TOPICS IN ANALYSIS

MANJUNATH KRISHNAPUR

Contents

CHAPTER 1

Questions of approximation

1. Weierstrass' approximation theorem

Unless we say otherwise, all our functions are allowed to be complex-valued. For e.g., $C[0,1]$ means the set of complex-valued continuous functions on $[0,1]$. When equipped with the sup-norm $\|f\|_{\sup} := \max\{|f(x)| : x \in [0,1]\}$, it becomes a Banach space. Weierstrass showed that polynomials are dense in $C[0,1]$.

Theorem 1 (Weierstrass). *If $f \in C[0,1]$ and $\varepsilon > 0$ then there exists a polynomial P such that $\|f - P\|_{\sup} < \varepsilon$. If f is real-valued, we may choose P to be real-valued.*

Bernstein's proof. Define $B_n^f(x) := \sum_{k=0}^n f(k/n)\binom{n}{k}x^k(1-x)^{n-k}$, called the Bernstein polynomial of degree n for the function f. Make the following observations about the coefficients $p_{n,x}(k) = \binom{n}{k}x^k(1-x)^{n-k}$.

$$\sum_{k=0}^n p_{n,x}(k) = 1, \quad \sum_{k=0}^n kp_{n,x}(k) = nx, \quad \sum_{k=0}^n (k-nx)^2 p_{n,x}(k) = nx(1-x),$$

all of which can be easily checked using the binomial theorem. In probabilistic language, $p_{n,x}$ is a probability distribution on $0, 1, \ldots, n$ whose mean is nx and standard deviation is $nx(1-x)$. From these observations we immediately get

$$\sum_{k:|\frac{k}{n}-x|\ge\delta} p_{n,x}(k) \le \frac{1}{\delta^2 n^2}\sum_{k=0}^n (k-nx)^2 p_{n,x}(k) = \frac{x(1-x)}{n\delta^2}.$$

Thus, denoting $\omega_f(\delta) = \sup\limits_{|x-y|\le\delta}|f(x)-f(y)|$, we get

$$\begin{aligned}
|B_n^f(x) - f(x)| &\le \sum_{k\,:\,|\frac{k}{n}-x|<\delta} |f(x)-f(k/n)|p_{n,x}(k) + \sum_{k\,:\,|\frac{k}{n}-x|\ge\delta} |f(x)-f(k/n)|p_{n,x}(k) \\
&\le \omega_f(\delta) \sum_{k\,:\,|\frac{k}{n}-x|<\delta} p_{n,x}(k) + 2\|f\|_{\sup}\frac{x(1-x)}{n\delta^2} \\
&\le \omega_f(\delta) + \frac{1}{2n\delta^2}\|f\|_{\sup}.
\end{aligned}$$

First pick $\delta > 0$ so that $\omega_f(\delta) < \varepsilon/2$ and then pick $n > \frac{\|f\|_{\sup}}{\varepsilon\delta^2}$ to get $\|B_n^f - f\|_{\sup} < \varepsilon$. ■

Here is another proof of Weierstrass' theorem, probably closer to the original proof. The idea is that real analytic functions (on an open neighbourhood of $[0,1]$) are obviously uniformly approximable by polynomials (by truncating their power series around $1/2$ to finitely many terms), hence it suffices to show that any continuous function can be approximated uniformly by a real-analytic function. The key idea is of convolution.

Definition 2. For $f, g : \mathbb{R} \mapsto \mathbb{R}$, define $(f * g)(x) := \int_{\mathbb{R}} f(x-t)g(t)dt$, whenever the integral exists.

When f, g are positive integrable functions, $(f * g) : \mathbb{R} \to \mathbb{R}_+ \cup \{+\infty\}$ is well-defined, and Fubini's theorem shows that

$$\int (f * g)(x)dx = \left(\int f(u)du\right)\left(\int g(t)dt\right)$$

which is finite. In particular this means that $(f*g)(x)$ is finite for almost every x, and the resulting function in L^1. One can assume less about one of the functions and more about the other, to ensure that $f * g$ is well-defined. In fact, the key thing to remember about convolution is that it has the combined niceness of the two functions. The following exercise gives a few important special conditions.

Exercise 3.
1. If f is bounded and measurable and g is (absolutely) integrable, then $f * g$ and $g * f$ are well-defined and equal. Further, $f * g$ is bounded and integrable.
2. If f is bounded and measurable and g is smooth, then $f * g$ is smooth.
3. If f is bounded and measurable and g is real-analytic and integrable, then $f * g$ is real-analytic.

An effective way to approximate a function by a nicer function is to convolve it with a sequence of probability densities that concentrate their mass closer and closer to zero. In the following exercise below, execute this plan to give another proof of Weierstrass' theorem.

Exercise 4. Take all functions to be real-valued and defined on whole of the real line.
1. A real-analytic function can be uniformly approximated on compact sets by polynomials.
2. If φ is a real-analytic probability density, then so is $\varphi_\sigma(x) := \frac{1}{\sigma}\varphi(x/\sigma)$.
3. If f is a compactly supported continuous function, then $f * \varphi_\sigma$ is real analytic.
4. As $\sigma \to 0$, we have $f * \varphi_\sigma \to f$ uniformly on compact sets.
5. Deduce Weierstrass' theorem.

There are many examples of real-analytic probability densities. For example, (1) $\varphi(x) = \frac{1}{\pi(1+x^2)}$ (Cauchy density) and (2) $\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ (normal density).

Remark 5. If the above densities are used, then the approximating functions $f * \varphi_\sigma$ used in the above exercise has a more special meaning.

1. The Cauchy density $\varphi(x) = \frac{1}{\pi(1+x^2)}$. In this case, $(f * \varphi_y)(x) = u(x, y)$ where $u : \overline{\mathbb{H}} \to \mathbb{R}$ is the unique function that solves the Dirichlet problem[1] on the upper-half plane $\mathbb{H} :=$

[1] The Dirichlet problem asks for a continuous function on the closure of the domain (here upper half plane) that is harmonic in the interior and equal to a given continuous function on the boundary (here real line or more precisely $\mathbb{R} \cup \{\infty\}$).

$\{(x,y) : y > 0\}$ with boundary condition f. What this means is that (a) u is continuous on $\bar{\mathbb{H}}$, (b) $u(\cdot,0) = f(\cdot)$, (c) $\Delta u = 0$ on $\mathbb{H}$.

The point is that $(f * \varphi_y)$ is just u restricted to the line with y-co-ordinate equal to y and approaches f (at least pointwise) when $y \to 0$.

(2) The normal density $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. In this case $(f * \varphi_t)(x) = u(x,t)$ where u solves the *heat equation* with initial condition f. What this means is that (a) u is continuous on $\mathbb{R} \times \bar{\mathbb{R}}_+$, (b) $u(\cdot,0) = f(\cdot)$, (c) $\frac{\partial}{\partial t} u(x,t) = \frac{1}{2} \frac{\partial^2}{\partial x^2} u(x,t)$ on $\mathbb{R} \times \mathbb{R}_+$.

Again, $(f * \varphi_{\sqrt{t}}) = u(\cdot,t)$ is the function ("temperature") at time t, and approaches the initial condition f (at least pointwise) as t approaches 0.

Some questions to think about. What about polynomials in m variables? Are they dense in the space $C(K)$ for $K \subseteq \mathbb{R}^m$? What about polynomials in one complex variable? Are they dense in the space $C(\bar{\mathbb{D}})$ where $\bar{\mathbb{D}}$ is the closed unit disk in the complex plane?

A somewhat challenging exercise on approximation of functions on the whole line.

Exercise 6. If $f : [0,\infty) \mapsto \mathbb{R}$ is continuous and $f(x) \to 0$ as $x \to +\infty$, then show that for any $\varepsilon > 0$, there is polynomial p such that $|f(x) - p(x)e^{-x}| < \varepsilon$ for all $x \geq 0$.

2. Fejér's theorem

Let S^1 denote the unit circle which we may identify with $[-\pi,\pi)$ using the map $\theta \mapsto e^{i\theta}$. Continuous functions on S^1 may be identified with continuous functions on $I = [-\pi,\pi]$ such that $f(-\pi) = f(\pi)$ or equivalently, with 2π-periodic continuous functions on $\mathbb{R}$.

Let $e_k(t) = e^{ikt}$ for $t \in [-\pi,\pi)$ (these are 2π-periodic as k is an integer). A supremely important fact is that e_k are orthonormal in $L^2(I, dt/2\pi)$, i.e., $\int_I e_k(t)\bar{e}_\ell(t)\frac{dt}{2\pi} = \delta_{k,\ell}$. The question of whether this is a complete orthonormal basis is answered to be "yes" by the following theorem. Note that the L^2 norm is dominated by the sup-norm, hence a dense subset of $C(S^1)$ is also dense in $L^2(S^1)$.

Theorem 1 (Fejér). *Given any $f \in C(S^1)$ and $\varepsilon > 0$, there exists a trigonometric polynomial $P(e^{it}) = \sum_{k=-N}^{N} c_k e^{ikt}$ such that $\|f - P\|_{\sup} < \varepsilon$.*

Proof. Define $\hat{f}(k) = \int_I f(t)e^{-ikt}\frac{dt}{2\pi}$ and set

$$\begin{aligned}\sigma_N f(t) &= \sum_{k=-N}^{N} \left(1 - \frac{|k|}{N+1}\right) \hat{f}(k) e^{ikt} \\ &= \sum_{k=-N}^{N} \left(1 - \frac{|k|}{N+1}\right) e^{ikt} \int_I f(s) e^{-iks} \frac{ds}{2\pi} \\ &= \int_I f(s) K_N(t-s) ds\end{aligned}$$

where the *Fejér kernel* K_N is defined as

$$K_N(u) = \sum_{k=-N}^{N} \left(1 - \frac{|k|}{N+1}\right) e^{iku} = \frac{1}{N+1} \frac{\sin^2\left(\frac{N+1}{2}u\right)}{\sin^2\left(\frac{u}{2}\right)}$$

The key observations about K_N (use the two forms of K_N whichever is convenient)

$$K_N(u) \geq 0 \text{ for all } u, \qquad \int\limits_I K_N(u)\frac{du}{2\pi} = 1, \qquad \int\limits_{I\setminus[-\delta,\delta]} K_N(u)\frac{du}{2\pi} \leq \frac{1}{N+1}\frac{1}{\sin^2(\delta/2)}.$$

In probabilistic language, $K_N(\cdot)$ is a probability density on I which puts most of its mass near 0 (for large N). Therefore,

$$\begin{aligned} |\sigma_N f(t) - f(t)| &\leq \int_{-\delta}^{\delta} |f(t)-f(s)|K_N(t-s)ds + \int_{I\setminus[-\delta,\delta]} |f(t)-f(s)|K_N(t-s)ds \\ &\leq \omega_f(\delta) + 2\|f\|_{\sup}\frac{1}{N+1}\frac{1}{\sin^2(\delta/2)}. \end{aligned}$$

Pick δ so that $\omega_f(\delta) < \varepsilon/2$ and then pick $N+1 > \frac{4\|f\|_{\sup}}{\varepsilon\sin^2(\delta/2)}$ to get $\|\sigma_N f - f\|_{\sup} < \varepsilon$. ■

Some applications: In the following exercise, derive Weierstrass' theorem from Fejér's theorem.

Exercise 2. Let $f \in C_{\mathbb{R}}[0,1]$.

(1) Construct a function $g : [-\pi,\pi] \to \mathbb{R}$ such that (a) g is even, (b) $g = f$ on $[0,1]$ and (c) g vanishes outside $[-2,2]$.

(2) Invoke Fejér's theorem to get a trigonometric polynomials T such that $\|T - g\|_{\sup} < \varepsilon$.

(3) Use the series $e^z = \sum_{k=0}^{\infty} \frac{1}{k!}z^k$ to replace the exponentials that appear in T by polynomials. Be clear about the uniform convergence issues.

(4) Conclude that there exists a polynomial P with *real* coefficients such that $\|f - P\|_{\sup} < 2\varepsilon$..

A more interesting application is the theorem of Weyl that the set $\{nx \pmod 1\}$ is equidistributed in $[0,1]$ whenever x is irrational. You are guided to prove this statement in the following exercise.

Exercise 3. Let $x \in [0,1]$. Let $x_n = e^{2\pi i n x}$ and $S = \{x_1, x_2, \ldots\}$.

(1) Show that S is dense in S^1 if and only if x is irrational.

(2) If $f \in C(S^1)$, show that $\frac{1}{n}\sum_{k=1}^{n} f(x_k) \to \int_0^1 f(e^{it})\frac{dt}{2\pi}$. **[Hint:** First do it for $f(e^{it}) = e^{2\pi i p t}$ for some $p \in \mathbb{Z}$]

(3) For any arc $I = \{e^{it} : a < t < b\}$, show that as $n \to \infty$,

$$\frac{1}{n}\#\{k \leq n : x_k \in I\} \to \frac{b-a}{2\pi}.$$

The point is that the points $x_1, x_2, ...$ spend the same amount of "time" in any arc of a given length. This is what we mean by *equidistribution*.

From Fejér's theorem, it follows that $\{e_n : n \in \mathbb{Z}\}$ is an orthonormal basis for $L^2(S^1)$, hence $S_n^f \to f$ in L^2 for every $f \in L^2$ and

$$\int_0^{2\pi} |f(t)|^2 \frac{dt}{2\pi} = \sum_{n\in\mathbb{Z}} |\hat{f}(n)|^2 \qquad \text{(Plancherel identity)}.$$

But for $f \in C[0,1]$, the convergence need not be uniform (or even pointwise). But with extra smoothness assumption on f, one can achieve uniform convergence.

Exercise 4. Let $f \in C^2(S^1)$ (i.e., as a 2π-periodic function on $\mathbb{R}$, f is twice differentiable and f'' is continuous and 2π-periodic). Then, show that $S_n^f \to f$ uniformly. [**Hint:** Express Fourier coefficients of f' in terms of Fourier coefficients of f]

Question: Could we have proved this exercise first, and then used the density of $C^2(S^1)$ in $C(S^1)$ (in fact $C^\infty(S^1)$ is also dense in $C(S^1)$) to get an alternate proof of Fejér's theorem?

A brief history of Fejér's theorem: This is a cut-and-dried history, possibly inaccurate, but only meant to put things in perspective!

(1) The vibrating string problem is an important PDE that arose in mathematical physics, and asks for a function $u : [a,b]\times\bar{\mathbb{R}}_+ \to \mathbb{R}$ satisfying $\frac{\partial^2}{\partial t^2}u(x,t) = \frac{\partial^2}{\partial x^2}u(x,t)$ for $(x,t) \in (a,b)\times\mathbb{R}_+$ and satisfying the initial conditions $u(x,0) = f(x)$ and $\frac{\partial}{\partial t}u(x,t)\Big|_{t=0} = g(x)$, where f and g are specified initial conditions.

(2) Taking $[a,b] = [-\pi,\pi]$ (without loss of generality), it was observed that if $f(x) = e^{ikx}$ and $g(x) = e^{i\ell x}$, then $u(x,t) = \cos(kt)e^{ikx} + \frac{1}{\ell}\sin(\ell t)e^{i\ell x}$ solves the problem.

(3) Linearity of the system meant that if f and g are trigonometric polynomials, then by taking linear combinations of the above solution, one could obtain the solution to the vibrating string problem.

(4) Thus, the question arises, whether given f and g we can approximate them by trigonometric polynomials (and hopefully the corresponding solutions will be approximate solutions).

(5) Fourier made the fundamental observation that $e_k(\cdot)$ are orthonormal on $[-\pi,\pi]$ and deduced that if the notion of approximation is in mean-square sense (i.e., in the L^2 distance $\sqrt{\int |f-g|^2}$), then the best degree-n trigonometric polynomial approximation to f is

$$S_n f(x) := \sum_{k=-n}^{n} \hat{f}(k)e^{ikx}.$$

(6) Before Fejér, it was an open question whether $\|S_n f - f\|_{L^2} \to 0$ as $n \to \infty$. In other words, is $\{e_k\}_{k\in\mathbb{Z}}$ a complete orthonormal set for $L^2([-\pi,\pi])$?

(7) Since continuous functions are dense is $L^2[-\pi,\pi]$, it suffices to show that continuous functions can be uniformly approximated by trigonometric polynomials.

(8) In $C(S^1)$, it is no longer the case that $S_n f$ is the best approximation (in sup-norm sense). Fejér's innovative idea was to consider averages of $S_n f$, i.e., $\sigma_n f := \frac{1}{2n+1}\sum_{k=0}^{2n} S_k f$ (the same trigonometric polynomials that appeared in the proof) and show that they do converge to f uniformly.

(9) Both $S_n f$ and $\sigma_n f$ can be written as convolutions. Indeed, we saw that $\sigma_n f = f \star K_n$ while $S_n f = f \star D_n$ with $D_n(t) = \sin((n+\frac{1}{2})t)/\sin(\frac{1}{2}t)$. The key properties of K_n, that (a) $K_n \geq 0$, (b) $\int K_n = 1$ and (c) $\int_{[-\delta,\delta]^c} K_n \to 0$ as $n \to \infty$, ensured that $f \star K_n \to f$ in $C(S^1)$. Hence K_n is called an *approximate identity* (one can make up many approximate identities but in Fejér's theorem it was essential that K_n is a trigonometric polynomial). The Dirichlet kernel D_n is also a trigonometric polynomial and has total integral 1. But it is not positive, and more importantly, $\int |D_n(t)|dt \to \infty$ as $n \to \infty$ (in fact it grows like $\log n$). This is why it does not act as an approximate identity, and $S_n f$ does not converge to f uniformly.

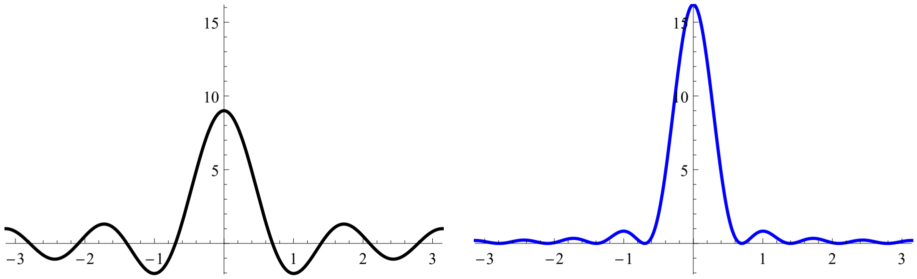

FIGURE 1. The Dirichlet and Fejér kernels for $n = 4$

(10) Here is an argument why $S_n(f)$ converges to f uniformly in general. If uniform convergence were to hold for all $f \in C(S^1)$, then by the uniform boundedness principle applied to the linear transformations $S_n : C(S^1) \mapsto C(S^1)$, it would follow that S_n is point-wise bounded and hence uniformly bounded, $\|S_n f\| \leq \kappa \|f\|$ for all $f \in C(S^1)$ and for all n, for a constant κ. However, if we take $f \in C(S^1)$ such that $-1 \leq f \leq 1$ and $f = \mathrm{sgn}(D_n)$ except on intervals of length $\varepsilon/2n$ around each zero of D_n, then $S_n f(0) \geq (\int |D_n|) - \varepsilon$ while $\|f\|_{\sup} = 1$. Thus, $\|S_n\| = \int |D_n| \asymp \log n$, which is unbounded.

References and further reading: Weierstrass' theorem is generalized to more abstract forms such as Stone-Weierstrass theorem.

Theorem 5 (Stone Weierstrass theorem)**.** *Let X be a compact Hausdorff space and let $\mathcal{A} \subseteq C_{\mathbb{R}}(X)$. If $\mathcal{A}$ is (a) a real vector space, (b) closed under multiplication, (c) contains constant functions and (d) separates points of X. Then, $\mathcal{A}$ is dense in $C(X)$ in sup-norm.*

The separation condition means that for any distinct points $x, y \in X$, there is some $f \in \mathcal{A}$ such that $f(x) \neq f(y)$. If that were not the case, then no function that gives distinct values to x and y could be approximated by elements of $\mathcal{A}$. A proof based on Weierstrass' theorem can be found in most analysis books. Observe that the Stone-Weierstrass theorem aimplies Fejér's theorem too.

Weyl's equidistribution theorem mentioned here is the simplest one. Weyl showed also that for any real polynomial $p(\cdot)$, the sequence $\{e^{2\pi i p(n)} : n \geq 1\}$ is equidistribted in S^1 whenever at least one of the coefficients of p (other than the constant coefficient) is irrational.

Some references:

(1) B.Sury, *Weierstrass' theorem - leaving no stone unturned*, a nice expository article on Weierstrass' theorem available at http://www.isibang.ac.in/~sury/hyderstone.pdf.

(2) Rudin, *Principles of mathematical analysis* or Simmon's *Topology and modern analysis* for a proof of Stone-Weierstrass' theorem.

(3) Katznelson, *Harmonic analysis* or many other book on Fourier series for basics of Dirichlet and Fejér kernels.

3. Müntz-Szasz theorem in L^2

Theorem 1. *Let $0 \leq n_1 < n_2 < \ldots$ be unbounded and let $W = span\{x^{n_j} : j \geq 1\}$. Then, W is dense in $L^2[0,1]$ if and only if $\sum_j \frac{1}{n_j} = \infty$.*

Almost exactly the same criterion is necessary and sufficient for W to be dense in $C[0,1]$, except that for uniform approximation we must take $n_1 = 0$ (otherwise functions not vanishing at 0 cannot be approximated). In the above theorem, n_j are not required to be integers. From the above theorem, it is easy to deduce that if $\sum_j \frac{1}{n_j} < \infty$, then W cannot be dense in $C[0,1]$. This is simply beacause $\|f\|_{L^2} \leq \|f\|_{\sup}$ for any $f \in C[0,1]$.

Some preliminaries in linear algebra: Let V be an inner product space over $\mathbb{R}$ and let $v_1, \ldots, v_k$ be elements of V. The Gram matrix of these vectors is the $k \times k$ matrix $A := (\langle v_i, v_j \rangle)_{i,j \leq k}$ whose entries are inner products of the given vectors. If $V = \mathbb{R}^k$ itself (the same k as the number of vectors), then $A = B^t B$ where $B = [v_1 \ldots v_k]$ is the $k \times k$ matrix whose columns are the given vectors. In this case, $\det(A) = \det(B)^2$ which is the squared volume of the parallelepiped formed by $v_1, \ldots, v_k$ (because $\det(B)$ is the signed volume of this parallelepiped). Convince yourself that even for general V, $\det(A)$ has the same meaning (but $\det(B)$ need not make sense, for example, if $V = \mathbb{R}^m$ with some $m > k$).

Now, let $u, v_1, \ldots, v_k$ be vectors in V. Let A be the Gram matrix of these $k+1$ vectors, and let B be the Gram matrix of $v_1, \ldots, v_k$. Using the above-mentioned volume interpretation of the determinants and the formula "volume = base volume × height" formula, we see that

$$\det(A) = \det(B) \times \text{dist.}^2(u, \text{span}\{v_1, \ldots, v_k\}).$$

Here the dist.^{2} term just means $\|P_W^{\perp} u\|^2$ where $P_W^{\perp}$ is the orthogonal projection to $W^{\perp}$ where $W = \text{span}\{v_1, \ldots, v_k\}$.

Example 2. Let $n_0, n_1, \ldots, n_k$ be distinct positive numbers and let $u = x^{n_0}, v_1 = x^{n_1}, \ldots, v_k = x^{n_k}$, all regarded as elements of $L^2[0,1]$. Let $W_k = \text{span}\{v_1, \ldots, v_k\}$. Then,

$$\text{dist.}^2(u, W_k) = \frac{\det(A)}{\det(B)}$$

where $A = \left(\frac{1}{n_i+n_j+1}\right)_{0 \le i,j \le k}$ and $B = \left(\frac{1}{n_i+n_j+1}\right)_{1 \le i,j \le k}$. The matrices here are called Hilbert matrices, and their determinants can be evaluated explicitly.

Cauchy determinant identity: Let $x_1, \ldots, x_k$ be distinct and $y_1, \ldots, y_k$ be distinct (we take them to be real numbers, but the same holds over any field). Then,

$$\det\left(\frac{1}{x_i + y_j}\right)_{1 \le i,j \le k} = \frac{\prod\limits_{i<j}(x_i - x_j)(y_i - y_j)}{\prod\limits_{i,j}(x_i + y_j)}.$$

To see this, observe that

$$\prod_{i,j}(x_i + y_j) \det\left(\frac{1}{x_i + y_j}\right)_{1 \le i,j \le k}$$

is a polynomial in x_is and y_js of degree at most $n^2 - n$, and vanishes whenever two of the x_is are equal or two of the y_js are equal. Hence, the polynomial is divisible by $\prod\limits_{i<j}(x_i - x_j)(y_i - y_j)$. The latter is a polynomial of degree $n(n-1)$, hence we conclude that

$$\prod_{i,j}(x_i + y_j) \det\left(\frac{1}{x_i + y_j}\right)_{1 \le i,j \le k} = C \prod_{i<j}(x_i - x_j)(y_i - y_j)$$

for some constant C. How to see that $C = 1$?

Proof of Theorem 1. Let $0 \le n_0 \notin \{n_1, n_2, \ldots\}$ and set $u = x^{n_0}, v_1 = x^{n_1}, \ldots, v_k = x^{n_k}$, all regarded as elements of $L^2[0,1]$. As already explained,

$$\text{dist.}^2(u, W_k) = \frac{\det(A)}{\det(B)}$$

where $A = \left(\frac{1}{n_i+n_j+1}\right)_{0\le i,j\le k}$ and $B = \left(\frac{1}{n_i+n_j+1}\right)_{1\le i,j\le k}$. Cauchy's identity applies to both these determinants (take $x_i = y_i = n_i + \frac{1}{2}$) and hence, after canceling a lot of terms,

$$\begin{aligned}\text{dist.}^2(u, W_k) &= \frac{\prod_{i=1}^k (n_0 - n_j)^2}{(2n_0+1)\prod_{j=1}^k (n_0+n_j+1)^2} = \frac{1}{2n_0+1}\prod_{j=1}^k \left(1 - \frac{n_0}{n_j}\right)^2 \left(1 + \frac{n_0+1}{n_j}\right)^{-2} \\ &= \frac{1}{2n_0+1}\prod_{j=1}^k \left(1 - \frac{A}{n_j} + O\left(\frac{1}{n_j^2}\right)\right)\end{aligned}$$

where $A \ne 0$. Recall that for $0 < x_j < 1$, the infinite product $\prod_{j=1}^\infty (1-x_j)$ is positive if $\sum_j x_j < \infty$ and zero if $\sum_j x_j = \infty$. From this, it immediately follows that

$$\lim_{k\to\infty} \text{dist.}^2(u, W_k) = 0 \text{ if and only if } \sum_j \frac{1}{n_j} = \infty.$$

Thus, if $\sum_j \frac{1}{n_j} < \infty$, then if we take $n_0 \notin \{n_1, n_2, \ldots\}$, it follows that x^{n_0} is not in the closed span of $W = \{x^{n_j} : j \ge 1\}$. In particular, W is not dense in $L^2[0,1]$.

Conversely, if $\sum_j \frac{1}{n_j} = \infty$, then for any $n_0 > 0$, we see that $x^{n_0} \in \bar{W}$. Thus all polynomials are in $\bar{W}$ which shows that $\bar{W} = L^2[0,1]$. ■

Hilbert arrived at the Hilbert matrix in studying the following question. How closely can x^n be approximated (in $L^2[0,1]$) by polynomials of lower degree? This just means finding

$$r_n = \text{dist.}(x^n, \text{span}\{1, x, \ldots, x^{n-1}\}).$$

The corresponding question in $C[0,1]$ is much deeper and was (asked and) answered by Chebyshev. We shall see it later.

Exercise 3. Find an explicit form of r_n. How big or small (i.e., the order of decay/growth) is it?

4. MÜNTZ-SZASZ THEOREM IN $C[0,1]$

Theorem 1. *Let $0 = n_0 < n_1 < n_2 < \ldots$ be unbounded and let $W = span\{x^{n_j} : j \ge 0\}$. Then, W is dense in $C[0,1]$ if and only if $\sum_{j\ge1} \frac{1}{n_j} = \infty$.*

The natural approach, would be analogous to the one we gave for L^2 approximation. The difference is that L^2 is a Hilbert space (self-dual) and $C[0,1]$ is a Banach space. Hence, the denseness in $C[0,1]$ of a subspace (here span$\{x^{n_j} : j \ge 0\}$ where $0 = n_0 < n_1 < n_2 < \ldots$) is detected by the absence of non-zero bounded linear functionals that vanish on the subspace. We give a sketch of a proof along these lines later. First we give an elementary, but possibly less natural, proof.

Another proof (from notes of Andreu Ferre Moragues). Fix $m > 0$ and let $f(x) = x^m$ so that $f'(x) = mx^{m-1}$ vanishes at 0. Fix j_0 so that $n_{j_0} > 1$ and set $W' = \text{span}\{x^{n_j-1} : j \ge j_0\}$. By the L^2 version of Müntz-Szasz theorem, W' is dense in $L^2[0,1]$, and hence we can find $Q_m \in W'$ such that $\|f' - Q_r\|_2 \to 0$ as $r \to \infty$.

Set $P_r(x) = \int_0^x Q_r(t)dt$. If $Q_r(t) = \sum_{j=j_0}^{j_0+d_r} c_{r,j} x^{n_j - 1}$, then $P_n(x) = \sum_{j=j_0}^{j_0+d_r} \frac{c_{r,j}}{n_j} x^{n_j}$, hence $P_n \in W$. Note that P_n and f both vanish at 0, hence, for any $x \in [0,1]$,

$$|f(x) - P_n(x)| = |\int_0^x (f'(t) - Q_n(t))dt| \le \int_0^1 |f'(t) - Q_n(t)|dt \le \sqrt{\int_0^1 |f'(t) - Q_n(t)|^2 dt}$$

by Cauchy-Schwarz inequality. Thus, $\|f - P_n\| \le \|f' - Q_n\|_2 \to 0$. Thus $f \in \overline{W}$, showing that W is dense in $C[0,1]$. ■

In the notes of Andreu Ferre Moragues referred to at the end, there is yet another proof of the above, which does not rely on the L^2 version.

4.1. **Sketch of a second proof.** Here we sketch the other natural approach that was alluded to earlier. We need the following ingredients:

(1) A subspace W of a Banach space X is dense if and only if there does not exist any bounded linear functional $L : X \to \mathbb{C}$ such that $L \neq 0$ but $L|_W = 0$.

(2) The space of real-valued continuous functions, $C[0,1]$, is a Banach space and its dual is $\mathcal{M}[0,1]$, the space of signed Borel measures on $[0,1]$. Elements of $\mathcal{M}_{\mathbb{C}}[0,1]$ are precisely of the form $\mu = \mu_1 - \mu_2$, where μ_j are finite, positive, Borel measures on $[0,1]$. If $\mu = \mu_1 - \mu_2$, then it acts on $C[0,1]$ by $f \mapsto \int f d\mu_1 - \int f d\mu_2$.

(3) The representation of a signed measure μ as $\mu_1 - \mu_2$ is not unique, as we can also write it as $(\mu_1 + \nu) - (\mu_2 + \nu)$ for any measure ν. Uniqueness of representation[2] can be obtained by imposing the further condition that $\mu_1 \perp \mu_2$ (singular measures). If we use this minimal representation, it is easy to check that the dual norm of μ is precisely $\mu_1([0,1]) + \mu_2([0,1])$.

Sketch of the proof. Let $W = \text{span}\{x^{n_j} : j \ge 0\}$. Recall that W is not dense in $C[0,1]$ if and only if there is a non-zero bounded linear functional on $C[0,1]$ that vanishes on W (by Hahn-Banach theorem). We know that the dual of $C[0,1]$ is the space of all complex Borel measures on $[0,1]$, acting by $f \mapsto \int_{[0,1]} f d\mu$ (one of F. Riesz's many representation theorems). Thus, W is not dense if and only if we can find a complex Borel measure μ on $[0,1]$ such that $\int t^{n_j} d\mu(t) = 0$ for all j.

For any μ, consider the function $F_\mu(z) = \int t^z d\mu(t)$. This is a holomorphic function on the right half-plane. A question is whether it can vanish at n_j for all j, without μ being identically zero. A holomorphic function can have no accumulation points inside the domain of holomorphicity, but there is no restriction on vanishing at a sequence of points that go to the boundary (or infinity). However, if there are some bounds on the growth of the holomorphic function, then its sequence of zeros must approach the boundary sufficiently fast.

We skip details for now, but what it amounts to is that when $\sum_j \frac{1}{n_j} = \infty$, such functions do not exist. Consequently W is dense in $C[0,1]$. ■

[2] This is analogous to how any function f can be written as a difference of two positive functions in many ways, but the minimal way is to write it as $f_+ - f_-$.

5. Mergelyan's theorem

On a compact subset K of the complex plane, what functions can be uniformly approximated by polynomials? Two examples to show what can go wrong.

Let $K = \bar{\mathbb{D}} = \{z : |z| \le 1\}$. Then $\bar{z}$ cannot be uniformly approximated by polynomials. This is because a uniform limit of polynomials must be holomorphic in the open disk $\mathbb{D}$. Thus not all continuous functions can be uniformly approximated by polynomials.

What about analytic functions? If $K = 2\bar{\mathbb{D}} \setminus \mathbb{D} = \{z : 1 \le |z| \le 2\}$, then the function $1/z$ cannot be uniformly approximated on K by polynomials. This is because polynomials integrate to zero on contours in the interior of K, but $\int_\gamma \frac{1}{z} dz \neq 0$ if γ has non-zero winding around 0.

Mergelyan's theorem gives the complete answer to the question. Let $\mathcal{A}(K)$ be the space of continuous functions on K that are holomorphic in the interior of K. Endow it with the sup-norm on K.

Theorem 1 (Mergelyan)**.** *Let K be a compact subset in the complex plane such that $\mathbb{C} \setminus K$ has finitely many connected components. Choose points $p_1, \dots, p_m$, one in each of the bounded components of $\mathbb{C} \setminus K$. Let $\mathcal{R}$ be the collection of all rational functions whose poles are contained in $\{p_1, \dots, p_m\}$.*

Then $\mathcal{R}$ is dense in $\mathcal{A}(K)$. In particular, if $\mathbb{C} \setminus K$ is connected, then polynomials are dense in $A(K)$.

For example, if K is the closure of a bounded simply connected region, then all continuous functions that are holmorphic in the interior can be approximated uniformly by polynomials. If $K = [0, 1]$ (or any curve $\gamma : [0, 1] \mapsto \mathbb{C}$ that is injective), then the interior is empty and $\mathbb{C} \setminus K$ is connected. Hence the analyticity condition is superfluous and all continuous functions are approximable by polynomials. If $K = S^1$, again the interior is empty but $\mathbb{C} \setminus K$ has one bounded component $\mathbb{D}$. Taking $p = 0$ and applying Mergelyan's theorem gives us Fejér's theorem.

As the proof of Mergelyan's theorem uses certain advanced theorems in complex analysis, we shall postpone it to later.

6. Chebyshev's approximation question

For $f \in C[0, 1]$, $n \ge 1$ and $a < b$, let

$$\gamma(f, n, a, b) := \inf\{\|f - p\|_{\sup} : p \text{ is a polynomial of degree at most } n\}.$$

Weierstrass' theorem is the statement that $\gamma(f, n) \to 0$ as $n \to \infty$. But what is the rate at which it goes to zero? Equivalently, for a given n, how good is the approximation? Try to work out the bound you get from the proofs we gave of Weierstrass' theorem. In a landmark paper, Chebyshev showed that $\gamma(x^n, n-1, -1, 1) = 2^{-n+1}$.

For instance, if we use x^{n-1} to approximate x^n, then,

$$\|x^n - x^{n-1}\|_{\sup[-1,1]} \ge \left(1 - \frac{1}{n}\right)^n - \left(1 - \frac{1}{n}\right)^{n-1} = \left(1 - \frac{1}{n}\right)^{n-1} \frac{1}{n} \sim e^{-1} n^{-1}.$$

But we shall see that it is possible to find degree $n-1$ polynomials that are exponentially close to x^n. To do this, he introduced an immortal class of polynomials, now known as *Chebyshev polynomials* (of the first kind).

In basic trigonometry, we see that

$$\cos(2\theta) = 2\cos^2(\theta) - 1, \quad \cos(3\theta) = 4\cos^2(\theta) - 3\cos(\theta).$$

It is not hard to see that in general, $\cos(n\theta)$ is a polynomial of degree n in $\cos(\theta)$. Thus, $\cos(n\theta) = T_n(\cos\theta)$, where T_n is defined to be the nth Chebysev polynomial (of the first kind). For instance, $T_2(x) = 2x^2 - 1$ and $T_3(x) = 4x^3 - 3x$.

By the identity $\cos((n+1)\theta) + \cos((n-1)\theta) = 2\cos(\theta)\cos(n\theta)$, we see that $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$. In fact, this recursion, together with the specification $T_0(x) = 1$ and $T_1(x) = x$, could be taken as an alternative definition of the Chebyshev polynomials.

Some easy properties: T_n has degree n. The coefficient of x^k in T_n is zero unless $n-k$ is even. The highest coefficient of T_n is 2^{n-1}. Lastly $\|T_n\|_{\sup[-1,1]} = 1$.

Consequently, $|x^n - p(x)| \le 1$ for all $x \in [-1,1]$, where $p(x) = x^n - 2^{-n+1}T_n(x)$ is a polynomial of degree $n-1$. Therefore, $\gamma(x^n, n-1, -1, 1) \le 2^{-n+1}$. Chebyshev's theorem is that $2^{-n+1}T_n$ is the best approximation to x^n among lower degree polynomials. We shall prove it shortly.

A less obvious looking property of Chebyshev polynomials is that they are orthogonal w.r.t. the arcsine measure $d\mu(x) = \frac{1}{\pi\sqrt{1-x^2}}dx$ on $[-1,1]$. That is,

$$\int_{-1}^{1} T_n(x)T_m(x)\frac{1}{\pi\sqrt{1-x^2}}dx = 0 \quad \text{if } m \neq n.$$

To do this without calculations, define the map $\varphi : S^1 \mapsto [-1,1]$ by $\varphi(e^{i\theta}) = \cos\theta$. The arcsine measure is precisely the push-forward of the normalized Lebesgue measure on S^1 under φ. Further, $T_k \circ \varphi = \text{Re}\{e_k\}$. From this, it easily follows that

$$\int_{-1}^{1} T_m(x)T_n(x)\frac{1}{\pi\sqrt{1-x^2}}dx = \begin{cases} 1 & \text{if } m = n = 0 \\ \frac{1}{2} & \text{if } m = n \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Relevance to the approximation question: Any $x \in [-1,1]$ can be written as $x = \cos\theta$, hence $T_n(x) = \cos(n\theta) \in [-1,1]$. Thus, $\|T_n\|_{\sup} = 1$. The monic polynomial $2^{-n+1}T_n$ has sup-norm 2^{-n+1} in $[-1,1]$. Write $p(x) = x^n - 2^{-n+1}T_n(x)$, a polynomial of degree $n-2$. Then,

$$\|x^n - p_n(x)\|_{\sup[-1,1]} = 2^{-n+1}\|T_n\|_{\sup[-1,1]} = 2^{-n+1}.$$

This is much smaller than the $1/n$ that we got earlier by approximating x^n by x^{n-1}. We next show that this is the best possible.

The oscillation idea: Let $f \in C[-1,1]$, let p be a polynomial of degree n, and $\delta > 0$. Suppose $f - p$ oscillates between $\pm\delta$ many times, say m. By this we mean that there exist $x_1 < x_2 < \ldots < x_{m+1}$ in $[-1,1]$ such that $|f(x_j) - p(x_j)| \geq \delta$ for all j and $\operatorname{sgn}(f(x_j) - p(x_j))$ alternates between $+1$ and -1 as j runs from 1 to $m+1$.

Now suppose q is another polynomial of degree n and $\|f - q\|_{\sup} < \delta$. Then, $\operatorname{sgn}(q(x_j) - p(x_j))$ alternates between $+1$ and -1 as j runs from 1 to $m+1$. Indeed, suppose $f(x_1) - p(x_1) \geq \delta$ and $f(x_2) - p(x_2) \leq -\delta$. Then $q(x_1) > f(x_1) - \delta \geq p(x_1)$ and $q(x_2) < f(x_2) + \delta \leq p(x_2)$. This shows that $q - p$ must have at least m roots, one in (x_j, x_{j+1}) for each $1 \leq j \leq m$.

If $m \geq n+1$, this is not possible, as $q - p$ has degree at most n. The way out of the contradiction is that $\|f - q\|_{\sup[-1,1]} \geq \delta$ for every degree n polynomial q. We collect this conclusion as a lemma below.

Lemma 1. *If $f \in C[0,1]$, p is a polynomial of degree n, and there exist $n+2$ points $x_1 < x_2 < \ldots < x_{n+2}$ in $[-1,1]$ such that $|f(x_j) - p(x_j)| \geq \delta$ for all j and $\operatorname{sgn}(f(x_j) - p(x_j))$ alternates between $+1$ and -1 as j runs from 1 to $n+2$. Then, for any polynomial q of degree n or less, $\|f - q\|_{\sup[-1,1]} \geq \delta$.*

Chebyshev polynomial is the best approximation to the monomial: Let $f(x) = x^n$, $p(x) = x^n - 2^{-n+1}T_n(x)$ (a degree $n-1$ polynomial) and $\delta = 2^{-n+1}$. Note that $f(x) - p(x) = 2^{-n+1}T_n(x)$. Write $x = \cos\theta$ and let θ range over $[0,\pi]$ (so x runs through $[-1,1]$). Recall that $T_n(\cos\theta) = \cos(n\theta)$, take $\theta_k = k\pi/n$ for $k = 0, 1, \ldots, n$, and note that $T_n(\cos\theta_k) = (-1)^k$. Thus, $f - p$ alternates $n+1$ times between $\pm\delta$. From Lemma 1, we conclude that for any polynomial q of degree $n-1$ or less, $\|f - q\|_{\sup[-1,1]} \geq 2^{-n+1}$.

An application: We have found the best way to approximate x^n by a polynomial of lower degree. In principle, replacing the highest power in the approximating polynomial by a lower degree Chebyshev polynomial, and continuing, it should be possible to reduce the degree of the approximating polynomial while keeping a reasonable level of approximation. This raises the question,[3] how small a degree m can we take and still approximate x^n (on $[-1,1]$) by a degree m polynomial?

First we find the expansion of x^n in terms of Chebyshev polynomials (this is obviously possible since T_k has degree k for each k). If

$$x^n = \sum_{k=0}^{n} c_{n,k} T_k(x),$$

[3] This part of the notes is taken from Nisheet Vishnoi's notes which contains more on these problems and their uses in algorithms. The derivation of expansion of x^n in terms of Chebyshev polynomials given here was suggested by Chaitanya Tappu, and is more natural than what we did in class.

then we must have

$$c_{n,0} = \int_{-1}^{1} x^n T_0(x) \frac{dx}{\pi\sqrt{1-x^2}} dx,$$
$$c_{n,k} = 2\int_{-1}^{1} x^n T_k(x) \frac{dx}{\pi\sqrt{1-x^2}} dx \quad \text{for } k \geq 1,$$

by the orthogonality of T_ks with respect to the arcsine measure. Make the change of variables $x = \cos\theta$ to write

$$\begin{aligned}\int_{-1}^{1} x^n T_k(x) \frac{dx}{\sqrt{\pi(1-x^2)}} dx &= \frac{1}{\pi}\int_0^{\pi} \cos^n\theta \times \cos(k\theta) d\theta \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \left(\frac{e^{i\theta}+e^{-i\theta}}{2}\right)^n \left(\frac{e^{ik\theta}+e^{-ik\theta}}{2}\right) d\theta \\ &= \frac{1}{2^n}\binom{n}{\frac{n+k}{2}}.\end{aligned}$$

To see the last inequality, expand $(e^{i\theta}+e^{-i\theta})^n$ and observe that only the terms with $e^{\pm ik\theta}$ give a non-zero integral. There are two of them, each with the binomial coefficient. If we write $p_{n,k} = 2^{-n}\binom{n}{(n+k)/2}$ for $k = n, n-2, \ldots, -n+2, -n$, then $c_{n,0} = p_{n,0}$ and $c_{n,k} = p_{n,k} + p_{n,-k}$.

The basic observation is that $(p_{n,k})_k$ is the Binomial probability distribution (the distribution of $\xi_1 + \ldots + \xi_n$, where ξ_k are independent and equal to ± 1 with probability $1/2$ each). Most of the mass of this distribution is concentrated in $|k| \lesssim \sqrt{n}$. For example, the Chebyshev inequality gives

$$\sum_{k:|k|\geq d} p_{n,k} \leq \frac{n}{4d^2}$$

which becomes small when $d \gg \sqrt{n}$. A better bound is the following

$$\text{Bernstein/Chernoff bound: } \sum_{k:|k|\geq d} p_{n,k} \leq 2e^{-d^2/2n}.$$

Thus, if we set $P_{n,d}(x) = \sum_{k=0}^{d} c_{n,k} T_k(x)$, then

$$\begin{aligned}\|x^n - P_{n,d}\|_{\sup[-1,1]} &= \Big|\sum_{k=d+1}^{n} c_{n,k} T_k(x)\Big| \\ &\leq 2\sum_{k:k>d} p_{n,k} \qquad (\text{as } \|T_k\|_{\sup[-1,1]} = 1) \\ &\leq 2e^{-d^2/2n}\end{aligned}$$

by the Chernoff bound. Thus, we can approximate x^n well by polynomials of degree d provided d is much larger than $\sqrt{n}$. For example, if $d = \sqrt{2Bn\log n}$, then $\|x^n - P_{n,d}\|_{\sup[-1,1]} \leq n^{-B}$.

This finishes our discussion of approximation questions. Much more can be found in the references we mentioned earlier.

Exercise 2. Show that $T_n(x) = \frac{(-1)^n}{1\times 3\times\ldots\times(2n-1)}\sqrt{1-x^2}\frac{d^n}{dx^n}(1-x^2)^{n-\frac{1}{2}}$.

One can also take this as the definition and prove the other properties (recursion, orthogonality, etc.). The following exercise introduces Chebyshev polynomials of the second kind.

Exercise 3. Argue that $\frac{\sin((n+1)\theta)}{\sin\theta}$ is a polynomial of $\cos\theta$. Hence define the polynomials U_n, $n \geq 0$ by $U_n(\cos\theta) = \frac{\sin((n+1)\theta)}{\sin\theta}$. Show that (1) $U_n(x) = \frac{1}{n+1}T'_{n+1}(x)$, (2) $\int_{-1}^{1} U_n(x)U_m(x)\sqrt{1-x^2}dx = 0$ if $m \neq n$.

CHAPTER 2

Equidistribution

1. Weyl's equidistribution theorem

Let $0 < \alpha < 1$ and consider the sequence $x_n = \overline{n\alpha}\}$, where (for this section), $\overline{x}$ denotes x (mod 1), i.e., $x - \lfloor x \rfloor$. How does this sequence inside $[0,1)$ look? If α is a rational number, than after a $x_n = 0$ for some n and then the entire sequence repeats periodically. For irrational α, it is an easy exercise to show that $x_n \neq x_m$ for all $n \neq m$, and that the sequence $\{x_n : n \geq 1\}$ is dense in $[0,1)$. Much more is true.

Theorem 1 (Weyl's equidistribution - the linear case)**.** *Suppose $\alpha \notin \mathbb{Q}$. For any $0 \leq a < b < 1$, as $n \to \infty$,*

$$\frac{1}{n}|\{k \leq n : x_k \in [a,b]\}| \to b - a.$$

In words, the sequence is uniformly distributed over the interval $[0,1)$. It is also worth noting that since we are working with $[0,1)$ with addition modulo 1, it is the same as the circle group S^1 with the isomorphism $x \mapsto e^{2\pi i x}$. Therefore, an equivalent formulation is that the sequence $\{z_k := e^{2\pi i n\alpha} : n \geq 1\}$ is equidistributed on S^1, in the sense that the proportion of $k \leq n$ for which z_k is in the arc $\{e^{i\theta} : a \leq \theta \leq b\}$, converges to $b - a$, as $n \to \infty$. We freely move between the two notations (eg., we may write $f(x)$ or $f(e^{ix})$).

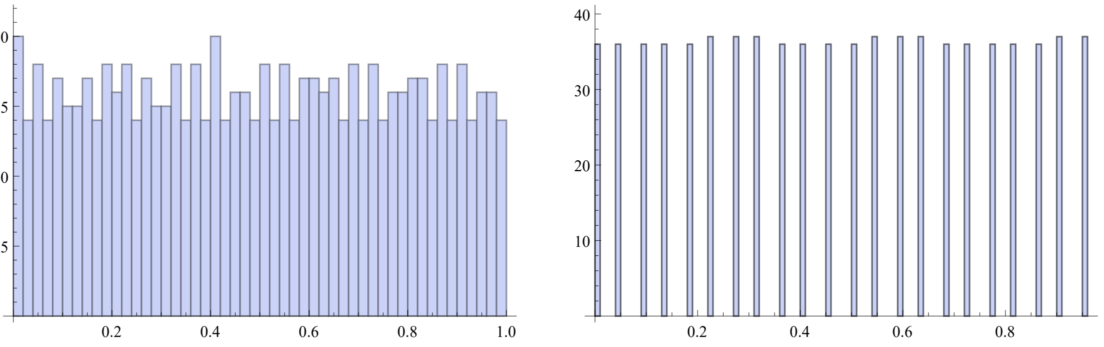

FIGURE 2. Histogram of $\overline{k\alpha}$, $1 \leq k \leq 800$, for $\alpha = 1/\pi$ and $\alpha = 7/22$.

Now we proceed to the proof of the theorem. First of all, it is good to note the virtual impossibility of getting a direct handle on the quantity on the left.

Step-1: The statement of Theorem 1 can be written as

$$\tag{1} \frac{1}{n}\sum_{k=1}^{n} f(\overline{k\alpha}) \to \int_0^1 f(x)dx.$$

where $f = \mathbf{1}_{[a,b]}$ is the indicator function of the interval $[a, b]$. This suggests the question: are there other functions for which one can prove the statement and then deduce it for indicators?

Step-2: Let $e_m(x) = e^{2\pi imx}$ for some $m \in \mathbb{Z}$ (if $m \notin \mathbb{Z}$, then $f(0) \neq f(1)$), then we get lucky because $e^{2\pi im\overline{x}} = e^{2\pi imx}$ and the annoying "modulo 1" operation can be removed. For $m \neq 0$, the sum on the left side of (1) is a geometric series:

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n} e^{2\pi im\overline{k\alpha}} &= \frac{1}{n}\sum_{k=1}^{n} e^{2\pi imk\alpha} \\ &= \frac{1}{n} \times e^{2\pi im\alpha}\frac{1 - e^{2\pi imn\alpha}}{1 - e^{2\pi im\alpha}}.\end{aligned}$$

The last step is possible because $e^{2\pi im\alpha} \neq 1$ (this is the only place where irrationality of α is used). Now, the entire quantity on the right is bounded in absolute value by

$$\frac{1}{n|1 - e^{2\pi im\alpha}|}$$

which goes to 0 as $n \to \infty$. Further, if $f = e_0$, then the sum on the left side of (1) is equal to 1 for any n. Since

$$\int_0^1 e_m(x)dx = \delta_{m,0}$$

we see that (1) is true when $f = e_m$ for any $m \in \mathbb{Z}$. Clearly, it then holds for any finite linear combinations[4] of e_m.

