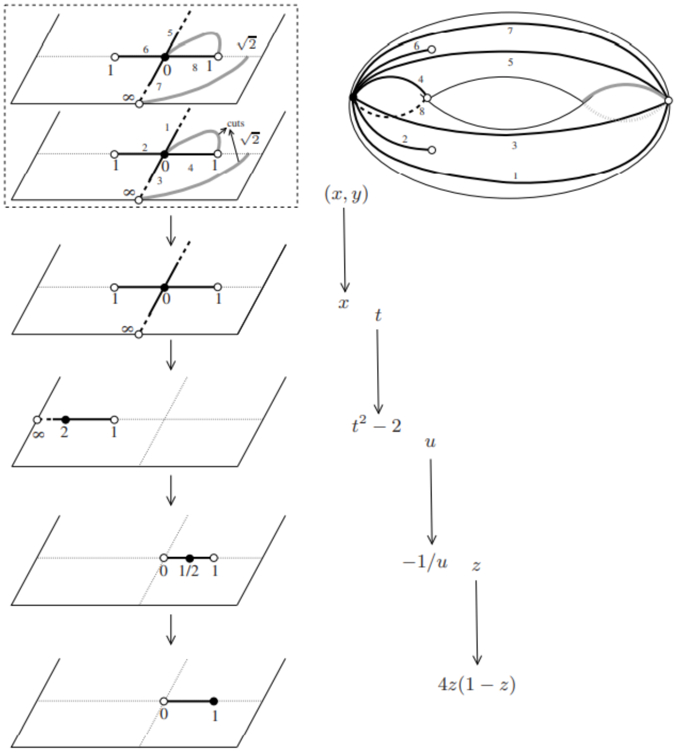


Figure 10: The procedure of finding the dessin

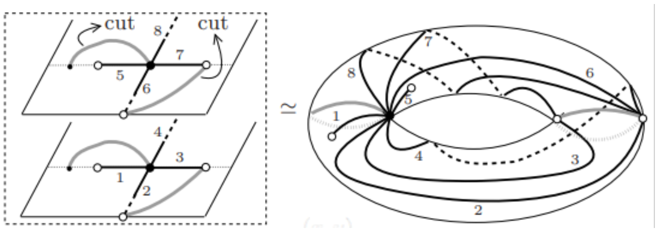


Figure 11: Dessin $\mathcal{D}^{\sigma}$

7 Applications

7.1 Minimal degree of $f^3 - g^2$

To state our main result in simpler terms, we introduce the following definitions.

Convention 7.1.1. All graphs are thought to be simple in this section, and by a *binary tree* we mean a tree whose vertices are all of degree 1 or 3.

Example 7.1.2. The graph in Figure 12 is a binary tree in the sense of 7.1.1.

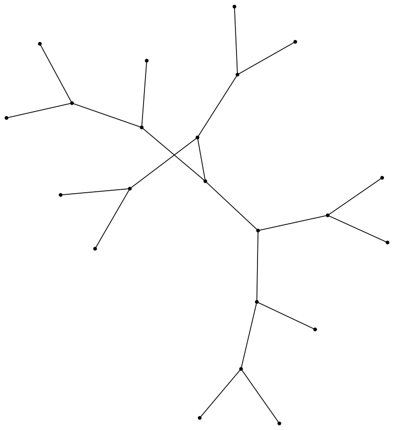


Figure 12: A binary tree

Definition 7.1.3 (Tacosad)**.** A graph Γ is a *tacosad*, if there is a surjection of graphs

$$\pi : \coprod_{i \in I} \gamma_i \to \Gamma$$

that is a bijection on the set of edges, where I is a finite index set, and each γ_i is isomorphic to K_3, such that any loop of Γ is generated by $\{\pi(\gamma_i)\}_{i \in I}$. We call these $\pi(\gamma_i)$ *faces* of the tacosad Γ.

Remark 7.1.4. It is straight-forward to check that any tacosad is the dual graph of a binary tree, and that the dual graph of any binary tree is a tacosad.

Example 7.1.5. The graph in Figure 13 is a tacosad. In fact, it is the dual graph of the binary tree in Example 7.1.2.

Definition 7.1.6 (Orientation of a tacosad)**.** An *orientation* of a tacosad Γ is a map that associates each face of Γ an orientation.

Definition 7.1.7 (Oriented binary tree)**.** An *oriented binary tree* is a pair $(\mathcal{T}, \varphi)$, where $\mathcal{T}$ is a binary tree, and φ is an orientation of the dual graph of $\mathcal{T}$, in the sense of Definition 7.1.6.

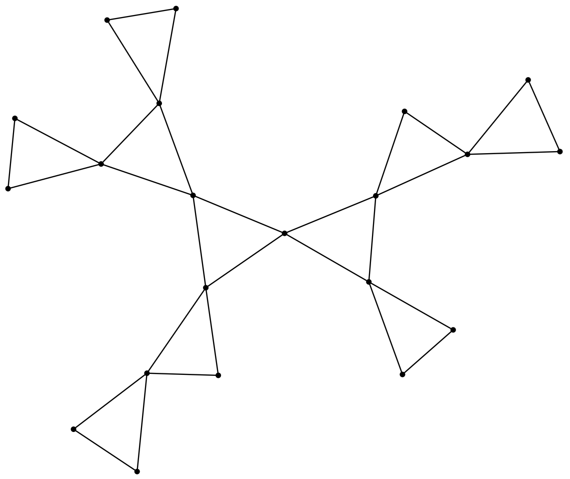


Figure 13: A tacosad

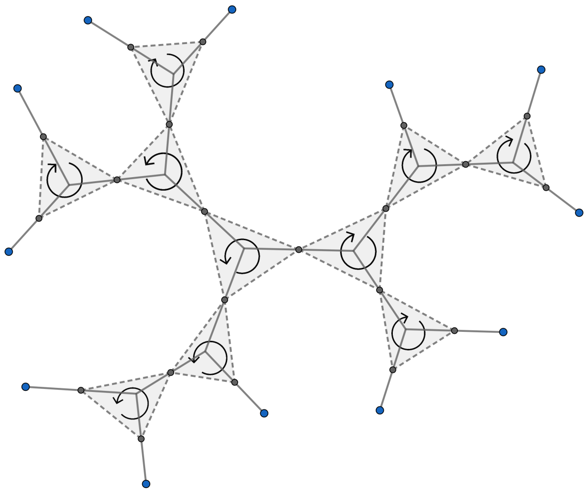


Figure 14: An oriented binary tree

Example 7.1.8. The structure in Figure 14 is an oriented binary tree. In fact, its underlying binary tree is the same as in Figure 12.

Remark 7.1.9. The motivation of Definition 7.1.7 is to keep track of the orientation of a dessin by adding more structure to the graph. More concretely speaking, the φ in the definition allow us to draw the tree on a oriented plane (or sphere to be more precise) in a unique way, namely we flip the faces to the "correct" position according to the orientation. In this manner we see that, in this specific case for binary trees, the notion of isomorphism for such clean dessins is actually coherent with the notion of isomorphism for $\mathfrak{S}_3$-trees, i.e., the topological data is of combinatorial nature.

The combinatorial structure Zannier used in his paper [13] is (essentially) the following more artificial analog of Definition 7.1.7.

Definition 7.1.10 ($\mathfrak{S}_3$-tree)**.** An $\mathfrak{S}_3$*-tree* is a pair $(\mathcal{T}, \sigma)$, where $\mathcal{T}$ is a binary tree, and σ is a map from the set of degree-3 vertices to the set of permutations on the vertices of $\mathcal{T}$, such that $\sigma(v)$ non-trivially permutes its three neighbours cyclically and fixes the rest vertices, for each vertex v of degree 3.

The main result of this section is the following:

Theorem 7.1.11. *Let f, g be two coprime complex coefficient polynomials of degree $2n, 3n$ respectively. Then*

(1) $\deg(f^3 - g^2) \geq n + 1$*;*

(2) The equality can be reached for each positive integer n;

(3) The number μ_n of linearly-equivalent classes of pairs (f, g) obtaining the minimal degree $\deg(f^3 - g^2) = n + 1$ *equals the number ϕ_n of isomorphism classes of oriented binary trees on $2n$ vertices; and*

(4) We have the following estimation for μ_n:

$$c_1 n^{-5/2} 4^n < \mu_n < c_2 n^{-3/2} 4^n$$

for some positive reals c_1, c_2 independent of n.

Remark 7.1.12. The first claim of Theorem 7.1.11 is actually a direct corollary of the *abc* conjecture for polynomials (a.k.a. Mason–Stothers theorem), which has a short and elementary proof. However, our approach here proves both the inequality and the criterion for equality at one strike.

Proof of Theorem 7.1.11. Let

$$r : \mathbb{P}^1_{\mathbb{C}} \to \mathbb{P}^1_{\mathbb{C}}$$

be given by the rational function $\frac{f^3}{f^3-g^2}$, then r is of degree $6n$, and $r-1$ is given by the rational function $\frac{g^2}{f^3-g^2}$. Now applying Riemann-Hurwitz formula to r gives

$$\begin{aligned} -2 = -12n + \sum_{x\in\mathbb{P}^1_{\mathbb{C}}} \nu_r(x) &\geq -12n + \sum_{\substack{x\in\mathbb{P}^1_{\mathbb{C}} \\ r(x)\in\{0,1,\infty\}}} \nu_r(x) \\ &= 6n - (\#r^{-1}(0) + \#r^{-1}(1) + \#r^{-1}(\infty)) \\ &\geq 6n - (2n + 3n + (\deg(f^3-g^2)+1)) \\ &= n - 1 - \deg(f^3-g^2), \end{aligned}$$

i.e., $\deg(f^3-g^2) \geq n+1$. The first claim is proved.

From the chain of inequalities above we see that, $\deg(f^3-g^2) = n+1$ if and only if the following conditions hold:

(a) r is only ramified above $0, 1, \infty$, i.e., r is a Belyi morphism;

(b) $\#r^{-1}(0) = 2n$, i.e., f has no multiple roots;

(c) $\#r^{-1}(1) = 3n$, i.e., g has no multiple roots;

(d) $\#r^{-1}(\infty) = \deg(f^3-g^2)+1$, i.e., f^3-g^2 has no multiple roots.

Now suppose r satisfy the above conditions, then the graph associated to the clean Belyi morphism r is essentially a tree on $2n$ vertices (discarding the middle-points and the self-loops), all of which are of degree 1 or 3. Recall that Theorem 3.2.3 gives us a bijection between the set of isomorphism classes of clean Belyi morphisms and the set of abstract clean dessins, and that each isomorphism class of clean Belyi morphisms is just a $\mathrm{PSL}_2(\mathbb{C})$-orbit of the natural action of $\mathrm{PSL}_2(\mathbb{C})$ on the set of clean Belyi morphisms, therefore non-linearly-equivalent pairs gives non-isomorphic Belyi morphisms. In conclusion, we have a bijection between the set of linearly-equivalent classes of pairs (f, g) obtaining the minimal degree $\deg(f^3-g^2) = n+1$ and the set of isomorphism classes of oriented binary tree on $2n$ vertices, in view of Remark 7.1.9. This completes the proof of the second and third claim.

As for the fourth claim, note that we've shown in the above paragraph that $\mu_n = \phi_n$, thus by applying the estimation B.0.4 for ϕ_n in Appendix B we obtain the desired estimation. □

7.2 The Mordell conjecture is as easy as the *abc* conjecture[1]

We will prove that the *abc* conjecture implies the Mordell conjecture in this section. The upshot is that validity of the *abc* conjecture gives us bound on ramifications, forcing the number of rational points to be finite, via Belyi's theorem A.0.1.

[1] This amusing equivalent formulation of Elkies' original title "ABC implies Mordell" is noted by Don Zagier, as pointed out by Elkies in [1].

Convention 7.2.1. Whenever we use the capital H for heights, we mean the exponential of the corresponding height h defined in Appendix C. Moreover, recall that we have a homomorphism

$$\mathrm{WCl}(X) \to \mathrm{Pic}(X),\ D \mapsto \mathcal{O}_X(D)$$

for any variety X, which is an isomorphism if X is non-singular. We will use H_D (resp. h_D) to denote the Weil heights $H_{\mathcal{O}(D)}$ (resp. $h_{\mathcal{O}(D)}$). We will also use $D_0(f)$ (resp. $D_\infty(f)$) to denote the divisor of zeros of f (resp. the divisor of poles of f), i.e., we have

$$\mathrm{div}(f) = D_0(f) - D_\infty(f),$$

where $D_0(f)$ and $D_\infty(f)$ are effective.

We first recall the *abc* conjecture and the Mordell conjectures.

Definition 7.2.2. (Conductor on $\mathbb{P}^n$) Let K be a number field. Then the *conductor* $N(x)$ of an algebraic point $x = (x_0, \dots, x_n) \in \mathbb{P}_K(\overline{K})$ is given by

$$N(x) = N(x_0, \dots, x_n) = \prod_{\substack{v \in M_{K,f} \text{s.t.} \\ |x_i/x_j|_v > 1 \\ \text{for some } i,j}} N(\wp_v).$$

We will use $N_0(r)$ (resp. $N_1(r), N_\infty(r)$) to denote the products of the absolute norms of the prime ideals at which r (resp. $r-1$, $1/r$) has positive valuation, then

$$N(r, -1, 1-r) = N_0(r) \cdot N_1(r) \cdot N_\infty(r)$$

by direct verification.

Conjecture 7.2.3 (*abc* conjecture over K)**.** *Let K be a number field. Then*

$$N(a, b, c) \gg_\epsilon H(a, b, c)^{1-\epsilon}$$

for all $a, b, c \in K^\times$ with $a + b + c = 0$, for each $\epsilon > 0$.

Theorem 7.2.4 (Mordell conjecture over K, proved by G. Faltings)**.** *Let K be a number field, C a curve of genus $g > 1$ over K. Then the set $C(K)$ of rational points of C is finite.*

Remark 7.2.5. Fatings' proof of the Mordell conjecture makes an essential use of height theory, which is beyond the scope of Appendix C. The idea is that you can define heights of (*not* on) any abelian variety, via two approach: one is directly using the integral of a differential on the abelian variety, obtaining the so-called Faltings height; the other is to view abelian varieties as points of the Siegel modular variety, which is the moduli space of abelian varieties added some extra data to shrink the automorphism group. The upshot is that these two heights are the same up to $O(1)$, giving us Northcott property (see Appendix C for definitions) for the Faltings height, therefore deduce the finiteness of a certain set of abelian varieties associated to each abelian variety, which completes the proof of the Mordell conjecture.

We also have an *effective* version of Mordell conjecture.

Conjecture 7.2.6 (Effective Mordell over K)**.** *Let X be a projective and smooth curve over $\mathbb{Q}$ of genus $g > 1$. Then for any $d \geq 1$, there exist constants $A(X, d)$ and $B(X, d)$ depending only on X and d such that for any finite extension K of $\mathbb{Q}$ of degree d, we have*

$$h(x) < A(X, d) \log |\Delta_K| + B(X, d),$$

for any $x \in X(K)$.

Proposition 7.2.7 (*abc* implies Mordell)**.** *For any number field K, the abc conjecture over K implies the Mordell conjecture over K.*

Remark 7.2.8. In fact, it can be shown that (some version of) the effective Mordell conjecture implies (some version of) the abc conjecture. Thus combining with (some version of) Proposition 7.2.7, these two conjectures are equivalent.

An essential part of the proof is the following observation.

Lemma 7.2.9. *Let C be any curve over K and $f \in K(C)$ be a rational function of degree d. Then for any rational point $x \in C(K) \setminus f^{-1}(0)$ we have*

$$\log N_0(f(x)) < (1 - \frac{b_f(0)}{d}) \log H(1, f(x)) + O(\sqrt{\log H(1, f(x))} + 1).$$

Proof. Write

$$D_0(f) = \sum_k m_k D_k,$$

where the D_k are distinct irreducible divisors of degrees d_k occurring with multiplicities m_k in $D_0(f)$. Then

$$d = \sum_k m_k d_k = \deg D \text{ and } b_f(0) = d - \sum_k d_k = \deg D_0(f) - \deg D_0(f)_{\text{red}},$$

where $D_0(f)_{\text{red}}$ denotes the divisor $\sum_{f(x)=0}(x)$, i.e. $D_0(f)$ with all multiplicities removed. We then have

$$\log H(1, f(x)) = h_{D_0(f)}(x) + O(1) = \sum_k m_k h_{D_k}(x) + O(1).$$

Now note that, a prime occurs in $N_0(f(x))$ if and only if it contributes to $h_{D_k}(x)$ for some k, except for the primes of bad reduction of C and the primes of good reduction at which f reduces to the identically zero function. But the total number of these "bad" primes is finite, giving

$$\log N_0(f(x)) < \sum_k h_{D_k}(x) + O(1) = h_{D_0(f)_{\text{red}}}(x) + O(1).$$

Thus it remains to show

$$h_{D_0(f)_{\mathrm{red}}}(x) = \frac{\deg D_0(f)_{\mathrm{red}}}{\deg D_0(f)} \cdot h_{D_0(f)}(x) + O(\sqrt{\log H(1, f(x))} + 1),$$

which is just

$$h_\Delta(x) = O(\sqrt{\log H(1, f(x))} + 1),$$

where Δ denotes the degree-zero divisor

$$\Delta = (\deg D_0(f))D_0(f)_{\mathrm{red}} - (\deg D_0(f)_{\mathrm{red}})D_0(f).$$

Recall that Theorem C.0.20 tells us this is true for any degree-zero divisor, thus invoking Theorem C.0.20 now completes the proof. □

We can now prove Proposition 7.2.7.

Proof of Proposition 7.2.7. By Belyi's theorem A.0.1, we may choose a rational function $f \in K(C)$ ramified only above $0, 1, \infty$. Then by Riemann-Hurwitz formula, we have

$$m := \#\{x \in C(\overline{\mathbb{Q}}) \mid f(x) \in \{0, 1, \infty\}\} = \deg(f) + 2 - 2g < \deg(f) =: d.$$

Now by summing the three inequalities obtained by applying Lemma 7.2.9 to $f, f-1, 1/f$ respectively, we see that

$$\log N(f(x), -1, 1 - f(x)) < \frac{m}{d} \log H(1, f(x)) + O(\sqrt{\log H(1, f(x))} + 1),$$

giving a counterexample to the *abc* conjecture over K for $\epsilon > 1 - (m/d)$ once $H(1, f(x)) = H_{D_0(f)}(x)$ is large enough, i.e. for all but finitely many x by Northcott property C.0.16. This completes the proof of *abc* implies Mordell. □

7.3 A characterization for finite image Galois representations

Notation 7.3.1. Let F be any field. For notational convenience, we use G_F to denote the absolute Galois group $\mathrm{Gal}(\overline{F}/F)$ over F, and $\mathcal{C}_F$ to denote the class

$$\{V \mid V \text{ appears as a subquotient of } \mathbb{Q}[\pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0, 1, \infty\}, 0_v)] \text{ for every } v\},$$

if no confusion arises.

The main result in this section the following characterization for finite image Galois representations.

Theorem 7.3.2. *Any continuous finite image representation*

$$\rho : G_F \to \mathrm{GL}_n(\mathbb{Q})$$

can be embedded into the space of locally constant functions

$$\mathrm{Func}^{\text{loc.const.}}(\pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0, 1, \infty\}, 0_v), \mathbb{Q}),$$

for any tangential base point 0_v.

A key ingredient of the proof of Theorem 7.3.2 is the following telescopic property of the fundamental group of $\mathbb{P}^1_{\overline{F}} \setminus \{0, 1, \infty\}$.

Proposition 7.3.3. *There exists an open subgroup*

$$\Gamma \subset \pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0, 1, \infty\}, 0_v)$$

stable under G_F-action and admitting a G_F-equivariant surjection

$$\Gamma \twoheadrightarrow \pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0, 1, \infty\}, 0_{v_1}) \times \pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0, 1, \infty\}, 0_{v_2})$$

for some tangential base points $0_{v_1}, 0_{v_2}$ at 0.

Proof. We make use of Belyi's explicit construction of Belyi morphism.

Consider the degree 3 finite morphism

$$f : \mathbb{P}^1_F \to \mathbb{P}^1_F,\ z \mapsto \frac{27}{4} z(z-1)^2,$$

then f is only ramified at $\frac{1}{3}, 1, \infty$. Since $f(1) = 0, f(\frac{1}{3}) = 1, f(\infty) = \infty$ the map f restricts to a finite étale cover

$$\mathbb{P}^1_F \setminus \{0, \frac{1}{3}, 1, \frac{4}{3}, \infty\} \to \mathbb{P}^1_F \setminus \{0, 1, \infty\}.$$

Moreover, since f is unramified at 0, we may choose a tangential base point 0_{v_1} for the truncated projective line $\mathbb{P}^1_F \setminus \{0, \frac{1}{3}, 1, \frac{4}{3}, \infty\}$ such that $f(0_{v_1}) = 0_v$.

Let Γ be defined as

$$\Gamma := f_*(\pi_1^{\text{ét}}(\mathbb{P}^1_F \setminus \{0, \frac{1}{3}, 1, \frac{4}{3}, \infty\}, 0_{v_1})) \subset \pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0, 1, \infty\}, 0_v).$$

Then the inclusion maps

$$i_1 : \mathbb{P}^1_F \setminus \{0, \frac{1}{3}, 1, \frac{4}{3}, \infty\} \to \mathbb{P}^1_F \setminus \{0, 1, \infty\},\ i_2 : \mathbb{P}^1_F \setminus \{0, \frac{1}{3}, 1, \frac{4}{3}, \infty\} \to \mathbb{P}^1_F \setminus \{0, \frac{1}{3}, \frac{4}{3}\}$$

induce a surjection

$$\Gamma \twoheadrightarrow \pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0, 1, \infty\}, 0_{v_1}) \times \pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0, \frac{1}{3}, \frac{4}{3}\}, 0_{v_1})$$

by the Seifert-Van Kampen theorem, this then completes the proof of Lemma 7.3.3 since $\pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0, \frac{1}{3}, \frac{4}{3}\}, 0_{v_1})$ can be identified with $\pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0, 1, \infty\}, 0_{v_2})$ via an automorphism of $\mathbb{P}^1_{\overline{F}}$, for some tangential base point 0_{v_2}. □

From Proposition 7.3.3 we deduce the following lemma to be used in the proof of the main theorem 7.3.2.

Lemma 7.3.4. *For a tangential base point 0_v supported at 0 there exist two other tangential base points 0_{v_1} and 0_{v_2} such that, if V_1 and V_2 are representations of G_F appearing as subquotients of $\mathbb{Q}[\pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0, 1, \infty\}, 0_{v_i})]^{G_F-\text{fin}}$, then $V_1 \otimes V_2$ is a subquotient of $\mathbb{Q}[\pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0, 1, \infty\}, 0_v)]^{G_F-\text{fin}}$.*

Proof. Note that the representation $V_1 \otimes V_2$ is a subquotient of

$$\mathbb{Q}[\pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0,1,\infty\}, 0_{v_1})]^{G_F-\text{fin}} \otimes \mathbb{Q}[\pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0,1,\infty\}, 0_{v_2})]^{G_F-\text{fin}} \subset \mathbb{Q}[\Gamma]^{G_F-\text{fin}},$$

and that $\mathbb{Q}[\Gamma]^{G_F-\text{fin}}$ is a quotient of $\mathbb{Q}[\pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0,1,\infty\}, 0_v)]^{G_F-\text{fin}}$ by Grothendieck's Galois theory formalism, therefore $V_1 \otimes V_2$ is a subquotient of $\mathbb{Q}[\pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0,1,\infty\}, 0_v)]^{G_F-\text{fin}}$ as desired. □

We now prove Theorem 7.3.2.

Proof of Theorem 7.3.2. Since the image of

$$\rho : G_F \to \mathrm{GL}_n(\mathbb{Q})$$

is finite, it factor through $\mathrm{Gal}(K/F)$ for a finite Galois extension K/F. Now recall that every faithful representation of a finite group G contains a faithful subrepresentation of dimension $\leq \#G$, thus if we can show that the space $\mathbb{Q}[\pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0,1,\infty\}, 0_v)]^{G_F-\text{fin}}$ has some faithful representation W_v of $\mathrm{Gal}(K/F)$ as a subquotient for every tangential base point 0_v, then we may choose the representations W_v in a way that they all belong to finitely many isomorphism classes, say $W_1, \ldots, W_N$. Then by repeatedly applying Lemma 7.3.4, we can conclude that $\mathbb{Q}[\pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0,1,\infty\}, 0_v)]^{G_F-\text{fin}}$ has a subquotient of the form $W_1^{\otimes a_1} \otimes \cdots \otimes W_N^{\otimes a_N}$ with $a_i \geq d$ for at least one i, for any $d \geq 0$. Finally, recall that any representation of a finite group is contained in a large enough tensor power of any faithful representation (†), we see that $\mathbb{Q}[\pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0,1,\infty\}, 0_v)]^{G_F-\text{fin}}$ has a subquotient of the form $W_1^{\otimes b_1} \otimes \cdots \otimes W_N^{\otimes b_N}$ with $b_i \geq d$ for *all* i, applying (†) again then completes the proof of Theorem 7.3.2. Thus it remains to show the existence of W_v for any v.

For this, we first choose a smooth proper geometrically connected curve C over K that does not descend to any proper subfield $K' \subset K$. By Belyi's theorem A.0.1 there exists a finite map $f : C \to \mathbb{P}^1_K$ that is étale over $\mathbb{P}^1_K \setminus \{0,1,\infty\}$. Denote by $U \subset C$ the preimage $f^{-1}(\mathbb{P}^1_K \setminus \{0,1,\infty\})$ of $\mathbb{P}^1_K \setminus \{0,1,\infty\}$. Choosing a tangential $\overline{F}$-base point x_w for $C \setminus U$ that lies above 0_v, we get an open subgroup $f_*(\pi_1^{\text{ét}}(U_{\overline{K}}, x_w)) \subset \pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{K}} \setminus \{0,1,\infty\}, 0_v)$. If an element $\sigma \in G_F$ stabilizes this subgroup, then the scheme $U_{\overline{K}}$ can be descended to the field $(\overline{F})^{\sigma=1}$. Our choice of C then forces the stabilizer of this subgroup to be contained inside $G_K \subset G_F$. In particular, there is a finite G_F-equivariant quotient $\pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0,1,\infty\}, 0_v) \twoheadrightarrow S$ such that the kernel of the action of G_F on S is contained in G_K. In conclusion, there exists a G_F-equivariant finite quotient $\pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0,1,\infty\}, 0) \to X$ such that the action of G_F on X factors through a faithful action of $\mathrm{Gal}(K/F)$. □

Remark 7.3.5. In the same manner we could prove a generalization of Theorem 7.3.2, in which the étale fundamental group is replaced with its pro-algebraic completion, and $\mathrm{Func}^{\text{loc.const.}}(\pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0,1,\infty\}, 0_v), \mathbb{Q})$ (or equivalently $\mathbb{Q}[\pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0,1,\infty\}, 0_v)]^{G_F-\text{fin}}$) is replaced by $\mathbb{Q}_p[\pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0,1,\infty\}, 0_v)]^{G_F-\text{fin}}$. In fact, it can be shown that *every* semi-simple representation coming from geometry appears as a subquotient of the space of functions on the pro-algebraic completion of $\pi_1^{\text{ét}}(\mathbb{P}^1_{\overline{F}} \setminus \{0,1,\infty\}, 0_v)$. This then gives an astounding reduction of the Fontaine-Mazur conjecture. For details, see [7].

A Appendix A: The "easy" part of Belyi's theorem[2]

We first recall Belyi's theorem:

Theorem A.0.1 (Belyi)**.** *A complex smooth projective curve X is defined over a number field, if and only if there exists a non-constant morphism $f : X \to \mathbb{P}^1_{\mathbb{C}}$ with at most 3 critical values.*

The "only if" part is proved in Section 3.1, the "if" part follows directly from a theorem of Weil (Theorem A.0.2), we will give a proof of this theorem in this section, for the sake of integrity.

Theorem A.0.2 (Weil)**.** *Let $X/\mathbb{C}$ be a projective smooth algebraic curve. If there is a morphism*

$$f : X \to \mathbb{P}^1_{\mathbb{C}}$$

such that all branch values lie in $\overline{\mathbb{Q}}$, then X is defined over $\overline{\mathbb{Q}}$.

Definition A.0.3 (Closed subgroup)**.** Let K be a field. A subgroup G of $\mathrm{Aut}(K)$ is *closed*, if there is a subfield k of K with $G = \mathrm{Aut}(K/k)$.

We will use the following elementary field-theoretic lemma:

Lemma A.0.4. *Let K/k be a field extension. Then,*

1. *Any automorphism of k can be extended to an automorphism of K. Furthermore, we have:*
$$K^{\mathrm{Aut}(K/k)} = k.$$
2. *Let G be a subgroup of $\mathrm{Aut}(K)$, and H be a subgroup of G of finite index. Then the field extension K^H/K^G is finite. If H is a normal subgroup of G or G is closed, then we have $[K^H : K^G] \leq [G : H]$. Moreover, the equality holds if H is closed.*

We also recall a fundamental result in ramification theory:

Proposition A.0.5. *Let S be a finite set of (closed) points of $\mathbb{P}^1_{\mathbb{C}}$, and $d \geq 1$ be a natural number. Then there are at most finitely many isomorphism classes of pairs (X, f) where $X/\mathbb{C}$ is a curve and $f : X \to \mathbb{P}^1_{\mathbb{C}}$ is a finite morphism of varieties over $\mathbb{C}$ of degree d whose branch values lie in S.*

Proof. By translating to the setting of Riemann surfaces, this reduces to the fact that there are at most finitely subgroups of index d of the fundamental group $\pi_1(\mathbb{P}^1_{\mathbb{C}} \setminus S)$, which is true since $\pi_1(\mathbb{P}^1_{\mathbb{C}} \setminus S)$ is finitely generated. □

[2] It should be noted that this part is neither easy nor direct. This reason this part is called easy is historical: It follows directly from a (hard) theorem of Weil (Theorem A.0.2), which is proved a lot earlier than the other part of Belyi's theorem.

Definition A.0.6 (Moduli field)**.** The *moduli field of* (X, f) is the field $M(X, f) := K^{U(X,f)}$ fixed by the group $U(X, f)$ consisting of all $\sigma \in \mathrm{Aut}(K)$ such that there exists an isomorphism $\sigma_X : X^\sigma \to X$ of varieties over K such that the following diagram commutes:

$$\begin{tikzcd}
X^\sigma \arrow[r, "\sigma_X"] \arrow[d, "f^\sigma"] & X \arrow[d, "f"] \\
(\mathbb{P}^1_K)^\sigma \arrow[r, "\mathrm{Proj}(\sigma)"] & \mathbb{P}^1_K.
\end{tikzcd}$$

When $f = 0$, we simply call the field $M(X) := M(X, 0)$ the *moduli field of* X.

Theorem A.0.2 is then the conjunction of the following two lemmas:

Lemma A.0.7. *Let $X/\mathbb{C}$ be a curve, let $f : X \to \mathbb{P}^1_\mathbb{C}$ be a finite morphism and let k be a subfield of $\mathbb{C}$ such that the branch values of f are k-rational. Then the moduli field of f is contained in a finite extension of k.*

Proof. For any $\sigma \in \mathrm{Aut}(\mathbb{C}/K)$, the branch values of $f(\sigma) : X^\sigma \xrightarrow{t^\sigma} (\mathbb{P}^1_\mathbb{C})^\sigma \xrightarrow{\mathrm{Proj}(\sigma)} \mathbb{P}^1_\mathbb{C}$ also lie in S, and $\deg f(\sigma) = \deg f$. So the $\mathrm{Aut}(\mathbb{C}/K)$-orbit of (the isomorphism class of) the pair (X, f) is finite, by Proposition A.0.5, therefore the stabilizer is of finite index in $\mathrm{Aut}(\mathbb{C}/K)$. Now note that the stabilizer is contained in $U(X, f)$, thus the moduli field $M(X, f) = \mathbb{C}^{U(X,f)}$ is contained in a finite extension of $\mathbb{C}^{\mathrm{Aut}(\mathbb{C}/K)} = K$, by Lemma A.0.4. □

Lemma A.0.8. *X and f are defined over a finite extension of $M(X, f)$.*

Proof. Choose a $\mathbb{Q}$-rational point $y_0 \in \mathbb{P}^1_K(\mathbb{Q}) \setminus S$, and a point x_0 in the fibre $f^{-1}(y_0)$. By Riemann-Roch, there is a meromorphic function $g \in K(X) \setminus K$ such that x_0 is the only pole of g. Then we have $K(X) = K(f, g)$ (as the field extension $K(X)/K(f, g)$ is a subextension of $K(X)/K(f)$ and of $K(X)/K(g)$, hence the corresponding morphism of curves is both unramified and totally ramified at x_0). We assume that we have chosen g in such a way that the order m of the pole is minimal. Then we have

$$T := \{f \in K(X) \mid \mathrm{ord}_{x_0}(f) \geq -m \text{ and } \mathrm{ord}_x(f) \geq 0 \text{ for all } x \in X \setminus \{x_0\}\} = K \oplus Kg;$$

since for any $f_1, f_2 \in T$ with $\mathrm{ord}_{x_0}(f_i) = -m$, $i = 1, 2$, there is a constant $\alpha \in K$ with $-\mathrm{ord}_{x_0}(f_1 - \alpha f_2) < m$, and then $f_1 - \alpha f_2$ is a constant function, as m is minimal. By the choice of y_0, the meromorphic function $f - y_0$ on X is a local parameter on X in x_0; if $K = \mathbb{C}$, this means, in the language of Riemann surfaces, that $f - y_0$ yields a chart of $X(\mathbb{C})$ in a neighborhood of x_0 which maps x_0 to 0. There is a unique function $g' \in T$ such that the leading coefficient and the constant coefficient in the Laurent expansion of g' with respect to the local parameter $f - y_0$ are equal to 1 and 0, respectively. We then assume that $g = g'$. We now claim that the minimal polynomial of g over $K(f)$ has coefficients in $k(f)$ where k is a finite extension of $M(X, t)$. Then, the field extension $K(X)/K(f)$ is defined over k. By the dictionary between curves and function fields, this means Lemma A.0.8 is proved.

As for the proof of the above claim, we denote by $U(X,f,x_0)$ the subgroup of $U(X,f)$ consisting of all $\sigma \in \mathrm{Aut}(K)$ such that there is an isomorphism $\sigma_X : X^\sigma \to X$ of curves over K such that the diagram

$$\begin{tikzcd}
X^\sigma \arrow[r, "\sigma_X"] \arrow[d, "f^\sigma"] & X \arrow[d, "f"] \\
(\mathbb{P}^1_K)^\sigma \arrow[r, "\mathrm{Proj}(\sigma)"] & \mathbb{P}^1_K
\end{tikzcd}$$

commutes and such that $\sigma_X(x_0^\sigma) = x_0$, where x_0^σ denotes the point on X^σ/K corresponding to x_0. Note that σ_X is unique since $\mathrm{Aut}(f)$ acts freely on the fibre $f^{-1}(y_0)$. Thus, mapping σ to the automorphism of the function field $K(X)$ induced by σ_X yields an action of $U(X,f,x_0)$ on $K(X)$ by K-semilinear field automorphisms which fix $f \in K(X)$. Being the stabilizer of $[x_0]$ under the action $(\sigma,[x_0]) \mapsto [\sigma_X(x_0^\sigma)]$ of $U(X,f)$ on $f^{-1}(x_0)/\mathrm{Aut}(f)$, the subgroup $U(X,f,x_0)$ has finite index in $U(X,f)$. The meromorphic function $g \in K(X)$ and hence the minimal polynomial of g over $K(f)$ are invariant under the action of $U(X,f,x_0)$ defined above since the image of g under $\sigma \in U(X,f,x_0)$ has the same three defining properties as g. Now applying Lemma A.0.4 completes the proof of the claim (thus also the proof of Lemma A.0.8). □

Remark A.0.9. From the proof we actually see that X and f are defined over $M(X,f)$ itself provided that f is Galois.

We can now prove Theorem A.0.2.

Proof of Theorem A.0.2. The theorem follows directly from Lemma A.0.7 and Lemma A.0.8. □

B Appendix B: Estimation of oriented binary trees

We establish some combinatorial estimations for getting the claimed estimation 7.1.11 in Section 7.1.

First we recall the definition of oriented binary trees for the sake of integrity.

Definition B.0.1 (Oriented binary tree)**.** An *oriented binary tree* is a pair $(\mathcal{T},\varphi)$, where $\mathcal{T}$ is a binary tree, and φ is an orientation of the dual graph of $\mathcal{T}$, in the sense of Definition 7.1.6.

We also have a notion of rooted oriented binary trees:

Definition B.0.2 (Rooted oriented binary tree)**.** An *rooted oriented binary tree* is a triple $(\mathcal{T},\varphi,v)$, where $(\mathcal{T},\varphi)$ form an oriented binary tree, and v is a degree-1 vertex.

Note that we have natural notions of isomorphism for both oriented binary trees and rooted oriented binary trees.

Now recall that we want to estimate the number of isomorphism classes of oriented binary trees on $2n$ vertices, as Theorem 7.1.11 states that this number is exactly the number of linearly-equivalent classes of pairs (f, g) obtaining the minimal degree $\deg(f^3 - g^2) = n+1$. For this, we first count the number F_n of isomorphism classes of *rooted* oriented binary trees on $2n$ vertices. It turns out that this number is easier to count directly.

Proposition B.0.3. *The number F_n of isomorphism classes of rooted oriented binary trees on $2n$ vertices satisfy the following asymptotic formula*

$$F_n \sim cn^{-3/2}4^n.$$

Proof. Note that we may split such a rooted oriented oriented binary tree into a pair of new (co-)rooted oriented oriented binary trees by taking out the old root, ordered according to the orientation attached to the face associated to the unique neighbour of the old root, giving rise to the following recurrence formula

$$F_n = \sum_{i+j=n} F_i F_j.$$

Clearly we have $F_0 = 0$, $F_1 = 1$ by definition. Therefore we have the following identity for the generating function $F(x) = \sum\limits_{n=0}^{\infty} F_n x^n$:

$$F(x) = x + F^2(x),$$

thus

$$F(x) = \frac{1}{2}(1 - \sqrt{1-4x}),$$

this tells us

$$F_n = (-1)^{n-1}\frac{1}{2}4^n\binom{1/2}{n} = \frac{\binom{2n}{n}}{2(2n-1)} \sim cn^{-3/2}4^n,$$

for a constant $c > 0$ independent of n. □

We can now estimate the number ϕ_n of isomorphism classes of oriented binary trees on $2n$ vertices.

Theorem B.0.4. *The number ϕ_n of isomorphism classes of $\mathfrak{S}_3$-trees on $2n$ vertices satisfy the following inequality*

$$c_1 n^{-5/2}4^n < \phi_n < c_2 n^{-3/2}4^n$$

for some positive reals c_1, c_2 independent of n.

Proof. Consider the surjective forgetful map from the set of isomorphic classes of *rooted* oriented binary trees on $2n$ to the set of isomorphism classes of oriented binary trees on $2n$ vertices, then the number of elements of each fiber is between 1 and $2n$ by definition, therefore

$$\frac{1}{2n}F_n \leq \phi_n \leq F_n,$$

applying Proposition B.0.3 then gives the desired inequality

$$c_1 n^{-5/2}4^n < \phi_n < c_2 n^{-3/2}4^n$$

for some positive reals c_1, c_2 independent of n. □

C Appendix C: Basic height theory revisited

We recite some basics of the height theory from Diophantine geometry in this section.

Generally speaking, heights are functions from the set of algebraic points on a variety to $\mathbb{R}$, that allow us to "count" algebraic points on varieties. It possesses a certain finiteness property which claims that a set of points with bounded height and degree is necessarily finite.

We first define heights on projective spaces $\mathbb{P}^n$.

Assumption C.0.1. We assume K is one of the following:

- Number field: i.e. a finite extension $K/\mathbb{Q}$;
- Function field: i.e. $K = k(B)$, where k is an arbitrary field, and B is a geometrically integral smooth projective curve over k.

Remark C.0.2. In some applications (e.g. Mordell-Weil theorem), we also require k is finite. Then Assumption C.0.1 is just saying K is a global field.

Notation C.0.3 (Valuations and places)**.**

- Number field: Let $K/\mathbb{Q}$ be a finite extension. We use $M_K, M_{K,f}, M_{K,\infty}$ to denote the set of places of K, the set of finite places of K, the set of infinite places of K, respectively. Then a finite place $v \in M_{K,f}$ corresponds to a prime ideal $\wp_v \subset \mathcal{O}_K$, and the *$v$-adic norm* is given by
$$|x|_v = N(\wp_v)^{-\operatorname{ord}_v(x)},$$
where
$$N(\wp_v) = \#(\mathcal{O}_K/\wp_v)$$
denotes the norm of $\wp_v$; and an infinite place v corresponds to an embedding $\sigma_v : K \to \mathbb{C}$, and the *$v$-adic norm* is given by the restriction of the complex absolute value.
- Function field: Let $K = k(B)$ be a function field over k. We use M_K to denote the set of places of K. Then a place $v \in M_K$ corresponds to a closed point of B, and the *v-adic norm* is given by
$$|x|_v = e^{-\deg(v)\cdot\operatorname{ord}_v(x)}$$
where
$$\deg(v) = [k(v) : k]$$
denotes the degree of the residue field at v.

Theorem C.0.4 (Product formula)**.** *Let K be a field satisfying Assumption C.0.1, then we have*
$$\prod_{v\in M_K} |x|_v = 1,$$
for any $x \in K^\times$.

Proof. We omit the proof and refer the interested reader to [6]. □

Remark C.0.5. The product formula C.0.4 plays a fundamental role in Diophantine Geometry, e.g., in the well-definedness of the naive height on projective space, the development of intersection theory on arithmetic varieties...