Step-3: On the other side, we observe that if (1) is proved for $f \in C(S^1)$ (this means $f \in C[0, 1]$ with the property that $f(0) = f(1)$), then the same follows for indicator functions. To see this, first show that given $[a, b]$ and any $\varepsilon > 0$, we can find $f, g \in C[0, 1]$ such that (a) $0 \leq f, g \leq 1$, (b) f is supported in $[a + \varepsilon, b - \varepsilon]$, (c) g is supported in $[a - \varepsilon, b + \varepsilon]$, (d) $f \leq \mathbf{1}_{a,b} \leq g$. Here all addition is modulo 1, as noted above. We leave it as an exercise to show that such f, g exist and that one can in fact take f, g to be smooth (we do not need that here).

Then,

$$\frac{1}{n}\sum_{k=1}^{n} f(\overline{k\alpha}) \leq \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{[a,b]}(\overline{k\alpha}) \leq \frac{1}{n}\sum_{k=1}^{n} g(\overline{k\alpha}) \quad \text{and} \quad \int f(x)dx \leq b - a \leq \int g(x)dx$$

[4] Functions of the form $f(x) = \sum_{m=-p}^{p} c_m e_m(x)$ for some p and some $c_m \in \mathbb{C}$, are called *trigonometric polynomials*.

By assumption, that (1) holds for continuous functions, we see that

$$\int f(x)dx \le \liminf_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{[a,b]}(\overline{k\alpha}) \le \limsup_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{[a,b]}(\overline{k\alpha}) \le \int g(x)dx.$$

But $\int g(x)dx - \int f(x)dx \le 4\varepsilon$ (since $g - f \le 1$ and equal to zero except on two intervals on length 2ε centered at a and b). Hence, the $\liminf$ and $\limsup$ in the above inequalities and the number $b - a$, all three are within 4ε of each other. As ε is arbitrary, the limit exists and is equal to $b - a$.

Step-4: From Step-2, we have the result for trigonometric polynomials and by Step-3 we would be done if we had the result for continuous functions on S^1. Another approximation is required, this time provided by Fejér's theorem, which states that if $C(S^1)$ is endowed with the sup-norm metric ($d(f, g) = \max |f(x) - g(x)|$), then trigonometric polynomials form a dense subset. This theorem is discussed at length in the chapter on approximation, but for now you may find it an exercise to derive it from the Stone-Weierstrass' theorem.

Given f, apply Fejér's theorem to find a trigonometric polynomial T such that $\|f - T\|_{\sup} < \varepsilon$. Then,

$$\left|\int f(x)dx - \int T(x)dx\right| < \varepsilon \quad \text{and} \quad \left|\frac{1}{n}\sum_{k=1}^{n} f(\overline{k\alpha}) - \frac{1}{n}\sum_{k=1}^{n} T(\overline{k\alpha})\right| < \varepsilon.$$

Therefore, letting $N \to \infty$ and using the result for T, we see that

$$\int f(x)dx - \varepsilon < \liminf_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} f(\overline{k\alpha}) \le \limsup_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} f(\overline{k\alpha}) < \int f(x)dx + \varepsilon.$$

As ε is arbitrary, the limit of $\frac{1}{n}\sum_{k=1}^{n} f(\overline{k\alpha})$ exists and is equal to $\int f(x)dx$. ■

Exercise 2. Let $\alpha, \beta \in [0, 1)$. Find conditions under which $\{\overline{\alpha n + \beta} : n \ge 1\}$ is equidistributed.

Summary: Steps 1, 3, 4 have general lessons in analysis. One is that for any statement about sets, it is good to find the analogous statement for functions, and vice versa. Second is of approaching problems by approximation - prove results for sufficiently nice functions, and then for more general functions by approximation.

But once we put aside these general lessons, the key step is the use of complex exponentials. This is the subject of Fourier series. More specifically, the idea of studying exponential sums to understand a sequence of numbers is a far reaching one. We shall see more of it in the next section. The use of characteristic functions to prove central limit theorem in probability may be also thought of as an outgrowth of the above use to show equidistribution.

2. WEYL'S EQUIDISTRIBUTION FOR POLYNOMIALS EVALUATED AT INTEGERS

Let $P(x) = \alpha_d x^d + \ldots + \alpha_1 x + \alpha_0$ be a polynomial with real coefficients. What can we say about the equidistribution of the sequence $\overline{P(n)}$, $n \ge 1$? If all the α_k are rational, then $P(n)$ is also rational, and in fact there are only a finite number of values taken by $P(n)$ as n varies over integers. No equidistribution can be hoped for. Further, α_0 is fairly irrelevant, since it shifts the entire sequence and cannot change the equidistribution property. Thus the question is what happens if one of $\alpha_1, \ldots, \alpha_d$ is irrational? One example is the sequence $\overline{\alpha n^2}$, $n \geq 1$.

Theorem 1 (Weyl's equidistribution for polynomials)**.** *Let $P(x) = \alpha_d x^d + \ldots + \alpha_1 x + \alpha_0$ where at least one of $\alpha_1, \ldots, \alpha_d$ is irrational. Then the sequence $\{\overline{P(n)} : n \geq 1\}$ is equidistributed in $[0, 1)$.*

We shall only prove a limited version which indicated the difficulties of the general case.

Theorem 2 (Weyl's equidistribution for quadratics)**.** *Let $P(x) = \alpha x^2 + \beta x + \gamma$ where α is irrational. Then the sequence $\{\overline{P(n)} : n \geq 1\}$ is equidistributed in $[0, 1)$.*

I cannot improve on the presentation of the proof in Sayantan Khan's notes. I recommend reading from those notes and the reference given there.

Presentation topic: Proof of equidistribution for polynomials.

3. Saying it in the language of weak convergence

What we are discussing here is convergence of measures. Recall that by Riesz's representation theorem, the dual of the Banach space $C(S^1)$ (with sup-norm) is the space of all complex Borel measures on S^1. In concrete terms, these are of the form $\mu = \mu_1 - \mu_2 + i\mu_3 - i\mu_4$, where μ_i are finite positive Borel measures (if you want a unique representation like this, then some conditions on singularity of μ_1 and μ_2, etc., must be imposed). It acts on $C(S^1)$ by integrating w.r.t μ, of course.

Recall the notion of weak-* convergence on the dual of a Banach space wherein $L_n \overset{w^*}{\to} L$ if $L_n(x) \to L(x)$ (in $\mathbb{C}$) for all x in the Banach space. On $C(S^1)^*$, this means that $\mu_n \to \mu$ (weakly) if and only if $\int f d\mu_n \to \int f d\mu$ for all $f \in C(S^1)$.

In general, if $\mu_n \to \mu$, it is not true that $\mu_n(A) \to \mu(A)$ (indicator functions are not continuous in general). In fact, it is not difficult to show that $\mu_n(A) \to \mu(A)$ if and only if $\mu(\partial A) = 0$. For example, if μ_n is the uniform probability measure on $[\frac{1}{2} - \frac{1}{n}, \frac{1}{2} + \frac{1}{n}]$, then $\mu_n \to \delta_{\frac{1}{2}}$ (the measure that puts mass 1 at the point $\frac{1}{2}$). However, $\mu_n[0, \frac{1}{2}] = \frac{1}{2}$ does not converge to $\mu[0, \frac{1}{2}] = 1$.

If $\zeta_1, \zeta_2, \ldots$ is a sequence in S^1, then we may form the *empirical measures* $\mu_n = \frac{1}{n}(\delta_{\zeta_1} + \ldots + \delta_{\zeta_n})$ that puts mass $1/n$ at each of the first n points of the sequence (with multiplicity if points are repeated). Equidistribution means $\mu_n \to m$, the normalized Lebesgue measure on S^1. Weak convergence implies that $\mu_n(I) \to m(I)$ for any arc I (as ∂I consists of at most two points, and m puts zero mass on points). Note that this is not true for general Borel sets. For example, if $A = \{\zeta_1, \zeta_2, \ldots\}$, then $\mu_n(A) = 1$ for all n but $m(A) = 0$.

In this language, what we have been doing for a sequence $x_1, x_2, \ldots$ is to consider the sequence of empirical measures μ_n and asking if they converge to uniform measure on $[0, 1)$. The main idea of Weyl is summarized as saying that this is true if and only if $\int e_m(t) d\mu_n(t) \to 0$ as $n \to \infty$ for any non-zero integer m. It is not necessary for μ_n to come from a sequence, this is true for any sequence of probability measures on S^1. In particular, we shall use the following version with triangular arrays replacing sequences. By a triangular array, we mean a collection $\{\{\xi_{n,1}, \ldots, \xi_{n,n}\} : n \geq 1\}$ of elements of S^1. We say that it is equidistributed if the empirical measures $\mu_n = \frac{1}{n}(\delta_{\xi_{n,1}} + \ldots + \delta_{\xi_{n,n}})$ converges to $m(\cdot)$, as $n \to \infty$. By the discussion so far, this is equivalent to convergence of exponential sums as in the following lemma. Observe that we do not need to consider exponential sums with negative powers since $\xi^{-1} = \overline{\xi}$ for $\xi \in S^1$.

Lemma 3. *A triangular array $\{\{\xi_{n,1}, \ldots, \xi_{n,n}\} : n \geq 1\}$ in S^1 is equidistributed if and only if*

$$\frac{1}{n}\sum_{k=1}^{n} \xi_{n,k}^{p} \to 0$$

as $n \to \infty$, for every integer $p \geq 1$.

4. Distribution of zeros of polynomials

Figure 4 shows that roots of certain random polynomials are concentrated close to the unit circle in the complex plane, and the angular distribution is roughly uniform. In this section we want to prove a theorem of this nature. Randomness will not play any role.

Consider a sequence of complex polynomials

$$\begin{aligned} f_n(z) &+ a_{n,n}z^n + a_{n,n-1}z^{n-1} + \ldots + a_{n,1}z + a_{n,0} \\ &= a_{n,n}(z - \zeta_{n,1}) \ldots (z - \zeta_{n,n}). \end{aligned}$$

Let $\zeta_{n,k} = r_{n,k}e^{2\pi i\theta_{n,k}}$ and let $\xi_{n,k} = e^{2\pi i\theta_{n,k}}$. To make precise the theorem suggested by the pictures, introduce the empirical measures $\mu_n = \frac{1}{n}(\delta_{\zeta_{n,1}} + \ldots + \delta_{\zeta_{n,n}})$. Let μ be the uniform measure

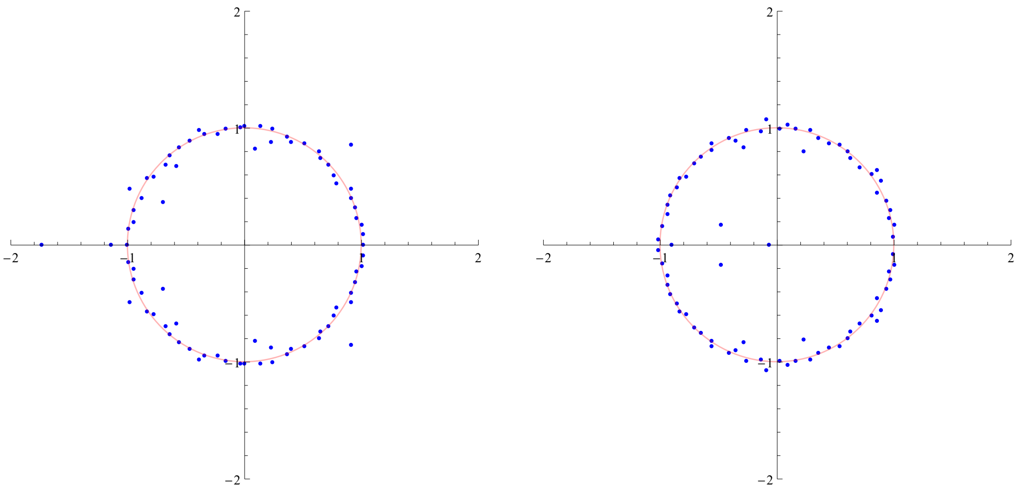

FIGURE 3. Zeros of two random polynomials of degree 80. Left: Coefficients are ±1 with equal probability. Right: Coefficients uniformly distributed in $[0, 1]$. The unit circle is shown in red.

on S^1, i.e., $\mu(A)$ is the normalized Lebesgue measure of $A \cap S^1$ for any Borel set $A \subseteq \mathbb{C}$. Then we show the following theorem.

Theorem 1. *Assume that there exist $0 < b < B < \infty$ such that $b \leq |a_{n,k}| \leq B$ for all $k \leq n$ and for all $n \geq 1$. Then $\mu_n \to \mu$ as $n \to \infty$.*

The convergence of μ_n to μ is exactly the same as the pair of statements below, taken together.

(1) (Radial distribution converges to δ_1): $\mu_n\{z : 1-\delta < |z| < 1+\delta\} \to 1$ as $n \to \infty$. This is clearly equivalent to weak convergence of the probability measures $\frac{1}{n}(\delta_{r_{n,1}} + \ldots + \delta_{r_{n,n}})$ on $\mathbb{R}$ to the degenerate measure δ_1.

(2) (Angular distribution converges to uniform on S^1): $\mu_n\{z : \alpha < \arg z < \beta\} \to \frac{\beta-\alpha}{2\pi}$ for any $0 \leq \alpha < \beta \leq 2\pi$. This is clearly equivalent to the equidistribution of the triangular array $\{\{\xi_{n,1}, \ldots, \xi_{n,n}\} : n \geq 1\}$.

We prove the theorem in steps as follows.

(1) Show that most of the roots have absolute value close to 1. This takes care of radial distribution.

(2) Show that $\frac{1}{n}\sum_{k=1}^{n} \zeta_{n,k}^m$ is small. This is easier because it is a symmetric polynomial of the roots and hence it can be expressed in terms of the coefficients of the polynomial.

(3) By the first step, $\xi_{n,k}$ and $\zeta_{n,k}$ are almost the same, for most zeros. Then use the second step to conclude that $\frac{1}{n}\sum_{k=1}^{n} \xi_{n,k}^m$ is small.

(4) Invoke Lemma 3 to conclude equidistribution.

Step-1: A first observation that we make is that all the zeros have absolute value between $b/(B+b)$ and $(B+b)/b$. Indeed, $|f_n(z)| \geq b - B(|z| + |z|^2 + \ldots + |z|^n) \geq b - B\frac{|z|}{1-|z|}$ which is strictly positive if $|z| < b/(B+b)$. Hence such a z cannot be a root.

For the same reason, the polynomial $f_n^*(z) := z^n f_n(1/z) = a_{n,n} + a_{n,n-1}z + \ldots + a_{n,0}z^n$ has roots with absolute value less than $b/(B+b)$. But the roots of f_n^* are the reciprocals of the roots of f_n, hence f_n has no roots with absolute value greater than $(B+b)/b$.

Step-2: The second observation is that for any $\delta > 0$, there are numbers M_δ such that f_n has at most M_δ roots whose absolute value are either less than $1-\delta$ or greater than $1+\delta$.

This is easy to see by a compactness argument. Let $\mathcal{H}$ denote the set of all power series with coefficients bounded between b and B in absolute value. Observe that the radius of convergence is equal to 1 for all $f \in \mathcal{H}$. On any subdisk $\mathbb{D}(0, 1-\delta)$, we have the uniform bound $|f(z)| \leq B/(1-|z|) \leq B/\delta$ for all $f \in \mathcal{H}$. hence by Montel's theorem[5], $\mathcal{H}$ is a normal family. Therefore, if there were a sequence of polynomials f_n in $\mathcal{H}$ with at least ℓ_n roots in $\mathbb{D}(0, 1-\delta)$, where $\ell_n \to \infty$, then by taking a subsequential limit we would get a power series $f \in \mathcal{H}$ that has infinitely many

[5] Montel's theorem is an overkill here. Can argue directly by taking subsequential limits of coefficients...

zeros in $\mathbb{D}(0, 1-\delta)$. But this is impossible as f is non-zero (since $|f(0)| \geq b$) and is holomorphic on the unit disk. This proves a uniform upper bound M_δ on the number of roots in $\mathbb{D}(0, 1-\delta)$ for any $f \in \mathcal{H}$.

Since the number of roots of f_n of absolute value greater than $1+\delta$ is the same the the number of roots of f_n^* in $\mathbb{D}(0, 1/(1+\delta))$, and $f_n^* \in \mathcal{H}$, we also get a similar uniform bound for the number of zeros outside $\mathbb{D}(0, 1+\delta)$.

Step-3: We now study the power sums of roots. For any $x_1, \ldots, x_n$, recall the *elementary symmetric polynomials* $e_k(x) = \sum_{i_1 < \ldots < i_k} x_{i_1} \ldots x_{i_k}$ and the *power symmetric polynomials* $p_k(x) = x_1^k + \ldots + x_n^k$. When applied to the roots of the polynomial f, the elementary symmetric polynomials are easily expressed in terms of the coefficients as $e_k(\zeta) = (-1)^k \frac{a_{n-k}}{a_n}$. What we want is to control $p_k(\zeta)$. For this, we must express p_k in terms of $e_1, \ldots, e_k$. For example, $p_1 = e_1$, $p_2 = e_1^2 - 2e_2$. More generally, one can see by induction that there are some universal polynomials Q_k (it is a homogeneous polynomial in k variables and has degree k) such that $p_k = Q_k(e_1, \ldots, e_k)$. It is important to note that the coefficients of Q_k do not depend on n at all[6].

Now $|e_k(\zeta)| \leq \frac{|a_{n-k}|}{|a_n|} \leq \frac{B}{b}$, we get the bounds (here we assume that $B \geq 1$, for if not, we can replace it by 1)

$$\frac{1}{n}|p_k(\zeta)| \leq \frac{C_k(B/b)^k}{n} \tag{1}$$

where C_k is the sum of absolute values of the coefficients of all monomials in Q_k.

Step-4: Next we study power symmetric sums of $\xi_k = \zeta_k/|\zeta_k|$, $1 \leq k \leq n$. We compare it to $\frac{1}{n}|p_k(\zeta)|$.

$$\Big|\frac{1}{n}p_k(\xi) - \frac{1}{n}p_k(\zeta)| \leq \frac{1}{n}\sum_{j=1}^{n} |1 - |\zeta_j|^k| \leq \frac{k((B+b)/b)^k}{n}\sum_{j=1}^{n} |1 - |\zeta_j||.$$

In the last step we used the result from Step-1 that $|\zeta_j| \leq \frac{B+b}{b}$ (and that the derivative of $x \mapsto x^k$ is kx^{k-1}). Fix any $\delta > 0$ and split the sum into terms with $1-\delta \leq |\zeta_j| \leq 1+\delta$ and the rest. The rest consists of at most M_δ terms (by Step-2) each of which is bounded by $(B+b)/b$ (by Step-1). The first summan has all terms bounded by δ. Hence,

$$\begin{aligned}\Big|\frac{1}{n}p_k(\xi)| &\leq |\frac{1}{n}p_k(\zeta)| + \delta + \frac{(B+b)M_\delta}{bn} \\ &\leq \frac{C_k(B/b)^k}{n} + \delta + \frac{(B+b)M_\delta}{bn}\end{aligned}$$

by (1). Let $n \to \infty$ and then $\delta \to 0$ to see that $\frac{1}{n}p_k(\xi) \to 0$ as $n \to \infty$, for every $k \geq 1$.

[6] While we do not need the explicit form, these relationships are expressed by *Newton's identities*: $p_k = (-1)^{k-1}ke_k - \sum_{j=1}^{k-1}(-1)^{k-j+1}e_{k-j}\,p_j$. See this Wikipedia article for more on these relationships.

In conclusion, Step-4 together with Lemma 3 shows that $\xi_{n,k}$, $n \geq k$, is equidistributed on S^1. By Step-2 we know that the empirical distribution of radii of zeros converges to δ_1. This completes the proof of the Theorem 1. ■

Theorem 1 "proves" the first picture in Figure 4 but not the second one!). This is a limitation of our method, but the point was not to derive the strongest results, but to illustrate the applicability of Weyl's method of using exponential sums. Here is a slight strengthening of the theorem, by being more quantitative in Step-2.

A quantitative bound for number of roots: Well-known theorems in complex analysis express the number of zeros of a holomorphic function in terms of certain integrals (eg., the argument principle). A convenient one is Jensen's formula which states that if f is holomorphic in a neighbourhood of $\mathbb{D}(0,R)$ and $f(0) \neq 0$, then

$$\int_0^{2\pi} \log|f(Re^{i\theta})|\frac{d\theta}{2\pi} - \log|f(0)| = \sum_{\zeta : f(\zeta)=0} \log_+\left(\frac{R}{|\zeta|}\right).$$

Here $\log_+ x = \max\{\log x, 0\}$. On the right zeros are counted with multiplicities, as always.

Apply this to polynomial $f(z) = a_n z^n + \ldots + a_1 z + a_0$ where $b \leq |a_k| \leq B$. Suppose $0 < r < R < 1$. If $n_f(r)$ is the number of zeros of f in $\mathbb{D}(0,r)$, then the right hand side is at least $n_f(r)\log(R/r)$, since each zero in $\mathbb{D}(0,r)$ contributes $\log(R/r)$ (and others contribute a non-negative amount). The left hand side is upper bounded by $\log(B/(1-R)) + \log(1/b)$. This is because $-\log|f(0)| \leq \log(1/b)$ and $|f(z)| \leq B/(1-|z|)$ for any $|z| < 1$. Thus, we arrive at

$$n_f(r)\log\frac{R}{r} \leq \log\frac{B}{b(1-R)}$$

This gives a quantitative bound for M_δ in terms of b and B.

Exercise 2. Use the quantitative bound on M_δ to strengthen Theorem 1 to allow B and $1/b$ to grow with n. Perhaps the condition $B_n + \frac{1}{b_n} = o(n^\varepsilon)$ for every $\varepsilon > 0$ suffices.

5. Erdös-Turan lemma

While weak convergence (which can be metrized, when restricted to probability measures) is the usual notion of convergence, many stronger metrics are sometimes used (perhaps on subsets of probability measures). Of course, convergence in these stronger metrics is a stronger result than convergence in weak sense. Here we introduce one of these distances and a quantitative version of Weyl's lemma.

For μ, ν Borel probability measures on S^1, let $\text{KS}(\mu,\nu) = \sup_I |\mu(I) - \nu(I)|$. Here the supremum is over all arcs in S^1. This is called the *Kolmogorov-Smirnov* distance.

Exercise 1. Show that there exist probability measures μ_n and μ on S^1 such that $\mu_n \to \mu$ weakly but not in Kolmogorov-Smirnov distance. However, when $\mu = m$, the normalized Lebesgue measure, then show that $\mu_n \to m$ in Kolmogorov-Smirnov distance if and only if $\mu_n \to m$ weakly.

Because of the second part of the exercise above, when talking of equidistribution on the circle, KS distance is equivalent to weak convergence. Erdös and Turán found a quantitative version of Weyl's lemma by giving a bound on the KS distance of a measure from the uniform measure, in terms of the Fourier coefficients of the measure. Note that even if two metrics are equivalent (in the sense that they induce the same topology), quantitative estimates in one metric do not automatically lead to quantitative estimates in the other metric.

Theorem 2. *[Erdös-Turán] Let μ be a probability measure on S^1 and let m denote the uniform measure on S^1. Then, KS$(\mu, m) \le 4\left[\sum\limits_{k=1}^{n} \frac{|\hat{\mu}(k)|}{k} + \frac{1}{n}\right]$ for all $n \ge 3$.*

The proof given here is from an unpublished note by Mikhail Sodin (personal communication). First we recall some facts about the Fejér kernel $K_N(u) = \frac{1}{N+1}\frac{\sin^2\left(\frac{N+1}{2}2\pi u\right)}{\sin^2\left(\frac{1}{2}2\pi u\right)}$ (all functions here are on S^1 or equivalently 1-periodic on $\mathbb{R}$). Then $K_N \ge 0$ and its integral over $[0,1]$ is 1. Further, $K_N(u) \le 1/(N+1)\sin^2(\pi u)$. Using the fact that $\sin(x)/x$ is decreasing on $[0, \pi/2]$ and hence $\sin(x) \ge 2x/\pi$, we see that for $\delta < \frac{1}{2}$

$$\begin{aligned}\int\limits_{[-\delta,\delta]^c} K_N(u)du &\le \frac{2}{4(N+1)}\int_\delta^{\frac{1}{2}} \frac{1}{u^2}du \\ &\le \frac{1}{2(N+1)\delta} \qquad (1)\end{aligned}$$

which is a better bound than what we used when proving Fejér's theorem (and the improvement will play a role below).

Proof of Theorem 2. Fix a probability measure μ on $S^1 = [0,1)$ and define the function $f : [0,1] \mapsto \mathbb{R}$ by $f(t) = t - \mu[0,t] - A$, where A is chosen so that $\hat{f}(0) = \int_0^1 f(t)dt = 0$ (clearly possible). Observe that $f(0) = f(1) = 0$ and extend f as an 1-periodic function on $\mathbb{R}$. Further note that for any $0 < s < 1$ and any t,

$$f(t+s) - f(t) = s - \mu(t, t+s] \le s.$$

Let t_0 be a point at which $|f(t_0)| = \|f\|$. Then by the above inequality,

(1) if $f(t_0) > 0$, then $f(t) \ge \|f\| - 2\delta$ for $t \in [t_0 - 2\delta, t_0]$,

(2) if $f(t_0) < 0$, then $f(t) \le -\|f\| + 2\delta$ for $t \in [t_0, t_0 + 2\delta]$.

We shall make the choice $\delta = 2/(N+1)$ later (since we need $\delta < \frac{1}{2}$ for the estimate (1), we assume $N \ge 3$ henceforth).

Next recall $\sigma_N f$ from the proof of Fejér's theorem:

$$\sigma_N f(t) = \sum_{k=-N}^{N}\left(1 - \frac{|k|}{N+1}\right)\hat{f}(k)e^{2\pi i k t} = \int_I f(s)K_N(t-s)ds.$$

In the first case (when $f(t_0) > 0$),

$$\begin{aligned}\sigma_N f(t_0 - \delta) &= \int_{[-\delta,\delta]} f(t_0 - \delta - s)K_N(s)ds + \int_{[-\delta,\delta]^c} f(t-s)K_N(s)ds \\ &\geq (\|f\| - 2\delta)\int_{[-\delta,\delta]} K_N(s)ds - \|f\|\int_{[-\delta,\delta]^c} K_N(s)ds \\ &= \|f\|\left(1 - 2\int_{[-\delta,\delta]^c} K_N(s)ds\right) - 2\delta \\ &\geq \|f\|\left(1 - \frac{1}{\delta(N+1)}\right) - 2\delta.\end{aligned}$$

Remembering that $\delta = 2/(N+1)$, we get $\|f\| \leq 2\sigma_N f(t_0 - \delta) + 2\delta$. This was the case when $f(t_0) > 0$. If $f(t_0) < 0$, then follow the same steps to get $\|f\| \leq 2\sigma_N f(t_0 + \delta) + 2\delta$. Overall, the conclusion is that in all cases,

$$\|f\| \leq 2\|\sigma_N\| + \frac{4}{N}.$$

Now use the series form of $\sigma_N f$ to see that

$$\|\sigma_N f\| \leq \sum_{k=-N}^{N} |\hat{f}(k)| = 2\sum_{k=1}^{N} |\hat{f}(k)|$$

where the last equality is because $\hat{f}(0) = 0$ (by choice of A) and $\hat{f}(-k) = \overline{\hat{f}(k)}$. For $k \geq 1$ we have

$$\begin{aligned}\hat{f}(k) &= \int_0^1 e^{-2\pi ikt}(\mu - m)[0,t]dt \qquad \text{(the integral against } A \text{ is 0)} \\ &= \int_0^1\int_0^1 e^{-2\pi ikt}\mathbf{1}_{[0,t]}(s)\ d(\mu - m)(s)\ dt \\ &= \int_0^1\int_0^1 e^{-2\pi ikt}\mathbf{1}_{[0,t]}(s)\ dt\ d(\mu - m)(s) \qquad \text{(justify the use of Fubini)} \\ &= \int_0^1 \frac{1 - e^{-2\pi iks}}{-2\pi ik} d(\mu - m)(s) \\ &= \frac{1}{2\pi ik}\hat{\mu}(k).\end{aligned}$$

In the last step we used the fact that $\int_0^1 d(\mu - m) = 0$ and $\int_0^1 e^{-2\pi ikt}dm(t) = 0$ (since $k \geq 1$). Plugging this into the bound on $\|\sigma_N f\|$ and using that in the bound for $\|f\|$, we arrive at

$$\|f\| \leq \frac{2}{\pi}\sum_{k=1}^{N} \frac{|\hat{f}(k)|}{k} \ + \ \frac{4}{N}.$$

Now for any $[a,b] \subseteq [0,1)$, we see that $|f(b) - f(a)| \leq 2\|f\|$. But $|f(b) - f(a)| = |\mu(a,b] - m(a,b]|$. Thus we get the bounds $|\mu(a,b] - (b-a)| \leq \frac{4}{\pi}\sum_{k=1}^{N}\frac{1}{k}|\hat{f}(k)| + \frac{4}{N}$. This completes the proof. ■

6. Matching sets of points on the torus

Let $\mathbb{T}^d$ be the d-dimensional torus that we identify with the cube $(-\pi,\pi]^d$. Given two sets of n distinct points $\{x_1,\ldots,x_n\}$ and $\{y_1,\ldots,y_n\}$, we look for a permutation $\pi \in \mathcal{S}(n)$ such that the *average matching cost* $\mathcal{C} := \frac{1}{n}\sum_{k=1}^n d(x_k, y_{\pi(k)})$ is minimized. This is one of many different cost-functions that can be imposed, for example, one can consider ℓ^p norms of $(d(x_1,y_{\pi(1)}),\ldots,d(x_n,y_{\pi(n)}))$, but the only one that we consider here.

The following is a deep and famous theorem

Theorem 1 (Ajtai–Komlos–Tusnady)**.** *Let x_i, y_i, $i \le n$, be chosen uniformly at random from $\mathbb{T}^d$. Then the average matching cost $\mathcal{C}_n$ satisfies*

$$\mathbf{E}[\mathcal{C}_n] \asymp \begin{cases} \frac{1}{\sqrt{n}} & \text{if } d=1, \\ \frac{\sqrt{\log n}}{\sqrt{n}} & \text{if } d=2, \\ \frac{1}{n^{1/d}} & \text{if } d\ge 3. \end{cases}$$

When we place n points in $\mathbb{T}^d$, typical inter-point distance is $n^{-1/d}$. This explains the $d \ge 3$ result above. Interestingly, when $d=2$, we have an extra factor of $\sqrt{\log n}$, which needs an explanation.

Our interest is not merely in proving this theorem for random points, but to establish a lemma similar to Erdös–Turan.

Definition 2. Let μ and ν be two probability measures on $\mathbb{T}^d$. The Kantorovich distance between them is defined as

$$\mathcal{W}_1(\mu,\nu) := \inf\left\{\int_{\mathbb{T}^d\times\mathbb{T}^d} d(x,y)\,d\theta(x,y) : \theta \text{ has marginals } \mu,\nu\right\}.$$

When μ is the uniform distribution on $\{x_1,\ldots,x_n\}$ and ν is the uniform distribution on $\{y_1,\ldots,y_n\}$, it is not hard to see that $\mathcal{W}_1(\mu,\nu)$ is precisely the average matching cost[7].

Recall that the Fourier coefficients of a probability measure μ on $\mathbb{T}^d$ are given by

$$\hat{\mu}(k) = \frac{1}{(2\pi)^d}\int_{\mathbb{T}^d} e^{-i\langle x,k\rangle} d\mu(x), \qquad \text{for } k \in \mathbb{Z}^d.$$

We establish the following lemma analogous to the Erdös-Turan lemma, except that the notion of distance is different now. Below, if $k=(k_1,\ldots,k_d)$, we write $|k|^2 = k_1^+ \ldots + k+_d^2$ (using any other norm on $\mathbb{R}^d$ will only change the constant C below).

Lemma 3 (Bobkov–Ledoux)**.** *Let μ,ν be probability measures on $\mathbb{T}^d$. Then*

$$\mathcal{W}_1(\mu,\nu) \le C\left[\sqrt{\sum_{1\le|k|\le N}\frac{1}{|k|^2}|\hat{\mu}(k)-\hat{\nu}(k)|^2 + \frac{1}{N}}\right]$$

Assuming this lemma, we can easily derive Theorem 1.

[7] The reason is that every doubly stochastic matrix is a convex combination of permutation matrices.

Proof of Theorem 1. The average matching cost is just $\mathcal{W}_1(\mu,\nu)$, where $\mu = \frac{1}{n}\sum_{k=1}^n \delta_{x_k}$ and $\nu = \frac{1}{n}\sum_{k=1}^n \delta_{y_k}$. From the lemma of Bobkov and Ledoux, and applying Cauchy-Schwarz, we get

$$\mathbf{E}[\mathcal{W}_1(\mu,\nu)] \le C\left[\sqrt{\sum_{1\le|k|\le N} \frac{1}{|k|^2}\mathbf{E}[|\hat{\mu}(k)-\hat{\nu}(k)|^2]} + \frac{1}{N}\right].$$

Fix k and write

$$\hat{\mu}(k)-\hat{\nu}(k) = \frac{1}{n}\sum_{p=1}^n e^{-i\langle x_p,k\rangle} - e^{-i\langle y_q,k\rangle}.$$

This is a sum of $2n$ uncorrelated random variables of unit variance, hence

$$\mathbf{E}[|\hat{\mu}(k)-\hat{\nu}(k)|^2] = \frac{4}{n}.$$

Therefore,

$$\mathbf{E}[\mathcal{W}_1(\mu,\nu)] \le C\left[\frac{1}{\sqrt{n}}\sqrt{\sum_{1\le|k|\le N} \frac{1}{|k|^2}} + \frac{1}{N}\right].$$

Now, by an elementary calculation (e.g., compare with the integral $\int_{1\le|x|\le R}\frac{1}{|x|^2}dx$),

$$\sum_{1\le|k|\le N}\frac{1}{|k|^2} \asymp \begin{cases} 1 & \text{if } d=1, \\ \log N & \text{if } d=2, \\ N^{d-2} & \text{if } d\ge 3. \end{cases}$$

In the resulting bound for $\mathbf{E}[\mathcal{W}_1(\mu,\nu)]$, the optimal choice of N is seen to be $N=\sqrt{n}$ (for $d=1$), $N = n/\sqrt{\log n}$ (for $d=2$) and $N=n^{1/d)}$ for $d\ge 3$, resulting in the bounds

$$\mathbf{E}[\mathcal{W}_1(\mu,\nu)] \lesssim \begin{cases} 1/\sqrt{n} & \text{if } d=1, \\ \sqrt{\log n}/\sqrt{n} & \text{if } d=2, \\ 1/n^{1/d} & \text{if } d\ge 3. \end{cases}$$

This completes the proof. ■

It remains to prove the Lemma of Bobkov and Ledoux. The key tool is a fundamental dual formulation of the Kantorovich distance. Recall that a Lip(c) function on a metric space is one that satisfies $|f(x)-f(y)|\le d(x,y)$.

Lemma 4 (Kantorovich–Rubinstein)**.** *Let μ and ν be probabilty measures on $\mathbb{T}^d$. Then*

$$\mathcal{W}_1(\mu,\nu) = \sup\left\{\int_{\mathbb{T}^d} u d\mu - \int u d\nu : u \in \mathit{Lip}(1)\right\}.$$

The infimum may also be taken over $u\in C^1$ with $|\nabla u|\le 1$.

Proof. Let u be smooth with $|\nabla u| \leq 1$. Then by the Plancherel/Parseval relation, we have

$$\int_{\mathbb{T}^d} u d\mu - \int_{\mathbb{T}^d} u d\nu = \sum_{k \in \mathbb{Z}^d} \hat{u}(k)(\hat{\mu}(k) - \hat{\nu}(k))$$

The smoothness of u ensures that the Fourier coefficients $\hat{u}(k) := \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} u(x) e^{-i\langle x,k\rangle} dx$ decay faster than any power of $|k|$, hence the above series is summable (as $\hat{\mu}, \hat{\nu}$ are bounded by 1). Observe that the term $k = 0$ vanishes, as $\hat{\mu}(0) = \hat{\nu}(0) = 1$. Hence, by Cauchy-Schwarz inequality,

$$\sum_{k \neq 0} \hat{u}(k)(\hat{\mu}(k) - \hat{\nu}(k)) \leq \left(\sum_{k \neq 0} |\hat{u}(k)|^2 |k|^2 \right)^{\frac{1}{2}} \left(\sum_{k \neq 0} \frac{1}{|k|^2} |\hat{\mu}(k) - \hat{\nu}(k)|^2 \right)^{\frac{1}{2}}.$$

As $\widehat{\partial_j u}(k) = -ik_j \hat{u}(k)$ and using Plancherel, we see that the first factor is equal to $\|\nabla u\|_2$, which is at most $(2\pi)^d$ as $|\nabla u| \leq 1$ pointwise.

The second factor is summable if we assume additional smoothness of μ and ν, but not in general. For general μ, ν, we replace them by $\mu' = \mu \star g$ and $\nu' = \nu \star g$ where g is a smooth probability density. As μ', ν' have smooth densities, we apply the bound above for them, and use the fact that $\hat{\mu'}(k) = \hat{\mu}(k)\hat{g}(k)$ and $\hat{\nu'}(k) = \hat{\nu}(k)\hat{g}(k)$ to get

$$\mathcal{W}_1(\mu', \nu') \leq C \left(\sum_{k \neq 0} \frac{|\hat{g}(k)|^2}{|k|^2} |\hat{\mu}(k) - \hat{\nu}(k)|^2 \right)^{\frac{1}{2}}.$$

By the obvious coupling $d\theta(x, y) = g(y - x) d\mu(x) dy$ (pick x according to μ and then pick $y - x$ according to g), we see that $\mathcal{W}_1(\mu, \mu') \leq \sigma_g$, where $\sigma_g^2 = \int_{\mathbb{T}^d} |x|^2 g(x) dx$ is the second moment of g. Similarly $\mathcal{W}_1(\mu, \mu') \leq \sigma_g$. Thus,

$$\mathcal{W}_1(\mu, \nu) \leq C \left(\sum_{k \neq 0} \frac{|\hat{g}(k)|^2}{|k|^2} |\hat{\mu}(k) - \hat{\nu}(k)|^2 \right)^{\frac{1}{2}} + 2\sigma_g.$$

We can optimize over all g. If we make the choice $g = K_N$, the d-dimensional Fejér kernel of order N, then we have seen (in $d = 1$, it is similar in higher dimensions) that $\sigma_g \asymp \frac{1}{N}$. Further $\hat{g}(k) = (1 - \frac{|k|}{N})_+$. Therefore, we arrive at

$$\mathcal{W}_1(\mu', \nu') \lesssim \left(\sum_{1 \leq |k| \leq N} \frac{1}{|k|^2} |\hat{\mu}(k) - \hat{\nu}(k)|^2 \right)^{\frac{1}{2}} + \frac{1}{N}.$$

This completes the proof. ■

7. How to distribute points uniformly on a square?

What is the best way to choose n points in the unit square so that they are as uniformly distributed as possible? If the underlying space was S^1, then the choice seems obvious, pick n equispaced points. But for the two-dimensional question, we need to be more precise about our criterion for "as uniformly distributed as possible". Changing the space and changing our measure of uniformity, one gets a variety of inequivalent problems, some of which are solved, some open. We stick to one specific choice in this brief introduction to this topic[8].

Notation: Let $Q = [0,1)^2$ (we use right-open, left-closed intervals and squares for usual reasons that we can partition them into smaller intervals or squares of the same kind). Always $\mathcal{P}_N$ (or simply $\mathcal{P}$) denotes a subset of N points in Q. Its *discrepancy* in any set $A \subseteq Q$ is defined as $\#(\mathcal{P} \cap A) - n|A|$. In particular, we write $D_{\mathcal{P}}(x,y)$ for the discrepancy of the set $[0,x) \times [0,y)$. The total discrepancy of $\mathcal{P}$ is defined as

$$D(\mathcal{P}) = \left(\int_Q |D_{\mathcal{P}}(x,y)|^2 \, dxdy \right)^{\frac{1}{2}}.$$

The goal is to find or get estimates on the lowest possible discrepancy. Note that the answer would be the same (up to constants) if we considered (as may seem more natural) all rectangles $[x_1, x_2) \times [y_1, y_2)$ and considered the L^2 norm in the four variables x_1, x_2, y_1, y_2.

Obvious generalizations (not considered here) include changing the space Q (eg., $[0,1)^d$ or sphere or disk etc.), changing the class of sets (eg., can allow rectangles with any orientation, the collection of all disks, or convex sets), and changing the criterion by which discrepancies of sets in the collection are combined to get the total discrepancy (eg., L^p norm in a suitable sense, in particular, the worst-case discrepancy corresponding to $p = \infty$).

The result that we shall prove is this.

Theorem 1 (Roth, Davenport)**.** *There exist constants $0 < c < C < \infty$ such that*

(1) $D(\mathcal{P}_N) \geq c \log N$ for any N-element subset $\mathcal{P}_N \subseteq Q$,

(2) There exists an N-element subset $\mathcal{P}_N^*$ such that $D(\mathcal{P}_N^*) \leq C \log N$.

Proof of the lower bound: We present Roth's proof[9] of the lower bound. Fix a set $\mathcal{P}$ with N elements. By Cauchy-Schwarz inequality, for any $f \in L^2(Q)$, we have

$$D(\mathcal{P}) \geq \frac{\langle D, f \rangle}{\sqrt{\langle f, f \rangle}}.$$

The strategy is to find a function f such that the right hand side is at least $c \log N$. But without knowing $\mathcal{P}$, how can one produce such a function? The idea is to consider not one, but a family of functions $\mathcal{F}$ such that for any N-point set $\mathcal{P}$, there is one $f \in \mathcal{F}$ that works. We introduce this class of functions now.

For a natural number p and $0 \leq k \leq 2^p - 1$, let I_k^p denote the dyadic interval $[k2^{-p}, (k+1)2^{-p})$. We refer to $I_k^p \times I_\ell^q$ as a (p,q)-dyadic rectangle. For an interval I, let $I(-)$ and $I(+)$ denote the left half and right half, respectively. Let h_k^p denote the *Haar function* supported on I_k^p and taking values

[8] The material here is taken largely from the book *Irregularities of distribution* by Beck and Chen.

[9] From the book of Beck and Chen, one gathers that later improvement, particularly by Schmidt, are incorporated here.

± 1 on $I_k^p(\pm)$. These form an orthogonal family (across p and k) in $L^2([0,1))$. The function $h_k^p \otimes h_\ell^q$ is supported on $I_k^p \times I_\ell^q$ and takes the values ± 1 in a checkerboard pattern in the four quarters of the dyadic rectangle. Now define

$$\mathcal{G}^{p,q} = \left\{ f : Q \mapsto \{+1,-1\} : f = \sum_{k=0}^{2^p-1} \sum_{\ell=0}^{2^q-1} \pm h_k^p \otimes h_\ell^q \right\}.$$

This family consists of $2^{2^{p+q}}$ functions by the choice of the signs. For $n \geq 0$, set

$$\mathcal{F}^n = \{ f : Q \mapsto \mathbb{R} : f = f_n + f_{n-1} + \ldots + f_0 \text{ with } f_p \in \mathcal{G}^{p,n-p} \}.$$

This is the family of functions mentioned in the outline above, for a suitable value of n (to break the suspense, $n \asymp \log N$).

Lemma 2. *Fix $n \geq 0$ and $0 \leq r < s \leq n$. Then $\mathcal{G}^{r,n-r} \perp \mathcal{G}^{s,n-s}$ in $L^2(Q)$. As a corollary, for any $f \in \mathcal{F}^n$, we have $\langle f, f \rangle = n+1$.*

Proof. Enough to show that $h_k^r \otimes h_\ell^{n-r} \perp h_{k'}^s \otimes h_{\ell'}^{n-s}$ for any k, ℓ, k', ℓ' (of course we mean $0 \leq k \leq 2^r - 1$, etc.). Fix the second co-ordinate y and integrate over the first co-ordinate x. But their inner product is just $\langle h_k^r, h_{k'}^s \rangle \langle h_\ell^{n-r}, h_{\ell'}^{n-s} \rangle$ (these inner products are in $L^2([0,1))$). As $r \neq s$, both factors vanish.

If $f = f_0 + \ldots + f_n$ with $f_p \in \mathcal{G}^{p,n-p}$, then by the orthogonality of f_ps and the fact that $\langle f_p, f_p \rangle = 1$ for each p (since f_p takes values ± 1 throughout Q), we conclude that $\langle f, f \rangle = n+1$. ■

Lemma 3. *Fix n such that $2^n \geq N$. Then for any N-point set $\mathcal{P} \subseteq Q$, there exists $f \in \mathcal{F}^n$ such that $\langle D_{\mathcal{P}}, f \rangle \geq (n+1) N 2^{-n-5}$.*

Proof. We shall first show that for each $0 \leq p \leq n$, there is some $f_p \in \mathcal{G}^{p,n-p}$ such that $\langle D_{\mathcal{P}}, f_p \rangle \geq N 2^{n-4}$. Setting $f = f_0 + \ldots + f_n$, we get the function as claimed in the lemma.

Now fix $0 \leq p \leq n-1$ and let $q = n - p$. We want f_p of the form $\sum_{k=0}^{2^p-1} \sum_{\ell=0}^{2^q-1} \pm h_k^p \otimes h_\ell^q$ that has all large an inner product with $D_{\mathcal{P}}(\cdot)$ as possible. First of all, choose the signs in the sum so that the inner product of $D_{\mathcal{P}}$ with each summand is non-negative. This allows us to drop terms without increasing $\langle D_{\mathcal{P}}, f \rangle$. In fact, we shall only consider those k, ℓ for which $\mathcal{P}$ has no points in $I_k^p \times I_\ell^q$. Note that there are at least $2^n - N$ such pairs (k, ℓ).

Take any such (k, ℓ). Since $\#(\mathcal{P} \cap [0,x) \times [0,y))$ stays constant over $x \in I_k^p$ if we fix $y \in I_\ell^q$, but $h_k^p(x)$ is $+1$ for $x \in I_k^p(+)$ and -1 for $x \in I_k^q(-)$, it follows that

$$\iint_{I_k^p \times I_\ell^q} \#(\mathcal{P} \cap [0,x) \times [0,y)) \, dx dy = 0.$$

Further, writing $h = 2^{-p-1}$ and $k = 2^{-q-1}$ for simplicity, we have

$$\begin{aligned}\iint_{I_k^p \times I_\ell^q} Nxy\, dxdy &= N \iint_{I_k^p(-)\times I_\ell^q(-)} \{xy - (x+h)y - x(y+k) + (x+h)(y+k)\}\ dxdy \\ &= N \iint_{I_k^p(-)\times I_\ell^q(-)} hk\, dxdy \\ &= Nh^2k^2.\end{aligned}$$

Now recall the definition of h, k and that $p + q = n$ to see that the last quantity is $N2^{-2n-4}$. Thus, in $\langle D_{\mathcal{P}}, f_p\rangle$, each empty $(p, n-p)$-dyadic rectangle having no points of $\mathcal{P}$ contributes this much, and the rest contribute a non-negative amount. Therefore,

$$\langle D_{\mathcal{P}}, f_p\rangle \geq (2^n - N)N2^{-2n-4} \geq N2^{-n-5}$$

since $2^n - N \geq 2^{n-1}$ by the choice of n. The proof is complete. ■

We put the ingredients together to get the lower bound in Theorem 1.

Take any N-point set $\mathcal{P}_N$ and choose n such that $2N \leq 2^n < 4N$. Then find $f \in \mathcal{F}^n$ such that $\langle D_{\mathcal{P}_N}, f\rangle \geq (n+1)N2^{-n-5}$. By the first lemma we know that $\langle f, f\rangle = n+1$. Therefore,

$$\|D_{\mathcal{P}_N}\|_{L^2} \geq \frac{\langle D_{\mathcal{P}_N}, f\rangle}{\sqrt{\langle f, f\rangle}} \geq \sqrt{n+1}\ N\ 2^{-n-5} \geq 2^{-9}\sqrt{\log N}.$$

This completes the proof of the lower bound. ■

Proof of the upper bound: We present Davenport's proof of the upper bound in Theorem 1. The first choice that comes to mind is to place N points at $(k/\sqrt{N}, \ell/\sqrt{N})$, $1 \leq k, \ell \leq \sqrt{N}$ (ok, $\sqrt{N}$ may not be an integer, but it should be clear that it is a silly point that can be fixed). However, that leaves long rectangles like $(\frac{1}{\sqrt{N}}, \frac{2}{\sqrt{N}}) \times (0,1)$ that have discrepancy of about $\sqrt{N}$. An idea would be to take this lattice arrangement, and in each horizontal line, shift the points by a different amount so as to "destroy" long empty rectangles. Clearly if we do the shifts in a regular manner, eg., $\frac{1}{\sqrt{N}}(k + k\alpha, \ell)$ for some number α (these numbers have to be considered modulo 1), then it is better to choose an irrational number α. complete this proof

CHAPTER 3

Isoperimetric iequality

1. Isoperimetric inequality

Isoperimetric inequality is a well-known statement in the following form: *Among all bodies in space (in plane) with a given volume (given area), the one with the least surface area (least perimeter) is the ball (the disk).*

Several things need to be made precise. The notion of volume in space or area in the plane are understood to mean Lebesgue measure on $\mathbb{R}^3$ or $\mathbb{R}^2$ or more generally on $\mathbb{R}^d$ (we denote it by $m_d(A)$). Then of course we restrict the notion of "bodies" to Borel sets (or Lebesgue measurable sets).

Still, in measure theory class we (probably!) did not study the notion of surface area of a Borel set in $\mathbb{R}^3$ or the perimeter of a Borel set in $\mathbb{R}^2$. We first need to fix this notion. And then state a precise theorem. First we state a form of the isoperimetric inequality which completely avoids the notion of surface area or perimeter.

Theorem 1 (Isoperimetric inequality)**.** *Let A be Borel subsets of $\mathbb{R}^d$ and let B be a closed ball such that $m_d(A) = m_d(B)$. Then, for any $\varepsilon > 0$, we have $m_d(A_\varepsilon) \geq m_d(B_\varepsilon)$ where $A_\varepsilon = \{x \in \mathbb{R}^d : d(x, y) \leq \varepsilon \text{ for some } y \in A\}$.*

How does this relate to the informally stated version above? If at all we can define the surface area of A, it must be the limit (or $\limsup$ or $\liminf$) of $(m_d(A_\varepsilon) - m_d(A))/\varepsilon$ as $\varepsilon \to 0$. For simplicity, let us define the surface area (or "perimeter") of a Borel set $A \subseteq \mathbb{R}^d$ as

$$\sigma_d(A) := \limsup_{\varepsilon \to 0} \frac{m_d(A_\varepsilon) - m_d(A)}{\varepsilon}$$

which is either a non-negative real number or $+\infty$. If A is a bounded set with smooth boundary, then the above definition agrees with our usual understanding of perimeter/surface area.

Theorem 1 clearly gives the following theorem as a corollary.

Theorem 2 (Isoperimetric inequality - standard form)**.** *Let A be Borel subsets of $\mathbb{R}^d$ and let B be a closed ball such that $m_d(A) = m_d(B)$. Then, $\sigma_d(A) \geq \sigma_d(B)$.*

In this sense, we are justified in saying that Theorem 1 is stronger than Theorem 2. In addition, note the great advantage of the former being easy to state for all Borel sets without having to define the notion of surface area. However, we have omitted a key point in the isoperimetric inequality which is the uniqueness of the surface-area-minimizing set.

Theorem 3 (Equality in isoperimetric inequality)**.** *In the setting of Theorem 1 assume that A is closed. If $m_d(A_\varepsilon) = m_d(B_\varepsilon)$ for some $\varepsilon > 0$, then $A = B(x, r)$ for some $x \in \mathbb{R}^d$.*

However, the analogous statement for Theorem 1 is false without further qualifications. For example, if A is the disjoint union of a closed disk and a closed line segment, then it has the same area and the same perimeter as the ball. But the uniqueness is "essentially true", for example, if one restricts to sets with smooth boundary or alternately by taking a more general notion of perimeter (which does distinguish a disk from a union of a disk and a line segment). We shall present two proofs of Theorem 1. A short one using the Brunn-Minkowski inequality and a longer but more natural one by Steiner symmetrization.

Exercise 4. Show that the isoperimetric inequality is equivalent to the following statement: If $A \subseteq \mathbb{R}^d$ is measurable, then $|A|^{\frac{d-1}{d}} \le C_d \sigma_d(A)$ where $C_d^{-1} = d^{1-\frac{1}{d}} \tau_d^{1/d}$ and $\tau_d = \frac{2\pi^{d/2}}{\Gamma(d/2)}$ is the surface area of the unit sphere S^{d-1}.

Here is a proof of isoperimetric inequality in the plane under some restrictions.

Exercise 5. Let $\gamma(t) = (x(t), y(t))$, $0 \le t \le L$ be a simple smooth curve in the plane, parameterized by its arc length, i.e., $\|\dot\gamma(t)\| = 1$ for all $t \in [0, 2\pi]$. Let A be the area enclosed by γ and let L be the length of γ.

(1) Show that the length of the curve is given by $L^2 = \int_0^{2\pi} |\dot\gamma(t)|^2 dt$ and $A = -\int_0^{2\pi} y(t)\dot x(t) dt$.

(2) WLOG assume that $\int_0^{2\pi} y(t)dt = 0$ and show that $\int_0^{2\pi} y(t)^2 dt \le \int_0^{2\pi} \dot y(t)^2 dt$. [**Hint:** Assume that the Fourier series $y(t) = \sum_{n \in \mathbb{Z}} \hat y_n e^{int}$ converges nicely and uniformly]

2. Brunn-Minkowski inequality and a first proof of isoperimetric inequality

For simplicity write $|A|$ for $m_d(A)$, the d-dimensional Lebesgue measure. For nonempty sets $A, B \subseteq \mathbb{R}^d$, define their Minkowski sum $A + B := \{a + b : a \in A, b \in B\}$.

Theorem 6 (Brunn-Minkowski inequality)**.** *If A, B are non-empty Lebesgue measurable subsets of $\mathbb{R}^d$, and if $A + B$ is also Lebesgue measurable, then,*

$$|A + B|^{1/d} \ge |A|^{1/d} + |B|^{1/d}.$$

The proof is very easy in one dimension. In fact, it is a continuous analogue of the following inequality that we leave as an exercise.

Exercise 7 (Cauchy-Davenport inequality)**.** Let A, B be non-empty finite subsets of $\mathbb{Z}$. Then $|A + B| \ge |A| + |B| - 1$ and the inequality cannot be improved (here $|A|$ denotes the cardinality of A).

Use the same idea to prove Brunn-Minkowski inequality for $d = 1$.

Proof of Theorem 1 using Brunn-Minkowski inequality. Assume $|A| = |rB|$ where B is the unit ball and $r > 0$. Then $A_\varepsilon = A + \varepsilon B$ and hence by Brunn-Minkowski

$$\begin{aligned} |A_\varepsilon|^{1/d} &\geq |A|^{1/d} + \varepsilon |B|^{1/d} \\ &= r|B|^{1/d} + \varepsilon |B|^{1/d} \\ &= |(r+\varepsilon)B|^{1/d}. \end{aligned}$$

Since $(rB)_\varepsilon = (r+\varepsilon)B$, we have proved that $|A_\varepsilon| \geq |(rB)_\varepsilon|$ as required. ■

Proof of Brunn-Minkowski inequality. The proof will proceed by proving it when the two sets are rectangles (parallelepipeds) with sides parallel to the co-ordinate, then for finite unions of rectangles, and finally

Step 1: Suppose $A = \mathbf{x} + [0, a_1] \times \ldots \times [0, a_d]$ and $B = \mathbf{y} + [0, b_1] \times \ldots \times [0, b_d]$ are any two closed parallelepipeds with sides parallel to the axes (we shall refer to them as standard parallelepipeds). Then $A + B = \mathbf{x} + \mathbf{y} + [0, a_1 + b_1] \times \ldots \times [0, a_d + b_d]$. Thus,

$$\begin{aligned} \frac{|A|^{1/d} + |B|^{1/d}}{|A+B|^{1/d}} &= \left(\prod_{k=1}^{d} \frac{a_k}{a_k + b_k} \right)^{1/d} + \left(\prod_{k=1}^{d} \frac{b_k}{a_k + b_k} \right)^{1/d} \\ &\leq \frac{1}{d} \sum_{k=1}^{d} \frac{a_k}{a_k + b_k} + \frac{1}{d} \sum_{k=1}^{d} \frac{b_k}{a_k + b_k} \qquad \text{(AM-GM inequality)} \\ &= 1. \end{aligned}$$

Step 2: Suppose $A = A_1 \sqcup \ldots \sqcup A_m$ and $B = B_1 \sqcup \ldots \sqcup B_n$ are finite unions of standard closed parallelepipeds with pairwise disjoint interiors. When $m = n = 1$ we have already proved the theorem. By induction on $m + n$, we shall prove it for all $m, n \geq 1$. This is the cleverest part of the proof.

Translating A or B does not change any of the quantities in the inequality, hence we may freely do so. Assume $m \geq 2$ without loss of generality (else interchange A and B).

Claim: There is at least one axis direction $j \leq d$ and a number $t \in \mathbb{R}$ such that each of the sets $A' := A \cup \{x : x_j \leq t\}$ and $A'' := A \cap \{x : x_j < t\}$ are both unions of atmost $m - 1$ standard parallelepipeds with pairwise disjoint interiors.

Proof of the claim: Let $R_1 = [a_1, b_1] \times \ldots \times [a_d, b_d]$ and $R_2 = [p_1, q_1] \times \ldots \times [p_d, q_d]$ be two among the parallelepipeds that comprise A. If $I_j = [a_j, b_j] \cap [p_j, q_j]$, then $I_1 \times \ldots \times I_d \subseteq R_1 \cap R_2$. But R_1 and R_2 have disjoint interiors, hence I_j must be empty or be a singleton for some j. This means $b_j \leq t \leq p_j$ or $q_j \leq t \leq a_j$, and we set $t = b_j$ or $t = q_j$ accordingly. The hyperplane $\{x : x_j = t\}$ will do the job, since R_1 will lie on one side of it and R_2 on the other (the boundary of both may intersect the hyperplane). The claim is proved.

Set $\lambda = |A'|/|A|$. By the above claim, $0 < \lambda < 1$ and each of A' and A'' is a disjoint union of at most $m - 1$ parallelepipeds (with sides parallel to the axes). Now translate B along the jth direction, i.e., for each s consider $B_s := B + s\mathrm{e_j}$ and let $B'_s = B_s \cap \{x : x_j \le t\}$ and $B''_s = B_s \cap \{x : x_j \ge t\}$. Choose a value of s such that $|B'_s| = \lambda |B|$ and set $B' = B'_s$ and $B'' = B''_s$.

By the induction hypothesis,

$$|A' + B'| \ge \left(|A'|^{1/d} + |B'|^{1/d}\right)^d = \lambda \left(|A|^{1/d} + |B|^{1/d}\right)^d,$$
$$|A'' + B''| \ge \left(|A''|^{1/d} + |B''|^{1/d}\right)^d = (1 - \lambda) \left(|A|^{1/d} + |B|^{1/d}\right)^d.$$

Further, observe that $A' + B' \subseteq \{x : x_j \le 2t\}$ and $A'' + B'' \subseteq \{x : x_j \ge 2t\}$ and hence $|(A' + B') \cap (A' + B')| = 0$, the intersection being contained in the hyperplane $\{x : x_j = t\}$. Therefore,

$$\begin{aligned}|A + B| &= |A' + B'| + |A'' + B''| \\ &= \lambda \left(|A|^{1/d} + |B|^{1/d}\right)^d + (1 - \lambda) \left(|A|^{1/d} + |B|^{1/d}\right)^d \\ &= \left(|A|^{1/d} + |B|^{1/d}\right)^d.\end{aligned}$$

This completes the proof when A, B are finite unions of standard parallelepipeds.

Step 3: Let A and B be compact sets. Let $Q = [-1, 1]^d$ and fix $\varepsilon > 0$. Observe that compactness of A implies that there exist $x_1, \ldots, x_n \in A$ (for some n) such that $A \subseteq A''$ where $A'' = \cup_{i=1}^n (x_i + \varepsilon Q)$. It is easy to see that $A'' \subseteq A_{\varepsilon\sqrt{d}}$ and that A'' may be written as a finite union of standard rectangles whose interiors are pairwise disjoint. Similarly find $B'' = \cup_{j=1}^m (y_j + \varepsilon Q)$ that is a union of standard rectangles whose interiors are pairwise disjoint and such that $B \subseteq B'' \subseteq B_{\varepsilon\sqrt{d}}$.

Then, observe that $A'' + B'' \subseteq (A + B)_{2\sqrt{d}\varepsilon}$. Since A'' and B'' are finite unions of standard parallelepipeds, by the previous case, we know that Brunn-Minkowski inequality applies to them. Thus,

$$\begin{aligned}|(A + B)_{2\sqrt{d}\varepsilon}| &\ge |A'' + B''| \\ &\ge (|A''|^{1/d} + |B''|^{1/d})^d \\ &\ge (|A|^{1/d} + |B|^{1/d})^d.\end{aligned}$$

This is true for every $\varepsilon > 0$. As $A + B$ is compact we see that $\cap_{\varepsilon > 0} (A + B)_{2\sqrt{d}\varepsilon} = A + B$ and hence $|(A + B)_{2\sqrt{d}\varepsilon}| \downarrow |A + B|$ as $\varepsilon \downarrow 0$. Therefore, Brunn-Minkowski inequality holds true when A and B are compact.

Step 4: Let A and B be general Borel sets. If either of A or B has infinite Lebesgue measure, there is nothing to prove. Otherwise, by regularity of Lebesgue measure, there are compact sets $A' \subseteq A$ and $B' \subseteq B$ such that $|A \setminus A'| < \varepsilon$ and $|B \setminus B'| < \varepsilon$. Then of course $A + B \supseteq A' + B'$ and hence

$$|A + B|^{1/d} \ge |A' + B'|^{1/d} \ge |A'|^{1/d} + |B'|^{1/d} \ge (|A| - \varepsilon)^{1/d} + (|B| - \varepsilon)^{1/d}.$$

Letting $\varepsilon \to 0$ we get the inequality for A and B. ■

Remark 8. If we do not assume that $A+B$ is measurable, then (see the last step) we still get

$$m_*(A+B)^{1/d} \geq |A|^{1/d} + |B|^{1/d}$$

where m_* is the inner Lebesgue measure, $m_*(C) := \sup\{m(K) : K \subseteq B,\ K \text{ compact}\}$. We shall see in the next section that $A+B$ is not necessarily measurable.

Exercise 9. Let K be a bounded convex set in $\mathbb{R}^d$. Fix a unit vector $u \in \mathbb{R}^d$ and let $K^t := \{x \in K : \langle x, u\rangle = t\}$ denote the sections of K for any $t \in \mathbb{R}$. Let $I = \{t : |K_t| > 0\}$ and let $f : I \mapsto \mathbb{R}$ be defined by $f(t) = |K_t|^{1/(n-1)}$. Show that I is an interval and that f is concave. **[Note:** Here $|K_t|$ denotes the $(d-1)$-dimensional Lebesgue measure of K_t in the hyperplane $\{x \in \mathbb{R}^d : \langle x, u\rangle = t\}$.**]**

3. Measurability questions

We want to exhibit measurable sets $A, B \subseteq \mathbb{R}$ such that $A+B$ is not measurable. In fact we shall produce an example with $B = A$. This construction is due to Sierpinski[10] and may also be taken simply as a construction of a non-measurable set (quite different from the one usually presented in measure theory class).

Step 1: Let $K \subseteq [0,1]$ be the usual $1/3$-set of Cantor. Then $K + K \supseteq [0,2]$.

To see this, recall that Cantor set consists of numbers whose ternary expansion has digits 0 and 2 (but not 1). Hence, if $x, y \in \frac{1}{2}\cdot K = \{u/2 : u \in K\}$, then $x = \sum_{i=1}^{\infty} \frac{x_i}{3^i}$ and $y = \sum_{i=1}^{\infty} \frac{y_i}{3^i}$ with $x_i, y_i \in \{0,1\}$. Now consider any $t \in [0,1]$ and write $t = \sum_{i=1}^{\infty} \frac{t_i}{3^i}$ where $t_i \in \{0,1,2\}$. Clearly, we can find $x_i, y_i \in \{0,1\}$ such that $x_i + y_i = t_i$ for each i. Thus, a given $t \in [0,1]$ can be written as $x+y$ with $x, y \in \frac{1}{2}\cdot K$ and hence a number in $[0,2]$ can be written as a sum of two elements of K.

Step 2: Regard $\mathbb{R}$ as a vector space over $\mathbb{Q}$. Then the first step says that the span of K is $\mathbb{R}$. Hence, by a standard application of Zorn's lemma, there exists a basis $B \subseteq K$ for the vector space.

Step 4: Define $E_0 = B \sqcup (-B) \sqcup \{0\}$ and $E_n = E_{n-1} + E_{n-1}$ for $n \geq 1$. From the previous step, it follows that

$$\bigcup_{n\geq 0}\bigcup_{q\geq 1} \frac{1}{q} E_n = \mathbb{R}.$$

Indeed, given $x \in \mathbb{R}$, write it as $x = r_1 b_1 + \ldots + r_n b_n$ with $n \geq 1$, $r_i \in \mathbb{Q}$, $b_i \in B$. Taking q to be the product of the denominator of r_is, we get $x = \frac{1}{q}(p_1 b_1 + \ldots + p_n b_n)$ with $p_i \in \mathbb{Z}$. Negating b_i if necessary (it will still be in E_0), we may assume $q \geq 1$ and $p_i \geq 1$.

Step 5: Let m be the smallest n for which $m^*(E_n) > 0$. Since E_0 is a subset of a set of zero measure, $m \geq 1$. Hence it makes sense to set $A = E_{m-1}$. Then A is Lebesgue measurable (since its outer measure is zero). We claim that $A + A = E_m$ is not Lebesgue measurable.

[10] We have taken this presentation from Rubel's paper *A pathological Lebesue measurable function*.

Indeed, if E_m was measurable, by Steinhaus' lemma, $E_m + E_m$ contains an open interval around 0 (since E_m is symmetric, $E_m + E_m$ is the same as $E_m - E_m$). But then E_{m+1} contains an interval around 0. Thus, given any $x \in \mathbb{R}$, we can find $q \geq 1$ so that $x/q \in E_{m+1}$. The conclusion is that every element of $\mathbb{R}$ can be written as a linear combination of at most 2^{m+1} distinct elements of B. But B is an infinite set (i.e., $\mathbb{R}$ has infinite dimensions over $\mathbb{Q}$), and hence $b_1 + \ldots + b_k \notin E_{m+1}$ if $k > 2^{m+1}$ and b_i are distinct elements of B. This contradiction can only be resolved by accepting that E_m cannot be measurable. ■

Remark 10. There is also an example to show that the Minkowski sum of Borel sets need not be Borel. However, the sum-set will necessarily be Lebesgue measurable.

Here is two simpler facts, one of which was used in the proof of Brunn-Minkowski inequality.

Exercise 11. Show that the Minkowski sum of two compact sets is $\mathbb{R}$ is necessarily compact. Show that the Minkowski sum of two closed sets in $\mathbb{R}$ need not be closed.

4. FUNCTIONAL FORM OF ISOPERIMETRIC INEQUALITY

It is a general idea that a statement about sets must have an analogous statement for functions and vice versa. When the function is taken to be the indicator of a set the functional inequality should reduce to the inequality for sets. This may not make sense immediately as there may be assumptions of smoothness etc., that are not satisfied by indicator functions, but in some approximate sense this should hold. Here is the functional analogue of the isoperimetric inequality.

Theorem 12 (Sobolev inequality). *Let $d \geq 2$ and $p = \frac{d}{d-1}$. Then, for every $f \in C_c^1(\mathbb{R}^d)$, we have*

$$\|f\|_{L^p} \leq \|\nabla f\|_{L^1}.$$

In what way is this analogous to the isoperimetric inequality? If f is the indicator of a bounded open set, its derivative is zero in interior of the set and in the interior of the complement. All change occurs at the boundary. As such a measure of the total change can be considered a measure of the boundary of the set. Transferring this to smooth functions, we would say that any inequality of the form $\|f\|_p \leq C_{d,p,q}\|\nabla f\|_q$ valid for all $f \in C_c^\infty(\mathbb{R}^d)$ and for some constant $C_{d,p,q}$, is a functional analogue of the isoperimetric inequality.

If such an inequality holds for some p and q, then start with $f \in C_c^\infty(\mathbb{R}^d)$ and set $f_s(x) = f(sx)$ for $s > 0$. Then,

$$\|f_s\|_p = s^{-\frac{d}{p}}\|f\|_p \quad \text{and} \quad \|\nabla f_s\|_{L^q} = s^{1-\frac{d}{q}}\|\nabla f\|_q.$$

Since the inequality must hold for f_s (with the same constant $C_{n,p,q}$), it follows that $s^{1-\frac{d}{q}+\frac{d}{p}}$ must be bounded as a function of s. Seeing the limits as s goes to 0 or ∞, we see that it is necessary to have $\frac{1}{q} - \frac{1}{p} = \frac{1}{d}$. Remarkably this condition is sufficient as the theorem below show. When $q = 1$, we get $p = d/(d-1)$, which is the special case of Sobolev's inequality above.

Theorem 13 (Gagliardo-Nirenberg-Sobolev inequality). *Suppose $p, q > 0$ satisfy $\frac{1}{q} - \frac{1}{p} = \frac{1}{d}$. Then $\|f\|_p \le C_{d,p,q} \|\nabla f\|_q$ valid for all $f \in C_c^\infty(\mathbb{R}^d)$*

We first give the proof of Sobolev's inequality and explain the modifications needed to obtain the more general case later. The idea of proof is explained easily when $d = 2$.

Proof for $d = 2$. Let f_i denote the ith partial derivative. Then, for any $(x, y) \in \mathbb{R}^2$, we have

$$f(x,y) = \int_{-\infty}^{x} f_1(s,y)ds \implies |f(x,y)| \le \int_{\mathbb{R}} |f_1(s,y)|ds.$$

$$f(x,y) = \int_{-\infty}^{y} f_2(x,t)dt \implies |f(x,y)| \le \int_{\mathbb{R}} |f_1(x,t)|dt.$$

Multiplying the two inequalities, we get

$$|f(x,y)|^2 \le \left(\int_{\mathbb{R}} |f_1(s,y)|ds \right) \left(\int_{\mathbb{R}} |f_2(x,t)|dt \right).$$

Integrate over x and y and observe that the right hand side factors

$$\int_{\mathbb{R}^2} |f(x,y)|^2 dxdy \le \left(\int_{\mathbb{R}^2} |f_1(s,y)|dsdy \right) \left(\int_{\mathbb{R}^2} |f_2(x,t)|dtdx \right)$$

Since $\|\nabla f\| = \sqrt{f_1^2 + f_2^2}$, we have $|f_1| \le \|\nabla f\|$ and $|f_2| \le \|\nabla f\|$, and therefore

$$\int_{\mathbb{R}^2} |f|^2 \le \left(\int_{\mathbb{R}^2} \|\nabla f\| \right)^2$$

which is precisely the claim of Sobolev inequality for $d = 2$. ■

The proof for $d \ge 3$ needs a little more work.

Proof for $d = 3$. We have as before

$$|f(x_1, x_2, x_3)| \le \int |f_1(s_1, x_2, x_3)|ds_1 =: I_1(x_2, x_3)$$

$$|f(x_1, x_2, x_3)| \le \int |f_2(x_1, s_2, x_3)|ds_2 =: I_2(x_1, x_3)$$

$$|f(x_1, x_2, x_3)| \le \int |f_3(x_1, x_2, s_3)|ds_3 =: I_3(x_1, x_2).$$

We again multiply them and write

$$|f(x_1, x_2, x_3)|^3 \le I_1(x_2, x_3) \times I_2(x_1, x_3) \times I_3(x_1, x_2).$$

But if we integrate with respect to x_1, x_2, x_3, the right side does not split as a product of integrals - this is the difference from the case $d = 2$. Note also that the power of f on the left is 3 while the desired inequality should have $3/2$. Take square roots and integrate over x_1 alone. We get

$$\begin{aligned}\int_{\mathbb{R}} |f(x_1,x_2,x_3)|^{3/2}dx_1 &\le I_1(x_2,x_3)^{1/2}\int_{\mathbb{R}} I_2(x_1,x_3)^{1/2} I_3(x_1,x_2)^{1/2}dx_1\\ &\le I_1(x_2,x_3)^{1/2}\left(\int_{\mathbb{R}} I_2(x_1,x_3)dx_1\right)^{1/2}\left(\int_{\mathbb{R}} I_3(x_1,x_2)dx_1\right)^{1/2}\\ &=: I_1(x_2,x_3)^{1/2}J(x_3)^{1/2}K(x_2)^{1/2}\end{aligned}$$

by Cauchy-Schwarz. Now integrate over x_2. The second factor is independent of x_2.We apply Cauchy-Schwarz again to get

$$\begin{aligned}\int_{\mathbb{R}^2} |f(x_1,x_2,x_3)|^{3/2}dx_1dx_2 &\le J(x_3)^{1/2}\left(\int_{\mathbb{R}} I_1(x_2,x_3)dx_2\right)^{1/2}\left(\int_{\mathbb{R}} K(x_2)dx_2\right)^{1/2}\\ &=: J(x_3)^{1/2}L(x_3)^{1/2}\left(\int_{\mathbb{R}} K(x_2)dx_2\right)^{1/2}\end{aligned}$$

Now integrate over x_3 and apply Cauchy-Schwarz again to the first two factors to get

$$\int_{\mathbb{R}^3} |f(x_1,x_2,x_3)|^{3/2}dx_1dx_2dx_3 \le \left(\int K(x_2)dx_2\right)^{1/2}\left(\int_{\mathbb{R}} J(x_3)dx_3\right)^{1/2}\left(\int_{\mathbb{R}} L(x_3)dx_3\right)^{1/2}.$$

It remains to observe that

$$\int_{\mathbb{R}} K(x_2)dx_2 = \int_{\mathbb{R}^3}|f_3|, \quad \int_{\mathbb{R}} J(x_3)dx_3 = \int_{\mathbb{R}^3}|f_2|, \quad \int_{\mathbb{R}} L(x_3)dx_3 = \int_{\mathbb{R}^3}|f_1|.$$

Each of the three is bounded by $\|\nabla f\|_{L^1}$ and hence,

$$\int_{\mathbb{R}^3} |f(x_1,x_2,x_3)|^{3/2}dx_1dx_2dx_3 \le \left(\int_{\mathbb{R}^3}\|\nabla f(x_1,x_2,x_3)\|dx_1dx_2dx_3\right)^{3/2}$$

This is the Sobolev inequality for $d=3$. ■

Exercise 14. Write the proof for general d. **[Hint:** Use Hölder's inequality in place of Cauchy-Schwarz where necessary.]

4.1. **Proof of the GNS inequality.** This section is written by Raghavendra Tripathi.

Assume that $\frac{1}{p} = \frac{1}{q} - \frac{1}{d}$. Let $f \in C_c^\infty(\mathbb{R}^d)$ and set $g := |f|^r$. Note that $|\nabla g| = r|f|^{r-1}|\nabla f|$. Apply Sobolev's inequality to g to get

$$\left(\int |f|^{\frac{dr}{d-1}}\right)^{\frac{d-1}{d}} \le r\int |f|^{r-1}|\nabla f|.$$

Since we want $|f|^p$ in the left integral, it is clear that we should set $r = \frac{p(d-1)}{d}$. With this choice of r, and applying Hölder's inequality on the right hand side, we get the following:

$$\left(\int |f|^p\right)^{\frac{d-1}{d}} \le r\left(\int |f|^{(r-1)\frac{q}{q-1}}\right)^{\frac{q-1}{q}}\|\nabla f\|_q. \tag{1}$$

Observe that $(r-1) = \frac{dp-p-d}{d}$, while $\frac{q}{q-1} = \frac{dp}{dp-d-p}$. Therefore (1) becomes

$$\|f\|_p^{\frac{p(d-1)}{d}} \leq r\|f\|_p^{\frac{q-1}{q}} \|\nabla f\|_q,$$

which gives the desired inequality with the constant $C_{d,p} = r = \frac{p(d-1)}{d}$. ■

5. Functional form of Brunn-Minkowski inequality

What is the functional form of the Brunn-Minkowski inequality? The latter involves volumes of A, B and $A+B$. That volume must be replaced by the integral of a function is clear. But what is the analogue of Minkowski sum for functions?

The first idea that comes to mind is the convolution. If A and B are bounded open sets, it is easy to see that $\mathbf{1}_A \star \mathbf{1}_B$ is positive on $A+B$ and zero on the complement. However, $\int \mathbf{1}_A \star \mathbf{1}_B = |A| \times |B|$ (in general $\int f \star g = \int f \times \int g$) and not $|A+B|$. Indeed, the convolution counts points of $A+B$ with multiplicity (meaning how many ways to write $z \in A+B$ as $x+y$ with $x \in A$ and $y \in B$), while $\mathbf{1}_{A+B}$ counts the same points without multiplicity. The following operation turns out to be the right analogue of Minkowski sum.

Fix $0 < t < 1$. For $f, g : \mathbb{R}^n \mapsto \mathbb{R}_+$, define

$$(f\#_t g)(z) = \sup\{f(x)^t g(y)^{1-t} : z = tx + (1-t)y\}.$$

These notations are for convenience and not the convention in any sense. Then observe that for any t and any sets $A, B \subseteq \mathbb{R}^n$, we have $\mathbf{1}_A \#_t \mathbf{1}_B = \mathbf{1}_{tA+(1-t)B}$. In this sense, $\#_t$ is a good functional form of Minkowski sum[11] With this, we present the functional analogue of the Brunn-Minkowski inequality.

Theorem 15 (Prékopa-Leindler inequality). *Let $f, g, h : \mathbb{R}^n \mapsto \mathbb{R}_+$ be measurable functions and let $0 < t < 1$ be fixed. Assume that $h \geq f\#_t g$ a.e. on $\mathbb{R}^n$. Then $\int h \geq (\int f)^t (\int g)^{1-t}$.*

Some remarks are in order.

(1) The reason why we simply don't write the theorem simply as $\int (f\#_t g) \geq (\int f)^t (\int g)^{1-t}$ is that $(f\#_t g)$ may not be measurable even though f and g are.

(2) The Prékopa-Leindler inequality is a kind of reverse to Hölder's inequality. Indeed, if we write $h_0 = f\#_t g$ (ignore the measuability question) and $h_1(z) = f(z)^t g(z)^{1-t}$, then it is clear that $h_0 \geq h_1$ (the supremum defining $h_0(z)$ contains the case $x = y = z$). What Hölder's inequality says is that $(\int f)^t (\int g)^{1-t} \geq \int h_1$, while Prékopa-Leindler says that $(\int f)^t (\int g)^{1-t} \leq \int h_0$.

(3) Brunn-Minkowski inequality can be deduced from Prékopa-Leindler. Indeed, we apply the latter to $f = \mathbf{1}_A$ and $g = \mathbf{1}_B$ and $h = \mathbf{1}_{tA+(1-t)B}$. For any $0 < t < 1$, the conditions are satisfied and we get $|tA + (1-t)B| \geq |A|^t |B|^{1-t}$. Apply this to $\frac{1}{t}A$ and $\frac{1}{1-t}B$ to see

[11] To think: Many others suggest themselves. Why not $\sup\{tf(x) + (1-t)f(y) : z = tx + (1-t)y\}$? I don't know what is the problem if we do this.

that $|A+B| \ge |A|^t|B|^{1-t}t^{-nt}(1-t)^{-n(1-t)}$. Optimize the right hand side over t (helps to take logarithms before differentiating!) to see that the optimal $t = \frac{|A|^{1/n}}{|A|^{1/n}+|B|^{1/n}}$ (hence $1-t = \frac{|B|^{1/n}}{|A|^{1/n}+|B|^{1/n}}$). A little calculation gives us $|A+B| \ge (|A|^{1/n}+|B|^{1/n})^n$.

Now we prove the Prékopa-Leindler inequality. There are two steps, firstly the statement for $n=1$, and secondly an induction step. The second step is almost trivial (the cleverness is already in the formulation of the statement).

Case $n=1$**:** Fix any $u>0$ and observe that if $f(x)>u$ and $g(y)>u$, then $h(tx+(1-t)y)>u$. In other notations, $t\{f>u\}+(1-t)\{g>u\} \subseteq \{h>u\}$. By the Cauchy-Davenport inequality (the one-dimensional Brunn-Minkowski), it follows that $|\{h>u\}| \ge |\{f>u\}|+|\{g>u\}|$. Integrate with respect to u from 0 to infinity to see that $\int h \ge t\int f + (1-t)\int g$. Here we used the fact that for a non-negative function f we have $\int f = \int_0^\infty |\{f>t\}|/$ But $tx+(1-t)y \ge x^t y^{1-t}$ (weighted AM-GM inequality). Thus the theorem is true in one dimension.

Case $n\ge 2$**:** Write $x,y,z \in \mathbb{R}^n$ as $(x',x''), (y',y''), (z',z'')$ with $x',y',z' \in \mathbb{R}^{n-1}$ and $x'',y'',z'' \in \mathbb{R}$. Given f,g,h as in the theorem, fix $x',y' \in \mathbb{R}^{n-1}$, set $z' = \theta x' + (1-\theta)y'$ and observe that the one-variable functions $f(x',\cdot)$, $g(y',\cdot)$, $h(z',\cdot)$ satisfy exactly the same assumptions. Therefore, by the case $n=1$, we conclude that

$$\int_{\mathbb{R}} h(\theta x' + (1-\theta)y', z'')\, dz'' \ge \left(\int_{\mathbb{R}} f(x',x'')dx''\right)^t \left(\int_{\mathbb{R}} g(y',y'')dy''\right)^{1-t}.$$

Now define $F,G,H : \mathbb{R}^{n-1} \mapsto \mathbb{R}$ by

$$F(x') = \int_{\mathbb{R}} f(x',x'')dx'', \quad G(y') = \int_{\mathbb{R}} g(y',y'')dy'', \quad H(z') = \int_{\mathbb{R}} h(z',z'')dz''.$$

What the previous inequality shows is that F,G,H satisfy the hypothesis of Prékopa-Leindler inequality. Inductively if we assume that the inequality has been proved for dimension $n-1$, then we conclude that

$$\int_{\mathbb{R}^{n-1}} H(z')dz' \ge \left(\int_{\mathbb{R}^{n-1}} F(x')dx'\right)^t \left(\int_{\mathbb{R}^{n-1}} G(y')dy'\right)^{1-t}.$$

But $\int_{\mathbb{R}^{n-1}} H(z')dz' = \int_{\mathbb{R}^n} h(z)dz$ and similarly for the other two integrals. Thus the above inequality is the same as what we set out to prove. ■

6. Proof of isoperimetric inequality by symmetrization

Using symmetrization techniques introduced by Steiner and induction on the dimension, we give a proof of the isoperimetric inequality[12].

[12] This proof is taken from the appendix to a paper of Figiel, Lindestrauss and Milman, where they prove the isoperimetric inequality on the sphere. They modeled it on a well-known proof for the Euclidean case which is written in many books but since I did not find one, I translated back their proof to the Euclidean case. Hence, there may be avoidable complications in the proof.

Theorem 16. *Let A be a compact subset of $\mathbb{R}^d$ and let B be a closed ball with $|A| = |B|$. Then, $|A_\varepsilon| \geq |B_\varepsilon|$ for all $\varepsilon > 0$.*

The theorem is obvious in one dimension. Indeed, if $M = \max A$ and $m = \min A$, then $A_\varepsilon \setminus A$ contains the intervals $(M, M+\varepsilon)$ and $(m-\varepsilon, m)$ and hence has measure at least 2ε. But B is an interval and clearly $|B_\varepsilon| = |B| + 2\varepsilon$. Thus, $|A_\varepsilon| \geq |B_\varepsilon|$ for all $\varepsilon > 0$.

Next we introduce the key notion of symmetrization. Given a unit vector $\mathbf{e} \in \mathbb{R}^{\mathbf{d}}$, let $\ell = \mathbb{R}\mathbf{e}$ (a line) and a set A, we define the sections of A as $A^t := A \cap (\ell^\perp + t\mathbf{e})$ for $t \in \mathbb{R}$ (the intersection of A with the affine hyperplane that orthogonal to ℓ and at a distance of t from the origin).

Definition 17. Given a line $\ell = \mathbb{R}\mathbf{e}$ and a compact set A, define the symmetrization of A with respect to ℓ as the set C such that: For any $t \in \mathbb{R}$, the section $C^t := C \cap (t + \ell^\perp)$ is the closed $(d-1)$-dimensional disk in the affine hyperplane $t + \ell^\perp$. Further, the center of C^t is on ℓ and the $(d-1)$-dimensional volume of C^t is the same as that of $A^t := A \cap (t + \ell^\perp)$. To be unambiguous, we adopt the following convention: If A^t is empty, then C^t is defined to be empty. If A^t is non-empty but has zero $(d-1)$-dimnesional Lebesgue measure, then C^t is defined to be a singleton. The resulting set C is denoted as $\sigma_\ell(A)$.

Exercise 18. Show that $\sigma_\ell(A)$ is compact for any compact A.

For compact A, let

$$\mathcal{M}(A) = \left\{ C \subseteq \mathbb{R}^d : C \text{ is compact, } |C| = |A|, \ |C_\varepsilon| \geq |A_\varepsilon| \text{ for all } \varepsilon > 0 \right\}.$$

These are the sets that are better than A in isoperimetric sense. Theorem 16 is equivalent to saying that $\mathcal{M}(A)$ contains a ball. The main idea of the proof is that the symmetrization of a set has better isoperimetric profile than the original set.

Lemma 19. *Let A be a compact subset of $\mathbb{R}^d$. Then, $\sigma_\ell(A) \in \mathcal{M}(A)$ for any line ℓ.*

Observe that $\sigma_\ell(A)$ is a set which is symmetric about the axis ℓ. Thus one may expect that by symmetrizing about various lines, the set becomes rounder and rounder, and approach a ball. The lemma assures us that the isoperimetric profile only gets better in the process. But a finite number of operations may not get to a ball. For a rigorous argument, we use an auxiliary functional on sets. Let the radius of a compact set be defined by

$$r(A) = \inf\{r > 0 : B(x, r) \supseteq A \text{ for some } x \in \mathbb{R}^d\}.$$

This will be used as follows.

Lemma 20. *If A is a compact subset that is not a ball, then there exist lines $\ell_1, \ldots, \ell_m$ (for some m) such that $\sigma_{\ell_1} \circ \ldots \circ \sigma_{\ell_m}(A)$ has strictly smaller radius than A.*

The isoperimetric inequality is an easy consequence of the previous two lemmas together with the next one.

Lemma 21. *Let A be a compact set. Then r attains its minimum on $\mathcal{M}(A)$.*

Proof of Theorem 16. Fix A and let B be a minimizer of r on $\mathcal{M}(A)$ (by Lemma 21). If B is not a ball, by Lemma 20 there is a sequence of symmetrizations that strictly reduce the radius. By Lemma 19, the resulting set is still in $\mathcal{M}(A)$, contradicting that B is a minimum of $r(\cdot)$ inside $\mathcal{M}(A)$. ■

It remains to prove the lemmas.

Proof of Lemma 20. Fix A and let $B = B(x, r(A))$ contain A. Take any line ℓ passing through x and symmetrize to get A_1. The ball remains fixed under symmetrization. Since $B \setminus A$ contains an open ball, $\partial B \setminus A_1$ contains a cap $C \subseteq \partial B$ (by cap, we mean a ball inside ∂B in spherical metric).Now pick a line ℓ_1 passing through x and a boundary point of C and symmetrize to get A_2. Then, draw a picture and convince yourself that $\partial B \setminus A_2$ contains a cap C' with radius double that of C. Continuing to reflect on further lines $\ell_2, \ell_3, \ldots$, in a finite number of steps we get to a set A_m such that $A_m \cap \partial B = \emptyset$. Then $r(A_m) < r(A)$. ■

Proof of Lemma 21. First we claim that r is continuous. In fact it is Lipschitz, i.e., $|r(A_1) - r(A_2)| \le d_H(A_1, A_2)$. This is because $\varepsilon > d_H(A_1, A_2)$ and $B(x, r) \supseteq A_1$ implies that $B(x, r+\varepsilon) \supseteq A_2$.

We next claim that $\mathcal{M}(A)$ is closed. To see this, let $C_n \in \mathcal{M}(A)$ and $C_n \to C$ in Hausdorff metric. Then $C_\varepsilon \supseteq C_n$ for large n showing that $|C_\varepsilon| \ge \limsup_{n\to\infty} |C_n| = |A|$. Put $\varepsilon = 1/k$ and note that $\cap_{k\ge 1} C_{1/k} = C$ (as C is compact) to get $|C| = \lim_{k\to\infty} |C_{1/k}| \ge |A|$. Further, for any $\delta > 0$ we have $C \subseteq (C_n)_\delta$ for large n and hence $|C| \le |(C_n)_\delta| \le |A_\delta|$. Now put $\delta = 1/k$ and use $A = \cap_k A_{1/k}$ to get $|C| \le |A|$. We have now proved that $|C| = |A|$. Next fix $\varepsilon > 0$ and $\delta > 0$ and observe that $|C_\varepsilon| \le \liminf |(C_n)_{\varepsilon+\delta}| \le |A_{\varepsilon+\delta}|$ since $C \subseteq (C_n)_\delta$ for large n. Put $\delta = 1/k$ and let $k \to \infty$ to get $|C_\varepsilon| \le |\bar{A}_\varepsilon|$, since $\cap_k A_{\varepsilon+1/k} = \bar{A}_\varepsilon$. Thus, $|C_\varepsilon| \le |\bar{A}_\varepsilon|$ for every $\varepsilon > 0$. Use this for $\varepsilon - 1/k$ and take union over k. Since $C_{\varepsilon - k^{-1}}$ increase to C_ε and $\overline{A_{\varepsilon-k^{-1}}}$ increase to A_ε, taking limits we get $|C_\varepsilon| \le |A_\varepsilon|$ for any $\varepsilon > 0$.

Since $A \in \mathcal{M}(A)$, in minimizing r we may restrict ourselves to $\{C \in \mathcal{M}(A) : r(C) \le r(A)\}$. Translation does not change isoperimetric profile, hence it suffices to $\mathcal{M}_0(A) = \{C \in \mathcal{M}(A) : C \subseteq B(0, r(A))\}$. But $\mathcal{M}_0(A)$ is a compact set (see Exercise 22) and r is continuous, there must be a minimum. ■

Exercise 22. Let (X, d) be a compact metric space. Then $(\mathcal{C}, d_H)$, the space of closed subsets endowed with Hausdorff metric, is also compact.

The following proof is easier understood with pictures, but I don't have time to draw some.

Some notation used in the following proof: Without loss of generality we shall take the line to be $\ell = \mathbb{R}e_d$ (where $e_d = (0, \ldots, 0, 1)$). For $t \in \mathbb{R}$, let $\tau_t(A) = A + te_d$ (translation in "vertical" direction). We use λ_d to denote the d-dimensional Lebesgue measure λ_{d-1} to denote the lower dimensional Lebesgue measure on any affine hyperplane in $\mathbb{R}^d$ (particularly on the hyperplane $\ell^\perp + te_d = \{x \in \mathbb{R}^d : x_d = t\}$).

Proof of Lemma 19. Fix A and ℓ and let $C = \sigma_\ell(A)$. Since $\lambda_{d-1}(C^t) = \lambda_{d-1}(A^t)$ for all t, it follows that $\lambda_d(A) = \lambda_d(C)$. This is because $\lambda_d(A) = \int_{\mathbb{R}} \lambda_{d-1}(A_t)dt$ and similarly for C.

It remains to compare $\lambda_d(C_\varepsilon)$ with $\lambda_d(A_\varepsilon)$. The sections of A_ε get contributions from many different sections of A. In fact,

$$(A_\varepsilon)^t = \bigcup_{s:|s-t|\le\varepsilon} (\tau_{t-s}[A^s])_{\sqrt{\varepsilon^2-(t-s)^2}}. \tag{2}$$

The notation does not show this, but the neighbourhoods on the right are taken inside the hyperplane $\ell^\perp + te_d$. Analogously,

$$(C_\varepsilon)^t = \bigcup_{s:|s-t|\le\varepsilon} (\tau_{t-s}[C^s])_{\sqrt{\varepsilon^2-(t-s)^2}}.$$

A key observation is that for fixed t, the sets on the right are concentric balls in H_t, hence there is at least one s for which $(\tau_{t-s}[C^s])_{\sqrt{\varepsilon^2-(t-s)^2}}$ is equal to the whole set of $(C_\varepsilon)^t$.

For that s, we use the inequality

$$\lambda_{d-1}((\tau_{t-s}[C^s])_{\sqrt{\varepsilon^2-(t-s)^2}}) \le \lambda_{d-1}((\tau_{t-s}[A^s])_{\sqrt{\varepsilon^2-(t-s)^2}}). \tag{3}$$

This inequality follows inductively (we assume the validity of Theorem 16 for dimension $d-1$) and using $|C^s| = |A^s|$ (which implies $|\tau_{t-s}[C^s]| = |\tau_{t-s}[A^s]|$, of course). In (3), the left side is equal to $\lambda_{d-1}((C_\varepsilon)^t)$ by the choice of s, while the right hand side is at most $\lambda_{d-1}((A_\varepsilon)^t)$ by (2). Thus,

$$\lambda_{d-1}((C_\varepsilon)^t) \le \lambda_{d-1}((A_\varepsilon)^t).$$

Integrate over t to get $\lambda_d(C_\varepsilon) \le \lambda_d(A_\varepsilon)$. Thus, we have proved that $C \in \mathcal{M}(A)$. ■

Remark 23. As remarked earlier, this proof is taken from a paper of Figiel, Lendenstrauss and Milman where they prove isoperimetric inequality in the sphere $\mathbb{S}^{n-1}$. Brunn-Minkowski inequality does not make sense in the sphere (there is no addition operation) but the above proof by symmetrization goes though virtually identically, with spherical metric replacing the Euclidean metric and symmetrization done w.r.t. great circles in place of straight lines. One difference is in the proof of Lemma 19, where $\sqrt{\varepsilon^2-(t-s)^2}$ has to be replaced by some function of ε, t, s (it is not required to know what this function precisely is!). A lesser point is that in the proof of Lemma 21, the whole collection $\mathcal{M}(A)$ is compact (since the sphere is itself compact), and there is no need to bring in $\mathcal{M}_0$.

Functional form of symmetrization: It is a valid an interesting question to ask about the analogue of symmetrization for functions. In the proof above, the symmetrization was done section by section. Here we simply consider symmetrization in a fixed dimension.

Let $f : \mathbb{R}^n \mapsto \mathbb{R}_+$ be a measurable function such that $|\{f > t\}| < \infty$ for any $t > 0$. For example, f could be a continuous function that vanishes at infinity. Now define a new function $f^* : \mathbb{R}^n \mapsto \mathbb{R}_+$ having the following properties:

(1) f^* is radial and decreasing. That is, $f^*(x) = g(|x|)$ where $g : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is a decreasing function.

(2) $|\{f^* > t\}| = |\{f > t\}|$ for all $t > 0$.

(3) g as defined above is a left-continuous function.

The last condition is simply put in to ensure uniqueness. The key point is that f^* is a radial, decreasing and has its super-level sets have the same measure as those of f. Such a function f^* does exist (why?).

As an example, if A is a bounded set with positive measure, then $\mathbf{1}_A^*$ is equal to the indicator of the closed ball $\overline{B(0,r)}$ where r is chosen so that $|A| = |B(0,r)|$. In this sense this is the right generalization to functions. Further, the key point of symmetrization, that the ball is isoperimetrically better than the original set, carries over to functions in the following sense.

Theorem 24. *Let f be a smooth function vanishing at infinity. Then, $\|\nabla f^*\|_{L^1} \le \|\nabla f\|_{L^1}$.*

This is only one of many *rearrangement inequalities* that are widely used in analysis. We shall not explore this topic further for lack of time[13].

7. Concentration of measure

Isoperimetric inequalities are closely related to a phenomenon called measure concentration - an important topic in probability, analysis and high dimensional geometry. We just introduce the basic notions here.

Let us assume the isoperimetric inequality on spheres. As remarked earlier, it can be proved by symmetrization. However, let us be precise about what the actual statement is.

Let $\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}$ denote the $(n-1)$-dimensional sphere. It is a metric space with metric d_n inherited from $\mathbb{R}^n$ and define A_ε as the ε-neighbourhood of A in the metric d_n. And balls are also defined with respect to d_n. One can also use the metric adapted to the sphere (measuring distance along geodesics), but for the purposes of this section the two are equivalent since they are monotone functions of each other. Apart from distance, we also need the right measure which we discuss next.

Orthogonal matrices of order $n \times n$ preserve $\mathbb{S}^{n-1}$. There is a unique probability measure μ_{n-1} on $\mathbb{S}^{n-1}$ that is invariant under the action of orthogonal matrices. That is, $\mu_{n-1}(A) = \mu_{n-1}(P.A)$ for any $A \in \mathcal{B}(\mathbb{S}^{n-1})$ and any $P \in O(n)$. Here $P.A = \{Px : x \in A\}$. We shall refer to it as the uniform measure on $\mathbb{S}^{n-1}$. It can be defined in many ways, but one quick way is as follows: Let γ_n be the Borel probability measure on $\mathbb{R}^n$ defined as $d\gamma_n(x) = \frac{1}{(2\pi)^{n/2}} e^{-|x|^2} dx$. Define the projection map $\Pi_n : \mathbb{R}^n \setminus \{0\} \mapsto \mathbb{S}^{n-1}$ as $\Pi_n(x) = x/\|x\|$. Then define $\mu_n = \gamma_n \circ \Pi_n^{-1}$ as the push-forward of γ_n under Π_n. Since $\gamma_n\{0\} = 0$, the measure μ_n has total mass 1. Its invariance under orthogonal transformations comes from the corresponding property of γ_n.

[13] The old book of Hardy, Littlewood and Pólya titled *Inequalities*, has a couple of chapters devoted to this topic. Another book is that of Lieb and Loss titled *Analysis*. This latter book has a remarkable selection of topics in analysis.

Exercise 25. Show that $\gamma_n(P.A) = \gamma_n(A)$ for all $A \in \mathcal{B}(\mathbb{R}^n)$ and for all $P \in O(n)$.

With these definitions, here is the statement of the isoperimetric inequality.

Theorem 26. *Let A is a measurable subset of $\mathbb{S}^{n-1}$ and let B be a ball. If $\mu_n(A) = \mu_n(B)$, then $\mu_n(A_\varepsilon) \geq \mu_n(B_\varepsilon)$ for any $\varepsilon > 0$.*

Now we come to the concentration of measure phenomenon. Suppose $A \subsetneq \mathbb{S}^{n-1}$ have $\mu_n(A) = \frac{1}{2}$. Then $\mu_n(A) = \mu_n(B)$ where $B = \{x \in \mathbb{S}^{n-1} : x_1 \leq 0\}$ (a hemisphere). Therefore, by the isoperimetric inequality, we see that $\mu_n(A_\varepsilon) \geq \mu_n(B_\varepsilon)$ for any $\varepsilon > 0$. The quantity $\mu_n(B_\varepsilon)$ can be calculated explicitly. Before doing the computation, observe that $\mu_n(B_\varepsilon^c)$ is the measure of those $x \in \mathbb{S}^{n-1}$ for which $x_1 \geq \varepsilon$. For large n, we should expect that most of the x_k are of order $1/\sqrt{n}$ (since $x_1^2 + \ldots + x_n^2 = 1$ and there is symmetry in co-ordinates), hence $\mu_n(B_\varepsilon^c)$ ought to be small. It is remarkably small, as we see now by explicit computation.

Indeed, by our definition of μ_n,

$$\begin{aligned} \mu_n(B_\varepsilon) &= \gamma_n\{x \in \mathbb{R}^n : x_1 < \varepsilon\} \\ &= 1 - \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \mathbf{1}_{x_1 \geq \varepsilon} e^{-\frac{1}{2}(x_1^2 + \ldots + x_n^2)} dx_1 \ldots dx_n \quad = 1 - \int_\varepsilon^\infty \end{aligned}$$

An elementary calculation shows that

$$\mu_n(B_\varepsilon) = 1 - \frac{1}{Z_n} \int_\varepsilon^1 (1 - t^2)^{\frac{1}{2}(n-2)}\, dt$$

where $Z_n = \int_{-1}^1 (1-t^2)^{\frac{1}{2}(n-1)} dt$. By a change of variable, one sees that $Z_n = \text{Beta}(\frac{1}{2}(n+1), \frac{1}{2}(n+1))$. Writing it in terms of Beta functions and using Stirlings' approximation, it is easy to see that $Z_n \geq n^{-\alpha}$ for some α. Consequently,

Another formulation of concentration of measure is in terms of Lipschitz functions. Suppose $F : \mathbb{S}^{n-1} \mapsto \mathbb{R}$ is a 1-Lipschitz function, i.e.,$|F(x) - F(y)| \leq d_n(x, y)$. Then the diameter of the image of F can be as large as 2 (since the diameter of the sphere is 2). However, from the point of view of measure μ_n, the function F is nearly constant! How is that? Define M such that $\mu_n\{F \geq M\} \geq \frac{1}{2}$ and $\mu_n\{F \leq M\} \geq \frac{1}{2}$ (such an M always exists, it is called a *median*). Then, let $A = \{F \geq M\}$ and observe that $A_\varepsilon \supseteq \{F \geq M - \varepsilon\}$ by the Lipschitz property of F. By the Lévy concentration inequality, we see that $\mu_n\{F \geq M - \varepsilon\} \geq 1 - Ce^{-cn\varepsilon^2}$ for some universal constants C, c. Applying this to $-F$, we see that $\mu_n\{F \leq M + \varepsilon\} \geq 1 - Ce^{-cn}$. Taking intersections, we find that

$$\mu_n\{x \in \mathbb{S}^{n-1} : M - \varepsilon \leq F \leq M + \varepsilon\} \geq 1 - Ce^{-cn\varepsilon^2}.$$

Thus, on most part of the sphere, F takes values very close to M.