Definition C.0.6 (Naive height on $\mathbb{P}^n$)**.** Let K be a field satisfying Assumption C.0.1, define the naive height

$$h = h_K : \mathbb{P}^n(\overline{K}) \to \mathbb{R}$$

by

$$h_K(x_0, \ldots, x_n) = \frac{1}{[K' : K]} \sum_{v \in M_{K'}} \log \max\{|x_0|_v, \ldots, |x_n|_v\},$$

where K' is a finite extension of K containing all the coordinates $x_0, \ldots, x_n \in \overline{K}$.

Remark C.0.7. By a direct computation we may verify that Definition C.0.6 is independent of the choice of K'. Independence of coordinates is guaranteed by the product formula C.0.4.

Proposition C.0.8. *Let $x \in \mathbb{P}^n(\overline{K})$ be an algebraic point, then*

(1) $h(x) \geq 0$;

(2) $h(x) = 0$ if and only if

(3) $h(x_0^m, \ldots, x_n^m) = |m| \cdot h(x_0, \ldots, x_n)$.

Proof. The first claim follows from product formula C.0.4. Number field case of the second claim is a well-known result, whereas the function field case is by direct verification, so is the third claim. □

The following result ensured that we could use height to "count" algebraic points.

Theorem C.0.9 (Northcott property)**.** *Let K be a global field. Then the set*

$$\{x \in \mathbb{P}^n(\overline{K}) \mid \deg(x) < c_1, h(x) < c_2\}$$

is finite, for any $c_1, c_2 \in \mathbb{R}$.

Proof. We omit the proof and refer the interested readers to [9]. □

We can now define heights on general projective varieties.

Convention C.0.10. In our context, a variety is an integral scheme which is separated and of finite type over the base field; a curve is a 1-dimensional variety; a surface is a 2-dimensional variety.

The upshot is that we embed projective varieties into $\mathbb{P}^n$, and use the heights induced by the heights C.0.6 on $\mathbb{P}^n$.

Definition C.0.11 (Naive height on projective varieties)**.** Let X/K be a projective variety, L an ample line bundle on X, and

$$i : X \to \mathbb{P}^N_K$$

a close immersion, with

$$i^*\mathcal{O}_{\mathbb{P}^N_K}(1) \cong mL$$

for some integer $m \geq 1$. Then *the height*

$$h_{(L,m,i)} : X(\overline{K}) \to \mathbb{R}$$

associated to the triple (L, m, i) is the composition of

$$X(\overline{K}) \xrightarrow{i} \mathbb{P}^N_K(\overline{K}) \xrightarrow{\frac{1}{m}h} \mathbb{R},$$

where h is the naive height on $\mathbb{P}^n$ defined in Definition C.0.6.

The following theorem shows that, the height $h_{(L,m,i)}$ in Definition C.0.11 only depends on L, up to $O(1)$. Moreover, non-ample line bundles can also induce heights in a somewhat natural way, giving lots of heights on a variety. For this reason, Theorem C.0.12 is called the height machine.

Theorem C.0.12 (Height machine)**.** *Let K be a field satisfying Assumption C.0.1, and X/K a projective variety. Then there exists a unique homomorphism*

$$\mathcal{H} : \mathrm{Pic}(X) \to \{\text{maps } X(\overline{K}) \to \mathbb{R}\}/\{\text{bounded maps}\}, L \mapsto \mathcal{H}_L,$$

such that

$$\mathcal{H}_L = h_{(L,m,i)} + O(1),$$

for any ample line bundle L on X, and close immersion

$$i : X \to \mathbb{P}^n$$

with

$$i^*\mathcal{O}_{\mathbb{P}^n}(1) \cong mL.$$

Remark C.0.13. The uniqueness of $\mathcal{H}$ is already contained in the statement of Theorem C.0.12, as any line bundle can be written as a difference of two very ample ones. The hard part is actually the existence of such an $\mathcal{H}$. As for a complete proof, we refer the interested readers to [9].

We call the heights given by the height machine C.0.12 *Weil heights*.

Definition C.0.14 (Weil height)**.** Let K and X/K be as above. For any line bundle L on X, any function

$$h_L : X(\overline{K}) \to \mathbb{R}$$

in the class $\mathcal{H}_L$ is called a *Weil height* associated to L.

We have a projection formula for Weil heights.

Corollary C.0.15 (Projection formula)**.** *Let K be a field satisfying Assumption C.0.1, and $f : X' \to X$ a morphism of varieties over K. Let $L \in \mathrm{Pic}(X)$ be a line bundle on X. Then*

$$h_{f^*L} = f^*h_L + O(1).$$

i.e., the following diagram

$$\begin{tikzcd} X'(\overline{K}) \arrow[rr, "f"] \arrow[dr, "h_{f^*L}"'] & & X(\overline{K}) \arrow[dl, "h_L"] \\ & \mathbb{R} & \end{tikzcd}$$

commutes up to $O(1)$.

We also have Northcott property for Weil heights.

Theorem C.0.16 (Northcott property)**.** *Let K be a global field, X/K a projective variety, L an ample line bundle on X, and $h_L : X(\overline{K}) \to \mathbb{R}$ the associated Weil height. Then the set*

$$\{x \in X(\overline{K}) \mid \deg(x) < c_1, h(x) < c_2\}$$

is finite, for any $c_1, c_2 \in \mathbb{R}$.

Proof. This follows directly from the Northcott property C.0.9 for the naive height on $\mathbb{P}^n$. □

The following interpretation of Weil heights as an intersection number justifies our definition.

Theorem C.0.17 (Height = intersection number)**.** *Let $K = k(B)$ be a function field, where B is a regular and geometrically integral projective curve. Let*

$$h : \mathbb{P}^n(\overline{K}) \to \mathbb{R}$$

be the naive height on $\mathbb{P}^n_{\overline{K}}$. Then

$$h(x) = \frac{1}{\deg(x)} \deg(\mathcal{O}_{\mathbb{P}^n_B}(1)|_{\tilde{x}}),$$

for any $x \in \mathbb{P}^n(\overline{K})$, where $\tilde{x} \subset \mathbb{P}^n_B$ is the zariski closure of the image of the composition of

$$\mathrm{Spec}\,\overline{K} \to \mathbb{P}^n_K \to \mathbb{P}^n_B.$$

Proof. By base change, we may assume $\deg(x) = 1, x \in \mathbb{P}^n(K)$, and $\tilde{x}$ is the section corresponding to x. Then the equation becomes

$$h(x) = \deg(\mathcal{O}_{\mathbb{P}^n_B}(1)|_{\tilde{x}}).$$

Note that $\deg(\mathcal{O}_{\mathbb{P}^n_B}(1)|_{\tilde{x}})$ is just the intersection number $H \cdot \tilde{x}$ for any hyperplane section H of $\mathbb{P}^n_B$, for which we have

$$H \cdot \tilde{x} = \sum_{\substack{v \in B \\ \text{closed point}}} m_v \deg(v),$$

where the $m_v \geq 0$ are the intersection multiplicities. Now write $x = (x_0, \ldots, x_n)$ with $x_i \in K$ and $x_0 \neq 0$. Then take $H = \mathrm{V}(x_0)$ in $\mathbb{P}^n_B$. By a direct computation we see that

$$m_v = \mathrm{ord}_v(x_v) - \min_{0 \leq i \leq n} \{\mathrm{ord}_v(x_i)\},$$

for any closed point $v \in B$. This then completes the proof of Theorem C.0.17. □

Corollary C.0.18 (General version of Theorem C.0.17)**.** *Let $K = k(B)$ be as in Theorem C.0.17, X/K a projective variety, and L a line bundle on X. Let $(\mathcal{X}, \mathcal{L})$ be an integral model of (X, L) over B. Then the function*

$$h_{\mathcal{L}} : X(\overline{K}) \to \mathbb{R},\ x \mapsto \frac{1}{\deg(x)} \deg(\mathcal{L}|_{\tilde{x}})$$

is a Weil height associated to L, where $\tilde{x}$ is the zariski closure of the image of the composition

$$\operatorname{Spec} \overline{K} \xrightarrow{x} X \to \mathcal{X}.$$

Remark C.0.19. In the number field (i.e. "arithmetic") case, by adding Hermitian metrics as part of the data "at ∞", and developing intersection theory in parallel, we may obtain a function (analogous to the one defined above)

$$h_{\overline{\mathcal{L}}} : X(\overline{K}) \to \mathbb{R},\ x \to \frac{1}{\deg(x)} \overline{\mathcal{L}} \cdot \tilde{x},$$

where $(\mathcal{X}, \overline{\mathcal{L}})$ is an "arithmetic model" of (X, L), and $\tilde{x}$ is the zariski closure of the image of the composition

$$\operatorname{Spec} \overline{K} \xrightarrow{x} X \to \mathcal{X}.$$

The upshot is that we can use this "height" function to develop height machine in the number field case in parallel to our original approach. In fact, this is exactly the fundamental idea of Arakelov theory.

We include the following result used in Section 7.2 for the sake of integrity.

Theorem C.0.20 (Néron)**.** *Let X be a non-singular projective variety, L and L_1 be two elements of* $\mathrm{Pic}(X)$*, with* $\deg L = 0$ *and L_1 ample. Then we have*

$$|h_L(x)| \leq O(\sqrt{h_{L_1}(x)} + 1).$$

Proof. See Serre's book [9]. □

D Appendix D: Étale fundamental group revisited

In the setting of algebraic geometry, the topological fundamental group of the underlying (Zariski) topological space contains very little information about the scheme, as Zariski topology is very "coarse" in some sense. The key is to encapsulate the essence of covering theory in a more categorical way, i.e. the so-called Grothendieck's Galois theory; and to realize that, (finite) étale morphisms in algebraic geometry are the analog of (finite) coverings in classical topology. However, many technical details arise. For example, there are no analog of universal covering in the setting of schemes, one would have to either choose to interpret the "algebraic fundamental group" of a scheme as an automorphism group of the category of finite étale morphisms over the scheme, or enlarge the category of schemes so that an "universal covering" may exist. Our approach here is the former.

We work under the framework of Grothendieck's Galois theory formalism. For details, see [11]. We first recall some notions from this framework, for the sake of integrity.

Definition D.0.1 (Galois category)**.** A *Galois category* is a pair $(\mathcal{C}, F)$, where $\mathcal{C}$ is a category and $F : \mathcal{C} \to Fin$ is a functor from $\mathcal{C}$ to the category of finite sets, such that

(1) $\mathcal{C}$ has finite limits and finite colimits;

(2) Every object in $\mathcal{C}$ is a finite coproduct of connected objects in $\mathcal{C}$;

(3) F detects isomorphisms and is exact.

Here an object X in $\mathcal{C}$ is called *connected* if $\mathrm{Aut}(X)$ acts freely on X. The functor F is usually addressed as the *fiber functor* of the Galois category $(\mathcal{C}, F)$.

Example D.0.2. Take $\mathcal{C}$ to be the category of finite coverings of a fixed pointed (locally simply connected) space, and F to be the literal "fiber" functor. Then $(\mathcal{C}, F)$ form a Galois category. This is one of the most import examples to keep in mind.

The "universal Galois group" $\mathrm{Aut}(F) := \varprojlim_X \mathrm{Aut}(F(X))$ in fact determines completely the structure of $\mathcal{C}$:

Theorem D.0.3. *Let $\mathcal{C}$ be a Galois category with fiber functor F, and $G = \mathrm{Aut}(F)$ be the profinite automorphism group of F. Then there is an equivalence of categories between $\mathcal{C}$ and the category $GFin$ of finite G-sets.*

After some easy checking, we may obtain:

Proposition D.0.4. *Let X be a connected scheme with a geometric point $\overline{x} : \mathrm{Spec}\,\overline{k} \to X$. Then the category $\mathbf{F\acute{E}t}_X$ of finite étale morphisms to X, together with the induced fiber functor*

$$F_{\overline{x}} : \mathbf{F\acute{E}t}_X \to \mathbf{Fin},\ (f : Y \to X) \mapsto |Y_{\overline{x}}|,$$

form a Galois category.

We can now define the étale fundamental group.

Definition D.0.5 (Étale fundamental group)**.** Let X be a connected scheme with a geometric point $\overline{x} : \operatorname{Spec} \overline{k} \to X$, and $F_{\overline{x}}$ be as above. Then the *étale fundamental group* $\pi_1^{\text{ét}}(X, \overline{x})$ *of* X *at* $\overline{x}$ is the automorphism group $\operatorname{Aut}(F_{\overline{x}})$ of $F_{\overline{x}}$.

By a routine reduction to curve case, in which we apply Grothendieck's comparison theorem along with a result of Shafarevich (see [11]), we see that:

Theorem D.0.6 (Grothendieck)**.** *Let X be a smooth projective scheme over an algebraically closed field k, and $\overline{x} : \operatorname{Spec} \overline{k} \to X$ be a geometric point. Then $\pi_1^{\text{ét}}(X, \overline{x})$ is topologically finitely generated as a profinite group.*

Remark D.0.7. The curve case can also be proved by deforming the curve to a curve of characteristic zero, for which the result follows from Lefschetz principle and the structure of the topological fundamental group of Riemann surfaces of finite type. Moreover, by applying de Jong's theory of alterations, we may only assume X is connected in Theorem D.0.6. Note that even properness of X is not necessary to deduce topologically finitely generatedness.

It is sometimes of great benefit to also have a notion of étale fundamental groups at certain "missed" points of a scheme, e.g., the "point" 0 of the truncated projective line $\mathbb{P}^1_{\overline{\mathbb{Q}}} \setminus \{0, 1, \infty\}$. This is the so-called "étale fundamental group with tangential basepoints", which we will now discuss.

Definition D.0.8 (Tangential basepoint)**.** Let X be an integral proper normal curve over a field k. A *k-rational tangential basepoint of X* is just a $k((t))$-point $x_v : \operatorname{Spec} k((t)) \to X$ of X.

Remark D.0.9. Let $\overline{X}$ be the compactification of X, then such a x_v in Definition D.0.8 gives a k-rational point $x : \operatorname{Spec} k \to \overline{X}$ along with a tangent vector $v \in T_x\overline{X}$ given by the parameter t, hence the name. These data allow us to recover more information from the fiber over x_v compared to x, such as the ramification indices over x.

The following theorem of Deligne allow us to apply Grothendieck's Galois theory formalism as in Definition D.0.5.

Theorem D.0.10 (Deligne)**.** *Let X be an integral normal curve over a field k of characteristic* 0*, and $x_v : \operatorname{Spec} k((t)) \to X$ be a k-rational tangential basepoint. Then the category $F\acute{E}t_X$ of finite étale morphisms to X, together with the induced fiber functor*

$$F_{x_v} : \mathbf{F\acute{E}t}_X \to \mathbf{Fin}, \ (f : Y \to X) \mapsto \{(y, e) \mid y \in \overline{f}^{-1}(x), e \in \overline{f}^{-1}(v) \cap T_y\overline{Y}\},$$

form a Galois category.

We can now make the following definition.

Definition D.0.11. Let X be an integral normal curve over a field k of characteristic 0, and $x_v : \operatorname{Spec} k((t)) \to X$ be a k-rational tangential basepoint. Then the *étale fundamental group* $\pi_1^{\text{ét}}(X, x_v)$ *of* X *at* x_v is the automorphism group $\operatorname{Aut}(F_{x_v})$ of F_{x_v}.

References

[1] N. D. Elkies, ABC implies Mordell. *International Mathematics Research Notices*, **1991**, no. 7, 99–109.

[2] E. Girondo and G. González-Diez, *Introduction to compact Riemann surfaces and dessins d'enfants*. London Math. Soc. Stu. Texts 79, Cambridge Univ. Press, 2012.

[3] A. Grothendieck, *Esquisse d'un Programme*, Preprint, 1985.

[4] B. Köck, Belyi's theorem revisited. *Beiträge Algebra Geom.*, **45** (2004), no.1, 253–265.

[5] W. Melanie, Belyi-extending maps and the Galois action on dessins d'enfants, *Publications of the Research Institute for Mathematical Sciences*, **42** (2006), 721–737.

[6] J. Neukirch, *Algebraic Number Theory*, Springer, 1999.

[7] A. Petrov, *Universality of the Galois action on the fundamental group of* $\mathbb{P}^1 \setminus \{0, 1, \infty\}$, arXiv: 2109.09301 (2021).

[8] L. Schneps, Dessins d'enfants on the Riemann sphere, in *The Grothendieck Theory of Dessins d'Enfants*, London Math. Soc. Lect. Note Ser. 200, Cambridge Univ. Press, 1994.

[9] J-P. Serre, *Lectures on the Mordell-Weil Theorem*, Springer, 1997.

[10] G. Shabat and V. Voevodsky, Drawing curves over number fields, in *The Grothendieck Festschrift, Volume III*, Springer, 2007.

[11] A. Shmakov, *Galois Representations in Étale Fundamental Groups and the Profinite Grothendieck-Teichmüller Group*, Preprint.

[12] X. Yuan, *Lectures on Arakelov geometry*, lecture notes, 2022.

[13] U. Zannier, On Davenport's bound for the degree of $f^2 - g^3$ and Riemann's Existence Theorem, *Acta Arithmetica*, **71** (1995), no. 2, 107–137.

[14] Y. Zhao, *Géométrie Algébrique et Géométrie Analytique*, unpublished notes, 2013.

[15] A. Zvonkine, Belyi functions: examples, properties, and applications. *Application of Group Theory to Combinatorics*, Jul 2007, Pohang, South Korea. pp. 161–180.

The Bott Periodicity Theorem For Complex vector Bundles

温家睿　数学科学学院　2000010859

2022 年 6 月 14 日

Bott Periodicity Theorem was first discribed by Raoul Bott in 1959, in which he established a periodicity in the homotopy groups of classical groups. And later in 1964, Atiyah and Bott noticed that the periodicity theorem is related with K-theory. It was then reformulated as $K(X \times S^2) = K(X) \otimes K(S^2)$, which is a fundamental theorem in K-theory. I'll introduce the statement of the periodicity theorem in terms of K-theory, and then offer the proof by Atiyah and Bott.

1 Preliminaries on vector bundle

As it's not the main part of our issue, I'll only present the definitions and theorems without proof that may be helpful in our discussion of the periodicity theorem.

Definition 1.1. *Let B be a topological space, a* ***complex vector bundle*** *over B is a topological space endowed with:*

(1)a continuous map $p : E \to B$ (called the projection)

(2)a finite dimensional complex vector space structure in each $E_b = p^{-1}(b), b \in B$. And these must satisfy the condition of ***local triviality****: For each point $b \in B$, there exists a neighborhood $U \subset B$, an integer $n \geq 0$, and a homeomorphism $h : U \times \mathbb{C}^n \to p^{-1}(U)$, so that for each $b \in U$, $x \to h(b, x)$ is an isomorphism between the vector space $\mathbb{C}^n$ and the vector space $p^{-1}(b)$. Moreover, if U can be chosen to be E itself, then E is called a* ***trivial bundle****.*

In a vector bundle, B is called the ***base space****, E is the* ***entire space****, and for each $b \in B$, $E_b := p^{-1}(b)$ is called the* ***fiber*** *over b.*

In this article, when we say ”vector bundle” we mean the ”complex vector bundle”.

Definition 1.2. *A* ***section*** *of a vector bundle E is a continuous map $s : B \to E$ with $ps = id$. The space of all sections of E is denoted by $\Gamma(E)$.*

Definition 1.3. *If E and F are two vector bundle over X then a **homomorphism** of E into F is a continuous map $\phi : E \to F$ commuting with the projections and inducing a vector space homomorphism $\phi_x : E_x \to F_x$ for each $x \in X$. If ϕ_x is an isomorphism for all x, then ϕ is an **isomorphism** of E to F. And we will denote the set of isomorphism classes od n-dimensional complex vector bundles over X by* $\mathrm{Vect}^n(B)$.

The union of all the vector space $\mathrm{Hom}(E_x, F_x)$ for each $x \in X$ has a natural topology making it into a vector bundle $\mathrm{Hom}(E, F)$, and a section of $\mathrm{Hom}(E, f)$ is a homomorphism of E into F. Similarly, we have $\mathrm{ISO}(E, F)$.

Hom is an example of natural operations on vector space carrying over to vector bundles. In addition we can define the direct sum $E \oplus F$,the tensor product $E \otimes F$, and the dual E^*. Canonical isomorphisms also go over to bundles, for instance, $\mathrm{Hom}(E, F) \simeq E^* \otimes F$. And we'll denote the iterated tensor product $E \otimes E \otimes \cdots \otimes E$($k$ times) by E^k. And if L is a ***line bundle***, i.e. a bundle of dimension one, we shall write L^{-1} for L^* and K^{-k} for $(L^*)^k$. Thus the line-bundles over X can form a multiplicative group with L^{-1} as the inverse of L, and the unit of this group is the trivial line-bundle $X \times \mathbb{C}$(denoted by 1).

Definition 1.4. *Let $f : Y \to X$ be a continuous map and let E be a vector bundle over X, the **induced bundle** or some called **pullback bundle** is a vector bundle $f^*(E)$ over Y that $f^*(E) = \{(y, v) \in Y \times E | f(y) = p(v)\}$ with projection $p'(y, v) = y$. It shows that* $f^* : \mathrm{Vect}^n(X) \to \mathrm{Vect}^n(Y)$ *is a pullback of $f : Y \to X$.*

Remark. *Note that if $f : Y \to X$ is the inclusion map of $Y \subset X$, then $f^*(Y) \simeq E|Y$.*

Here comes an important proposition:

Proposition 1.1. *Let Y be a compact space, $f_t : Y \to X$ a homotopy (*$0 \leq t \leq 1$*) and E a vector bundle over X. Then $f_0^* E \simeq f_1^* E$.*

Vector bundles are frequently constructed by a clutching construction. Let $X = X_1 \cup X_2$,$A = X_1 \cap X_2$, and suppose all those spaces are compact. Assume that E_i is a vector bundle over X_i $(i = 1, 2)$, and $\phi : E_1|A \to E_2|A$ is an isomorphism. Then we define the vector bundle $E_1 \cup_\phi E_2$ on X to be the quotient of $E_1 \sqcup E_2$ by the equivalence relation which identifies $e_1 \in E_1|A$ with $\phi(e_1) \in E_2|A$. It's easy to verify that $E_1 \cup_\phi E_2$ is a vector bundle over X.

Here are some properties of this construction:

Proposition 1.2. *If E is a bundle over X and $E_i = E|X_i$, then the identify defines an isomorphism $1_A : E_1|A \to E_2|A$, and $E_1 \cup_{1_A} E_2 \simeq E$.*

Proposition 1.3. *If $\beta_i : E_i \to E'_i$ are isomorphisms on X_i and $\phi'\beta_1 = \beta_2\phi$, then $E_1 \cup_\phi E_2 \simeq E'_1 \cup_{\phi'} E'_2$*

Proposition 1.4.

$$E_1 \cup_\phi E_2 \oplus E'_1 \cup_{\phi'} E'_2 \simeq E_1 \oplus E'_1 \cup_{\phi\oplus\phi'} E_2 \oplus E'_2,$$

$$E_1 \cup_\phi E_2 \otimes E'_1 \cup_{\phi'} E'_2 \simeq E_1 \otimes E'_1 \cup_{\phi\otimes\phi'} E_2 \otimes E'_2,$$

$$(E_1 \cup_\phi E_2)^* \simeq E^*_1 \cup_{(\phi^*)^{-1}} E^*_2.$$

And the following is the most important proposition in our proof of the periodicity theorem.

Proposition 1.5. *The iso morphism class of $E_1 \cup_\phi E_2$ depends only on the homotopy class of the isomorphism $\phi : E_1|A \to E_2|A$.*

And here are two facts we might use in our proof:

Proposition 1.6. *If P is a **projection operator** for a vector bundle E, namely, P is an endomorphism with $P^2 = P$, then PE and $(1 - P)E$ have an induced vector bundle structure and $PE \oplus (1 - P)E = E$.*

Proposition 1.7. *A **metric** on E is a section of* $\mathrm{End}(E)$ *which is positive define Hermitian for each $x \in X$. If X is a compact space, then the metric on E always exists.*

Finally, we come to the definition of K-group $K(X)$. Let X be a compact space. First we note that the isomorphic classes of vector bundles over X form an abelian semigroup with the option $\oplus$. By taking a generator a in a semigroup, and adding relations $[a] = [b] + [c]$ whenever $a = b + c$, we can obtain an abelian group from the semigroup we have. And the group we get from the isomorphic classes of vector bundles over X is denoted by $K(X)$. Further more, the operation *otimes* induces a multiplication in $K(X)$ turning it into a commutative ring. A continuous map $f : Y \to X$ also induces a ring homomorphism $f^* : K(X) \to K(Y)$ where $f^*[E] = [f^*E]$. And obviously, if X is a point then $K(X) \simeq \mathbb{Z}$.

2 Statement of the periodicity theorem

Atiyah and Bott actually proved a little further than $K(X) \otimes K(S^2) \simeq K(X \times S^2)$, the statement is established as follows.

If E is a vector bundle, then by deleting the 0-section and dividing out by the action of non-zero scalars we obtain a space $P(E)$ called the ***projection bundle*** of E. There is a natural map $P(E) \to X$ and the inverse image of $x \in X$ denoted by $P(E_x)$ is the complex projective space. If we assign to each $y \in P(E_x)$ the one-dimensional subspace of E_x which corresponds to it, we obtain a line-bundle over $P(E)$, and this line-bundle is denoted by $H*$, while its dual is defined by H. In addition, the projection $P(E) \to X$ induces a ring homomorphism $K(X) \to K(P(E))$, thus $K(P(E))$ becomes a $K(X)$-algebra. The periodicity theorem claims the structure of this algebra:

Theorem 2.1. *Let L be a line-bundle over a compact space X, and H be the line-bundle over $P(L \oplus 1)$ defined above. Then,as a $K(X)$-algebra, $K(P(L \oplus 1))$ is generated by $[H]$ with a single relation $([H] - [1])([L][H] - [1]) = 0$.*

Note that if X is a point, then $P(L \oplus 1) \simeq \mathbb{CP}^1 \simeq S^2$ and theorem 2.1 implies that $K(S^2)$ is generated by $[1]$ and $[H]$ with relation $([H] - 1)^2 = 0$. Hence in the case when L is trivial, $P(L \oplus 1) \simeq X \times S^2$ and we have:

Corollary 2.1. *Let $\pi_1 : X \times S^2 \to X$, $\pi_2 : X \times S^2 \to S^2$ denote the projections. Then the homomorphism*

$$f : K(X) \otimes K(S^2) \to K(X \times S^2)$$

$f(a \otimes b) = \pi_1^*(a)\pi_2^*(b)$ is a ring isomorphism.

The idea of the proof comes from a basic observation. Since S^2 can be written as the union of its upper and lower hemispheres D_+ and D_- with $D_+ \cap D_- =$. A clutching function on $D_- \cup D_+$ can construct a vector bundle on S_2, since D_- and D_+ are contractible hence having trivial vector bundle. And by proposition 1.2 such a clutching function only depends on the homotopy classes of itself, thus an element in the homotopy classes $[S^1, GL_n(\mathbb{C})]$ can define an element in $\mathrm{Vect}^n(S^2)$ and in fact it is an isomorphism.

We want to do the same for bundles over $P(L \oplus 1)$. For any x there is a natural embedding $L_x \to P(L \oplus 1)_x$ given by $y \to (y \oplus 1)$. In this way $P(L \oplus 1)_x$ can be regarded as the compactification of L_x by adding "the point at infinity". Therefore, we get an embedding of L in $P = P(L \oplus 1)$ by adding the "section at infinity". Since there's a metric in L, we can choose a unit circle bundle $S \subset L$ in this metric. We identify L with a subspace of P so that $P = P^0 \cup P^\infty$, $S = P^0 \cap P^\infty$, where P^0 is the closed disc bundle interior of S which contains the 0-section and P^∞ is the closed disc bundle exterior to S containing the ∞-section. And we'll denote the projections $S \to X$, $P^0 \to X$, $P^\infty \to X$ respectively by π,π_0,π_∞.

Suppose that E^0 and E^∞ are two vector bundles over X and that $f \in \operatorname{Iso}(\pi^* E^0, \pi^* e^\infty)$. Then we can construct the vector bundle $\pi_0^* E^0 \cup_f \pi_\infty^* E^\infty$ over P. And we'll denote this bundle by (E^0, f, E^∞). Actually, any vector bundles over P can be constructed in such way, and this is shown by the following lemma.

Lemma 2.1. *Let E be any vector bundle over P and let E^0, E^∞ be the vector bundles over X induced by the 0 -section and ∞-section respectively. Then there exists $f \in \operatorname{ISO}(\pi^* E^0, \pi^* E^\infty)$ such that $E \cong (E^0, f, E^\infty)$. And it's uniquely defined up to homotopy.*

Proof. Let $s_0 : X \to P^0$ be the 0 -section. Then $s_0 \pi_0$ is homotopic to the identity map of P^0, and so by Proposition 1.1 we have an isomorphism

$$f_0 : E \mid P^0 \to \pi_0^* E^0.$$

Similarly we can do the same for $E|P^\infty$. And the lemma follows from taking $f = f_\infty f_0^{-1}$. $\square$

The simplest bundle constructed in this way is $(F, 1, F)$, which is the bundle over P induced by projection $P \to X$. And written this in $K(P)$ we have $[(F, 1, F)] = [F][1]$, where the multiplication on the right-hand side is the module multiplication of $K(X)$ in $K(P)$.

In the remaining part of the proof, we'll try to simplify the clutching function as possible. The first thing we do is approaching it by a Laurent series.

3 Laurent cluntching function

When L is the trivial line-bundle over X, S is the trivial circle bundle $X \times S^1$ so the points of S are represented by pairs (x, z) with $x \in X$ and $z \in \mathbb{C}$ with $|z| = 1$. Thus z can be regarded as a function on S, so is z^{-1}. Generally, we define z to be a section in $\pi^*(L)$ defined by the inclusion $S \to L$. Then we may regard z^k as the section of $\pi^*(L^k)$. If $a_k \in \Gamma \operatorname{Hom}(L^k \otimes E^0, E^\infty)$, then $a_k z^k := \pi^*(a_k) z^k \in \Gamma \operatorname{Hom}(\pi^* E^0, \pi^* E^\infty)$ is a clutching function for (E^0, E^∞). Thus we can write a finite sum $f = \sum_{-n}^{n} a_k z^k \in \Gamma \operatorname{Hom}(\pi^* E^0, \pi^* E^\infty)$ and we call this a finite Laurent series for (E^0, E^∞). If $f \in \Gamma \operatorname{Iso}(\pi^* E^0, \pi^* E^\infty)$ then it defines a clutching function and we call this a *Laurent clutching function* for (E^0, E^∞). Now we'll consider a simplest clutching function $f = z$ and $(E^0, E^\infty) = (1, L)$.

Lemma 3.1. *H^* can be represented as $(1, z, L)$.*

Proof. Since for $y \in P(L \oplus 1)_x$, H_y^* is the complex line through y, we have $H_0^* = 0 \oplus 1 \cong 1$ and $H_\infty^* = L \oplus 0 \cong L$. Thus projection $L \oplus 1 \to 1$, defines an isomorphism $f_0 : H^*|P^0 \to \pi_0^*(1)$ and $L \oplus 1 \to L$ defines an isomorphism $f_\infty : H^*|P^\infty \to \pi_\infty(L)$. Hence $f = f_\infty f_0^{-1}$ is the clutching function for H^*, and we can see that f is exactly our section z. □

And from proposition 1.4 we have $H^k \cong (1, z^{-k}, L^{-k})$ for any integer k.

Suppose $f \in \Gamma \operatorname{Hom}(\pi^* E^0, \pi^* E^\infty)$ is any section, then we define Fourier coefficients $a_k \in \Gamma \operatorname{Hom}(L^k \otimes E^0, E^\infty)$ by $a_k(x) = \frac{1}{2\pi i} \int_{S_x} f_x z_x^{-k-1} dz_x$. Where f_x shall be regarded as the matrix so that the integral is reasonable. Now let $s_n = \sum\limits_{-n}^{n} a_k z^k$ and we define the Cesaro means $f_n = \frac{1}{n} \sum\limits_{0}^{n} s_k$ as we did in Fourier extension. Moreover, we can do the same as in Fourier extension thus we have

Lemma 3.2. *Let f be any clutching function for (E^0, E^∞), f_n the sequence of Cesaro means. Then f_n converges uniformly to f and hence is a Laurent clutching function homotopic to f for all sufficiently large n.*

This lemma enables us to consider only the case when f is a Laurent clutching function.

4 Linearization

We'll handle the case when the Laurent clutching function is a polynomial first. Let $p = \sum\limits_{k=0}^{n} a_k z^k$ be a polynomial clutching function for (E^0, E^∞). consider the homomorphism $\mathcal{L}^n(p) : \pi^*(\sum\limits_{k=0}^{n} L^k \otimes E^0) \to \pi^*(E^\infty \oplus \sum\limits_{k=1}^{n} L^k \otimes E^0)$ given by the matrix

$$\mathcal{L}^n(p) = \begin{pmatrix} a_0 & a_1 & a_2 & \dots & a_n \\ -z & 1 & & & \\ & -z & 1 & & \\ & & \ddots & \ddots & \\ & & & -z & 1 \end{pmatrix}$$

Since the elementary transformation of $\mathcal{L}^n(p)$ is homotopy to $\mathcal{L}^n(p)$, so we can find a homotopy from $\mathcal{L}^n(p)$ to matrix

$$\begin{pmatrix} p & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{pmatrix}$$

by doing elementary transformations. Along with proposition 1.3, we have

Proposition 4.1.

$$(E^0, p, E^\infty) \oplus (\sum_{k=1}^{n} L^k \otimes E^0, 1, \sum_{k=1}^{n} L^k \otimes E^0) \cong (\sum_{k=0}^{n} L^k \otimes E^0, \mathcal{L}^n(p), E^\infty \oplus \sum_{k=1}^{n} L^k \otimes E^0)$$

This proposition enables us to change a polynomial p into something linear in z. And for brevity we now write $\mathcal{L}^n(E^0, p, E^\infty)$ for the bundle $(\sum_{k=0}^{n} L^k \otimes E^0, \mathcal{L}^n(p), E^\infty \oplus \sum_{k=1}^{n} L^k \otimes E^0)$.

Proposition 4.2. *Let p be a polynomial clutching function of degree $\leq n$ for (E^0, E^∞). Then*

$$\mathcal{L}^{n+1}(E^0, p, E^\infty) \cong \mathcal{L}^n(E^0, p, E^\infty) \oplus (L^{n+1} \otimes E^0, 1, L^{n+1} \otimes E^0)$$

and

$$\mathcal{L}^{n+1}(L^{-1} \otimes E^0, zp, E^\infty) \cong \mathcal{L}^n(E^0, p, E^\infty) \oplus (L^{-1} \otimes E^0, z, E^0)$$

.

Proof. The proof is similar to the proof of proposition 4.1. Since

$$\mathcal{L}^{n+1}(p) = \begin{pmatrix} \mathcal{L}^n(p) & 0 \\ 0 \ldots -z & 1 \end{pmatrix},$$

we can do elementary transformations to kill the bottom z, and then we get a homotopy from $\mathcal{L}^{n+1}(p)$ to $\mathcal{L}^n(p) \oplus 1$.

Similarly,

$$\mathcal{L}^{n+1}(zp) = \begin{pmatrix} 0 & a_0 & a_1 & \dots & a_n \\ -z & 1 & & & \\ & -z & 1 & & \\ & & \ddots & \ddots & \\ & & & -z & 1 \end{pmatrix}$$

is homotopic to

$$\begin{pmatrix} 0 & a_0 & a_1 & \dots & a_n \\ -z & 0 & & & \\ & -z & 1 & & \\ & & \ddots & \ddots & \\ & & & -z & 1 \end{pmatrix}$$

. Namely, $\mathcal{L}^{n+1}(zp)$ is homotopic to $\mathcal{L}^n(p) \oplus -z$. And the proposition follows. □

From the above propositions we can establish a formula in $K(p)$. For convenience we write $[E^0, p, E^\infty]$ for $[(E^0, p, E^\infty)]$ in $K(P)$.

Proposition 4.3. *For any clutching function p for (E^0, E^∞) we have*

$$([E^0, p, E^\infty] - [E^0, 1, E^0])([L][H] - [1]) = 0$$

Proof. On one hand, by proposition 4.1,

$$\mathcal{L}^{n+1}(L^{-1} \otimes E^0, zp, E^\infty) \cong (L^{-1} \otimes E^0, zp, E^\infty) \oplus (\sum_{k=0}^{n} L^k \otimes E^0, 1, \sum_{k=0}^{n} L^k \otimes E^0).$$

And on the other hand, by proposition 4.2

$$\begin{aligned} \mathcal{L}^{n+1}(L^{-1} \otimes E^0, zp, E^\infty) &\cong \mathcal{L}^n(E^0, p, E^\infty) \oplus (L^{-1} \otimes E^0, z, E^0) \\ &\cong (E^0, p, E^\infty) \oplus (\sum_{k=1}^{n} L^k \otimes E^0, 1, \sum_{k=1}^{n} L^k \otimes E^0) \oplus (L^{-1} \otimes E^0, z, E^0). \end{aligned}$$

So we have

$$
\begin{aligned}
(L^{-1}\otimes E^0, zp, E^\infty)\oplus(\sum_{k=0}^{n} L^k\otimes E^0, 1, \sum_{k=0}^{n} L^k\otimes E^0) \\
\cong (E^0, p, E^\infty)\oplus(\sum_{k=1}^{n} L^k\otimes E^0, 1, \sum_{k=1}^{n} L^k\otimes E^0)\oplus(L^{-1}\otimes E^0, z, E^0)
\end{aligned}
$$

Write this in $K(P)$, and with lemma 3.1 we get

$$[L^{-1}][H^{-1}][E^0,p,E^\infty]+[E^0,1,E^0]=[E^0,p,E^\infty]+[L^{-1}][H^{-1}]\oplus[E^0,1,E^0]$$

and the result follows. □

Putting $E^0=1, p=z, E^\infty=L$ in the proposition, we obtain

Corollary 4.1.

$$([H]-[1])([L][H]-[1])=0$$

which is the relation we require in theorem 2.1.

5 Linear clutching functions

Though we've change the polynomial clutching function into a linear one, but we have to throw away the redundant parts we added to linearize the polynomial.

Now we note some facts in linear algebra. Suppose T is an endomorphism of a finite dimensional complex vector space E, and let S be a circle in the complex plain that does not pass through any eigenvalue of T. Then let

$$Q=\frac{1}{2\pi i}\int_S (z-T)^{-1}dz$$

be a projection operator in E which commutes with T (consider the integral acts on the eigenvalue of T). Thus, we get an decomposition $E=E_+\oplus E_-$ where $E_+=QE$ and $E_-=(1-Q)E$ is invariant under T, so we can decompose $T=T_+\oplus T_-$. Then T_+ has all eigenvalues inside S while T_- has all eigenvalues outside S.

We want to extend this spectral decomposition to vector bundles. Suppose $p=az+b$ is a linear clutching function for (E^0,E^∞). We have the following proposition:

Proposition 5.1. *Define* $Q^i \in \operatorname{End}(E^i)(i = 0, \infty)$ *by* $Q^0_x = \frac{1}{2\pi i}\int_{S_x} p_x^{-1} dp_x$ *and* $Q^\infty_x = \frac{1}{2\pi i}\int_{S_x} dp_x p_x^{-1}$*. Then* Q^0 *and* Q^∞ *are projection operators and* $pQ^0 = Q^\infty p$*. Let* $E_+ i = Q^i E^i, E^i_- = (1 - Q^i)E^i (i = 0, \infty)$*, so that* $E^i = E^i_+ \oplus E^i_-$*. And we have decomposition* $p = p^+ \oplus p^-$*, where* p_+ *is an isomorphism outside* S*, and* p_- *is an isomorphism inside* S*.*

Proof. We only need to vertify the statements point-wise for each $x \in X$. Since $p(z)$ is an isomorphism for $|z| = 1$ we can find a real number $\alpha > 1$ so that $p(\alpha) : E^0 \to E^\infty$ is an isomorphism. We can and do suppose that $p(\alpha) = 1$. Consider the conformal transformation $\omega = \frac{1-\alpha z}{z-\alpha}$ which preserve the unit disc. By simple calculation we have $p(z) = \frac{\omega - T}{\omega + \alpha}$ where $T = a + b\alpha \in \operatorname{End} E^0$. Hence

$$\begin{aligned} Q^0 &= \frac{1}{2\pi i}\int_{|z|=1} p^{-1} dp \\ &= \frac{1}{2\pi i}\int_{|\omega|=1} [-(\omega+\alpha)^{-1} + (\omega - T)^{-1}] d\omega \\ &= \frac{1}{2\pi i}\int_{|\omega|=1} (\omega - T)^{-1} d\omega \\ &= Q^\infty. \end{aligned}$$

So all the statements follow from what we have asserted above in the discussion of linear transformation.

□

Corollary 5.1. *For linear clutching function p, we have*

$$(E^0, p, E^\infty) \cong (E^0_+, z, L \otimes E^0_+) \oplus (E^0_-, 1, E^0_-).$$

Proof. Suppose $p = p_+ \oplus p_-$ and $p_+ = a_+ z + b_+, p_- = a_- z + b_-$. We construct homotopy $p^t = p^t_+ \oplus p^t_-$ where $p^t_+ = a_+ z + t b_+$, $p^t_- = t a_z + b_- \quad 0 \le t \le 1$. It's well defined homotopy by the last statement in proposition 5.1. And with proposition 1.4 and 1.3 we have

$$\begin{aligned} (E^0, p, E^\infty) &\cong (E^0, p^0, E^\infty) \\ &\cong (E^0_+, a_+ z, E^\infty_+) \oplus (E^0_-, 1, E^0_-) \\ &\cong (E^0_+, z, L \otimes E^0_+) \oplus (E^0_-, 1, E^0_-). \end{aligned}$$

□

Now we'll apply this corollary to $\mathcal{L}^(p)$. Let p be a polynomial clutching function of degree $\leqslant n$ for (E^0, E^∞) then $\mathcal{L}^n(p)$ is a linear clutching function for (V^0, V^∞) where

$$V^0 = \sum_{k=0}^{n} L^k \otimes E^0, V^\infty = E^\infty \oplus \sum_{k=1}^{n} L^k \otimes E^0.$$

Hence it defines a decomposition$V^0 = V_+^0 \oplus V_-^0$. Write $V_+^0 = V_n(E^0, p, E^\infty)$.