8. Gaussian isoperimetric inequality

Let γ_n be the standard Gaussian measure on $\mathbb{R}^n$ given by $d\gamma_n(x) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\|x\|^2} dx$ where dx. This is a Borel probability measure. Some of the most basic properties of γ_n are given in the exercise below.

Exercise 27. (1) Show that $\gamma_n(P.A) = \gamma_n(A)$ for all $A \in \mathcal{B}(\mathbb{R}^n)$ and for all $P \in O(n)$.

(2) If $\Pi_d : \mathbb{R}^n \mapsto \mathbb{R}^d$ is the projection $\Pi_d(x_1, \ldots, x_n) = (x_1, \ldots, x_d)$ (here $d \le n$), show that $\gamma_n \circ \Pi_d^{-1} = \gamma_d$.

The first part shows that the measure is invariant under a change of orthonormal basis. In that sense, it is not associated to a co-ordinate system but to the Hilbert space structure of $\mathbb{R}^n$ itself. The second fact, taken together with the first, shows that the marginal (push-forward under projection) on any d-dimensional subspace is the standard Gaussian measure on that subspace.

Now we state the Gaussian isoperimetric inequality. Recall that a half-space is a set of the form $\{x \in \mathbb{R}^n : \langle x, u\rangle \le t\}$ for some unit vector $u \in \mathbb{R}^n$ and some $t \in \mathbb{R}$. These take the place of balls as the optimal sets in the isoperimetric sense.

Theorem 28 (Borell, Sudakov-Tsirelson). *Let $A \in \mathcal{B}(\mathbb{R}^n)$ and the half-space H be such that $\gamma_n(A) = \gamma_n(H)$. Then $\gamma_n(A_\varepsilon) \ge \gamma_n(H_\varepsilon)$.*

While one can build proofs analogous to the proofs we gave for Euclidean space or the sphere, our intention is to introduce a useful idea that relates the Lebesgue measure on the sphere and Gaussian measure. Indeed, the original proofs of the Gaussian isoperimetric inequality were deduced in this way.

Lemma 29 (Maxwell? Poincaré?). *Let $\Pi_{n,d} : S^{n-1} \mapsto \mathbb{R}^d$ be defined by $\Pi_{n,d}(x_1, \ldots, x_n) = \sqrt{n}(x_1, \ldots, x_d)$. Let μ_n be the uniform probability measure on S^{n-1}. Then, $\mu_n \circ \Pi_{n,d}^{-1}(A) \to \gamma_d(A)$ for all $A \in \mathcal{B}(\mathbb{R}^d)$.*

The usual notion of convergence of Borel probability measures on $\mathbb{R}^d$ is weak convergence: $\nu_n \to \nu$ weakly if for any bounded continuous function $f \in C_b(\mathbb{R}^d)$, we have $\int f d\nu_n \to \int f d\nu$ as $n \to \infty$. This is equivalent to the statement that $\nu_n(A) \to \nu(A)$ for those $A \in \mathcal{B}(\mathbb{R}^d)$ for which $\nu(\partial A) = 0$. The sense of convergence in the lemma above is stronger. It comes from convergence of densities, as shown by the following general fact.

Lemma 30 (Scheffe's lemma). *Let ν_n, ν be Borel probability measures on $\mathbb{R}^d$ having densities f_n, f with respect to Lebesgue measure. If $f_n \to f$ a.e. (w.r.t. Lebesgue measure), then $\nu_n(A) \to \nu(A)$ for all $A \in \mathcal{B}(\mathbb{R}^d)$.*

Proof. Write $|f_n - f| = f_n - f + (f - f_n)_+$ (where $x_+ = \max\{x, 0\}$). Fix any $A \in \mathcal{B}(\mathbb{R}^d)$. Since $(f - f_n)_+$ is dominated by f and goes to zero $a.e.$, it follows by the dominated convergence theorem that $\int_A (f - f_n)_+ \to 0$ as $n \to \infty$. Therefore, $\int_A |f_n - f| \to 0$ (by the relationship with $(f - f_n)_+$ and the fact that $\int f_n = \int f = 1$). Thus $\int_A f_n \to \int_A f$ for any Borel set A. ■

One could state this lemma in greater generality on any measure space with probability measures ν_n, ν whose densitites with respect to some fixed measure λ converge almost everywhere. The proof is the same.

Proof of Lemma 29. Let f_n be the density of $\mu_n \circ \Pi_{n,d}^{-1}$. If $(t_1, \ldots, t_d) \in \mathbb{R}^d$, then the pre-image of it under $\Pi_{n,d}$ is the set $\{x \in S^{n-1} : x_i = \frac{t_i}{\sqrt{n}},\ i \le d\}$, which is a $n - 1 - d$ dimensional sphere with radius $\sqrt{1-\frac{1}{n}(t_1^2+\ldots+t_d^2)}$. From this, it is clear that (the first factor comes from the $\sqrt{n}$ scaling included in the projection $\Pi_{n,d}$)

$$f_n(t) = \frac{1}{n^{\frac{d}{2}}} \frac{\sigma_{n-d}}{\sigma_n} \left(1 - \frac{t_1^2+\ldots+t_d^2}{n}\right)^{\frac{1}{2}(n-1-d)}.$$

Plug in the standard formula $\sigma_d = d\pi^{d/2}/\Gamma(\frac{d}{2}+1)$. Now let $n \to \infty$ and use the fact that $\frac{\Gamma(\frac{n}{2}+1)}{\Gamma(\frac{n-d}{2}+1)} \to \frac{1}{2^{\frac{d}{2}}}$ (from Stirlings' formula, for example) to get

$$f_n(t) \to \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2}(t_1^2+\ldots+t_d^2)}.$$

By Scheffe's lemma, $\mu_n \circ \Pi_{n,d}^{-1}(A) \to \gamma_d(A)$ for all Borel sets A (or even measurable sets). ■

Exercise 31. Compute the exact density of x_1^2 and joint density of (x_1^2, x_2^2), etc., where $x = (x_1, \ldots, x_n)$ has the distribution μ_n on S^{n-1}.

Now we deduce the Gaussian isoperimetric inequality from the isoperimetric inequality on spheres.

Proof of Theorem 28. Let $A \in \mathcal{B}(\mathbb{R}^d)$ with $\gamma_d(A) = \alpha$. Assume $0 < \alpha < 1$ (the cases when $\alpha = 0$ and $\alpha = 1$) are trivial. Fix $\beta < \alpha$ and choose a half-space $H = \{x : x_1 \le t\}$ such that $\gamma_d(H) = \beta$. Define $A_n = \Pi_{n,d}^{-1}(A)$ and $H_n = \Pi_{n,d}^{-1}(H)$, subsets of S^{n-1}. By Lemma 29 we know that $\mu_n(A_n) \to \alpha$ and $\mu_n(H_n) \to \beta$ as $n \to \infty$. Consequently, $\mu_n(A_n) \ge \mu_n(H_n)$ for all large n. Further, $H_n = \{x \in S^{n-1} : x_1 \le \frac{t}{\sqrt{n}}\}$ is a ball in S^{n-1}. Consequently, by the isoperimetric inequality on the sphere, it follows that $\mu_n((A_n)_{\varepsilon/\sqrt{n}}) \ge \mu_n((H_n)_{\varepsilon/\sqrt{n}})$. Note that these enlargements are in the metric on the sphere (which we take to be the distance inherited from $\mathbb{R}^n$ in the standard embedding of S^{n-1} in $\mathbb{R}^n$). However, $\Pi_{n,d}$ is a Lipschitz map with Lipschitz constant $\sqrt{n}$. Therefore, $\Pi_{n,d}^{-1}(A_\varepsilon) \supseteq (A_n)_{\varepsilon/\sqrt{n}}$. Therefore,

$$\mu_n(\Pi_{n,d}^{-1}(A_\varepsilon)) \ge \mu_n((H_n)_{\varepsilon/\sqrt{n}}).$$

But then $(H_n)_{\varepsilon/\sqrt{n}} = \{x \in S^{n-1} : x_1 \le \frac{s}{\sqrt{n}}\}$ where s is related to t, ε by

■

9. Some recent developments

There are many aspects of the isoperimetric inequality that are still being studied. We mention just two that we are aware of.

Stability version: Suppose A is a compact set in $\mathbb{R}^n$ with the same volume as a ball B and having only an ε more surface area than B has. Is A necessarily close to a ball? Closeness could be measured in Hausdorff distance, for example.

Double bubble conjecture: Suppose positive numbers $a_1, \dots, a_k$ are specified. Among all collections of compact sets $A_1, \dots, A_k$ having disjoint interiors and having volumes $|A_j| = a_j$ for $1 \leq j \leq k$, is there one that minimizes the surface area of the union $A_1 \cup \ldots \cup A_k$? What is that configuration?

10. Alexandrov-Fenchel inequalities

If $K_1, \dots, K_n$ are convex bodies in $\mathbb{R}^n$, then

$$m_n(t_1 K_1 + \ldots + t_n K_n) = \sum_{i_1, \dots, i_n = 1}^{n} V[K_{i_1}, \dots, K_{i_n}] t_{i_1} \ldots t_{i_n}$$

where $V[K_1, \dots, K_n]$ are called mixed volumes. It is symmetric, positive and $V[K, \dots, K] = m_n(K)$. Alexandrov-Fenchel inequalities state that

$$V[K_1, K_2, K_3, \dots, K_n]^2 \geq V[K_1, K_1, K_3, \dots, K_n] V[K_2, K_2, K_3 \dots, K_n].$$

Many other equivalent forms, for example, $t \mapsto V[K_1 + tK_2, \dots, K_1 + tK_2, K_{m+1}, \dots, K_n]^{1/m}$ is concave on $\mathbb{R}_+$ (note that there are no $K_3, \dots, K_m$ in this expression). When $m = n$ this gives Brunn-Minkowski inequality.

CHAPTER 4

Matching theorem and some applications

1. THREE THEOREMS IN COMBINATORICS

1.1. **Hall's matching theorem.** We recall some basic notions. A *graph* $G = (V, E)$ is a set V ("vertex set") together with an edge-set E where E is a symmetric relation on the set V (i.e., $E \subseteq V \times V$ and $(u, v) \in E$ implies $(v, u) \in E$). We shall also assume that the relation is anti-reflexive, i.e., $(u, u) \notin E$ for any $u \in E$. In such a case, we shall be loose in our language and say that $\{u, v\}$ is an edge or write $u \sim v$ and say u is adjacent to v. Also, we say that the edge $\{u, v\}$ is incident to the vertices u and v.

When we talk about a *directed graph*, the symmetry condition on E is dropped (and the convention is to say that the edge (u, v) is directed from u towards v) but we shall still require it to be anti-reflexive. Clearly, any undirected graph can also be considered as a directed graph.

A *bipartite graph* is an undirected graph whose vertex set V can be partitioned into V_1 and V_2 such that if $u \sim v$, then $u \in V_1, v \in V_2$ or $u \in V_2, v \in V_1$.

A (complete) *matching* of a graph is a collection of edges such that every vertex is adjacent to exactly one edge in the collection.

In a graph, $G = (V, E)$, let $N(A) = \{v \in V : v \sim u \text{ for some } u \in A\}$ be the neighbourhood of A.

Theorem 1 (Hall's marriage theorem). *Let $G = (V, E)$ be a finite bipartite graph with parts V_1, V_2 of equal cardinality. Then G has a complete matching if and only if $|N(A)| \geq |A|$ for all $A \subseteq V_1$.*

We shall give multiple proofs of the marriage theorem. It is related to other theorems of a similar flavour (similar in the sense that the most natural obstacle to achieving something is shown to be the only obstacle) such as *Dilworth's theorem* and *Ford-Fulkerson max-flow min-cut theorem*, etc. We shall first state Dilworth's theorem and derive Hall's theorem from it and also give a direct proof of Hall's theorem. In a later section we shall derive it from the max-flow min-cut theorem. The latter theorem can be given a direct proof, but we shall derive it from a more sophisticated viewpoint which gives a chance to introduce minimax theorems that are of much importance in many fields and also to make connection to convexity and duality.

1.2. **Dilworth's theorem.** Let $\mathcal{P}$ be a partially ordered set. Recall that this means that there is a relation (denoted "$\leq$") on $\mathcal{P}$ that is reflexive ($x \leq x$ for all $x \in \mathcal{P}$), anti-symmetric ($x \leq y$ and $y \leq x$ imply $x = y$) and transitive ($x \leq y$ and $y \leq z$ imply $x \leq z$). It will be convenient to write the reversed relation as "$\geq$" (i.e., $x \geq y$ if $y \leq x$).

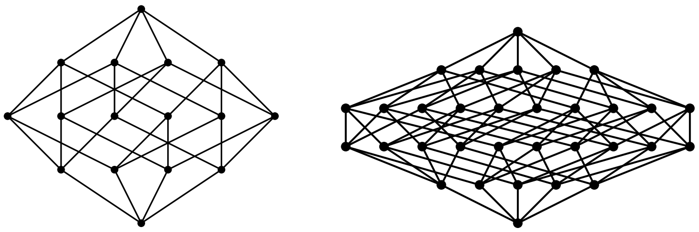

FIGURE 4. Poset of subsets of $\{1,2,3,4\}$ and of $\{1,2\ldots,5\}$.

A *chain* is a totally ordered subset in $\mathcal{P}$. An *anti-chain* (also called *independent set*) is a subset in which no two distinct elements are comparable. Suppose we write the poset as a union of chains C_j for $j \in J$ (some index set). If A is any anti-chain in $\mathcal{P}$, it can put at most one point in each of the chains C_j. Therefore, $|A| \le |J|$ (in these sections $|A|$ will denote the cardinality of A).

Example 2. The collection of all subsets of a given set is a poset with the order given by set inclusion is a poset. For example, if the given set is $\{1,2,3\}$, then $C_1 = \{\emptyset, \{1\}, \{1,2\}, \{1,2,3\}\}$, $C_2 = \{\{2\},\{2,3\}\}$ and $C_3 = \{\{3\},\{1,3\}\}$ are all chains and $C_1 \cup C_2 \cup C_3 = \mathcal{P}$.

Theorem 3. *If m is the maximal size of an anti-chain in a finite poset, then the poset can be written as a union of m chains.*

Use Dilworth's theorem to solve a famous problem first posed by Erdös and Szekeres.

Exercise 4. Let $N = mn+1$ and $a_1, \ldots, a_N$ be distinct real numbers. Then, there exists an increasing subsequence of cardinality $n+1$ or a decreasing subsequence of cardinality $m+1$ (or both).

In many interesting posets, it is hard to find the size of the maximal anti-chain. A very beautiful example is that of the Boolean poset consisting of subsets of $\{1,2,\ldots,n\}$, with ordering given by set-inclusion. It is clear that A_k, the collection of all subsets with a given cardinality k, is an anti-chain. Since $|A_k| = \binom{n}{k}$, among these anti-chains, the maximal size if $\binom{n}{\lfloor n/2 \rfloor}$. It is a beautiful result of Sperner that in fact this is the maximal size among all anti-chains of the Boolean poset. We outline it as an exercise.

The language used will be simpler if you imagine the Hasse diagram of the Boolean poset as shown in the figures above. At the bottom is the empty set and at the top is the whole set.

Exercise 5. Let an ant start at the empty set and moves upward by picking an element uniformly at random (among all elements in the layer immediately above that are connected by an edge in the Hasse diagram). At each step the picks are made independently. After n steps, the ant is at the top.

For any set $A \subseteq \{1,2,\ldots,n\}$, let $p(A)$ be the probability that the ant passes through the vertex A. Calculate $p(A)$. What can you say about $p(A_i)$s if $A_1,\ldots,A_m$ is an anti-chain?

Put these together to prove Sperner's lemma.

Here is another standard application of Dilworth's theorem. Let G be any graph. A *vertex cover* is any subset of vertices such that every edge is incident to at least one of the vertices in the subset. A *matching* (now we use it to mean incomplete matchings) is a collection of edges such that no vertex of the graph is incident to more than one edge in the collection.

Exercise 6. In a finite bipartite graph, show that the maximal number of edges in any matching is equal to the minimal number of vertices in any vertex cover. This is known as König's theorem.

1.3. **Proof of Hall's theorem and some consequences.** Here is how Dilworth's theorem implies Hall's theorem.

Proof of Hall's matching theorem. Given a bipartite graph as in the statement of Hall's theorem, define a partial order of V by declaring $u \le v$ if $u \in V_1$, $v \in V_2$ and $u \sim v$ in the graph. We claim that the maximal size of an anti-chain is $n := |V_1| = |V_2|$. Indeed, if A is an anti-chain, $N(A \cap V_1)$ is disjoint from $A \cap V_2$ (if not, there is a vertex in $A \cap V_1$ that is adjacent to a vertex in $A \cap V_2$, contradicting the anti-chain property). Their cardinalities sum to at most $|V_2| = n$. Thus, using Hall's condition, $|A \cap V_1| \le N(A \cap V_1)|$, we have

$$|A| = |A \cap V_1| + |A \cap V_2| \le |N(A \cap V_1)| + |A \cap V_2| \le n.$$

On the other hand, we do have anti-chains of cardinality n (eg., V_1 or V_2). Thus the size of a maximal anti-chain is precisely n.

By Dilworth's theorem, we can write V as a union of n chains. As $|V| = 2n$ and each chain has cardinality at most 2, this means that V is a union of n pairs $\{u, v\}$ with $u \le v$ (i.e., $u \in V_1$, $v \in V_2$ and $u \sim v$). That is precisely the matching that we want. ■

As a useful consequence of Hall's theorem, we derive a theorem of Birkoff and von Neumann that every doubly stochastic matrix is a convex combination of permutation matrices. Recall that a doubly stochastic matrix is a square matrix having non-negative entries and whose row and column sums are all equal to 1. The space DS_n of all $n \times n$ doubly stochastic matrices is easily seen to be a convex set. It is also compact (as a subset of $\mathbb{R}^{n^2}$). If K is a compact convex set in $\mathbb{R}^d$, then a point is said to be an extreme point of K if it cannot be written as a strict convex combination of two distinct points in K and the set of all extreme points of K is denoted by $E(K)$. In other words, $x \in E(K)$ if and only if $x \in K$ and $x = \alpha y + (1-\alpha)z$ for some $0 < \alpha < 1$ and $y, z \in K$ implies that $y = z$.

A well-known theorem of Krein and Milman states that for any non-empty compact convex set K in $\mathbb{R}^d$ is the convex hull of its extreme points. That is

$$K = \text{conv}(E(K)) := \left\{ \sum_{i=1}^{n} \alpha_i x_i : n \ge 1, x_i \in E(K), \alpha_i \ge 0 \text{ and } \sum_{i=1}^{n} \alpha_i = 1 \right\}.$$

In fact, Krein-Milman theorem is valid in general locally convex spaces, except that we must take the closure on the right. That is $K = \overline{\text{conv}(E(K))}$.

Example 7. The space of probability measures on $\mathbb{R}$ is a convex set whose extreme points are δ_a, $a \in \mathbb{R}$. The space of probability measure whose mean exists and is equal to 0 is also a convex set. Its extreme points are δ_0 and $\frac{b}{a+b}\delta_{-a} + \frac{a}{a+b}\delta_b$ for some positive a, b. In general, the set of measures with specified m moments will form a convex set. What are its extreme points?

The set DS_n is convex and compact. Its extreme points are precisely the set of permutation matrices (we had trouble justifying this in class, but it follows from the proof below) and then Krein-Milman would imply that all such matrices are convex combinations of permutation matrices. We show this directly, invoking Hall's theorem.

Theorem 8 (Birkoff-von Neumann theorem)**.** *Every doubly stochastic matrix is a convex combination of permutation matrices.*

Proof. Let $A \in DS_n$. Define a bipartite graph with V_1 being the set of rows of A and V_2 being the set of columns of A and put an edge between ith row and jth column if and only if $a_{i,j} > 0$. If $R_1, \ldots, R_k$ are any k rows, the $k \times n$ matrix formed by these rows has a total sum of k (each row sums to 1) and hence the sum of all the column sums is k. Since each column sum is at most 1, there must be at least k con-zero columns. Therefore, $|N(S)| \geq |S|$ for $S = \{R_1, \ldots, R_k\}$. This shows the validity of Hall's conditions, and hence there is a matching of rows and columns in this bipartite graph.

Denote the matching by $i \sim \pi(i)$ where π is a permutation. Let $\alpha = \min\{a_{i,\pi(i)} : i \leq n\}$ which is positive. If P_π denotes the permutation matrix with 1s at $(i, \pi(i))$, then the matrix $A - \alpha P$ has row and column sums equal to $1 - \alpha$. If $\alpha = 1$, then $A = P$ and we are done. If $\alpha < 1$, we can rescale it to a doubly stochastic matrix and write $A = \alpha P + (1-\alpha)B$ where $B \in DS_n$. Note that B has at least one more zero entry than A. Continue to write B as $\beta Q + (1-\beta)C$ where Q is a permutation and C is a doubly stochastic matrix, etc. The process must terminate as the number of zeros in the doubly stochastic matrix increases by at least 1 in each step. We end with a representation of A as a convex combination of permutation matrices. ■

1.4. **Proof of Dilworth's theorem.** The proof will be by induction on the cardinality of the poset. Check the base case yourself.

Let $\mathcal{P}$ be a finite poset and let $a_1, \ldots, a_m$ be an anti-chain of maximal cardinality in $\mathcal{P}$. Then, for any $x \in \mathcal{P}$, there is some i such that $x \leq a_i$ or $a_i \leq x$ (otherwise $\{a_1, \ldots, a_m, x\}$ would be a larger anti-chain). Hence, if we define

$$\mathcal{P}_- = \{x : x \leq a_i \text{ for some } i \leq m\}, \quad \mathcal{P}_+ = \{x : a_i \leq x \text{ for some } i \leq m\},$$

then $\mathcal{P} = \mathcal{P}_- \cup \mathcal{P}_+$. Both $\mathcal{P}_-$ and $\mathcal{P}_+$ are posets and $\{a_1, \ldots, a_m\}$ is an anti-chain in both. If we could argue that these two posets had strictly smaller cardinality than $\mathcal{P}$, then inductively we could write them as unions of m chains:

$$\mathcal{P}_+ = C_1^+ \cup \ldots \cup C_m^+, \quad \mathcal{P}_- = C_1^- \cup \ldots \cup C_m^-$$

where each $C_i^\pm$ is a chain and $a_i \in C_i^\pm$. Since a_i is a maximal element in $\mathcal{P}_-$ (and hence in C_i^-) and a minimal element in $\mathcal{P}_+$ (and hence in C_i^+), it follows that $C_i = C_i^+ \cup C_i^-$ is a chain in $\mathcal{P}$. This gives the decomposition $\mathcal{P} = C_1 \cup \ldots \cup C_m$ of the given poset into chains.

The gap in the proof is that $\mathcal{P}_+$ or $\mathcal{P}_-$ could be all of $\mathcal{P}$ (clear, but give an explicit example) and hence induction does not help.

To fix this problem, we first take a maximal chain C_0 in $\mathcal{P}$ and set $\mathcal{Q} = \mathcal{P} \setminus C_0$. Then $\mathcal{Q}$ is strictly smaller than $\mathcal{P}$.

Case 1: Suppose $\mathcal{Q}$ has an anti-chain of size m, say $\{a_1, \ldots, a_m\}$. Now take this anti-chain in the argument outlined earlier (for the poset $\mathcal{P}$, now we may forget $\mathcal{Q}$). The proof is now legitimate because $\mathcal{P}_+$ does not contain the minimal element of C_0 (else $C_0 \cup a_i$ would be a chain for some i and $a_i \notin C_0$). Similarly $\mathcal{P}_-$ does not contain the maximal element of C_0. Both $\mathcal{P}_+$ and $\mathcal{P}_-$ have strictly smaller cardinality and hence the induction hypothesis applies. We get a chain decomposition of $\mathcal{P}$ as above.

Case 2: Suppose that the maximal cardinality of a chain in $\mathcal{Q}$ is m' which is strictly smaller than m (then $m' = m - 1$ in fact). Then write $\mathcal{Q}$ (by induction hypothesis) as a union of m' chains. Together with C_0 this decomposes $\mathcal{P}$ into m chains.

This completes the proof. ■

1.5. **A direct proof of the marriage theorem.** This proof is taken from a paper of Halmos and Vaughan, and just like in that paper we shall state it in a more general form (which is not used in the sequel).

Theorem 9 (Hall's marriage theorem, general form). *Let G be a bipartite graph with V_1, V_2 being the two parts of the vertex set (these are now allowed to be infinite sets). Assume that $|N(A)| \geq |A|$ for all finite $A \subseteq V_1$. Then there is a matching of V_1 into V_2, i.e., there exists an injective function $f : V_1 \mapsto V_2$ such that $x \sim f(x)$ for all $x \in V_1$.*

Proof of the finite case. The proof is by induction on $n = |V_1|$. When $n = 1$, the hypothesis immediately gives a partner for the sole element of V_1 and hence the claim is true. Assume that the theorem is proved when $|V_1| < n$ and consider the case when $|V_1| = n$. The idea is to marry off a proper subset of V_1 first, and then marry of the remaining, using the induction hypothesis. Care is needed to make sure that Hall's condition is preserved in the reduced graphs.

Case 1: Assume that $|N(A)| \geq |A| + 1$ for all proper subsets $A \subsetneq V_1$. In this case, we pick an element $x \in V_1$, an element $y \in V_2$ such that $y \sim x$ and set $V_1' = V_1 \setminus \{x\}$ and $V_2' = V_2 \setminus \{y\}$. The subgraph G' of the given graph with vertex set $V_1' \sqcup V_2'$ is claimed to satisfy Hall's condition.

Granting that, the induction hypothesis applies and we find a matching $g : V_1' \mapsto V_2'$. Then define $f : V_1 \mapsto V_2$ by $f(x) = y$ and $f(t) = g(t)$ for $t \in V_1'$. Clearly a matching.

Case 2: There is a proper subset $A \subsetneq V_1$ such that $|N(A)| = |A|$. In this case, the subgraph with vertex set $A \sqcup N(A)$ must satisfy Hall's condition (since any subset of A has the same neighbourhood in the subgraph as in the original graph). Inductively there is a matching g of A into $N(A)$, but since the cardinalities are equal, g is in fact a bijection from A onto $N(A)$. Now consider the subgraph with vertex set $V_1' \sqcup V_2'$ where $V_1' = V_1 \setminus A$ and $V_2' = V_2 \setminus N(A)$. If $A_1 \subseteq V_1'$ and $N'(A_1)$ is its neighbourhood in the subgraph, then $N(A \sqcup A_1) = N(A) \sqcup N'(A_1)$. Since G satisfies Hall's condition, it follows that $|N'(A_1)| \geq |A_1|$. Thus Hall's condition is satisfied on the subgraph and a matching $h : V_1' \mapsto V_2'$ exists. Setting $f = g$ on A and $f = h$ on V_1' gives the matching of the original graph. ■

Proof of the infinite case. Let $\mathcal{F} = \{f : V_1 \mapsto V_2 : f(x) \sim x \text{ for all } x \in V_1\}$. One can clearly identify $\mathcal{F}$ as the Cartesian product of the sets $N(\{x\})$ as x varies over V_1. For $A \subseteq V_1$, let $\mathcal{F}_A$ be the collection of f that are injective on A. By the finite version of Hall's theorem (applied to the subgraph with vertex set $A \sqcup N(A)$ or if you are squeamish that $N(A)$ could be infinite, observe that you may choose a sufficiently large subset of $N(A)$ and still preserve Hall's condition), it follows that $\mathcal{F}_A$ is not empty for any finite A. The collection $\mathcal{F}_A$, $|A| < \infty$, has the finite intersection property (since $\mathcal{F}_A \cap \mathcal{F}_B = \mathcal{F}_{A \cup B}$). If we can find a topology on $\mathcal{F}$ in which all the sets $\mathcal{F}_A$ (for finite A) are compact, then the finite intersection property implies that the complete intersection of $\mathcal{F}_A$, over all $|A| < \infty$, is not empty. But such a function is clearly injective (for any $x, y \in V_1$, since $f \in \mathcal{F}_{\{x,y\}}$, it follows that $f(x) \neq f(y)$), and hence gives a matching.

Since $N(\{x\})$ is finite for each x, if we endow it with the discrete topology it becomes compact. By Tychonoff's theorem $\mathcal{F}$ is compact, and so are $\mathcal{F}_A$ for all finite A. ■

The following example shows why we had to assume that $N(\{x\})$ is finite for all x.

Example 10. Consider a bipartite graph with $V_1 = \{1, 2, 3, \ldots\}$, $V_2 = \{1', 2', 3', \ldots\}$ and edges between $k+1$ and k' for all $k \geq 1$ and between 1 and j' for all $j \geq 1$. In this case, Hall's conditions are satisfied. Indeed, if $A \ni 1$, then $N(A)$ is an infinite set and if $A \not\ni 1$, then $|N(A)| = |A|$. However, any attempt at matching forces $2 \mapsto 1'$, $3 \mapsto 2'$, $4 \mapsto 3', \ldots$, leaving no possible match for 1 even though he/she has taken great efforts to know everyone of the opposite sex.

Wherever Tychonoff's theorem is used, if we restrict to an appropriate countable setting, it can be replaced by a diagonal argument.

Exercise 11. Assume that V_1 is countable. Use diagonal argument in place of Tychonoff's theorem to prove the existence of a matching.

2. Haar measure on topological groups

Topological groups: A *topological group* is a group G endowed with a Hausdorff topology such that the operations $(xy) \mapsto xy$ (from $G \times G$ to G) and $x \mapsto x^{-1}$ (from G to G) are continuous.

As examples, we may take any finite or countable group (with discrete topology), the group $(\mathbb{R}^n, +)$, the group $GL(n, \mathbb{R})$ of $n \times n$ invertible matrices with real entries, similarly $GL(n, \mathbb{C})$, the unitary group $\mathcal{U}(n)$, the orthogonal group $O(n)$, various other subgroups of matrices (all with topology inherited from $\mathbb{R}^{n^2}$ or $\mathbb{C}^{n^2}$), the group $M_n(\mathbb{R})$ of isometries of $\mathbb{R}^n$ (which can be built from the translation group $\mathbb{R}^n$, the "rotation group" $O(n)$ and reflections $x \mapsto -x$), group of isometries of Hyperbolic space, groups constructed by taking products such as $(\mathbb{Z}/(2))^J$ for an arbitrary index set J, etc.

Another kind of example (for the sole purpose of giving an exercise): For a graph $G = (V, E)$, by an automorphism of G we mean a bijection $f : V \mapsto V$ such that $u \sim v$ if and only if $f(u) \sim f(v)$. The set of all such automorphisms $\text{Aut}(G)$ forms a group under composition. The graph is said to be *transitive* if for any $u, v \in V$, there exists $f \in \text{Aut}(G)$ such that $f(u) = v$. Examples of transitive graphs are $\mathbb{Z}^d$, lattices, regular trees, etc. Give the topology of pointwise convergence on $\text{Aut}(G)$. If the graph is rooted (i.e., one vertex is distinguished), then the automorphism is required to fix the root. They too form a group (an obvious subgroup of $\text{Aut}(G)$).

Exercise 12. If G is any transitive group with a countable vertex set where each vertex has finite degree , show that the group of automorphisms fixing the root vertex is compact.

Exercise 13. Identify the automorphism group of the rooted infinite binary tree shown in Figure 2.

Invariant measures: On a topological group we may talk of the Borel sigma-algebra and measures on it. We have seen some of these.

- On $\mathbb{R}^n$ we have the Lebesgue measure λ_n with the property that $\lambda_n(A + x) = \lambda_n(A)$ for all $A \in \mathcal{B}(\mathbb{R}^n)$ and for all $x \in \mathbb{R}^n$. Any constant multiple of λ_n also has this property of *translation-invariance*, and no other measure does.
- On $\mathbb{R}_+ = (0, \infty)$ with multiplication, define the measure $d\mu(x) = \frac{dx}{x}$. Check that $\mu(xA) = \mu(A)$ for all $A \in \mathcal{B}(\mathbb{R}_+)$ and for all $x \in \mathbb{R}_+$. For example,

$$\mu(a, b) = \int_a^b \frac{1}{x} dx = \log(b/a)$$

which is clearly the same as $\mu(2a, 2b)$.

- On $GL(n, \mathbb{R})$, define the measure $d\mu(X) = |\det X|^{-n} dX$ where dX denotes Lebesgue measure on $\mathbb{R}^{n^2}$ (of which $GL(n\mathbb{R})$ is an open set). Then if $A \in GL(n, \mathbb{R})$, the map $X \mapsto A.X$ has

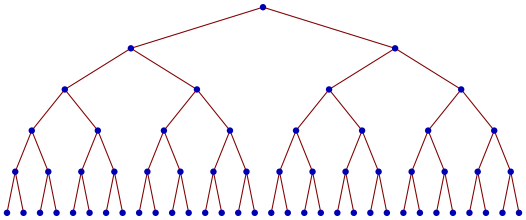

FIGURE 5. The rooted binary tree shown up to four levels. At the top is the root.

Jacobian determinant equal to $\det(A)^n$ (why?). Therefore, for a Borel set $\mathcal{S} \subseteq GL(n, \mathbb{R})$, we have

$$\begin{aligned}\mu(A\mathcal{S}) &= \int_{A\mathcal{S}} |\det X|^{-n} dX = \int_{\mathcal{S}} |\det(AX)|^{-n} d(AX) \\ &= \int_{\mathcal{S}} |\det(AX)|^{-n} \, |\det(A)|^n \, dX = \int_{\mathcal{S}} |\det(X)|^{-n} dX = \mu(\mathcal{S}).\end{aligned}$$

Thus μ is invariant under left multiplication. Check that it is also invariant under right multiplication.

► Let $G = (\mathbb{Z}/(2))^J$ where J is an arbitrary index set. Let $\mu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$ be the unique invariant measure on $\mathbb{Z}/(2)$. A basic theorem in probability theory (Kolmogorov's existence theorem) assures us that there is a unique "product measure" $\mu^{\otimes J}$ on G such that its projection to any finite number of co-ordinates $j_1, \ldots, j_n$ is precisely the n-fold product $\mu \otimes \ldots \otimes \mu$. This $\mu^{\otimes J}$ is an invariant measure on G (check!).

The general question is whether every topological group has an invariant measure.

Definition 14. Let G be a topological group. A nonzero, regular Borel measure μ is said to be a *left Haar measure* on G if $\mu(gA) = \mu(A)$ for all $g \in G$, $A \in \mathcal{B}(G)$. Similarly, a *right Haar measure* is one that satisfies $\mu(Ag) = \mu(A)$. If a measure is both left and right invariant, we call it a *Haar measure*.

Recall that regularity means that for any Borel set A,

$$\mu(A) = \inf\{\mu(U) : U \supseteq A, U \text{ open}\} = \sup\{\mu(K) : A \supseteq K, K \text{ compact}\}.$$

Some situations are problematic.

Example 15. Consider $\mathbb{Q}$ under addition. If μ is invariant, and the singleton $\{0\}$ has mass p, then every singleton must have mass p. If $p = 0$, the measure is identically zero, so we must take $p > 0$. Then μ is basically counting measure on $\mathbb{Q}$, and the only sets with finite measure are finite sets. It does not appear to be of much use, for instance, all nonempty open sets have infinite measure. Alternately, observe that the same measure would be an invariant measure for $\mathbb{Q}$ with discrete topology. The fact that we are taking a more interesting topology that respects addition on $\mathbb{Q}$ is not getting us a more interesting measure.

Example 16. Take any infinite dimensional normed space X, eg., ℓ^2 or $C[0,1]$. Addition is the group operation. If μ is a translation-invariant measure on X, that would be like Lebesgue measure in infinite dimensions - something looks suspicious! Here is one issue. Any such μ must give infinite measure to all open sets. To see this, observe that the unit ball contains countably many balls of identical radius $r > 0$ (intersection of the unit ball with each orthant is open). Since each of these smaller balls must have equal measure, either the unit ball has zero measure or infinite measure.

Example 17. Consider affine transformations on the real line, $f_{a,b}(x) = ax + b$, where $a > 0$ and $b \in \mathbb{R}$. These form a group under composition with the multiplication: $f_{a,b} \circ f_{c,d} = f_{ac,ad+b}$.

In searching for an invariant measure, we try $d\mu(a,b) = h(a,b)da\, db$. Push forward under left multiplication by $f_{A,B}$ to get $A^{-2}h(a/A,(b-B)/a)da\, db$. Invariance requires $h(a,b) = A^{-2}h(a/A,(b-B)/a)$ for almost all a, b and any A, B, which implies that $h(a,b) = a^{-2}$ (up to a constant).

Similarly, if we consider right multiplication by $f_{A,B}$, then the measure $h(a,b)da\ db$ pushes forward to $A^{-1}h(a/A, b - aB)db\, db$. Deduce that right invariance forces $h(a,b) = 1/a$ (again, up to constant factor).

This example shows that the right Haar measure and left Haar measure can both exist and be distinct.

Exercise 18. Consider the group of affine transformation $f_{A,b} : \mathbb{R}^n \mapsto \mathbb{R}^n$ defined by $f_{A,b}(x) = Ax + b$. Here $A \in GL(n, \mathbb{R})$ and $b \in \mathbb{R}^n$. Show that they form a group under composition and find the left and right Haar measures.

Now we are ready to state the results on existence and uniqueness of Haar measures.

Theorem 19 (André Weil)**.** *If G is a locally compact topological group, then it has a unique (up to multiplication by constants) left Haar measure. Similarly for right Haar measure.*

We shall not prove this. But we shall prove the theorem for compact groups.

Theorem 20 (Haar, von Neumann)**.** *If G is a compact topological group, then it has a unique (up to multiplication by constants) Haar measure.*

One can phrase invariance in terms of integrals instead of functions.

Exercise 21. Let G be a locally compact topological group. Let μ be a regular Borel measure on G. Show that the following are equivalent.

(1) μ is a left-Haar measure on G.

(2) For any $f \in C_c(G)$, we have $\int f(x)d\mu(x) = \int f(gx)d\mu(x)$ for all $g \in G$.

3. Proof of existence of Haar measure on compact groups

Let G be a compact group. We want to show the existence of a unique probability measure μ on G such that for any $f \in C(G)$ and any $y \in G$,

$$\int f(yx)d\mu(x) = \int f(xy)d\mu(x) = \int f(x)d\mu(x).$$

This μ is then the unique Haar measure on G.

The key idea: Distribute n points as spread out regularly as possible on G. Then the probability measure that puts mass $1/n$ at each of these points converges to a measure on G that is the Haar measure. For example, if $G = S^1$, it is clear that the most regular distribution of points is to take the nth roots of 1 (or rotate them all by one element of S^1).

There is a starting issue with this plan - what is the meaning of a well-distributed set of points? For simplicity of presentation of the key ideas, we first make the following assumption and remove it later.

Assumption: The topology of G is induced by an invariant metric d, i.e., $d(zx, zy) = d(x, y)$ for all $x, y \in G$.

Once we have a metric, we can talk about ε-nets. Recall that an ε-net is a set $A \subseteq G$ such that every point of G is at distance less than ε of some point in A. Since G is compact, finite ε-nets exists for every $\varepsilon > 0$. Let N_ε be the smallest cardinality of any ε-net. The following lemma has the key idea which makes the proof work.

Lemma 22. *If $A = \{x_1, \dots, x_{N_\varepsilon}\}$ and $B = \{y_1, \dots, y_{N_\varepsilon}\}$ are two ε-nets of minimal cardinality for G, then there is a permutation π such that $d(x_i, y_{\pi(i)}) < 2\varepsilon$ for every $i \leq N_\varepsilon$.*

Proof. Define a bipartite graph with parts A and B (even if a point is common to A and B, it corresponds to two vertices in this graph) with edges $x_i \sim y_j$ if $d(x_i, y_j) < 2\varepsilon$.

Suppose $A' \subseteq A$ and let $N(A') \subseteq B$ be its neighbourhood in the graph. Let $C = (A \setminus A') \cup N(A')$. We claim that C is an ε-net. To show this, take any $z \in G$, and find i, j such that $d(x_i, z) < \varepsilon$ and $d(y_j, z) < \varepsilon$. Then $d(x_i, y_j) < 2\varepsilon$, hence $x_i \sim y_j$. Therefore, either (1) $x_i \in A \setminus A'$ in which case $x_i \in C$ or (2) $x_i \in A'$ in which case $y_j \in N(A') \subseteq C$. Thus every point of G is within ε of a point of C, showing that C is an ε-net. Therefore $N_\varepsilon \leq |C| = N_\varepsilon - |A'| + |N(A')|$. In other words, $|N(A')| \geq |A'|$.

Thus, Hall's conditions are satisfied, and we get a matching of the bipartite graph. That is precisely the permutation π. ■

For a finite set $A = \{x_1, \dots, x_n\}$, let $L_A : C(G) \mapsto \mathbb{R}$ be defined by

$$L_A f = \frac{1}{n} \sum_{k=1}^{N} f(x_k) = \int f d\mu_A$$

where $\mu_A = \frac{1}{n}\sum_{k=1}^{n}\delta_{x_k}$. For any $f \in C(G)$, we define its modulus of continuity $\omega_f(\varepsilon) = sup\{|f(x) - f(y)| : d(x,y) \leq \varepsilon\}$. Then $\omega_f(\varepsilon) \to 0$ as $\varepsilon \to 0$. The above lemma easily implies that if A and B are two minimal cardinality ε-nets, then $|L_A f - L_B f| \leq \omega_f(2\varepsilon)$. We now extend this comparision to nets for different ε.

Lemma 23. *Let A (and B) be minimal cardinality ε-net (respectively δ-net) for G. Then for any $f \in C(G)$, we have $|L_A f - L_B f| \leq \omega_f(2\varepsilon) + \omega_f(2\delta)$.*

Proof. Let $A = \{x_1, \ldots, x_n\}$ and $B = \{y_1, \ldots, y_m\}$. Let $C = A.B = \{x_i y_j : i \leq n, j \leq n\}$. We can write $C = \bigcup_{i \leq m} x_i B = \bigcup_{j \leq m} A y_j$. Thus,

$$L_C f = \frac{1}{n}\sum_{i=1}^{n} L_{x_i B} f = \frac{1}{n}\sum_{j=1}^{m} L_{A y_j} f.$$

But Ay_j is a minimal cardinality ε net for each $j \leq m$, hence the numbers $L_{Ay_j} f$ are all within $\omega_f(2\varepsilon)$ of $L_A f$. Therefore $L_C f$ (being an average of $L_{Ay_j} f$, $j \leq m$, is also within $\omega_f(2\varepsilon)$ of $L_A f$. By an analogous argument, $|L_C f - L_B f| \leq \omega_f(2\delta)$. Putting these together, we see that $|L_A f - L_B f| \leq \omega_f(2\delta) + \omega_f(2\varepsilon)$. ■

Lemma 24. *For each $\varepsilon > 0$, fix a minimal cardinality ε-net A_ε. Then, $\lim_{\varepsilon \to 0} L_{A_\varepsilon} f$ exists for every $f \in C(G)$. The limit does not depend on the choice of the nets A_ε.*

Proof. For $f \in C(G)$. For $\varepsilon > 0$ let K_ε be the collection of all numbers $L_A f$, where A varies over all minimal-cardinality δ-nets for any $\delta < \varepsilon$. Clearly $K_\varepsilon \subseteq [-\|f\|_{\sup}, \|f\|_{\sup}]$. Further, dia$(K_\varepsilon) \leq 2\omega_f(2\varepsilon)$. Therefore, it follows that $\cap \bar{K}_\varepsilon$ is a singleton $\{c\}$, and that number is the limit of $L_{A_\varepsilon} f$ along any sequence of minimal-cardinality ε-nets (as $\varepsilon \to 0$). ■

Proof of Theorem 20 under Assumption 3. For each $f \in C(G)$, let Lf be the number given by the previous lemma, i.e., $Lf = \lim_{\varepsilon \to 0} L_{A_\varepsilon} f$ along any sequence of minimal cardinality ε-nets A_ε. Linearity and positivity of L is obvious. Also $L(\mathbf{1}) = 1$.

For any $g \in G$, let $\tau_g f(x) = f(gx)$. Then,

$$L(\tau_g f) = \lim_{\varepsilon \to 0} L_{A_\varepsilon}(\tau_g f) = \lim_{\varepsilon \to 0} L_{gA_\varepsilon} f = Lf$$

where the last equality follows from the fact that gA_ε is also a minimal cardinality ε-net. Similarly, $L(\sigma_g f) = Lf$ where $\sigma_g f(x) = f(xg)$.

By Riesz's representation theorem, $Lf = \int_G f d\mu$ for a probability measure μ. Invariance of L implies that this measure satisfies the second condition in Exercise 21. Hence, it is a bi-invariant probability measure on G. ■

Removing the assumption 3: If we don't assume that an invariant metric exists, then we cannot talk of ε-nets, but we shall simply consider the net of neighbourhoods of the identity[14].

Given a neighbourhood V of identity, xVy, $x, y \in G$, is an open cover for G. Hence it has a finite sub cover. A *blocking set* for V is a set of minimal cardinality that intersects every one of the sets xAy for $x, y \in G$. Write $a \sim b$ (w.r.t. V) if there is some x, y such that xVy contains both a and b. A *blocking set* is a set that intersects each of the sets xVy for $x, y \in G$. Minimum cardinality blocking sets will replace minimal cardinality ε-nets in our proof. We prove the analogous lemmas.

Note added later: An important missing point in this discussion was pointed out by Abu Sufian. We must show that finite blocking sets exist. When you consider a metric space and ε balls, in finding a blocking set we would need to consider $\varepsilon/2$ balls. The analogue of this without the metric is the following.

Fact: Let V be a neighbourhood of the identity in a topological group G. Then, there exists a neighbourhood W of the identity such that $W.W.W := \{xyz : x, y, z \in W\}$ is contained in V.

It is easy to see this from the fact that the map $(x, y, z) \mapsto xyz$ from $G \times G \times G$ to G is continuous, hence the pull back of V is an open set containing (e, e, e), where e is the identity of the group.

We leave it as an exercise to work out the existence of a finite blocking set using this observation.

For $f \in C(G)$, we define $\omega_f(V) = \sup\{|f(x) - f(y)| : x \sim y\}$.

Lemma 25. *If A and B are blocking sets of minimal cardinality (w.r.t V), then $|L_A f - L_B f| \leq \omega_f(V)$.*

Proof. We shall apply Hall's marriage theorem to say that there is a bijection π between $A = \{a_1, \ldots, a_n\}$ and $B = \{b_1, \ldots, b_n\}$ such that $a_i \sim b_{\pi(i)}$ (w.r.t. V) for all $i \leq n$. Once that is done,

$$|L_A f - L_B f| \leq \frac{1}{n} \sum_{k=1}^{n} |f(a_k) - f(b_{\pi(k)})| \leq \omega_f(V).$$

To check Hall's condition, let $A' \subseteq A$ and $N(A') = \{b \in B : b \sim a \text{ for some } a \in A'\}$. Set $C = (A \setminus A') \cup N(A')$. Show that C is a blocking set. But its cardinality is $|A| - |A'| + |N(A')|$ which shows that $|N(A')| \geq |A'|$. ■

Lemma 26. *Let V, W be two neighbourhoods of identity in G. Let A and B be minimal cardinality blocking sets w.r.t. V and W, respectively. Then $|L_A f - L_B f| \leq \omega_f(V) + \omega_f(W)$.*

Proof. Let $A = \{a_1, \ldots, a_n\}$ and $B = \{b_1, \ldots, b_m\}$. Compare $L_A f$ and $L_B f$ with

$$\frac{1}{mn} \sum_{i \leq n} \sum_{j \leq m} f(a_i b_j). \tag{1}$$

This can be written alternately as $\frac{1}{m} \sum_{j=1}^{m} L_{Ab_j}$ or as $\frac{1}{n} \sum_{i=1}^{n} L_{a_i B}$. Since each Ab_j is a minimal cardinality blocking set, $|L_{Ab_j} f - L_A f| \leq \omega_f(V)$ for all $j \leq m$. Similarly, $|L_{a_i B} f - L_B f| \leq \omega_f(W)$.

[14] A *net* is a partially ordered set in which given any two elements, there is a common element greater than or equal to both. For example, the collection of all neighbourhoods of a point x_0 in a topological space, endowed with the reverse inclusion, is a net. Given U, V, we have $U \cap V$ lying above U and above V.

This shows that the quantity in (1) is with $\omega_f(V)$ of L_Af and within $\omega_f(W)$ of L_Bf. Therefore, $|L_Af - L_Bf| \le \omega_f(V) + \omega_f(W)$. ■

Exercise 27. Let $f \in C(G)$. Given $\varepsilon > 0$, show that there exists a neighbourhood V of identity such that for all neighbourhoods $e \in W \subset V$ and all minimal blocking sets A, we have $\omega_f(W) \le \varepsilon$.

We put all these to prove the existence of Haar measure.

Proof of Theorem 20. Fix $f \in C(G)$ and for V, a neighbourhood of identity, define

$$K_V = \{L_Af : A \text{ is a minimal cardinality blocking set w.r.t. } W \text{ for some } W \subseteq V,\ W \ni e\}.$$

Then, all elements of K_V are within $2\omega_f(V)$ of each other, hence $\text{dia}(K_V) \le 2\omega_f(V)$ which goes to zero by the exercise above. Further, $K_V \supseteq K_W$ if $V \supseteq W$. Hence, the sets $\bar{K_V}$ have finite intersection property since $K_{V_1} \cap \ldots \cap K_{V_m} \supseteq K_W$ where $W = V_1 \cap \ldots \cap V_m$. From this, it follows that $\bigcap_V \bar{K_V}$ is a singleton that we denote as $\{Lf\}$. Another way to say this is that if V_i is a net of neighbourhoods that converge to $\{e\}$, then $\lim L_{A_i}f$ exists and is independent of the choice of the minimal cardinality blocking sets A_i chosen.

The mapping $L : C(G) \mapsto \mathbb{R}$ is linear, positive and $L(\mathbf{1}) = 1$. Hence $Lf = \int f d\mu$ for some probabilit measure μ (Riesz's representation theorem). Fix any $x_0 \in G$ and consider $g(x) = f(x_0 x)$. Then, for any blocking set A, it is clear that $L_Ag = L_{x_0A}f$. Since A is a minimal cardinality blocking set w.r.t. V if and only is x_0A is, it follows that $Lf = Lg$. In other words, L is left-invariant. Similarly it is also right invariant. That is, for any $f \in C_c(G)$, we have $\int f(x)d\mu(x) = \int f(gx)d\mu(x)$ for all $g \in G$. Hence μ is a Haar measure. ■

The uniqueness question: We have constructed a bi-invariant probability measure μ. Suppose ν is another left-invariant probability measure on G. Define the measure $\theta = \mu \star \nu$ by (the right hand side is a positive linear functional of f, hence represented by a measure)

$$\int f(x)d\theta(x) = \iint f(xy)d\mu(x)d\nu(y)$$

for $f \in C(G)$. By the right-invariance of μ, the inner integral is $\int f d\mu$ for every $y \in G$ (a constant independent of y). Integrating w.r.t ν gives us that $\int f d\theta = \int f d\mu$.