Do decomposition on the two sides of the formulas in proposition 4.2, we have

$$V_{n+1}(E^0, p, E^\infty) \cong V_n(E^0, p, E^\infty),$$

$$V_{n+1}(L^{-1} \otimes E^0, zp, E^\infty) \cong V_n(E^0, p, E^\infty) \oplus (L^{-1} \otimes E^0)$$

and the second formula can be expressed as

$$V_{n+1}(E^0, zp, L \otimes E^\infty) \cong L \otimes V_n(E^0, p, E^\infty) \oplus E^0.$$

Finally, we apply corollary 5.1 to proposition 4.1 in $K(P)$ and obtain

$$[E^0, p, E^\infty] + \{\sum_{k=1}^{n} [L^k \otimes E^0]\}[1] = [V_n(E^0, p, E^\infty)][H^{-1}] + \{\sum_{k=0}^{n} [L^k \otimes E^0] - [V_n(E^0, p, E^\infty)]\}[1]$$

and hence

$$[E^0, p, E^\infty] = [V_n(E^0, p, E^\infty)]([H^{-1}] - [1]) + [E^0][1].$$

6 Proof of Theorem 2.1

Finally we come to the proof of theorem 2.1. Since we have the relation in corollary 4.1, we can define a $K(X)$-algebra homomorphism

$$\mu : K(X)[t]/(t-1)([L]t-1) \to K(P)$$

by mapping $t \to [H]$. We only need to show that μ is an isomorphism. We'll do this by constructing an inverse.

Let f be a clutching function for (E^0, E^∞) and f_n be the Cesaro means of its Fourier series. Let $p_n = z^n f_n$. Then for sufficiently large n, p is a polynomial clutching function of degree $\leq 2n$ for $(E^0, L^n \otimes E^\infty)$. We define

$$\nu_n(f) = [V_{2n}(E^0, p_n, L^n \otimes E^\infty)](t^{n-1} - t^n) + [E^0]t^n \in K(X)[t]/(t-1)([L]t-1).$$

Note that for sufficiently large n, p_{n+1} and zp_n are homotopic. And by the formulas we acquired in the last section, we have

$$\begin{aligned} V_{2n+2}(E^0, p_{n+1}, L^{n+1} \otimes E^\infty) &\cong V_{2n+2}(E^0, zp_n, L^{n+1} \otimes E^\infty) \\ &\cong V_{2n+1}(E^0, zp_n, L^{n+1} \otimes E^\infty) \\ &\cong L \otimes V_{2n}(E^0, p_n, L^n \otimes E^\infty) \oplus E^0 \end{aligned}$$

Hence

$$\nu_{n+1}(f) = \{[L][V_{2n}(E^0, p_n, L^n \otimes E^\infty)] + [E^0]\}(t^n - t^{n+1}) + [E^0]t^{n+1} = \nu_n(f).$$

Thus $\nu_n(f)$ is independent of n for sufficiently large n, and we can write it as $\nu(f)$. Since f uniquely determines the bundle E in the sense of homotopy, we can write $\nu(E) = \nu(f)$. Since $\nu(E)$ is additive for $\oplus$, it induces a homomorphism

$$\nu : K(P) \to K(X)[t]/(t-1)([L]t-1)$$

. Now we only need to check that $\mu\nu$ and $\nu\mu$ are identity of $K(P)$ and $K(X)[t]/(t-1)([L]t-1)$.

$$\begin{aligned} \mu\nu[E] &= \mu\{[V_{2n}(E^0, p_n, L^n \otimes E^\infty)](t^{n-1} - t^n) + [E^0]t^n\} \\ &= [V_{2n}(E^0, p_n, L^n \otimes E^\infty)]([H]^{n-1} - [H^n]) + [E^0][H]^n \\ &= [E^0, p_n, L^n \otimes E^\infty][H]^n \\ &= [E^0, f_n, E^\infty] \\ &= [E] \end{aligned}$$

And for $\nu\mu$ we only need to check the generators t^n.

$$\begin{aligned}\nu\mu(t^n) &= \nu[H^n] \\ &= \nu[1, z^{-1}, L^{-n}] \\ &= [V_{2n}(1,1,1)](t^{n-1} - t^n) + [1]t^n \\ &= f^n\end{aligned}$$

This completes the proof of Theorem 2.1.

Reference

[1] Atiyah M, Bott R. On the periodicity theorem for complex vector bundles[J]. Acta Mathematatica, 1964, 112(1): 229-247

[2] Hatcher A. Vector bundles and K-theory[J]. Im Internet under https://pi.math.cornell.edu/ hatcher/VBKT/VB.pdf, 2017

Bott Periodicity Theorem and Introduction to Topological K-Theory

Zheng Zhi

June 2022

1 Introduction

This article is a reading report of the paper *On the Periodicity Theorem for Complex Vector Bundles* by M.Atiyah and R.Bott, which provides a quite accessible proof of the complex case of Bott periodicity theorem. Bott's original proof[1] uses Morse theory to state that $\pi_k(U) = \pi_{k+2}(U), \pi_k(O) = \pi_{k+4}(Sp), \pi_k(Sp) = \pi_{k+4}(O)$, while this paper uses K-theory and polynomial approximation and then applies basic linear algebra.

2 Some basic definitions and lemmas

We first review some basic notions of vector bundles.

Definition 2.1. A *complex vector bundle* is a map $p : E \to X$ together with a complex vector space structure on $p^{-1}(x)$ for each $x \in X$, which is locally isomorphic to the product of the base space with a complex vector space. It is called a *line bundle* if the dimension of its fiber is 1. A *section* $\Gamma(E)$ is a continuously map $s : X \to E$ with $ps = \text{id}$.

We can extend some constructions on vector spaces to vector bundles. Given two vector bundles $E \to X, F \to X$, we can define $\hom(E, F)$ (resp. $E \oplus F, E \otimes F, E^*$) as the union of $\hom(E_x, F_x)$ (resp. $E_x \oplus F_x, E_x \otimes F_x, E_x^*$) for all $x \in X$, and define the natural topology on them using local trivializations. Besides, given a bundle $E \to X$ and a map $Y \to X$, we define the *pullback bundle* $f^*(E) \to Y$ as the pullback of $E \longrightarrow X \longleftarrow Y$.

Let $X = X_1 \cup X_2, A = X_1 \cap X_2$, and assume we have two bundles $E_1 \to X_1, E_2 \to X_2$ with an isomorphism $\varphi : E_1|A \to E_2|A$. Then we can glue them together by the *clutching construction* $E_1 \cup_f E_2$, which is the quotient of $E_1 \sqcup E_2$ by the equivalence relation induced by φ.

The direct sum makes the set of vector bundles of X into a commutative semi-group, and we denote the corresponding *Grothendieck group* as $K(X)$. In fact, the tensor product makes

[1] R.Bott, The Stable Homotopy of the Classical Groups, *Ann. of Math.*, 70(1959), 313-337

$K(X)$ into a ring, so $K(-)$ is a contravariant functor from the category of topological spaces to the category of rings. The kernel of $K(X) \to K(x_0)$ is denoted by $\widetilde{K}(X)$.

3 Statement of the periodicity theorem

Associated with a bundle $E \to X$, there is a projective (fiber) bundle $P(E) \to X$, whose fiber as $x \in X$ is the space of all lines through the origin of E_x. If we assign to each $y \in P(E_x)$ the $1 - \dim$ subspace E_x corresponding to it, we obtain a canonical line bundle over $P(E)$, which is denoted by H^*.

We can present the main theorem now.

Theorem 3.1. *Let L be a line bundle over the compact space X, H^* the canonical line bundle over $P(L \oplus 1)$. Then, as a $K(X)$-algebra, $K(P(L \oplus 1))$ is generated by $[H]$ subject to the relation*

$$([H] - [1])([L][H] - [1]) = 0,$$

i.e.

$$K(P(L \oplus 1)) \cong K(X)[H]/([H] - 1)([L][H] - [1]).$$

In particular, if X is a point, the theorem implies $K(S^2) \cong \mathbb{Z}[H]/([H]^2 - [1])$. Let L be the trivial bundle, we deduce the following important result.

Corollary 3.2. *For any compact space X,*

$$K(X) \otimes K(S^2) \cong K(X \times S^2).$$

Now we turn to the proof of the theorem.

4 Clutching functions

For any $x \in X$, $P(L \oplus 1)_x$ is the compactification of L_x by adding the point at infinity, so $P = P(L \oplus 1)$ is the compactification of L by adding the section at infinity. Choosing an arbitrary metric in L and let $S \in L$ be the unit circle bundle, so that

$$P = P^0 \cup P^\infty, S = P^0 \cap P^\infty,$$

where $P^0(P^\infty)$ is the closed disk bundle interior (exterior) to S. Denote the projection $S \to X, P^0 \to X, P^\infty \to X$ by π, π_0, π_∞ respectively.

Let E^0, E^∞ be two vector bundles over X and $f \in \mathrm{Iso}(\pi^* E^0, \pi^* E^\infty)$. Then we can form the vector bundle $\pi_0^* E^0 \cup_f \pi_\infty^* E^\infty$ over P, denoted by (E^0, f, E^∞), and we denote its equivalent class in $K(P)$ by $[E^0, f, E^\infty]$. In fact, every bundle on P is of this form.

Proposition 4.1. *Let E be any vector bundle over P, and E^0, E^∞ is induced by the* 0*-section and* ∞*-section. Then there exists $f \in \mathrm{Iso}(\pi^* E^0, \pi^* E^\infty)$ such that*

$$E \cong (E^0, f, E^\infty).$$

the isomorphism being the obvious one on the 0*-section and* ∞*-section. Moreover f is unique up to homotopy.*

Proof. Let $s_0 : X \to P^0$ be the 0-section. Then $s_0\pi_0$ is homotopic to the identity map of P_0, so we have an isomorphism $f_0 : E|P^0 \to \pi_0^* E^0$, which is unique up to homotopy. Similarly we have $f_\infty : E|P^\infty \to \pi_\infty^* E^\infty$, and we can simply take $f = f_\infty f_0^{-1}$. □

Remark 4.2. Let F be a vector bundle over X, obviously $(F, 1, F)$ is the pullback bundle induced from the projection $P \to X$, which can be written as $[F, 1, F] = [F][1]$.

The idea of the proof of the theorem is to simplify the clutching function f. First we focus on a simple function z and polynomials of z.

When L is the trivial line bundle $X \times \mathbb{C}$, S is $X \times S^1$, so points of S are represented by pairs $(x, z), x \in X, |z| = 1$. In other words, z is a function on S, so is z^{-1} and finite Laurent series $\sum_{k=-n}^{n} a_k(x) z^k$.

For example, consider the most simple case: L is the trivial line bundle over a single point. Then $P = \mathbb{C}P^1 = S^2$, and $S = S^1$. Every point of P is denoted by the equivalent class $[z_0, z_1] \in \mathbb{C}P^1$ or $z = \frac{z_0}{z_1} \in \mathbb{C} \cup \{\infty\} = S^2$. Therefore every point in the disk D_0^2 can be expressed by $[z, 1]$ with $|z| \leq 1$, and every point in the disk D_∞^2 can be expressed by $[1, z^{-1}]$ with $|z^{-1}| \leq 1$. Let E be the canonical line bundle over $\mathbb{C}P^1$. Over D_0^2 a section of E is given by $[z, 1] \mapsto (z, 1)$, and over D_∞^2 a section of E is given by $[1, z^{-1}] \mapsto (1, z^{-1})$. Then over their common boundary S, we can pass from D_∞^2 to D_0^2 by multiplying by z.

When L is not trivial, there are only some notational difficulties. Recall that we have the projection $\pi : S \to X$, so $\pi^*(L)$ is a line bundle over S. We can define a section by $S \to \pi^*(L), (x, z) \mapsto (x, z, z)$ (using local coordinates), and we denote this section by z. Similarly, z^k is a section of $\pi^*(L^k)$. If $a_k \in \Gamma \hom(L^k \otimes E^0, E^\infty)$, then

$$a_k z^k := \pi^*(a_k) \otimes z^k \in \Gamma \hom(\pi^* E^0, \pi^* E^\infty),$$

and we can define

$$f = \sum_{k=-n}^{n} a_k(x) z^k \in \Gamma \hom(\pi^* E^0, \pi^* E^\infty).$$

When $f \in \mathrm{Iso}(\pi^* E^0, \pi^* E^\infty)$, we call it a Laurent clutching function.

In fact, the bundle $(1, z, L)$ is just H^* defined in the theorem. Recall for each $y \in P(L \oplus 1)_x$, H_y^* is a subspace of $(L \oplus 1)_x$ and $H_y^* = L_x \oplus 0$ when $y = \infty$, $H_y^* = 0 \oplus 1_x$ when $y = 0$. So there are isomorphisms $f_0 : H^*|P^0 \to \pi_0^*(1)$ and $f_\infty : H^*|P^\infty \to \pi_\infty^*(L)$. Moreover, the clutching function is $f = f_\infty f_0^{-1} = z : \pi^*(1) \to \pi^*(L)$, because for $y \in S_x$, H_y^* is the subspace of $L_x \oplus 1_x$ spanned by $y \oplus 1$. Thus

$$H^* \cong (1, z, L)$$

and

$$H^k \cong (1, z^{-k}, L^{-k}).$$

The first simplification is to approximate an arbitrary clutching function f by Laurent clutching functions, which is an exercise of complex analysis. Define

$$a_k(x) = \frac{1}{2\pi i}\int_{S_x} f_x z_x^{-k-1} dz_x,\ \ s_n = \sum_{-n}^{n} a_k z^k.$$

Then define the Cesaro means

$$f_n = \frac{1}{n}\sum_{0}^{n} s_k.$$

Lemma 4.3. *f_n converges uniformly to f and is a Laurent clutching function for sufficiently large n.*

5 Linearization

In this section we describe a linearization procedure for a polynomial clutching function $p = \sum_{k=0}^{n} a_k z^k$. Consider

$$\mathcal{L}^n(p) : \pi^*(\sum_{k=0}^{n} L^k \otimes E^0) \to \pi^*(E^\infty \oplus \sum_{k=1}^{n} L^k \otimes E^0)$$

given by

$$\mathcal{L}^n(p) = \begin{pmatrix} a_0 & a_1 & . & . & . & a_n \\ -z & 1 & & & & \\ & -z & 1 & & & \\ & & . & . & & \\ & & & . & . & \\ & & & & -z & 1 \end{pmatrix}.$$

It is linear in z. By elementary matrix operations, it is equivalent to $diag(p, 1, \ldots, 1)$. More specifically, $\mathcal{L}^n(p) = (1 + N_1)(p \oplus 1)(1 + N_2)$, where N_1 is strictly upper triangular and N_2 is strictly lower triangular. Since $1 + tN$ gives a homotopy of isomorphisms for N nilpotent, we have the following proposition.

Proposition 5.1. *$\mathcal{L}^n(p)$ and $p \oplus 1$ defines isomorphic bundles on P, i.e.*

$$(E^0, p, E^\infty) \oplus (\sum_{k=1}^{n} L^k \otimes E^0, 1, \sum_{k=1}^{n} L^k \otimes E^0) \cong (\sum_{k=0}^{n} L^k \otimes E^0, \mathcal{L}^n(p), E^\infty \oplus \sum_{k=1}^{n} L^k \otimes E^0).$$

For brevity, write $\mathcal{L}^n(E^0, p, E^\infty)$ for the bundle defined in the proposition. By elementary matrix operations we have the following proposition.

Proposition 5.2. *Let p be a polynomial clutching function of degree $\leq n$ for (E^0, E^∞). Then*

$$\mathcal{L}^{n+1}(E^0, p, E^\infty) \cong \mathcal{L}^n(E^0, p, E^\infty) \oplus (L^{n+1} \otimes E^0, 1, L^{n+1} \otimes E^0),$$
$$\mathcal{L}^{n+1}(L^{-1} \otimes E^0, zp, E^\infty) \cong \mathcal{L}^n(E^0, p, E^\infty) \oplus (L^{-1} \otimes E^0, z, E^0).$$

Proof. We have

$$\mathcal{L}^{n+1}(p) = \begin{pmatrix} & & & 0 \\ & \mathcal{L}^n(p) & & \vdots \\ & & & 0 \\ 0 & \dots & -z & 1 \end{pmatrix} \simeq \begin{pmatrix} & & & 0 \\ & \mathcal{L}^n(p) & & \vdots \\ & & & 0 \\ 0 & \dots & 0 & 1 \end{pmatrix} = \mathcal{L}^n(p) \oplus 1.$$

Similarly,

$$\mathcal{L}^{n+1}(zp) = \begin{pmatrix} 0 & a_0 & a_1 & \cdot & \cdot & \cdot & a_n \\ -z & 1 & & & & & \\ & -z & 1 & & & & \\ & & \cdot & \cdot & & & \\ & & & \cdot & \cdot & & \\ & & & & \cdot & \cdot & \\ & & & & & -z & 1 \end{pmatrix} \simeq \begin{pmatrix} 0 & a_0 & a_1 & \cdot & \cdot & \cdot & a_n \\ -z & 0 & & & & & \\ & -z & 1 & & & & \\ & & \cdot & \cdot & & & \\ & & & \cdot & \cdot & & \\ & & & & \cdot & \cdot & \\ & & & & & -z & 1 \end{pmatrix} = \mathcal{L}^n(p) \oplus -z.$$

□

Using the above two propositions, we can establish a simple algebraic formula in $K(P)$.

Proposition 5.3. *For any polynomial clutching function p for (E^0, E^∞), we have*

$$([E^0, p, E^\infty] - [E^0, 1, E^0])([L][H] - [1]) = 0.$$

Proof. By the previous lemma,

$$(L^{-1} \otimes E^0, zp, E^\infty) \oplus (\sum_{k=0}^{n} L^k \otimes E^0, 1, \sum_{k=0}^{n} L^k \otimes E^0) \cong \mathcal{L}^{n+1}(L^{-1} \otimes E^0, zp, E^\infty) \cong$$

$$\mathcal{L}^n(E^0, p, E^\infty) \oplus (L^{-1} \otimes E^0, z, E^0) \cong (E^0, p, E^\infty) \oplus (\sum_{k=1}^{n} L^k \otimes E^0, 1, \sum_{k=1}^{n} L^k \otimes E^0) \oplus (L^{-1} \otimes E^0, z, E^0).$$

Passing to $K(P)$, we have

$$[L^{-1}][H^{-1}][E^0, p, E^\infty] + [E^0, 1, E^0] = [E^0, p, E^\infty] + [L^{-1}][H^{-1}][E^0, 1, E^0],$$

and the result follows. □

Let $E^0 = 1, p = z, E^\infty = L$, we obtain part of the main theorem:

$$([H] - [1])([L][H] - [1]) = 0.$$

6 Linear clutching functions

We review some basic facts about linear algebra. Suppose $T : E \to E$ is a linear transformation with no eigenvalue on S^1. Then there is a unique decomposition $E = E_+ \oplus E_-$, such that the eigenvalues of $T|E_+$ are outside S^1 and the eigenvalues of $T|E_-$ are inside S^1. Explicitly,

$$Q = \frac{1}{2\pi \mathrm{i}} \int_{S_1} (z - T)^{-1} dz$$

is a projection operator in E and commutes with T, and we can take $E = E_+ \oplus E_-, E_+ = QE, E_- = (1 - Q)E$. Now we extend this decomposition to vector bundles.

Proposition 6.1. *Given $(E^0, p, E^\infty), p = az+b$, there are decompositions $E^0 = E^0_+ \oplus E^0_-, E^\infty = E^\infty_+ \oplus E^\infty_-, p = p_+ \oplus p_-$, such that p_+ is an isomorphism outside S, and p_- is an isomorphism inside S.*

Proof. Define endomorphisms Q^0, Q^∞ of E^0, E^∞ by

$$Q^0_x = \frac{1}{2\pi \mathrm{i}} \int_{S_x} p_x^{-1} dp_x, Q^\infty_x = \frac{1}{2\pi \mathrm{i}} \int_{S_x} p_x dp_x^{-1}.$$

We want to show they are the projection operators we needed, and we only need to verify it pointwisely. In other words, we may assume X is a point, $L = \mathbb{C}$ and z is just a complex number. Since $p(z)$ is an isomorphism for $|z| = 1$, we can find $\alpha > 1$ such that $p(\alpha) : E^0 \to E^\infty$ is an isomorphism, and we identify them by this isomorphism, i.e. $p(\alpha) = 1$. Let $w = \frac{1-\alpha z}{z-\alpha}$, then $p(z) = \frac{w-T}{w+\alpha}$, where $T = a + \alpha b$ is a linear transformation of E^0. Then

$$Q^0 = \frac{1}{2\pi \mathrm{i}} \int_{|z|=1} p^{-1} dp = \frac{1}{2\pi \mathrm{i}} \int_{|w|=1} (-(w+\alpha)^{-1} dw + (w-T)^{-1} dw) = \frac{1}{2\pi \mathrm{i}} \int_{|w|=1} (w-T)^{-1} dw.$$

Similarly,

$$Q^\infty = \frac{1}{2\pi \mathrm{i}} \int_{|w|=1} (w-T)^{-1} dw.$$

Then the conclusion follows from the discussion at the beginning of the section. □

Proposition 6.2. *Using the notation in the previous lemma, we have*

$$(E^0, p, E^\infty) \cong (E^0_+, z, L \otimes E^0_+) \oplus (E^0_-, 1, E^0_-).$$

Proof. Write $p_+ = a_+z+b_+, p_- = a_-z+b_-$, and define $p^t_+ = a_+z+tb_+, p^t_- = ta_-z+b_-, 0 \le t \le 1$. By the previous lemma, p^t_+ and p^t_- are isomorphisms. Therefore p is homotopic to $a_+z \oplus b_-$, i.e.

$$(E^0, p, E^\infty) \cong (E^0_+, a_+z, E^\infty_+) \oplus (E^0_-, b_-, E^\infty_-) \cong (E^0_+, z, L \otimes E^0_+) \oplus (E^0_-, 1, E^0_-),$$

since $a_+ : L \otimes E^0_+ \to E^\infty_+, b_- : E^0_- \to E^\infty_-$ are isomorphisms. □

Let p be a polynomial clutching function of degree $\leq n$ for (E^0, E^∞). Then $\mathcal{L}^n(p)$ is a linear clutching function for (V^0, V^∞), where

$$V^0 = \sum_{k=0}^{n} L^k \otimes E^0, V^\infty = E^\infty \oplus \sum_{k=1}^{n} L^k \otimes E^0.$$

Then we have a decomposition $V^0 = V^0_+ \oplus V^0_-$. To express the dependence of V^0_+ on p and n, we write

$$V^0_+ = V_n(E^0, p, E^\infty).$$

Using the results from proposition 5.2, we have:

Proposition 6.3.

$$V_{n+1}(E^0, p, E^\infty) \cong V_n(E^0, p, E^\infty),$$
$$V_{n+1}(E^0, zp, L \otimes E^\infty) \cong L \otimes V_n(E^0, p, E^\infty) \oplus E^0.$$

Besides, by proposition 5.1,

$$[E^0, p, E^\infty] + \sum_{k=1}^{n} [L^k \otimes E^0][1] = [V_n(E^0, p, E^\infty)][H^{-1}] + (\sum_{k=0}^{n} [L^k \otimes E^0] - [V_n(E^0, p, E^\infty)])[1].$$

Then we can establish another formula in $K(P)$, which shows that elements in $K(P)$ with polynomial clutching functions can be generated by $[H^{-1}]$ with coefficients in $K(X)$.

Proposition 6.4. *With the above notation,*

$$[E^0, p, E^\infty] = [V_n(E^0, p, E^\infty)]([H] - [1]) + [E^0][1].$$

7 Proof of the main theorem

Finally we prove the main theorem. We already have an homomorphism

$$\mu : K(X)[t]/(t-1)([L]t-1) \to K(P), t \mapsto H.$$

Now we construct an inverse v.

Let f be any clutching function and f_n be its Cesaro means. Define $p_n = z^n f_n$. Then p_n is a polynomial clutching function. Define

$$\nu_n(f) = [V_{2n}(E^0, p_n, L^n \otimes E^\infty)](t^{n-1} - t^n) + [E^0]t^n \in K(X)[t]/(t-1)([L]t-1).$$

For sufficiently large n, the linear segment joining p_{n+1} and zp_n provides a homotopy of polynomial clutching function of degree $\leq 2(n+1)$. Then by proposition 6.3,

$$V_{2n+2}(E^0, p_{n+1}, L^{n+1} \otimes E^\infty) \cong V_{2n+2}(E^0, zp_n, L^{n+1} \otimes E^\infty) \cong V_{2n+1}(E^0, p_{n+1}, L^{n+1} \otimes E^\infty)$$

$$\cong L \otimes V_{2n}(E^0, p_n, L^n \otimes E^\infty) \oplus E^0.$$

Therefore

$$\nu_{n+1}(f) = ([L][V_{2n}(E^0, p_n, L^n \otimes E^\infty)] + [E^0])(t^n - t^{n+1}) + [E^0]t^{n+1} = \nu_n(f)$$

because $(t-1)([L]t-1) = 0$.

Therefore, $\nu_n(f)$ is independent of n when n is large, so we can write it as $\nu_(f)$. Besides, if f and g are sufficiently close, then for sufficiently large n, f_n and g_n are also homotopic, so $\nu(f) = \nu_n(f) = \nu_n(g) = \nu(g)$. So $\nu(f)$ only depends on the homotopy class of f, and we can define $\nu(E) = \nu(f)$, where f is a clutching function of E. By definition, ν is additive and only depend on the isomorphism class of E.

Now we check that ν is the inverse of μ. On the one hand, by proposition 6.4,

$$\mu\nu[E] = [V_{2n}(E^0, p_n, L^n \otimes E^\infty)]([H]^{n-1} - [H]^n) + [E^0][H]^n = [E^0, p_n, L^n \otimes E^\infty][H]^n = [E^0, f_n, E^\infty]$$

$$= [E^0, f, E^\infty] = [E].$$

On the other hand,

$$\nu\mu(t^n) = \nu[1, z^{-n}, L^{-n}] = [V_{2n}(1,1,1)](t^{n-1} - t^n) + [1]t^n = t^n.$$

Then we prove the theorem.

8 More about topological K-theory

K-theory forms a generalized cohomology theory as follows. Given a pair of compact space (X, A), the natural map sequence $A \to X \to X/A$ induces an exact sequence $\widetilde{K}(X/A) \to \widetilde{K}(X) \to \widetilde{K}(A)$. Moreover, we have the coexact Puppe sequence

$$A \to X \to X/A \to SA \to SX \to \dots,$$

which gives rise to a long exact sequence

$$\dots \to \widetilde{K}(SX) \to \widetilde{K}(SA) \to \widetilde{K}(X/A) \to \widetilde{K}(X) \to \widetilde{K}(A).$$

In particular, consider the pair $(X \times Y, X \vee Y)$. Since we have a map

$$\widetilde{K}(X \vee Y) = \widetilde{K}(X) \oplus \widetilde{K}(Y) \to \widetilde{K}(X \times Y), (a, b) \mapsto p_1^*(a) + p_2^*(b),$$

where p_1, p_2 are projections, thus the sequence splits, i.e.

$$\widetilde{K}(X \times Y) = \widetilde{K}(X \wedge Y) \oplus \widetilde{K}(X) \oplus \widetilde{K}(Y).$$

We can also consider the external product

$$\widetilde{K}(X) \otimes \widetilde{K}(Y) \to \widetilde{K}(X \times Y), (a,b) \mapsto a * b := p_1^*(a)p_2^*(b).$$

In fact, $a * b$ lies in $\widetilde{K}(X \wedge Y)$, which defines the external product $\widetilde{K}(X) \otimes \widetilde{K}(Y) \to \widetilde{K}(X \wedge Y)$. This is because $p_1^*(a)$ is zero in $K(Y)$ and $p_2^*(b)$ is zero in $K(X)$. It can also be deduced from the unreduced external product $K(X) \otimes K(Y) \to K(X \times Y)$ since

$$K(X) \otimes K(Y) = (\widetilde{K}(X) \otimes \widetilde{K}(Y)) \oplus \widetilde{K}(X) \oplus \widetilde{K}(Y) \oplus \mathbb{Z}$$

and

$$K(X \times Y) = \widetilde{K}(X \wedge Y) \oplus \widetilde{K}(X) \oplus \widetilde{K}(Y) \oplus \mathbb{Z}.$$

Proposition 8.1. *For any compact space X,*

$$\widetilde{K}(X) \cong \widetilde{K}(S^2X).$$

In particular,

$$\widetilde{K}(S^{2n+1}) = 0, \widetilde{K}(S^{2n}) \cong \mathbb{Z}.$$

Proof. Notice that S^nX is homotopy equivalent to its quotient space $\Sigma^nX = S^n \wedge X$, and $\widetilde{K}(S^2) = \mathbb{Z}$ is generated by $[H]-1$, thus by the periodicity theorem we have the isomorphism

$$\widetilde{K}(X) \cong \widetilde{K}(S^2) \otimes \widetilde{K}(X) \cong \widetilde{K}(S^2 \wedge X) \cong \widetilde{K}(S^2X).$$

□

Define $\widetilde{K}^{-n}(X) = \widetilde{K}(S^nX)$ and $\widetilde{K}^{-n}(X,A) = \widetilde{K}(S^n(X/A))$. By the periodicity theorem, we have $\widetilde{K}^{2i}(X) = \widetilde{K}(X)$ and $\widetilde{K}^{2i+1}(X) = \widetilde{K}(SX)$. Then there is a six-term exact sequence:

$$\begin{tikzcd}
\widetilde{K}^0(X,A) \arrow[r] & \widetilde{K}^0(X) \arrow[r] & \widetilde{K}^0(A) \arrow[d] \\
\widetilde{K}^1(A) \arrow[u] & \widetilde{K}^1(X) \arrow[l] & \widetilde{K}^1(X,A) \arrow[l]
\end{tikzcd}$$

The unreduced version is similar by defining $K^n(X) = \widetilde{K}^n(X_+)$, where X_+ is X with a disjointed basepoint.

Define $\widetilde{K}^*(X) = \widetilde{K}^0(X) \oplus \widetilde{K}^1(X)$. It has a ring structure defined by $\widetilde{K}^*(X) \otimes \widetilde{K}^*(X) \to \widetilde{K}^*(X \wedge X) \to \widetilde{K}^*(X)$, where the second map is induced by the diagonal map $X \to X \wedge X, x \mapsto (x,x)$. We have the following two proposition.

Proposition 8.2. $\alpha\beta = (-1)^{ij}\beta\alpha$ *for* $\alpha \in \widetilde{K}^i(X)$ *and* $\beta \in \widetilde{K}^j(X)$.

Proposition 8.3. *There is an exact sequence of $\widetilde{K}^*(X)$-modules:*

$$\begin{tikzcd}
\widetilde{K}^*(X, A) \arrow[rr] & & \widetilde{K}^*(X) \arrow[dl] \\
& \widetilde{K}^*(A) \arrow[ul] &
\end{tikzcd}$$

The above two propositions also hold for the unreduced version $K^*(X)$.

We list some further theorems and applications in topological K-theory.

Theorem 8.4 (Leray-Hirsch theorem). *Let $p : E \to B$ be a fiber bundle with E and B compact and with fiber F such that $K^*(F)$ is free. Suppose that there exist classes $c_1, \ldots, c_k \in K^*(E)$ that restrict to a basis for $K^*(F)$ in each fiber F. If either*

(a) B is a finite cell complex, or

(b) F is a finite cell complex having all cells of even dimension,

then $K^*(E)$, as a module over $K^*(B)$, is free with basis $\{c_1, \ldots, c_k\}$.

Theorem 8.5 (the splitting principle). *Given a vector bundle E over compact space X, there is a compact space $F(E)$ and a map $p : F(E) \to X$ such that $p^* : K^*(X) \to K^*(F(E))$ is injective and $p^*(E)$ splits as a sum of line bundles.*

Theorem 8.6. *The following statements are true only for $n = 1, 2, 4, 8$:*

(a) $\mathbb{R}^n$ is a division algebra.

(b) S^{n-1} is parallelizable, i.e. there exist $n-1$ tangent vector fields to S^{n-1} which are linearly independent at each point, i.e. the tangent bundle to S^{n-1} is trivial.

Remark 8.7. The Frobenius theorem states that there are only three finite-dimensional associative division algebras over $\mathbb{R}$: $\mathbb{R}, \mathbb{C}, \mathbb{H}$. For the non-associative case, Hopf proved that the dimension must be a power of 2. He also proved that every finite-dimensional real commutative division algebra is either 1- or 2-dimensional, and no direct algebraic proof is known. The result in the previous theorem was independently proved by Michel Kervaire and John Milnor in 1958.

9 Reference

[1] M.Atiyah and R.Bott, On the Periodicity Theorem for Complex Vector Bundles, *Acta Math.*, 112 (1964), 229-247

[2] A.Hatcher, Vector Bundles and K-Theory

Volume Conjecture for Reshetikhin-Turaev Invariants

Xu Xiaoyu
Peking University

June 2022

Contents

1 Introduction

Volume conjecture is a profound topic in the research of 3-manifolds, which is related to a combination of low dimensional topology, hyperbolic geometry, quantum field theory etc.

This paper will provide a brief insight in one of the motivations of the Chen-Yang's Reshetikhin-Turaev type Volume Conjecture (Conjecture 5.2) [2]. The motivation is based on the important way of constructing and studying 3-manifolds, called Dehn surgery, which can be realized in both a topological point of view and a geometrical point of view.

The key relation between these two point of views is Proposition 4.5, which states that the completion of an incomplete hyperbolic manifold under the conditions given by Hyperbolic Glueing Equation (4.1) and (4.2) is topologically equivalent to a Dehn filling (surgery).

Furthermore, Dehn surgery along links in S^3 studied in the topological method can induce quantum topological invariants, while studied in the geometrical method can induce hyperbolic geometrical invariants. The connection between these two invariants of 3-manifolds is given by Volume conjecture.

In this paper, we will review a calculative proof due to T. Ohtsuki [6] that deals with special cases of Chen-Yang's volume conjecture. The part of Ohtsuki's result which is discussed in this paper is stated in Theorem 6.1, and our proof will be based on complex analysis methods, including Poisson Summation and Saddle Point Method. The most valuable result of this proof is that the critical function coincides with the Hyperbolic Glueing Equation (4.3), demonstrated in Proposition 6.2.

The author would like to thank Professor Yang Tian for his report at Hua Loo-Keng Key Laboratory of Mathematics in 2021.

2 Topological Backgrounds

2.1 Dehn Surgery

We first consider the topological method to construct and study 3-manifolds, a typical method of obtaining new 3-manifolds from given 3-manifolds is Dehn surgery, relative works are mainly due to Lickorish, Wallace, and Kirby. Reference [8] provides a collection of these results.

Definition 2.1 (Dehn Surgery)**.** *Suppose M is a 3-manifold, and N is an embedded solid torus. Let N' be a copy of N, a Dehn Surgery on M along N is a new manifold M' obtained through:*

$$M' = (M \setminus \mathring{N}) \cup_h N'$$

where h is a orientation-preserving homeomorphism: $h : \partial N' \to \partial N$

Proposition 2.1. *Under the assumption of Definition 2.1, suppose m is a meridian of N', and l is one of the longitudes of N', then:*

(1) The image of m and l in $H_1(\partial N)$ determines h up to isotopy

(2) The image of m in $H_1(\partial N)$ determines M' up to homeomorphism

Proof:

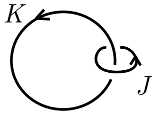
(1) Suppose ∂N and $\partial N'$ are both given by universal covering $\mathbb{R}^2/\mathbb{Z}^2$.
By Map Lifting Theorem, we can lift h to a homeomorphism $\tilde{h} : \mathbb{R}^2 \to \mathbb{R}^2$.

Obviously, $\tilde{h}$ is istopic to a standard map with the image of $\mathbb{Z}^2$ fixed, through translation on each coordinate.

(2) Suppose h and h' are two orientation preserving homeomorphisms between $\partial N'$ and ∂N.
Take a basis $\bar{m}$, $\bar{l}$ of $H_1(\partial N)$, and suppose $h_* \begin{pmatrix} m \\ l \end{pmatrix} = \begin{pmatrix} p & q \\ r & s \end{pmatrix} \begin{pmatrix} \bar{m} \\ \bar{l} \end{pmatrix}$, $h'_* \begin{pmatrix} m \\ l \end{pmatrix} = \begin{pmatrix} p & q \\ r' & s' \end{pmatrix} \begin{pmatrix} \bar{m} \\ \bar{l} \end{pmatrix}$.
$p, q, r, s, r', s' \in \mathbb{Z}$.
Since h, h' are orientation preserving homeomorphisms, we have $\begin{vmatrix} p & q \\ r & s \end{vmatrix} = \begin{vmatrix} p & q \\ r' & s' \end{vmatrix} = 1$.
Then p and q are coprime, hence $\exists k \in \mathbb{Z}$, such that $\begin{pmatrix} p & q \\ r' & s' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ k & 1 \end{pmatrix} \begin{pmatrix} p & q \\ r & s \end{pmatrix}$.
Notice that there exists a self-homeomorphism $\varphi : N' \to N'$ such that $\varphi_*(l) = km + l$.
Then $(h \circ \varphi)_* \begin{pmatrix} m \\ l \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ k & 1 \end{pmatrix} \begin{pmatrix} p & q \\ r & s \end{pmatrix} \begin{pmatrix} \bar{m} \\ \bar{l} \end{pmatrix} = \begin{pmatrix} p & q \\ r' & s' \end{pmatrix} \begin{pmatrix} \bar{m} \\ \bar{l} \end{pmatrix}$.
By (1), $h \circ \varphi$ and h' are isotopic.
Hence $(M \setminus \mathring{N}) \cup_h N' \cong (M \setminus \mathring{N}) \cup_{h \circ \varphi} N' \cong (M \setminus \mathring{N}) \cup_{h'} N'$.

□

Especially, it is natural to consider Dehn Surgery in S^3, where the solid tori are given by regular neighbourhoods of components of a link.

Definition 2.2 (Linking Number). *Suppose J, K are two embedded (oriented) S^1 in S^3*
By Alexander Duality, $H_1(S^3 \setminus K) \cong \mathbb{Z}$ $lk(J,K) :=$ image of $[J]$ in $H_1(S^3 \setminus K) \cong \mathbb{Z}$
The linking number may vary between a positive or negative sign, which can be determined by the orientation given: 

, $lk(J,K) = 1$

Proposition 2.2. *By Tubular Neighbourhood Theorem, every embedded S^1 in S^3 has a tubular neighbourhood N.*
There is a unqiue meridian and a unique preferred longitude in $H_1(\partial N)$
The meridian $\bar{m}$ is the meridian of the solid torus N, and the longitude $\bar{l}$ is the only one of the longitudes with linking number 0 with the core.

Proof: The conclusion is obvious since $\bar{l}$ can be replaced by $k\bar{m} + \bar{l}$, and $\bar{m}$ is a generator of $H_1(S^3 \setminus K)$. □

Definition 2.3 (Dehn Surgery Coefficients). *Suppose $L = K_1 \sqcup ... \sqcup K_n$ are embedded S^1's in S^3.*
N_i are their disjoint tubular neighbourhoods.
$\bar{m}_i$, $\bar{l}_i$ are their meridians and preferred longitudes, oriented so that they have $+1$ linking number.
$M' = S^3 \setminus (\mathring{N}_1 \cup ... \cup \mathring{N}_n) \cup_{h_1,...,h_n} (N'_1 \cup ... \cup N'_n)$ is a Dehn surgery.
Suppose m_i are the meridians of N'_i, and $h_{i*}([m_i]) = p_i[\bar{m}_i] + q_i[\bar{l}_i]$.
(p_i, q_i) are co-prime integer pairs.
Then M' is uniquely determined up to homeomorphism by the link L and the surgery coefficients $\frac{p_1}{q_1}, \ldots, \frac{p_n}{q_n}$.
The surgery coefficients $\frac{p}{q}$ is a rational number including ∞.

Remark 2.1. *The surgery coefficients are independent from the orientation of the link, because* $\frac{-p}{-q} = \frac{p}{q}$.

We have a more simple explanation for Dehn surgery with integral coefficients.

Proposition 2.3. *If M is obtained through Dehn surgery along a link in S^3 with* ***integral coefficients****, then M can be realized through the boundary of the manifold constructed through pasting $2-$handles to D^4*

Proof: Suppose $X = D^4 \cup_{f_1,\ldots,f_n} (D^2 \times D^2 \sqcup \ldots \sqcup D^2 \times D^2)$
$f_i : D^2 \times S^1 \to \partial D^4$, such that:

- $f : D^2 \times S^1 \to N_i$ is a homeomorphism
- $f\,|_{S^1 \times S^1} : S^1 \times S^1 \to \partial N$
  maps $S^1 \times \{0\}$ to $\bar{m}_i$ and maps $\{0\} \times S^1$ to $p\bar{m}_i + \bar{l}_i$

Then we take $M = \partial X$. □

Dehn surgery along links in S^3 is a complete explanation of closed, orientable 3-manifolds, which is because of the theorem provided by Lickorish and Wallace which states that all closed, orientable 3-manifolds can be obtained through this method.

Theorem 2.1 (Lickorish[1962], Wallace[1960])**.** *Every closed, orientable, connected 3-manifold can by obtained through Dehn Surgery with integral coefficients on a link in S^3.*
Moreover, the surgery presentation can be constructed such that all components of the link are unknotted, and all surgery coefficients are ± 1.