Apply Fubini's theorem (applicable since the integrand is bounded) to write

$$\int f(x)d\theta(x) = \iint f(xy)d\nu(y)d\mu(x) = \iint f(y)d\nu(y)d\mu(x)$$

by the left-invariance of ν. The inner integral is independent of x and we simply get $\int f d\theta = \int f d\nu$.

Thus, $\int f d\mu = \int f d\nu$ for all $f \in C(G)$, whence it follows that $\mu = \nu$.

Remark 28. What was all this? If you are comfortable with probability language, let X and Y be independent random variables with distribution μ and ν, respectively. Bi-invariance of μ means that gX, Xg and X all have the same distribution μ, for any fixed $g \in G$. Left-invariance of ν means that gY has the same distribution as Y, for any $g \in G$.

Now consider $Z = XY$. Using independence, we can argue that Z has the same distribution as X (condition on Y) and that Z has the same distribution as Y (condition on X). Hence X and Y have the same distribution, i.e., $\mu = \nu$.

4. MATCHING THEOREM VIA CONVEX DUALITY AND FLOWS

In this section we revisit the matching theorem and give a new proof from a more sophisticated view point. In the process, we introduce some deep ideas of importance in themselves.

The matching theorem will be derived from the max-flow, min-cut theorem. To state this theorem, we introduce the notion of a *flow* on a directed graph. Let $G = (V, E)$ be a finite directed graph. This means that V is a finite set and E is a subset of $V \times V \setminus \{(u, u) : u \in V\}$. An element $(u, v) \in E$ will be interpreted as an edge going from u to v. Note that both (u, v) and (v, u) may be present (thus an undirected graph can be made into a directed graph by replacing an undirected edge by directed edges in both directions). For every vertex $u \in V$, its incoming edges are those of the form (x, u) and outgoing edges are those of the form (u, x).

Fix two vertices $s, t \in V$ (called source and sink, respectively). A *flow* on G from s to t is a function $\theta : E \mapsto \mathbb{R}_+$ such that (a) for any $u \in V \setminus \{s, t\}$ we have $\sum_x \theta(x, u) = \sum_x \theta(u, x)$ (the left sum is over incoming edges and the right sum is over outgoing edges), (b) $\theta(x, s) = 0$ if $(x, s) \in E$, (c) $\theta(t, x) = 0$ if $(t, x) \in E$. The strength of a flow is defined as $\|f\| := \sum f(s, x)$. Since

$$\sum_{u \in V} \left[\sum_x f(u, x) - \sum_x f(x, u) \right] = \sum f(s, x) - \sum f(x, t)$$

as all other contributions cancel, it follows that the strength of f is also equal to $\sum_x f(x, t)$.

A *cut-set* (for s, t) is a subset Π of V such that every directed path from s to t passes through some vertex of Π (i.e., if $s = u_0, u_1, \ldots, u_{m-1}, u_m = t$ are vertices such that $(u_i, u_{i+1}) \in E$ for all i, then there is some k such that $u_k \in \Pi$). The capacity of a cut-set is defined as

Theorem 29 (Ford-Fulkerson max-flow min-cut theorem). *In the setting above, the maximum strength over all flows is equal to the minimum capacity of a cutset.*

Hall's marriage theorem can be derived from the max-flow min-cut theorem, somewhat similarly to the derivation from Dilworth's theorem.

Proof of the matching theorem from the max-flow min-cut theorem. From the given bipartite graph, create a directed graph with vertex set $\{s\} \sqcup V_1 \sqcup V_2 \sqcup \{t\}$ (here s and t are two new vertices). The edges of the directed graph are of three kinds: (a) (s, x) for all $x \in V_1$, (b) (x, y) for all $x \in V_1$, $y \in V_2$ with $x \sim y$ in the given bipartite graph, (c) (y, t) for all $y \in V_2$. We give capacity 1 to all edges.

What is the maximum flow from s to t? For this, we claim that $\Pi_0 = \{(s, x) : x \in V_1\}$ is a minimal cut-set. Indeed, if Π is any cut-set ■

4.1. **Convex functions, Legendre transformation and a minimax theorem.** Let $f : X \times Y \mapsto \mathbb{R}$ be a function. Then it is true in general that

$$\sup_{x\in X} \inf_{y\in Y} f(x,y) \le \inf_{y\in Y} \sup_{x\in X} f(x,y).$$

Indeed, fix any $x \in X$ and $y \in Y$, and observe that $\inf\limits_{y'\in Y} f(x,y') \le f(x,y) \le \sup\limits_{x'\in X} f(x',y)$. Therefore the supremum (over $x \in X$) of the left hand side is bounded above by the infimum (over $y \in Y$) of the right hand side, which is what the above inequality says. Results that provided conditions (on the spaces and the function) under which the above inequality is actually an inequality are called *minimax theorems*. We shall prove one such theorem. Convexity plays a key role, hence we recall some aspects of it.

If E is a topological vector space and $f : E \mapsto \mathbb{R} \cup \{+\infty\}$, then f is said to be a convex function if $f(tx + (1-t)y) \le tf(x) + (1-t)f(y)$ for any $x, y \in E$ and any $0 \le t \le 1$. We shall assume that $f(x) < +\infty$ for at least one $x \in E$. The *Legendre transform* (or convex dual or convex conjugate or Legendre-Fenchel dual...) is the function $f^* : E^* \mapsto \mathbb{R} \cup \{+\infty\}$ defined as

$$f^*(L) = \sup_{x\in E} L(x) - f(x) \qquad \text{for } L \in E^*.$$

If $x_0 \in E$ is a point such that $f(x_0) < +\infty$, then $f^*(L) \ge L(x_0) - f(x_0)$, hence f^* takes values in $\mathbb{R} \cup \{+\infty\}$.

Example 30. Let $E = \mathbb{R}^n$ and let $f(x) = \frac{1}{p}\sum_i |x_i|^p$. For $p \ge 1$, this is a convex function. Then $E^* = \mathbb{R}^n$ and $f^*(\lambda) = \sup\{\langle \lambda, x\rangle - \frac{1}{p}\|x\|^p : x \in \mathbb{R}^n\}$. By a simple calculation, this is found to be $\frac{1}{q}\sum_i |\lambda_i|^q$ where q is defined by $\frac{1}{p} + \frac{1}{q} = 1$. This is the source of all the intimate connection between the p-norm and the q-norm.

Exercise 31. Find the convex dual of (1) $f(x) = \max_i |x_i|$ on $\mathbb{R}^n$, (2) $f(x) = |x|$ on $\mathbb{R}$, (3) $f(x) = e^x$ on $\mathbb{R}$.

Lemma 32. *f^* is lower semi-continuous in the weak-* topology on E^*.*

Proof. Suppose $f^*(L) > t$ for some $t \in \mathbb{R}$. Then there is some $x \in E$ such that $L(x) - f(x) > t$. Hence if L' is close to L (recall that the weak-* topology on X^* is the smallest topology in which $L \mapsto L(x)$ is continuous for each x), then $L'(x) - f(x) > t$ for the same x, and therefore $f^*(L) > t$. Thus, $\{f^* > t\}$ is open, implying that f^* is lower semi-continuous. ■

One of the key properties of the convex dual is that it is a dual - in other words $f^*{*} = f$. But to make sense of this, we must recall that f^{**} is a function on E^{**} and E is naturally embedded inside E^{**}. The question is whether the restriction of f^{**} to E is equal to f. In finite dimensional spaces this is always true, but in infinite dimensional spaces some condition on the continuity of f is needed.

Theorem 33 (Fenchel-Moreau). *Let E be a locally convex and let $f : E \mapsto \mathbb{R} \cup \{+\infty\}$ be a convex function that is finite somewhere. Assume that f is lower semi-continuous. Then, $f^{**}\big|_E = f$.*

Proof. First note that $f^{**}(x) = \sup\{L(x) - f^*(L) : L \in E^*\}$ for $x \in E$. Since $L(x) - f^*(L) \leq f(x)$ for all $x \in E$ and all $L \in E^*$ (by the definition of f^*), it follows that $f^{**}(x) \leq f(x)$. ■

CHAPTER 5

Asymptotics of integrals

1. Some questions

Consider the sequence $n!$, which, by definition, it is the product of the first n positive integers. Do we understand how large it is? For example, it is easy to see that $2^{n-1} \le n! \le n^{n-1}$. Both sides of this inequality are quantities we are more familiar with and can work with easily. However, they are quite far from each other. We can sharpen the bounds as follows. Write $\log n! = \sum_{k=1}^{n} \log k$ and hence $\int_{k-1}^{k} \log x dx \le \log k \le \int_{k}^{k+1} \log x dx$. Therefore,

$$\int_0^n \log x\, dx \le \log n! \le \int_1^{n+1} \log x\, dx$$

giving $n\log n - n \le \log n! \le n\log(n+1) - n + \log(n+1)$. Thus,

$$n^n e^{-n} \le n! \le n^{n+1} e^{-n}(n+1).$$

The ratio of the upper and lower bounds is only of order n^2 now. Can we sharpen it further and get an elementary expression $f(n)$ such that[15] $n! \sim f(n)$? Stirling's formula asserts that $n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}$. We shall prove this later.

Similarly, one is often interested in the magnitudes of various quantities such as

(1) Asymptotics of the Bell numbers B_n, the number of ways to partition the set $\{1, 2, \dots, n\}$.

(2) Asymptotics of $p(n)$, the number of ways to partition the number n.

(3) Asymptotics of $H_n(x)$ (fixed x, large n), where H_n is the nth Hermite polynomial.

One could list many more. We shall see some basic techniques to get the asymptotics of such quantities. We shall restrict ourselves to quantities that can be expressed as integrals of certain kinds. Fortunately, this covers many examples.

(1) $n! = \int_0^\infty x^n e^{-x} dx$.

(2) $B_n = n! \frac{1}{2\pi i} \int_C e^{e^z - 1} z^{-n-1} dz$ where C is a simple closed contour enclosing the origin in the complex plane.

(3) $H_n(x) = (-1)^n n! e^{x^2} \frac{1}{2\pi i} \int_\gamma \frac{1}{(z-x)^{n+1}} e^{-z^2} dz$.

In general, if we have a sequence (a_n), and its generating function $F(z) = \sum_{n=0}^\infty a_n z^n$ or exponential generating function $G(z) = \sum_{n=0}^\infty a_n z^n / n!$ has a positive radius of convergence, we can write

[15] Common notation: (1) $a_n \sim b_n$ means $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$, (2) $a_n \asymp b_n$ means that $cb_n \le a_n \le Cb_n$ for some constants c and C, (3) $a_n \approx b_n$ means $\log a_n \sim \log b_n$. Similar interpretation for $f(x) \sim g(x)$ etc.

a_n as

$$a_n = \frac{1}{2\pi i}\int_\gamma F(z)z^{-n-1}dz, \quad a_n = n!\frac{1}{2\pi i}\int_\gamma G(z)z^{-n-1}dz.$$

We will work out a few examples in the next sections[16]. The method is important, more than than the statements of the results.

2. Laplace's method

Let us return to the factorial function. We shall derive its asymptotics (Stirling's formula) using the integral representation

$$n! = \int_0^\infty e^{-x}x^n dx.$$

Here is a quick sketch of the idea. The integrand is $\exp\{-x + n\log x\}$. The exponent (and hence the integrand) is maximized at $x = n$. Near this point, the second order Taylor expansion of the exponent is (derivative term vanishes because we are at a maximum)

$$-x + n\log x = (-n + n\log n) - \frac{1}{2n}(x-n)^2.$$

If we blindly replace the exponent by this, we get

$$e^{-n+n\log n}\int_0^\infty e^{-\frac{1}{2n}(x-n)^2}dx = n^n e^{-n}\int_{-\sqrt{n}}^\infty e^{-\frac{1}{2}t^2}dt.$$

For large n, the integral can be extended to the whole line without affecting the value significantly, hence we get

$$n^n e^{-n}\int_{-\infty}^\infty e^{-\frac{1}{2}t^2}dt = n^n e^{-n}\sqrt{2\pi n}$$

which is precisely Stirling's approximation! We have not yet justified the steps, or shown in what precise sense this approximates $n!$, but the idea described here is general: The contribution to the integral comes from a certain neighbourhood (here of order $\sqrt{n}$ in length) of the point where the integrand is maximized (here n).

A general theorem: We now try a general integral of the form $I(\lambda) = \int_{\mathbb{R}} e^{-\lambda f(x)}g(x)dx$. Make the following assumptions.

1. Let $f : \mathbb{R} \mapsto \mathbb{R}_+$ be C^2, with a unique minimum at 0. Assume that $f(0) = 0$ (without loss of generality) and that $f''(0) > 0$. For $\delta > 0$, assume that $m_\delta = \inf_{|x|\ge\delta} f(x)$ is strictly positive.
2. Let $g : \mathbb{R} \mapsto \mathbb{R}$ be continuous and assume that $g(0) > 0$ (if $g(0) < 0$, replace g by $-g$).
3. Assume that the integral defining $I(\lambda)$ converges absolutely for all λ.

Theorem 1. *With the above assumptions, we have (as $\lambda \to \infty$)*

$$I(\lambda) \sim \frac{\sqrt{2\pi}g(0)}{\sqrt{f''(0)}\sqrt{\lambda}}.$$

[16] Much of this material is taken from a very well-written old book of de Bruijn titled *Asymptotic methods in analysis*.

In one line, the idea is that most of the contribution to the integral comes from a neighbourhood of 0, and the contribution there is go by a second order Taylor expansion of f (and continuity of g) which leads to a standard Gaussian integral whose value is given in the statement of the theorem.

Proof. Let $\delta > 0$ (we may later allow it to depend on λ) and write $I(\lambda) = I_1(\lambda) + I_2(\lambda) + I_3(\lambda)$ where

$$I_1(\lambda) = \int_{-\delta}^{\delta} e^{-\lambda f(x)} g(x)dx, \quad I_2(\lambda) = \int_{\delta}^{\infty} e^{-\lambda f(x)} g(x)dx, \quad I_3(\lambda) = \int_{-\infty}^{-\delta} e^{-\lambda f(x)} g(x)dx.$$

The contribution from I_2 and I_3 are small. Indeed, we can write

$$\begin{aligned} |I_2(\lambda)| + |I_3(\lambda)| &\leq \int_{[-\delta,\delta]^c} e^{-\lambda f(x)} |g(x)| dx \\ &\leq e^{-\lambda m_\delta} \int_{\mathbb{R}} |g(x)| dx. \end{aligned}$$

Now we turn to $I_1(\lambda)$. We have $\varepsilon(\delta)$ that goes to zero as δ goes to zero, such that (from our assumptions $f(0) = f'(0) = 0$) for all $x \in [-\delta, \delta]$,

$$\frac{1}{2} f''(0)(1-\varepsilon)x^2 \leq f(x) \leq \frac{1}{2} f''(0)(1+\varepsilon)x^2, \quad g(0)(1-\varepsilon) \leq g(x) \leq g(0)(1+\varepsilon)$$

where we have written ε for $\varepsilon(\delta)$ so as to not add to the ugliness in this world. We can write the errors multiplicatively as $g(0)(1 \pm \varepsilon)$ and $f''(0)(1 \pm \varepsilon)$ because of the assumption that $g(0) > 0$ and $f''(0) > 0$. Thus,

$$(1) \qquad g(0)(1-\varepsilon) e^{-\frac{1}{2}\lambda f''(0)(1+\varepsilon)x^2} \leq g(x) e^{-\lambda f(x)} \leq g(0)(1+\varepsilon) e^{-\frac{1}{2}\lambda f''(0)(1-\varepsilon)x^2}, \quad \text{for } |x| \leq \delta.$$

From basic facts about the Gaussian integral[17], for any $\tau > 0$ we know that

$$\frac{\sqrt{2\pi}}{\sqrt{\tau}} \left(1 - \frac{2}{\sqrt{2\pi}\delta} e^{-\frac{1}{2}\tau\delta^2}\right) \leq \int_{[-\delta,\delta]} e^{-\frac{1}{2}\tau x^2} dx \leq \frac{\sqrt{2\pi}}{\sqrt{\tau}}$$

Integrating all sides of (1) and using these inequalities gives

$$\frac{\sqrt{2\pi} g(0)(1-\varepsilon)}{\sqrt{\lambda f''(0)(1+\varepsilon)}} \left(1 - \frac{2}{\sqrt{2\pi}\delta} e^{-\frac{1}{2}\lambda f''(0)(1+\varepsilon)\delta^2}\right) \leq I_1(\lambda) \leq \frac{\sqrt{2\pi} g(0)(1+\varepsilon)}{\sqrt{\lambda f''(0)(1-\varepsilon)}}$$

If $\delta \to 0$ as $\lambda \to 0$, then also $\varepsilon \to 0$ and we can simply write

$$I_1(\lambda) \sim \frac{\sqrt{2\pi} g(0)}{\sqrt{\lambda f''(0)}}.$$

Observe that this behaves as $1/\sqrt{\lambda}$ while our bound for $|I_2(\lambda)| + |I_3(\lambda)|$ was $e^{-\lambda m_\delta}$. Hence the latter is negligible compared to $I_1(\lambda)$ and we arrive at $I(\lambda) \sim \frac{\sqrt{2\pi} g(0)}{\sqrt{\lambda f''(0)}}$ as claimed. ∎

[17] These inequalities follow by recalling that $\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} dx = \sqrt{2\pi}$ and for any $a > 0$, we have the estimate $\int_a^\infty e^{-\frac{1}{2}x^2} dx \leq \int_a^\infty \frac{x}{a} e^{-\frac{1}{2}x^2} dx = \frac{1}{a} e^{-\frac{1}{2}a^2}$. Rescaling the variable by $\sqrt{\tau}$ gives the estimates in the form we need.

Remark 2. To simplify notation we assumed that the minimum of f occurs at 0 and the value taken is zero. If the minimum occurs at a unique point x_0, then by applying the theorem with $f(x-x_0)-f(x_0)$, we get the asymptotic formula

$$I(\lambda)\sim\frac{\sqrt{2\pi}g(x_0)e^{-\lambda f(x_0)}}{\sqrt{\lambda f''(x_0)}}.$$

Remark 3. In most cases (eg., if we assume additional smoothness on f and g), it is reasonable to expect that $\varepsilon(\delta)\asymp\delta$ and that $m_\delta\asymp\delta^2$ as $\delta\downarrow 0$. Going through the proof, the ratio between $I(\lambda)$ and the asymptotic form give is of the form $1\pm\delta\pm e^{-\lambda\delta^2/2}$. Taking $\delta=C\sqrt{\log\lambda}/\sqrt{\lambda}$ for a suitably large constant C optimizes the error and we get $1\pm\frac{\sqrt{\log\lambda}}{\sqrt{\lambda}}$. Observe that it is important to take $\delta>>\frac{1}{\sqrt{\lambda}}$ to get the full Gaussian integral in $I_1(\lambda)$. Later we show that by expanding f and g to higher orders, one can improve the asymptotics so well that the error is reduced to $O(\lambda^{-p})$ for any p.

Example 4. Now we are ready to make precise the derivation of Stirlings' approximation. We start with

$$\begin{aligned}n!&=\int_0^\infty x^ne^{-x}dx=\int_0^\infty(nx)^ne^{-nx}ndx\\&=n^{n+1}\int_0^\infty e^{-n[x-\log x]}dx.\end{aligned}$$

This is in the form of the integral considered in the theorem, with $\lambda=n$, $g(x)=1$ and $f(x)=x-\log x$ (the interval of integration is $(0,\infty)$ instead of $\mathbb{R}$, but that does not make a difference as long as the global minima of f are in the interior. Alternately, set $g=0$ on $(-\infty,0)$). Since $f'(x)=1-\frac{1}{x}$, we see that f attains its unique minimum at $x_0=1$ and that $f''(1)=1$. Therefore, by the theorem (or the remark following it)

$$n!\sim n^{n+1}\frac{\sqrt{2\pi}g(1)e^{-nf(1)}}{\sqrt{n}\sqrt{f''(1)}}=n^{n+\frac{1}{2}}e^{-n}\sqrt{2\pi}.$$

This is Stirlings' formula.

Exercise 5. Show that $\int_0^\pi x^n\sin x dx\sim\frac{\pi^{n+2}}{n^2}$.

When the maximum of the integral occurs at and end of the interval of integration, the same methods can be followed to get a different answer.

Exercise 6. Let $I(\lambda)=\int_0^\infty e^{-\lambda f(x)}dx$ where $f:\mathbb{R}_+\mapsto\mathbb{R}_+$ has a unique minimum at zero. Make appropriate assumptions and show that $I(\lambda)\sim\frac{-1}{f'(0)\lambda}e^{\lambda f(0)}$.

3. On asymptotic expansions

We have shown that $\frac{n!}{n^{n+\frac{1}{2}}e^{-n}\sqrt{2\pi}}=1+o(1)$. Now we want to find the error term and improve the asymptotics. Or more generally, how to get better approximants to $I(\lambda)=\int_{\mathbb{R}}g(x)e^{-\lambda f(x)}dx$?

Let us work in the language of the theorem in the previous section, but with some extra assumptions. Let $g(x) = 1$ for simplicity, and assume that f has additional smoothness as required. We still assume that f has a minimum at 0 and $f(0) = 0$.

As before we write $I(\lambda)$ as $I_1(\lambda)+I_2(\lambda)+I_3(\lambda)$ splitting the integral over $[-\delta,\delta]$, over (δ,∞) and over $(-\infty,-\delta)$. Recall that $|I_2(\lambda)| + |I_3(\lambda)| = O(e^{-m_\delta \lambda})$ while $I_1(\lambda) \asymp \frac{1}{\sqrt{\lambda}}$. Consequently, as long as $m_\delta \lambda \geq C \log \lambda$, the contribution from I_2 and I_3 can be safely ignored. Further, since $m_\delta \asymp \delta^2$, this just means that we should make sure that $\delta^2 \lambda 1$. Thus we concentrate only on I_1. By Taylor expansion, we see that with $\varepsilon = \varepsilon(\delta)$ that goes to 0 as $\delta \to 0$, we have for all $x \in [-\delta,\delta]$,

$$f(x) \overset{\varepsilon}{=} \frac{1}{2}x^2 f_2 + \frac{1}{3!}x^3 f_3 + \frac{1}{4!}x^4 f_4$$

where $f_k = f^{(k)}(0)$. Here $h(x) \overset{\varepsilon}{=} k(x)$ means that $h(x)/k(x)$ is between $1-\varepsilon$ and $1+\varepsilon$ (for $x \in [-\delta,\delta]$). Consequently, $I_1(\lambda)$ is bounded between $J(\lambda(1 \pm \varepsilon))$ where

$$\begin{aligned} J(\mu) &= \int_{-\delta}^{\delta} e^{-\mu[\frac{1}{2}f_2 x^2 f_2 + \frac{1}{3!}x^3 f_3 + \frac{1}{4!}x^4 f_4]} dx \\ &= \int_{-\delta\sqrt{\mu}}^{\delta\sqrt{\mu}} e^{-\frac{1}{2}x^2} e^{-\frac{f_3}{3!}\frac{x^3}{\sqrt{\mu}} - \frac{f_4}{4!}\frac{x^4}{\mu}} dx. \end{aligned}$$

Write the series expansion

$$e^{-\frac{1}{3!}\frac{x^3}{\sqrt{\mu}} - \frac{1}{4!}\frac{x^4}{\mu}} = 1 - \frac{f_3}{3!}\frac{x^3}{\sqrt{\mu}} + \frac{1}{\mu}\left[\frac{f_3^2 x^6}{2 \times (3!)^2} - \frac{f_4 x^4}{4!}\right] + \frac{1}{\mu^{\frac{3}{2}}}\left[-\frac{f_3^3 x^9}{3! \times (3!)^3}\right] + [\ldots]$$

We shall integrate these terms now. Three points:

(1) The integral will be extended to the whole line. To estimate the error, one must use bounds of the form

$$\int_a^\infty x^p e^{-\frac{1}{2}x^2} dx \leq C_p a^p e^{-\frac{1}{2}a^2}$$

for some constant C_p and valid for all $a \geq 1$. We leave you to figure out how such a bound can be derived (we showed it earlier for $p = 0$).

(2) Check that $\int_{-\infty}^{\infty} x^p e^{-\frac{1}{2}x^2} dx = \sqrt{2\pi}(p-1) \times (p-3) \ldots \times 3 \times 1$ if p is even and zero if p is odd. We only need the cases $p \leq 6$ for the calculation shown here.

(3) Show rigorously that the terms denoted by $[\ldots]$ contribute only $O(\mu^{-2})$ after integration.

Once these points are made, we have the asymptotic relation

$$J(\mu) \sim \frac{\sqrt{2\pi}}{\sqrt{\mu}\sqrt{f_{''(0)2}}}\left(1 - \frac{1}{\mu}\left[\frac{5f_3^2}{24} - \frac{f_4}{8}\right] + O(\mu^{-2})\right)$$

Writing these for $\mu = \lambda(1 \pm \varepsilon)$ and using that $\varepsilon \to 0$ as $\delta \to 0$, we finally arrive at

$$I(\lambda) \sim I_1(\lambda) \sim \frac{\sqrt{2\pi}}{\sqrt{\lambda}\sqrt{f_2}}\left(1 - \frac{1}{\lambda}\left[\frac{5f_3^2}{24} - \frac{f_4}{8}\right] + O(\lambda^{-2})\right)$$

Remark 1. As before, if the minimum of f is attained at some x_0, we get

$$I(\lambda) \sim \frac{\sqrt{2\pi}e^{-\lambda f(x_0)}}{\sqrt{\lambda}\sqrt{f''(x_0)}}\left(1-\frac{1}{\lambda}\left[\frac{5f^{(3)}(x_0)^2}{24}-\frac{f^{(4)}(x_0)}{8}\right]+O\left(\frac{1}{\lambda^2}\right)\right).$$

Example 2. Let us apply this in the case of $n! = n^{n+1}\int_0^\infty e^{-nf(x)}dx$ with $f(x) = x - \log x$. Then $x_0 = 1$, $f^{''}(0) = 1$, $f^{(3)}(0) = 2$, $f^{(4)}(0) = 6$. Therefore,

$$\begin{aligned} n! &\sim n^{n+1}\frac{\sqrt{2\pi}e^{-n}}{\sqrt{n}\sqrt{1}}\left(1+\frac{1}{n}\left[\frac{5\times 2^2}{24}-\frac{6}{8}\right]+O(\frac{1}{n})\right) \\ &= n^{n+\frac{1}{2}}e^{-n}\sqrt{2\pi}\left(1+\frac{1}{12n}+O\left(\frac{1}{n^2}\right)\right). \end{aligned}$$

It is clear that there is nothing to stop us (except boredom and reluctance) from getting further terms. If one does that, one arrives at the notion of *asymptotic expansion*. For a function F, if we can find number $b_0, b_1, \ldots$ such that

$$F(x) = b_0 + \frac{b_1}{x} + \ldots + \frac{b_n}{x^n} + O(x^{-n-1})$$

as $x \to \infty$, for any n, then we say that F has an asymptotic expansion and write

$$F(x) \overset{\text{asy.}}{=} b_0 + \frac{b_1}{x} + \frac{b_2}{x^2} + \ldots$$

One must be careful to not think of this as equality for fixed x. In fact, for any fixed x, the series on the right (usually) diverges! But truncations of the series give excellent approximations as $x \to \infty$, and the more terms we keep, better the asymptotic approximation.

Exercise 3. Work out the next term in the asymptotic expansion of $I(\lambda) = \int_{\mathbb{R}} e^{-\lambda f(x)}dx$ (under the usual assumptions). Deduce that

$$n! = n^{n+\frac{1}{2}}e^{-n}\sqrt{2\pi}\left(1+\frac{1}{12n}+\frac{1}{288n^2}+O\left(\frac{1}{n^3}\right)\right).$$

Exercise 4. Find the expansion of $I(\lambda) = \int_{\mathbb{R}} g(x)e^{-\lambda f(x)}dx$ (make appropriate assumptions) to improve the approximation

$$I(\lambda) \sim \frac{\sqrt{2\pi}g(x_0)e^{-\lambda f(x_0)}}{\sqrt{\lambda f''(x_0)}}.$$

to the next term.

4. Asymptotics of Bell numbers

Let B_n denote the number of set partitions of the set $[n] = \{1, 2, \ldots, n\}$. For example, $B_3 = 5$, because the possible partitions are

$$\{\{1,2,3\}\}, \quad \{\{1,2\},\{3\}\} \quad \{\{1,3\},\{2\}\} \quad \{\{2,3\},\{1\}\} \quad \{\{1\},\{2\},\{3\}\}.$$

Let us emphasize that we disregard ordering of the blocks, or the ordering of elements within blocks. Equivalently, we may adopt the convention that the blocks are ordered according by the smallest element they contain, and that elements are ordered increasingly within each block.

Question: What are the asymptotics of B_n for large n?

Our approach will be to express B_n as an integral and then find the asymptotics of the integral. Various choices made here are explained later under the general rubric of the *saddle point method* or the *method of steepest descent*.

The first step is that although there is no explicit formula for B_n, there is a weighted sum for which there is. The power series in the lemma below is known as the *exponential generating function* of the sequence $(B_n)_{n\geq 0}$. This is in contrast to the ordinary generating function $\sum_{n=0}^{\infty} B_n z^n$. Which of these two is convenient to work with depends on the sequence.

Lemma 1. *Set $B_0 = 1$. Then $\sum_{n=0}^{\infty} B_n \frac{z^n}{n!} = e^{e^z-1}$ for all $z \in \mathbb{C}$.*

Proof. Observe the recurrence (for $n \geq 1$)

$$B_n = \sum_{k=1}^{n} \binom{n-1}{k-1} B_{n-k}$$

that results from fixing the block containing the element 1, and then partitioning the remaining elements. Therefore,

$$\begin{aligned} \sum_{n=1}^{\infty} B_n \frac{z^{n-1}}{(n-1)!} &= \sum_{n=1}^{\infty} \frac{z^{n-1}}{(n-1)!} \sum_{k=1}^{n} \binom{n-1}{k-1} B_{n-k} \\ &= \sum_{k=1}^{\infty} \frac{z^{k-1}}{(k-1)!} \sum_{n=k}^{\infty} B_{n-k} \frac{z^{n-k}}{(n-k)!}. \end{aligned}$$

Writing $B(z)$ for the exponential generating function, the above equation reads $B'(z) = B(z)e^z$. But e^{e^z-1} satisfies this differential equation and has the value 1 at $z = 0$, just like $B(z)$. Hence $B(z) = e^{e^z-1}$. ■

Exercise 2. In writing the above proof we did not bother about convergence issues. Fill the gaps.

Applying Cauchy's formula for derivatives, we arrive at the following formula

$$B_n = \frac{n!}{2\pi i} \int_{\gamma} e^{e^z-1} z^{-n-1} dz$$

where γ is any contour that winds around the origin once (anti-clockwise). We choose the following contour (the choices will be explained later):

(1) Let $u > 0$ be the unique solution to $ue^u = n + 1$. Observe that $u \sim \log n - \log\log n$ as $n \to \infty$ (to see this, check that $xe^x > n + 1$ if $x = (1 + \delta)(\log n - \log\log n)$ and $xe^x < n + 1$ if $x = (1 - \delta)(\log n - \log\log n)$).

(2) Let $\delta > 0$ be fixed (we may also let it depend on n after seeing various error terms). Then let $\gamma = \gamma_0 + \gamma_1$, where γ_0 is the straight line segment from $u - i\delta$ to $u + i\delta$ and γ_1 is the circular arc centered at the origin and going from $u + i\delta$ to $u - i\delta$ (in the anti-clockwise direction)[18].

We shall show that the integral over γ_1 is negligible and that the integral over γ_0 comes from the behaviour of the integrand at the point u alone. First some preparation.

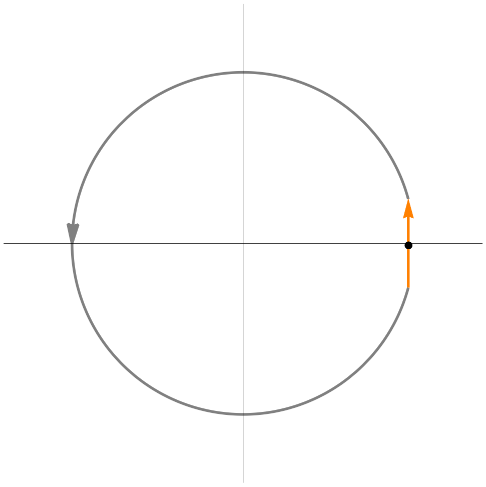

FIGURE 6. The contour $\gamma = \gamma_0 + \gamma_1$ used for estimating B_n. Most of the contribution comes from a neighbourhood of the saddle point marked in black - the point $(u, 0)$ where $ue^u = n + 1$.

Integral over γ_0: Parameterize $\gamma_0(t) = u + it$ for $-\delta \le t \le \delta$ to see that the integral is (we have set aside the $n!/2\pi$ factor)

$$I_0(n) = \int_{-\delta}^{\delta} e^{\psi(t)} dt$$

where $\psi(t) = e^{u+it} - 1 - (n+1)\log u - (n+1)\log(1 + \frac{it}{u})$ (here we choose the branch of logarithm on the right half-plane that is equal to 0 at 1). Since $|t| \le \delta$ and $u \sim \log n - \log\log n$ is going to ∞, the following Taylor expansion is valid uniformly over $[-\delta, \delta]$:

$$\psi(t) = \psi(0) + t\psi'(0) + \frac{1}{2}t^2\psi''(0) + [\ldots]$$

Note that

(1) $\psi(0) = e^u - 1 - (n+1)\log u$,

(2) $\psi'(0) = i[e^u - \frac{n+1}{u}]$ which is zero by the choice of u (this explains the choice!).

(3) $\psi''(0) = -[e^u + \frac{n+1}{u^2}] = -e^u(1 + \frac{1}{u})$ (second equality is by the choice of u).

[18] This part was differently in class. There we followed de Bruijn and transformed the curve to the vertical straight line $t \mapsto u + it$, $-\infty < t < \infty$. But doing it the way written here, it seems that one can take care of the error term more easily. Please check the arguments carefully and critically!

Although the integrand is complex-valued, the first two derivatives are real and we can use Laplace's method and get

$$\int_{-\delta}^{\delta} e^{\psi(t)}dt \sim e^{\psi(0)} \int_{-\delta}^{\delta} e^{-\frac{1}{2}\psi''(0)t^2}dt$$
$$\sim \sqrt{2\pi}e^{\psi(0)}\frac{1}{\sqrt{-\psi''(0)}} \quad \sim \quad \sqrt{2\pi}\frac{e^{e^u-1-\frac{1}{2}u}}{u^{n+1}}.$$

Integral over γ_1: Let $u+i\delta = re^{i\eta}$ in polar coordinates, so $r=\sqrt{u^2+\delta^2}$. Then $\gamma_1(\theta)=re^{i\theta}$ for $\eta \le \theta \le 2\pi-\delta$. For $z=re^{i\theta}$ on γ_1 we have $\operatorname{Re} e^z \le |e^z| = e^{\operatorname{Re} z} \le e^u$ (since γ_1 lies to the left of the vertical line through u). Therefore, the absolute value of the integrand

$$\left|\frac{e^{e^z-1}}{z^{n+1}}\right| = \frac{e^{\operatorname{Re}\{e^z\}-1}}{r^{n+1}} \le \frac{e^{e^u-1}}{u^{n+1}}\left(\frac{u}{r}\right)^{n+1}.$$

Now observe that $\frac{u}{r} = \left(1+\frac{\delta^2}{u}\right)^{-1} \le e^{-c_\delta/u}$ for some constant c_δ (for example, one may use the elementary inequality $1+x \ge e^{x/2}$ valid for $0 \le x \le \frac{1}{2}$) and hence $(u/r)^{n+1} \le e^{-c_\delta\frac{n+1}{u}} = e^{-c_\delta e^u}$. Thus we see that

$$I_1(n) \le 2\pi r\,\frac{e^{e^u-1}}{u^{n+1}}\,e^{-c_\delta e^u}$$
$$\le 4\pi u e^{-ce^u+\frac{1}{2}u}\,\frac{e^{e^u-1-\frac{1}{2}u}}{u^{n+1}}.$$

Here we used $r \le 2u$ which is obvious. The last factor is the asymptotic expression that we got for $I_0(n)$ (up to constants). The remaining factor is clearly $O(e^{-\frac{1}{2}e^u})$. Thus we see that $\frac{I_1(n)}{I_0(n)} = O(e^{-cn/\log n})$ (since $e^u \asymp \frac{n}{\log n}$).

Putting everything together, we arrive at

$$B_n = \frac{n!}{2\pi}(I_0(n)+I_1(n))$$
$$\sim \frac{n!}{\sqrt{2\pi}}\,\frac{e^{e^u-1-\frac{1}{2}u}}{u^{n+1}}$$

where $ue^u = n+1$.

Exercise 3. Write the asymptotic expression in terms of n (without using the implicitly defined u).

Remark 4. What were the key steps? The first was of course getting the explicit formula for the generating function that allowed us to write B_n as a contour integral. After that, the key point was to choose the right curve, so that the contour integral comes entirely from contributions close to one point. In the case at hand, it was important that the curve passed through the point u on the real line satisfying $ue^u = n+1$, and also that it passed through the point in the vertical direction. What is special about u is that it is a saddle point of the function $\psi(z) = e^z - 1 - \operatorname{Log} z$ (but not the only saddle point). At any saddle there are two perpendicular directions, one in which the value of ψ increases (when moving away from the saddle) and one in which the values of ψ decreases.

The vertical line is the line of steepest descent in our example. The general idea of using saddle points to find a convenient contour is explained in the next section.

5. The saddle point method - generalities

Here we are interested in evaluating integrals of the form $I(\lambda) = \int_{[A,B]} g(z)e^{\lambda f(z)}dz$ where f, g are holomorphic functions on some region and A, B are points in the region. The parameter λ is real and will go to infinity. We want the asymptotic behaviour of $I(\lambda)$.

In our examples, we shall take f and g to be entire functions. The idea consists of two steps:

(1) By holomorphicity, the integral does not change if we deform the contour of integration (keeping end points fixed at A, B). The first step is to choose the right contour, by which we mean whatever will make the second step work!

(2) Once the contour is chosen, write $I(\lambda)$ as an integral over an interval in the real line, $I(\lambda) = \int_a^b g(\gamma(t))e^{\lambda f(\gamma(t))}\dot{\gamma}(t)dt$. If the contour is well-chosen, Laplace's method (or some other) could apply to this integral and we could calculate the asymptotics.

How do we choose a good contour? Here are some guidelines. They are not guaranteed to work, but often do.

Guidelines: Consider the absolute value of $e^{\lambda f(z)}$ which is $e^{\lambda u(z)}$ where $u = \operatorname{Re} f$. For Laplace method to apply in the second step, we would like $e^{\lambda u(\gamma(t))}$ to be peaked at one point t_0 so that the entire contribution to the integral comes from a neighbourhood of t_0 (for large λ). Then $u(\gamma(t))$ should achieve a maximum at t_0. Let us assume t_0 is not an endpoint, then u achieves its maximum on γ at $\gamma(t_0)$.

But since u is harmonic, it has no maxima (or minima) in the plane. Therefore, $\gamma(t)$ must be a saddle point of u. Working backwards, we see that a good choice of γ is one that passes through one of the saddle points of u, and it should pass through the saddle point in such a way that the maximum of u on the curve is attained at this saddle point.

Example 1. If $f(z) = z^2$, then $u(x, y) = x^2 - y^2$ (where $z = x + iy$) and $\nabla u(z) = (2x, -2y)$. The only saddle point is $(0, 0)$. Along the x-axis, this is a minimum of u, and along the y-axis, this is a maximum of u. We could choose our curve to pass through 0 in the direction of the y-axis. Other choices are possible. In fact, $u < u(0)$ in the two sectors $|x| > |y|$ and $u < u(0)$ in the two sectors $|x| < |y|$. We can take any curve that stays strictly within the latter sectors, for example, $t + 2it$ is such a curve. Difficulties arise if the curve passes through the saddle point (here 0) tangentially to the boundary of the sector - hence we used the phrase 'strictly inside'.

In general, if $f = u + iv$ is holomorphic, then $f' = u_x + iv_x = u_x - iu_y$ (Cauchy-Riemann equations). Thus, saddles of u are precisely the zeros of f'. Further, the Taylor expansion of f near ζ looks like $f(z) = f(\zeta) + \frac{1}{2}(z - \zeta)^2 f''(\zeta) + \ldots$. We shall always assume that $f''(\zeta) \neq 0$. One can handle the cases where this is violated (we refer to them as degenerate saddles) by going to

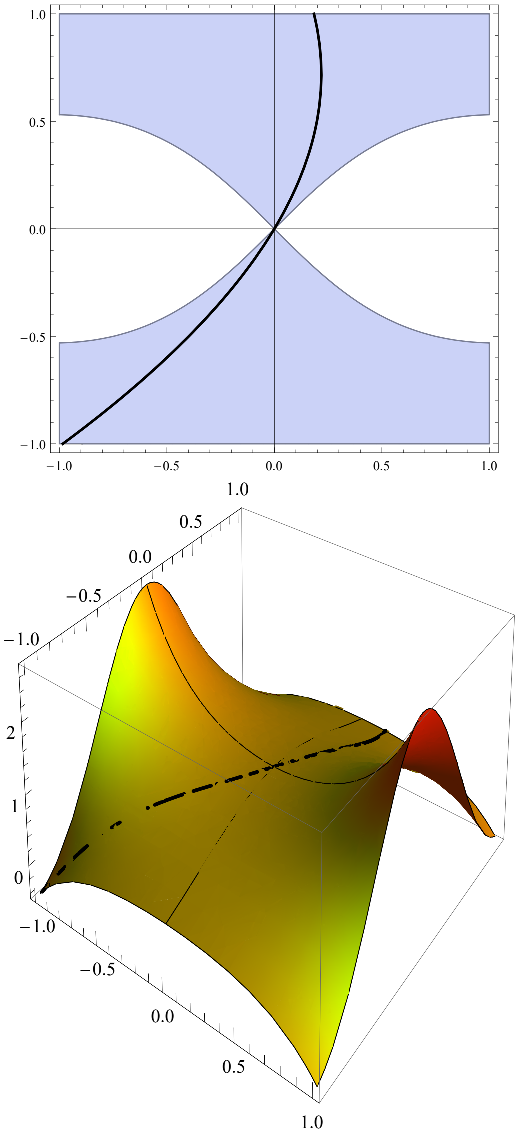

FIGURE 7. The function e^{z^2} has a saddle at zero. The blue region is where the value is lower than the value at the saddle point. What is shown in black is an admissible curve for the saddle point method. The y-axis is the curve of steepest descent.

higher derivatives, but that only increases the complications. For the applications we give, there will only be non-degenerate saddles. Then, if $z = \zeta + re^{i\theta}$ (small r) and $f''(\zeta) = Re^{i\alpha}$, then

$$u(z) = u(\zeta) + \frac{1}{2}r^2 R\cos(2\theta + \alpha) + \dots$$

The directions of steepest descent (respectively ascent) are the two values of θ (differing by π from each other) for which $\cos(2\theta + \alpha) = -1$ (respectively $+1$). The straight line through the saddle point in the direction of the steepest descent is called the *axis of the saddle*. In other words, it is the line of z such that $(z - \zeta)^2 f''(\zeta)$ is real and negative. The line of ascent is orthogonal to the line of descent. Further, the curves of constancy of u are given by $\cos(2\theta+\alpha)=0$, and these are a pair of curves passing through ζ at $\pi/4$ angle to the directions of ascent and descent.

Exercise 2. Show the same in an alternative way by going through the Hessian of u given by

$$Hu(\zeta)=\begin{bmatrix} u_{x,x} & u_{x,y} \\ u_{y,x} & u_{y,y} \end{bmatrix}$$

and the fact that the direction of the steepest descent is the direction of the eigenvector corresponding to the negative eigenvalue of Hu.

6. Integer partitions - preliminarie - 1

Let $p(n)$ denote the number of integer partitions of n. For example,

$$p(1)=1,\ p(2)=2,\ p(3)=3,\ p(4)=5,\ p(5)=7,\ p(6)=11,\ p(7)=15,\ldots$$

and Mathematica gives $p(100)=190569292$ (because it uses the formula of Hardy-Ramanujan-Rademacher!). The goal in this section is to get the asymptotic formula for $p(n)$ as $n\to\infty$.

Like with Bell numbers, we shall first find the generating function of the sequence $p(\cdot)$, and then express $p(n)$ as a contour integral in which n is a parameter. Unlike before, the generating function is now analytic in the unit disk, and has singularities as one approaches any root of unity on the circle. The *circle method* is to deform the curve in such a way that the contributions of theses singularities are extracted. Some technical details will be skipped[19].

6.1. Euler's generating function for $p(\cdot)$.

Let $p(0)=1$. Euler showed that

$$\sum_{n=0}^{\infty}p(n)q^n=\prod_{n=1}^{\infty}(1-q^n)^{-1}.$$

The product converges uniformly on compact subsets of the open unit disk $\mathbb{D}$ and hence the generating function is holomorphic in $\mathbb{D}$.

Proof. If we expand the right hand side formally, we get

$$\prod_{n=1}^{\infty}\sum_{k=0}^{\infty}q^{kn}=\sum_{\ell_1,\ell_2,\ldots}z^{\ell_1+2\ell_2+3\ell_3+\ldots}$$

where the last sum is over all sequences $\ell_1,\ell_2,\ldots$ that are eventually zero. The term q^n occurs as many times as there are such sequences with $\ell_1+2\ell_2+\ldots=n$. But that is precisely $p(n)$, by

[19] The result about $p(n)$ is due to Hardy and Ramanujan, but the circle method itself is attributed to Hardy and Littlewood. What we present is from a later refinement of the method and result is due to Rademacher, as presented in his paper *On the expansion of the partition function in a series*. I learned many of the points here from conversations with Surya Ramana of HRI, whose lecture gives a beautiful overview, with many details, of the topic. Some parts of the proof are taken from the paper *The circle method and non lacunarity of modular functions* by Sanoli Gun and Joseph Oesterlé.

identifying the sequence with the partition of n that consists of ℓ_1 ones, ℓ_2 twos, etc. We leave it as an exercise to fill in the details to make this argument rigorous. ■

6.2. **Contour integral representation for $p(n)$.** From the generating function, it follows that

$$p(n) = \frac{1}{2\pi i}\int_C q^{-n-1}\prod_{n=1}^{\infty}(1-q^n)^{-1}\,dq$$

for any closed curve C inside the disk that winds around the origin once. We shall write everything on the upper half plane $\mathbb{H}$ by composing with the map $z \mapsto q := e^{2\pi i z}$ from $\mathbb{H}$ to $\mathbb{D}\setminus\{0\}$. This map is surjective but not injective. In fact $q(z) = q(z+1)$ for all z and each vertical strip $\{z \in \mathbb{H} : t \leq \operatorname{Re} z < t+1\}$ gets mapped to the whole of $\mathbb{D}\setminus\{0\}$ under this mapping. Define $F : \mathbb{H} \mapsto \mathbb{C}$ by $F(z) = \prod\limits_{n=1}^{\infty}(1-q^n)^{-1}$. Then the integral formula becomes (note that $\frac{dq}{dz} = 2\pi i q$)

$$p(n) = \int_i^{i+1} q^{-n}\prod_{n=1}^{\infty}(1-q^n)^{-1}\,dz.$$

Under the map $z \mapsto q$ (which also maps $\mathbb{R}$ to S^1), the pre-image of roots of unity are precisely rational numbers in $\mathbb{R}$. By the mapping properties of q listed above, it suffices to look at the strip $0 \leq \operatorname{Re} z \leq 1$ and rational numbers in $[0,1]$ which we proceed to do next.

6.3. **Ford's geometric picture of fractions.** Rational numbers will be written as p/q with $q \geq 1$ and $\gcd(p,q) = 1$, unless otherwise stated. Let $C_{p/q}$ denote the circle of diameter $1/q^2$ whose lowest point is p/q. Thus, $C_{p/q} \subseteq \overline{\mathbb{H}}$ and its center is $\frac{p}{q} + i\frac{p}{2q^2}$.

Observation: The interiors of $C_{p/q}$ and $C_{r/s}$ are disjoint from each other. Two distinct circles are tangential to each other if and only if $ps - qr = \pm 1$.

Let us write $\frac{p}{q} \sim \frac{r}{s}$ if $C_{p/q}$ and $C_{r/s}$ are tangential to each other and say that $\frac{p}{q}$ is adjacent to $\frac{r}{s}$.

Proof. The square of the distance between the centers is

$$\left(\frac{p}{q}-\frac{r}{s}\right)^2 + \frac{1}{4}\left(\frac{1}{q^2}-\frac{1}{s^2}\right)^2 = \frac{1}{4q^4s^4}\left\{4q^2s^2(ps-rq)^2 + (s^2-q^2)^2\right\}$$

while the square of the sum of their radii is

$$\left(\frac{1}{2q^2}+\frac{1}{2s^2}\right)^2 = \frac{1}{4q^4s^4}(s^2+q^2)^2 = \frac{1}{4q^4s^4}\left\{4q^2s^2 + (s^2-q^2)^2\right\}.$$

Comparing the two expressions, it is clear that the circles can intersect if and only if $(ps-qr)^2 \leq 1$. If $ps - qr = 0$, the circles are the same, while if $ps - qr = \pm 1$, the circles are tangential (because the two quantities are then equal). ■

The *Farey series* of order n is the finite increasing sequence of all rational numbers p/q with $0 \leq p \leq q \leq n$.

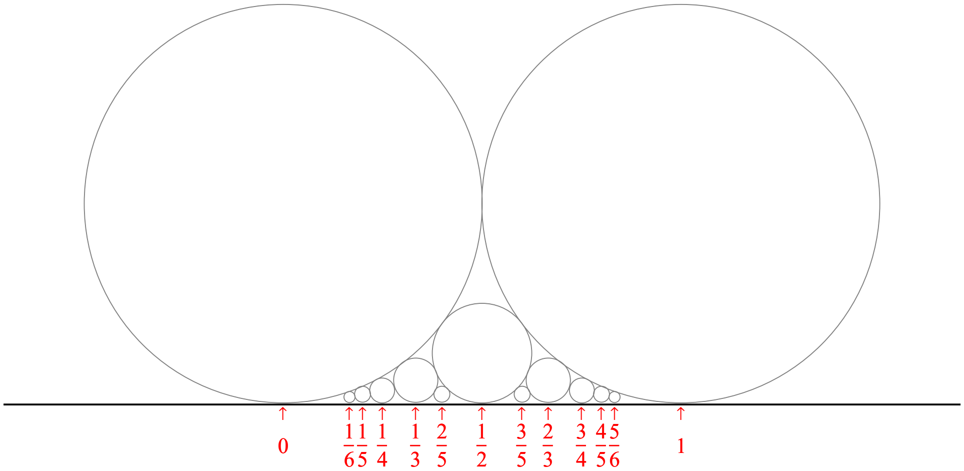

FIGURE 8. A picture of the Ford circles for fractions with $0 \le p \le q \le 6$. The fractions shown form the Farey series of order 6.