Proof: Suppose M is a closed, orientable 3-manifold. We consider a Heggard Splitting of M: $M = H \cup_f H'$, where H and H' are handle-bodies of genus g.
Let $\Sigma = \partial H$ and $\Sigma' = \partial H'$, $f : \Sigma \to \Sigma'$ is an orientation-preserving homeomorphism.
Then there exists a standard homeomorphism $f_0 : \Sigma \to \Sigma'$, such that $H \cup_{f_0} H' = S^3$.
According to Lickorish Twist Lemma, $h = {f_0}^{-1} \circ f : \Sigma \to \Sigma$, up to isotopy, is a finite composition of twist homeomorphisms along the $3g-1$ curves demonstrated in Figure 1.

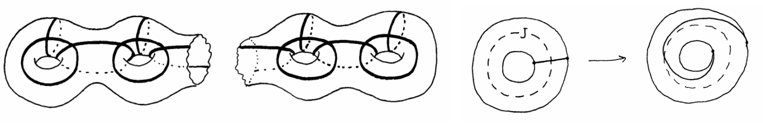


Figure 1: Twist Homeomorphisms along $3g-1$ Curves [8]

$f = f_0 \circ h$ can be extended to a neighbourhood $\Sigma \times [0,1]$, and obviously, as is shown in Figure 2, each twist operation can be realized up to a homeomorphism of the 3-manifold through a Dehn surgery of coefficient ± 1 along a tubular neighbourhood of a paralled copy of one of the $3g-1$ curves.

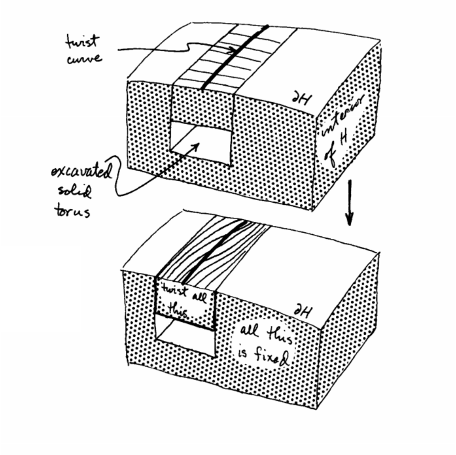


Figure 2: Twist Operation Realized through Dehn Surgery [8]

Moreover, the $3g-1$ curves either bounds a disk in H or H' hence is an unknot in S^3. □

2.2 Kirby Moves

Theorem 2.1 shows that every closed orientable 3-manifold has a surgery representation, it is natural to consider when two surgery representation will provide two homeomorphic manifolds. Relative results are due to Kirby.

Theorem 2.2 (Kirby[1976])(Dehn Surgery Version)**.** *Two surgery presentations of two links yield homeomorphic 3-manifolds if and only if one of the surgery presentations can be transferred into the other through finite steps of Kirby Moves:*

(1) Add or delete an unknotted component with coefficient ∞

(2) Find an unknotted component L_i *and perform* t *right-hand twists to its complement* ($t \in \mathbb{Z}$)*, changing the coefficients by* $r_i{}' = \frac{1}{t+\frac{1}{r_i}}$*, and* $r_j' = r_j + t(lk(L_i, L_j))^2$ ($j \neq i$)

Proof: Sufficiency: The first Kirby move is simply pasting back a solid torus the same way as it is removed, hence yields a homeomorphic manifold.
The second Kirby move is to re-picture a cylinder area of the manifold, shown in Figure 3, and tracking the image of meridians directly yields the change of the coefficients.

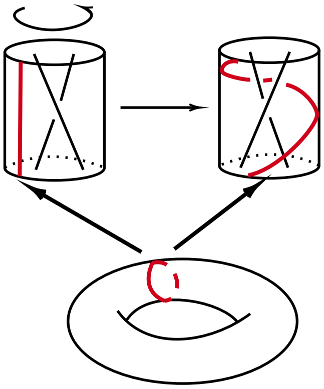


Figure 3: The Second Kirby Move

Necessity: The necessity part is due to Cerf's Theory based on characterizing singularities of the Morse function. Readers can check [4] for more details. □

Theorem 2.3 (Kirby[1978])(Handle Version)**.** *Two surgery presentations with integral coefficients of two links yield homeomorphic 3-manifolds if and only if one of the surgery presentations can be transferred into the other through finite steps of Kirby Handle Moves:*

(1) Add or delete an unknotted component with coefficient ± 1, where this unknotted component lies in a smoothly embedded B^3 disjoint from other components.

(2) Choose two different components L_i, L_j, and let L'_j be the curve (up to isotopy) on the boundary of the tubular neighbourhood of L_j to which the meredian of the replacing solid torus is attached. Replace L_i with a band sum of L_i and L'_j, and change the coefficient to $r'_i = r_i + r_j \pm 2lk(L_i, L_j)$, $\pm$ depends on the choice of the orientation of the band sum.

Proof: Sufficiency: The first Kirby handle move is a connected sum with the lens space $L(1, \pm 1) \cong S^3$, which yields a homeomorphic manifold.
The second Kirby handle move is based on the model of pasting handles onto D^4. In fact, the second Kirby move is the move on the boundary of a handle slid, demonstrated in Figure 4.

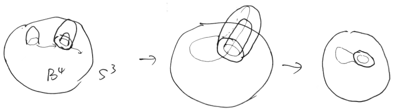


Figure 4: Handle Slid

Necessity: The necessity part is again due to Cerf's Theory. Readers can check [4] for more details.
□

Remark 2.2. *In this section, we have successfully established a one-to-one correspondence:*

$$\boxed{\frac{\textit{Links in } S^3 + \textit{Surgery Coefficients} \in \mathbb{Q}\cup\{\infty\}\,(\mathbb{Z})}{\textit{Kirby (handle) moves}}} \xleftrightarrow{1:1} \boxed{\textit{Closed, orientable 3-manifolds}}$$

3 Knot (Link) Invariants

Through Section 2, we have completely transferred the study of closed, oriented 3-manifolds to the study of links (with coefficients, or equivalently a framing) in S^3. It is natural to follow similar ideas in algebraic topology, which is to find invariants for links (or framed links) in S^3.
Classical topological invariants of links in S^3 include complement space, knot group, Seifert surface, Alexander invariants, etc.
However, the link invariants which come up to have close relationship with hyperbolic volume of manifolds are quantum topological invariants, which originate mainly from quantum topological group theory and physics.

Example 3.1. *Jones Polynomial (Jones[1985]) and Colored Jones Polynomial*

$$\{\textit{Knots (or Oriented Links)}\} \to \mathbb{Z}\left[A^{\pm}\right]$$

Example 3.2. *Witten's Quantum Invariants (Witten[1989])*
A part of its mathematical requirements include:

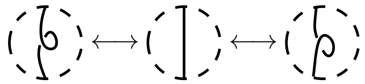
$$F:\ \{\textit{Closed, Oriented 3-Manifolds}\} \to \mathbb{C}\ \textit{(or } \mathbb{C}^{\mathbb{N}})$$
$$\begin{array}{lc} (1) & F(M_1\#M_2) = F(M_1)\cdot F(M_2) \\ (2) & F(-M) = \overline{F(M)} \\ (3) & F(S^3) = 1 \end{array}$$

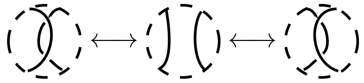
This paper will focus on Reshetikhin-Turaev invariants, which is one of the mathematical realizations of Witten's quantum invariants.
A skein theoretical approach of constructing Reshetikhin-Turaev invariants is based on a cabling method of Kauffman brackets, which is a generalization of the approach in constructing the colored Jones polynomials.
The definition of Reshetikhin-Turaev invariants is based on the approach due to Masbaum and his co-authors [1].

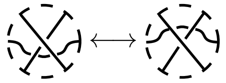
3.1 Reidemeister Move and Kauffman Bracket

Theorem 3.1 (Reidemeister[1927])**.** *Two links in S^3 are isotopy equivalent, if and only if a regular projection of one link can be obtained from a regular projection of the other link through finite steps of Reidemeister move I, II, III:*

Reidemeister move I:



Reidemeister move II: 

Reidemeister move III: 

Definition 3.1 (Framed Link)**.** *A framed link $\tilde{L}$ is an embedding $S^1 \times [0,1] \sqcup ... \sqcup S^1 \times [0,1] \hookrightarrow S^3$. The two boundary components of each annulus, $S^1 \times \{0\}$ and $S^1 \times \{1\}$ are denoted by K_i and K_i' respectively. K_i' is called the paralleled copy of K_i.*
The writhe of K_i is defined by $lk\,(K_i, K_i')$.

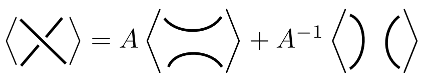
Note 3.1. *Isotopy between framed links denotes the isotopy between the embedding maps.*

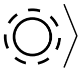
Definition 3.2 (Blackboard Framing)**.** *A regular projection of a framed link is called standard or blackboard framing, if each K_i' is paralleled to K_i in the projection image.*

Definition 3.3. *Kauffman Bracket is a map:*

$$\langle \cdot \rangle : \textit{Link Diagrams} \to \mathbb{Z}\left[A^{\pm}\right]$$

defined through skein theoretical relations:

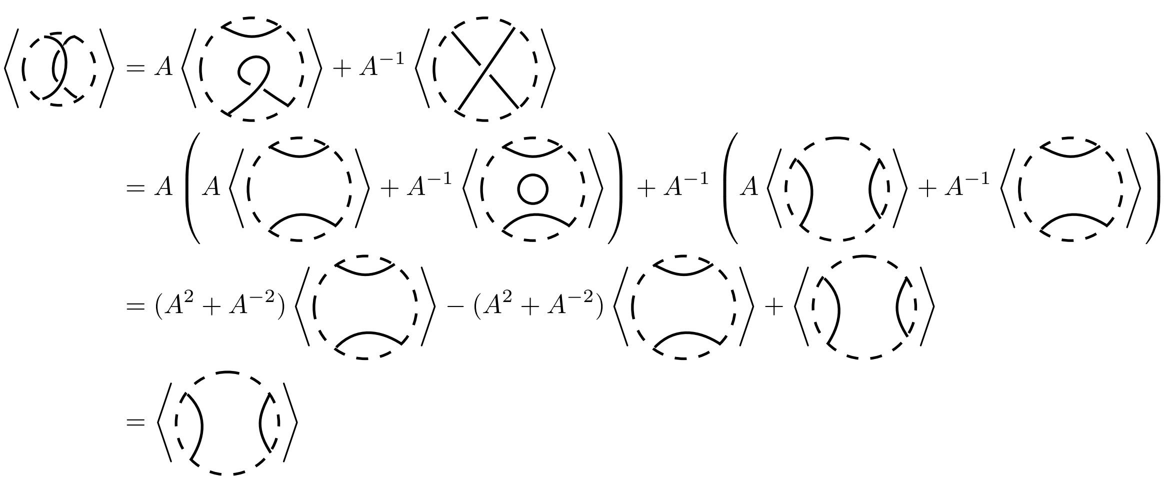
(i) 

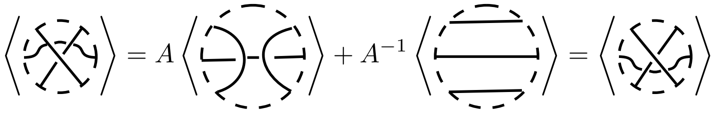
(ii) $\Big\langle L \sqcup$  $\Big\rangle = (-A^2 - A^{-2})\,\langle L \rangle$

(iii) $\langle \bigcirc \rangle = -A^2 - A^{-2}$

Theorem 3.2 (Kauffman[1987])**.** *Kauffman Bracket is invariant under Reidemeister move II and III.*

Proof:

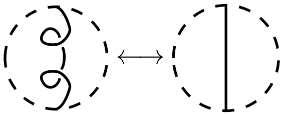




□

Remark 3.1. *Unfortunately, Kauffman bracket is not invariant under Reidemeister move I, so it is not an invariant of links under isotopy equivalence.*
However, it is noticable that Reidemeister move I is different from Reidemeister move II and III when we consider framed links.
Reidemeister move II and III yield isotopic blackboard framing of the link diagram, but Reidemeister move I change the linking number between the component and its blackboard paralleled copy by ±1.
It is natural to suggest that Reidemeister move I can be completely reflected by a choice of framing.
The following Proposition gives an explanation for the relation between the special Reidemeister move I and framed links.

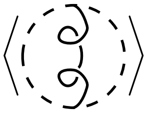
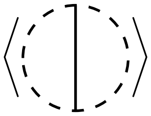
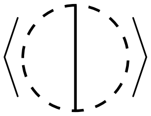
Proposition 3.1. *We introduce a Move I' (paired Reidemeister move I) defined by:*

Move I': 

Then: two framed links in S^3 are isotopy equivalent, if and only if a standard (blackboard) regular projection of one framed link can be obtained from a standard (blackboard) regular projection of the other framed link through finite steps of Move I',and Reidemeister move II, III.

Proof: Sufficiency is obvious, we will now show the necessity.
By isotopy equivalence of the two framed links, the $S^1 \times \{0\}$ components are also isotopy equivalent.
According to Theorem 3.1, one diagram can be obtained from the other through finite steps of Reidemeister move I, II, III.
For each component, pick a local part that is not influenced by any Reidemeister move I operations. Then for each Reidemeister move I, we replace it by a Move I', retain the half part consistent with the Reidemeister move I, contract the reversed part to be small enough, and push the reversed part along the component towards the local part we have chosen that is not influenced by any Reidemeister move I operations.
It is obvious that the pushing process only include finite steps of Reidemeister move II and III. Thus through Move I, Reidemeister move II and III, we obtained a link diagram with only a slight difference with the target: for every component, there is a local part including some extra twists. Notice that the blackboard framing of the two link diagrams are isotopic, thus the linking number of the link components and their paralleled copies are consistent. A twist in one or the other direction will change the linking number between the component and its blackboard paralleled copy by ±1. Hence the number twists in the two directions are equal for each local part of the components. These twist can be combined into pairs one by one from inside to the outside, and can be cancelled out by finite steps of Move I'. □

Proposition 3.2. *Kauffman Bracket is invariant under Move I'.*

Proof:  $= (-A^3)(-A^{-3})$  $=$  □

Corollary 3.1. *Kauffman bracket is an invariant for framed links.*

Proof: This is a direct corollary from Proposition 3.1, Theorem 3.2 and Theorem 3.2. □

Definition 3.4 (Framing Presentation of Dehn Surgery)**.** *A Dehn surgery along a link in S^3 with integral coefficients can be denoted by a framed link, in which the paralleled copy of each component coincides with the image of the pasted meridian in the boundary of its tubular neighbourhood.*

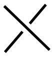
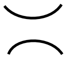
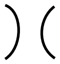
3.2 Coloring of Framed Links

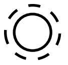
In order to obtain an invariant of Dehn surgery, it is equivalent to consider invariants of framed links which are stable under Kirby moves. A natural method of obtaining more information is to add patterns on the annuli defined by the framed link, and derive invariants for the patterned framed link. This method is called coloring.

The natural method is to cable some new knots onto the annuli, hence we would like to first study the knot diagrams in an annulus (or equivalently, knots in $S^1 \times [0,1] \times [0,1]$)

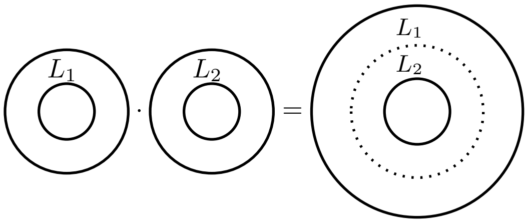
Definition 3.5. *By tensor product with the coefficient ring,*
$\mathbb{Z}[A^{\pm}] \otimes \{$*isotopy classes of link diagrams in an annulus*$\}$*, quotient the following relations, the isotopy classes of link diagrams in an annulus form a commutative $\mathbb{Z}[A^{\pm}]$-algebra, denoted by $\mathscr{B}$. The relations are:*

- $\mathbb{Z}[A^{\pm}]$*-module relations:*

(1)  $= A$  $+ A^{-1}$ 

(2) $L \sqcup$  $= (-A^2 - A^{-2})L$

- *Multiplication relations:*



Remark 3.2. *$\mathscr{B}$ is commutative, because switching L_1 and L_2 in the annulus requires only finite steps of Reidemeister move II and III (switching line segments and striding over crossings).*
And Reidemeister move II and III will not change the element quotient the $\mathbb{Z}[A^{\pm}]$-module relations (1) and (2), for completely the same reason of Theorem 3.2.

Remark 3.3. *The empty diagram $\varnothing$ is the identity element of $\mathscr{B}$.*

Proposition 3.3.

$$\mathscr{B} \cong \mathbb{Z}\left[A^{\pm}\right][z]$$

Where z defines an essential curve in the annulus.

Proof: We first define $\Phi : \mathscr{B} \to \mathbb{Z}\left[A^{\pm}\right][z]$.
By $\mathbb{Z}\left[A^{\pm}\right]$-module relations, any link diagram in the annulus can be disassembled as follows:
First, cancel out all the crossings according to relation (1).
Then we get a link diagram with no crossings, every component must be a simple S^1 in the projected diagram.
In the annlus, every simple S^1 is either essential or contractible.
We can remove all contractible simple S^1's one by one from inside to the outside according to relation (2).
All remainings are a $\mathbb{Z}\left[A^{\pm}\right]$-linear combination of link diagrams consisting of finitely many simple essential curves.
Since every essential curve divides the annulus into two path components, all these essential curves can be arranged up to isotopy one by one from inside to the outside.
Let z denote a simple essential curve, then the aforementioned diagram is denoted by z^n according to the multiplication relation, as is shown in Figure 5.

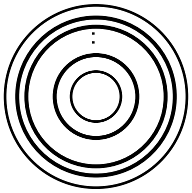
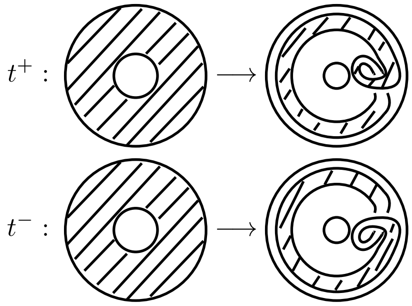


Figure 5: The image of z^n

Thus, we have defined the image of a link diagram in the annulus to be a $\mathbb{Z}\left[A^{\pm}\right]$-linear combination of z^n's.
Conversely, we define $\Psi : \mathbb{Z}\left[A^{\pm}\right][z] \to \mathscr{B}$, by directly assigning each z^n to n paralled copies of essential curves.
Clearly, Φ and Ψ are $\mathbb{Z}\left[A^{\pm}\right]$-algebra homomorphisms, and that $\Phi \circ \Psi = id$, $\Psi \circ \Phi = id$, hence $\mathscr{B} \cong \mathbb{Z}\left[A^{\pm}\right][z]$ □

The Kauffman bracket, or the $\mathbb{Z}\left[A^{\pm}\right]$-module relations are designed to be stable under Reidemeister move II and III. The remaining question is how to deal with Reidemeister move I. In the framed link model (annulus link diagram model), Reidemeister move I is depicted by twist operations defined as follows.

Definition 3.6. $t^{\pm}$ *defines a* $\mathbb{Z}[A^{\pm}]$*-linear map* $\mathscr{B} \to \mathscr{B}$*, called twist operation:*



In order to define an invariant under Reidemeister move I, we would like to find elements that are fixed under the twist operations. But unfortunately, no elements except for $\mathbb{Z}[A^{\pm}]$ (elements without z) are fixed.
For a substitute, we would like to find the eigenvectors for $t^{\pm}$.

Definition 3.7 (Second Type Chebyshev Polynomials)**.**

$$e_0 = 1 \quad e_1 = z$$
$$e_{n+1} = ze_n - e_{n-1}$$

defines $e_n \in \mathbb{Z}[z]$*, which is called the second type Chebyshev polynomials.*

Proposition 3.4. *The second type Chebyshev polynomials* $e_n \in \mathbb{Z}[z]$ *are eigenvectors for twist operation.*

$$t^+(e_n) = (-1)^n A^{n^2+2n} \cdot e_n$$
$$t^-(e_n) = (-1)^n A^{-n^2-2n} \cdot e_n$$

Remark 3.4. *Proposition 3.4 is not essential in our following proofs, readers can check [1] for its proof.*

On the other hand, if we want to construct invariants for Dehn surgery, we have to consider invariants for framed links under Kirby moves.
The first Kirby move is relatively easy. The main problem lies in the second Kirby move, which includes a common twist around a link and an unknotted component.
Hence, we would like to calculate the Kauffman bracket of link (with coefficients in $\mathbb{Z}[A^{\pm}]$) derived from cabling elements of $\mathscr{B}$ onto a framed link in S^3, so that the common twist operation is relatively simple.
In fact, we will construct a coloring such that the Kauffman bracket of the twisted link is simply derived from multiplying an item to the original Kauffman bracket.
Unfortunately, this cannot be realized unless A is an element of finite order. Thus, we will take A to be a root of unity. Under this assumption, such requirements can be realized through the Kirby coloring.

Definition 3.8 (Kirby Coloring)**.** *The Kirby coloring is defined as:*

$$\omega_r = \sum_{n=0}^{r-2} (-1)^n \frac{A^{2n+2} - A^{-2n-2}}{A^2 - A^{-2}} e_n \in \mathscr{B}$$

ω_r is also called the quantum Chebyshev polynomial.

The skein theoretical theorem of Masbaum [1] guarantees a simple formula for the common twist operation.

Theorem 3.3 (Masbaum[1991])**.** *If $A = e^{\frac{\pi i}{r}}$, and $r \geq 3$ is an odd integer, then ω_r is an orthogonnal vector satisfying:*

- *$\langle t^{\pm}(O_{\omega_r}), t^{\pm}(K_b)\rangle = \langle t^{\pm}(O_{\omega_r})\rangle \langle K_b\rangle$, $\forall b \in \mathscr{B}$, for O and K being the two components of a standard 0-framing of a Hopf link.*
- *$\langle t^{\pm}(O_{\omega_r})\rangle \neq 0$, where O denotes the standard 0-framing of an unknot.*

the Kauffman bracket calculates the link diagram consisting of the annulus link diagrams cabled on a blackboard framing of the links.

Note 3.2. *The adoption of blackboard framing is unnecessary, since it can be replaced by calculating the Kauffman bracket of framed links cabled in the thickened original framed link.*
Yet, for simplicity of link diagrams, we will not mess the diagram up with all the paralleled copies of link components, and instead always refer to a blackboard framing, and mark the coloring (cabling) of each framed component with a polynomial beside.

Remark 3.5. *The proof for Theorem 3.3 is rather complicated, readers may check [1] for details.*

3.3 Reshetikhin-Turaev Invariants

Based on these results, Masbaum and his co-authors [1] were able to define an invariant of framed links under Kirby moves. which is consistant with the Reshetikhin-Turaev invariants defined by Reshetikhin and Turaev in reference [7].

Definition 3.9 (Linking Matrix)**.** *Suppose $L = K_1(K_1') \sqcup ... \sqcup K_n(K_n')$ is an oriented framed link in S^3*
The linking matrix $LK(L)$ is defined to be:

$$\begin{aligned} &LK(L) \in \mathbb{Z}^{n\times n} \\ &LK_{i,j} = lk(K_i, K_j) \\ &LK_{i,i} = lk(K_i, K_i') \end{aligned}$$

Remark 3.6. *Obviously, $LK(L)$ is a symmetric matrix with entries in $\mathbb{Z}$.*

Definition 3.10 (Reshetikhin-Turaev Invariants[1991])**.** *Suppose M is a closed, oriented 3-manifold obtained through Dehn surgery along a link L in S^3 with integral surgery coefficients.*
The link L is oriented according to the orientation of M.
L can be framed so that the paralleled copy of each component correspond to the image of the pasted meridian in the boundary of its tubular neighbourhood.
Suppose the link diagram of L is a blackboard framing, without loss of generality.
Let $r \geq 3$ be an odd number, then:

$$\begin{aligned} &RT_r(M) = \langle \omega_r, ..., \omega_r\rangle_L \left\langle t^{+}(O_{\omega_r})\right\rangle^{-b_+} \left\langle t^{-}(O_{\omega_r})\right\rangle^{-b_-} \\ &b_+,\ b_- \text{ are the number of positive and negative eigenvalues of } LK(L) \end{aligned}$$

$RT_r(M)$ valued at $A = e^{\frac{\pi i}{r}}$, defines an invariant in $\mathbb{C}$ for M.

Proof: Notice that according to Theorem 3.3, $\langle t^+(O_{\omega_r})\rangle$ and $\langle t^-(O_{\omega_r})\rangle$ are non-zero, thus the division is well defined.
We will now show that $RT_r(M)$ is an invariant.
First, according to Corollary 3.1, the kauffman bracket is independent from the choice of the standard (blackboard) regular projection diagram.
Second, we will show $RT_r(M)$ is consistent under the first and second Kirby handle moves.
By the first Kirby handle move, the change in the linking matrix will be $\begin{pmatrix} LK(L) & 0 \\ 0 & \pm 1 \end{pmatrix} \longleftrightarrow \big(LK(L)\big)$, which is to say the number of positive (or negative) eigenvalues is altered by one. And the change in the Kauffman bracket will be $\big\langle L_{\omega_r,\ldots,\omega_r} \sqcup O^{\pm}_{\omega_r}\big\rangle = \langle L_{\omega_r,\ldots,\omega_r}\rangle \cdot \big\langle O^{\pm}_{\omega_r}\big\rangle$, where $\sqcup$ denotes a seperation between two link parts by a smooth S^2. The extra $\big\langle O^{\pm}_{\omega_r}\big\rangle$ is cancelled out with the change of the number of positive (or negative) eigenvalues.
The second Kirby handle move can be decomposed into four steps:

(a) Introduce an unknotted component seperated by a smooth S^2.

(b) Twist one component with the added component in one direction.

(c) Twist another component with the added component in the other direction.

(d) Remove the unknot.

According to the result we have proved for the first Kirby handle move, step (a) and (d) altogether will not change the number of positive and negative eigenvalues, and will keep $RT_r(M)$ consistent, hence also keeps the Kauffman bracket consistent.
The change in the linking matrix through the entire second Kirby handle move given by a band sum and $r_i' = r_i + r_j \pm 2lk(L_i, L_j)$ can be expressed in matrix by $\begin{pmatrix} 1 & \pm 1 & \\ & 1 & \\ & & \ddots \end{pmatrix} LK(L) \begin{pmatrix} 1 & & \\ \pm 1 & 1 & \\ & & \ddots \end{pmatrix} \longleftrightarrow LK(L)$. This will not change the number of positive and negative eigenvalues.
Step (b) and (c) will not alter the kauffman bracket when valued at $A = e^{\frac{\pi i}{r}}$, which is directly obtained from Theorem 3.3. □

Proposition 3.5. *The Reshetikhin-Turaev Invariants coincide with the mathematical requirements for Witten's quantum invariants in Example 3.2.*

Proof:

(1) $RT_r(M_1 \# M_2) = RT_r(M_1) \cdot RT_r(M_2)$
Suppose $M_{1,2}$ are obtained from Dehn surgery with integral coefficients along links $L_{1,2}$ in S^3. Then $M_1 \# M_2$ is obtained from Dehn surgery with the same coefficients along links $L_1 \sqcup L_2$. The result is direct by noticing $\langle L_{1\omega_r} \sqcup L_{2\omega_r}\rangle = \langle L_{1\omega_r}\rangle \cdot \langle L_{2\omega_r}\rangle$ and
$LK(L_1 \sqcup L_2) = \begin{pmatrix} LK(L_1) & \\ & LK(L_2) \end{pmatrix}$.

(2) $RT_r(-M) = \overline{RT_r(M)}$
To change the orientation of M, it suffices to change the orientation of S^3 when performing the Dehn surgery, which is equivalent with substituting A with A^{-1} in the skein theoretical

calculation.
Notice that A is the only complex valued item in the calculation and when $A = e^{\frac{\pi i}{r}}$, $A^{-1} = \overline{A}$

(3) $RT_r(S^3) = 1$
Notice that S^3 is obtained through Dehn surgery along an empty link.

□

Remark 3.7. *Through section 2 and 3, we transferred closed, oriented 3-manifolds into a representation of framed links, and use the results inspired by quantum topology, and finally obtained a series of complex valued invariants for framed links under Kirby moves, or equivalently, for closed, oriented 3-manifolds.*

4 Hyperbolic Geometry and Dehn Filling

Besides topological and quantum topological methods in studying 3-manifolds, there is another method which endues and studies geometrical structures on 3-manifolds. Relative results are mainly due to Thurston (reference [9]).

4.1 Motivation

Theorem 4.1 (Jaco–Shalen–Johannson Decomposition[1979]). *Suppose that 3-manifold M is orientable, compact, irreducible, and $\partial-$irreducible with boundary consisting of tori. Take a disjoint collection $\{T_1, ..., T_n\}$ of canonical tori in M, such that no two of them are parallel and the collection is maximal.*
Then M is fibered by the tori into $M = (\sqcup H_i) \sqcup (\sqcup E_j)$, with H_i hyperbolic and E_j Seifert-fibered, and the decomposition is unique.

Piecewise linear, irreducible knot or link complements in S^3 are typical examples of 3-manifolds satisfying the conditions of JSJ-Decomposition. So it is natural to study the geometrical structures, especially hyperbolic geometrical structures, of kont or link complements.

Corollary 4.1 (Thurston[1979]). *A piecewise linear knot complement is hyperbolic if and only if it contains no essential spheres or essential tori.*

Proof: Since there are no essential spheres, the connected sum decomposition of the knot complement is trivial, hence the knot complement is irreducible.
Then we adapt JSJ-decomposition to the knot complement, there are no essential tori, hence the JSJ-decomposition is also trivial.
Hence the knot complement has only one piece of geometrical structure, obviously it cannot be seifert fibered, hence is hyperbolic. □

Definition 4.1 (Dehn Filling). *Suppose M is a 3-manifold with a torus boundary T.*
Then $M \cup_f D^2 \times S^1$, $f : S^1 \times S^1 \to T$ homeomorphism, implies a Dehn filling.

Example 4.1. *Riley [1979] calculated that the fundamental group of the figure-8 knot complement can be embedded into $PSL(2,\mathbb{C})$, which indicates that the figure-8 knot complement has a hyperbolic structure.*
An important type of examples of knot complement yielding hyperbolic structure on its complement $S^3 \setminus K$ is whitehead doubled knots with twist number $\neq\ 0$

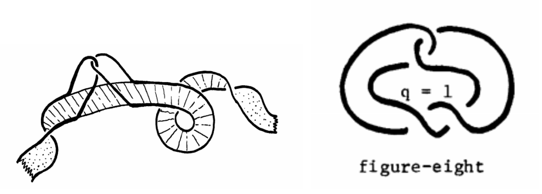


Figure 6: Whitehead Doubled Knots

Figure-8 knot is an easiest example of whitehead doubled knots.

4.2 Hyperbolic Manifolds

Now we turn to the process of constructing hyperbolic maifolds.
There are two equivalent definitions for a hyperbolic $n-$manifold

Definition 4.2. *A metric space (M, d) is a hyperbolic $n-$manifold, if and only if $\forall x \in M$, $\exists U$ an open neighbourhood of x, such that U is isometric with an open disc $D \subset \mathbb{H}^n$*

Definition 4.3. *A topological space M is a hyperbolic $n-$manifold, if there exists an atlas $\{(U_i, \phi_i)\}_i$, such that:*

(1) $\phi_i : U_i \to D_i$ is a homeomorphism, where $D_i \subset \mathbb{H}^n$ is an open disc

(2) $\bigcup\limits_i U_i = M$

(3) Transition maps $\phi_j \phi_i^{-1}$ are isometries

Proposition 4.1. *Definition 4.2 and 4.3 are equivalent.*

Proof: If M satisfies Definition 4.2, we can directly choose the atlas $\{(U_x, \phi_x) | x \in M, \phi_x : U_x \to D_x\}$, where $\phi_x : U_x \to D_x \subset \mathbb{H}^n$ is the isometry map.
For each $U_i \cap U_j \neq \varnothing$, $\phi_j {\phi_i}^{-1}$ is an isometry map restricted on $\phi_i(U_i \cap U_j) \subset \mathbb{H}^n$.
The isometry map on an open sub set of $\mathbb{H}^n$ can be extended to the whole space $\mathbb{H}^n$ according to the classification of isometry maps in $\mathbb{H}^n$.
If M satisfies Definition 4.3, for any piecewise smooth path $\alpha \subset M$, there is a finite covering of the atlas: $\bigcup\limits_{1 \leq i \leq n} U_i \supset \alpha$. Devide α into segments $\alpha = \bigsqcup\limits_{1 \leq j \leq m} \alpha_j$, such that $\alpha_j \subset U_{i(j)}$.
Define $\|\alpha\| = \sum\limits_{1 \leq j \leq m} Len_{\mathbb{H}^n}(\phi_{i(j)}(\alpha_j))$.
Since the transition maps are isometry, $\|\alpha\|$ is independent with the choice of U_i and division α_j.
For $\forall x, y \in M$, we define $d(x,y) = \underset{\alpha :\, x \to y}{inf} \|\alpha\|$.

This is a metric on M which induces small isometric discs near any point. □

Corollary 4.2. *The universal covering of a hyperbolic manifold is a hyperbolic manifold.*

Corollary 4.3. *If M is a hyperbolic n-manifold, then lifting to its universal covering $\tilde{M}$ enduces a devloping map and a holonomy with covariant isometric action:*

$$\begin{tikzcd}
Dev: & \tilde{M} \arrow[r] & \mathbb{H}^n \\
Hol: & \pi_1(M) \arrow[r] \arrow[u, bend right] & Isom(\mathbb{H}^n) \arrow[u, bend right]
\end{tikzcd}$$

Especially, Dev is a local isometry.

Proof: Corollary 4.2 and 4.3 are direct results from Definition 4.3. □

Definition 4.4. *A hyperbolic manifold is called complete, if it is a complete metric space according to Definition 4.2.*

Remark 4.1. *A compact hyperbolic manifold is obviously complete.*

Corollary 4.4. *If M is a complete hyperbolic n-manifold. Then the developing map Dev is an isometry, and the universal covering $\tilde{M}$ is a subset of $\mathbb{H}^n$. As a consequence, $M \cong Dev(\tilde{M}) \Big/ Hol(\pi_1(M))$.*

Proof: Local isometry map of a connected complete metric space is an isometry.
Thus, Dev is an isometry, and as a consequence an injection and an open map.
Hence $\tilde{M} \cong Dev(\tilde{M})$ is a subset of $\mathbb{H}^n$. □

4.3 Hyperbolic Ideal Tetrahedrons

We will follow a topological method of constructing hyperbolic 3-manifolds, which is to combine simplexes through pasting maps.
The simplexes of hyperbolic 3-manifolds are ideal hyperbolic tetrahedrons.
Ideal hyperbolic triangles and ideal hyperbolic tetrahedrons (with geodesic boundaries) have special rigid properties.

Proposition 4.2. *Hyperbolic ideal triangle is rigid, i.e. orientation preserving isometry between any two hyperbolic ideal triangles exists and is unique up to a rotation of vertices.*

Proof: Take the standard ideal triangle to be the one with ideal vertices $0, 1, \infty \in \hat{\mathbb{C}}$ in the upper-half space model.
On one hand, for any ideal triangle with ideal vertices $a, b, c \in \hat{\mathbb{C}}$, there is an orientation-preserving isometry that maps a, b, c to $0, 1, \infty$ respectively.
First, take an elliptic element (a rotation) f_1 to map c to ∞.
Then, take a parabolic element (a translation on $\mathbb{C}$) f_2 to map $f_1(a)$ to 0.
Finally, take a loxodromic element (a similarity transformation on $\mathbb{C}$) f_3 which fixes 0 and ∞ to map $f_2(f_1(b))$ to 1.

$f_3 f_2 f_1$ is an orientation-preserving isometry that maps the ideal triangle to the standard one.
On the other hand, it suffices to show that if f is an orientation-preserving isometry that preserves the standard ideal triangle (with ideal vertices fixed), f can only be the identity map.
This is because f preserves ∞, so it must be a similarity transformation on $\mathbb{C}$. f preserves 0 and 1, so the similarity coefficient must be 1, hence an isometry on $\mathbb{C}$. The orientation-preserving isomentry on the plane has two fixed points, and hence must be identity. □

Definition 4.5. *Hyperbolic ideal tetrahedron is characterised by one complex parameter, called the dihedral angle parameter, defined as follows:*
Choose one geodesic edge of the hyperbolic ideal tetrahedron.
Then, map one of the adjacent facet of this edge to the standard ideal triangle, such that: the chosen edge is map to the geodesic connecting 0 *and* ∞*, and the fourth ideal vertex is mapped to* $w \in \mathbb{C}$ *with* $\mathrm{Im} w > 0$*.*

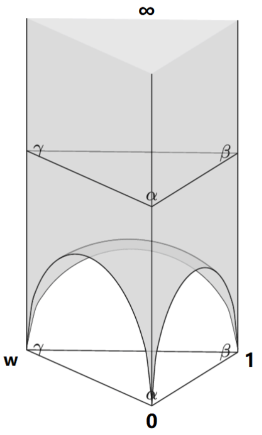


Figure 7: Hyperbolic Ideal Tetrahedron with Dihedral Angle Parameter w

Remark 4.2. *The dihedral angle parameter is well-defined, i.e. independent with which of the two ideal vertexes of the chosen edge is mapped to* ∞*, and which one is mapped to* 0*.*
This is because an orientation-preserving isometry induced thourgh Poincaré extension of $z \mapsto \frac{w}{z}$ *alters the two cenarios.*

Remark 4.3. *One can also refer the dihedral angle parameter to the complex cross ratio of the four ideal vertices in* $\hat{\mathbb{C}}$*.*

Proposition 4.3. *(1) The dihedral angle parameters of the three edges around one ideal vertex are* w, $\frac{1}{1-w}$, $\frac{w-1}{w}$ *counterclockwise.*

(2) The dihedral angle parameters of opposite edges are equal.

Proof:

(1) Calculate directly the quotient of the vectors connecting 0, 1 and w.

(2) This is directly derived from (1), for example: $\frac{\frac{1}{1-w}-1}{\frac{1}{1-w}} = w$.

□

4.4 Hyperbolic Gluing Equations

Theorem 4.2. *Pasting hyperbolic ideal tetrahedrons through hyperbolic isometry between facets yields a hyperbolic 3-manifold if and only if the Thurston's Equation holds:*

$$\prod z_i = 1 \tag{4.1}$$

which denotes that the product of all dihedral angle parameters around each edge equals 1

Proof: The neighbourhood of each interior point of the ideal tetrahedrons is obviously hyperbolic manifold. The only issue relates with points on the boundary or edges of the ideal tetrahedrons.
The conditions are necessary and sufficient, if the dihedral angles around each edge sum up to 2π, and each point on the facet are precisely maped back to themselves through a series of pasting isometry maps.
Placing the tetrahedrons in the upper-half space model, and without loss of generality, suppose the considered edge is the geodesic connecting 0 and ∞.
Then the conditions hold if and only if the isometry mappings yield the vertical-viewed image shown in Figure 8.

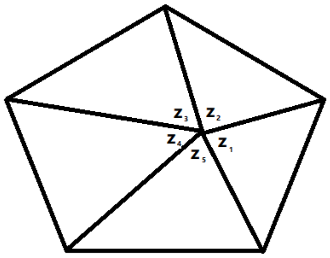


Figure 8: Thurston's Equation — Vertical View

This is equivalent the product of all dihedral angle parameters around each edge equals 1. □

Though Thurston's equation provides an equivalent condition that ensures the pasting process yields a manifold, it is not guaranteed whether the manifold is complete.
We provide a 2-dimensional example to demonstrate this subtle difference:

Example 4.2. *Figure 9 shows complete and incomplete hyperbolic 2-manifolds.*

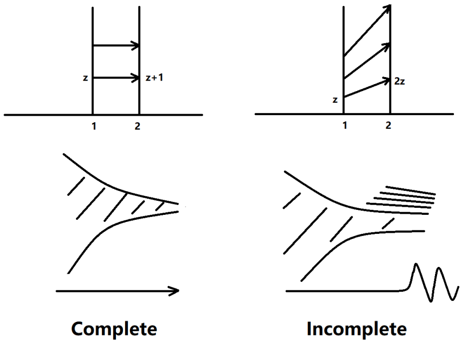


Figure 9: Complete and Incomplete Hyperbolic 2-manifolds

Notice that the compact neighbourhoods of any interior point is obviously complete, so incompleteness originates from the metric structure near the ideal vertices.
We first provide a topological view of the manifold near the ideal vertices.

Proposition 4.4 (Structure of the Link of Ideal Points)**.**

(1) If an orientable hyperbolic 3-manifold is obtained by gluing finitely many hyperbolic tetrahedrons, then we can lift the manifold to its universal covering and visualize as its image under the developing map in $\mathbb{H}^3$, and move the ideal vertex v to ∞

(2) The link of the ideal point $Lk(v)$ is topologically an orientable closed surface. It is realized by gluing a fundamental domain (which is a polygon in $\mathbb{C}$) by side pairings equal to the hyperbolic isometry maps between tetrahedron facets given by similarity transformations on $\mathbb{C}$

(3) The hyperbolic structure is complete, if and only if for any ideal point, the side pairings are isometric transformations on $\mathbb{C}$

Proof: On one hand, when the side pairings are isometric transformations on $\mathbb{C}$, then all levels of link are flat, and can pave the full plane.
Thus, the developing map is an injection near v, and all geodesics in the neighbourhood of v goes straight towards ∞, hence is complete.
On the other hand, when there exists a side pairing f with similarity coefficient $\neq 1$. We may assume the similarity coefficient of f is less than one, otherwise we consider f^{-1}. Take a point a which is the projection of a point in the fundamental domain on $\mathbb{C}$. Then $f^n(a)$ is a Cauchy sequence, it converges to a point $z_0 \in \mathbb{C}$.
Then the geodesic connecting z_0 and ∞ is a set of limit points, i.e. a set of points outside the devloping image of the universal covering, but with Cauchy sequences in the manifold converging to it.
Hence, the manifold is incomplete. □

It is natural to consider the completion of an incomplete hyperbolic manifold to obtain a compact space, but it is not always guaranteed that the completion yields a manifold (it is possible to obtain a one-point compactification which there is no hyperbolic structure near the infinity point,

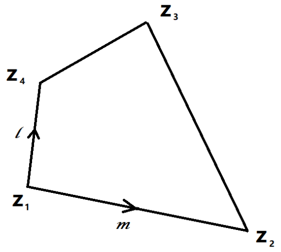


Figure 10: The Fundamental Domain

and it is also possible to obtain an orbifold).
However, the following conditions ensures that the completion yields a hyperbolic manifold, and furthermore provides results closely related to Dehn filling or Dehn surgery introduced in Section 2.

Proposition 4.5. *Let M be the hyperbolic manifold obtained through pasting ideal tetrahedrons by isometry maps between facets.*
If the link of an ideal vertex v of the hyperbolic manifold M is a torus with incomplete hyperbolic structure, and v is moved to ∞ in the developing image. Pick a meridian m and a longitude l of the torus and a fundamental domain with sides m and l (possibly broken segments instead of a straight segment).
Let $H'(m)$ and $H'(l)$ be the complex similarity coefficient (its absolute value being the real scaling coefficient and the argument corresponding to a rotation) of the side pairing translation along m and l.
If the Dehn Filling Equation holds:

$$H'(m)^p H'(l)^q = 1 \quad (p,q) \textit{ is a comprime integer pair} \tag{4.2}$$

Then the completion of M near v yields a complete hyperbolic 3-manifold, the image of the regular neighbourhood under covering is topologically a solid torus, and the completion near this vertex is topologically equivalent to a Dehn filling with coefficients (p,q):

$$M \setminus N(v) \cup_f D^2 \times S^1$$

Proof: The projection of the fundamental domain on $\mathbb{C}$ is a quadrilateral with vertices z_1, z_2, z_3, z_4 as shown in Figure 10.
Since there are non-$\mathbb{C}$-isometries in the side pairings, the quadrilateral is not a parallelogram.
Then the holonomy $H(l)$ is a linear function, mapping $z_1 \mapsto z_4$, $z_2 \mapsto z_3$. Hence the fixed point of $H(l)$ is $z_0 = \frac{z_1 z_3 - z_2 z_4}{z_1 + z_3 - z_2 - z_4}$.
Similarly, the holonomy $H(m)$ is a linear function, mapping $z_1 \mapsto z_2$, $z_4 \mapsto z_3$. Hence the fixed point of $H(m)$ is also $z_0 = \frac{z_1 z_3 - z_2 z_4}{z_1 + z_3 - z_2 - z_4}$.
Without loss of generality, we may assume $z_0 = 0$.
Consider the group $G = \langle H(m), H(l) \rangle < Isom^+(\mathbb{H}^3)$.
Then $H(m)$ and $H(l)$ are commutative, and satisfy a relation $H(m)^p H(l)^q = id$.

By Bezout's Theorem, $\exists u, v \in \mathbb{Z}$, such that $up + vq = 1$.
Let $\phi = H(m)^{-v} H(l)^{-u}$
Then ϕ is a loxodromic element and a generator of G, $\phi^{-q} = H(m)$, $\phi^{p} = H(l)$.
G is a loxodromic type elementary group, and $z_0 = 0$ is an isolated limit point.
The image of the fundamental domain under translation through side pairings are depicted in Figure 11.

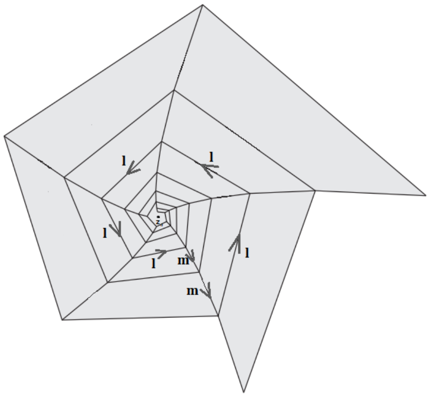


Figure 11: Side Parings of the Fundamental Domain and the Meridian Loop

It is clear that in $\mathbb{H}^3$, the geodesic connecting $z_0 = 0$ and ∞ is the set of limit points. And the neighbourhood of the ideal vertex v is given by $\mathbb{H}^3 \Big/ G$.

Completion of the manifold is equivalent to the completion of its universal covering. Thus, the completion process is to paste the image of a regular neighbourhood of the $0 - -\infty$ geodesic to the original manifold.

The regular neighbourhood of the $0 - -\infty$ geodesic is a standard cone N in $\mathbb{H}^3$, shown in Figure 12. The G-action on N is a loxodromic action, hence $N \Big/ G \cong D^2 \times S^1$.

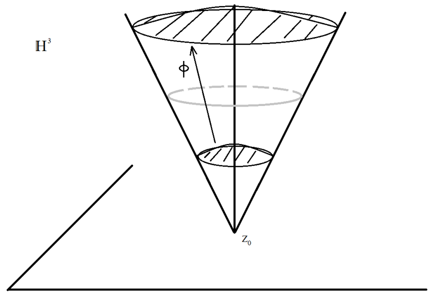
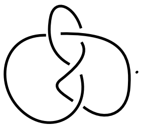


Figure 12: The Regular Neighbourhood of the Limit Set

The completion process is to paste the $D^2 \times S^1$ to the neighbourhood of the ideal vertex, which is obviously a Dehn filling.
The Dehn filling coefficients are shown in Figure 11 by tracking the image of the meridian $S^1 \times 0$, the meridian is consistent up to homotopy with $m^p l^q$. It remains to clarify that the loop $m^p l^q$ is a simple closed curve with winding number ±1 respect to $z_0 = 0$.
First, the winding number cannot be zero, since the paving pattern of the fundamental domain in a simply connect domain without $z_0 = 0$ is flat, $m^p l^q$ yields a closed curve if and only if $p = q = 0$, a contradiction.
Second, the pictured curve $m^p l^q$ must be a simple curve, for if it intersects itself, the intersection must occur on the vertices. Hence, we obtained two integers $p', q' \in \mathbb{Z}$, $|p'| \leq |p|$, $|q'| \leq |q|$, the equivalence cannot hold simultaneously, such that $H(m)^{p'} H(l)^{q'} = id$.
But $H(m) = \phi^{-q}$, $H(l) = \phi^p$, and ϕ being a loxodromic element has infinite order.
Thus $-qp' + pq' = 0$.
Because p and q are coprime, there exists $k \in \mathbb{Z}$ such that $p' = kp$, $q' = kq$.
By the inequalities, $k = 0$, a contradiction.
Hence, $m^p l^q$ is a simple closed curve, and a simple closed curve in a plane has winding number at most 1 respect to any point. □

Remark 4.4. *The equations (4.1) and (4.2) together are named the Hyperbolic Gluing Equations.*

Example 4.3. *If a knot (link) complement has a hyperbolic structure given by pasting ideal tetrahedrons, then the neighbourhood of the ideal vertices are the tubular neighbourhoods of the knot (link) components, the link of the ideal vertices are tori.*
Choose meridians and preferred longitudes for the link components, then if the Hyperbolic Gluing Equations (4.1) and (4.2) hold, then a completion of this hyperbolic structure yields a Dehn filling to the knot (link) complement.
This is equivalent to a Dehn surgery along the knot (link) in S^3.
Hence, the completion of incomplete hyperbolic structure of knot (link) complements under conditions of Hyperbolic Gluing Equations is a geometrical realization of Dehn Surgery introduced in Section 2.

Example 4.4. *We will calculate a specific example of the figure-8 knot:* .

The figure-8 knot complement $S^3 \setminus K$ is heomeomorphic to the manifold obtained by gluing two ideal tetrahedrons as shown in Figure 13.

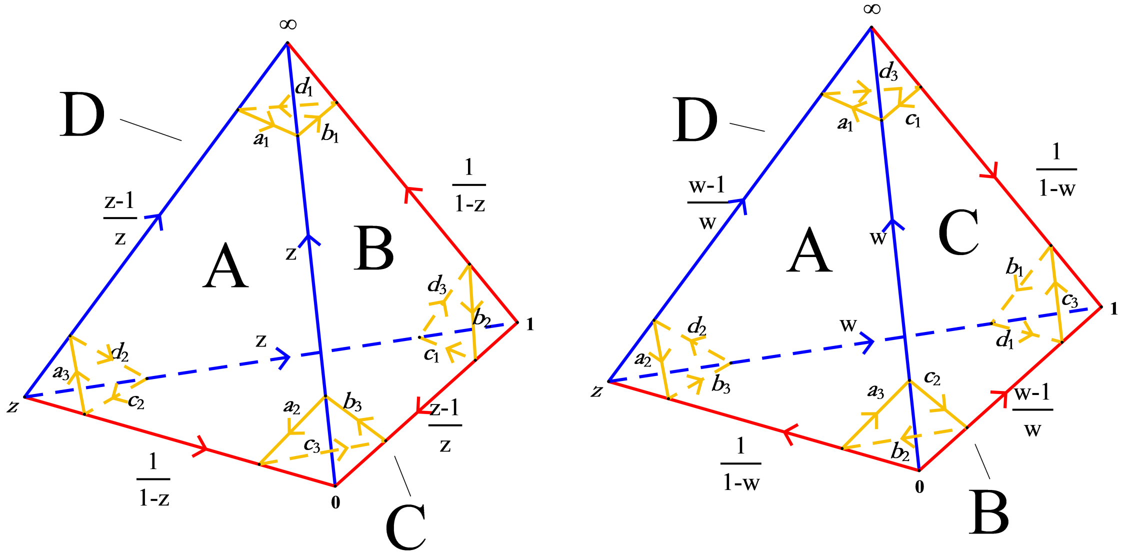


Figure 13: $S^3 \setminus K$ consists of two ideal tetrahedrons

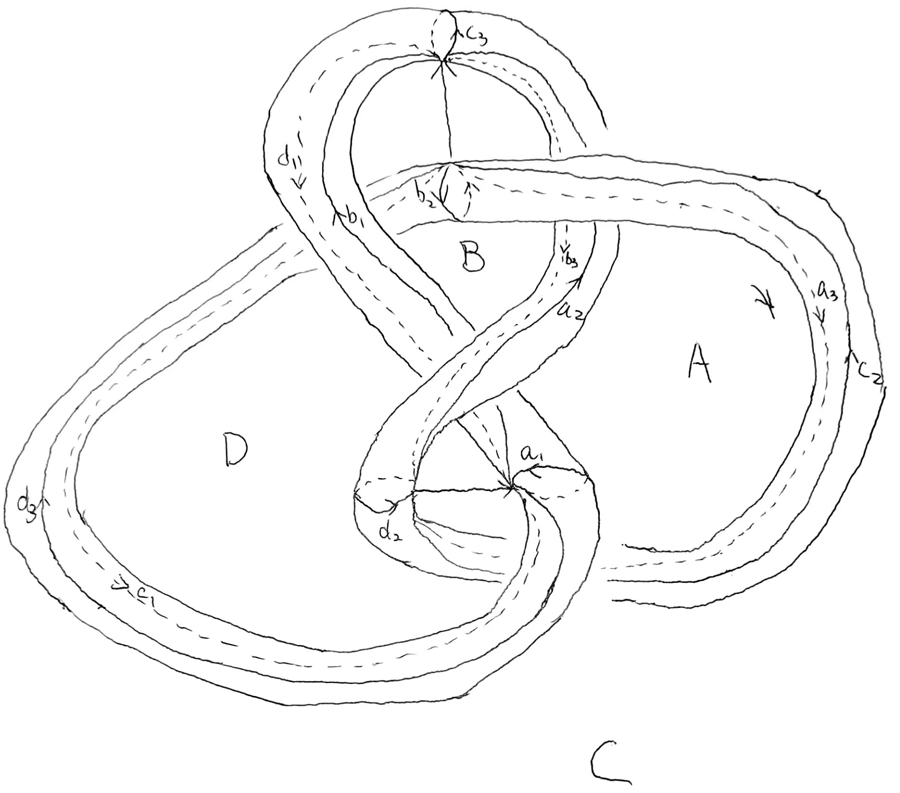


Figure 14: Hyperbolic Structure of Figure-8 Knot Complement

The link of the ideal vertex is a torus which is triangulated as is shown in Figure 15.

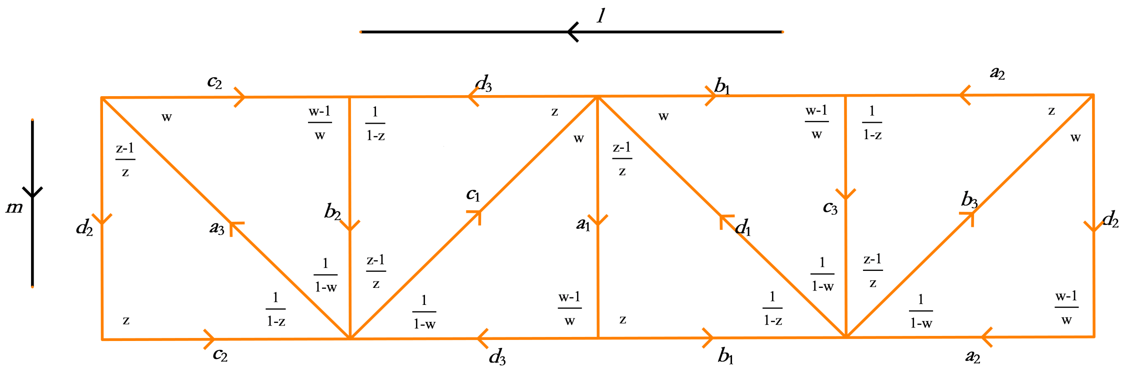


Figure 15: Triangulation of the Link of the Ideal Vertex

The preferred longitude is marked by $a_2b_1^{-1}d_3c_2^{-1}$*, and the meridian is marked by* d_2*. If we define a hyperbolic structure on the knot complement* $S^3 \setminus K$*, by determining the dihedral angle parameters of the two ideal tetrahedrons, then we can calculate* $H'(l) = \left(\frac{1}{w(1-w)}\right)^2$*,* $H'(m) = \frac{1}{w(1-z)}$*.*
The Hyperbolic Gluing Equation for a $\frac{p}{q}$ *Dehn surgery is:*

$$\begin{array}{rl} \textit{Thurston's Equation:} & z(z-1)w(w-1) = 1 \\ \textit{Dehn Filling Equation:} & (\frac{1}{w(1-z)})^p(\frac{1}{w(1-w)})^{2q} = 1 \end{array} \tag{4.3}$$

Theorem 4.3 (Thurston[1979])**.** *The Hyperbolic Gluing Equation (4.3) has a unique solution* (z, w)*, with* $\text{Im}(z) > 0$*,* $\text{Im}(w) > 0$*, for* $q = 1$*,* $p \geq 5$*,* $p \in \mathbb{Z}$

Corollary 4.5. *3-manifolds obtained through Dehn surgery along the figure-8 knot in* S^3 *with surgery coefficient* $p \geq 5$*,* $p \in \mathbb{Z}$*, admits a complete hyperbolic structure.*

Remark 4.5. *The proof for Theorem 4.3 is based on direct calculation, readers may check [9] for details.*

4.5 Hyperbolic Volume as an invariant

Obtaining a geometrical explanation of the 3-manifold enables us to obtain a geometrical invariant of the 3-manifold.
The most typical result of geometrical invariants is the hyperbolic volume.

Theorem 4.4 (Mostow's Rigidity Theorem[1973])**.** *Suppose* $n \geq 3$*,* M_1^n *and* M_2^n *are complete hyperbolic n-manifolds with finite total volume. If* $\pi_1(M_1)$ *and* $\pi_1(M_2)$ *are isomorphic, then* M_1 *and* M_2 *are isometric.*

Corollary 4.6. *Homeomorphic complete hyperbolic 3-manifolds are isometric, hence hyperbolic volume is an invariant for complete hyperbolic 3-manifolds.*

The hyperbolic volume can be generalized to deal with 3-manifolds that are not (as a whole) a hyperbolic manifold, the ideas are closely related with Thurston's Geometrization Conjecture.

Definition 4.6. *Suppose that M is an orientable, closed 3-manifold. We define the Gromov Norm of M be:*

$$\|M\| = \frac{1}{v_3}\sum Vol(H_i)$$

Where H_i are all the hyperbolic components of M obtained by connected-sum decomposition and JSJ-decomposition, v_3 is the volume of the regular hyperbolic ideal tetrahedron.
The Gromov Norm is an invariant for orientable, closed 3-manifolds.

Proposition 4.6.

$$\begin{array}{rccc} F: & \{\textit{Closed, Oriented 3-Manifolds}\} & \to & \mathbb{C} \\ & M & \mapsto & e^{\|M\|} \end{array}$$

Satisfies Witten's mathematical requirements in Example 3.2.

Proof: This is direct from $\|S^3\| = 0$, and $Vol(B^3) = 0$. □

According to our methods in constructing hyperbolic 3-manifolds, we are essentially concerned with the hyperbolic volume of the ideal tetrahedrons.
We provide a direct formula in calculating the hyperbolic volume of ideal tetrahedrons.

Theorem 4.5 (Hyperbolic Volume of Ideal Tetrahedron)**.**
Suppose z is one dihedral angle parameter of an ideal tetrahedron Δ, then:

$$\begin{aligned} Vol(\Delta) &= \Lambda(arg(z)) + \Lambda(arg(\frac{1}{1-z})) + \Lambda(arg(\frac{z-1}{z})) \\ &= D_2(z) = D_2(\frac{1}{1-z}) = D_2(\frac{z-1}{z}) < \infty \end{aligned}$$

$\Lambda(\theta) = -\int_0^\theta \log|2\sin t|\mathrm{d}t$ *is the Lobachevsky function.*
$D_2(z) = \mathrm{ImLi}_2(z) + arg(1-z)\log|z|$ *is the Bloch-Wigner function.*

Proof: Through an orientation-preserving isometry, we may suppose one of the ideal vertices of Δ is ∞.
We divide Δ into six parts, shown in Figure 16, without loss of generality $O = 0$ is the circumcenter of the three ideal vertices.
By a hyperbolic isometry (similarity transformation on $\mathbb{C}$), we suppose the radius of the circumcenter is 1.
Write $\alpha = \arg(z)$, then:

$$\begin{aligned} Vol(V_1) &= \int_{x=0}^{\cos\alpha}\int_{y=0}^{x\tan\alpha}\int_{z=\sqrt{1-x^2-y^2}}^{+\infty}\frac{\mathrm{d}x\mathrm{d}y\mathrm{d}z}{z^3} \\ &= \int_{x=0}^{\cos\alpha}\int_{y=0}^{x\tan\alpha}\frac{\mathrm{d}x\mathrm{d}y}{2(1-x^2-y^2)} \\ &= \int_{x=0}^{\cos\alpha}\int_{\theta=0}^{\alpha}\frac{x\frac{1}{\cos^2\theta}\mathrm{d}x\mathrm{d}\theta}{2\left(1-\frac{x^2}{\cos^2\theta}\right)} \end{aligned}$$

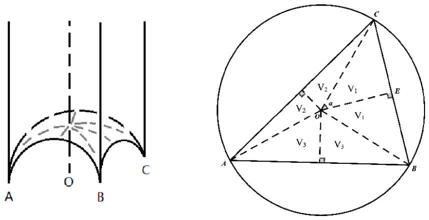


Figure 16: Division of Δ into Six Parts

$$
\begin{aligned}
&=\int_{x=0}^{\cos\alpha}\int_{\theta=0}^{\alpha}\frac{x\mathrm{d}x\mathrm{d}\theta}{2\left(x^2-\cos^2\theta\right)}\\
&=-\frac{1}{4}\int_{\theta=0}^{\alpha}\left(\log\left|\cos^2\theta-\cos^2\alpha\right|-\log\left|\cos^2\theta\right|\right)\mathrm{d}\theta\\
&=-\frac{1}{4}\int_{\theta=0}^{\alpha}\left(\log\left|4\sin\frac{\theta+\alpha}{2}\sin\frac{\theta-\alpha}{2}\cos\frac{\theta+\alpha}{2}\cos\frac{\theta-\alpha}{2}\right|-2\log\left|\cos\theta\right|\right)\mathrm{d}\theta\\
&=-\frac{1}{4}(2\int_{0}^{\alpha}\log 2\mathrm{d}\theta+2\int_{\frac{\alpha}{2}}^{\alpha}\log\left|\sin\theta\right|\mathrm{d}\theta+2\int_{-\frac{\alpha}{2}}^{0}\log\left|\sin\theta\right|\mathrm{d}\theta+2\int_{\frac{\alpha}{2}}^{\alpha}\log\left|\cos\theta\right|\mathrm{d}\theta\\
&\quad+2\int_{-\frac{\alpha}{2}}^{0}\log\left|\cos\theta\right|\mathrm{d}\theta-2\int_{0}^{\alpha}\log\left|\cos\theta\right|\mathrm{d}\theta)\\
&=-\frac{1}{2}\int_{0}^{\alpha}\log\left|2\sin\theta\right|\mathrm{d}\theta\\
&=\frac{1}{2}\Lambda(\alpha)\\
&=\frac{1}{2}\Lambda(\arg(z))
\end{aligned}
$$

Similarly, $Vol(V_2)=\frac{1}{2}\Lambda(\arg(\frac{1}{1-z}))$, and $Vol(V_3)=\frac{1}{2}\Lambda(\arg(\frac{z-1}{z}))$.
Thus, $Vol(\Delta)=2Vol(V_1)+2Vol(V_2)+2Vol(V_3)=\Lambda(arg(z))+\Lambda(arg(\frac{1}{1-z}))+\Lambda(arg(\frac{z-1}{z}))$.
Next, we consider the Bloch-Wigner function, suppose $z=re^{i\theta}$, $\theta\in(0,\pi)$, then:

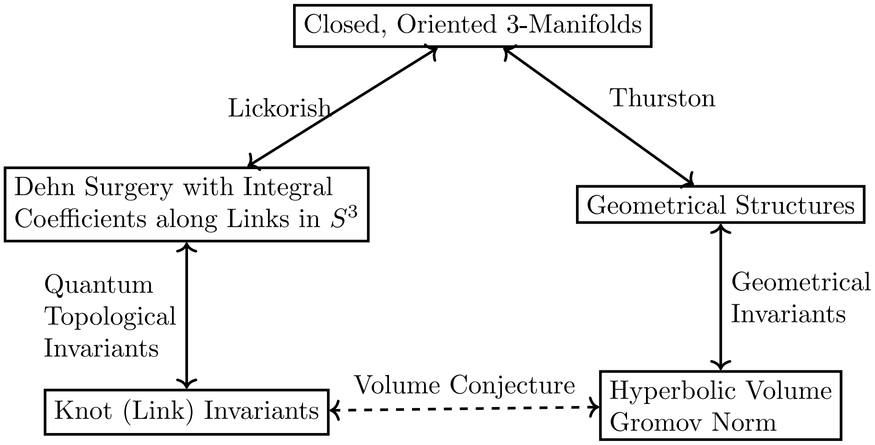
$$
\begin{aligned}
\mathrm{Im}\mathrm{Li}_2(z)=&-\mathrm{Im}\int_0^r\frac{\log(1-\xi e^{i\theta})}{\xi}\mathrm{d}\xi\\
=&\int_0^r\arctan\left(\frac{\xi\sin\theta}{1-\xi\cos\theta}\right)\frac{\mathrm{d}\xi}{\xi}\\
&\left(\text{substitute: }\frac{\xi\sin\theta}{1-\xi\cos\theta}=t\right)\\
=&\int_0^{\frac{r\sin\theta}{1-r\cos\theta}}\arctan(t)\left(\frac{1}{t}-\frac{1}{t+\tan\theta}\right)\mathrm{d}t\\
&(\text{substitute: } t=\tan\phi\text{ , and let }\omega=-arg(1-z))
\end{aligned}
$$

$$\begin{aligned}
&=\int_0^{\omega} \phi \frac{\sin\theta}{\sin\phi\sin(\phi+\theta)}\mathrm{d}\theta \\
&=\int_0^{\omega} \phi\, \mathrm{d}\left(\log\sin\phi - \log\sin(\phi+\theta)\right) \\
&=\omega\log\frac{\sin\omega}{\sin(\omega+\theta)} - \int_0^{\omega}\log\sin\phi\mathrm{d}\phi + \int_0^{\omega}\log\sin(\phi+\theta)\mathrm{d}\phi \\
&=\omega\log\frac{\sin\omega}{\sin(\omega+\theta)} - \int_0^{\omega}\log(2\sin\phi)\mathrm{d}\phi + \int_0^{\omega}\log(2\sin(\phi+\theta))\mathrm{d}\phi \\
&=\omega\log r + \Lambda(\omega) - \Lambda(\theta+\omega) + \Lambda(\theta) \\
&=-arg(1-z)\log|z| + \Lambda(\omega) + \Lambda(\pi-\theta-\omega) + \Lambda(\theta) \\
&=-arg(1-z)\log|z| + \Lambda(arg(z)) + \Lambda(arg(\frac{1}{1-z})) + \Lambda(arg(\frac{z-1}{z})) \\
&=-arg(1-z)\log|z| + Vol(\Delta)
\end{aligned}$$

Hence, $Vol(\Delta) = D_2(z)$.
Similarly, $Vol(\Delta) = D_2(z) = D_2(\frac{1}{1-z}) = D_2(\frac{z-1}{z})$. □

5 Volume Conjectures

We have now introduced a quantum topological method and a geometrical method in obtaining invariants for closed, oriented three manifolds. The relationship between these two invariants is the volume conjecture.


Closed, Oriented 3-Manifolds
Lickorish
Thurston
Dehn Surgery with Integral Coefficients along Links in S^3
Geometrical Structures
Quantum Topological Invariants
Geometrical Invariants
Knot (Link) Invariants
Volume Conjecture
Hyperbolic Volume Gromov Norm

The most famous volume conjecture is due to Kashaev, which relates colored Jones polynomials with the hyperbolic volume of knot complements.

Conjecture 5.1 (Kashaev[1997], H. Murakami-J. Murakami[2001])**.**

$$\underset{n\to\infty}{lim}\frac{2\pi}{n}\log|V_n(K, e^{\frac{2\pi i}{n}})| = v_3\cdot\|S^3\setminus K\|$$

***Note**: If K is a hyperbolic knot, then $v_3 \cdot \|S^3 \setminus K\| = Vol(S^3 \setminus K)$*

In this paper, we will focus on Chen-Yang's volume conjecture (reference [2]) , which relates Reshetikhin-Turaev invariants with the hyperbolic volume of closed, oriented 3-manifolds.
The complete version of Chen-Yang's Volume Conjecture is as follows:

Conjecture 5.2 (Chen-Yang[2018])**.** *Suppose M is a closed, oriented 3-manifold. Let $RT_r(M)$ denote the r-th Reshetikhin-Turaev invariant evaluated at $q = e^{\frac{2\pi i}{r}}$, then when r is an odd positive integer:*

$$RT_r(M) = \frac{C_r}{2} \frac{1}{\sqrt{Tor(M, Ad_\rho)}} e^{\frac{r}{4\pi}(v_3 \cdot \|M\| + iCS(M))}(1 + O(\frac{1}{r}))$$

C_r is a constant of norm 1 independent of M.

In this paper, we will only deal with a simple version of this conjecture:

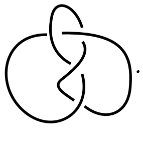
Conjecture 5.3 (Simple Version of Chen-Yang's volume conjecture)**.** *Suppose M is a closed, oriented hyperbolic 3-manifold. Let $r \geq 3$ be an odd integer, $RT_r(M)$ denotes the r-th Reshetikhin-Turaev invariant evaluated at $A = e^{\frac{\pi i}{r}}$, then:*

$$\lim_{\substack{r \to +\infty \\ r\ odd}} \frac{4\pi}{r} \log |RT_r(M)| = Vol(M)$$

Remark 5.1. *Ohtsuki's method is probably effective when applied to manifolds obtained through Dehn surgery along Whitehead twisted knots.*
However, existing proofs for special cases of Chen-Yang's volume conjecture are all, to some extent, based on specific calculations to express the Reshetikhin-Turaev invariants.
Since the topological essence of Kauffman Bracket is still not clearly explained, Chen-Yang's volume conjecture becomes further more difficult concerning links or 3-manifolds whose Kauffman bracket is difficult to compute.
However, Chen-Yang's volume conjecture contains some information beyond Witten's quantum group, which is still inspiring in the field of physics.

6 Ohtsuki's Method on Figure-8 Knot

In this section, we will sketch the framework of Ohtsuki's work on special cases of Chen-Yang's volume conjecture (reference [6]). We will only review a part of Ohtsuki's results. This part of proof includes all the essential methods used by Ohtsuki, and can possibly be generalized to the cases of Whitehead twisted knots.
We will give a proof of the following theorem:

Theorem 6.1 (T. Ohtsuki[2018])**.** *For integers $p \geq 6$, let M_p denote the 3-manifold obtained through Dehn surgery with coefficient p along the figure-8 knot $K =$* .

Let $RT_r(M_p)$ denote the Reshetikhin-Turaev invariant of M_p valued at $t = A^4 = e^{\frac{4\pi i}{r}}$. Then M_p is a closed, orientable hyperbolic 3-manifold, and Chen-Yang's Conjecture holds:

$$\lim_{\substack{r\to+\infty \\ r\ odd}} \frac{4\pi}{r} \log |RT_r(M_p)| = Vol(M_p)$$

Notation 6.1. *In Section 6, we will use the following quantum notaions:*

$$t = e^{\frac{4\pi i}{r}}$$
$$[\,n\,] = \frac{t^{\frac{n}{2}} - t^{-\frac{n}{2}}}{t^{\frac{1}{2}} - t^{-\frac{1}{2}}}$$
$$\{\,n\,\} = t^{\frac{n}{2}} - t^{-\frac{n}{2}}$$
$$\{\,n\,\}! = \prod_{k=1}^{n} \{k\}$$
$$(t)_n = \prod_{k=1}^{n} (1 - t^k)$$

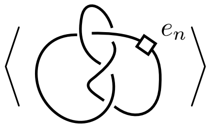
Notation 6.2. *In Section 6, Kauffman bracket for cabled (colored) links implicitly denotes the cabling on the blackboard framing to avoid over-complicated link diagrams.*

Notation 6.3. *In Section 6, we will use $O(\cdot)$ to indicate remainders that can be uniformly controlled on any compact subset.*

6.1 Direct Calculation of the Reshetikhin-Turaev Invariants

We first calculate the Reshetikhin-Turaev invariant of M_p.

Lemma 6.1 (Habiro's Formula (K. Habiro[2000])).



$$= \frac{(-1)^n}{\{1\}} \sum_{m=0}^{n} \frac{\{n+1+m\}!}{\{n-m\}!}$$

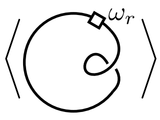
The knot diagram denotes a Chebyshev cabling on the figure-8 knot.

Remark 6.1. *For a skein theoretical proof of Lemma 6.1, readers can check up in reference [5].*

Corollary 6.1. *We can obtain a concrete expression of $RT_r(M_p)$ from Lemma 6.1:*

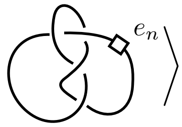
$$RT_r(M_p) = {c_r}' \sum_{n=-\frac{r-2}{2}}^{\frac{r-2}{2}} \sum_{m=|n|}^{\frac{r-2}{2}} [n] t^{\frac{pn^2}{4} - mn - \frac{n}{2}} \frac{(t)_{r-1-m-n}}{(t)_{m-n}}$$

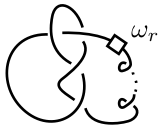
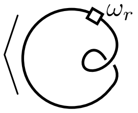
The coefficient $|{c_r}'| = |\frac{c_r}{\{1\}}|$, and $c_r =$

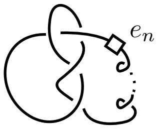


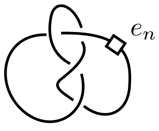
$$^{-1}$$

Proof: When $t = e^{\frac{4\pi i}{r}}$, then $\{r\} = 0$, and we have:

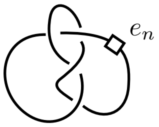
$\Big\langle$  $\Big\rangle = \frac{(-1)^{n+1}}{\{1\}} \sum_{m=0}^{min\{n,r-2-n\}} t^{-(n+1)(m+\frac{1}{2})} \frac{(t)_{n+1+m}}{(t)_{n-m}}$

$RT_r(M_p) = \Big\langle$  $\Big\rangle \Big\langle$  $\Big\rangle^{-1}$

$= c_r \sum_{n=0}^{r-2} (-1)^n [n+1] \Big\langle$  $\Big\rangle$

$= c_r \sum_{n=0}^{r-2} (-1)^n [n+1] \Big\langle (t^+)^p \Bigg($  $\Bigg) \Big\rangle$

(By Proposition 3.4)

$= c_r \sum_{n=0}^{r-2} (-1)^n [n+1] ((-1)^n t^{\frac{n^2+2n}{4}})^p \Big\langle$  $\Big\rangle$

$$= \frac{c_r}{\{1\}} \sum_{n=0}^{r-2} \sum_{m=0}^{min\{n,r-2-n\}} (-1)^{np+1} [n+1] t^{\frac{pn(n+2)}{4} - (n+1)(m+\frac{1}{2})} \frac{(t)_{n+1+m}}{(t)_{n-m}}$$

(Substitute $n' = \dfrac{r-2}{2} - n$, $m' = \dfrac{r-2}{2} - m$)

$$= \frac{c_r}{\{1\}} \sum_{n'=-\frac{r-2}{2}}^{\frac{r-2}{2}} \sum_{m'=|n|}^{\frac{r-2}{2}} (-1)^{\frac{r-2}{2}p - n'p + 1} \frac{t^{\frac{\frac{r}{2}-n'}{2}} - t^{-\frac{\frac{r}{2}-n'}{2}}}{t^{\frac{1}{2}} - t^{-\frac{1}{2}}} t^{\frac{pn'^2}{4} - m'n' - \frac{n'}{2}}$$

$$t^{\frac{p}{4}\frac{r^2-4}{4} - \frac{r(r-1)}{4}} (-1)^{-n'p+2} \frac{(t)_{r-1-m'-n'}}{(t)_{m'-n'}}$$

$$= c_r{}' \sum_{n=-\frac{r-2}{2}}^{\frac{r-2}{2}} \sum_{m=|n|}^{\frac{r-2}{2}} [n] t^{\frac{pn^2}{4} - mn - \frac{n}{2}} \frac{(t)_{r-1-m-n}}{(t)_{m-n}}$$

□

Lemma 6.2.

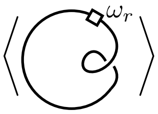
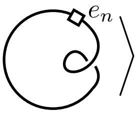
$$\underset{\substack{r\to\infty\\ r\ odd}}{lim}\frac{1}{r}\log|{c_r}'|=0$$

Proof: First, by induction on the Chebyshev polynomials, it is obvious that:

(i) $e_n(-z)=(-1)^n e_n(z)$

(ii) $e_n(2\cos\theta)=\frac{\sin((n+1)\theta)}{\sin\theta}$

Hence, when r is odd:

$\Bigg\langle$  $\Bigg\rangle = \sum_{n=0}^{r-2}(-1)^n\,[n+1]\,\Bigg\langle$  $\Bigg\rangle$

$$\begin{aligned}
&=\sum_{n=0}^{r-2}(-1)^n\,[n+1]\,(-1)^n t^{\frac{n^2+2n}{4}}\,\langle e_n\rangle\\
&=\sum_{n=0}^{r-2}(-1)^n\,[n+1]\,(-1)^n t^{\frac{n^2+2n}{4}}\,e_n(-e^{\frac{2pi}{r}}-e^{-\frac{2pi}{r}})\\
&=\sum_{n=0}^{r-2}[n+1]\,t^{\frac{n^2+2n}{4}}e_n(-2\cos\frac{2\pi}{r})\\
&=\sum_{n=0}^{r-2}(-1)^n\,[n+1]\,t^{\frac{n^2+2n}{4}}\frac{\sin\frac{2(n+1)\pi}{r}}{\sin\frac{2\pi}{r}}\\
&=\frac{1}{4(\sin\frac{2\pi}{r})^2}\sum_{n=0}^{r-1}(-1)^n(t^{\frac{n}{2}}-t^{-\frac{n}{2}})^2t^{\frac{n^2-1}{4}}\\
&=\frac{t^{-\frac{1}{4}}}{4(\sin\frac{2\pi}{r})^2}\sum_{n=0}^{r-1}(-1)^{n^2}\left(e^{\frac{\pi i}{r}(n-2)^2}t^{-1}-2e^{\frac{\pi i}{r}n^2}+e^{\frac{\pi i}{r}(n+2)^2}t^{-1}\right)\\
&=\frac{t^{-\frac{1}{4}}(t^{-1}-1)}{2(\sin\frac{2\pi}{r})^2}\sum_{n=0}^{r-1}(-e^{\frac{\pi i}{r}})^{n^2}\\
&=\frac{t^{-\frac{1}{4}}(t^{-1}-1)}{2(\sin\frac{2\pi}{r})^2}(-i)^{\frac{r-1}{2}}\sqrt{r}
\end{aligned}$$

$$\lim_{\substack{r\to\infty\\ r\text{ odd}}}\frac{1}{r}\log|c_r|=\lim_{\substack{r\to\infty\\ r\text{ odd}}}\frac{-\frac{3}{2}\log r}{r}=0$$

$$\lim_{\substack{r\to\infty\\ r\text{ odd}}}\frac{1}{r}\log|\{1\}|=\lim_{\substack{r\to\infty\\ r\text{ odd}}}\frac{-\log r}{r}=0$$

Hence, $\underset{\substack{r\to\infty\\ r\text{ odd}}}{lim}\frac{1}{r}\log|{c_r}'|=0$ □

Now, what we are concerned with $RT_r(M_p)$ is essentially the sum of quantum values at half integer points, Ohtsuki's method is aimed to estimate such expressions.

Ohtsuki's method can be divided into three steps:

Step 1. Write $RT_r(M_p)$ as sum of values of a holomorphic function at (half) integral points:

$$RT_r(M_p) = c_1 \sum_{(m,n)\in\mathbb{Z}^2} f(m,n) + c_2$$

Step 2. Use **Poisson Summation Formula** to write $RT_r(M_p)$ in the form of integrations:

$$\sum_{(m,n)\in\mathbb{Z}^2} f(m,n) = \sum_{(m,n)\in\mathbb{Z}^2} \hat{f}(m,n)$$

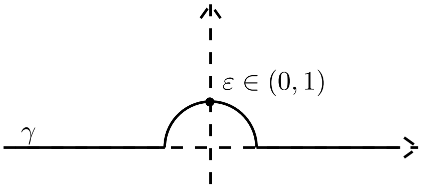
Step 3. Use **Saddle Point Method** to calculate the limit of the integral through its critical value.

6.2 Step 1 of Ohtsuki's Method — Sum of values at integral points

For Step 1, we introduce Faddev's quantum dilogarithm function to simplify the expressions.

Definition 6.1 (Quantum Dilogarithm Function (Faddev[1995], Kashaev[2001])).

$$\varphi_r(z) = \frac{4\pi i}{r}\int_\gamma \frac{e^{(2z-\pi)x}}{4x\sinh(\pi x)\sinh(\frac{2\pi x}{r})}\mathrm{d}x \quad (r \geq 3)$$
$$\gamma = (-\infty,-\varepsilon] \cup \{z\in\mathbb{C}|\mathrm{Im}(z)\geq 0,\ |z|=\varepsilon\} \cup [\varepsilon,+\infty)$$



$\varphi_r(z)$ is holomorphic in $\left\{z\in\mathbb{C} \mid -\frac{\pi}{r} < \mathrm{Re}z < \pi + \frac{\pi}{r}\right\}$, by comparing the exponents.

Lemma 6.3 (Functional Properties of Quantum Dilogarithm Function).

(1) For $0 < \mathrm{Re}z < \pi$:

$$1 - e^{2iz} = e^{\frac{r}{4\pi i}(\varphi_r(z-\frac{\pi}{r})-\varphi_r(z+\frac{\pi}{r}))}$$

(2) For $-\frac{\pi}{r} < \mathrm{Re}z < \frac{\pi}{r}$:

$$1 + e^{riz} = e^{\frac{r}{4\pi i}(\varphi_r(z)-\varphi_r(z+\pi))}$$

Proof: Since the functions are all holomorphic, it suffices to prove the situation when $z\in\mathbb{R}$.