Exercise 1. Show that $\frac{p}{q} \sim \frac{r}{s}$ if and only if p/q and r/s are consecutive terms in the Farey series of some order.

Exercise 2. If $\frac{p}{q} \sim \frac{r}{s}$, then the set of all rationals $\frac{t}{u}$ adjacent to $\frac{p}{q}$ are of the form $\frac{r+np}{s+nq}$ for some $n \in \mathbb{Z}$. As $n \to \infty$ (respectively $n \to -\infty$), these circles approach $\frac{p}{q}$ from the right (respectively left).

There is a feature of the picture that we have not shown, that is, the region $\{z \in \mathbb{H} : \operatorname{Im} z < 1\}$ is made up of 'circular triangles'. This is established if we show that if $\frac{p}{q} < \frac{r}{s}$ are adjacent, then there is a (necessarily unique) $\frac{t}{u}$ between $\frac{p}{q}$ and $\frac{r}{s}$ and adjacent to both. To see that use the exercise to see that $\frac{p}{q}$ and $\frac{r}{s}$ are adjacent to each other in the Farey series of order n for some $n \ge 1$. Find the smallest $m > n$ for which there is a rational number between the two given rationals. We claim that this fraction is unique. Indeed, if more than one such rational, then we can find two that are adjacent to each other. But then they must be of the form $\frac{t}{m}$ and $\frac{t+1}{m}$ (in reduced form) in which case they cannot be adjacent to each other (because $(t+1)m - tm = m \ge 2$).

A new contour for integration: Let $N \ge 1$ and let $0 = x_1 < \ldots < x_{k_N} = 1$ be the Farey series of order N. As observed above, $C_{x_{k-1}}$ touches C_{x_k} for $2 \le k \le k_N$. Let this point of intersection be denoted $\zeta_{N,k}$. Make the convention that $\zeta_{N,0} = i$ and $\zeta_{N,k_N+1} = i+1$. Let $\gamma_{N,i}$ denote the a cirve that traces the upper arc of C_{x_i} connecting $\zeta_{N,i-1}$ to $\zeta_{N,i}$. Then $\gamma^N := \gamma_{N,0} + \ldots + \gamma_{N,k_N}$ is a curve from i to $i+1$ that is homotopic to the straight line $[i, i+1]$. Hence we certainly have

$$p(n) = \int_{\gamma_N} F(z) e^{-2\pi i n z} dz.$$

We want to let $N \to \infty$. Every rational number $x \in \mathbb{Q} \cap [0,1)$ belongs to some Farey series and hence enters the picture for sufficiently large N. Further, the arc of the circle that is part of $\gamma^{(N)}$

increases to $C_x^* := C_x \setminus \{x\}$. To see this, just note that there are infinitely many Ford circles that touch C_x on the right and on the left, and since these have to be smaller and smaller, their points of contact with C_x gets closer and closer to x. In other words, in the limit, γ_N becomes the union of all Ford circles (with their lowest points excluded). We claim that

$$p(n) = \sum_{x \in \mathbb{Q} \cap [0,1)} \int_{C_x^*} F(z) e^{-2\pi i n z} dz.$$

This can be justified by the dominated convergence theorem, provided we check that

$$\sum_{x \in \mathbb{Q} \cap [0,1)} \int_{C_x^*} |F(z)| |e^{-2\pi i n z}| |dz| < \infty.$$

Here $\int_\gamma |f(z)||dz|$ means $\int_a^b |f(\gamma(t))| \, |\dot{\gamma}(t)| dt$. We skip the justification for now, because it involves getting some estimates on F and that will be easier after we have the transformation rules for F. We just state for now that the integral on C_x^* will be bounded by (a constant times) $q^{-5/2}$ where $x = p/q$. Since there are at most q rational numbers in $[0,1)$ having denominator q, the sum above can be bounded by $\sum\limits_{q \ge 1} q^{-3/2}$ which is finite.

The geometric picture of rational numbers is quite useful. We show here a consequence for rational approximation[20]. The rest of this section are not necessary for the problem of finding asymptotics of $p(n)$ and may be safely skipped.

Fix any $x \in \mathbb{R}$ and let L_x denote the vertical line $x + iy$, $y > 0$. We claim that if $x \notin \mathbb{Q}$, then L_x intersects infinitely many circles. Clearly this is false if $x \in \mathbb{Q}$. However, when $x \notin \mathbb{Q}$, traverse L_x downwards. It must leave every circle that it enters (otherwise x must be the bottom point of the circle). But when it exits a circle, it must then enter a mesh triangle made up of three circles (one of which it just left) and then enter the smaller of the two other circles. As this repeats indefinitely, we have proved the claim. But now observe that L_x interesects $C_{p/q}$ if and only if $|x - \frac{p}{q}| \le \frac{1}{2q^2}$ (the projection of the circle to the x-axis must contain the point x). Thus, we have proved that for any irrational number x, there are infinitely many rational numbers $\frac{p}{q}$ satisfying $|x - \frac{p}{q}| \le \frac{1}{2q^2}$. This is (a slightly stronger form) of a theorem of Dirichlet.

Question: Fix $\lambda > 0$. Is it true that for every irrational x, there are infinitely many rational numbers $\frac{p}{q}$ such that $|x - \frac{p}{q}| \le \frac{\lambda}{q^2}$?

We have see that the answer is 'yes' for $\lambda = \frac{1}{2}$. Here is the optimal result.

Theorem 3 (Hurwitz). *The answer to the above question is 'yes' if $\lambda \ge \frac{1}{\sqrt{5}}$ and 'no' if $\lambda < \frac{1}{\sqrt{5}}$.*

Let $C_{p/q}^\lambda$ denote the circle concentric with $C_{p/q}$ and having radius λ/q^2. If we show that L_x intersects infinitely many of the circles $C_{p/q}^\lambda$ for any irrational number x, then we get that $|x - \frac{p}{q}| \le \frac{\lambda}{q^2}$ for infinitely many $\frac{p}{q}$. In particular, if we show that when L_x is in a mesh triangle, this inequality

[20] The proof is essentially in Ford's paper, although the presentation is not identical

holds for one of the three rational numbers whose circles bound the triangle, then we are done. This leads us to the following geometric consideration.

A geometric consideration: Consider three mutually tangential circles C_1, C_2, C_3 in the upper half plane and tangential to the real line at the points $0, 2x, 2$ respectively (where $0 < x < 1$) and having radii r, s, t respectively. First we show that r is the only free parameter and write the formulas for x, t, s in terms of r. Observe that s is the smaller of the three radii and without loss of generality we take $s \le r \le t$.

Indeed, the tangency of C_1 and C_2 forces $(r-s)^2 + 4x^2 = (r+s)^2$ which implies $rs = x^2$. Similarly, we must have $st = (1-x)^2$ (tangency of C_2 and C_3) and $rt = 1$ (tangency of C_1 and C_3).

$$x = \frac{\sqrt{r}}{\sqrt{r}+\sqrt{t}} = \frac{r}{r+1}, \qquad s = \frac{1}{(\sqrt{r}+\sqrt{t})^2} = \frac{r}{(1+r)^2}.$$

Now let $\hat{C}_i$ denote the circle C_i with the center fixed and radius scaled by λ. Then the projections of the $\hat{C}_i$ onto the x-axis are $[-\lambda r, \lambda r]$, $[2x - \lambda s, 2x + \lambda s]$ and $[2 - \lambda t, 2 + \lambda t]$ respectively. We wish to find a condition on λ such that (for any r), the union of these three intervals contains $[0, 1]$. This happens if and only if at least one of the following happens.

(1) The first and third interval overlap. This happens if and only if $2 - \lambda t \le \lambda r$ or equivalently, $\lambda \ge \frac{2}{r+\frac{1}{r}}$.

(2) The second interval overlaps with the first as well as the third. This happens if and only if $2x - \lambda s \le \lambda r$ and $2x + \lambda s \ge 2 - \lambda t$. In other words,

$$\lambda \ge \frac{2x}{r+s} = \frac{2}{(r+1)+\frac{1}{r+1}} \quad \text{ and } \quad \lambda \ge \frac{2(1-x)}{s+t} = \frac{2}{\frac{r+1}{r}+\frac{r}{r+1}}.$$

The function $u \mapsto \frac{2}{u+\frac{1}{u}}$ decreases from ∞ to 1 as u increases from 0 to 1. Since we assumed $r \le t$ we must have $r \le 1 \le t$ (because $rt = 1$), and hence $\frac{r}{r+1} \le \frac{1}{r+1}$. Therefore of the two inequalities here, the second one is the more stringent one.

In conclusion, the condition is that λ must be at least $\min\{\frac{2}{r+\frac{1}{r}}, \frac{2}{\frac{r+1}{r}+\frac{r}{r+1}}\}$. The first term is decreasing and the second is increasing for $0 < r < 1$, hence the minimum is largest when the two terms are equal. That is when $r = \frac{1}{r+1}$ or $r^2 + r - 1 = 0$ which gives $r = \frac{\sqrt{5}-1}{2}$. The minimum at this point is $\frac{2}{r+\frac{1}{r}} = \frac{2}{\sqrt{5}}$. Thus the desired condition is $\lambda \ge \frac{2}{\sqrt{5}}$.

Returning to Hurwitz's theorem, the above geometrical consideration applies to any three mutually tangential circles. Since the Ford circles at 0 and 1 is of radius $\frac{1}{2}$ (not 1 as in the geometrical consideration above), we see that the in Hurwitz's theorem, for $\lambda \ge \frac{1}{\sqrt{5}}$ the claim holds. The above considerations also suggest why it fails for $\lambda < \frac{1}{\sqrt{5}}$. complete this

6.4. **The modular group.** In complex analysis class you would likely have seen the fact that $SL_2(\mathbb{Z})$ acts on the upper half plane $\mathbb{H} = \{z : \operatorname{Im} z > 0\}$ by $g.z = \frac{az+b}{cz+d}$, where $g = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. Since

g and $-g$ act exactly the same way, one often thinks of the action of $PSL_2(\mathbb{R}) = SL_2(\mathbb{R})/\{\pm I\}$. This action is faithful. The subgroup $SL_2(\mathbb{Z})$ becomes $PSL_2(\mathbb{Z}) = SL_2(\mathbb{Z})/\{\pm I\}$ under quotienting. This is called the *Modular group*.

Example 4. $T = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ and $S = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ act as $T.z = z + 1$ and $S.z = \frac{-1}{z}$.

Lemma 5. *S and T generate the modular group.*

The basic idea is that of the Euclidean algorithm[21]

Proof. Let $g = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. Observe that

$$Sg = \begin{bmatrix} -c & -d \\ a & b \end{bmatrix} \quad \text{and} \quad T^{-q}g = \begin{bmatrix} a - qc & b - qd \\ c & d \end{bmatrix}.$$

Our goal will be to keep multiplying on the left by S or powers of T till we get to $\pm I$. That of course shows that g is in the group generated by S and T. As we proceed through the steps, we shall maintain our matrix in such a way that the $(1, 1)$ entry is at least as large as the $(2, 1)$ entry, in absolute value. Here is the first step:

Two cases, (a) $|c| > |a|$, (b) $|c| \le |a|$. In the first case we multiply g on the left by S. In the second case we multiply by ST^{-q} where q is chosen so thats $0 \le a - qc < |c|$. The new matrix is

$$\begin{bmatrix} -c & -d \\ a & b \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} -c & -d \\ a - qc & b - qd \end{bmatrix}.$$

In both cases observe that the requirement on the entries of the first column is maintained.

Applying this step repeatedly, after a finite number of steps we arrive (recall the Euclidean algorithm which ends with the remainder becoming 0 at some stage) at a matrix of the form

$$\begin{bmatrix} x & y \\ 0 & z \end{bmatrix}.$$

Since this is an $SL_2(\mathbb{Z})$ matrix, we must have $x = z = 1$ or $x = z = -1$. If $y = 0$, we have reached our goal. Otherwise multiply by $T^{\pm y}$ on the left to get $\pm I$. ■

6.5. **Transformation property of $F(z)$ under the action of the modular group.** The function F transforms in a nice way under the action of the modular group. Throughout this section, whenever we encounter the square root of a complex number, it is defined to be the branch of $\sqrt{w}$ on the complement of the negative real axis with the property that the root of a positive number is positive (explicitly, $\sqrt{w} = \sqrt{|w|}e^{i\frac{1}{2}arg(w)}$ where $-\pi < \arg(w) < \pi$.

[21] There are some tricky points one must pay attention to. I learned this proof from Soumya Bhattacharya.

Lemma 6. *$F(z+1) = F(z)$ and $F\left(-\frac{1}{z}\right) = \frac{1}{\sqrt{-iz}} e^{-\frac{i\pi}{12}(z+\frac{1}{z})} F(z)$. More generally,*

$$F(g.z) = \bar{\varepsilon}(g) \frac{1}{(cz+d)^{\frac{1}{2}}} e^{\frac{i\pi}{12}(g.z-z)} F(z)$$

where $\varepsilon(g)$ is a 12th root of unity (can be written explicitly but we shall not).

Since the action of g can be got by composing powers of S and T, it is clear that the third statement should follow from the first two. However, as we did not give an explicit expression for g in terms of S and T this is not entirely obvious. An alternative is to check that the set of g for which the transformation formula hold forms a group. Since it holds for S and T, it holds for all g in the group generated by S, T, which is the whole modular group. To carry this out, the explicit form of $\varepsilon(g)$ will be needed, of course.

Of the first two identities, the first one, $F(z+1) = F(z)$ is clear, since q is itself 1-periodic. It remains to prove the identity for $F(-1/z)$. There are two main ingredients.

- Euler's pentagonal number theorem: For $|q| < 1$, we have the expansion
$$\frac{1}{F(z)} = \sum_{n \in \mathbb{Z}} (-1)^n q^{\frac{1}{2}n(3n-1)}.$$
- Poisson summation: For $a, b \in \mathbb{C}$ with $\mathrm{Re}(b) > 0$, let $H(a,b) := \sum_{n \in \mathbb{Z}} e^{-\pi(bn^2+2an)}$. Then,
$$H(a,b) = \frac{1}{\sqrt{b}} e^{\frac{\pi a^2}{b}} H\left(\frac{ia}{b}, \frac{1}{b}\right).$$

These are explained later. Assuming these, here is how we derive the transformation rule relating $F(-1/z)$ to $F(z)$.

Proof. By the pentagonal number theorem

$$\begin{aligned}\frac{1}{F(z)} &= \sum_{n \in \mathbb{Z}} e^{i\pi n} e^{-\pi(-3izn^2+izn)} = H\left(\frac{1}{2}i(z-1), -3iz\right) \\ &= \frac{1}{\sqrt{-3iz}} e^{-\frac{i\pi(z-1)^2}{12z}} H\left(\frac{i(1-z)}{6z}, \frac{i}{3z}\right).\end{aligned}$$

In the second line we used the Poisson summation formula. On the other hand, substituting $-1/z$ in the first line, we see that

$$\frac{1}{F(-\frac{1}{z})} = H\left(3i, \right)$$

■

The group $SL_2(\mathbb{R})$ consists of 2×2 real matrices $g = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ satisfying $ad - bc = 1$. It has the subgroup $SL_2(\mathbb{Z})$ of matrices with integer entries with determinant one. Now, $SL_2(\mathbb{R})$ acts on the upper half plane $\mathbb{H}$ by

$$g.z = \frac{az+b}{cz+d}.$$

The action is holomorphic with derivative $g'.z = \frac{1}{(cz+d)^2}$ (the notation $g'.z$ is not meaningful, but we use it nevertheless).

Observe that $g.z = h.z$ if and only if $g = \pm h$, hence the quotient group $PSL_2(\mathbb{R}) := SL_2(\mathbb{R})/\{\pm I\}$ acts faithfully on $\mathbb{H}$ (meaning that distinct group elements do not act identically). The function $F(z)$ has a remarkable transformation formula under the action of the subgroup $SL_2(\mathbb{Z})$ (also known as the *modular group*). To state it in terms of a more familiar function that turns up often, we define the *Dedekind eta function*

$$\eta(z) = q^{\frac{1}{24}} \prod_{n=1}^{\infty} (1 - q^n)$$

which is the same as $q^{\frac{1}{24}} F(z)$. A function closely related to η is $\Delta(z) = (2\pi)^{12}\eta(z)^{24}$. The lemma below can be stated in standard jargon as "Δ is a modular form (in fact a cusp form) of weight 12".

Lemma 7. $(g'.z)^{-12}\Delta(g.z) = \Delta(z)$ *for all* $z \in \mathbb{H}$*, for all* $g \in SL_2(\mathbb{Z})$*. Consequently,*

$$F(e^{2\pi i \frac{p}{q} - 2\pi \frac{z}{q^2}}) = \omega_{p,q}\sqrt{\frac{z}{q}}e^{\frac{\pi}{12z} - \frac{\pi z}{12q^2}} F\left(e^{2\pi i \frac{p'}{q} - \frac{2\pi}{z}}\right)$$

where $pp' = -1$ *(mod q) and* $\omega_{p,q}$ *is a root of unity (which can be made explicit).*

Proof. ■

7. Integer partitions - preliminaries 2

For an integrable function $f : \mathbb{R} \mapsto \mathbb{C}$, define its Fourier transform $\hat{f} : \mathbb{R} \mapsto \mathbb{C}$ as $\hat{f}(\lambda) = \int_{\mathbb{R}} f(x)e^{-2\pi i \lambda x}dx$.

Theorem 1 (Poisson summation formula)**.** *Assume that* f *is smooth (*C^2 *suffices) and that* f, f' *decay fast (*$O(|x|^{-2})$ *suffices) at* $\pm\infty$*. Then*

$$\sum_{n\in\mathbb{Z}} f(n) = \sum_{n\in\mathbb{Z}} \hat{f}(n).$$

Proof. Define $g : \mathbb{R} \mapsto \mathbb{C}$ by $g(x) = \sum_{n\in\mathbb{Z}} f(x+n)$. Since $\|f\|_{\sup[n,n+1]}$ is summable by assumption on the decay of f, the series defining g converges uniformly and thus g is a continuous 1-periodic function. Further, the same series with f' in place of f also converges uniformly for the same reason (assumption on the decay of f'). This allows us to differentiate term by term and see that $g'(x) = \sum_{n\in\mathbb{Z}} f'(x+n)$ and that it is also a continuous, 1-periodic function.

As $\{e^{2\pi nt} : n \in \mathbb{Z}\}$ forms an orthonormal basis for $L^2(S^1)$, we have the L^2 expansion

$$g(x) \overset{L^2}{=} \sum_{n\in\mathbb{Z}} \tilde{g}(n)e^{2\pi i n x}, \quad \text{where } \tilde{g}(n) = \int_0^1 g(x)e^{-2\pi i n x}dx.$$

However, g is continuously differentiable, and by an integration by parts one sees that $\tilde{g'}(n) = 2\pi in\tilde{g}(n)$. Since Fourier coefficients of g' are square summable, we must have $\sum_n |\tilde{g}(n)|^2 n^2 < \infty$. By Cauchy-Schwarz inequality,

$$(\sum_{n\neq 0} |\tilde{g}(n)|)^2 \leq \left(\sum_{n\neq 0} n^2|\tilde{g}(n)|^2\right)\left(\sum_{n\neq 0}\frac{1}{n^2}\right)$$

and thus $\tilde{g}(n)$ is absolutely summable. Consequently, the Fourier series of g converges to g uniformly (Weierstrass' M-test). In particular, for each $x \in \mathbb{R}$, we have

$$\sum_{n\in\mathbb{Z}} \tilde{g}(n)e^{2\pi innx} = g(x). \tag{1}$$

Now we claim that $\tilde{g}(n) = \hat{f}(n)$. Indeed, using Fubini's to interchange sum and integral,

$$\begin{aligned}\tilde{g}(n) &= \sum_{k\in\mathbb{Z}} \int_0^1 f(x+k)e^{-2\pi inx}dx = \sum_{k\in\mathbb{Z}} \int_k^{k+1} f(x)e^{-2\pi inx}dx \\ &= \int_{-\infty}^{\infty} f(x)e^{-2\pi inx}dx\end{aligned}$$

which is just $\hat{f}(n)$. Substituting this and the definition of g in (1) we get

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

Set $x = 0$ to get Poisson summation formula. ■

Corollary 2. *For $a, b \in \mathbb{C}$ with $\mathrm{Re}(b) > 0$, let $H(a,b) = \sum_{n\in\mathbb{Z}} e^{-\pi(bn^2+2an)}$. Then,*

$$H(a,b) = \frac{e^{\pi\frac{a^2}{b}}}{\sqrt{b}} H\left(\frac{ia}{b}, \frac{1}{b}\right).$$

Proof. Let $f(x) = e^{-\pi(bx^2+2ax)}$. Then

$$\begin{aligned}\hat{f}(\lambda) &= \int_{\mathbb{R}} e^{-\pi(bx^2+2ax)-2\pi i\lambda x}dx \\ &= e^{\pi b(\frac{a+i\lambda}{b})^2}\int_{\mathbb{R}} e^{-\pi b(x+\frac{a+i\lambda}{b})^2} \quad = \frac{e^{\pi\frac{a^2}{b}}}{\sqrt{b}}e^{-\pi(\frac{1}{b}\lambda^2+\frac{2ia}{b}\lambda)}.\end{aligned}$$

Applying the Poisson summation formula, we get the identity in the corollary. ■

Example 3. Theta function: An immediate application of the corollary is the transformation rule for the *theta function* $\theta(z) = \sum_{n\in\mathbb{Z}} q^{n^2}$ with $q = e^{2\pi iz}$. Then $\theta(z+1) = \theta(z)$ obviously. Further, by the Poisson summation formula (see the corollary with $b = 1$

Exercise 4. Define the *theta function* $\theta(z,w) = \sum_{n\in\mathbb{Z}} e^{i\pi wn^2}e^{2\pi izn}$. Then show that $\theta(z+1,w) = \theta(z)$ and $\theta(-\frac{1}{z}, \frac{w}{z}) =$

8. INTEGER PARTITIONS - ASYMPTOTICS VIA THE CIRCLE METHOD

We first give an overview of the whole proof and then deal with the individual steps.

(1) Find the generating function of $p(n)$ as $\sum_{n=0}^{\infty} p(n)q^n = \prod_{n=1}^{\infty}(1-q^n)^{-1} =: f(q)$. Use Cauchy's integral formula and make the substitution $q = e^{2\pi iz}$ to get the contour integral formula

$$p(n) = \frac{1}{2\pi i}\int_i^{i+1} F(z)e^{-2\pi i(n+1)z}dz$$

where $F(z) = f(q)$ for $z \in \mathbb{H}$.

(2) Consider the Farey series of order N, say $0 = x_1 < \ldots < x_{k_N} = 1$ (these x_is are all the fractions with denominator at most N). Deform the contour $[i, i+1]$ to $\gamma_{x_1,N} + \ldots + \gamma_{x_{k_N},N}$, where $\gamma_{x_k,N}$ is the upper arc on the Ford circle C_{x_k} traversed clockwise, from the point of intersection of this circle with $C_{x_{k-1}}$ to the point of intersection with $C_{x_{k+1}}$. This is the description for all except the first and last. For the first, the starting point of $\gamma_{0,N}$ is taken to be i and for the last the ending point of $\gamma_{1,N}$ to be $i + 1$. Conclude that

$$p(n) = \frac{1}{2\pi i}\sum_{j=1}^{k_N}\int_{\gamma_{x_j,N}} F(z)e^{-2\pi iz(n+1)}dz.$$

(3) Let $N \to \infty$ and observe that the arc γ_x, N increases to $C_x^* := C_x \setminus \{x\}$ (the whole circle except the bottom point) for $x \in \mathbb{Q} \cap (0, 1)$. Further, $\gamma_{0,N}$ increases to the right half of C_0 from i to 0 and $\gamma_{1,N}$ increases to the left half of C_1 from 1 to $1 + i$. Argue that

$$p(n) = \frac{1}{2\pi i}\sum_{x\in\mathbb{Q}\cap[0,1)}\int_{C_x^*} F(z)e^{-2\pi iz(n+1)}dz.$$

Here, we have used the identity $F(z) = F(z + 1)$ to replace the arc of C_1 from 1 to $1 + i$ by the arc from 0 to i on C_0, which combines with the arc from i to 0 to give C_0^*.

(4) Establish the transformation formulas: $F(z + 1) = F(z)$ which is trivial and

$$F(z) = \sqrt{-iz}\, e^{\frac{i\pi}{12}\left(z+\frac{1}{z}\right)}F\left(-\frac{1}{z}\right).$$

Here $w = -iz$ is in the right half plane, and the square root is defined by $\sqrt{w} = \sqrt{|w|}e^{\frac{i}{2}\text{Arg}(w)}$ where $-\pi < \text{Arg}(w) < \pi$.

(5) Consider the action of $G = SL_2(\mathbb{Z})$ on $\mathbb{H}$ given by $g.z = \frac{az+b}{cz+d}$ (in fact this defines an action by $SL_2(\mathbb{R})$). Show that the action of any element can be written as a composition of the maps $z \mapsto z + 1$ and $z \mapsto \frac{-1}{z}$. Consequently, we get the more general transformation formula

$$F(g.z)e^{-\frac{i\pi g.z}{12}} = \varepsilon(g)(cz + d)^{-\frac{1}{2}}F(z)e^{-\frac{i\pi z}{12}}.$$

where $\varepsilon(g)$ is a certain 12th root of unity (can be written more explicitly).

(6) Now fix a fraction $0 \leq x = \frac{a}{c} < 1$ and find integers b, d such that $ad - bc = 1$ (take a rational number smaller than a/c whose Ford circle touches the Ford circle of a/c. Then by writing $g = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ and writing

$$g.z = \frac{az+b}{cz+d} = \frac{a}{c} - \frac{1}{c^2(z + \frac{d}{c})},$$

we see that the action of g maps C^*_∞ onto C^*_x (recall that $C_\infty = (\mathbb{R} + i) \cup \{\infty\}$, by C^*_∞ we just mean $\mathbb{R} + i$). The derivative of this action is $\frac{d}{dz} g.z = \frac{1}{(cz+d)^2}$. Therefore,

$$\begin{aligned} \int_{C^*_x} F(z) e^{-2\pi i z(n+1)} dz &= \int_{C^*_\infty} F(g.z) e^{-2\pi i (n+1) g.z} \frac{1}{(cz+d)^2} dz \\ &= \varepsilon(g) \int_{C^*_\infty} F(z) e^{\frac{i\pi}{12}(g.z - \frac{1}{z})} e^{-2\pi i (n+1) g.z} \frac{1}{(cz+d)^{\frac{5}{2}}} dz \end{aligned}$$

(7) Put together the previous steps to see that

$$p(n) = \frac{1}{2\pi i} \sum_{x = \frac{a}{c} \in [0,1) \cap \mathbb{Q}} \varepsilon(g) \int_{C^*_\infty} F(z) e^{\frac{i\pi}{12}(g.z - \frac{1}{z})} e^{-2\pi i (n+1) g.z} \frac{1}{(cz+d)^{\frac{5}{2}}} dz$$

.

CHAPTER 6

Moment problems

1. Moment problems

If μ is a measure on $\mathbb{R}$, the number $\alpha_k = \int x^k d\mu(x)$ is said to be its kth moment, if it exists. Throughout this section, we work with measures for which all moments do exist. In particular, all measures will be finite, and often we normalize them to be probability measures.

The moment problem: Given a sequence $\alpha = (\alpha_0, \alpha_1, \ldots)$ of real numbers, does there exist a Borel measure on $\mathbb{R}$ whose nth moment is α_n? Is it unique? What if the measure is restricted to the half-line $[0, \infty)$ or to an interval?[22]

Necessary condition: The integral of a positive function against a measure is positive. Suppose α is the moment sequence of a measure μ whose support is the closed set $I \subseteq \mathbb{R}$. Then, for any (real) polynomial $p(x) = \sum_{j=0}^{n} c_j x^j$ such that $p(x) \geq 0$ for all $x \in I$, we must have $\int p(x) d\mu(x) \geq 0$. Since $\int p(x) d\mu(x) = \sum_{j=0}^{n} c_j \alpha_j$, writing $L(p) := \sum_{j=0}^{n} c_j \alpha_j$, we see that

$$L(p) \geq 0 \quad \text{whenever} \quad p(x) \geq 0 \text{ for all } x \in I. \tag{1}$$

The first main theorem is that this condition is also sufficient.

Theorem 1 (Existence part of the moment problem)**.** *Let I be a closed subset of $\mathbb{R}$ and let $\alpha = (\alpha_0, \alpha_1, \ldots)$ be a sequence of real numbers. There exists a measure μ on I such that $\int x^k d\mu(x) = \alpha_k$ for all $k \geq 0$ if and only if the positivity condition* (1) *holds.*

We shall prove this in the next section. For now, we take I to be an interval and find more tractable conditions for (1) to hold. The question is, what polynomials are positive on I?

The whole line: Let $I = \mathbb{R}$. Write

$$p(x) = a_n \prod_{j=1}^{k} (x - t_j) \prod_{j=1}^{\ell} (x - z_j)(x - \bar{z}_j)$$

where $k, \ell \geq 0$ and $t_j \in \mathbb{R}$ and $z_j \in \mathbb{C} \setminus \mathbb{R}$. Let $q(x) = \prod_{j=1}^{\ell} (x - z_j)$ (a complex polynomial), so that $p(x) = a_n |q(x)|^2 \prod_{j=1}^{k} (x - t_j)$. Thus p is positive on $\mathbb{R}$ if and only if $a_n > 0$ (take $x \to +\infty$ to see

[22] There are various names, such as the Hamburger moment problem, the Stieltjes' moment problem, Hausdorff moment problem, etc.

this) and each distinct real root of p occurs with even multiplicity. Then, letting $q = q_1 + iq_2$ and $r(x)^2 = \prod_{j=1}^k (x - t_j)$,

$$p = (\sqrt{a_n} r q_1)^2 + (\sqrt{a_n} r q_2)^2.$$

Conversely, any such polynomial is positive on $\mathbb{R}$.

Thus, in condition (1) it suffices to take p to be the square of another polynomial and thus the condition becomes $L(p^2) \geq 0$ for all $p \in \mathcal{P}$. Writing $p(x) = \sum_{j=0}^n c_j x^j$, this can be written in terms of the sequence α as

$$0 \leq \sum_{j,k=0}^{n} c_j c_k \alpha_{j+k} \quad \text{for all } n \geq 1 \text{ and } c_0, \dots, c_n \in \mathbb{R}. \tag{2}$$

This can also be phrased as saying that the infinite matrix $H_\alpha = (\alpha_{i+j})_{i,j \geq 0}$ is positive semi-definite (meaning that $\det[(H_\alpha(i,j))_{0 \leq i,j \leq n}] \geq 0$ for all $n \geq 0$).

Half-line: Let $I = [0, \infty)$. Going by the same logic as before, we see that if p is positive on $[0, \infty)$, then all its real roots in $(0, \infty)$ must have even multiplicity, but the negative roots are not restricted. Hence

$$p(x) = q(x) \prod_{j=1}^{m} (x + t_j)$$

where $q(x) \geq 0$ for all $x \in \mathbb{R}$. Expanding the product further, and writing $q = q_1^2 + q_2^2$, we see that $p(x)$ is a positive linear combination of polynomials of the form $x^k r(x)^2$ where r is a real polynomial and $k \geq 0$. Since even powers of x can be absorbed into r, we see that any polynomial positive on $[0, \infty)$ is a linear combination (with positive coefficients) of polynomials of the form r^2 and $xr^2(x)$. Thus the condition (1) is equivalent to

$$L(p^2) \geq 0 \text{ and } L(xp^2(x)) \geq 0 \quad \text{for all} \quad p \in \mathcal{P}.$$

Again, writing $p(x) = \sum_{j=0}^n c_j x^j$, we can write these conditions as

$$0 \leq \sum_{j,k=0}^{n} c_j c_k \alpha_{j+k} \quad \text{and} \quad 0 \leq \sum_{j,k=0}^{n} c_j c_k \alpha_{j+k+1} \quad \text{for all } n \geq 1 \text{ and } c_0, \dots, c_n \in \mathbb{R}.$$

This is equivalent to positivity of the determinants of $(H_\alpha(i,j))_{0 \leq i,j \leq n}$ and $(H_\alpha(i,j))_{0 \leq i \leq n, 1 \leq j \leq n+1}$.

Compact interval: Let $I = [0, 1]$. We claim that if $p \geq 0$ on I, then it can be written as a positive linear combination of the polynomials $x^k(1-x)^\ell$ for $k, \ell \geq 0$. Accepting this claim, the condition (1) becomes equivalent to

$$L(x^k(1-x)^\ell) \geq 0 \quad \text{for all } k, \ell \geq 0.$$

Expanding $(1-x)^\ell$, this is the same as

$$\sum_{j=0}^{\ell} \binom{\ell}{j} (-1)^j \alpha_{k+j} \geq 0$$

There is a nice way to express this. Define the difference operator from sequences to sequences by $(\Delta c)_k = c_{k+1} - c_k$. Then the above conditions can also be written succinctly as $(-1)^p(\Delta^p \alpha)_k \geq 0$ for all p, k (the original sequence is positive, the differences are negative, second differences are positive, etc.).

Exercise 2. (1) Prove the claim. (2) Prove the equivalence of the derived condition to alternating signs of successive differences.

2. Some theorems similar in spirit to the moment problems

There are other theorems that one sees in analysis that are similar in spirit to the moment problem. We mention a couple of them in this section. These will not be required in future sections.

Theorem 3 (Riesz's representation theorem.)**.** *Let X be a locally compact Hausdorff space. Let $L : C_c(X) \mapsto \mathbb{R}$ be a linear functional. Then, there exists a (regular) Borel measure μ on X such that $Lf = \int f d\mu$ for all $f \in C_c(X)$ if and only if L is positive (i.e., $L(f) \geq 0$ whenever $f \geq 0$).*

Presumably Riesz had the solutions to the moment problems in mind when he formulated this theorem. But the solutions to the moment problems cannot be deduced directly from Riesz's representation.

Other questions with the same flavour as the moment problem are as follows: We are given a linear functional on a subspace of continuous functions. The problem being to determine if it comes from a measure. Here are two concrete problems of interest.

Example 4. For a measure μ on S^1, we define its Fourier coefficients $\hat{\mu}(k) = \int e_{-k} d\mu$. The question "what sequences of complex numbers can arise as the Fourier coefficients of a measure?" is clearly very similar in spirit to the moment problem.

Given $\alpha = (\alpha_k)_{k \in \mathbb{Z}}$, a necessary condition for α to be the Fourier coefficients of a measure is that for any trigonometric polynomial $p = \sum_{k=-n}^{n} c_k e_k$,

$$0 \leq \sum_{j,k=-n}^{n} c_j \bar{c}_k \alpha_{j-k}.$$

We leave it to you to figure why. The non-trivial point is that these conditions are also sufficient. Here uniqueness of the measure comes for free!

Example 5. For a finite measure μ on $\mathbb{R}$, define its Fourier transform $\hat{\mu} : \mathbb{R} \mapsto \mathbb{C}$ by $\hat{\mu}(t) = \int_{\mathbb{R}} e_{-t} d\mu$ where $e_t(x) = e^{itx}$. Given a function $f : \mathbb{R} \mapsto \mathbb{C}$, is it the Fourier transform of a finite measure? Two necessary conditions are

(1) f is continuous

(2) $\sum_{j,k=1}^{n} c_j \bar{c}_k \hat{\mu}(t_j - t_k) \geq 0$ for all $n \geq 1$, all $c_1, \ldots, c_n \in \mathbb{C}$ and all $t_1, \ldots, t_n \in \mathbb{R}$.

Bochner's theorem asserts that these conditions are also sufficient. Again, uniqueness holds, in contrast to moment problem on the whole line.

Given the similarity between the moment problems, the theorems on Fourier transforms, and the Riesz representation theorem, one might prefer a more abstract statement that captures the general situation. We give one such theorem due to M. Riesz.

3. Marcel Riesz's extension theorem

Before we state the theorem, recall that a cone in a real vector space is a set that is closed under multiplication by positive scalars. A convex cone is a cone that is closed under convex combination of its elements. This is the same as a cone that is closed under addition of its elements. For example, the first quadrant is a convex cone in $\mathbb{R}^2$.

Theorem 6 (M. Riesz's extension theorem)**.** *Let W be a subspace of a real vector space X. Let K be a convex cone in X such that ~~$span(K) + W = X$~~ $K + W = X$. Let $L : W \mapsto \mathbb{R}$ be a linear functional such that $L(v) \geq 0$ for all $v \in W \cap K$. Then, there exists a linear functional $\tilde{L} : X \mapsto \mathbb{R}$ such that $\tilde{L}(v) = L(v)$ for all $v \in W$ and $\tilde{L}(v) \geq 0$ for all $v \in K$.*

The proof will be reminiscent of the proof of Hahn-Banach theorem that you have seen in Functional analysis class. Historically, perhaps both that proof and this were arrived at in stages, by polishing and making more abstract the solutions of the moment problems.

Proof. If $K \subseteq W$, then $W = X$ and there is nothing to prove. Otherwise, pick $u \in K \setminus W$ and let $W' = W + \mathbb{R}u$, a subspace strictly larger than W. We show that it is possible to extend L to W' so that it is positive on $K \cap W'$. There is no choice but to define the extension as $L'(w + \alpha u) = L(w) + \alpha t$ for some $t \in \mathbb{R}$. The only freedom is in t, and we must choose it so that $L(w) + \alpha t \geq 0$ whenever $w \in W$ and $w + \alpha u \in K$. It is enough to check this condition for $\alpha = \pm 1$, since K and W are both closed under multiplication by $|\alpha|$. Thus, the conditions for positivity of L' are precisely that

$$L(w) + t \geq 0 \text{ for } w \in W \cap (K - u), \quad \text{and} \quad L(w) - t \geq 0 \text{ for } w \in W \cap (K + u).$$

We may rewrite this as

$$-L(w_1) \leq t \leq L(w_2) \text{ for all } w_1 \in W \cap (K - u) \text{ and } w_2 \in W \cap (K + u).$$

Such a choice of t is possible if and only if $-L(w_1) \leq L(w_2)$ for all $w_1 \in W \cap (K + u)$ and $w_2 \in W \cap (K - u)$. But if $w_1 \in W \cap (K - u)$ and $w_2 \in W \cap (K + u)$, then $w_2 + w_1 \in K \cap W$ and hence $L(w_2) + L(w_1) = L(w_2 + w_1) \geq 0$ by the positivity of L. This completes the proof that a positive (on K) linear functional on a subspace can be extended to a subspace got by adding one new element.

Gap in the proof: We have not checked if $W \cap (K + u)$ is non-empty (note that $W \cap (K - u)$ cannot be empty as it contains the zero vector). If $W \cap (K + u)$ is empty, it may happen that the value of t obtained above is $+\infty$. This is possible if we only assume that $\text{span}(K) + W = X$ as we originally did. But if we assume that $K + W = X$, then $-u = w + k$ for some $w \in W$ and $k \in K$, and hence, $w \in W \cap (K + u)$, ensuring the non-emptyness of $W \cap (K + u)$.

The rest is the usual Zorn's lemma ritual. Consider all positive (on K) extensions of L, i.e., tuples $(\hat{W}, \hat{L})$ such that $\hat{W}$ is a subspace containing W, $\hat{L}$ is a positive (on K) linear functional on $\hat{W}$ that extends L. This set is partially ordered by the order $(W_1, L_1) \leq (W_2, L_2)$ if $W_1 \subseteq W_2$ and L_2 is an extension of L_1. Given a totally ordered subset (a chain) $\{(W_i, L_i)\}$, it is clear that $(\cup_i W_i, \vee_i L_i)$ (how are they defined?) is a maximal element of the chain. Applying Zorn's lemma, we get a maximal element (W_0, L_0). If $W_0 \neq X$, then as above, it is possible to extend L_0 to a strictly larger subspace while preserving positivity on K, contradicting the maximality of (W_0, L_0). Thus, $W_0 = X$ and the theorem is proved. ■

Remark 7. Here is a standard example to show that the hypothesis $K + W = X$ cannot be replaced by span$(K) + W = X$. Let $X = \mathbb{R}^2$, $W = \{(x, 0) : x \in \mathbb{R}\}$, and $K = \{(x, y) : y > 0\}$. Let $L(x, 0) = x$, a linear functional positive on K (tautologically, since $W \cap K = \emptyset$). Any extension must look like $\tilde{L}(x, y) = x + ty$ for some $t \in \mathbb{R}$. But then $\tilde{L}(-2t, 1) = -t$ while $\tilde{L}(0, 1) = t$, showing that both cannot be positive although $(0, 1)$ and $(-2t, 1)$ are both in K.

Remark 8. (For those who feel a pang of uneasiness when using Zorn's lemma). If there are countably many elements $u_1, u_2, \ldots$ in K such that $X = W + \text{span}\{u_1, u_2, \ldots\}$, then a simple induction argument may be used in place of Zorn's lemma. In many applications this suffices.

As a corollary, we derive solutions to the moment problems. Just to illustrate the idea, we first deal with the case when I is compact.

Theorem 9. *Let $I \subseteq \mathbb{R}$ be a non-empty compact set, and let $\alpha = (\alpha_0, \alpha_1, \ldots)$ be a sequence of real numbers such that if a polynomial $p(x) = \sum_{j=0}^{n} c_j x^j \geq 0$ for all $x \in I$, then $\sum_{j=0}^{n} \alpha_j c_j \geq 0$. Then, there is a Borel measure μ on I such that α_n is the nth moment of μ for every n.*

Proof. Let $X = C[0, 1]$, $W = \mathcal{P}$, $K = \{f \in C[0, 1] : f \geq 0\}$. It is clear that W is a subspace of X and K is a convex cone. To see that $W + K = X$, write any $f \in C[0, 1]$ as $(f + \|f\|_{\sup(I)}) - \|f\|_{\sup(I)}$. The first summand is in K while the second is in $\mathcal{P}$ (being a constant!). Hence L extends to a positive linear functional on $C[0, 1]$, which, by Riesz's representation theorem is represented by integration with respect to a Borel measure on $[0, 1]$. ■

Remark 10. In the above theorem, uniqueness of the measure is easy to prove. This is because polynomials are dense in $C(I)$, by Weierstrass' theorem. Hence, the extension has to be unique (two bounded linear functionals on a Banach space that agree on a dense subset must agree everywhere). Uniqueness is not true in general, and not easy to prove when it is, for non-compact domains.

The use of Riesz's representation was a little extravagant, but employed to make the point quickly. We now give a direct argument that works for unbounded sets also.

Theorem 11. *If I be a closed subset of $\mathbb{R}$. Let $\alpha = (\alpha_0, \alpha_1, \ldots)$ be a real sequence. Define $L(p) = \sum_{j=0}^{n} c_j \alpha_j$ for any polynomial $p(x) = \sum_{j=0}^{n} c_j x^j$. If $L(p) \geq 0$ whenever $p \geq 0$ on I, then there exists a Borel measure μ such that $\alpha_n = \int x^n d\mu(x)$ for all n.*

Proof of the Theorem for the case $I = \mathbb{R}$. **Step 1:** To apply M. Riesz's extension theorem, let (here $\mathbf{1}_{(-\infty,\infty]}$ just means the constant function $\mathbf{1}$)

$$W = \mathcal{P}, \;\; V = \text{span}\{\mathbf{1}_{(-\infty,b]} : b \in \mathbb{R} \cup \{\infty\}\}, \;\; X = W + V, \;\; K = \{f \in X : f \geq 0\},$$

and $L : W \mapsto \mathbb{R}$ as in the statement of the theorem. To apply the extension theorem, we need to check that $W + K = X$. If $f \in X$, write $f = p + g$ with $g = \sum_{i=1}^{m} a_i \mathbf{1}_{A_i}$ where A_i are disjoint (left-open, right-closed) intervals, possibly including intervals of the form $(-\infty, b]$ and (b, ∞). Let $a = \min\{a_1, \ldots, a_m\}$ and write $f = (p + a) + (g - a)$. Clearly $g - a \in K$ and $p + a \in W$. Thus $W + K = X$. Consequently, L extends to all of X as a positive linear functional. We continue to denote it by L.

Step 2: To get a measure, define $G(t) = L(\mathbf{1}_{(-\infty,t]})$. If $s < t$, then $0 \leq \mathbf{1}_{(s,t]} = \mathbf{1}_{(-\infty,t]} - \mathbf{1}_{(-\infty,s]}$ and hence $G(s) \leq G(t)$ by the positivity of L. Thus, G is an increasing function on $\mathbb{R}$. It is also clear that $0 \leq G(t) \leq \alpha_0$ for all t because $0 \leq \mathbf{1}_{(-\infty,t]} \leq 1$.

We claim that $G(-\infty) = 0$ and $G(+\infty) = \alpha_0$. To see this, use the Chebyshev-like idea and write $\mathbf{1}_{(-\infty,-b]}(t) \leq t^2/b^2$ for $b > 0$. Applying L, we get $G(-b) \leq \alpha_2/b^2$ which shows that $G(t) \to 0$ as $t \to -\infty$. Similarly, show that $\alpha_0 - G(b) = L(\mathbf{1}_{(b,\infty)}) \leq \alpha_2/b^2$ to see that $G(b) \to \alpha_0$ as $b \to +\infty$. The claim is proved.

It would be clean if we could show that G is right-continuous, but I was not able to (is it false in general?). But we can easily modify it to be right continuous by defining $F : \mathbb{R} \mapsto \mathbb{R}_+$ by

$$F(t) = \inf\{G(s) : s \in \mathbb{Q}, s > t\}.$$

Clearly F is increasing and right-continuous. It also satisfies $F(+\infty) = \alpha_0$ and $F(-\infty) = 0$. Therefore, there exists a unique Borel measure[23] μ on $\mathbb{R}$ such that $\mu(a, b] = F(b) - F(a)$ for any $a < b$.

Let D be the set of continuity points of G. Then D^c is countable (since G is increasing) and hence D is dense. We note for future use that $F(t) = G(t)$ for all $t \in D$.

Step 3: We make some estimates on the tails of μ. Using $\mathbf{1}_{|x| \geq b} \leq b^{-2k}|x|^{2k}$ and positivity of L, we get

$$L(\mathbf{1}_{(-\infty,-b]}) + L(\mathbf{1}_{[b,\infty)}) \leq b^{-2k}\alpha_{2k}$$

for every $k \geq 1$. From this, it easily follows that (at least when $b \in D$)

$$\mu(-\infty, b] + \mu[b, \infty) \leq \alpha_{2k} b^{-2k}.$$

[23] A quick proof if you have not seen this before. Define $H : (0, \alpha_0) \mapsto \mathbb{R}$ by $H(u) = \inf\{t \in \mathbb{R} : F(t) \geq u\}$. Then, if λ is the Lebesgue measure on $(0, \alpha_0)$, define $\mu = \lambda \circ H^{-1}$. Check that $H(u) \leq t$ if and only if $u \leq F(t)$ or in other notation $H^{-1}(-\infty, t] = (0, F(t)]$. Therefore, $\mu(-\infty, t] = \lambda(0, F(t)] = F(t)$.

From this we bound the tails of integrals with respect to μ as follows[24].

$$\int |u|^n \mathbf{1}_{|u|>b} d\mu(u) = \int_0^\infty \mu\{|u|^n \mathbf{1}_{|u|\geq b} \geq t\} dt.$$

Observe that $|u|^n\mathbf{1}_{|u|\geq b} \geq t$ if and only if $|u|^n \geq t$ and $|u| \geq b$. For $t \leq b^n$ this means $|u| \geq b$ while for $t > b^n$ this means $|u| \geq t^{1/n}$. Therefore,

$$\begin{aligned}\int |u|^n \mathbf{1}_{|u|>b} d\mu(u) &= \int_0^{b^n} \mu\{u : |u| \geq b\} du + \int_0^{b^n} \mu\{u : |u| \geq t^{1/n}\} du \\ &= b^n \mu([-b^n, b^n]^c) + \int_{b^n}^\infty \mu([-t^{1/n}, t^{1/n}]^c) dt\end{aligned}$$

Using the usual Chebyshev idea, we write the bounds $\mu([-s,s]^c) \leq \alpha_{2m} s^{-2m}$ valid for any m, apply it with $m = 1$ for the first term and $m = n$ for the second term to get

$$\begin{aligned}\int |u|^n \mathbf{1}_{|u|>b} d\mu(u) &\leq \alpha_2 b^{-n} + \alpha_{2n} \int_{b^n}^\infty t^{-2} dt \\ &= (\alpha_2 + \alpha_{2n}) b^{-n}. \qquad (3)\end{aligned}$$

Moral: When in distress, remember Chebyshev's inequality or the idea behind it: $\mathbf{1}_{[b,\infty)}(t) \leq t/b$ or more generally $\mathbf{1}_{[b,\infty)}(t) \leq f(t)/f(b)$ for an increasing function f.

Step 4: Now fix a large M and N and let $-M = t_0 < t_1 < \ldots < t_N = M$ be closely spaced points (quantification later). Let n be odd so that u^n is increasing on the whole line. Therefore,

$$u^n \mathbf{1}_{(-\infty,-M]}(u) + \sum_{j=0}^{N-1} t_j^n \mathbf{1}_{(t_j,t_{j+1}]}(u) \leq u^n \leq \sum_{j=0}^{N-1} t_j^n \mathbf{1}_{(t_j,t_{j+1}]}(u) + u^n \mathbf{1}_{[M,\infty)}(u) \tag{4}$$

Integrate w.r.t μ and use (3) to get

$$-\frac{\alpha_{2n}}{M} + \sum_{j=0}^{N-1} t_j^n (F(t_{j+1}) - F(t_j)) \leq \int u^n d\mu(u) \leq \sum_{j=0}^{N-1} t_{j+1}^n (F(t_{j+1}) - F(t_j)) + \frac{\alpha_{2n}}{M} \tag{5}$$

Similarly, we want to get an inequality by applying L to (4). But $u^n\mathbf{1}_{[M,\infty)}$ is not in X, hence we bound it by u^{n+1}/M. Similarly $u^n\mathbf{1}_{(-\infty,-M]}(u) \geq -u^{n+1}/M$. Thus,

$$-\frac{1}{M}u^{n+1} + \sum_{j=0}^{N-1} t_j^n \mathbf{1}_{(t_j,t_{j+1}]}(u) \leq u^n \leq \sum_{j=0}^{N-1} t_j^n \mathbf{1}_{(t_j,t_{j+1}]}(u) + \frac{1}{M}u^{n+1}$$

Now we can apply L and use positivity to get

$$-\frac{\alpha_{n+1}}{M} + \sum_{j=0}^{N-1} t_j^n (G(t_{j+1}) - G(t_j)) \leq \alpha_n \leq \sum_{j=0}^{N-1} t_{j+1}^n (G(t_{j+1}) - G(t_j)) + \frac{\alpha_{n+1}}{M}$$

Compare this with (5). By taking M large, we can make the $1/M$ terms as small as we like. Then by taking N large, we can make sure that $t_{j+1}^n - t_j^n$ are small. By perturbing the points slightly as

[24] If (X,μ) is a measure space and $f : X \mapsto \mathbb{R}_+$ is a positive function, then $\int_X f(x) d\mu(x) = \int_0^\infty \mu\{f > t\} dt$ by a simple Fubini argument applied to the double integral $\iint_{X\times\mathbb{R}_+} \mathbf{1}_{0<t<f(x)} dt d\mu(x)$. Some people call this the "bath-tub principle". In probability it is often written in the form $\mathbf{E}[X] = \int_0^\infty \mathbf{P}\{X > t\} dt$ for a positive random variable X.

needed, we may assume that $t_j \in D$ for all j, and hence $F(t_j) = G(t_j)$. Now it is clear that α_n and $\int u^n d\mu(u)$ are sandwiched between two numbers that are very close to each other, and hence must be equal.

So far we assumed that n was odd. For even n, a very similar argument can be given if one is not too tired by now. ■

We stated the last theorem only for intervals. What about general closed sets I? Observe that if $L(p) \geq 0$ for p that is positive on I, then it is certainly the case that $L(p) \geq 0$ for p that is positive on the whole line. From the above proof, we get a measure μ supported on $\mathbb{R}$ whose moments are α_n. To argue that it is supported on I is an exercise.

Exercise 12. Suppose $[a, b] \subseteq I^c$. Argue that in the above proof, when L is extended to X, the resulting functional satisfies $G(a) = G(b)$. Deduce that $\mu(I^c) = 0$.

If you understood the above proof, the following should be easier.

Exercise 13. Prove Riesz's representation theorem for $C_c(\mathbb{R})$: If L is a positive linear functional on $C_c(\mathbb{R})$, then there exists a Borel measure μ such that $L(f) = \int f d\mu$ for all $f \in C_c(\mathbb{R})$.

Remark 14. Can we prove Riesz's representation theorem for general locally compact Hausdorff spaces? Presumably it will work, by extending L from $C_c(X)$ to $C_c(X) + W$ where W is the span of indicators of all compact sets. Then we must define the measure μ by taking $\mu(A) = \sup\{L(\mathbf{1}_K) : K \subseteq A \text{ and } K \text{ is compact}\}$. But then one must show that μ is a measure, it is outer regular etc., and that it agrees with L on $C_c(X)$. This starts looking like the lengthy proof in Rudin's *Real and complex analysis*. The proof is simpler for $X = \mathbb{R}$ (and in the moment problem above), because we assumed the existence of Lebesgue measure and that an increasing right continuous function is the CDF of a measure got by pushing forward the Lebesgue measure...

4. Measures, Sequences, Polynomials, Matrices

To be more precisely, we should have titled this section as "Measures on the line having all moments, positive semi-definite sequences, orthogonal polynomial sequences and Jacobi matrices". All these objects are intimately connected to each other and to the moment problem. This will also lead to the resolution of the uniqueness part of the moment problem, but we may not completely discuss it. Let us introduce all the four objects in the title.

(1) Measures. By this, in this section we shall mean positive Borel measures on the line whose moments are all finite. It is convenient to consider two cases separately. *Case 1:* The measure is has infinite support, *Case 2:* The measure is supported on finitely many points, i.e., $\mu = p_1\delta_{\lambda_1} + \ldots + p_n\delta_{\lambda_n}$, where λ_i are distinct real numbers and p_i are strictly positive numbers.

(2) Positive semi-definite sequences. By this we mean a sequence $\alpha = (\alpha_0, \alpha_1, \ldots)$ such that the infinite matrix $H_\alpha = (\alpha_{i+j})_{i,j\geq 0}$ is a positive semi-definite matrix. This just means that for any $n \geq 1$, and any real numbers $c_0, \ldots, c_n$,

$$\sum_{i,j=0}^{n} c_i c_j \alpha_{i+j} \geq 0$$

Case 1: The sequence is positive definite. That is, equality holds above if and only if all c_is vanish. *Case 2:* The sequence is positive semi-definite but not positive definite. There is a smallest n for which equality holds above for some c_is, not all zero.

(3) Orthogonal polynomial sequence. By this we mean a sequence of polynomial $\varphi_0, \varphi_1, \ldots$ such that -

(a) The degree of φ_j is exactly j for every $j \geq 0$.

(b) If φ_j are declared to be an orthonormal set in $\mathcal{P}$, then in the resulting inner product space, the multiplication operator $M : \mathcal{P} \mapsto \mathcal{P}$ defined by $Mf(x) = xf(x)$ is symmetric: $\langle Mf, g\rangle = \langle f, Mg\rangle$ for all $f, g \in \mathcal{P}$.

For example, $\varphi_j(x) = x^j$ is not a valid choice for an orthogonal polynomial sequence, because $\langle M\varphi_1, \varphi_2\rangle = 1$ while $\langle \varphi_1, M\varphi_2\rangle = 0$.

What we describe so far is *Case 1*. *Case 2* is when we have a finite sequence $\varphi_0, \ldots, \varphi_{n-1}$ such that ... (details later)

(4) Jacobi matrix. A tridiagonal matrix is a finite or infinite matrix whose (i, j) entry is zero unless $|j-i| \leq 1$ (only the main diagonal and the the diagonals immediately above and below it, can contain non-zero entries). A Jacobi matrix is a tridiagonal matrix that is symmetric and has strictly positive entries on the super-diagonal (hence also the sub-diagonal). The main diagonal entries will be labelled $a_0, a_1, \ldots$ while the super-diagonal entries will be labelled $b_0, b_1, \ldots$. *Case 1:* Infinite Jacobi matrix $T = T(a, b) = (t_{i,j})_{i,j\geq 0}$ such that $t_{i,i} = a_i$ and $t_{i,i+1} = t_{i+1,i} = b_i$ for $i \geq 0$. *Case 2:* Finite Jacobi matrix $T_{n\times n}$ whose main diagonal has $a_0, \ldots, a_{n-1}$ and super-diagonal has $b_0, \ldots, b_{n-1}$.

What we shall see is that these objects are very closely linked and almost (but not quite!) in one-one correspondence with each other. The objects in *Case 1* are related to each other and the objects in *Case 2* are related to each other. Rather than carrying the two cases all the time, let us first describe the connections in the first case. Later we shall discuss the second case.

5. Measures, Sequences, Polynomials, Matrices: Case 1

5.1. **Measure to sequence.** Given a measure μ whose support is not finite, let $\alpha_n = \int x^n d\mu(x)$ be the nth moment of μ. We claim that the moment sequence $\alpha = (\alpha_0, \alpha_1, \ldots)$ is positive definite. This is because

$$\sum_{i,j=0}^{m} c_i c_j \alpha_{i+j} = \sum_{i,j=0}^{m} c_i c_j \int x^{i+j} d\mu(x) = \int \left(\sum_{i=1}^{m} c_i x^i\right)^2 d\mu(x) \geq 0.$$

Equality holds if and only if the polynomial $\sum_{i=0}^{n} c_i x^i$ vanishes on the support of μ. As the latter is an infinite set, this forces $c_i = 0$ for all i. Thus, α is positive definite.

An alternate way to say the same thing is that the matrix $H_\alpha = (\alpha_{i+j})_{i,j\geq 0}$ is positive definite, meaning that all finite principal submatrices of H_α have strictly positive determinant.

5.2. **Sequence to polynomial sequence.** Given a positive semi-definite α, we can define an inner product on $\mathcal{P}$ by defining $\langle x^i, x^j\rangle = \alpha_{i+j}$ for $i, j \geq 0$ and extending by linearity. That is

$$\left\langle \sum_{i=0}^{n} c_i x^i, \sum_{j=0}^{m} d_j x^j \right\rangle = \sum_{i=0}^{n}\sum_{j=0}^{m} c_i d_j \alpha_{i+j}.$$

The bilinearity and symmetry are clear while the positive definiteness of α ensures that $\langle p, p\rangle > 0$ for any $p \neq 0$.

Apply Gram-Shmidt process to $x^0, x^1, x^2, \ldots$ (in that order) to get $\varphi_0, \varphi_1, \ldots$, an orthonormal set that spans the whole space $\mathcal{P}$. It is also clear that φ_j is a polynomial of degree j and that it has positive leading coefficient. There is another property of this sequence

Observation: Let $M : \mathcal{P} \mapsto \mathcal{P}$ be defined by $(Mp)(x) = xp(x)$. Then,

$$\langle Mx^i, x^j\rangle = \langle x^{i+1}, x^j\rangle = \alpha_{i+1+j}, \quad \langle x^i, Mx^j\rangle = \langle x^i, x^{j+1}\rangle = \alpha_{i+j+1}$$

showing that M is symmetric: $\langle Mp, q\rangle = \langle p, Mq\rangle$ for all $p, q \in \mathcal{P}$.

By an *orthogonal polynomial sequence* we mean a sequence of polynomials $\varphi_0, \varphi_1, \varphi_2, \ldots$ such that φ_j has degree j, has positive leading coefficient, and such that if an inner product on $\mathcal{P}$ is defined by declaring φ_js to be orthonormal, then the multiplication operator is symmetric.

Remark 15. (Rameez) The symmetry of M can be equivalently stated as the condition that the Gram matrix of $x^0, x^1, x^2, \ldots$ is a Hankel matrix. This shows that $x^0, x^1, x^2, \ldots$ is not an orthogonal polynomial sequence (because the identity matrix is not Hankel!).

5.3. **Polynomial sequence to Jacobi matrix.** Let $\varphi_0, \varphi_1, \ldots$ be an orthogonal polynomial sequence. Let $\langle \star, \star\rangle$ denote the inner product on $\mathcal{P}$ got by declaring $\langle \varphi_j, \varphi_k\rangle = \delta_{j,k}$ and extending by linearity (possible since the span of $\{\varphi_j\}$ is all of $\mathcal{P}$). Two simple observations: $\langle Mp, q\rangle = \langle p, Mq\rangle$ (by definition of orthogonal polynomials) and $\langle \varphi_k, p\rangle = 0$ if p has degree less than k.

For $k \geq 0$, $M\varphi_k$ has degree $k+1$ and hence there is a unique way to write it as $M\varphi_k = \sum_{j=0}^{k+1} c_{k,j}\varphi_j$. For $j < k-1$, by the symmetry of M, we see that $c_{k,j} = \langle \varphi_k, M\varphi_j\rangle = 0$ since $M\varphi_j$ has degree less than than k. Further,

$$c_{k,k+1} = \langle M\varphi_k, \varphi_{k+1}\rangle = \langle \varphi_k, M\varphi_{k+1}\rangle = c_{k+1,k}.$$

Writing $a_k = c_{k,k}$ and $b_k = c_{k,k+1}$, we see that (with the convention that $b_{-1} = 0$)

$$b_{k-1}\varphi_{k-1} + a_k\varphi_k + b_k\varphi_{k+1} = M\varphi_k.$$

It will be convenient to collect the coefficients a_k, b_ks as an infinite tridiagonal matrix

$$T = T(a;b) = \begin{bmatrix} a_0 & b_0 & 0 & \dots & \dots \\ b_0 & a_1 & b_1 & \dots & \dots \\ 0 & b_1 & a_2 & b_2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}.$$

To be a Jacobi matrix, we also require $b_k > 0$ for all k. To see this, let A_k be the leading coefficient of φ_k and observe that $\varphi_{k+1} - \frac{A_{k+1}}{A_k} M\varphi_k$ has degree less than k and hence is orthogonal to φ_{k+1}. Thus,

$$b_k = \langle M\varphi_k, \varphi_{k+1} \rangle = \frac{A_{k+1}}{A_k} > 0.$$

Remark 16. In terms of the Jacobi matrix, the three term recurrence can be written in the matrix form

$$T\varphi_\bullet(x) = x\varphi_\bullet(x)$$

where $\varphi_\bullet(x) = (\varphi_0(x), \varphi_1(x), \varphi_2(x), \dots)^t$. Formally, this looks like an eigenvalue equation. The appearance is more than skin deep.

5.4. **From Jacobi matrix to orthogonal polynomial sequence.** Let $T = T(a,b)$ be a finite or infinite tridiagonal matrix with $T_{i,i} = a_i \in \mathbb{R}$ and $T_{i,i+1} = b_i > 0$. We want to recover the orthogonal polynomial sequence. The short answer is that we solve the "eigenvalue equation" $T\mathbf{v} = \lambda\mathbf{v}$ for any $\lambda \in \mathbb{R}$ and write the eigenvector as $\mathbf{v} = (\varphi_0(\lambda), \varphi_1(\lambda), \dots)^t$. These φ_ks are the orthogonal polynomials.

Let us examine this in more detail. Fix any $\lambda \in \mathbb{R}$, set $v_0 = 1$ and recursively solving for $v_1, v_2, \dots$ from the equations

$$\begin{aligned} a_0v_0 + b_0v_1 &= \lambda v_0 \\ b_0v_0 + a_1v_1 + b_1v_2 &= \lambda v_1 \\ b_1v_1 + a_2v_2 + b_2v_3 &= \lambda v_2 \quad \dots \quad \dots \end{aligned}$$

As $b_k > 0$ for all k, this is possible and we get a vector $\mathbf{v} = (v_0, v_1, \dots)$ that satisfies $T\mathbf{v} = \lambda\mathbf{v}$. Now let us change notation and show the dependence on λ by writing v_k as $\varphi_k(\lambda)$, starting with $\varphi_0(\lambda) = 1$. It is clear from the recursions that φ_k is a polynomial of degree k and that it has positive leading coefficient. We must check one last point: If φ_k are declared to be orthonormal, then the multiplication M must be symmetric. That is indeed the case, as we can check that $\langle M\varphi_k, \varphi_\ell \rangle = \langle \varphi_k, M\varphi_\ell \rangle$ from the recursions

$$M\varphi_k = b_{k-1}\varphi_{k-1} + a_k\varphi_k + b_k\varphi_{k+1}.$$

Important observation: We want to say that this mapping from Jacobi matrices to OP-sequences and the mapping of the previous section from OP-sequences to Jacobi matrices are inverses of each other. This is almost correct in that we recover the orthogonal polynomial sequence up to an overall constant factor. Indeed, in the recovery, we always get $\varphi_0 = 1$. This is easily seen by staring for a minute at the three-term recurrence.

It would have been cleaner if we had assumed all measures to be probability measures, all positive definite sequences to have $\alpha_0 = 1$, all OP sequences to have $\varphi_0 = 1$. Then the above mapping would have been exactly the inverse of the one from OP sequences to Jacobi matrices. *Henceforth, let us adopt this convention.*

5.5. **From orthogonal polynomials to positive definite sequence.** Given an orthogonal polynomial sequence $\varphi_0, \varphi_1, \dots$ and the associated inner product, we construct a positive definite sequence as follows. There is a unique way to write $x^k = \sum_{j=0}^{k} c_{k,j}\varphi_j$ from which we get

$$\langle x^k, x^\ell\rangle = \sum_{j=0}^{k\wedge\ell} c_{k,j}c_{\ell,j}.$$

Since we already know that M is symmetric in this inner product, it follows that the above quantity must depend only on $k+\ell$. Denote this number by $\alpha_{k+\ell}$. This is a positive definite sequence because H_α is the Gram matrix of $x^0, x^1, x^2, \dots$ and these are linearly independent.

It is also a easy to see that this mapping is the inverse of the mapping that we gave earlier from positive definite sequences to orthogonal polynomial sequences.

It may look a little unsatisfactory that the mapping given here is not explicit. It can be made explicit. Fix $k \geq 0$ and write $\alpha_k = \langle x^k, x^0\rangle = c_{k,0}$, the constant term in φ_k. This can be put in a more interesting form in terms of the Jacobi matrix (recall that we already know how to move between OP-sequences and Jacobi matrices).

5.6. **From Jacobi matrix to positive definite sequence.** Let $T = T(a,b)$ be a Jacobi matrix. Define $\beta_k = \langle T^k e_0, e_0\rangle$ for $k \geq 0$. We claim that this is a positive semi-definite sequence and that this is the inverse of the mapping we have see from positive semi-definite sequence to tridiagonal matrices (via orthogonal polynomial sequence and three term recurrence).

As we have seen how to recover orthogonal polynomials from T, let us write

$$T\varphi_\bullet(x) = x\varphi_\bullet(x)$$

where $\varphi_\bullet(x) = (\varphi_0(x), \varphi_1(x), \varphi_2(x), \dots)^t$. Therefore, $T^k\varphi_\bullet(x) = x^k\varphi_\bullet(x)$ or in terms of the coordinate vectors $e_0, e_1, \dots$

$$\sum_{j=0}^{\infty} \varphi_j(x) T^k e_j = \sum_{j=0}^{\infty} x^k \varphi_j(x) e_j.$$

Take inner product with e_0 (this is inner product in ℓ^2) to get

$$\sum_{j=0}^{\infty} \varphi_j(x)\langle T^k e_j, e_0\rangle = x^k.$$

This gives the expansion of x^k in terms of the orthogonal polynomials that we needed above (it should not worry you that the sum here is infinite, indeed $\langle T^k e_j, e_0\rangle = 0$ as can be seen from the tridiagonal structure of T). In particular $c_{k,0} = \langle T^k e_0, e_0 \rangle$. Combining with the previous observation of how to recover the α_ks from $c_{k,0}$, this shows that $T \mapsto (\beta_0, \beta_1, \ldots)$ mapping Jacobi matrices into positive definite sequence is the inverse of the mapping in the other direction that we have seen earlier (going through OP-sequences).

Remark 17. A better way as pointed out by Sayantan Khan in class. The mapping $\varphi_j \leftrightarrow e_j$, $j \geq 0$, is an isomorphism with $\mathcal{P}$ (with $\{\varphi_j\}$ as ONB) and $V = \text{span}\{e_0, e_1, \ldots\}$ where e_j is the vector with 1 at the jth place and 0s elsewhere. Under this isomorphism, $M : \mathcal{P} \mapsto \mathcal{P}$ becomes $T : V \mapsto V$. As we saw earlier, $\alpha_k = \langle x^k, x^0 \rangle = \langle M^k \varphi_0, \varphi_0 \rangle$ which, by the isomorphism, equals $\langle T^k e_0, e_0 \rangle$.

5.7. **The picture so far.**

$$\text{Measure} \to \text{PD sequence} \leftrightarrows \text{OP sequence} \leftrightarrows \text{Jacobi matrix}$$

The key question was whether a positive definite sequence is the moment sequence of a unique measure. We have not touched that question but introduced two other objects that are in one-one correspondence with positive definite sequences. We shall return to this question after talking about some nice consequences of the rich interactions between these objects in the next two sections.

5.8. **Exercises.** In these exercises, the relationship between the positive definite sequences, orthogonal polynomials and Jacobi matrices is further strengthened.

Exercise 18. Let α be a positive definite sequence. Let $D_m = \det\left(\alpha_{i+j}\right)_{0 \leq i,j \leq m-1}$. Show that the corresponding orthogonal polynomials are given by $\varphi_0(x) = 1$ and for $m \geq 1$,

$$\varphi_m(x) = \frac{(-1)^{m-1}}{\sqrt{D_{m-1} D_m}} \det \begin{bmatrix} 1 & x & \ldots & x^m \\ \alpha_0 & \alpha_1 & \ldots & \alpha_m \\ \vdots & \vdots & \vdots & \vdots \\ \alpha_{m-1} & \alpha_m & \ldots & \alpha_{2m-1} \end{bmatrix}$$

Exercise 19. Let $T = T(a, b)$ be the Jacobi matrix corresponding to the positive definite sequence α. Let D_n be the determinant of $(\alpha_{i+j})_{0 \leq i,j \leq n-1}$. Show that

$$b_k = \frac{\sqrt{D_{k-1} D_{k+1}}}{D_k}$$

for $k \geq 1$ and $b_0 = \frac{\sqrt{D_1}}{D_0}$. If $\alpha_n = 0$ for all odd n (for eg., if it is the moment sequence of a symmetric measure), show that $a_n = 0$ for all n. [**Remark:** There is also a formula for a_ns in general, but we skip it for now]

Exercise 20. Let the OP sequence $\varphi_0, \varphi_1, \ldots$ correspond to the Jacobi matrix $T = T(a, b)$. If T_n is the top $n \times n$ sub-matrix of T, show that φ_n is (up to a constant) the characteristic polynomial of T_n. Deduce that,

(1) The roots of φ_n are all real and distinct.

(2) The roots of φ_n and φ_{n-1} interlace.

6. Quadrature formulas

Let μ be a measure on the line with all moments finite. Assume that μ is not supported on finitely many points. Fix $n \geq 1$. We seek n distinct points $\lambda_1, \ldots, \lambda_n$ and poisitive weights $w_1, \ldots, w_n$ such that

$$\int Q(x) d\mu(x) = \sum_{k=1}^{n} w_k Q(\lambda_k)$$

for as many polynomials Q as possible. Since we have a choice of $2n$ parameters for the points and weights, we may expect that this can be done for all polynomials of degree $2n-1$ or less (it has $2n$ coefficients).

Why care? It has to do with numerical integration. Once we fix n and choose λ_is and w_is, given any $f : \mathbb{R} \mapsto \mathbb{R}$, we numerically compute its integral with respect to μ by

$$\int f(x) d\mu(x) \approx \sum_{k=1}^{n} w_k f(\lambda_k).$$

This gives the exact answer for polynomials of degree up to $2n-1$. Hence, if f is nice enough that it is well approximated by its Taylor expansion to order $2n-1$, then the above approximation gives a reasonably close answer to $\int f d\mu$.

How to find the points and weights? Note that what we are asking for is a measure $\mu_n = \sum_{k=1}^{n} w_k \delta_{\lambda_k}$ whose first $2n-1$ moments agree with those of μ.

Assume that μ is a probability measure, without loss of generality. From μ, we go to the infinite tridiagonal matrix $T = T(a, b)$ (via moments, orthogonal polynomials and the three-term recurrence). Let T_n be the top $n \times n$ principal submatrix of T. Let μ_n be the measure corresponding to T_n, i.e., the spectral measure of T_n at the vector e_0. Recall that this is given by

$$\mu_n = \sum_{k=1}^{n} w_k \delta_{\lambda_k}$$

where λ_k are the eigenvalues of T_n and $w_k = Q_{1,k}^2$, where $T_n = Q \Lambda Q^t$ is the spectral decomposition of T_n, with $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ and Q an orthogonal matrix.

Recall that the moment sequence α can be recovered from the tridiagonal matrix μ by the equations $\alpha_k = \langle T^k e_0, e_0 \rangle$. This is just the $(0,0)$ entry of T^k, which can also be written as

$$\sum_{i_1, \ldots, i_{k-1} \geq 0} T_{0,i_1} T_{i_1,i_2} \ldots T_{i_{k-1},0}.$$

Since T is tridiagonal, the non-zero terms must have $i_1, \ldots, i_{k-1} \leq \lfloor k/2 \rfloor$. Hence,

$$\langle T^k e_0, e_0 \rangle = \langle (T_n)^k e_0, e_0 \rangle$$

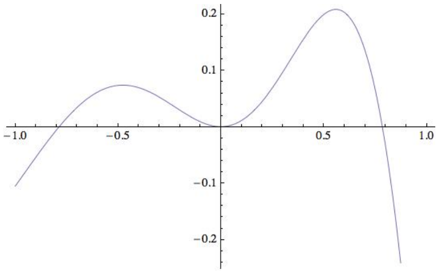

FIGURE 9. Plot of the function $e^x \cos(2x)\log(1+x^2)$

for $k \le n-1$. This shows that the first $2n-1$ moments of μ_n and μ are identical.

Remark 21. It is also possible to express the points and weights in terms of the orthogonal polynomials for μ. Indeed, $\lambda_1, \dots, \lambda_n$ are the roots of φ_n, and $(\varphi_0(\lambda_k) \dots \varphi_{n-1}(\lambda_k))^t$ is an eigenvector corresponding to the eigenvalue λ_k. After normalizing, this becomes the kth column of R. Hence,

$$w_k = \frac{\varphi_0(\lambda_k)^2}{\sum_{j=0}^{n-1} \varphi_j(\lambda_k)^2} = \frac{1}{\sum_{j=0}^{n-1} \varphi_j(\lambda_k)^2}.$$

An example - Lebesgue measure on $[-1,1]$**:** If μ is the Lebesgue measure on $[-1,1]$, then the corresponding orthogonal polynomials are called Legendre polynomials. There are explicit formulas to express them. The zeros of the Legendre polynomial can be computed and so can the weights. This gives us a way to numerically integrate functions over $[-1,1]$. Just to illustrate, here is an example:

If we want four points, the points and weights are given by (computations on Mathematica)

$$\lambda = (-0.861136, -0.339981, 0.339981, 0.861136), \quad w = (0.347855, 0.652145, 0.652145, 0.347855).$$

The function $f(x) = e^x \cos(2x)\log(1+x^2)$ (chosen without fear or favour) has integral 0.0350451 and the numerical approximation using the above points and weights gives 0.036205. With five points, it improves to 0.0348706 and with 10 points, the agreement is up to 7 decimal places! In contrast, with equispaced points and equal weights, the approximations are 0.0255956 for 100 points, 0.0327198 for 1000 points and 0.0349526 for 10000 points. In general, what is the error like? Let $f \in C^{(n)}$ (on an open set containing $[-1,1]$) and write $f(x) = Q_n(x) + R_n(x)$, where Q_n is the $2n-1$ order Taylor expansion of f. The remainder term R_n can be estimated by

$$\sup_{x \in [-1,1]} |R_n(x)| \le \frac{1}{(2n)!} \|f^{(2n)}\|_{\sup[-1,1]}.$$

Since $\sum_{j=1}^n w_j Q_n(\lambda_j) = \int_{-1}^1 Q_n(x)dx$, we get

$$\begin{aligned} \Big|\int_{-1}^1 f(x)dx - \sum_{k=1}^n w_k f(\lambda_k)\Big| &\le \int_{-1}^1 |R_n(x)|dx + \frac{1}{n}\sum_{k=1}^n w_k R_n(\lambda_k) \\ &\le \frac{2}{(2n)!}\|f^{(2n)}\|_{\sup[-1,1]} \end{aligned}$$

For instance, if the derivatives are uniformly bounded in $[-1,1]$ (or grow at most exponentially etc.) then the error term is $O(e^{-cn\log n})$. In contrast, for n equispaced points, the error goes down like $O(1/n)$ and for n randomly chosen points the error goes down like $1/\sqrt{n}$.

Similarly, one uses zeros of Chebyshev polynomials, Hermite polynomials (OPs for Gaussian measure), Laguerre polynomials (OPs for $e^{-x}dx$ on $\mathbb{R}_+$), etc., to integrate against $\frac{1}{\sqrt{1-x^2}}$, e^{-x^2}, e^{-x}, respectively. They carry names such as Chebyshev quadrature, Gaussian quadrature etc.

This may be a good occasion to say something explicit about orthogonal polynomials for special measures. The few examples are, the uniform measure (Legendre polynomials), the Gaussian measure (Hermite polynomials), Exponential measure (Laguerre polynomials), arcsine measure (Chebyshev polynomials). The uniform and arcsine fall into the family of Beta measures (whose orthogonal polynomials are called Jacobi polynomials) and the exponential is part of the Gamma family of distributions.

Exercise 22. Define $P_n(x) = \frac{d^n}{dx^n}(1-x^2)^{2n}$. Show that P_n are orthogonal on $[-1,1]$ with respect to Lebesgue measure. Find c_n so that c_nP_n become orthonormal. These are the Legendre polynomials.

The expression for Legendre polynomials in the exercise is called Rodrigues' formula. Similarly, one can show that

$$H_n(x) := e^{x^2/2}\frac{d^n}{dx^n}e^{-x^2/2}$$

are orthogonal with respect to the standard Gaussian measure on $\mathbb{R}$.

7. Another proof that positive semi-definite sequences are moment sequences

Let α be a positive semi-definite sequence with $\alpha_0 = 1$ (without loss of generality). We want to show that there is a measure μ such that $\alpha_n = \int x^n d\mu(x)$ for all n.

The idea of this proof is to solve a sequence of problems approximating our problem, and then extract a limit solution that will solve the actual problem. At this level of generality, this is a very repeatable (and natural) idea. In addition, it will illustrate one of the theorems we learned in functional analysis class.

If α is positive semi-definite but not positive definite, we shall see a simple proof that it is the moment sequence of a unique measure which in fact has finite support. Hence, let us assume that α is positive definite below.

Step-1: For any n, there exists a measure μ_n such that $\alpha_k = \int x^k d\mu_n(x)$ for $0 \le k \le n-1$.

We saw this in the previous section. From α, construct the OP sequence and then the Jacobi matrix T. Let μ_n be the spectral measure at e_0 of T_n, where T_n is the top $n \times n$ principal submatrix of T, that is, $\mu_n = \sum_{j=1}^n Q_{1,j}^2 \delta_{\lambda_j}$ where $T_n = Q\Lambda Q^t$ is the spectral decomposition of T_n with $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$. Then, μ_n has the first $2n-1$ moments equal to those of μ.

Step-2: There is a subsequence n_k such that μ_{n_k} converges weakly to a probability measure μ.

This is a direct consequence of Helly's theorem,[25] since $\mu_n(\mathbb{R}) = \alpha_0$ for all n.

Step-3: We claim that μ has the moment sequence α.

Fix an even number $2k$ and write

$$\begin{aligned}\int x^{2k} d\mu_n(x) &= \int_0^\infty \mu_n\{x : x^{2k} > t\} dt \\ &= \int_0^\infty \mu_n(-\infty, -t^{1/2k}) dt + \int_0^\infty \mu_n(t^{1/2k}, \infty) dt.\end{aligned}$$

Consider the first integral, take $n = n_j$ and let $j \to \infty$. For a.e. x (according to Lebesgue measure), the integrand converges to $\mu(-\infty, -t^{1/2k})$. If we can justify the hypothesis of DCT, it follows that the integral converges to $\int_0^\infty \mu(-\infty, -t^{1/2k}) dt$. Similarly for the second integral. Taking the sum, we get $\int x^{2k} d\mu(x)$, showing that the even moments of μ_{n_j} converge to those of μ. But for every k, the kth moment of μ_{n_j} is α_k for large enough j. Therefore, the $2k$th moment of μ is α_{2k}. Argue similarly for odd moments.

To justify DCT, use the bounds (Chebyshev again!)

$$\mu_n(-\infty, -t^{1/2k}) \le \frac{1}{t^2} \int x^{4k} d\mu_n(x) = \alpha_{4k} t^{-2},$$

the last inequality being for large enough n. Of course we also have the bound α_0 for the left side, which we use for $t < 1$. Thus, the integrand is dominated by $\alpha_0 + \alpha_{4k} t^{-2} \mathbf{1}_{t \ge 1}$ which is integrable. This completes the proof.

Remark 23. We shall have occasion to use Helly's theorem again. It is a compactness criterion for measures on the line (with the topology of weak convergence). It is instructive to compare it and its proof with other compactness theorems that you have seen, like the Arzela-Ascoli theorem or Montel's theorem in complex analysis.

Helly's theorem can be seen as a special case of Banach-Aloglu theorem as follows: The space $C_0(\mathbb{R})$ of continuous functions vanishing at infinity is a Banach space under the sup-norm, and its dual is the space of all signed measures that are Radon. The weak-* topology on the dual is precisely the topology of weak convergence. Thus, a sequence $\{\mu_n\}$ as in Helly's theorem is contained in a ball in $(C_0(\mathbb{R}))^*$ and hence pre-compact.

In general, compactness does not imply sequential compactness (note that the weak-* topology is not metrizable in general), but the separability of $C_0(\mathbb{R})$ can be used to show the sequential

[25] *Helly's theorem:* If μ_n is a sequence of Borel measures on $\mathbb{R}$ such that $\mu_n(\mathbb{R}) \le A$ for some A for all n, then there is a subsequence n_k and a measure μ such that $\mu_n[a, b] \to \mu[a, b]$ for all a, b such that $\mu\{a, b\} = 0$.

Proof: For each x, the sequence $\mu_n(-\infty, x]$ has a subsequential limit. Enumerate rationals in a sequence, take subsequences of subsequences etc., and use a diagonal argument to get a single subsequence along which $G(x) := \lim_{k\to\infty} \mu_{n_k}(-\infty, x]$ exists for all $x \in \mathbb{Q}$. Now define $F(x) = \inf\{G(y) : y > x\}$, an increasing, right-continuous, bounded function. Let μ be the measure such that $\mu(-\infty, x] = F(x)$ for all x. Check that $\mu_{n_k}[a, b] \to \mu[a, b]$ at least if $\mu\{a\} = \mu\{b\} = 0$.

compactness as required in Helly's theorem. The best way to understand this is to assume that a Banach space is separable and prove Banach-Alaoglu theorem for its dual by imitating the proof of Helly's theorem (take a countable dense set in X...).

8. Some special orthogonal polynomials

We have talked about general measures and not actually worked out any examples. Here we present a few.

Gaussian measure: Let $d\mu(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}dx$ on the line. The odd moments are zero while the even moments are

$$\begin{aligned}\alpha_{2n} &= \frac{2}{\sqrt{2\pi}}\int_0^\infty x^{2n}e^{-x^2/2}dx = \frac{2}{\sqrt{2\pi}}\int_0^\infty (2t)^n e^{-t}\frac{dt}{\sqrt{2t}} = \frac{2^n}{\sqrt{\pi}}\Gamma(n+\frac{1}{2}) \\ &= 2^n(n-\frac{1}{2})(n-1-\frac{1}{2})\dots(1-\frac{1}{2}) = (2n-1)\times(2n-3)\times\dots\times 3\times 1.\end{aligned}$$

This has the nice interpretation as the number of matchings of the set $\{1,2,\dots,2n\}$ into n pairs[26].

I do not know how to derive the orthogonal polynomials using Gram-Schmidt or the determinant formula that we gave in an earlier exercise. We simply define for $n \geq 0$,

$$H_n(x) = (-1)^n e^{x^2/2}\frac{d^n}{dx^n}e^{-x^2/2}$$

which is clearly a polynomial of degree n. Also,

$$\begin{aligned}\int H_n H_m d\mu &= \frac{(-1)^{m+n}}{\sqrt{2\pi}}\int_{-\infty}^\infty \left[\frac{d^n}{dx^n}e^{-x^2/2}\right]H_m(x)dx \\ &= \frac{(-1)^{m+2n}}{\sqrt{2\pi}}\int_{-\infty}^\infty e^{-x^2/2}\left[\frac{d^n}{dx^n}H_m(x)\right]dx\end{aligned}$$

by integrating by parts n times (the boundary terms vanish because of the rapid decay of $e^{-x^2/2}$). If $m < n$, the integrand is zero (since H_m has degree m) and if $m = n$, we observe that $H_n(x) = x^n + \dots$ to see that $\frac{d^n}{dx^n}H_n(x) = n!$. The rest of the integral is one, and we arrive at $\int H_n^2(x)d\mu(x) = n!$, from which we get the OPs as

$$\varphi_n(x) = \frac{1}{\sqrt{n!}}H_n(x).$$

These form an ONB for $L^2(\mathbb{R},\mu)$. As a corollary, $\frac{1}{\sqrt[4]{2\pi}\sqrt{n!}}\varphi_n(x)e^{-x^2/4}$, $n \geq 0$, form an ONB for $L^2(\mathbb{R})$. Completeness may require an argument.

The Jacobi matrix corresponding to this is given by $T = T(a,b)$ where

$$a_n = \int x\varphi_n(x)^2 d\mu(x), \quad b_n = \int x\varphi_n(x)\varphi_{n+1}(x)d\mu(x).$$

[26] This is not mere numerology. If $(X_1,\dots,X_{2n})$ are jointly Gaussian with zero means and covariance $\mathbf{E}[X_iX_j] = \sigma_{i,j}$, then $\mathbf{E}[X_1\dots X_{2n}]$ is equal to $\sum_M w(M)$, where the sum is over all matchings of $\{1,2\dots,2n\}$ and the weight of a matching $M = \{\{i_1,j_1\},\dots,\{i_n,j_n\}\}$ is given by $w(M) = \sigma_{i_1,j_1}\sigma_{i_2j_2}\dots\sigma_{i_nj_n}$. This is sometimes called *Wick formula* or *Feynman diagram formula*.

It is easy to see that H_n (and hence φ_n) is even or odd according as n is even or odd. Hence $a_n = 0$, being the integral of an odd function. Further, if we write $x\varphi_n(x) = C_n\varphi_{n+1}(x) +$ [lower order terms], then it is clear that $b_n = C_n$. But it is easy to work out that

$$C_n = \frac{[x^n]\varphi_n(x)}{[x^{n+1}]\varphi_{n+1}(x)} = \frac{1/\sqrt{n!}}{1/\sqrt{(n+1)!}} = \sqrt{n+1}.$$

Thus, the Jacobi matrix for this measure is

$$T = \begin{bmatrix} 0 & \sqrt{1} & 0 & \dots & \dots \\ \sqrt{1} & 0 & \sqrt{2} & \dots & \dots \\ 0 & \sqrt{2} & 0 & \sqrt{3} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

Uniform measure: Let μ be the uniform probability measure on $[-1,1]$. The moments are

$$\alpha_n = \begin{cases} 0 & \text{if } n \text{ is odd,} \\ \frac{1}{n+1} & \text{if } n \text{ is even.} \end{cases}$$

Observe that the Hankel matrix H_α is very similar to the Hilbert matrix. Again, rather than working out the orthogonal polynomials, we simply present the answer. Define the *Legendre polynomials*

$$P_n(x) := \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n, \text{ for } n \ge 0.$$

Clearly P_n is a polynomial of degree n and $[x^n]P_n(x) = \frac{(2n)!}{2^n(n!)^2}$. We leave it as an exercise to check that

$$\int_{-1}^{1} P_n(x)P_m(x)\frac{dx}{2} = \frac{1}{2n+1}\delta_{n,m}.$$

Thus, $\varphi_n(x) = \sqrt{2n+1}P_n(x)$, $n \ge 0$, are the orthogonal polynomials.

To get the Jacobi matrix, we compute a_n and b_n as before. Again, φ_n are alternately odd and even, hence $a_n = 0$. Further,

$$b_n = \int_{-1}^{n} x\varphi_n(x)\varphi_{n+1}(x)\frac{dx}{2} = \frac{[x^n]\varphi_n(x)}{[x^{n+1}]\varphi_{n+1}(x)} = \frac{n+1}{\sqrt{2n+1}\sqrt{2n+3}}.$$

Exercise 24. Let $d\mu(x) = e^{-x}dx$ on $\mathbb{R}_+$. Find the moments, orthogonal polynomials and the Jacobi matrix corresponding to this measure.

Hint: Consider the Laguerre polynomials

$$L_n(x) = \frac{1}{n!}e^x\frac{d^n}{dx^n}[x^n e^{-x}].$$

Other special orthogonal polynomials: In a similar fashion, it is possible to obtain explicitly the orthogonal polynomials and the Jacobi matrix for the Beta family of distributions (that includes the uniform measure and also the arcsine measure) and the Gamma family of distributions (a special case being the exponential measure $e^{-x}dx$ on $\mathbb{R}_+$). The corresponding orthogonal polynomials are called Jacobi polynomials and generalized Laguerre polynomials. In addition to the general properties shared by all orthogonal polynomials, these special ones also satisfy differential equations, recursions involving the polynomials and the derivatives etc. They arise in a variety of problems. For example, the Legendre polynomials arise naturally in the representation theory of the orthogonal group.

9. The uniqueness question: some sufficient conditions

Now suppose we have a positive definite sequence α. We also have the associated OP sequence $\varphi_0, \varphi_1, \dots$ and the Jacobi matrix $T = T(a,b)$. The question is whether there is a unique measure having moment sequence α? If not, what are all the measures that have this moment sequence?

First we give examples to show that uniqueness need not always hold. A standard example is the measure $d\mu(t) = f(t)dt$ where

$$f(t) = \frac{1}{\sqrt{2\pi}t} e^{-\frac{1}{2}(\log t)^2} dt \quad \text{for } t > 0.$$

In probabilistic language, if X has $N(0,1)$ distribution, then e^X has density f. The moments of μ are given by

$$\alpha_n = \int_0^\infty t^n f(t)dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{nx} e^{-\frac{1}{2}x^2} dx = e^{\frac{1}{2}n^2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{1}{2}(x-n)^2} dx = e^{\frac{1}{2}n^2}.$$

To get some other measures, consider the sum

$$\sum_{k\in\mathbb{Z}} e^{-\frac{1}{2}(k-n)^2} = e^{-\frac{1}{2}n^2} \sum_{k\in\mathbb{Z}} e^{-\frac{1}{2}k^2} e^{kn}.$$

The left hand side does not depend on n (index the sum by $k-n$ instead of k)! Denoting it as Z and $p_k = e^{-\frac{1}{2}k^2}/Z$, we see that

$$\sum_{k\in\mathbb{Z}} p_k e^{kn} = \alpha_n.$$

Thus, the discrete measure $\nu = \sum_{k\in\mathbb{Z}} p_k \delta_{e^k}$ and μ have the same moment sequence. Instead of summing k over integers, if we sum over $\mathbb{Z} + t$ for some $t \in (0,1)$, we would get other measures with the same moment sequence.

Here is another kind of example. Observe that

$$\begin{aligned}\int_0 t^n f(t) \sin(2\pi \log t)dt &= \int_{-\infty}^\infty e^{nx} e^{-\frac{1}{2}x^2} \sin(2\pi x)dx \\ &= e^{\frac{1}{2}n^2} \int_{-\infty}^\infty e^{-\frac{1}{2}(x-n)^2} \sin(2\pi x)dx \\ &= e^{\frac{1}{2}n^2} \int_{-\infty}^\infty e^{-\frac{1}{2}x^2} \sin(2\pi x)dx\end{aligned}$$

where the last line used $\sin(2\pi(x+n)) = \sin(2\pi x)$. The above integral is zero because the integrand is an odd function. Thus if $g_c(t) = f(t)(1 + c\sin(2\pi t))$, with $|c| \le 1$, then $g_c \ge 0$ and

$$\int t^n g_c(t)dt = \int t^n f(t)dt \quad \text{for all } f.$$

Exercise 25. Fix $0 < \lambda < 1$. For a suitable choice of β, show that $\int x^n e^{-|x|^\lambda} \sin(\beta|x|^\lambda \text{sgn}(x))dx = 0$ for all n. Produce many measures having a common moment sequence.

Sufficient conditions for uniqueness: For practical purposes, it is useful to have sufficient conditions for recovery. Here are three (progressively stronger) sufficient conditions that are sufficient for most purposes.

(1) If μ is compactly supported, it is determined by its moment sequence. In terms of the moment sequence, this is equivalent to $\alpha_{2n} \le C^n$ for some $C < \infty$ (i.e., $\limsup\limits_{n\to\infty} \alpha_{2n}^{1/2n} < \infty$).

(2) If μ has finite Laplace transform in a neighbourhood of zero, i.e., if $\mathcal{L}_\mu(t) = \int e^{tx}d\mu(x) < \infty$ for $t \in (-\delta, \delta)$ for some $\delta > 0$, then μ is determined by its moment sequence. This condition is equivalent to $\alpha_{2n} \le (Cn)^n$ for some $C < \infty$ (i.e., $\limsup\limits_{n\to\infty} \frac{1}{n}\alpha_{2n}^{1/2n} < \infty$).

(3) If $\sum_{n=1}^\infty \alpha_{2n}^{-1/2n} = \infty$ (and α is positive definite), there is a unique measure whose moment sequence is α. This is known as *Carleman's condition*.

We just justify the first condition. First, observe that if μ is supported on $[-M, M]$, then $\alpha_{2n} \le M^{2n}$. Conversely, if $\alpha_{2n} \le C^{2n}$, observe that $\mu([-M, M]^c) \le M^{-2n}\alpha_{2n}$ which goes to zero as $n \to \infty$, provided $M > C$. Thus μ is supported on $[-C, C]$.

Now if a moment sequence α satisfying $\alpha_{2n} \le M^{2n}$ is given, and μ and ν are two measures on $[-M, M]$ having the moment sequence α, we see that $\int p(x)d\mu(x) = \int p(x)d\nu(x)$ for all polynomials p. Use Weierstrass' approximation to conclude that $\int f d\mu = \int f d\nu$ for all $f \in C[-M, M]$. For any $[a, b] \subseteq [-M, M]$, it is easy to find continuous functions that decrease to $\mathbf{1}_{[a,b]}$. Monotone convergence theorem implies that $\mu[a, b] = \nu[a, b]$ and thus $\mu = \nu$.

10. The uniqueness question: Finitely supported measures

In this section, we consider finitely supported measures, positive semi-definite sequences (that are not positive definite), finite sequences of orthogonal polynomials, and finite Jacobi matrices. As before, we show how to go from one to the next, but crucially, we can also go back from Jacobi matrices to finitely supported measures, completing the cycle. This will also motivate our next discussion on the importance of the spectral theorem in going from a positive semi-definite sequence to a measure. Since the steps are analogous, we keep this account brief.

Let $\mu = p_1\delta_{\lambda_1} + \ldots + p_n\delta_{\lambda_n}$ where $n \ge 1$, $\lambda_1 < \ldots < \lambda_n$ and $p_i > 0$ with $p_1 + \ldots + p_n = 1$ be a measure supported on finitely many points of the real line.

The kth moment of μ is $\alpha_k = \int x^k d\mu(x) = \sum_{j=1}^n p_j \lambda_j^k$. Clearly $\alpha_0 = 1$. As before, the matrix $H_\alpha = (\alpha_{i+j})_{i,j\geq 0}$ is positive semi-definite, because for any $m \geq 0$ and $c_0, \ldots, c_m \in \mathbb{R}$,

$$0 \leq \int \Big| \sum_{i=0}^m c_i x^i \Big|^2 d\mu(x) = \sum_{i,j=1}^N c_i c_j \int x^{i+j} d\mu(x) = \sum_{i,j=1}^N c_i c_j \alpha_{i+j}.$$

Equality holds in the above inequality if and only if $\sum_{i=0}^m c_i x^i = 0$ $a.e.[\mu]$ which is the same as saying that $\sum_{i=0}^m c_i \lambda_k^i = 0$ for $1 \leq k \leq n$. Writing in matrix form, this is equivalent to

$$\begin{bmatrix} \lambda_1^0 & \lambda_1^1 & \ldots & \lambda_1^m \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_n^0 & \lambda_1^1 & \ldots & \lambda_n^m \end{bmatrix} \begin{bmatrix} c_0 \\ \vdots \\ c_m \end{bmatrix} = \mathbf{0}.$$

If $m = n-1$, the matrix on the left is square and has determinant $\prod_{i<j}(\lambda_j - \lambda_i)$ which is non-zero as the λ_is are distinct. For $m \geq n$, clearly there exist c_js such that the equation is satisfied, since the matrix has rank at most n. Thus, H_α has rank n, and more specifically, its top $k \times k$ principal sub-matrix is non-singular for $k \leq n-1$ and singular for $k \geq n$. Thus, α is positive semi-definite but not positive definite.

Applying Gram-Schmidt to $1, x, x^2, \ldots$, we get polynomials $\varphi_0, \varphi_1, \ldots, \varphi_{n-1}$, where φ_j has degree j. We cannot proceed further, as x^n is linearly dependent on $1, x, \ldots, x^{n-1}$ in the given inner product (i.e., in $L^2(\mu)$). This is the orthogonal polynomial sequence.

To get the three term recurrence, we again write, for $k \leq n-2$,

$$x\varphi_k(x) = c_{k,k+1}\varphi_{k+1}(x) + \ldots + c_{k,0}\varphi_0(x).$$

Using the inner product of $L^2(\mu)$ (since $M : L^2(\mu) \mapsto L^2(\mu)$ is symmetric), we reason as before that $c_{k,j} = 0$ for $j \leq k-2$, $c_{k,k+1} = c_{k+1,k}$ and writing $a_k = c_{k,k}$ and $b_k = c_{k,k+1}$ (this is positive, why?) thus get the three term recurrence

$$\begin{aligned} x\varphi_0(x) &= a_0\varphi_0(x) + b_0\varphi_1(x), \\ x\varphi_k(x) &= b_{k-1}\varphi_{k-1}(x) + a_k\varphi_k(x) + b_k\varphi_{k+1}(x) \quad \text{for } 1 \leq k \leq n-2. \end{aligned}$$

Lastly, it also holds that (we leave the reasoning to you)

$$x\varphi_{n-1}(x) \overset{L^2(\mu)}{=} b_{n-2}\varphi_{n-2}(x) + a_{n-1}\varphi_{n-1}(x).$$

Equality in $L^2(\mu)$ means that the difference has zero norm in $L^2(\mu)$, or equivalently, equality holds for $x \in \{\lambda_1, \ldots, \lambda_n\}$. The equality cannot be for all x as the left side is a polynomial of degree n but the right side has lower degree.

The three term recurrences can be written in matrix form as

$$\begin{bmatrix} a_0 & b_0 & 0 & \ldots & 0 \\ b_0 & a_1 & b_1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \ldots & 0 & b_{n-2} & a_{n-1} \end{bmatrix} \begin{bmatrix} \varphi_0(x) \\ \vdots \\ \varphi_{n-1}(x) \end{bmatrix} \overset{L^2(\mu)}{=} x \begin{bmatrix} \varphi_0(x) \\ \vdots \\ \varphi_{n-1}(x) \end{bmatrix}$$

The equality is in L^2 because the very last equation holds only when $x \in \{\lambda_1, \ldots, \lambda_n\}$. Let T_n denote the Jacobi matrix on the left.

The previous equality holds for $x \in \{\lambda_1, \ldots, \lambda_n\}$, showing that λ_k is an eigenvalue of T_n with eigenvector $(\varphi_0(\lambda_k), \ldots, \varphi_{n-1}(\lambda_k))^t$. Thus, if we are given the Jacobi matrix, we recover the support of μ, it is precisely the spectrum of T_n. We can also recover the weights as follows (we have seen very similar reasoning earlier). Observe that T is the matrix for the multiplication operator on $L^2(\mu)$. Therefore,

$$\langle T^m e_0, e_0 \rangle = \langle x^m, x^0 \rangle_{L^2(\mu)} = \alpha_m = \sum_{k=1}^{n} \lambda_k^m p_k.$$

On the other hand, the spectral decomposition of the Jacobi matrix is $T_n = Q\Lambda Q^t$ where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ and

$$Q_{i,j} = \frac{\varphi_i(\lambda_j)^2}{\sum_{\ell=0}^{n-1} \varphi_i(\lambda_j)^2}.$$

Therefore, it is clear that

$$\langle T_n^m e_0, e_0 \rangle = \langle Q\Lambda^m Q^t e_0, e_0 \rangle = \sum_{j=1}^{n} \lambda_j^m Q_{0,j}^2.$$

Equating with the earlier identity, we have recovered the measure as

$$\mu = \sum_{k=1}^{n} p_k \delta_{\lambda_k}$$

where λ_k are the eigenvalues of T_n and $p_k = Q_{0,k}^2$ are the squared entries of the first row of the eigenvector matrix of T_n.

Conclusion: From a finitely supported measure, we can compute its moments. From the moment sequence we can recover the measure by going first to the orthogonal polynomials, then to the Jacobi matrix describing the three-term recurrence, and from there to the measure, via the *spectral decomposition of the Jacobi matrix*. In summary, the measure is just the *spectral measure of the Jacobi matrix at the vector* e_0.

11. The uniqueness question: Connection to spectral theorem

If α is a positive definite sequence, we construct its Jacobi matrix $T = T(a, b)$. Going by the finite support case, we may expect that the measure (or one measure) with moment sequence α can be recovered from T by taking spectral measure at e_0. There are many subtleties on the way.