(1) By definition:

$$\frac{r}{4\pi i}(\varphi_r(z-\frac{\pi}{r})-\varphi_r(z+\frac{\pi}{r}))=-\int_\gamma \frac{e^{(2z-\pi)x}}{2x\sinh(\pi x)}\mathrm{d}x$$

Notice that, by Dominated Convergence Theorem:

$$\limsup_{N\to\infty}\left|\int_{\substack{|x|=N+\frac{1}{2}\\ \mathrm{Im}(x)>0}} \frac{e^{(2z-\pi)x}}{2x\sinh(\pi x)}\mathrm{d}x\right| = \limsup_{N\to\infty}\left|\int_0^\pi \frac{e^{(2z-\pi)(N+\frac{1}{2})e^{i\theta}}}{2\sinh(\pi(N+\frac{1}{2})e^{i\theta})}\mathrm{d}\theta\right| = 0$$

Take $\Gamma_N=[-(N+\frac{1}{2}),-\varepsilon]\cup\{z\in\mathbb{C}|\mathrm{Im}(z)\geq 0,\,|z|=\varepsilon\}\cup[\varepsilon,N+\frac{1}{2}]\cup\big\{z\in\mathbb{C}|\mathrm{Im}(z)\geq 0,\,|z|=N+\frac{1}{2}\big\}$. Then, by Cauchy's Theorem and Residue Theorem:

$$\begin{aligned}\int_\gamma \frac{-e^{(2z-\pi)x}}{2x\sinh(\pi x)}\mathrm{d}x =& \lim_{N\to\infty}\int_{\Gamma_N}\frac{-e^{(2z-\pi)x}}{2x\sinh(\pi x)}\mathrm{d}x\\ =&2\pi i\sum_{n=1}^\infty Res_{ni}\left(\frac{-e^{(2z-\pi)x}}{2x\sinh(\pi x)}\right)\\ =&-2\pi i\sum_{n=1}^\infty\frac{e^{(2z-\pi)ni}}{2ni\pi\cosh(\pi ni)}\\ =&-2\pi i\sum_{n=1}^\infty\frac{e^{2zin}(-1)^n}{2ni\pi(-1)^n}\\ =&-\sum_{n=1}^\infty\frac{e^{2zin}}{n}\\ =&\log(1-e^{2iz})\end{aligned}$$

(2) By definition:

$$\frac{r}{4\pi i}(\varphi_r(z)-\varphi_r(z+\pi))=-\int_\gamma\frac{e^{2zx}}{2x\sinh(\frac{2\pi x}{r})}\mathrm{d}x$$

Again, by Dominated Convergence Theorem:

$$\limsup_{N\to\infty}\left|\int_{\substack{|x|=\frac{r}{2}(N+\frac{1}{2})\\ \mathrm{Im}(x)>0}}\frac{e^{2zx}}{2x\sinh(\frac{2\pi x}{r})}\mathrm{d}x\right|=0$$

Hence:

$$\begin{aligned}\int_\gamma\frac{-e^{2zx}}{2x\sinh(\frac{2\pi x}{r})}\mathrm{d}x =&2\pi i\sum_{n=1}^\infty Res_{\frac{r}{2}ni}\left(\frac{-e^{2zx}}{2x\sinh(\frac{2\pi x}{r})}\right)\\ =&-2\pi i\sum_{n=1}^\infty\frac{e^{zrin}}{rni\frac{2\pi}{r}\cosh(\pi ni)}\end{aligned}$$

$$
\begin{aligned}
&=\sum_{n=1}^{\infty}\frac{(-1)^{n+1}e^{rizn}}{n}\\
&=\log(1+e^{riz})
\end{aligned}
$$

□

Remark 6.2. *The functional properties indicate an analytic continuation for φ_r.*
φ_r is a meromorphic function on $\mathbb{C}$, with poles: $(a+1)\pi+\frac{b\pi}{r}$ and $-a\pi-\frac{b\pi}{r}$ $(a,b\in\mathbb{N},\ b$ odd).

Corollary 6.2. *For $r\geq 3$ an odd number, and $0\leq n\leq r-1$*

$$(t)_n=e^{\frac{r}{4\pi i}(\varphi_r(\frac{\pi}{r})-\varphi_r(\frac{2n\pi}{r}+\frac{\pi}{r}))}$$

Proof:

$$
\begin{aligned}
(t)_n&=\prod_{k=1}^{n}(1-t^k)\\
&=\prod_{k=1}^{n}(1-e^{\frac{4ik\pi}{r}})\\
&=\prod_{k=1}^{n}e^{\frac{r}{4\pi i}(\varphi_r(\frac{2k\pi}{r}-\frac{\pi}{r})-\varphi_r(\frac{2k\pi}{r}+\frac{\pi}{r}))}\\
&=e^{\frac{r}{4\pi i}(\varphi_r(\frac{\pi}{r})-\varphi_r(\frac{2n\pi}{r}+\frac{\pi}{r}))}
\end{aligned}
$$

Notice that the condition r is odd ensures that we will not encounter singularities in this calculation.

□

Corollary 6.3. *Let $r\geq 3$ be an odd number.*

$$RT_r(M_p)=\frac{2{c_r}'}{\sin\frac{2\pi}{r}}\sum_{n=-\frac{r-2}{2}}^{\frac{r-2}{2}}\sum_{m=|n|}^{\frac{r-2}{2}}g_r(\frac{2n\pi}{r},\frac{2m\pi}{r})$$

The function g_r is defined as follows:

$$
\begin{aligned}
g_r(x,y)&=\sin x\ e^{-ix}e^{\frac{r}{4\pi i}V_r(x,y)}\\
V_r(x,y)&=-px^2+4xy+\varphi_r(y-x+\frac{\pi}{r})-\varphi_r(\pi-x-y-\frac{\pi}{r})
\end{aligned}
$$

Write ${c_r}''=\frac{2{c_r}'}{\sin\frac{2\pi}{r}}$

Proof: According to Corollary 6.1, and Corollary 6.2:

$$RT_r(M_p)={c_r}'\sum_{n=-\frac{r-2}{2}}^{\frac{r-2}{2}}\sum_{m=|n|}^{\frac{r-2}{2}}[n]t^{\frac{pn^2}{4}-mn-\frac{n}{2}}\frac{(t)_{r-1-m-n}}{(t)_{m-n}}$$

$$=\frac{{c_r}'}{\sin\frac{2\pi}{r}}\sum_{n=-\frac{r-2}{2}}^{\frac{r-2}{2}}\sum_{m=|n|}^{\frac{r-2}{2}}\sin(\frac{2n\pi}{r})e^{\frac{4\pi i}{r}(\frac{pn^2}{4}-mn-\frac{n}{2})}e^{\frac{r}{4\pi i}(\varphi_r(\frac{2(m-n)\pi}{r}+\frac{\pi}{r})-\varphi_r(\frac{2(r-1-m-n)\pi}{r}+\frac{\pi}{r}))}$$

$$=\frac{{c_r}'}{\sin\frac{2\pi}{r}}\sum_{n=-\frac{r-2}{2}}^{\frac{r-2}{2}}\sum_{m=|n|}^{\frac{r-2}{2}}\sin(\frac{2n\pi}{r})e^{\frac{4\pi i}{r}(\frac{pn^2}{4}-mn-\frac{n}{2})}e^{\frac{r}{4\pi i}(\varphi_r(\frac{2(m-n)\pi}{r}+\frac{\pi}{r})-\varphi_r(\frac{(r-2-2m-2n)\pi}{r}+\frac{\pi}{r}))}$$

$$(1+e^{ri(\frac{(r-2-2m-2n)\pi}{r}+\frac{\pi}{r})})$$

$$=\frac{2{c_r}'}{\sin\frac{2\pi}{r}}\sum_{n=-\frac{r-2}{2}}^{\frac{r-2}{2}}\sum_{m=|n|}^{\frac{r-2}{2}}\sin(\frac{2n\pi}{r})e^{\frac{4\pi i}{r}(\frac{pn^2}{4}-mn-\frac{n}{2})}e^{\frac{r}{4\pi i}(\varphi_r(\frac{2(m-n)\pi}{r}+\frac{\pi}{r})-\varphi_r(\frac{(r-2-2m-2n)\pi}{r}+\frac{\pi}{r}))}$$

$$=\frac{2{c_r}'}{\sin\frac{2\pi}{r}}\sum_{n=-\frac{r-2}{2}}^{\frac{r-2}{2}}\sum_{m=|n|}^{\frac{r-2}{2}}g_r(\frac{2n\pi}{r},\frac{2m\pi}{r})$$

□

Lemma 6.4. *For* $0<\mathrm{Re}z<\pi$*:*

$$\varphi_r(z)=Li_2(e^{2iz})+O(\frac{1}{r^2})$$

Proof: Suppose $z=a+bi$, $a,b\in\mathbb{R}$, $0<a<\pi$.
Then:

$$\begin{aligned}
Li_2(e^{2iz})=&-\int_0^{e^{2iz}}\frac{\log(1-\xi)}{\xi}\mathrm{d}\xi\\
=&\int_b^{+\infty}(-2)\log(1-e^{2i(a+yi)})\mathrm{d}y\\
=&\int_b^{+\infty}(-2)\frac{r}{4\pi i}(\varphi_r(a+yi-\frac{\pi}{r})-\varphi_r(a+yi+\frac{\pi}{r}))\mathrm{d}y\\
=&\frac{r}{2\pi}\lim_{R\to+\infty}\left(\int_{a-\frac{\pi}{r}+bi}^{a-\frac{\pi}{r}+Ri}\varphi_r(w)\mathrm{d}w-\int_{a+\frac{\pi}{r}+bi}^{a+\frac{\pi}{r}+Ri}\varphi_r(w)\mathrm{d}w\right)\\
&\text{(By Cauchy's Theorem)}\\
=&\frac{r}{2\pi}\lim_{R\to+\infty}\left(\int_{z-\frac{\pi}{r}}^{z+\frac{\pi}{r}}\varphi_r(w)\mathrm{d}w-\int_{a-\frac{\pi}{r}+Ri}^{a+\frac{\pi}{r}+Ri}\varphi_r(w)\mathrm{d}w\right)\\
&\text{(By Dominated Convergence Theorem)}\\
=&\frac{r}{2\pi}\int_{z-\frac{\pi}{r}}^{z+\frac{\pi}{r}}\varphi_r(w)\mathrm{d}w\\
=&\varphi_r(z)+O(\frac{1}{r^2})
\end{aligned}$$

Because the order-1 term in the Taylor expansion are cancelled out, and the remainder is controlled on any compact subset. □

Remark 6.3. *Lemma 6.4 is the reason why φ_r is referred to as the quantum dilogarithm function.*

Lemma 6.5. *For $\theta \in \mathbb{R}$:*

$$\mathrm{Im}\mathrm{Li}_2(e^{2i\theta}) = 2\Lambda(\theta)$$

Proof:

$$\begin{aligned}
\mathrm{Im}\mathrm{Li}_2(e^{2i\theta}) =& \mathrm{Im}\left(-\int_0^{e^{2i\theta}} \frac{\log(1-\xi)}{\xi}\mathrm{d}\xi\right)\\
=& \mathrm{Im}\left(-\int_0^{1} \frac{\log(1-\xi)}{\xi}\mathrm{d}\xi - \int_1^{e^{2i\theta}} \frac{\log(1-\xi)}{\xi}\mathrm{d}\xi\right)\\
=& \mathrm{Im}\int_0^{\theta} 2i\log(1-e^{2it})\mathrm{d}t\\
=& 2\mathrm{Re}\int_0^{\theta} \log(1-e^{2it})\mathrm{d}t\\
=& 2\mathrm{Re}\int_0^{\theta} \log|2\sin t|\mathrm{d}t\\
=& 2\Lambda(\theta)
\end{aligned}$$

□

Corollary 6.4.

$$\mathrm{Im}V_r(x,y) = 2\Lambda(y+x) + 2\Lambda(y-x) + O(\frac{1}{r^2})$$

Proof: This is direct from Lemma 6.4 and 6.5. □

6.3 Step 2 of Ohtsuki's Method — Poisson Summation

Now we are able to move on to Step 2, which is to transform $RT_r(M_p)$ through Poisson summation. First, we can further simplify the summation.

Notation 6.4. *For $\delta \geq 0$, define:*
$D_\delta = \left\{(x,y) \in \mathbb{R}^2 | \delta < x+y < \frac{\pi}{2} - \delta,\ \delta < y-x < \frac{\pi}{2} - \delta\right\}$
$D_\delta{}' = \left\{(x,y) \in \mathbb{R}^2 | \delta < x+y < \frac{\pi}{2} - \delta,\ \pi+\delta < y-x < \frac{3\pi}{2} - \delta\right\}$
$D_\delta{}'' = \left\{(x,y) \in \mathbb{R}^2 | \pi+\delta < x+y < \frac{3\pi}{2} - \delta,\ \delta < y-x < \frac{\pi}{2} - \delta\right\}$

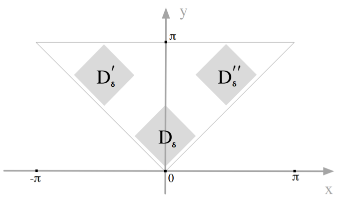


Figure 17: Image of D, D' and D'' [10]

Lemma 6.6. *If $p \geq 6$, then for fixed $\varepsilon > 0$, there exists $\delta > 0$, such that:*

$$\mathrm{Im}V_r(\frac{2\pi n}{r}, \frac{2\pi m}{r}) < Vol(M_p) - \varepsilon$$

For $\forall(\frac{2\pi n}{r}, \frac{2\pi m}{r}) \notin D_\delta \cup {D_\delta}' \cup {D_\delta}''$ and $\forall r >> 0$

Proof: According to Corollary 6.4, when $(\frac{2\pi n}{r}, \frac{2\pi m}{r}) \notin D_\delta \cup {D_\delta}' \cup {D_\delta}''$, then

$$\begin{aligned}\mathrm{Im}V_r(\frac{2\pi n}{r}, \frac{2\pi m}{r}) &= 2\Lambda\left(\frac{2\pi m}{r} + \frac{2\pi n}{r}\right) + 2\Lambda\left(\frac{2\pi m}{r} - \frac{2\pi n}{r}\right) + O\left(\frac{1}{r^2}\right)\\ &< 2\Lambda(\frac{\pi}{6}) + O\left(\frac{1}{r^2}\right)\\ &= \frac{1}{2}Vol(S^3 \setminus K) + O\left(\frac{1}{r^2}\right)\\ &< Vol(M_p) - \varepsilon + O\left(\frac{1}{r^2}\right)\end{aligned}$$

□

Notation 6.5. *Write $\Omega_\delta = D_\delta \cup {D_\delta}' \cup {D_\delta}''$*
Now we take a real C^∞ bump function ψ on $\mathbb{R}^2$, such that:

$$\begin{cases}\psi(x,y) = 1 & \text{if } (x,y) \in \overline{\Omega_\delta}\\ \psi(x,y) \in (0,1) & \text{if } (x,y) \in \Omega_{\delta/2} \setminus \overline{\Omega_\delta}\\ \psi(x,y) = 0 & \text{if } (x,y) \notin \Omega_{\delta/2}\end{cases}$$

And set $f_r(x,y) = \psi(x,y)g_r(x,y)$
Then $f_r(x,y)$ is in the Schwarz space of $\mathbb{R}^2$

Proposition 6.1 (Poisson Summation of $RT_r(M_p)$)**.**

$$RT_r(M_p) = {c_r}'' \sum_{(a,b)\in\mathbb{Z}^2} \hat{f}_r(a,b) + O(e^{\frac{r}{4\pi}(Vol(M_p)-\varepsilon)})$$

$\hat{f}_r$ is defined to be:

$$\hat{f}_r(a,b) = (-1)^{a+b}(\frac{r}{2\pi})^2 \int_{\Omega_{\delta/2}} \psi(x,y)\sin x\, e^{-ix} e^{\frac{r}{4\pi i}(V_r(x,y)-4\pi a x-4\pi b y)}\mathrm{d}x\mathrm{d}y$$

Proof: By Lemma 6.6 and Poisson Summation Formula for functions in the Schwarz space:

$$\begin{aligned}
RT_r(M_p) &= \frac{2{c_r}'}{\sin\frac{2\pi}{r}} \sum_{n=-\frac{r-2}{2}}^{\frac{r-2}{2}} \sum_{m=|n|}^{\frac{r-2}{2}} g_r(\frac{2n\pi}{r}, \frac{2m\pi}{r}) \\
&= {c_r}'' \sum_{(m,n)\in(\mathbb{Z}+\frac{1}{2})^2} f_r(\frac{2n\pi}{r}, \frac{2m\pi}{r}) + O(e^{\frac{r}{4\pi}(Vol(M_p)-\varepsilon)}) \\
&= {c_r}'' \sum_{(a,b)\in\mathbb{Z}^2} \hat{f}_r(a,b) + O(e^{\frac{r}{4\pi}(Vol(M_p)-\varepsilon)})
\end{aligned}$$

□

6.4 Step 3 of Ohtsuki's Method — Saddle Point Method

We are now facing a typical form of integration:

$$I(r) = \int_\Omega g(\mathbf{x})e^{rf(\mathbf{x})}\mathrm{d}\mathbf{x}$$

We are going to estimate the behaviour of $I(r)$ as $r \to +\infty$, this leads us to Step 3 to adapt the Saddle Point Method (reference [3]).

Theorem 6.2 (Saddle Point Method — Real Version)**.** *Suppose Ω is a bounded, closed region in $\mathbb{R}^n$, $f, g \in C^3(\Omega \to \mathbb{R})$.*
If $\mathbf{x}_0 \in \mathring{\Omega}$, such that:

(i) $f'(\mathbf{x}_0) = 0$

(ii) $Hess\, f(\mathbf{x}_0)$ *is non-singular*

(iii) $f(\mathbf{x}) < f(\mathbf{x}_0)$, $\forall \mathbf{x} \in \Omega \setminus \{\mathbf{x}_0\}$

(iv) $g(\mathbf{x}_0) \neq 0$

Then:

$$\int_\Omega g(\mathbf{x})e^{rf(\mathbf{x})}\mathrm{d}\mathbf{x} = \left(\frac{2\pi}{r}\right)^{\frac{n}{2}} \frac{1}{\sqrt{-det\, Hess\, f(\mathbf{x}_0)}} g(\mathbf{x}_0)e^{rf(\mathbf{x}_0)} \left(1 + O\left(\frac{1}{r}\right)\right)$$

Proof: For any $\epsilon > 0$, $\int_{\Omega\setminus\bar{B}(\mathbf{x}_0,\epsilon)} g(\mathbf{x})e^{rf(\mathbf{x})}\mathrm{d}\mathbf{x} = O(e^{rM})$, where $M = \sup_{\Omega\setminus\bar{B}(\mathbf{x}_0,\epsilon)} f$, is a lower term.
Hence, it suffices to consider $\Omega = \bar{B}(\mathbf{x}, \epsilon)$, with $0 < \epsilon << 1$.
Write $\Sigma = Hess\, f(\mathbf{x}_0)$.

$$\begin{aligned}
\int_\Omega g(\mathbf{x})e^{rf(\mathbf{x})}\mathrm{d}\mathbf{x} &= \int_\Omega g(\mathbf{x}_0)(1+O(\epsilon))e^{rf(\mathbf{x}_0)}e^{\frac{r}{2}(\mathbf{x}-\mathbf{x}_0)^T\Sigma(\mathbf{x}-\mathbf{x}_0)+rO(|\mathbf{x}-\mathbf{x}_0|^3)}\mathrm{d}\mathbf{x} \\
&= \frac{e^{rf(\mathbf{x}_0)}}{\sqrt{r}^n} \int_{\bar{B}(0,\sqrt{r}\epsilon)} g(\mathbf{x}_0)(1+O(\epsilon))e^{\frac{1}{2}\mathbf{x}^T\Sigma\mathbf{x}+\frac{1}{\sqrt{r}}O(|\mathbf{x}|^3)}\mathrm{d}\mathbf{x}
\end{aligned}$$

$$\text{(Take } \epsilon \sim \frac{1}{r} \to 0\text{, and by Gauss Integration)}$$
$$= \left(\frac{2\pi}{r}\right)^{\frac{n}{2}} \frac{1}{\sqrt{-det\,\Sigma}} g(\mathbf{x}_0) e^{rf(\mathbf{x}_0)} \left(1 + O\left(\frac{1}{r}\right)\right)$$

□

Theorem 6.3 (Saddle Point Method — Complex Version)**.** *Suppose S is a smoothly embedded closed disk D^n of real dimension n in $\mathbb{C}^n$, f, f_r, g are holomorphic in a neighbourhood of S.*
If $\mathbf{z}_0 \in \mathring{S}$, such that:

(i) $f'(\mathbf{z}_0) = 0$

(ii) $Hess\,f(\mathbf{z}_0)$ *is non-singular*

(iii) $\mathrm{Re} f(\mathbf{z}) < \mathrm{Re} f(\mathbf{z}_0)$, $\forall \mathbf{z}$ *in a neighbourhood of* $S \setminus \{\mathbf{z}_0\}$

(iv) $g(\mathbf{z}_0) \neq 0$

(v) $f_r(\mathbf{z}) = f(\mathbf{z}) + \frac{v_r(\mathbf{z})}{r^2}$, *where* $|v_r(\mathbf{z})|$ *is uniformly bounded in a neighbourhood of* S

Then:

$$\int_S g(\mathbf{z}) e^{rf_r(\mathbf{z})} \mathrm{d}\mathbf{z} = \left(\frac{2\pi}{r}\right)^{\frac{n}{2}} \frac{1}{\sqrt{-det\,Hess\,f(\mathbf{z}_0)}} g(\mathbf{z}_0) e^{rf(\mathbf{z}_0)} \left(1 + O\left(\frac{1}{r}\right)\right)$$

Proof: By a change of coordinates, we may assume $Hess\,f(\mathbf{z}_0) = -I$.
Then the neighbourhood N of $S \setminus \{\mathbf{z}_0\}$, such that $\mathrm{Re} f(\mathbf{z}) < \mathrm{Re} f(\mathbf{z}_0)$, $\forall \mathbf{z} \in N$, can be pictured as shown in Figure 18.

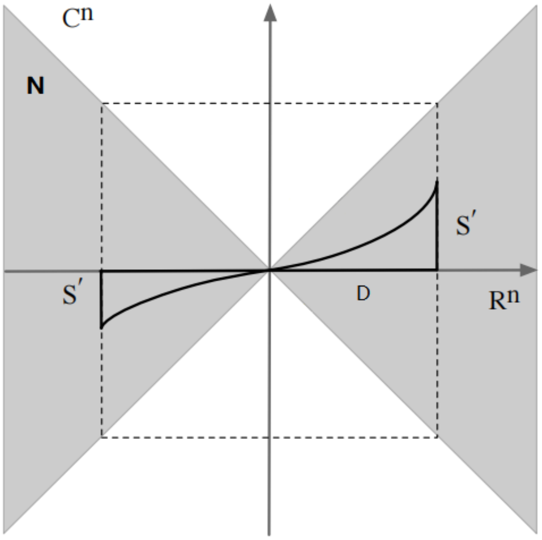


Figure 18: Deformation of Integration Area [10]

Choose a real closed ϵ-disk $D \subset \mathbb{R}^n$, $0 < \epsilon << 1$, and let $S_0 \subset S$ be the connected component of $\mathbf{z}_0$ in $\{\mathbf{z} \in S | \mathrm{Re}(\mathbf{z}) \in D\}$, such that S_0 has a positive distance with ∂S.
We replace S by $S \setminus S_0 \cup D \cup S'$, where $S' = \{\mathrm{Re}\mathbf{z} + ti\mathrm{Im}\mathbf{z} | \mathbf{z} \in S_0,\ \mathrm{Re}\mathbf{z} \in \partial D,\ t \in [0,1]\}$
Then $S' \setminus D$ has a positive distance δ with $\mathbf{z}_0$.
By Cauchy's Theorem, the deformation from S to S' will not change the result of the integration, and we can adapt Theorem 6.2 to the integration on D.
Hence:

$$\begin{aligned}
\int_S g(\mathbf{z}) e^{rf_r(\mathbf{z})} \mathrm{d}\mathbf{z} &= \int_S g(\mathbf{z}) e^{rf(\mathbf{z})} \mathrm{d}\mathbf{z} \left(1 + O\left(\frac{1}{r}\right)\right) \\
&= \int_S' g(\mathbf{z}) e^{rf(\mathbf{z})} \mathrm{d}\mathbf{z} \left(1 + O\left(\frac{1}{r}\right)\right) \\
&= \left(\int_D g(\mathbf{z}) e^{rf(\mathbf{z})} \mathrm{d}\mathbf{z} + \int_{S' \setminus D} g(\mathbf{z}) e^{rf(\mathbf{z})} \mathrm{d}\mathbf{z}\right) \left(1 + O\left(\frac{1}{r}\right)\right) \\
&\quad \text{(By Theorem 6.2)} \\
&= \left(\left(\frac{2\pi}{r}\right)^{\frac{n}{2}} \frac{1}{\sqrt{-det\, Hess\, f(\mathbf{z}_0)}} g(\mathbf{z}_0) e^{rf(\mathbf{z}_0)} + O(e^{r \sup_{N'} \mathrm{Re} f}\right) \left(1 + O\left(\frac{1}{r}\right)\right) \\
&= \left(\frac{2\pi}{r}\right)^{\frac{n}{2}} \frac{1}{\sqrt{-det\, Hess\, f(\mathbf{z}_0)}} g(\mathbf{z}_0) e^{rf(\mathbf{z}_0)} \left(1 + O\left(\frac{1}{r}\right)\right)
\end{aligned}$$

Where N' is a compact subset of N containing $S \setminus \bar{B}(\mathbf{z}_0, \delta)$. □
Use saddle point method, we can estimate each Fourier coefficient $\hat{f}(a,b)$.

Notation 6.6. *Define* $V(x,y) = -px^2 + 4xy - Li_2\left(e^{-2i(y+x)}\right) + Li_2\left(e^{2i(y-x)}\right)$
Then according to Lemma 6.4, $V_r(x,y) = V(x,y) + O\left(\frac{1}{r^2}\right)$
Set $V_{a,b}(x,y) = V(x,y) - 4\pi ax - 4\pi bx$
Then,

$$\begin{aligned}
\hat{f}_r(a,b) =& (-1)^{a+b} (\frac{r}{2\pi})^2 \int_{\Omega_{\delta/2}} \psi(x,y) \sin x\, e^{-ix} e^{\frac{r}{4\pi i}(V_r(x,y) - 4\pi ax - 4\pi by)} \mathrm{d}x \mathrm{d}y \\
=& (-1)^{a+b} (\frac{r}{2\pi})^2 \int_{\Omega_\delta} \sin x\, e^{-ix} e^{\frac{r}{4\pi i}(V_r(x,y) - 4\pi ax - 4\pi by)} \mathrm{d}x \mathrm{d}y + O\left(e^{\frac{r}{4\pi}(Vol(M_p) - \varepsilon')}\right) \\
=& (-1)^{a+b} (\frac{r}{2\pi})^2 \int_{\Omega_\delta} \sin x\, e^{-ix} e^{\frac{r}{4\pi i}(V_{a,b}(x,y) + O\left(\frac{1}{r^2}\right)} \mathrm{d}x \mathrm{d}y + O\left(e^{\frac{r}{4\pi}(Vol(M_p) - \varepsilon')}\right)
\end{aligned}$$

To apply Saddle Point Method, we need to calculate the critical point and critical values, and to find an appropriate integration area.
The most surprising result is that the critical equation of $V_{a,b}$ is completely consistent with the hyperbolic gluing equation (4.3) for the Figure-8 knot with surgery coefficient p, yet this phenomenon still cannot be clearly explained.

Proposition 6.2. *If (x_0, y_0) is a critical point of $V_{a,b}$, then by a change of variables:*

$$1 - e^{2i(y_0 - x_0)} = \frac{1}{z}\ , \qquad 1 - e^{-2i(y_0 + x_0)} = \frac{1}{1-w}$$

And an appropriate choice of solution such that $(\mathrm{Re}x, \mathrm{Re}y) \in D_\delta$ and $\mathrm{Im}z, \mathrm{Im}w > 0$, then the critical equation for (x_0, y_0) is equivalent to the hyperbolic gluing equation (4.3) of figure-8 knot with coefficient $(p, 1)$.
Specifically:

(1) If (x_0, y_0) is an appropriate critical point of $V_{a,b}$
then z and w are the solution for $\begin{cases} z(z-1)w(w-1) = 1 \\ (\frac{1}{w(1-z)})^p(\frac{1}{w(1-w)})^2 = 1 \end{cases}$

(2) If z and w are the solution for $\begin{cases} z(z-1)w(w-1) = 1 \\ (\frac{1}{w(1-z)})^p(\frac{1}{w(1-w)})^2 = 1 \end{cases}$
then $(x_0, y_0) = \left(\frac{\log\frac{w-1}{w} - \log\frac{z-1}{z}}{4i}, \frac{\log\frac{w-1}{w} + \log\frac{z-1}{z}}{4i}\right)$ by choosing the main branch of logarithm (argument $\in (0, 2\pi)$), is a critical point for $V_{0,0}$.

Proof: (1) The critical equation is

$$\frac{\partial V_{a,b}}{\partial x} = -2px + 4y + 2i\log\left(1 - e^{2i(y-x)}\right) - 2i\log\left(1 - e^{-2i(y+x)}\right) - 4\pi a = 0$$
$$\frac{\partial V_{a,b}}{\partial y} = 4x - 2i\log\left(1 - e^{2i(y-x)}\right) - 2i\log\left(1 - e^{-2i(y+x)}\right) - 4\pi b$$

Notice that $e^{4iy_0} = \frac{w-1}{w}\frac{z-1}{z}$ and $e^{4ix_0} = \frac{w-1}{w}\frac{z}{z-1}$

$$\begin{aligned}
\frac{\partial V_{a,b}}{\partial y} = 0 &\Rightarrow \exp\left(i\frac{\partial V_{a,b}}{\partial y}\right) = 1 \\
&\Leftrightarrow z(z-1)w(w-1) = 1 \ \ \text{(Thurston's Equation)} \\
&\Rightarrow e^{2ix} = w(1-z) \\
\frac{\partial V_{a,b}}{\partial x} = 0 &\Rightarrow \exp\left(i\frac{\partial V_{a,b}}{\partial x}\right) = 1 \\
&\Leftrightarrow (w(1-z))^p\,\frac{w(w-1)}{z(z-1)} = 1 \\
&\Leftrightarrow \left(\frac{1}{w(1-z)}\right)^p\left(\frac{1}{w(w-1)}\right)^2 = 1 \ \text{(Dehn Filling Equation)}
\end{aligned}$$

(2) follows from (1) by checking the arguments. □

Now the critical point of V is related with the dihedral parameters of the hyperbolic geometrical structure of M_p, it is natural that the critical value is also relavent with the geometry.

Proposition 6.3. *The critical value of $V_{0,0}$ coincides with the hyperbolic volume of M_p, specifically:*

$$\mathrm{Im}V_{0,0}(x_0, y_0) = Vol(M_p)$$

Proof:

$$\begin{aligned}
\mathrm{Im}V_{0,0}(x_0,y_0) =&\mathrm{Im}\left(-px_0^2+4x_0y_0-Li_2\left(e^{-2i(y_0+x_0)}\right)+Li_2\left(e^{2i(y_0-x_0)}\right)\right)\\
&=D(\frac{z-1}{z})-D(\frac{w}{w-1})\\
&=D(\frac{z-1}{z})+D(\frac{w-1}{w})\\
&=Vol(\Delta_1)+Vol(\Delta_2)\\
&=Vol(M_p)
\end{aligned}$$

Δ_1 and Δ_2 are the two hyperbolic ideal tetrahedrons making up M_p under completion ,i.e. two hyperbolic ideal tetrahedrons satisfying the Hyperbolic Gluing Equation (4.3).
Notice that in the completion process, the limit points added into the manifold form a 1-dimensional sub-object which is the image of the "z-axis" under the covering map.
Thus, considered as a metric space (with local hyperbolic metrics), the limit points do not contribute to the hyperbolic volume.
Thus $Vol(M_p)=Vol(\Delta_1)+Vol(\Delta_2)$. □
In order to apply Saddle Point Method, we should have the integration area pass through the critical point.

Notation 6.7. *We define a new integration area for the Fourier coefficient $\hat{f}_r$, the new area is constructed by pushing the original D_δ towards an imaginary direction to pass the critical point.*

$$\begin{aligned}
&S_{top}=\{(x+i\mathrm{Im}x_0,y+i\mathrm{Im}y_0)\mid(x,y)\in D_\delta\}\\
&S_{side}=\{(x+ti\mathrm{Im}x_0,y+ti\mathrm{Im}y_0)\mid(x,y)\in\partial D_\delta,t\in[0,1]\}\\
&S=S_{top}\cup S_{side},\ \ \partial S=\partial D_\delta
\end{aligned}$$

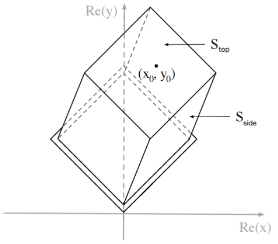


Figure 19: Deformed Integration Area [10]

Lemma 6.7. *(1)* $\mathrm{Im}V(x,y)$ *is strictly concave on* S_{top}

(2) $\mathrm{Im}V(x,y)$ *is strictly convex on* S_{side}

(3) $det\,Hess\,V(x_0,y_0)\neq 0$

Proof: By direct calculation:

$$Hess_{(\mathrm{Re}(x),\mathrm{Re}(y))}(\mathrm{Im}V) = \begin{pmatrix} -\frac{4\mathrm{Im}e^{2i(y+x)}}{|1-e^{2i(y+x)}|^2} - \frac{4\mathrm{Im}e^{2i(y-x)}}{|1-e^{2i(y-x)}|^2} & -\frac{4\mathrm{Im}e^{2i(y+x)}}{|1-e^{2i(y+x)}|^2} + \frac{4\mathrm{Im}e^{2i(y-x)}}{|1-e^{2i(y-x)}|^2} \\ -\frac{4\mathrm{Im}e^{2i(y+x)}}{|1-e^{2i(y+x)}|^2} + \frac{4\mathrm{Im}e^{2i(y-x)}}{|1-e^{2i(y-x)}|^2} & -\frac{4\mathrm{Im}e^{2i(y+x)}}{|1-e^{2i(y+x)}|^2} - \frac{4\mathrm{Im}e^{2i(y-x)}}{|1-e^{2i(y-x)}|^2} \end{pmatrix}$$
$$= -\begin{pmatrix} 2 & -2 \\ 2 & 2 \end{pmatrix} \begin{pmatrix} \frac{\mathrm{Im}e^{2i(y+x)}}{|1-e^{2i(y+x)}|^2} & 0 \\ 0 & \frac{\mathrm{Im}e^{2i(y-x)}}{|1-e^{2i(y-x)}|^2} \end{pmatrix} \begin{pmatrix} 2 & 2 \\ -2 & 2 \end{pmatrix}$$

When $(\mathrm{Re}(x), \mathrm{Re}(y)) \in \Omega_0$, $\frac{\mathrm{Im}e^{2i(y+x)}}{|1-e^{2i(y+x)}|^2} > 0$, $\frac{\mathrm{Im}e^{2i(y-x)}}{|1-e^{2i(y-x)}|^2} > 0$.
Thus $\mathrm{Im}V$ is strictly concave about $(\mathrm{Re}(x), \mathrm{Re}(y))$.
Furthermore, because $\mathrm{Im}V$ is a harmonic function, so $\mathrm{Im}V$ is strictly convex about $(\mathrm{Im}(x), \mathrm{Im}(y))$.
Especially, S_{top} only concerns the $(\mathrm{Re}(x), \mathrm{Re}(y))$ part, and S_{side} only concerns the $(\mathrm{Im}(x), \mathrm{Im}(y))$-part.
Hence (1) and (2) holds. And (3) is obvious due to the above calculation. □

Corollary 6.5.

$$\left| \int_{D_\delta} \sin x \, e^{-ix} e^{\frac{r}{4\pi i}(V_r(x,y))} \mathrm{d}x\mathrm{d}y \right| = P(r) e^{\frac{r}{4\pi} Vol(M_p)} \left(1 + O\left(\frac{1}{r}\right)\right)$$

$P(r)$ is a rational function of r.

Proof: Because $\sin x \, e^{-ix} e^{\frac{r}{4\pi i}(V(x,y) - 4\pi a x - 4\pi b y)}$ is a holomorphic function for $(\mathrm{Re}x, \mathrm{Re}y) \in \Omega_\delta$, by Cauchy's Theorem:

$$\left| \int_{D_\delta} \sin x \, e^{-ix} e^{\frac{r}{4\pi i}(V_r(x,y))} \mathrm{d}x\mathrm{d}y \right| = \left| \int_{S} \sin x \, e^{-ix} e^{\frac{r}{4\pi i}(V_r(x,y))} \mathrm{d}x\mathrm{d}y \right|$$

Lemma 6.7 (1) and (2) guarantess that $\mathrm{Im}V(x_0, y_0)$ is a maximal point on S.
Lemma 6.7 (3) guarantess that the Hessian matrix is non-singular.
Obviously, $sin(x_0)e^{-ix_0} \neq 0$.
By Notation 6.6, $V_r(x,y) = V(x,y) + O\left(\frac{1}{r^2}\right)$.
Hence, all conditions are satisfied for applying Saddle Point Method (Theorem 6.3).

$$\left| \int_{S} \sin x \, e^{-ix} e^{\frac{r}{4\pi i}(V_r(x,y))} \mathrm{d}x\mathrm{d}y \right| = P(r) e^{\frac{r}{4\pi} Vol(M_p)} \left(1 + O\left(\frac{1}{r}\right)\right)$$

$P(r)$ is a rational function of r. □

Lemma 6.8 (Functional property of Li_2)**.** *For $(\mathrm{Re}x, \mathrm{Re}y) \in \Omega_0$:*

$$-Li_2\left(e^{-2i(y\pm x)}\right) = Li_2\left(e^{2i(y\pm x)}\right) + \frac{\pi^2}{6} + \frac{1}{2}\left(\log\left(-e^{2i(y\pm x)}\right)\right)^2$$
$$= Li_2\left(e^{2i(y+x)}\right) + \frac{\pi^2}{6} - \frac{1}{2}\left(2(y \pm x) - \pi\right)^2$$

Proof: Consider $F(z) = Li_2(z) + Li_2(z^{-1}) + \frac{\pi^2}{6} + \frac{1}{2}\left(\log(-z)\right)^2$.
Then $F'(z) = \frac{1}{z}\left(\log\left(1-\frac{1}{z}\right) - log(1-z) + log(-z)\right) = 0$.
And $F(1) = 2Li_2(1) + \frac{\pi^2}{6} + \frac{1}{2}\left(\log(-z)\right)^2 = 2\frac{\pi^2}{6} + \frac{\pi^2}{6} - \frac{1}{2}\pi^2 = 0$.
Hence $F \equiv 0$, so the first equation holds.
The second equation is derived from the first equation by checking the arguments. □

Corollary 6.6.

$$\hat{f}_r(0,0) = \hat{f}_r(1,0)(1 + O\left(e^{\frac{1}{r}}\right))$$

Proof: Change the variable from x to $-x$, and this is direct from Lemma 6.8. □

Proposition 6.4.

$$\left|\int_{D'_\delta} \sin x\, e^{-ix} e^{\frac{r}{4\pi i}(V_r(x,y) - 4\pi ax - 4\pi by)} \mathrm{d}x\mathrm{d}y\right| = O\left(e^{\frac{r}{4\pi}(Vol(M_p) - \overline{\varepsilon})}\right) \ \forall (a,b)$$

$$\left|\int_{D''_\delta} \sin x\, e^{-ix} e^{\frac{r}{4\pi i}(V_r(x,y) - 4\pi ax - 4\pi by)} \mathrm{d}x\mathrm{d}y\right| = O\left(e^{\frac{r}{4\pi}(Vol(M_p) - \overline{\varepsilon})}\right) \ \forall (a,b)$$

$$\left|\int_{D_\delta} \sin x\, e^{-ix} e^{\frac{r}{4\pi i}(V_r(x,y) - 4\pi ax - 4\pi by)} \mathrm{d}x\mathrm{d}y\right| = O\left(e^{\frac{r}{4\pi}(Vol(M_p) - \overline{\varepsilon})}\right) \ (a,b) \neq (0,0), (0,1)$$

Proof: By completely the same method, Ohtsuki [6] calculated that the relavent critical values are smaller for $(a,b) \neq (0,0), (0,1)$ or D_δ is replaced by ${D_\delta}', {D_\delta}''$.
Hence the result is given by Saddle Point Method (Theorem 6.3). □

Proof of Theorem 6.1:
Because the Fourier transformation on the Schwarz space converges locally uniformly, and a polynomial order of sum of lower exponential terms provide lower exponential terms.
Hence,

$$\begin{aligned}
&\lim_{\substack{r\to+\infty \\ rodd}} \frac{4\pi}{r} \log\left|RT_r(M_p)\right| \\
=&\lim_{\substack{r\to+\infty \\ rodd}} \frac{4\pi}{r} \log\left|{c_r}'' \sum_{(a,b)\in\mathbb{Z}^2} \hat{f}_r(a,b) + O(e^{\frac{r}{4\pi}(Vol(M_p)-\varepsilon)})\right| \\
=&\lim_{\substack{r\to+\infty \\ rodd}} \frac{4\pi}{r} \log\left|R(r) e^{\frac{r}{4\pi}Vol(M_p)}\left(1 + O\left(\frac{1}{r}\right)\right)\right| \\
=&Vol(M_p)
\end{aligned}$$

Q.E.D.

Bibliography

[1] C. Blanchet, N. Habegger, G. Masbaum, and P. Vogel. Three-manifold invariants derived from the kauffman bracket. *Topology*, 30(685-699), 1992.

[2] Qingtao Chen and Tian Yang. A volume conjecture for a family of turaev-viro type invariants of 3-manifolds with boundary. *arXiv:1503.02547*, 2015.

[3] Vadim Kaplunovsky. Notes for quantum field theory: Saddle point method. 2011.

[4] R. C. Kirby. A calculus for framed links in S^3. *Invent. Math.*, 45, 1978.

[5] Gregor Masbaum. Skein-theoretical derivation of some formulas of habiro. *Algebraic and Geometric Topology*, 2013.

[6] Tomotada Ohtsuki. On the asymptotic expansion of the quantum SU(2) invariant at $q = exp(4\pi\sqrt{-1}/N)$ for closed hyperbolic 3-manifolds obtained by integral surgery along the figure-eight knot. *Algebraic and Geometric Topology*, 2018.

[7] Nicolai Reshetikhin and Vladimir Turaev. Invariants of 3-manifolds via link polynomials and quantum groups. *Inventiones Mathematicae*, 1991.

[8] D. Rolfson. *Knots and Links.* American Mathematical Society, 2003 ed.

[9] W. P. Thurston. The geometry and topology of three manifolds. 1979.

[10] Ka Ho Wong and Tian Yang. On the volume conjecture for hyperbolic dehn-filled 3-manifolds along the figure-eight knot. *arXiv:2003.10053*, 2020.

McShane's Identity and Integration over Moduli Space

Qiu Ruicen, Yin Shun

May 2022

Contents

1 Abstract

We'll introduce an identity on a hyperbolic once punctured torus given by summing a function of the length over all closed simple geodesics. McShane discovered the constant in his Doctoral thesis and later gave a generalization about it and Mirzakhani generalize the identity to apply it to calculate the Weil-Petersson volume of the moduli space of curves with marked points. We will introduce the proof of the identities basically following McShane and Mirzakhani and then study the integration over moduli space. Ultimately we give a few examples to calculate their volumes to illustrate in general case how to operate.

2 Introduction

In the celebrated paper by Mirzakhani, a generalized McShane's identity is obtained and applied to calculate the volumes of moduli space. That is an amazing method that provide an effective way to calculate the volume of all moduli spaces of hyperbolic surfaces with marked points without finding the fundamental domain which is much more difficult.

That paper is divided to two parts. One is to generalize McShane's amazing identity from punctured hyperbolic surface to hyperbolic surface with geodesic boundaries which has finite area. The other is to give a method to integrate over moduli space of curves with marked points. Then combining the two aspects an algorithm can be attained to calculate the volumes by not very complicate integral.

Our article is a reading report of Mirzakhani's paper and we also get some ideas from the papers of McShane's. It also consist of the two parts and at last we calculate some examples to illustrate how to combine the two aspects.

Remark 2.0.1 (Remark of the McShane's identity). There are some important factors in the proof of the identities. The behavior of geodesics near geodesic boundaries is just like a Cantor set with isolated points. The first is the Birman-Series theorem. Then the proof of all forms of the identities consist of the structure theorem of the simple geodesics to classify them in the above intersection. Then we calculate the lengths of the gaps to get the formula.

On the structure of this article, in §3 we state the notations and review the basic facts in hyperbolic geometry and have an intuitive impression about a minimal lamination. It's crucial to realize the structure of pants and the geodesics on it. In §4 we state 3 versions of McShane's identities and prove the main theorem about the structure of the geodesics then obtain the identity. In §5 we recall the concepts about the Teichmüller space and introduce forms on some space which have close relations to our problem. In §6 we give an intuitive explanation of the formula of some functions integrating over moduli space. In §7 we combine the generalized McShane's identity and the integration formula to calculate the basic examples.

3 Preliminaries

To prove the identity, first we should study the structure of a hyperbolic punctured torus, especially the simple geodesics on it.

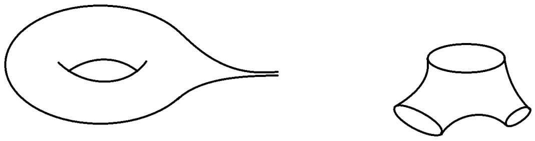
3.1 hyperbolic surface

Definition 3.1. A *hyperbolic surface* M is a surface admitting a hyperbolic structure, which is to say, M has an atlas $\mathcal{A} = \{(U, \phi_U : U \to \mathbb{H}^2)\}$ so that the transition map $\phi_U \phi_V^{-1}$ is isometry on $\phi_V(U \cap V)$.

Remark 3.1.1. a hyperbolic surface has universal covering $\mathbb{H}^2$ since we can construct it explicitly. So it admits a canonical complex structure. So a hyperbolic surface is exactly a Riemann surface with uniformization $\mathbb{H}^2$.

Remark 3.1.2. We sometimes also treat a hyperbolic surface with geodesic boundary as a hyperbolic surface.

Remark 3.1.3. We call M is complete if it is complete as a Riemannian manifold.



Punctured torus on the left. Pair of pants on the right.

Definition 3.2. A *cusp* region of a hyperbolic surface M is a subsurface P of M isometric to the quotient $\mathbb{H}^2 \supset \{z : \mathrm{Im} z > 1\}/[z \mapsto z + l]$, $l \in \mathbb{R}_+$ is the length of the *horocycle* (boundary of the cusp region).

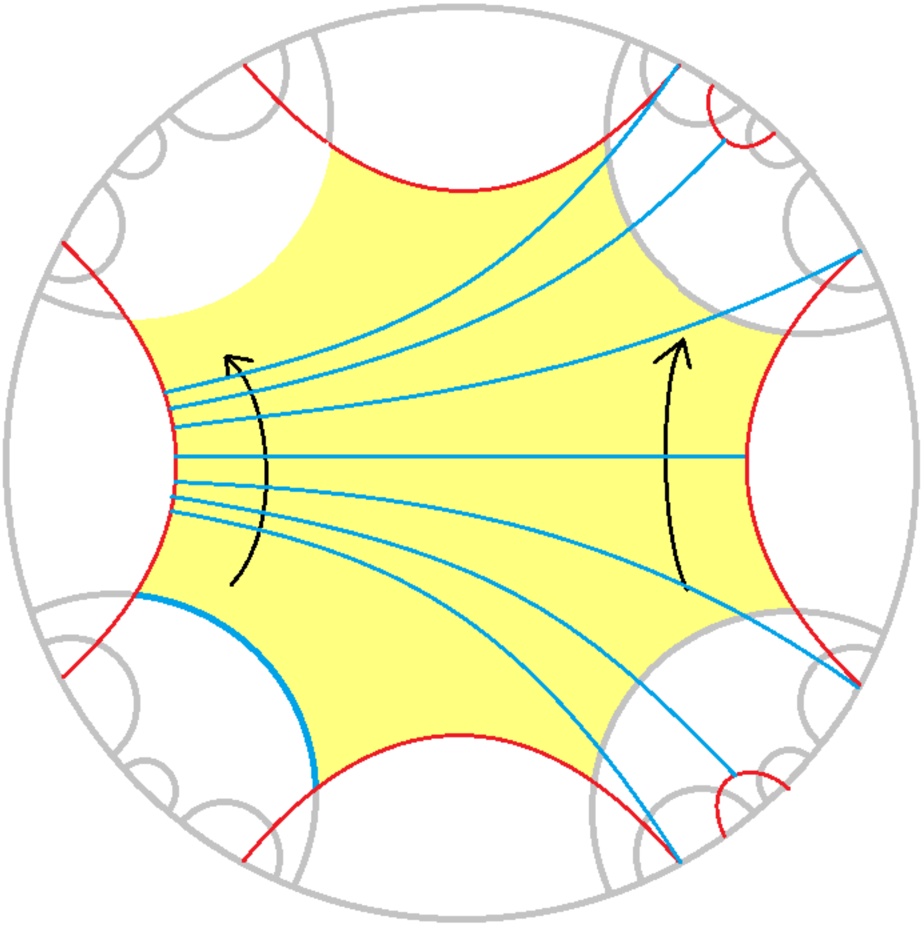
3.2 pair of pants

Pants decomposition is powerful in the classification of hyperbolic surfaces. A closed hyperbolic surface of genus g has at most $3g-3$ disjoint simple geodesics. And They cut the surface into $2g-2$ pair of pants.

Definition 3.3. A hyperbolic *pair of pants* is a hyperbolic surface which is topologically homeomorphic to a 3 punctured sphere with 3 geodesic boundaries of length $|\alpha|, |\beta|, |\gamma|$.

Remark 3.3.1. Such pants is unique up to isometry. In fact, it can be uniquely decomposed into 2 identical hyperbolic right-angle hexagon with 3 edges known, which is unique by easy calculation.

Remark 3.3.2. when $|\alpha| = 0$, such boundary becomes a cusp. We also regard these surface as pair of pants with some boundary length 0.

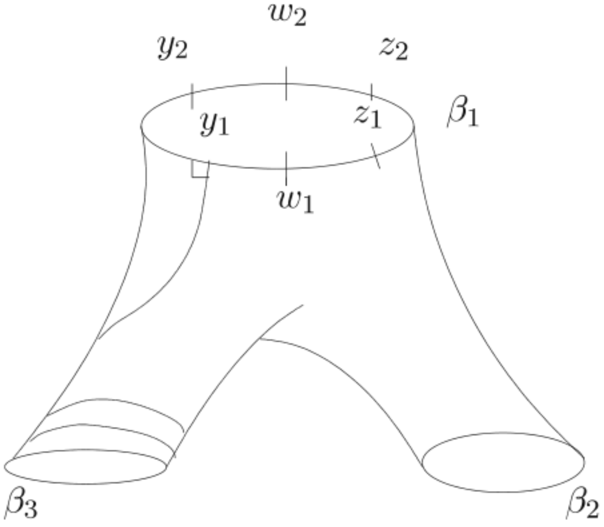


Remark 3.3.3. Only finitely many geodesics containing in a hyperbolic pair of pants are simple. In particular, There are exactly 8 simple geodesics starting from one of the boundary components, which can be seen in the figure above.

Marking the three boundary with $\beta_1, \beta_2, \beta_3$ and their lengths are x_1, x_2, x_3. The geodesic perpendicular to β_1 twice intersect β_1 at point w_1, w_2. The two geodesics spiral to β_3 is perpendicular to β_1 at points y_1, y_2 respectively. The two geodesics spiral to β_2 is perpendicular to β_1 at points z_1, y_z respectively.

Geometry on pants. In [Mir07a], Mirzakhani introduce two length function playing crucial pole in the calculation of the volumes. As the picture showed.

Define $\mathcal{D}(x_1, x_2, x_3)$ as the geodesic length of the interval (y_1, y_2) which contains the points w_1, w_2. Let $\mathcal{R}(x_1, x_2, x_3)$ denote the length of the sum of two interval (y_1, z_1) and (y_2, z_2) with equal length. It is easy to see $x_1 + \mathcal{R}(x_1, x_2, x_3) = \mathcal{D}(x_1, x_2, x_3) + \mathcal{D}(x_1, x_3, x_2)$.



Actually, from hyperbolic geometry on the universal covering, the explicit formula can get:

$$\mathcal{D}(x, y, z) = 2\log\left(\frac{e^{\frac{x}{2}} + e^{\frac{y+z}{2}}}{e^{-\frac{x}{2}} + e^{\frac{y+z}{2}}}\right).$$

$$\mathcal{R}(x, y, z) = x - \log\left(\frac{\cosh(\frac{y}{2}) + \cosh(\frac{x-z}{2})}{\cosh(\frac{y}{2}) + \cosh(\frac{x+z}{2})}\right).$$

with derivatives $\frac{\partial}{\partial x}\mathcal{D}(x, y, z) = H(y + z, x), \frac{\partial}{\partial x}\mathcal{R}(x, y, z) = \frac{1}{2}(H(z, x + y) + H(z, x - y))$ where

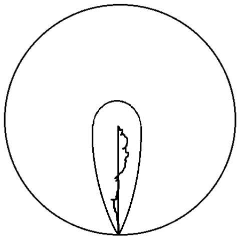
$$H(x, y) = \frac{1}{1 + e^{\frac{x+y}{2}}} + \frac{1}{1 + e^{\frac{x-y}{2}}}.$$

3.3 quasi-geodesic

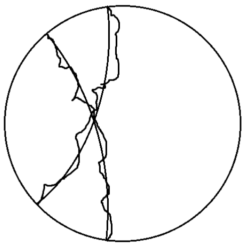
Definition 3.4. A (λ, c) *quasi-geodesic* from x to y in complete Riemannian manifold M is a path $\gamma : [0, t] \to M$, parameterized by arclength, such that for any $u, v \in [0, t]$,

$$\frac{1}{\lambda}d(\gamma(u), \gamma(v)) - c \leq |u - v| \leq \lambda d(\gamma(u), \gamma(v)) + c.$$

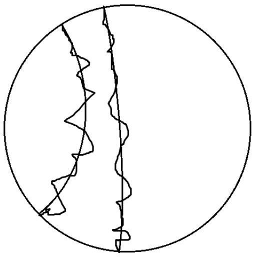
Remark 3.4.1. In hyperbolic space $\mathbb{H}^n$, a lemma of Morse indicates that a quasi geodesic cannot goes too far away from a geodesic. There's a constant $D = D(\lambda, c)$ such that a (λ, c) quasi-geodesic from x to y is contained in the D-neighborhood of the geodesic $[x, y]$.



D-neighbourhood of
a geodesic



not simple



simple quasi-geodesic
=> simple geodesic

Therefore, we can prove that a quasi-geodesic ray (has infinite length and locally AC) in $\mathbb{H}^n$ must converge into a unique point of $\partial\mathbb{H}^n$. As a corollary, a closed piecewise geodesic curve (not contractable or peripheral) γ in a hyperbolic surface M lift to some quasi-geodesic lines in $\mathbb{H}^2$. Each component corresponds to an unique geodesic, which is invariant under the action of $[\gamma] \in \pi_1(M)$. So it project to a closed geodesic $\tilde{\gamma}$. So we have the lemma

Lemma 3.5. *A closed piecewise geodesic curve (not contractable or peripheral) γ on a hyperbolic surface is free homotopic to a unique closed geodesic $\tilde{\gamma}$. Furthermore, if γ is simple, then $\tilde{\gamma}$ is also simple.*

3.4 geodesic lamination

To study simple geodesics on a hyperbolic plane, Thurston introduced the conception of *lamination* in [Thu] and its structure can be classified into several classes.

Definition 3.6. A *geodesic lamination* λ on a hyperbolic surface M is the union of some disjoint complete simple geodesics, which is closed as a subset of M.

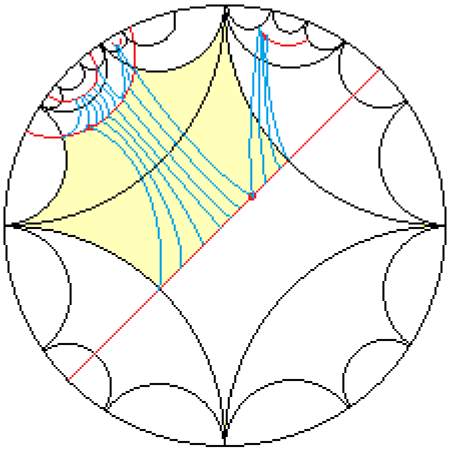
Remark 3.6.1. For convenience, we will use lamination instead of geodesic lamination on this article, and our lamination is always on a hyperbolic surface. A minimal lamination is a lamination which has no nontrivial sublamination. It is known that

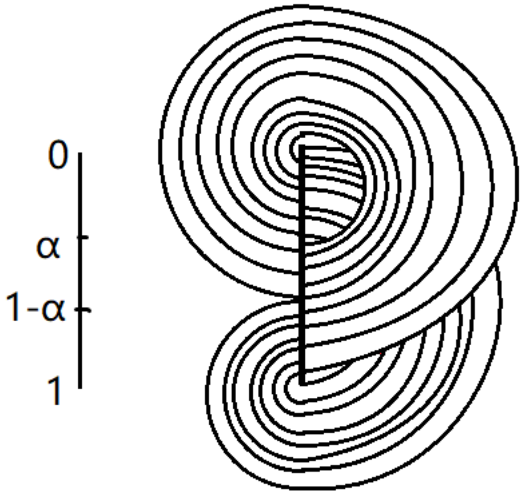
Lemma 3.7. *a minimal lamination λ is either*

1. *a complete simple geodesic or*
2. *has uncountably many leaves(geodesics) and for any transverse segment τ, $\tau \cap \lambda$ is a Cantor set.*

Example 3.8. A minimal lamination which is not a simple geodesic.

First, we construct a series of closed simple curve Γ. As in the picture, choose a hexagonal fundamental domain D of a standard punctured torus. Red edges are the lift of a simple closed geodesic c. Assume $|c| = 1$. Let α be a irrational number, $x + \alpha$ be the translation of length α along c. γ_x be a simple geodesic other than c from $x \in c$ to $x + \alpha$. Then γ_x can be moved continuously to γ_y for any $y \in [x, x+1) \in c(\gamma_y$ may not be continuous at $x)$. Then $\Gamma = \{\gamma_y : y \in c\}$ is what we wanted. (blue segments on the picture)





Then Γ lift to a series of complete quasi-geodesic $\Gamma' = \{\gamma'_y\}$ crossing D in $\mathbb{H}^2$. Each γ'_y represents an unique geodesic μ_y in $\mathbb{H}^2$. Since γ'_y's are simple and disjoint, we see μ_y's are also simple and disjoint. It's easy to see the closure of each μ_y is exactly the union of all the μ_y's. Thus this is a nontrivial minimal lamination.

Remark 3.8.1. I constructed this example by referring [Bon98]. He introduced a more general method to construct a family of nontrivial laminations using interval exchange maps.

4 Generalized McShane's Identity

In this section, we consider the general case when M is a hyperbolic surface with geodesic boundaries $\beta_1, \cdots, \beta_n$ of length $l_1, \cdots, l_n$ which has finite area. If $l_i = 0$, then β_i represents a cusp.

4.1 3 versions of Mcshane's identity

The first identity was introduced by McShane in his doctoral thesis [Mcs91].

Theorem 4.1. *Let M be a once punctured torus then*

$$\sum_{\gamma} \frac{1}{1 + \exp|\gamma|} = \frac{1}{2},$$

where the sum is over all closed simple geodesics γ on M.

And the second identity was generalized by McShane himself in [Mcs98].

Theorem 4.2. *Let M be hyperbolic surface which has finite area with a cusp and without boundary,then*

$$\sum \frac{1}{1+\exp\frac{1}{2}(|\alpha|+|\gamma|)} = \frac{1}{2},$$

where the sum is over all pairs of closed simple geodesics α, γ which bound an embedded pair of pants containing the cusp point.

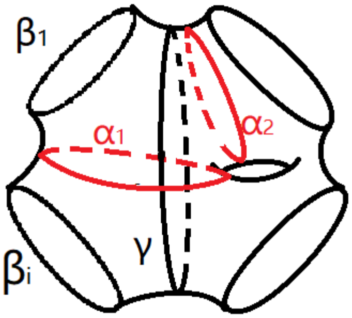
Mirzakhani further generalized the identity in general hyperbolic surface.

And introduced the viewpoint of coarse geometry to explain and generalize the structure theorem of McShane in [Mir07a]. Based on the boarder structure theorem, she gave the further generalized version of Mcshane's identity

Theorem 4.3. *For any hyperbolic surface M with n geodesic boundary components $\beta_1, \cdots, \beta_n$ of lengths $L_1, \cdots, L_n$, we have*

$$\sum_{\{\alpha_1,\alpha_2\}\in\mathcal{F}_1} \mathcal{D}(L_1, l_{\alpha_1}(M), l_{\alpha_2}(M)) + \sum_{i=2}^{n}\sum_{\gamma\in\mathcal{F}_{1,i}} \mathcal{R}(L_1, L_i, l_\gamma(M)) = L_1.$$

where $\mathcal{F}_1$ consists unordered pairs of simple closed geodesics (possibly same) which bound a pair of pants with β_1; $\mathcal{F}_{1,i}$ consists simple closed geodesics which bounds a pair of pants with β_1, β_i. And for a geodesic α on hyperbolic surface M, function $\ell_\alpha(M)$ denote the hyperbolic length of α on M. Notice that because $\mathcal{D}(x, y, z)$ is symmetric on y, z the left side of above identity is well-defined.



To prove the identity, first we will study simple geodesics which intersects the boundary β_i perpendicularly (If β_i represents a cusp, then it should intersect the sufficiently small cusp region).

4.2 simple geodesics

4.2.1 case of only one cusp

This case is studied by Mcshane in [Mcs98].

Remark 4.3.1. In our context, simple geodesic in hyperbolic surface with geodesic boundary is a geodesic with both end either complete or intersects boundary component perpendicularly and also simple.

Simple geodesics was classified in [CEG87]. However, here we have a slightly different version.

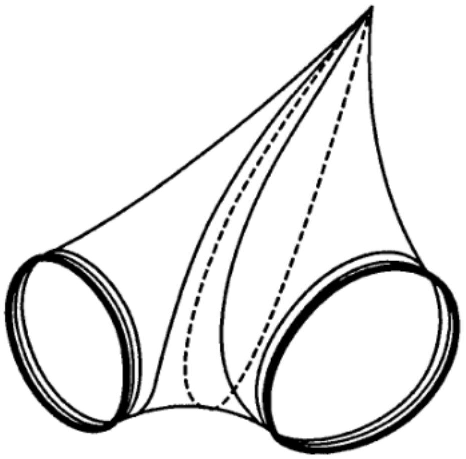
Lemma 4.4. *Complete simple geodesics has 3 types*

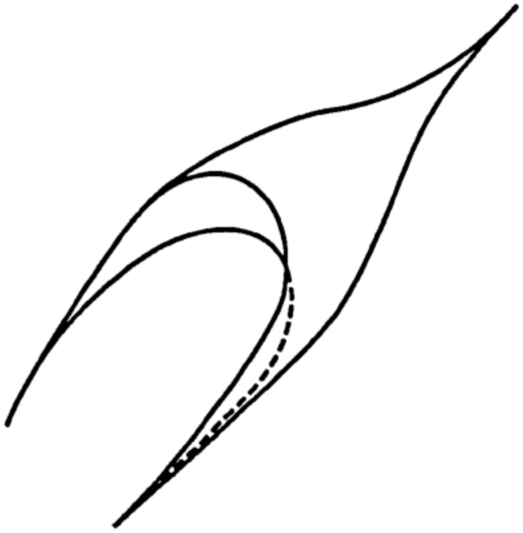
1. *it has both end up a cusp or intersects boundary;*
2. *it has one end up a cusp or intersects boundary and another end spiraling to a compact lamination;*
3. *it is a leaf of a compact lamination.*

Recall a compact lamination has 2 types (3.7), simple geodesics which has one end up the cusp or intersects boundary thus has 3 types

Lemma 4.5. *Complete simple geodesics which intersects the cusp region has 3 types: the other end of the geodesic*

1. *up a cusp or intersects boundary;*
2. *spirals to a closed geodesic inside the surface;*
3. *spirals to a minimal lamination consisting of uncountably many leaves;*
4. *spirals to a closed geodesic on the boundary.*

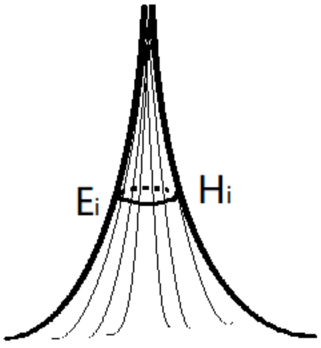




On the left is a simple geodesic with one end spiralling to a simple geodesic.On the right is a punctured torus cutting a nontrivial minimal lamination, which has two geodesic boundary component of length infinity.

Remark 4.5.1. In this article, we call a geodesic is of type 1/2/3/4 if it falls into the 1/2/3/4 class above.

From *Lemma 4.3.1*, we can choose a small cusp region so that a complete simple geodesics which intersect the cusp region must have one end up a cusp. Then it has the classification above. We'll try to describe the structure of these geodesics.



On such a horocycle(boundary of the cusp region) H_i, we define a set $E_i \subset H_i$ to be all the points which a simple geodesic go through. Thus each point of E_i exactly correspond to a simple geodesic with one end up to the cusp. It is actually a Cantor set unions countably many isolated points. By a theorem of Birman and Series, E_i has measure zero.

Theorem 4.6 (Birman Series)**.** *Let G be the set of all simple geodesics on a hyperbolic surface. The set S of points which lie on a geodesic $\gamma \in G$ has Hausdorff dimension 1.*

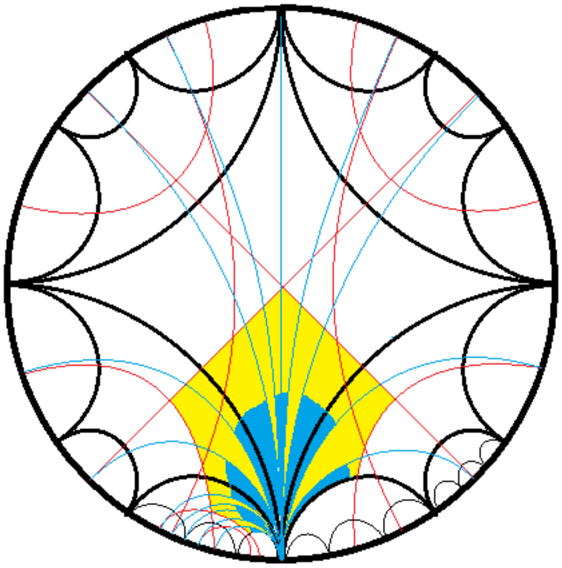
Points in $H_i - E_i$ represents a geodesic which intersects itself, thus transversely intersects itself. So a small perturbation preserve the property. Therefore $H_i - E_i$ is open, and thus decomposed into countably many open intervals, called *gap*.

Now we can state our structure theorem of the case of cusp.

Theorem 4.7 (Structure Theorem of cusp)**.**

1. *$x \in E_i$ is isolated if and only if x is of type 1 or 4;*
2. *x is boundary of a gap and not isolated if and only if x is of type 2;*
3. *x is not a boundary of a gap if and only x is of type 3;*

Furthermore, every gap has one endpoint of type 1 and another endpoint of type 2 or 4.



Remark 4.7.1. Here is an example of the structure of the geodesic when M is a punctured torus. The yellow region is a fundamental domain and the blue region are some gaps. Red lines are simple geodesics.

Each gap is beside an isolated point in E_i, that is, a geodesic of type 1. And each isolated point of E_i besides 2 gaps.

Therefore, the area of the cusp region can be calculated by summing the length of all the gaps, or summing the length of gaps beside all the geodesics of type 1.

Our next proposition shows that each isolated point corresponds to a pants with one of the boundaries a cusp, which let us convert the length of gaps into the length of simple closed geodesics.

Proposition 4.8. *Let M be a surface with at least one cusp and let γ be a geodesic with both ends up the cusp then γ is contained in a unique embedded pair of pants.*

Remark 4.8.1. It's easy to see that for each pair of pants with one of the boundaries a cusp, there's exactly one geodesic of type 1 contained in it. So our calculation may over all pants of this kind.

4.2.2 geodesic boundary case

This case is basically same as the previous one, which has studied simple geodesics in general hyperbolic surface. It's somehow much easier than cusp case.

We consider simple geodesics γ_x which intersects boundary component β_i at x perpendicularly. Let H_i just be β_i and E_i be the points x in H_i which the geodesic γ_x perpendicularly start at is simple, and if γ_x intersects boundary components at the other end, then it should intersect perpendicularly. It has the similar structure theorem

Theorem 4.9 (Structure Theorem)**.**

1. *$x \in E_i$ is isolated if and only if the other end of γ_x intersects or spirals to a boundary component (No matter β_i itself or not);*

2. *x is boundary of a gap and not isolated if and only if the other end of γ_x spirals to a non-boundary closed geodesic;*

3. *x is not a boundary of a gap if and only the other end of γ_x spirals to a nontrivial minimal lamination (uncountably many leaves);*

Furthermore, every gap has one endpoint of the first type and another endpoint of the second type.

Remark 4.9.1. We also call the points of E_i or geodesics of type 1/2/3 if the geodesic satisfies the 1/2/3 conditions respectively.

In this case, the proposition which let us convert the length of gaps into the length of closed geodesics should split into 2 parts

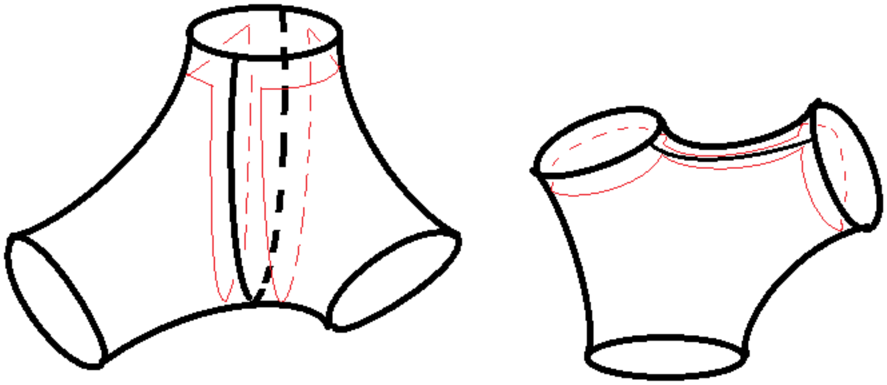
Proposition 4.10. *Let M be a surface with at least one geodesic boundary and let γ be a geodesic with*

1. *both ends intersect boundary component β_i perpendicularly; or*

2. *two ends intersect different boundary component β_i, β_j perpendicularly;*

then γ is contained in a unique embedded pair of pants.

Remark 4.10.1. In the proposition, cusp can be regarded as a geodesic boundary of length 0, and the statement of intersect the geodesic boundary perpendicularly should be translated into up the cusp. Thus it's a generalized version of 4.8.



Proof. In case 1, as shown in the left figure, γ_x divide the boundary component β_i into 2 parts. each part with γ_x form a simple closed curve, which is obviously a quasi-geodesic. So it is free homotopic to a simple closed geodesic.

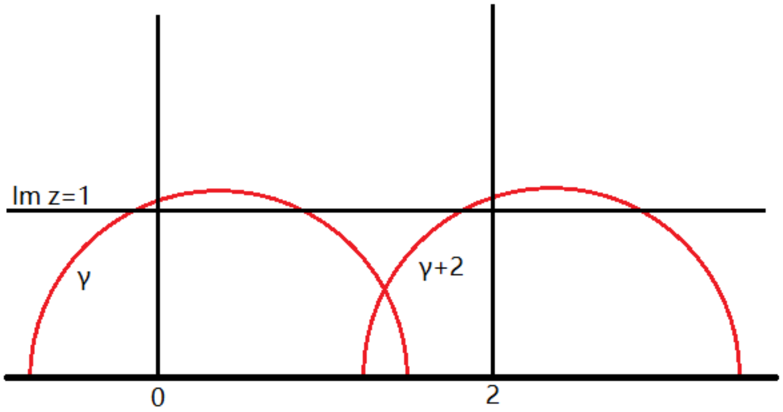
The 2 quasi-geodesics are disjoint with γ_x, β_i, so the corresponding 2 geodesics are disjoint with them. So from the homotopy type, the subsurface bound inside the 3 closed geodesic is a pair of pants P containing β_i, which is what we required.

In case 2, it's similar. we can also construct such a simple closed quasi-geodesic as shown in the right figure. □

4.3 structure theorem

We will study the 3 types of geodesics respectively. First we should complete our classification of simple geodesics crossing cusp region.

Lemma 4.11. *Let M be a surface with a cusp; such a surface has a cusp region of area 2. The portion of a complete simple geodesic lying in a cusp region of area less than 2 always meets the horocyclic foliation perpendicularly.*



Proof. Let $H \cong \{z : \text{Im}z > 1\}/[z \mapsto z+2]$ be the cusp region of area 2. It has a fundamental domain $D = \{x + yi : |x| < 1, y > 1\}$. If γ is a complete geodesic crossing H, then it can be lifted to $\tilde{\gamma}$, a geodesic in $\mathbb{H}^2$ crossing D.

If $\tilde{\gamma}$ has one end ∞, then it intersects ∂H perpendicularly.

Otherwise, as we showed in the figure above, $\tilde{\gamma}$ extend to a semicircle centered on the real axis with radius $r > 1$ (because it crosses D). $\tilde{\gamma}$ and $\tilde{\gamma} + 2$ intersects at some point in $\mathbb{H}^2$.

Thus γ will intersect itself at some point in the surface M. This is because if M is a hyperbolic surface without boundary, then our geodesics are complete. So it intersects itself. However, if M has boundary, we only consider the geodesics with both end complete or intersects boundary perpendicularly. Thus the claim also remains true. □

Remark 4.11.1. Now when we say a sufficiently small cusp region always has area less than 2, so that satisfies the condition above.

Now we will prove our main theorem. (definition of type is in 4.9.1)

Theorem 4.12 (Structure Theorem)**.**

1. $x \in E_i$ is isolated if and only if x is of type 1;

2. x is boundary of a gap and not isolated if and only if x is of type 2;

3. x is not a boundary of a gap if and only x is of type 3;

Furthermore, every gap has one endpoint of type 1 and another endpoint of type 2.

We only prove the case when the boundary component β_i doesn't represent a cusp.

4.3.1 type 1

Lemma 4.13 (type 1 part)**.** *If $x \in E_i$ is a point which lies on a geodesic γ_x of type 1, then x is an isolated point.*

Proof. If the other end perpendicularly intersects boundary, by *Proposition 4.10*, we can find a pants P containing the geodesic γ_x starts from x. A slight translation of γ_x to γ_x' let the other end deviate from right angle. So it is not a simple geodesic in our context (see 4.3.1).

If the other end up a cusp, then by *Proposition 4.8*, we can also find a pants P containing the geodesic γ_x across x. Then from the universal covering of pants, for any points $y \in H_i$ near x, γ_y is contained in the pants and intersects cusp region at the other end not perpendicularly. So by *Lemma 4.11* it can't be simple.

If the other end spirals to a boundary β_j, then by 4.10, there's a pants P bounding β_i and β_j. It's easy to see that little perturbation of γ_x to one side let it intersect β_j, while that to another side let it intersect itself on P. So x is also isolated. □

4.3.2 type 2

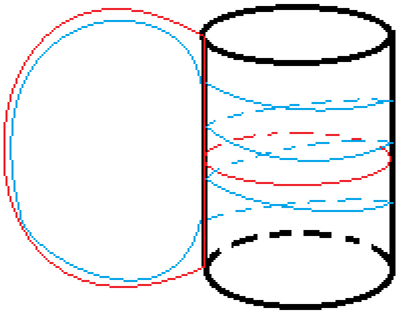
Lemma 4.14 (type 2 part)**.** *If $x \in E_i$ is a point which lies on a geodesic of type 2, then x is not an isolated point and one of the end of a gap.*

Proof. Let γ_x be the geodesic perpendicularly starting at x, γ_x spirals to a simple geodesic $\Omega = \Omega(\gamma_x)$.

Similar to the proof of *Proposition 4.10*, we know that Ω and β_i bounds a pair of pants. So one side of x in H_i is a gap.

Now we will construct a series of closed simple geodesics Ω_i, to which γ_{y_i} spirals, such that $\lim\limits_{n\to\infty} y_i = x$, whose existence can be also guranteed by the existence of pants bounding a pair of pants with β_i.

Case 1 When $M - \Omega$ has only 1 components, let u, v be two points close enough lies on the two sides of Ω. Minimal geodesic $\alpha_{u,v}$ between u, v can be enclosed into a simple closed quasi-geodesic and then homotopic to a simple geodesic α intersecting Ω at only 1 point z.



Now Let α_n (blue line) be a simple curve starts at u, goes along a n round twist $\phi_n(u,v)$ of geodesic $[u,v]$ to v and then goes along $\alpha_{u,v}$ back to u.

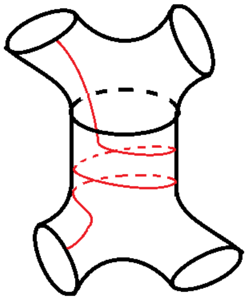
We now construct the twist ϕ_n. Choose a collar neighborhood $U \cong S^1 \times I$ of Ω. We might assume u, v is on $S^1 \times \{0\}, S^1 \times \{1\}$, respectively. Then $(x,y) \mapsto (x + 2\pi ny, y)$ conjugates to ϕ_n in U.

Since α_n is simple, it is free homotopic to a simple closed geodesic Ω_n, to which γ_{y_i} spirals. We will next prove that y_i converges to x.

Denote $[\alpha_n]$ to be the class of α_n in the fundamental group $\Gamma = \pi_1(M,u)$. Then Γ acts discretely on a convex subsurface N of $\mathbb{H}^2$, that is, the universal covering of M, with fundamental domain $D \ni \tilde{u}$, where $\tilde{u}$ is the base point of N. Then $[\alpha_n] = [\Omega]^n[\alpha_0]$. As a loxodromic element in $\mathrm{Aut}(\mathbb{H}^2)$, the fix point of $[\alpha_n]$ on $\partial\mathbb{H}^2$ converges to the fix points of $[\Omega]$. Thus the other end of γ_{y_n} in N, which is exactly the attracting fix point of $[\alpha_n]$, converges to the attracting fix point of $[\Omega]$, which means $y_i \to x$.

Case 2 If $M - \Omega$ has 2 components P_1, P_2, then assume β_i is in P_1.

Case 2.1 If P_2 contains a boundary component β_j, then consider the shortest path from β_i to β_j, denote it α. α intersects Ω at only 1 point. So twist it by ϕ_n as we did in case 1, and we obtain $\alpha_n = \phi_n(\alpha)$. α'_n be the geodesic free homotopic to α_n.



Denote $[\alpha_n] = [\alpha'_n]$ to be an loxodromic element in the fundamental group $\Gamma = \pi_1(M,u)$ acting on the universal covering $N \subset \mathbb{H}^2$ of M, such that $[\alpha'_n]$ send $y_i = \alpha'_n \cap \beta_i$ to $\alpha'_n \cap \beta_j$ and preserves the complete line α_n. Then $[\alpha_n] = [\Omega]^n[\alpha_0]$. So the attracting and repelling fixed point converges to the attracting and repelling fixed point of $[\Omega]$, respectively, which means $y_i \to x$.

Case 2.2 If P_2 contains no boundary components, since P_2 is not a disk, there's a simple geodesic α' inside P_2, which is not homotopic to Ω. Now let α be a geodesic starts from β_i, spiraling to α'. Then we can assume α has only one intersection with Ω.

Just like we did in **Case 2.1**, we can twist it to be $\alpha_n = \phi_n(\alpha)$ and canonically define an $[\alpha_n]$ as a loxodromic element in $\Gamma = \pi_1(M, u)$ acting on the universal covering $N \subset \mathbb{H}^2$ of M. We have $[\alpha_n] = [\Omega]^n[\alpha_0]$ and then $y_i \to x$. □

4.3.3 type 3

To prove the type 3 part, we need more delicate analysis to find out approximations $y_i \to x$ on two side of x. The example of a nontrivial lamination is given in 3.8. Here we quote a lemma from Mirzakhani ([Mir07a] *Lemma 4.7*).

Lemma 4.15. *Assume that*

$$c \cap \{\gamma_x(t) | 0 \leq t < t_2\} = \emptyset.$$

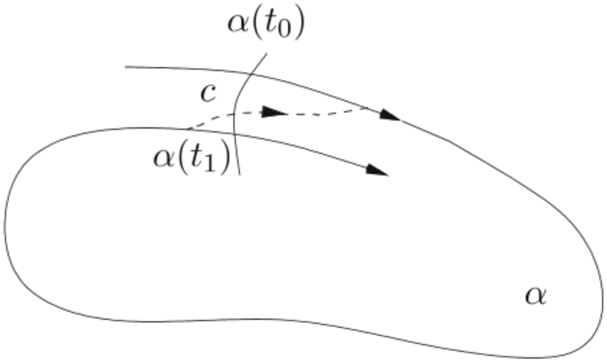
For any $\epsilon > 0$ there exist $\delta, L > 0$ such that if (α, t_0, t_1, c) is a δ-good geodesic segment and $L \leq t_1 - t_0$, then η is a simple quasi geodesic. Let $\hat{\eta}$ denote the geodesic representative of η and $y = \hat{\eta} \cap \beta_i$. Then

$$d(y, x) < \epsilon.$$

Furthermore, y lies on the right (left) side of x if and only if (α, t_0, t_1, c) is positive (negative).

Remark 4.15.1. Here γ_x is the parameter representation of geodesic ray starting from $x = \gamma_x(0)$. A δ-good geodesic (α, t_0, t_1, c) is to say that

1. $l(c) \leq \delta$
2. The angle between c and α at each endpoint is in $(\frac{\pi}{2} - \delta, \frac{\pi}{2} + \delta)$.
3. c meets the α in only two endpoints.



Such a δ-good geodesic can be obtained by the density of any leaf in the non-trivial lamination.

η is a piecewise geodesic curve starting at x, going along $\gamma_x(t)$ for $t \le t_2 = \inf\{t : \gamma_x(t) \in c\}$, ($\gamma_x(t_2)$ is the first intersection of c and γ_x), and spirals to the simple closed curve consisting of $\alpha([t_0, t_1])$ and c.

Now the y's constructed by this lemma instantly proved our last part of type 3 geodesics.

Lemma 4.16 (type 3)**.** *If $x \in E_i$ is a point which lies on a geodesic of type 3, then x is not a boundary of a gap, or that is to say, x can be approximated from both sides in E_i.*

Till now, we have already proved the structure theorem.

4.4 Proof of the Generalized McShane's Identity

Now we prove the Generalized McShane's Identity. At first, we describe the structure of E_i. We need the collar lemma to describe the behavior of geodesics near a closed geodesic.

Lemma 4.17 (Collar lemma)**.** *Let γ be a close simple geodesic on the hyperbolic surface S, then the collar neighborhood*

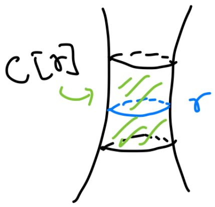
$$\mathcal{C}[\gamma] = \{p : dist(p, \gamma) \le w(\ell_\gamma(S))\}$$

is an embedded annulus, where

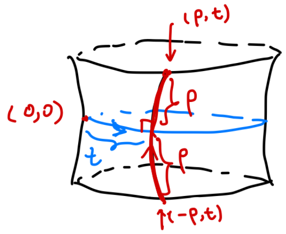
$$w(x) = \sinh^{-1}(\frac{1}{\sinh(x)})$$

is the width.

In the collar neighborhood every point has a unique geodesic passing it and perpendicular to γ, hence we can give it a coordinate describing by the oriented distance to γ and the perpendicular foot on the geodesic which called the *Fermi coordinate*, as the picture shows. More details are available in [Bus92].



Collar



Fermi coordinates

Theorem 4.18. *The set E_i is measure zero.*

Proof. By theorem 4.6, E is a null set on the surface. We choose a copy of the surface and past the two surfaces along the boundary component β_i. Then in the half collar neighborhood of β_i on the surface, we have measure relation

$$\mu(E \cap \mathcal{C}[\beta_i]) = \sinh w \times \mu(E_i).$$

Hence $\mu(E_i) = 0$ from $\mu(E) = 0$. □

Theorem 4.19. *The intersection set E_i is a set and homemorphic to the Cantor set union countably many isolated points.*

Proof. There are at most countably many isolated points in E_i, let I_i be the set of the isolated points, which by theorem 4.12 are the intersections of the geodesics with the other end approaches to boundary components, i.e., spiraling to it or perpendicular to it. The $E_i - I_i$ has no isolated points, and it is closed from the argument following 4.6. Because this set has measure zero and closed hence its complementary is open dense subset of β_i which leads that E_i is totally disconnected.

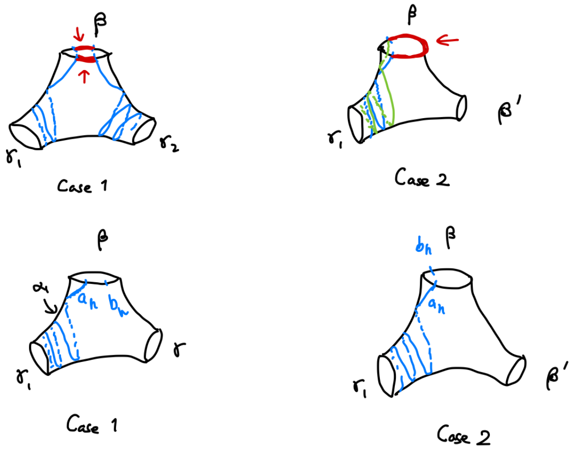
Then the result follows from any perfect totally disconnected compact metric space is homemorphic to Cantor set. □

Now we prove the theorem 4.3.

Proof. Let H_i be the set of connected components in $G_i = I_i \cup (\beta_i - E_i)$. By the Theorem 4.12, we can assume that $\beta_i - (E_i - I_i) = \bigcup_{h \in H_i} (a_h, b_h)$, where a_h, b_h are the ends of h. By theorem 4.18,

$$L_1 = \sum_{h \in H_i} |b_h - a_h|,$$

where $|b_h - a_h|$ denotes the geodesic length on the surface.



We state the correspondence between pair of pants and H_i.

Case 1. For an embedded pair of pants bounding by the boundary β and two non-peripheral geodesics γ_1, γ_2. There are two intervals in this pants, namely, the two parts bounding by one geodesic spiraling to γ_1 and one geodesic spiraling to γ_2,corresponding the interval (y_1, z_1) and (y_2, z_2) in 3.2 and the two red parts in above graph.

The ends of the intervals spiral to non-peripheral geodesics hence they are boundary points. There are no boundary points in the intervals otherwise the geodesic passing one boundary point entending off the pants forcing it must be perpendicular to the boundary which is impossible. Hence this intervals are intervals in the corresponding H_i indeed.

Case 2. For an embedded pair of pants bounding by the boundaries β and β' and one non-peripheral geodesic γ_1. The two points which the passing geodesics spiraling to γ_1 are boundary points and the interval between them intersecting with the geodesic perpendicular to both γ and γ' is indeed an interval because and boundary points on it corresponding to a geodesic extending off the pants which must intersect with γ_1 and hence perpendicular to it and this is impossible.

This corresponds ti the interval (y_1, y_2) containing w_1, w_2 in 3.2 and the right red part in the above graph.

For interval (a_h, b_h) on β, suppose a_h corresponds to a geodesic α_1 spiraling to a non-peripheral closed geodesic γ_1. There exists a unique embedded pair of pants containing γ_1 and α_1 with another boundary. We have:

Case 1. The another boundary is a nonperipheral geodesic γ. By the above correspondence, a_h is in an interval (a_h, b') and because (a_h, b_h) is an interval hence either $(a_h, b_h) \subset (a_h, b')$ or $(a_h, b') \subset (a_h, b_h)$ which deduce $(a_h, b_h) = (a_h, b')$ otherwise we will find a boundary point in an interval, a contradiction.

If we assume $\beta, \gamma_1, \gamma_2$ have lengths L, ℓ_1, ℓ_2. This case states that such an embedded pants corresponds to the sum of two interval, which is $\mathcal{D}(L, \ell_1, \ell_2)$.

Case 2. The another boundary is a different boundary. In this case, the geodesics passing a_h, b_h spiral to γ_1 and because there exists an interval ending them, the interval in above construction must be (a_h, b_h), having length $\mathcal{R}(L, \ell_1, \ell_2)$.

With these observation, and the length sum of β_i, we get the desired identity. □

5 Moduli space

In this section we introduce some basic notations about Teichmüller spaces and Fenchel-Nielsen coordinates then we introduce the multi curves and their symmetry group.

5.1 Teichmüller space

The moduli problem for a class of Riemann surfaces is to find a set of isometry invariants which determine the surface up to isometry. Teichmüller space is a space that parametrizes complex structures on a surface up to the action of homeomorphisms that are isotopic to the identity homeomorphism.

To be exactly, let S be a topological space of genus g, which is the base space.

Definition 5.1. A *marked Riemann surface* is a pair (X, f), where X is a complete hyperbolic surface and $f : S \to X$ is a homeomorphism.

We need the concept of isotopy class.

Definition 5.2. Two homeomorphisms of topological spaces $f : A \to B, g : A \to B$ are *isotopic* if there exists a continuous map $\Phi : [0,1] \times A \to B$ such that $\Phi(0,.) = f$ and $\Phi(1,.) = g$ and for every $s \in (0,1)$, $\Phi(s,.)$ is a homeomorphism.

Definition 5.3. (X, f) and (X', f') are *marking equivalent* if $f^{-1}f' : X \to X'$ is isotopic to a conformal map. Teichmüller space is the equivalence classes of all marked Riemann surfaces,namely,

$$\mathcal{T}(S) = \{ \text{ marked Riemann surfaces } \}/\text{marking equivalence.}$$

In this article, we are interesting about those hyperbolic surfaces with geodesic boundary components of fixed length, hence we consider the similar definition.

Let $A = \partial S$ consist of finitely many boundary components and $L = (L_\gamma) \in \mathbb{R}_+^{|A|}$ where γ is one of the boundary components of S and $|A|$ equals to the number of boundary components.

Definition 5.4. The Techimüller space $\mathcal{T}(S, L)$ consists of the remarked Riemann surfaces with geodesic components of given lengths, i.e., for any boundary component β in A,

$$\ell_\beta(X) = L_\beta.$$

If $S_{g,n}$ is an oriented connected surface of genus g with n boundary components $(\beta_1, \cdots, \beta_n)$. We fixed these notations. We write

$$\mathcal{T}_{g,n}(L) = \mathcal{T}_{g,n}(L_1, \cdots, L_n) = \mathcal{T}(S_{g,n}, L_1, \cdots, L_n)$$

as the Teichmüller space of hyperbolic structures on $S_{g,n}$ with geodesic boundary components of lengths $L_1, \cdots, L_n$.

Particularly, a boundary geodesic of length 0 degenerate to a cusp.

Next we introduce the mapping class group

Definition 5.5. The *mapping class group* $\mathrm{Mod}(S)$ of surface S is the isotopy classes of orientation preserving homemorphisms $h : S \to S$ such that $h(\beta_i) = \beta_i$ for all i. We write $\mathrm{Mod}_{g,n} = \mathrm{Mod}(S_{g,n})$.

For $h \in \mathrm{Mod}(S)$, h acts on $\mathcal{T}(S) \to \mathcal{T}(S)$ via the rule

$$(X, f) \mapsto (X, f \circ h).$$

Definition 5.6. The quotient space of the Teichmüller space with mapping class group

$$\mathcal{M}_{g,n}(L) = \mathcal{M}(S_{g,n}, \ell_{\beta_i} = L_i) = \mathcal{T}_{g,n}(L)/\mathrm{Mod}_{g,n}$$

is the *moduli space* of Riemann surfaces homeomorphic to $S_{g,n}$ with n geodesic boundary components of of length $\ell_{\beta_i} = L_i$.

Remark 5.6.1. For $\mathcal{T}(S)$ and $\mathrm{Mod}(S)$ defined above, we will omit the symbol S when no confusion.

Example 5.7. Consider an example of pair of pants. Let $S = S_{0,3}$ be a pair of pants.

From hyperbolic geometry, we know for every 3 positive reals $\beta_1, \beta_2, \beta_3$, there exists unique hyperbolic hexagon with the lengths of the 3 disjoint sides equaling to $\beta_1, \beta_2, \beta_3$ respectively and all its angles are right angle. we can past the hexagon with a copy of it along the 3 other sides to get a pair of pants with geodesic boundary components of lengths $\beta_1, \beta_2, \beta_3$. For a pair of pants, it can decompose to two same hexagon by cutting along the geodesics perpendicular to two boundary components. Hence the lengths of the boundary components uniquely determine a pair of pants.

By remarking, we can easily see that $\mathcal{T}_{0,3}(S) = \mathbb{R}_+^3$. And the mapping class group denote some symmetry, which is $\mathrm{Mod}_{0,3}(S) = \mathfrak{S}_3$. The moduli space $\mathcal{M}_{0,3}(S) = \mathcal{T}_{0,3}(S)/\mathrm{Mod}_{0,3}(S)$ and its fundamental domain is $\mathcal{F} = \{(x, y, z) : 0 < x \leq y \leq z\}$.

5.2 The Fenchel-Nielsen coordinates

A hyperbolic surface of genus g with n geodesic boundary components has pants decomposition to $2g-2+n$ pair of pants by cutting along $3g-3+n$ closed simple curves. Fix such a system of pants decomposition of $S_{g,n}$ and let $\mathcal{P} = \{\alpha_i\}^k$ be the closed simple curves, where $k = 3g-3+n$. The *Fenchel-Nielsen coordinates* associated to $\mathcal{P}$ is consist of the k length parameters ℓ_{α_i} and k twisting parameters τ_{α_i}, we have isomorphism

$$\mathcal{T}_{g,n}(L) \cong \mathbb{R}_+^{\mathcal{P}} \times \mathbb{R}^{\mathcal{P}}$$

via the rule

$$X \mapsto (\ell_{\alpha_i(X)}, \tau_{\alpha_i}(X)).$$

Remark 5.7.1. The concrete content about the Fenchel-Nielsen coordinates could be found in many textbooks, such as [Bus92].

We want to explain the twisting parameters clearly which is important in the integration formula later.

5.2.1 pasting

Consider two hyperbolic surfaces S and S' with two closed geodesics boundaries γ, γ' of the same length, and the two geodesics are have the same orientation, i.e., the S and S' are either both on the left hand side or on the right hand side. Assume $\gamma : \mathbb{S}^1 \to S$ and $\gamma' : \mathbb{S}^1 \to S'$ be the parametrizations. For $\alpha \in \mathbb{R}$, we past the two surfaces along the geodesics with marked twisting parameter α by identifying points on the boundary

$$\gamma(t) \equiv \gamma'(\alpha - t) := \gamma^\alpha(t), t \in S^1,$$

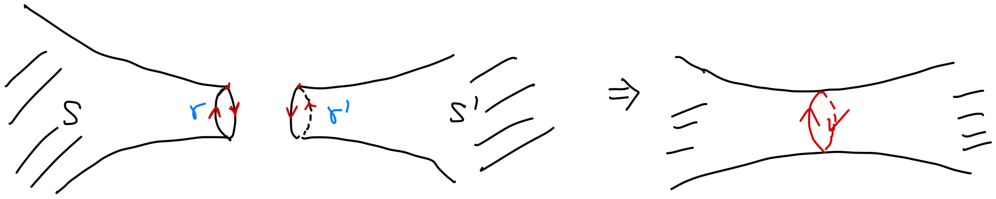
and let

$$F^\alpha = S + S'(\text{mod identification on } \gamma \text{ and } \gamma')$$

be the surface we get after pasting and let

$$\pi^\alpha : S \sqcup S' \to F$$

be the canonical projection. In particular, π denotes the trivial homeomorphism and $F = F^0$ be the trivial pasting.



5.2.2 twisting in the collar neighborhood

Recall in the collar neighborhood, we have Fermi coordinate (ρ, t). Let $\mathcal{C}[\gamma]$ and $\mathcal{C}[\gamma^\alpha]$ be the collar neighborhood in F and F^α. Let's construct a twist in the collar neighborhood.

Consider a homemorphism $T^\alpha : \mathcal{C}[\gamma] \to \mathcal{C}[\gamma^\alpha]$

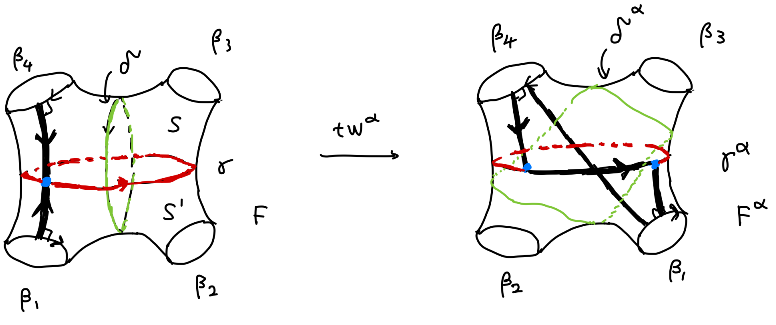
$$T^\alpha(\rho, t) = (\rho, t + \alpha \frac{w + \rho}{2w})$$

where w is the width of collar.

Consider twist homomorphism $\mathrm{tw}^\alpha_\gamma : F \to F^\alpha$ as

$$\mathrm{tw}^\alpha_\gamma(p) = \begin{cases} T^\alpha, \text{ if } p \in \mathcal{C}[\gamma], \\ \pi^\alpha \circ \pi^{-1}(p), \text{ if } p \in F - \mathcal{C}[\gamma]. \end{cases}$$

When $\alpha = \ell_\gamma(F)$, the twist homeomorphism $\phi_\gamma = \mathrm{tw}^\alpha_\gamma$ is called an elmentary *Dehn twist*. See the following graph.



5.3 multi curve

Definition 5.8. Let $\eta = \sum_{i=1}^{t} c_i\eta_i$ be a multi curve where η_i's are disjoint, essential and non-peripheral closed simple curves of $S_{g,n}$, namely each one is not homotopic to the other or any boundary components, and $c_i \geq 0$ for $1 \leq i \leq t$. And Let $\Gamma = (\eta_1, \cdots, \eta_t)$ be the marking. Fixed these notations.

Notice that we must have $t \leq 3g - 3 + n$.

Cutting $S_{g,n}$ along $\eta_1, \cdots, \eta_t$ we get the cut surface $S_{g,n}(\eta)$ which is possible disconnected. Each η_i produces two boundary components η_i^1, η_i^2 on $S_{g,n}(\eta)$, hence $S_{g,n}(\eta)$ has $n + 2t$ boundary components $\beta_1, \cdots, \beta_n, \eta_1^1, \eta_1^2, \cdots, \eta_t^1, \eta_t^2$.

Suppose $S_{g,n}(\eta)$ has s connected components S_{g_i,n_i} in which let A_i be the boundary of S_{g_i,n_i} with $n_i = |A_i|$ and g_i denote the genus of S_{g_i,n_i}. For $\vec{a} \in \mathbb{R}_+^t$, let

$$\mathcal{M}(S_{g,n}(\eta), \ell_\Gamma = \vec{a}, \ell_\beta = L)$$

be the moduli space of hyperbolic Riemann surface homemorphic with fixed lengths $\ell_{\eta_i} = a_i, \ell_{\beta_j} = L_j$. Obviously, this space is actually the product of the moduli space of the connected componnets, namely,

$$\mathcal{M}(S_{g,n}(\eta), \ell_\Gamma = \vec{a}, \ell_\beta = L) = \prod_{i=1}^{s} \mathcal{M}(S_{g_i,n_i}, L_{A_i})$$

where L_{A_i} is the length tuples of the boundary of S_{g_i,n_i}.

Remark 5.8.1. We can regard the η_i's as geodesics on $S_{g,n}$ if we set the base space as hyperbolic Riemann surface.

Later we will consider the moduli space which have fixed lengths on its boundaries and the multi curve.

5.4 Symmetry group of multi curve

For a multi curve $\eta = \sum_{i=1}^{t} c_i\eta_i$ as above. We consider the symmetry group of η that essentially preserve the η_i's as follows.

Definition 5.9. Let C be the homotopy classes of some closed simple curves on $S_{g,n}$, we define

$$\mathrm{Stab}(C) = \{h \in \mathrm{Mod}_{g,n} : h \cdot C = C\}$$

to be the subgroup of the mapping class group that preserving the homotopy classes.

Let

$$\mathrm{Sym}(\eta) = \mathrm{Stab}(\eta) / \bigcap_{i=1}^{t} \mathrm{Stab}(\eta_i)$$

be the symmetry group of η.

Remark 5.9.1. This group is similar with the permutation group of $\eta_1, \cdots, \eta_t$. For example, if $c_1 = c_2 = 1$, we can see that $|\mathrm{Sym}(\eta_1 + \eta_2)| = 2$ iff there exists $h \in \mathrm{Mod}_{g,n}$ satisfying $h \cdot \eta_1 = \eta_2, h \cdot \eta_2 = \eta_1$, i.e., a homemorphism between $S_{g,n}(\eta_1)$ and $S_{g,n}(\eta_2)$.

5.5 $\mathrm{Mod}_{g,n}$-orbit of simple closed curves

Assume $[\gamma]$ be the homotopy class of a simple closed curve γ on $S_{g,n}$. Define the $\mathrm{Mod}_{g,n}$-orbit of γ as

$$\mathcal{O}_\gamma = \{[h \cdot \gamma] : h \in \mathrm{Mod}_{g,n}\},$$

which is determined by γ.

For a multi curve $\eta = \sum_{i=1}^{t} c_i \eta_i$, let $\Gamma = (\eta_1, \cdots, \eta_t)$. We can define the $\mathrm{Mod}_{g,n}$-orbit of Γ in the same way, namely, let

$$\mathcal{O}_\Gamma = \{[h \cdot \Gamma] : h \in \mathrm{Mod}_{g,n}\},$$

where $[h \cdot \Gamma] = ([h \cdot \gamma_1], \cdots, [h \cdot \gamma_t])$.

For a function $f : \mathbb{R}_+ \to \mathbb{R}_+$, we set

$$f_\eta(X) = \sum_{[\alpha] \in \mathrm{Mod} \cdot [\eta]} f(\ell_\alpha(X)),$$

where $\ell_\alpha(X) = \sum_{i=1}^{t} c_i \ell_{\gamma_i}(X)$. This defines a function $f_\eta : \mathcal{M}_{g,n}(L) \to \mathbb{R}_+$.

6 Integration over moduli space

In this section, through the preparation in previous section, we can deduce the integration formula.

Follow the notation in Section 5.3. Let $\mathrm{Vol}_{g,n}(\Gamma, \vec{x}, \beta, L)$ denote the volume of $\mathcal{M}(S_{g,n}(\eta), \ell_\Gamma = \vec{a}, \ell_\beta = L)$. We write $\mathrm{Vol}_{g,n}(L)$ when we consider only boundary components.

The main theorem of this part followed [Mir07a].

Theorem 6.1. *The integral of f_η over moduli space $\mathcal{M}_{g,n}(L)$ with respect to the Weil-Petersson volume form is given by*

$$\int_{\mathcal{M}_{g,n}(L)} f_\eta(X) dX = \frac{2^{-M(\eta)}}{|Sym(\eta)|} \int_{\vec{x} \in \mathbb{R}_+^t} f(|\vec{x}|)\, Vol_{g,n}(\Gamma, \vec{x}, \beta, L) \vec{x} d\vec{x},$$

where $|\vec{x}| = \sum_{i=1}^{t} c_i x_i$ *and*

$$M(\eta) = |\{i : \eta_i \textit{ separates off a one-handle from } S_{g,n}\}|.$$

Remark 6.1.1. Follow the notation in 5.3, we can see that naturally we should have

$$\mathrm{Vol}_{g,n}(\Gamma, \vec{x}, \beta, L) = \prod_{i=1}^{s} \mathrm{Vol}_{g_i,n_i}(L_{A_i}).$$

We will see in the calculation section this formula reduces the volume of moduli space to the volume of "smaller" space.

6.1 Symplectic sructure over Teichmüller space

Definition 6.2. Let M be a smooth manifold M. A *nondegenerate 2-form* on M is a 2-form ω such that ω_p is a nondegenerate 2-covector for each $p \in M$. A *symplectic form* on M is a closed nondegnerate 2-form. Such a form sometimes is called a symplectic structure.

Remark 6.2.1. We must remind ourselves that in this article we will not deal strictly with this concept.

Example 6.3. With standard coordinates on $\mathbb{R}^{2n}$ denoted by $(x^1, \cdots, x^n, y^1, \cdots, y^n)$, the standard sympletic form on $\mathbb{R}^{2n}$ is

$$\omega = \sum_{i=1}^{n} \mathrm{d}x^i \wedge \mathrm{d}y^i.$$

In his celebrated paper [Wol82], Wolpert proves that the Weil-Petersson symplectic structure has a simple form in Fenchel-Nielsen coordinates.

Theorem 6.4 (Wolpert). *The Weil-Petersson sympletic form is given by*

$$\omega_{wp} = \sum_{i=1}^{k} d\ell_{\alpha_i} \wedge d\ell_{\tau_i}.$$

6.2 Integration under covering maps

Let $\pi : X \to Y$ be a covering maps and ω is a volume form on Y then we can define a volume form on X by pushing back. To be exactly, $\eta = \pi^*\omega$, i.e.,

$$(\pi^*\omega)_p = d\pi_p^*(\omega_{F(p)}),$$

which acts on a vector $v \in T_pM$ by

$$(\pi^*\omega)_p(v) = \omega_{F(p)}(dF_p(v)).$$

For $f \in L^1(X, \eta)$, the push forward

$$(\pi_* f)(x) = \sum_{y \in \pi^{-1}x} f(y)$$

defines a function in $L^1(Y, \omega)$ and by integrating on atlases we get

$$\int_X f \mathrm{d}\eta = \int_Y (\pi_* f) \mathrm{d}\omega.$$

6.3 Coverings and volume forms of the $\mathcal{M}_{g,n}(L)$'s

Recall $\mathcal{O}_\Gamma = \{[h \cdot \Gamma] : h \in \mathrm{Mod}_{g,n}\}$. To prove Theorem 6.1, it is supposed that we should extend the moduli space $\mathcal{M}_{g,n}(L)$. To view this, we see that the function f_η is defined as sum over $\mathrm{Mod}_{g,n}$-orbit of multi-curve η, namely, we have

$$f_\eta(X) = \sum_{h \in \mathrm{Mod}_{g,n}/\mathrm{Stab}(\eta)} f(\ell_{h\cdot\eta}(X)).$$

Recall the symmetry group of η, $\mathrm{Sym}(\eta)$, which means some extra symmetry on η. If $g \in \mathrm{Sym}(\eta)$, then

$$\eta = \sum_{i=1}^{t} c_i \eta_i = \sum_{i=1}^{t} c_i g \cdot \eta_i,$$

hence when $\eta_j = g \cdot \eta_i$, $c_i = c_j$. From this, the value of c_i's in an orbit of the action of the cyclic group generated by g is constant. So for $g \in \mathrm{Sym}(\eta)$ and $h \in \mathrm{Mod}_{g,n}/\mathrm{Stab}(\eta)$ the formula $\ell_{gh\cdot\eta}(X) = \ell_{h\cdot\eta}(X)$ holds.

We can rewrite the integral formula from above analysis and the definition 5.9 of $\mathrm{Sym}(\eta)$

$$\begin{aligned}\int_{\mathcal{M}_{g,n}(L)} f_\eta(X)\mathrm{d}X &= \int_{\mathcal{M}_{g,n}(L)} \sum_{[\alpha]\in \mathrm{Mod}\cdot[\eta]} f(\ell_\alpha(X))\mathrm{d}X \\ &= \int_{\mathcal{M}_{g,n}(L)} \sum_{h \in \mathrm{Mod}_{g,n}/\mathrm{Stab}(\eta)} f(\ell_{h\cdot\eta}(X))\mathrm{d}X \\ &= \frac{1}{|\mathrm{Sym}(\eta)|} \int_{\mathcal{M}_{g,n}(L)} \sum_{h \in \mathrm{Mod}_{g,n}/\bigcap_{i=1}^{t} \mathrm{Stab}(\eta_i)} f(\ell_{h\cdot\eta}(X))\mathrm{d}X\end{aligned}$$

Notice that the sum of integral part is over the $\mathrm{Mod}_{g,n}$-orbit of $\Gamma = (\eta_1, \cdots, \eta_t)$. Once we calculate the integral by choosing $\vec{x}$ and collecting (X, h) with $\ell_{h\cdot\eta_i}(X) = x_i$ the integral is transformed to calculate some special moduli space with fixed lengths on boundary components and $\mathcal{O}_\Gamma$, which is less complicated than original integral.

There exists a unique simple geodesic in the homotopy class of a simple closed curve, thus we define

Definition 6.5. Set $\mathcal{M}_{g,n}(L)^\Gamma$ be the set of pairs

$$\{(X, \gamma) : X \in \mathcal{M}_{g,n}(L), \gamma = (\gamma_1, \cdots, \gamma_t) \in \mathcal{O}_\Gamma, \gamma_i\text{'s are closed geodesics on X}\}.$$

Let $\pi^\Gamma : \mathcal{M}_{g,n}(L)^\Gamma \to \mathcal{M}_{g,n}(L)$ be the natural projection $\pi^\Gamma(X, \gamma) = X$.

By the above analysis, this space is actually the marked hyperbolic structures on X carried by

$$G_\Gamma = \mathrm{Mod}_{g,n}/\bigcap_{i=1}^{t} \mathrm{Stab}(\eta_i),$$

hence we have

$$\mathcal{M}_{g,n}(L)^{\Gamma} = \mathcal{T}_{g,n}(L)/G_{\Gamma}.$$

Next, we will reduce this volume to the volume of "smaller" space. We give an intuitive explanation of the process.

Consider the length function $\mathcal{L}_{\Gamma} : \mathcal{M}_{g,n}(L)^{\Gamma} \to \mathbb{R}^t$ defined as

$$\mathcal{L}_{\Gamma}(X, \gamma) = (\ell_{\gamma_1}(X), \cdots, \ell_{\gamma_t}(X)).$$

In view of integration, we consider the level set $\mathcal{M}_{g,n}(L)^{\Gamma}[\vec{a}] := \mathcal{L}_{\Gamma}^{-1}(\vec{a})$.

We have a map

$$\pi : \mathcal{M}_{g,n}(L)^{\Gamma}[\vec{a}] \to \mathcal{M}(S_{g,n}(\eta), \ell_{\Gamma} = \vec{a}, \ell_{\beta} = L)$$

defined in such way:

Given $(X, \eta) \in \mathcal{M}_{g,n}(L)^{\Gamma}$,we take a lift $\tilde{X} \in \mathcal{T}_{g,n}(L)$ of X and maps it to the point corresponding to its cut surface in $\mathcal{M}(S_{g,n}(\eta), \ell_{\Gamma} = \vec{a}, \ell_{\beta} = L)$.

For every twist $\phi_{\alpha}^{t} = \mathrm{tw}_{\alpha}^{t \cdot \ell_{\alpha}(X)}$ the equation $\ell_{\alpha}(X) = \ell_{\alpha}(\mathrm{tw}_{\alpha}^{t}(X))$ holds. Hence the twisting along each components of η induced $(S^1)^k$ action on the level set. Under this action the Weil-Petersson form is invariant hence induced a symplectic structure on the quotient space of the action $\pi : \mathcal{M}_{g,n}(L)^{\Gamma}[\vec{a}]$ and the images of the action under π is actually invariant. In fact, these two spaces are essentially the same up to simplecteomorphism.

The actions of $\mathbb{S}^1$-actions on η mean normalizing the length parameters and to collect the twisting parameters determining different hyperbolic structures. In general, we can see the twisting parameter of η_i are between 0 and ℓ_i and when η_i separates a torus of genus 1, the twsiting parameter is between 0 and $\dfrac{\ell_i}{2}$. The exceptions form a closed set of measure 0 in $\mathcal{M}_{g,n}(L)^{\Gamma}[\vec{a}]$.

Thus for any open set U in $\mathcal{M}(S_{g,n}(\eta), \ell_{\Gamma} = \vec{a}, \ell_{\beta} = L)$, we have

$$\mathrm{Vol}(\pi^{-1}(U)) = 2^{-M(\eta)} \mathrm{Vol}(U) a_1 \cdots a_t.$$

Hence we can deduce that

$$\begin{aligned}
\int_{\mathcal{M}_{g,n}(L)} f_{\eta}(X) \mathrm{d}X &= \frac{1}{|\mathrm{Sym}(\eta)|} \int_{\mathcal{M}_{g,n}(L)} \sum_{h \in \mathrm{Mod}_{g,n} / \bigcap_{i=1}^{t} \mathrm{Stab}(\eta_i)} f(\ell_{h \cdot \eta}(X)) \mathrm{d}X \\
&= \frac{1}{|\mathrm{Sym}(\eta)|} \int_{\vec{x} \in \mathbb{R}_{+}^{t}} \left(\int_{\mathcal{M}_{g,n}(L)^{\Gamma}[\vec{x}]} dX \right) f(|\vec{x}|) d\vec{x} \\
&= \frac{1}{|\mathrm{Sym}(\eta)|} \int_{\vec{x} \in \mathbb{R}_{+}^{t}} f(|\vec{x}|) 2^{-M(\eta)} \mathrm{Vol}_{g,n}(\Gamma, \vec{x}, \beta, L) x_1 \cdots x_t d\vec{x} \\
&= \frac{2^{-M(\eta)}}{|\mathrm{Sym}(\eta)|} \int_{\vec{x} \in \mathbb{R}_{+}^{t}} f(|\vec{x}|) \mathrm{Vol}_{g,n}(\Gamma, \vec{x}, \beta, L) \vec{x} d\vec{x},
\end{aligned}$$

as desired.

7 Examples of Calculation

In [Mir07a], Mirzakhani derives the recursive formula of the border surface $V_{g,n}(L)$. In this section, we provide a few examples to realize the recursion process instead of stating the complete result. Notice $\mathrm{Vol}_{0,3}(L) = 1$.

7.1 $\mathcal{M}_{1,1}(L)$

There is only one boundary component on $S_{1,1}$, hence we consider only $\mathcal{F}_1$ in the generalized McShane's identity, and when $\{\alpha_1, \alpha_2\} \in \mathcal{F}_1$, it is supposed that $\alpha_1 = \alpha_2$. Hence the generalized McShane's identity reduces to

$$\sum_{\alpha} \mathcal{D}(L, \ell_\alpha(X), \ell_\alpha(X)) = L,$$

where the sum is over all simple closed geodesics on $X \in \mathcal{M}_{1,1}(L)$. Actually,they are the $\mathrm{Mod}_{1,1}(L)$-orbit of the non-peripheral closed geodesics and cutting along the geodesic we get a pair of pants with two equaled boundary components.

Integrating the formula over moduli space, we get

$$L\mathrm{Vol}_{1,1}(L) = \int_{\mathcal{M}_{1,1}(L)} \sum_{\alpha} \mathcal{D}(L, \ell_\alpha(X), \ell_\alpha(X))dX,$$

Take the derivative of L, and by integration formula 6.1,

$$\begin{aligned}
\frac{\partial}{\partial L} L\mathrm{Vol}_{1,1}(L) &= \int_{\mathcal{M}_{1,1}(L)} \sum_{\alpha} \frac{\partial}{\partial L}\mathcal{D}(L, \ell_\alpha(X), \ell_\alpha(X))dX \\
&= \int_0^\infty \frac{\partial}{\partial L}\mathcal{D}(L,t,t)\mathrm{Vol}_{0,3}(t,t,L)dt \\
&= \int_0^\infty t(\frac{1}{1+e^{t+\frac{L}{2}}} + \frac{1}{1+e^{t-\frac{L}{2}}})dt \\
&= \int_{\frac{L}{2}}^\infty \frac{t-\frac{L}{2}}{1+e^t}dt + \int_{\frac{-L}{2}}^\infty \frac{t+\frac{L}{2}}{1+e^t}dt \\
&= \int_0^\infty \frac{2t}{1+e^t}dt + \int_0^{\frac{L}{2}} (\frac{L}{2}-t)(\frac{1}{1+e^t} + \frac{1}{1+e^{-t}})dt \\
&= \int_0^\infty \frac{2t}{1+e^t}dt + \int_0^{\frac{L}{2}} (\frac{L}{2}-t)dt \\
&= \frac{\pi^2}{6} + \frac{L^2}{8},
\end{aligned}$$

hence

$$\mathrm{Vol}_{1,1}(L) = \frac{\pi^2}{6} + \frac{L^2}{24}.$$

Notice from the above calculation we can get $\int_0^\infty xH(x,L)dx = \frac{2}{3}\pi^2 + \frac{L^2}{2}$.

7.2 $\mathcal{M}_{0,4}(L)$

Consider the hyperbolic surface X homemorphic to $S_{0,4}$ with geodesic boundary components $\beta_1, \beta_2, \beta_3, \beta_4$ with lengths $L = (L_1, L_2, L_3, L_4)$.

Before we apply the identity, we should study the set $\mathcal{F}_1$ and $\mathcal{F}_{1,i}$.

For $\{\alpha_1, \alpha_2\} \in \mathcal{F}_1$, if cutting along them the cut surface has at most two connected components. Two connected components will force the original surface has genus more than zero, a contradiction. If there are two components, the two components must have 3 and 2 boundary components corresponding to a pair of pants and an embedded annulus, a contradiction from the fact that there is no hyperbolic structure on annulus.

Hence we consider only the sets $\mathcal{F}_{1,i}$. Fixed i, the set $\mathcal{F}_{1,i}$ is a complete homotopy class, i.e., $\mathcal{F}_{1,i} = \mathrm{Mod}_{0,4} \cdot [\gamma]$ for $\gamma \in \mathcal{F}_{1,i}$. And the cut surface is obvious a pair of pants.

Similarly,

$$\sum_{i=2}^{4} \sum_{\gamma \in \mathcal{F}_{1,i}} \mathcal{R}(L_1, L_i, l_\gamma(X)) = L_1,$$

and

$$\begin{aligned}
\frac{\partial}{\partial L_1} L_1 \mathrm{Vol}_{0,4}(L) &= \int_{\mathcal{M}_{0,4}(L)} \sum_{i=2}^{n} \sum_{\gamma \in \mathcal{F}_{1,i}} \frac{\partial}{\partial L_1} \mathcal{R}(L_1, L_i, l_\gamma(X)) dX \\
&= \int_0^\infty \sum_{i=2}^{4} \frac{\partial}{\partial L_1} \mathcal{R}(L_1, L_i, t) \mathrm{Vol}_{0,3}(t, L_j, L_k) t dt \\
&= \int_0^t \sum_{i=2}^{4} \frac{1}{2} (H(t, L_1 + L_i) + H(t, L_1 - L_i)) t dt \\
&= \sum_{i=2}^{4} \frac{1}{2} (\frac{2}{3}\pi^2 + \frac{1}{2}(L_1 + L_i)^2 + \frac{2}{3}\pi^2 + \frac{1}{2}(L_1 - L_i)^2) \\
&= 2\pi^2 + \frac{3}{2} L_1^2 + \sum_{i=2}^{4} \frac{1}{2} L_i^2,
\end{aligned}$$

thus

$$\mathrm{Vol}_{0,4}(L) = 2\pi^2 + \frac{1}{2}(L_1^2 + L_2^2 + L_3^2 + L_4^2).$$

7.3 $\mathcal{M}_{1,2}(L)$

Let's consider a little more complicate example, the moduli space of $S_{1,2}(L)$ with boundary components β_1, β_2 of fixed lengths $L = (L_1, L_2)$.

As usual, we consider the sets $\mathcal{F}_1$ and $\mathcal{F}_{1,i}$.

Case 1. For set $\mathcal{F}_1$, suppose $\{\alpha_1, \alpha_2\} \in \mathcal{F}_1$. Obviously $\alpha_1 \neq \alpha_2$ and cutting along them we get one connected component with 3 boundary components or two connected components with genus 1 and 3 boundary components which is impossible. And similarly, $\mathcal{F}_1$ is a complete homotopy class.

Hence cutting along the two geodesics, we get a pair of pants. And by remark 5.9.1, the symmetry group of the multi curve added by the two geodesics has two elements because cutting along any one of them we get a surface homemorphic to $S_{0,4}$.

Case 2. For set $\mathcal{F}_{1,2}$. Cutting along $\gamma \in \mathcal{F}_{1,2}$, we get an $Y \in \mathcal{M}_{1,1}$ and a pair of pants. That means the geodesic separates off a torus with one boundary.

Hence write down the identity,

$$\sum_{\{\alpha_1,\alpha_2\}\in\mathcal{F}_1} \mathcal{D}(L_1, l_{\alpha_1}(X), l_{\alpha_2}(X)) + \sum_{\gamma\in\mathcal{F}_{1,2}} \mathcal{R}(L_1, L_2, l_\gamma(X)) = L_1.$$

and

$$\begin{aligned}\frac{\partial}{\partial L_1} L_1 \mathrm{Vol}_{1,2}(L) &= \int_{\mathcal{M}_{1,2}(L)} \sum_{\{\alpha_1,\alpha_2\}\in\mathcal{F}_1} \frac{\partial}{\partial L_1}\mathcal{D}(L_1, l_{\alpha_1}(X), l_{\alpha_2}(X)) + \sum_{\gamma\in\mathcal{F}_{1,2}} \frac{\partial}{\partial L_1}\mathcal{R}(L_1, L_2, l_\gamma(X)) dX \\ &= \frac{1}{2}\int_{\vec{x}\in\mathbb{R}_+^2} H(|\vec{x}|, L_1) x_1 x_2 d\vec{x} + \frac{1}{2}\int_0^\infty \frac{\partial}{\partial L_1}\mathcal{R}(L_1, L_2, t)\mathrm{Vol}_{1,1}(t) t dt \\ &= \frac{1}{2}\int_0^\infty H(t, L_1)\frac{1}{6}t^3 dt + \int_0^\infty \frac{1}{4}(H(t, L_1+L_2) + H(t, L_1 - L_2))(\frac{\pi^2}{6} + \frac{t^2}{24}) t dt\end{aligned}$$

And we see that it suffices to calculate for every integer $k \geq 0$, the following integral

$$I_{2k+1}(t) = \int_0^\infty x^{2k+1} \cdot H(x,t) dx.$$

This will reduce to calculate the integral

$$I_{2k+1} = I_{2k+1}(t, 0) = \int_0^\infty \frac{2x^{2k+1}}{1 + e^{\frac{x}{2}}} dx,$$

which relates to $\zeta(2k+2)$. In fact, an explicit formula is appeared in [Mir07a] as

$$\frac{I_{2k+1}(t)}{(2k+1)!} = \sum_{i=0}^{k+1} \zeta_{2i}(2^{2i+1} - 4)\frac{t^{2k+2-2i}}{(2k+2-2i)!},$$

where $k \geq 0$. Recall $\zeta(s)$ has unique pole at 1 and has value $-\dfrac{1}{2}$ at 0.

From this, we can deduce by easy calculation that

$$\mathrm{Vol}_{1,2}(L) = \frac{1}{192}(4\pi^2 + L_1^2 + L_2^2)(12\pi^2 + L_1^2 + L_2^2).$$

7.4 Summary of Calculation

From the above calculation, we can see the only concerned multi curves are those in $\mathcal{F}_1$ and $\mathcal{F}_{1,i}$. For $\mathcal{F}_{1,i}$, it is a complete homotopy class, we just count how many torus with one boundary it separates off.

For $\mathcal{F}_1$, we must notice the symmetry group and the separating property which completely determined by the topological structure on the base space $S_{g,n}$. And when we cut along these multi curves we see that the remaining smaller space has less genus or less boundary components and the sum of their numbers is strictly decreasing. Hence this afford an algorithm to calculate the volume of moduli space of bordered surfaces.

In fact, the volume $\mathrm{Vol}_{g,n}(L)$ is a symmetric polynomial of $L_1, L_2, \cdots, L_n$.

8 More about moduli space and its volumes

In a next paper [Mir07b], Mirzakhani applies the recursive formula of $\mathrm{Vol}_{g,n}(L)$ to the intersection theory on the moduli space of curves which can deduce the Witten conjecture. In fact, these information are stored in the coefficients of the polynomials of the volume which in this article are not mentioned.

From the last section of calculation and the pants decomposition, we can feel that sometimes the structure of moduli space is something more or less like combinatorics somehow. Actually, in his proof of Witten conjecture, Kontsevich[Kon92] studied a combinatorial model $\mathcal{M}_{g,n}^{\mathrm{comb}}$ for the moduli space of curves. In [etc10], the authors study the Kontsevich geometry of the combinatorial Theichmüller space and attain many results including a formula parallel to the recursive formula of Mirzakhani's, which is very interesting. [Do] is a wonderful survey of the related topics.

References

[Bon98] Francis Bonahon. Geodesic laminations on surfaces. 1998.

[Bus92] Peter Buser. Geometry and spectra of compact riemann surfaces. 1992.

[CEG87] R.D. Canary, D.B.A Epstein, and P. Green. *notes on notes of thurston*. Cambridge University Press, 1987.

[Do] Norma Do. Moduli spaces of hyperbolic surfaces and their weil–petersson volumes. `arXiv:1103.4674v1`.

[etc10] Jørgen Ellegaard Andersen etc. On the kontsevich geometry of the combinatorial theichmüller space. 2010.

[Kon92] M. Kontsevich. Intersection theory on the moduli space of curves and the matrix airy function. 1992.

[Mcs91] G. Mcshane. *A remarkable identity for lengths of curves.* PhD thesis, University of Warwick, 1991.

[Mcs98] G. Mcshane. Simple geodesics and a series constant over teichmüller space. *Invent. Math*, pages 607–632, 1998.

[Mir07a] M. Mirzakhani. Simple geodesics and weil-petersson volumes of moduli spaces of bordered riemann surfaces. *Invent. Math*, pages 179–222, 2007.

[Mir07b] M. Mirzakhani. Weil-petersson volumes and intersection theory on the moduli space of curves. 2007.

[Thu] William.P. Thurston. The geometry and topology of 3-manifolds. To Appear.

[Wol82] S. Wolpert. The fenchel-nielsen deformation. *Ann. Math. 115*, pages 501–528, 1982.