First w regard T as an operator on sequences by mapping $(Tx)_n = b_{n-1}x_{n-1} + a_n x_n + b_n x_{n+1}$. While this is well-defined for any $x \in \mathbb{R}^{\mathbb{N}}$, to talk about spectral theorem, we must work inside a Hilbert space. Here the natural Hilbert space is $\ell^2 = \{(x_0, x_1, x_2, \ldots) : \sum_n x_n^2 < \infty\}$.

A special case is when the entires of T are bounded. In this case, by Cauchy-Schwarz inequality

$$|(Tx)_n|^2 \le (a_n^2 + b_{n-1}^2 + b_n^2)(x_{n-1}^2 + x_n^2 + x_{n+1}^2)$$

and hence $\|Tx\|_{\ell^2}^2 \le 3M^2\|x\|_{\ell^2}^2$ where M is a bound for $|a_n|$s and $|b_n|$s. Thus $T : \ell^2 \mapsto \ell^2$ is a bounded operator. It also satisfies $\langle Tx, y\rangle = \langle x, Ty\rangle$ for all $x, y \in \ell^2$, which makes it self-adjoint.

The spectral theorem for bounded self-adjoint operators tells us that we can write $T = \int \lambda dE(\lambda)$ where E is a *projection valued measure*. This representation is also unique etc. As a consequence, there is a measure μ (it is defined by $\mu(A) = \langle E(A)e_0, e_0\rangle$ for $A \in \mathcal{B}_{\mathbb{R}}$) such that or any m,

$$\langle T^m e_0, e_0\rangle = \int x^m d\mu(x).$$

Recall that the positive definite sequence can be recovered from the Jacobi matrix as $\alpha_n = \langle T^n e_0, e_0\rangle$ to see that μ has the moment sequence α.

If the entries of T are not bounded, it is no longer the case that T defines a bounded linear operator on ℓ^2. By restricting the domain to $D = \{x \in \ell^2 : x_n = 0 \text{ eventually}\}$, we see that $T : D \mapsto \ell^2$ is linear. Since D is dense in ℓ^2, if we had $\|Tx\|_{\ell^2} \le C\|x\|_{\ell^2}$ for $x \in D$, then it would extend to all of ℓ^2 as a bounded linear operator. That is not the case when the entries of T are unbounded. These operators are called *unbounded operators*

Example 26. Let $a_n = n$ and $b_n = 0$. Then of T acts on D by $(Tx)_n = nx_n$. We could also have defined T on a larger domain $D_1 = \{x : \sum nx_n^2 < \infty\}$ because then T clearly maps D_1 into ℓ^2. It is better to denote tis operator as T_1 and regard it as an extension of T.

In functional analysis class one learns how to associate an adjoint operator T^* which is defined on another proper subspace D^*. For our T, the symmetry of the Jacobi matrix forces that $D^* \supseteq D$ and that $T^*\big|_D = T$. We say that T is *symmetric*. This is not sufficient to get a spectral decomposition. What one needs is *self-adjointness*, i.e., for D and D^* to coincide and T and T^* to coincide. Once T is self-adjoint, spectral theorem can be proved in full force (the only difference is that when T is bounded, the projection valued measure E and the spectral measure μ are both compactly supported, while they need no be so now).

Example 27. Take $T : D \mapsto \ell^2$ and $T_1 : D_1 \mapsto \ell^2$ as in the previous example. It is easy to work out that $D^* = D_1^* = D_1$. Further, T^*, T_1^* and T coincide on D_1. Thus (in the language introduced next), T is symmetric, while T_1 is self-adjoint.

To achieve this one tries to extend T to a larger domain $D_1 \supseteq D$ and get an operator $T_1 : D_1 \mapsto \ell^2$. Then it turns out that $D_1^* \subseteq D^*$ and $T_1^* : D_1^* \mapsto \ell^2$ is the adjoint of T_1. General theorems assert the existence of self-adjoint extensions (at least for our Jacobi matrices), but there can be several self-adjoint extensions. This has repercussions in the moment problem.

Theorem 28. *Let T be the Jacobi matrix of a positive definite sequence α (with $\alpha_0 = 1$). Regard it as an operator $T : D \mapsto \ell^2$ where D is the set of sequences that are eventually 0.*

(1) If $\tilde{T} : \tilde{D} \mapsto \ell^2$ is a self-adjoint extension of T, then the spectral measure of $\tilde{T}$ at the vector e_0 is a probability measure whose moment sequence is α. If the self-adjoint extension is unique, this is the unique measure having this moment sequence.

(2) *If there are distinct self-adjoint extensions, then they have distinct spectral measures at* e_0*, each having the same moment sequence* α*.*

12. Convexity approach to generalized moment problems

12.1. **The generalized moment problem on an interval.** Let E be a metric space and let u_i, $i \in I$, be (complex-valued) continuous functions on E. Given $\alpha : I \to \mathbb{C}$, does there exist a Borel measure μ on E such that $\int_E u_i d\mu = \alpha(i)$ for all $i \in I$.

Many problems are of this nature.

(1) *Classical moment problems.* Let $E = [0,1]$ and $u_k(x) = x^k$, $k \in \mathbb{N}$. This is the Hausdorff moment problem. Changing E to $\mathbb{R}$ or $[0,\infty)$ lead to the Hamburger and Steiltjes' moment problems.

(2) *Trigonometric moment problem.* Let $E = \mathbb{S}^1 \cong [0,2\pi)$ and $u_k(\theta) = e^{ik\theta}$, $k \in \mathbb{Z}$. This is the problem of characterizing Fourier series of measures on $\mathbb{S}^1$, i.e., what functions $\alpha : \mathbb{Z} \to \mathbb{C}$ are of the form $\alpha(k) = \int e^{ik\theta} d\mu(\theta)$? If one prefers real-valued functions, one may replace the pair u_k, u_{-k} by $\cos(k\theta), \sin(k\theta)$, for $k \geq 0$.

(3) *Bochner's theorem.* $E = \mathbb{R}$ and $u_\lambda(x) = e^{i\lambda x}$, for $\lambda \in \mathbb{R}$. The question here is of characterizing Fourier transforms of measures on $\mathbb{R}$. That is, what functions $\alpha : \mathbb{R} \to \mathbb{C}$ are of the form $\alpha(\lambda) = \int e^{ix\lambda} d\mu(x)$? Bochner's theorem provides the answer.

(4) *Riesz's representation theorem.* $\{u_i\} = C(E)$. The answer is that a functional $\alpha : C(E) \to \mathbb{R}$ is attainable if and only if it is a positive linear functionals on $C(E)$ (when E is locally compact).

(5) *Poisson integral formula for positive harmonic functions on the disk.* Let $E = \mathbb{S}^1$ and $u_z(\lambda) = \operatorname{Re} \frac{\lambda+z}{\lambda-z} = \frac{1-r^2}{1-2r\cos(\theta-t)+r^2}$ where $z = re^{i\theta} \in \mathbb{D}$ and $\lambda = e^{-it} \in \mathbb{S}^1$. This is basically the Poisson kernel (when thought of as a function of (z,λ)). The question here is to determine functions $\alpha : \mathbb{D} \to \mathbb{R}$ that are of the form $\alpha(z) = \int_{\mathbb{S}^1} u_z(t) d\mu(t)$ for a probability measure μ on $\mathbb{S}^1$. The answer turns out to be precisely positive harmonic functions.

(6) *Herglotz representation theorem.* Let $E = \mathbb{S}^1$ and $u_z(\lambda) = \frac{\lambda+z}{\lambda-z}$ for $z \in \mathbb{D}$ (the unit disk), $\lambda \in \mathbb{S}^1$. In this case, the functions $\alpha : \mathbb{D} \to \mathbb{C}$ that are representable in the form $\alpha(z) = \int_{\mathbb{S}^1} u_z(\lambda) d\mu(\lambda)$ turn out to be precisely the collection of holomorphic functions from $\mathbb{D}$ into the right half-plane $\mathbb{H}_+ = \{w : \operatorname{Re} w > 0\}$ with the additional condition that $\alpha(0) \in \mathbb{R}_+$.

(7) *Nevanlinna-Pick interpolation problem*: Given $z_1, \ldots, z_n \in \mathbb{D}$ and $w_1, \ldots, w_n \in \mathbb{D}$, does there exist a holomorphic function $f : \mathbb{D} \to \mathbb{D}$ such that $f(z_i) = w_i$, for $i \leq n$? This does not look like it has anything to do with measures, but because of the Herglotz representation above (and the fact that we have an explicit conformal equivalence of $\mathbb{H}_+$ with $\mathbb{D}$), we can realize this as a generalized moment problem!

12.2. **Finite index case.** The main points are most easily conveyed when the index set is finite, so we take $u_0, \ldots, u_n : E \to \mathbb{R}$ (the complex case can be handled by separating each u_i into its real and imaginary parts) and $\alpha_0, \ldots, \alpha_n \in \mathbb{R}$. Let $W = \text{span}\{u_0, \ldots, u_n\}$ be the space of all linear combinations and let $W_+ = \{u \in W : u \geq 0 \text{ on } E\}$ be the positive ones.

Assume that $\{u_0, \ldots, u_n\}$ are linearly independent. Then there is a unique linear functional $L : W \to \mathbb{R}$ such that $L(u_k) = \alpha_k$, $0 \le k \le n$.

In some cases, there may be linear dependence among u_is, but then α_i must obviously satisfy the corresponding linear constraints, and we can reduce the problem to the above case by taking a maximal linearly independent subset of $\{u_0, \ldots, u_n\}$.

Theorem 29. *For linearly independent $u_0, \ldots, u_n \in C(E)$, and $\alpha_0, \ldots, \alpha_n \in \mathbb{R}$, the following are equivalent.*

(1) There exists a measure μ such that $\int u_k d\mu = \alpha_k$ for $0 \le k \le n$.

(2) $L(u) \ge 0$ for all $u \in W_+$.

12.3. **Some ingredients from convex analysis.** The heart of the matter is convexity. This was implicit in the other approach, where we used Hahn-Banach like theorems, but is made more explicit in this approach.

Definition 30. Let $K \subseteq \mathbb{R}^n$. We say that K is convex if $tx + (1-t)y \in K$ whenever $x, y \in K$ and $0 < t < 1$. We say that K is conical if $tx \in K$ whenever $x \in K$ and $t > 0$. A wedge is a closed convex conical set.

In other words, a closed set K is a wedge if and only if it is closed under addition of vectors and under multiplication by positive scalars. A fundamental concept is that of the convex dual.

Definition 31. Let K be a wedge. Its (convex) dual is the set

$$K^{\dagger} := \{a \in \mathbb{R}^n : \langle a, x\rangle \ge 0 \; \forall x \in K\}.$$

It is easy to see that $K^{\dagger}$ is a wedge (this is true even if K is an arbitrary set, not necessarily a wedge). In fact $K^{\dagger}$ is the intersection of all H_x, $x \in K$, where H_x is the closed half-space of vectors that have positive inner product with x. The crucial fact, and the reason for the word "dual", is as follows.

Proposition 32. *If K is a wedge, then $(K^{\dagger})^{\dagger} = K$.*

Proof. By definition of $K^{\dagger}$, we have $\langle a, x\rangle \ge 0$ for all $a \in K^{\dagger}$, $x \in K$. Hence it follows that $K \subseteq (K^{\dagger})^{\dagger}$. If the inclusion was strict, then let $y \in (K^{\dagger})^{\dagger} \setminus K$. We can find an affine linear functional $L(x) = \langle b, x\rangle - c$ with $c \in R$, $b \in \mathbb{R}^n$, such that $L(x) \ge 0$ for $x \in K$ and $L(y) < 0$ (this is the separating hyperplane theorem, you may call it Hahn-Banach too). This means that $\langle y, b\rangle < c \le \langle x, b\rangle$ for all $x \in K$. Taking $x = 0$ (which belongs to K by closedness) in the second inequality, $c \le 0$. Then $\langle b, x\rangle \ge 0$ for all $x \in K$, showing that $b \in K^{\dagger}$. But $y \in (K^{\dagger})^{\dagger}$, hence $\langle y, b\rangle \ge 0$, contradicting the first inequality. ■

12.4. **Analysis of the generalized moment problem.** Define the curve $U : E \to \mathbb{R}^{n+1}$ by $U(x) = (u_0(x), \ldots, u_n(x))$. Let K be the smallest wedge containing the image of U (i.e., the intersection of all wedges that contains the image). We need a description of K and $K^{\dagger}$.

(1) K: Clearly, K must contain all finite convex combinations of the form $p_1 U(x_1) + \ldots + p_m U(x_m)$, $m \geq 1$ and $p_i \geq 0$ with $p_1 + \ldots + p_m = 1$. Approximating general probability measures by discrete ones, and the closedness of K, we see that K must also contain $\int U(x) d\mu(x)$ for any Borel probability measure μ on E. On the other hand, the collection of all vectors $\int U d\mu$, as μ varies, is a wedge. Therefore, $K = \{\int U d\mu : \mu \in \mathcal{P}(E)\}$.

(2) $K^\dagger$: If $a = (a_0, \ldots, a_n) \in K^\dagger$, then $\langle a, v\rangle \geq 0$ for $v \in \text{Image}(U)$, i.e., $a_0 u_0(x) + \ldots + a_n u_n(x) \geq 0$ for all $x \in E$. Conversely, if $a_0 u_0(x) + \ldots + a_n u_n(x) \geq 0$ for all $x \in E$, then the wedge (in fact a half-space) $\{v : \langle a, v\rangle \geq 0\}$ contains the image of U, and hence the whole of K. Thus, we may identify $K^\dagger$ with W_+, i.e., $K^\dagger = \{(a_0, \ldots, a_n) : a_0 u_0 + \ldots + a_n u_n \in W_+\}$.

Now the duality tells us that $(K^\dagger)^\dagger = K$. In other words, the following are equivalent for $\alpha = (\alpha_0, \ldots, \alpha_n) \in \mathbb{R}^{n+1}$:

(1) $\alpha \in K$: This means that $\alpha_k = \int u_k d\mu$ for all $0 \leq k \leq n$, for some $\mu \in \mathcal{P}(E)$.

(2) $\alpha \in (K^\dagger)^\dagger$: This means that $\sum_k \alpha_k a_k \geq 0$, whenever $\sum_k a_k u_k \in W_+$. In other words, $L(u) \geq 0$ for $u \in W_+$.

This completes the proof of Theorem 29.

12.5. **Identifying W_+ in special cases.** The solution to the generalized moment problem given in Theorem 29 is the first step. The main weakness is that the collection of positive functions W_+ is not very explicit, making the checking of the condition $L(u) \geq 0$ for $u \in W_+$ near-impossible. In a few specific examples, one can find an explicit description of W_+, thus leading to a more usable solution to the problem.

Here is the representation of positive polynomials on intervals of the real line.

Proposition 33. *Let $p(x)$ be a real polynomial of degree n.*

(1) If $p \geq 0$ on $\mathbb{R}$, then p is a sum of squares of polynomials.

(2) If $p \geq 0$ on $[0, \infty)$, then p is a sum of polynomials of the form $q^2(x)$ and $xq^2(x)$.

(3) If $p \geq 0$ on $[0, 1]$, then p is a sum of polynomials of the form $q^2(x)$ and $x(1-x)q^2(x)$.

One can be more explicit about how many summands are needed, but that is not necessary for our purpose.

Proof. As p is real, we can factorize it as $p(x) = C \prod_{i=1}^{\ell} (x - t_i)^{m_i} \prod_{i=1}^{k} (x - w_i)(x - \bar{w}_i)$ where t_i are distinct real and $\text{Im}\, w_i > 0$.

(1) Assume $p \geq 0$ on $\mathbb{R}$. If any m_i is odd, then $p(x)$ changes sign at t_i, hence cannot be positive. Writing $m_i = 2n_i$, we get $p(x) = |Q(x)|^2$ with $Q(x) = \sqrt{C} \prod_{i=1}^{m} (x - s_i)^{n_i} \prod_{i=1}^{k} (x - w_i)$. Here note that $C > 0$ as $p(x) \sim Cx^{\ell+2k}$ as $x \to \infty$. If $Q = q + ir$, then $|Q|^2 = q^2 + r^2$, making p a sum of two squares.

(2) If $p \geq 0$ on $[0, \infty)$, then the real zeros with odd multiplicity must all be in $(-\infty, 0]$, hence we may write $p(x) = |Q(x)|^2 \prod_{i=1}^{\ell'} (x + s_i)$ where $s_i \geq 0$ are distinct (if s_i has multiplicity $2k+1$,

then k of them are absorbed into $Q(x)$). Expand $\prod(x+s_i)$ and collect the even and odd terms separately (all coefficients are positive as they are sums of products of $s_1,\dots,s_{\ell'}$). A term of the form $cx^{2k}|Q(x)|^2 = (\sqrt{c}x^k q(x))^2 + (\sqrt{c}x^k r(x))^2$ is a sum of squares, while a term of the form $cx^{2k+1}|Q(x)|^2 = x(\sqrt{c}x^k q(x))^2 + x(\sqrt{c}x^k r(x))^2$ are of the form x times a square polynomial. Thus p has the claimed representation, with at most four summands.

(3) If $p \geq 0$ on $[0,1]$, then the real zeros with odd multiplicity are in $(-\infty,0]$ or in $[1,\infty)$, hence we may write $p(x) = |Q(x)|^2 \prod_{i=1}^{k}(x+s_i)\prod_{j=1}^{\ell}(u_i+1-x)$, where $s_i, u_i \geq 0$. Expand this product and observe that each term (keep the $1-x$ term intact in the second factor) is a product of a square polynomial with one of 1 or x or $1-x$ or $x(1-x)$. But $xq^2(x) = x(1-x)q^2(x) + (xq(x))^2$ and $(1-x)q^2(x) = x(1-x)q^2(x) + ((1-x)q(x))^2$, hence we only have summands of the form $q^2(x)$ and $x(1-x)q^2(x)$. Clearly six summands suffice. ■.

For trigonometric polynomials, we have the following analogous result which has an even simpler form.

Proposition 34. *Let $T(\theta) = \sum_{k=-n}^{n} a_k e^{ik\theta}$ be a trigonometric polynomial such that $T(\theta) \geq 0$ for $\theta \in [0,2\pi)$. Then $T(\theta) = |S(\theta)|^2$ for a trigonometric polynomial S.*

Proof. Writing $T(e^{i\theta})$ instead of $T(\theta)$, we see that T is the restriction to the unit circle of the rational function $T(z) = \sum_{k=-n}^{n} a_k z^k$. Note that $a_0 \in \mathbb{R}$ and $a_{-k} = \bar{a}_k$ for T to be real-valued on the unit circle. As $\overline{T(1/\bar{z})} = T(z)$ for $z = e^{i\theta}$, the identity holds for all $z \in \mathbb{C}$. This shows that the zeros of the polynomial $z^n T(z)$ are either on the unit circle, or occur in pairs of the form $\{w, \frac{1}{\bar{w}}\}$. The zeros on the unit circle must occur with even multiplicity, otherwise $T(e^{i\theta})$ would change sign at such a zero. Unpacking all this gives us the representation $z^n T(z) = a_n \prod_{i=1}^{k}(z-w_i)(z-\frac{1}{\bar{w}_i})\prod_{i=1}^{\ell}(z-\xi_i)^{2m_i}$ where ξ_i are distinct points on the unit circle. Thus $z^n T(z) = |Q(z)|^2$ for $Q(z) = \prod_{i=1}^{k}(z-w_i)\prod_{i=1}^{\ell}(z-\xi_i)^{m_i}$. TO COMPLETE ■

12.6. **Solution to the moment problem in special cases.** In the special cases, we shall be able to express the condition of positivity of L on W_+ in terms of positive semi-definiteness of certain matrices. Everywhere, an infinite matrix $(a_{i,j})_{i,j\geq 0}$ is said to be positive semi-definite if every finite principal submatrix is. What this means is that $a_{i,j} = \bar{a}_{j,i}$ for all i,j and

$$\sum_{i,j=0}^{n} c_i \bar{c}_j a_{i,j} \geq 0 \qquad \text{for any } n \geq 0 \text{ and } c_1,\dots c_n \in \mathbb{C}.$$

When $a_{i,j}$ are real, it suffices to take $c_i \in \mathbb{R}$.

Theorem 35. *Let $\alpha_k \in \mathbb{R}$, with $\alpha_0 = 1$. Then there exists a probability measure μ on $\mathbb{R}$ such that $\int x^k d\mu(x) = \alpha_k$ for all k if and only if the infinite matrix $(\alpha_{j+k})_{j,k\geq 0}$ is positive semi-definite.*

Proof. Fix finite n and consider the moment problem with $u_k(x) = x^k$, $0 \leq k \leq n$. Here W is the space of polynomials of degree at most n $L : W \to \mathbb{R}$ is defined by $L(u_k) = \alpha_k$ for $0 \leq k \leq n$. and any $p \in W_+$ can be written as a sum of squares of polynomials. Hence, the condition $L(u) \geq 0$ for $u \in W_+$, is equivalent to $L(q^2) \geq 0$ for all $q^2 \in W$. But if $q(x) = c_0 + c_1 x + \ldots + c_m x^m$, then

$$L(q^2) = L\left(\sum_{j,k=0}^{m} c_j c_k u_{j+k}\right) = \sum_{j,k=0}^{m} c_j c_k \alpha_{j+k}.$$

Thus, the positivity of L on W_+ is equivalent to positive semi-definiteness of $(\alpha_{j+k})_{0 \leq j,k \leq m}$.

This proves the necessity of the positive semi-definiteness of $(\alpha_{j+k})_{j,k\geq 0}$.

To see sufficiency, by Theorem 29 it follows that for each n, there is a probability measure μ_n on $\mathbb{R}$ such that $\int x^k d\mu_n(x) = \alpha_k$ for $0 \leq k \leq n$. By the fact that the second moments of $\int x^2 d\mu_n(x) = \alpha_2$ are bounded, we get the tightness of μ_n, and hence we can get a subsequential limit μ. As $\int x^{2p} d\mu_n(x) = \alpha_{2p}$ is also bounded (as n varies), by a similar argument, we conclude that the moments of μ_n along the subsequence converge to those of μ. In other words, $\int x^k d\mu(x) = \alpha_k$ for all $k \geq 0$. ■

Moment problem on $\mathbb{R}_+$: By the same method of proof, but using the description of W_+ as given in Proposition 34, we arrive at the following theorem.

Theorem 36. *Let $\alpha_k \in \mathbb{R}$ with $\alpha_0 = 1$. There exists a probability measure on $\mathbb{R}_+$ whose moments are α_ks if and only if the infinite matrices $(\alpha_{j+k})_{j,k\geq 0}$ and $(\alpha_{j+k+1})_{j,k\geq 0}$ are positive semi-definite.*

Proof. Given the description of elements of W_+ as sums of $q^2(x)$ and $xq^2(x)$, writing $q(x) = c_0 + \ldots + c_m x^m$, we see that

$$L(q^2(x)) = \sum_{j,k=0}^{m} c_j c_k \alpha_{j+k}, \qquad L(xq^2(x)) = \sum_{j,k=0}^{n} c_j c_k \alpha_{j+k+1}.$$

The rest of the proof is identical to the proof of Theorem 35. ■

Moment problem on $[0,1]$: Again, using the description of W_+ as given in Proposition 34, we arrive at the following theorem.

Theorem 37. *Let $\alpha_k \in \mathbb{R}$ with $\alpha_0 = 1$. There exists a probability measure on $[0,1]$ whose moments are α_ks if and only if the infinite matrices $(\alpha_{j+k})_{j,k\geq 0}$ and $(\alpha_{j+k+1} - \alpha_{j+k+2})_{j,k\geq 0}$ are positive semi-definite.*

Proof. of W_+ are sums of polynomials of the form $q^2(x)$ and $x(1-x)q^2(x)$. Writing $q(x) = c_0 + \ldots + c_m x^m$, we see that

$$L(q^2(x)) = \sum_{j,k=0}^{m} c_j c_k \alpha_{j+k}, \qquad L(x(1-x)q^2(x)) = \sum_{j,k=0}^{m} c_j c_k (\alpha_{j+k+1} - \alpha_{j+k+2}).$$

The rest of the proof is identical to the previous cases. ■

Trigonometric moment problem:

Theorem 38. *Let $\gamma_k \in \mathbb{C}$ with $\gamma_0 = 1$ and $\gamma_{-k} = \bar{\gamma}_k$. Then there exists a probability measure μ on $\mathbb{S}^1$ with Fourier coefficients γ_k if and only if $(\gamma_{j-k})_{j,k \geq \in \mathbb{Z}}$ is positive semi-definite.*

Proof. By Proposition 34, positive trigonometric polynomials are of the form $|Q|^2$. Writing $Q(\theta) = \sum_{j=-m}^{m} c_j u_j$ where $u_j(\theta) = e^{ij\theta}$, we see that

$$L(|Q|^2) = \sum_{j,k=-m}^{m} c_j \bar{c}_k L(u_{j-k}) \; = \sum_{j,k=-m}^{m} c_j \bar{c}_k \gamma_{j-k}.$$

Thus the positivity of L on W_+ is the same as positive definiteness of $(\gamma_{j-k})_{j,k \geq \in \mathbb{Z}}$. The rest of the proof is similar to the previous cases: Consider $\{u_j\}_{|j| \leq m}$, get a measure μ_m with $\int u_j d\mu = \gamma_j$, use compactness of $\mathbb{S}^1$ to get a subsequential limit μ and argue that its Fourier coefficients are γ_ks. ■

Nevanlinna-Pick interpolation problem: Let $\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$ and $\mathbb{H}_+ = \{z : \operatorname{Re} z > 0\}$.

Theorem 39. *Let $z_i \in \mathbb{D}$ and $w_i \in \mathbb{H}_+$, $1 \leq i \leq n$. Then there exists a holomorphic function $f : \mathbb{D} \to \mathbb{H}_+$ that maps $f(z_i) = w_i$ for $1 \leq i \leq n$ if and only if the matrix $\left(\frac{w_j + \bar{w}_k}{1 - z_j \bar{z}_k}\right)_{1 \leq j,k \leq n}$ is positive semi-definite.*

A priori this does not look related to the generalized moment problem as we are not asking for a measure. That connection comes from the following representation of holomorphic functions from the disk to the right half-plane.

Theorem 40 (Herglotz representation theorem). *Let $f : \mathbb{D} \to \mathbb{H}_+$ be holomorphic. Then there is a unique probability measure μ on $\mathbb{S}^1 \cong [0, 2\pi)$ and $b \in \mathbb{R}$ and $a > 0$ such that*

$$f(z) = ib + a \int_{\mathbb{S}^1} \frac{e^{it} + z}{e^{it} - z} d\mu(t).$$

Conversely, any function of the above form is a holomorphic function from $\mathbb{D}$ to $\mathbb{H}_+$.

Observe that $a = \operatorname{Re} f(0)$ and $b = \operatorname{Im} f(0)$ in this representation. As usual, the converse part is easy, since for each $t \in [0, 2\pi)$, the function $\varphi_t(z) = \frac{e^{it}+z}{e^{it}-z}$ maps $\mathbb{D}$ to $\mathbb{H}_+$. The forward implication is proved using Poisson integral representation for the harmonic function $u = \operatorname{Re} f$. It is convenient to do this first assuming that f is holomorphic in a neighbourhood of $\overline{\mathbb{D}}$ so that boundary function is nice. This can be done by approximating f by $f_r(z) = f(rz)$ as $r \uparrow 1$. But it is also illuminating to see the Herglotz representation itself as an expression of the generalized moment problem! This heuristic is outlined in the remark below.

Remark 41. Let $\mathcal{F}$ denote the space of holomorphic function on $\mathbb{D}$ that are real on the real line. Let K be the smallest wedge in $\mathcal{F}$ containing φ_t, $t \in [0, 2\pi)$. This is easily seen to be all functions of the form $a \int_{\mathbb{S}^1} \varphi_t d\mu$, where $a > 0$ and μ is a probability measure on $\mathbb{S}^1$. We need to identify $K^\dagger$ (as we are in infinite dimensional space, the $K^\dagger$ should be defined as the collection of all linear functionals that are positive on K). It is easy to see that $f \mapsto \operatorname{Re} f(z) = f(z) + f(\bar{z})$ is a linear functional that is positive on K, for each $z \in \mathbb{D}$. If we can argue that these are all (some closure is needed, in fact), then we essentially identify $K^\dagger$ with $\overline{\mathbb{D}_+} = \{z : |z| \le 1, \operatorname{Im} z \ge 0\}$. By the duality $(K^\dagger)^\dagger = K$ (again, this is a heuristic, as we have not established duality in infinite dimensional spaces), we get that holomorphic functions in $\mathcal{F}$ that map $\mathbb{D}$ to $\mathbb{H}_+$ (which are the elements of $(K^\dagger)^\dagger$) are precisely those with the Herglotz integral representation (which are the elements of K).

Proof of Theorem 39. Without loss of generality, we may assume $z_1 = 0$ and $w_1 = 1$. This can be done by replacing f by $\psi \circ f \circ \varphi$, where φ (respectively ψ) is a linear fractional transformation that map $\mathbb{D}$ onto itself (respectively, $\mathbb{H}_+$ onto itself). Then the Herglotz representation says that $f(z) = \int \varphi_t(z) d\mu(t)$ for some probability measure μ on $\mathbb{S}^1$. We need to show the existence of a μ such that $\int \varphi_t(z_k) d\mu(t) = w_k$, for $1 \le k \le n$.

Consider the generalized moment problem on $\mathbb{S}^1$, with (a) $u_0 = 1$ and $u_k(t) = \varphi_t(z_k)$, $1 \le k \le n$ and (b) $L(u_k) = w_k$ Here we are working with complex valued functions. Positive functions (elements of W_+) are those of the form $\sum_k a_k u_k \ge 0$. Thus, the existence of a measure μ is equivalent to "If $\sum_k a_k \varphi_t(z_k) \ge 0$ for all $t \in \mathbb{S}^1$, then $\sum_k a_k w_k \ge 0$". ■

CHAPTER 7

Asymptotics of the eigenvalues of the Laplacian

1. Weyl's law

The Laplacian, $\Delta := \sum_{i=1}^{n} \partial_i^2$, where $\partial_i = \frac{\partial}{\partial x_i}$, is perhaps the most important linear operator in mathematics. It shows up in many contexts in physics. For example, the law connection the charge distribution $\rho(\cdot)$ to the electirc potential generated by it is $\Delta\varphi = \rho$, the same if ρ is interpreted as mass distribution and φ as the gravitational potential. One could give many other examples. Just to mention two-

(1) Wave equation: $\frac{\partial^2}{\partial t^2} u(x,t) = \Delta u(x,t)$. Here $u(x,t)$ represents the displacement of a stretched membrane (say $x \in \Omega$, a domain in $\mathbb{R}^2$), eg. a drum, where the ends are tied down, $u(x,t) = 0$ for $x \in \partial\Omega$.

(2) Heat equation: $\frac{\partial}{\partial t} u(x,t) = \Delta u(x,t)$, where $u(x,t)$ is the temperature at location x at time t.

In mathematics, the importance of the Laplacian comes from its symmetry with respect to rotations and translations. If $f, g : \mathbb{R}^n \mapsto \mathbb{R}$ are smooth functions such that $g(x) = f(Ax+b)$ where $A_{n\times n}$ is an orthogonal matrix and $b \in \mathbb{R}^n$, then

$$(\Delta g)(x) = (\Delta f)(Ax+b).$$

In other words, the Laplacian commutes with isometries of $\mathbb{R}^n$. It is only natural when a system is described by second order derivatives, and there is symmetry of translation and rotation, that the Laplacian should make an appearance.

Here we are interested in eigenvalues and eigenfunctions of the Laplacian on bounded regions of the Euclidean space. The setting is that we have a nice bounded region $\Omega \subseteq \mathbb{R}^n$ with piecewise smooth boundary, and we consider functions $f : \Omega \mapsto \mathbb{R}$ satisfying $f\big|_{\partial\Omega} = 0$ and some smoothness requirements (eg., C^2) inside Ω. Let us see some examples.

Example 1. $\Omega = (0, L)$ in $\mathbb{R}$. Here $\Delta = \frac{d^2}{dx^2}$. Clearly, $\varphi_n(x) = \sin(\pi n x/L)$ satisfy $\Delta\varphi_n = -\pi^2 L^{-2} n^2 \varphi_n$ and also $\varphi_n(0) = \varphi_n(1) = 0$. We could say that φ_n, $n \geq 1$, are eigenfunctions of the Laplacian on $[0,1]$ with Dirichlet boundary conditions.

There are no other eigenfunctions (we are not saying why, as yet). We see that the nth largest eigenvalue of $-\Delta$ is $\pi^2 n^2 L^{-2}$. Equivalently, if $N(\lambda)$ is the number of eigenvalues not exceeding λ, then $N(\lambda) \sim (L/\pi)\sqrt{\lambda}$.

Example 2. $\Omega = (0,a) \times (0,b)$ in $\mathbb{R}^2$. It is clear that $\varphi_{n,m}(x,y) = \sin(\pi n x/a)\sin(\pi m y/b)$ satisfies $\Delta\varphi_{n,m} = -\pi^2(\frac{n^2}{a^2} + \frac{m^2}{b^2})\varphi_{n,m}$.

What is $N(\lambda)$? It is equal to the number of lattice points (m,n), $m,n \geq 1$, that lie inside the ellipse

$$\frac{x^2}{a^2/\pi^2} + \frac{y^2}{b^2/\pi^2} = \lambda.$$

Hence, $N(\lambda)$ is close to one-quarter of the area of the ellipse, which is $(ab/\pi)\lambda = (|\Omega|/\pi)\lambda$. More precisely, $N(\lambda) \sim (ab/4\pi)\lambda = \frac{1}{(2\pi)^2}|\Omega|\lambda$.

Based on such calculations, and perhaps a few more explicit examples, physicists conjectured (late 1800s) that the asymptotics of eigenvalues depends only on the volume of the domain (and the dimension). That was proved by Weyl. Let ω_d denote the volume of the unit ball in $\mathbb{R}^d$.

Theorem 3 (Weyl). *Let Ω be a domain in $\mathbb{R}^d$ with piecewise smooth boundary. Let $N(\lambda)$ be the number of eigenvalues of $-\Delta$ on Ω, with Dirichlet boundary conditions. Then,*

$$N(\lambda) \sim (2\pi)^{-d}\omega_d|\Omega|\lambda^{d/2}.$$

This is only the most basic version of the theorem, in one setting. It can be extended by finding further corrections. And similar theorems exist for other boundary conditions (eg., Neumann boundary conditions), to related operators (eg., the Schrodinger operator $-\Delta + V$), to the Laplace-Beltrami operator on closed Riemannian manifolds, etc.

2. The spectrum of the Laplacian

There are three ingredients: the domain, the operator and the boundary condition[27].

The domain: We shall assume that Ω is a bounded, open set in $\mathbb{R}^d$ whose boundary is a union of finitely many piecewise smooth closed curves. Let $B = \partial\Omega$. Let $n(x)$ denote the unit outward normal to Ω at $x \in B$ (it exists except at finitely many points).

Boundary conditions: Standard boundary conditions are as follows

(1) Dirichlet: $u = 0$ on B.

(2) Neumann: $\frac{\partial}{\partial n}u = 0$ on B (except at the finitely many points where the normal is not well-defined).

(3) Mixed: Fix a nice (continuous/smooth) function $\sigma : B \mapsto \mathbb{R}$ and ask for $\frac{\partial u}{\partial n} + \sigma u = 0$ on B.

The operator: For $u \in C^2(\Omega)$, we define $\Delta u(x) = \sum_{i=1}^d \partial_i^2 u(x)$ for $x \in \Omega$.

What we want are eigenvalues and eigenfunctions of $-\Delta$. In principle, this must simply be a function $u \in C_c^2(\bar{\Omega})$ (continuous on $\bar{\Omega}$ and smooth in Ω) and a number $\lambda \in \mathbb{R}$ such that $-\Delta u = \lambda u$

[27] This section will be very sketchy, but gives an overview of many important ideas required to make sense of the eigenvalues and eigenfunctions of the Laplacian.

inside Ω and also require that u is not identically zero and that satisfies the boundary conditions imposed.

But as we know, to talk about spectral theorem, we require the setting of a Hilbert space, although the operator need not be defined on all of the space. Also, a reasonable spectral theorem exists only for self-adjoint, or at least normal, operators. What is this Hilbert space for the Laplacian? Although all this can be made sense of, we shall change the setting slightly and work with quadratic forms. First, we introduce the required Hilbert spaces.

A simple integration by parts shows that for $f, g \in C_c(\mathbb{R}^2)$, we have

$$\int_\Omega (-\Delta f)(x)g(x)dx = \int_\Omega \nabla f(x).\nabla g(x)dx.$$

The right side is the required quadratic form, or more precisely, $\int_\Omega |\nabla f|^2$. When we work in a bounded open set Ω, then for $f, g \in C_c^2(\Omega)$, the above identity is still valid. However, if f, g are merely smooth (say on a neighbourhood of $\bar{\Omega}$), then we must be careful about the boundary terms and the identity changes to

$$\int_\Omega (-\Delta f)(x)g(x)dx = \int_\Omega \nabla f(x).\nabla g(x)dx - \int_{\partial\Omega} g\frac{\partial f}{\partial n}.$$

For simplicity of language, let us stick to 2-dimensions. If f satisfies, Neumann boundary condition, then the second term above vanishes. Thus, if f, g are smooth and f satisfies either the Dirichlet or the Neumann boundary condition, then $\int(-\Delta f)g = \int \langle \nabla f, \nabla g \rangle$. This leads us to study the quadratic form

$$Q[f, g] = \int_\Omega \nabla f.\nabla g.$$

What is the right class of functions for which this makes sense? It looks like we must require ∇f to be in L^2. This can be made sense of by the notion of weak derivative.

Weak derivative: If $f, g_i : \mathbb{R}^d \mapsto \mathbb{R}$ are locally integrable functions such that

$$\int f\partial_i g = -\int g_i \varphi$$

for all $\varphi \in C_c^\infty(\mathbb{R}^d)$, then we say that g_i is the weak ith partial derivative of f. If $f \in C^1(\mathbb{R}^d)$, then this is satisfied with $g_i = \partial_i f$, the usual definition of derivative (integration by parts formula). In general, if it exists, it is well defined $a.e.$ If all the weak partial derivative $g_1, \ldots, g_d$ exist, we say that $(g_1, \ldots, g_d)$ is the weak gradient of f.

Now we are ready to define the spaces that we want. Let Ω be a bounded open set in $\mathbb{R}^2$ (for simplicity, stick to $d = 2$ henceforth).

$$H^1(\Omega) = \{f \in L^2(\Omega) : \nabla f \text{ exists in the weak sense and belongs to } L^2(\Omega)\}.$$

On $H^1(\Omega)$, define the inner product $(f, g) = \langle f, g \rangle + \langle \nabla f, \nabla g \rangle$.

Fact: $H^1(\Omega)$ is complete under this inner product.

Define $H^1_0(\Omega)$ to be the closure of $C_c^\infty(\Omega)$ in $H^1(\Omega)$. Then, $H^1_0(\Omega)$ is also a Hilbert space with the same inner product. Functions in H^1_0 are the ones that are meant to satisfy Dirichlet boundary condition.

By the earlier discussion, once we move to the level of quadratic forms (instead of the Laplacian), the boundary condition is no longer required in the Neumann problem. In short, the quadratic form Q, when restricted to $H^1(\Omega)$ and to $H^1_0(\Omega)$, represent the quadratic forms induced by the Laplacian with the Neumann and Dirichlet boundary conditions, respectively.

Definition of eigenvalues and eigenvectors: Define

$$\mu_1 = \min_{f \in H^1, \|f\|=1} Q[f,f].$$

It is clear that $\mu_1 = 0$ (since $Q[f,f] \geq 0$ for all f and $Q[\mathbf{1},\mathbf{1}] = 0$). The minimum is attained by constant functions. Let ψ_1 be one such, normalized in $L^2(\Omega)$. We refer to μ_1 and ψ_1 as the first eigenvalue and the first eigenfunction of the Neumann-Lapacian, respectively.

For $k \geq 1$, let

$$\mu_k = \min_{\substack{f \in H^1, \|f\|=1 \\ f \perp \psi_1, \ldots, \psi_{k-1}}} Q[f,f].$$

It is true, but no longer obvious, that the minimum is attained. Let ψ_k be a minimizer (normalize it in $L^2(\Omega)$). We refer to μ_k and ψ_k as an eigenvalue-eigenfunction pair. Assuming the existence of minimizers, we proceed inductively and obtain $\mu_1 \leq \mu_2 \leq \ldots$ and $\psi_1, \psi_2, \ldots$. By definition, $\{\psi_1, \psi_2, \ldots\}$ is an orthonormal set in $L^2(\Omega)$. Observe also that

$$Q[\psi_k, \psi_j] = \begin{cases} 0 & \text{if } k \neq j, \\ \mu_k & \text{if } k = j. \end{cases}$$

The second is clear by definition of ψ_k and μ_k. The first is also easy (if $j > k$, observe that $Q[\psi_k + t\psi_j, \psi_k + t\psi_j] \leq Q[\psi_k, \psi_k]$ for all t, since $\psi_k + t\psi_j$ is also considered in the minimum. Use that to show that $Q[\psi_k, \psi_j] = 0$).

Another important fact (requires proof) is that $\mu_k \to \infty$ as $k \to \infty$. This ensures that the eigenfunctions form an orthonormal basis for $L^2(\Omega)$ (why?).

We have left two facts unproved: (a) Existence of minimizers and (b) That eigenvalues increase without bound.

In a similar fashion, one can work with the same quadratic form on $H^1_0(\Omega)$ and define $0 < \lambda_1 \leq \lambda_2 \ldots$ and an orthonormal basis $\{\varphi_1, \varphi_2, \ldots\}$ for $L^2(\Omega)$ such that

$$\lambda_k = \min_{\substack{f \in H^1_0, \|f\|=1 \\ f \perp \varphi_1, \ldots, \varphi_{k-1}}} Q[f,f].$$

These are defined to be the eigenvalues of the Dirichlet-Laplacian and the minimizers are the eigenfunctions.

The above definition is essentially the Rayleigh-Ritz formulas that we are familiar with in the case of symmetric matrices. We shall need the min-max theorem (actually max-min theorem, but that sounds odd!) for these eigenvalues.

Theorem 4 (Min-Max theorem)**.** *Let Ω be as above. Then*

$$\mu_k = \max_{\substack{W\subseteq H^1 \\ dim(W)\le k-1}} \min_{\substack{f\in H^1, \|f\|=1 \\ f\perp W}} Q[f,f] \quad \textit{and} \quad \lambda_k = \max_{\substack{W\subseteq H^1_0 \\ dim(W)\le k-1}} \min_{\substack{f\in H^1_0, \|f\|=1 \\ f\perp W}} Q[f,f].$$

3. Proof of Weyl's law using the min-max theorem

Let Ω be as before. Let $N_0(\lambda)$ be the number of Dirichlet eigenvalues in the interval $[0,\lambda]$ and let $N'(\lambda)$ be the number of Neumann eigenvalues in the same interval. Now we are ready to prove Weyl's law. We shall stick to the simplest version of it only.

Theorem 5. *$N(\lambda) \sim (2\pi)^{-d/2}\omega_d|\Omega|\lambda^{d/2}$ and as $\lambda \to \infty$ where $N(\lambda) = N_0(\lambda)$ or $N'(\lambda)$.*

The proof consists of three steps.

(1) Show the theorem for rectangles. This can be done because the eigenvalues of the Laplacian under both Dirichlet and Neumann conditions can be computed explicitly.

(2) Show the theorem for a finite union of standard rectangles. This can be done by comparison theorems using the min-max criteria. The essential point is to show that $N(\lambda)$ is nearly additive in the domain, i.e., $N_{\Omega_1\sqcup\Omega_2}(\lambda) \approx N_{\Omega}(\lambda) + N_{\Omega_2}(\lambda)$. That makes the appearance of $|\Omega|$ transparent.

(3) For a general Ω, sandwich it from inside and outside by regions that are finite unions of standard rectangles. Again invoke comparison theorems.

Remaining notes to be written. No time now!

4. A discrete approach

Let $G = (V,E)$ be a graph with vertex set V and edge set E. We assume that it is undirected, simple, has no loops and is locally finite. The $d(v)$ denote the degree of vertex v and write $u \sim v$ to mean that u and v are connected by an edge. The Laplacian $L : \mathbb{R}^V \to \mathbb{R}^V$ is defined by

$$Lf(v) = d(v)f(v) - \sum_{u:u\sim v} f(u).$$

Let $W \subseteq V$ be a finite set. There are multiple ways to "restrict" the Laplacian to $\mathbb{R}^W$.

(1) *Dirichlet Laplacian:* Define $D_W : \mathbb{R}^W \to \mathbb{R}^W$ by $D_W = L\circ j_W$, where $j_W : \mathbb{R}^W \to \mathbb{R}^V$ extends a function $f : W \to \mathbb{R}$ to a function $j_W(f) : V \to \mathbb{R}$ by defining it to be zero on $V \setminus W$.

(2) *Neumann Laplacian:* Regard W as a graph in its own terms and let N_W denote the corresponding Laplacian from $\mathbb{R}^W$ to $\mathbb{R}^W$.

We write $\ell^2(W)$ for $\mathbb{R}^W$ to emphasize that we endow it with the usual inner product $\langle f,g\rangle = \sum_{u\in W} f(u)g(u)$. It is easy to see that D_W and N_W are symmetric operators on $\ell^2(W)$. In fact, they are positive-semidefinite as shown by the following exercise.

Exercise 6. For $f \in \ell^2(W)$, show that

$$\langle N_W f, f\rangle = \frac{1}{2}\sum_{u,v\in W:u\sim v}(f(u)-f(v))^2,$$

$$\langle D_W f, f\rangle = \frac{1}{2}\sum_{u,v\in W:u\sim v}(f(u)-f(v))^2 + \sum_{u\in W, v\notin W:u\sim v} f(u)^2.$$

The sum is written over ordered pairs u,v, hence the factor of $1/2$. Alternately, we can write both as $\sum_{e=\{u,v\}}(f(u)-f(v))^2$, where the sum is over all edges connecting two vertices of W for the Neumann Laplacian, and over all edges with at least one vertex in W for the Dirichlet Laplacian (in addition to setting the function to be zero outside W).

As symmetric positive-semidefinite operators on a finite dimensional Hilbert space, the spectral theorem tells us that with $n = |W|$, there are $0 \le \lambda_1 \le \ldots \le \lambda_n$ and $0 \le \mu_1 \le \ldots \le \mu_n$ and orthonormal bases $\{\varphi_1,\ldots,\varphi_n\}$ and $\{\psi_1,\ldots,\psi_n\}$ of $\ell^2(W)$ such that

$$D_W = \lambda_1\varphi_1\varphi_1^t + \ldots + \lambda_n\varphi_n\varphi_n^t,$$

$$N_W = \mu_1\psi_1\psi_1^t + \ldots + \mu_n\psi_n\psi_n^t.$$

When necessary, we shall indicate the dependence of λ_i, φ_i etc. on W.

Example 7. Let $G = \mathbb{Z}^d$ (with the usual edge-structure) so $Lf(m) = 2df(m) - \sum_{j=1}^d f(m + e_j) + f(m-e_j)$ for $m \in \mathbb{Z}^d$, where e_j is the jth co-ordinate vector. Observe that if $\varphi_x(m) = \exp\{2\pi i\langle m,x\rangle\}$ for $x \in \mathbb{R}^d$, then

$$L\varphi_x = 2\left(d - \sum_{j=1}^d \cos(2\pi x_j)\right) e_\lambda \;=\; \left(4\sum_{j=1}^d \sin^2(\pi x_j)\right) e_x.$$

Thus φ_x is *formally* an eigenvector of L with eigenvalue $\lambda_x = 4\sum_{j=1}^d \sin^2(\pi x_j)$. Observe that replacing x_j by $-x_j$ does not change the eigenvalue, i.e., for any $\varepsilon.x = (\varepsilon_1 x_1,\ldots,\varepsilon_d x_d)$ where $\varepsilon\in\{-1,1\}^d$. Hence, their linear combinations are also formal eigenfunctions of L.

Let $W = \{1,\ldots,n-1\}^d$. If ψ_x is a formal eigenfunction of L such that $\psi_x(n) = 0$ if one of the n_j is equal to 0 or n, then $D_W\psi_x = \lambda_x\psi_x$. To get such vanishing, consider

$$\psi_x = \sum_{\varepsilon\in\{-1,1\}^d}\prod_{j=1}^d \varepsilon_j\varphi_{\varepsilon.x}$$

Let us recall the following variational characterization of eigenvalues. We leave it as an exercise (or consult a linear algebra book).

Exercise 8. Let $A_{n\times n}$ be a symmetric matrix with eigenvalues $\lambda_1 \leq \ldots \leq \lambda_n$ and corresponding orthonormal eigenvectors $u_1, \ldots, u_n$ (of course, there is a choice in eigenvectors, unless the eigenvalues are distinct). Then

(1) *Rayleigh-Ritz:* $\lambda_k = \min\{\langle Au, u\rangle : \|u\| = 1,\ u \perp \{u_1, \ldots, u_{k-1}\}$ and u_k attains this minimum.

(2) *Max-Min formula:* $\lambda_k = \max\limits_{S:\dim(S)=k-1} \min\{\langle Au, u\rangle : \|u\| = 1,\ u \perp S\}$. The max-min is attained by choosing $S = \text{span}\{u_1, \ldots, u_{k-1}\}$ and $u = u_k$.

Here is a quick illustration of the power of the max-min formula over the Rayleigh-Ritz.

Lemma 9. *If A, B are $n \times n$ real symmetric matrices and $B - A$ is positive semi-definite, then $\lambda_k^A \leq \lambda_k^B$.*

While $\lambda_1^A \leq \lambda_2^A$ following from Rayleigh-Ritz, for $k \geq 2$ it is helpless as the minimization is over different collections of vectors for A and for B.

Proof. Fix any subspace S with dimension $k-1$. As $\langle Au, u\rangle \leq \langle Bu, u\rangle$ for all u, it follows that $\min\{\langle Au, u\rangle : \|u\| = 1,\ u \perp S\} \leq \min\{\langle Bu, u\rangle : \|u\| = 1,\ u \perp S\}$. Take maximum over all S and use the max-min formula to see that $\lambda_k^A \leq \lambda_k^B$. ■

We collect a few observations about the Neumann and Dirichlet eigenvalues. Below, W, W', W_j etc. denote finite non-empty subsets of the vertex set with cardinality n, n', n_j etc.

- From the quadratic forms, it is clear that $D_W \geq N_W$. Hence

(1) $$\lambda_k^W \geq \mu_k^W \qquad \text{for } 1 \leq k \leq n.$$

- If $W \subseteq W'$ with cardinalities n, n', then[28]

(2) $$\lambda_k^{W'} \leq \lambda_k^W \qquad \text{for } 1 \leq k \leq n.$$

  This is because λ_k^W minimizes $\langle Lf, f\rangle$ over functions that vanish outside W, and such functions also vanish outside W'. Minimum over a larger class of functions is smaller. (Caution: This comparison is *not true* for Neumann eigenvalues!)

- Let $W \supseteq W_1 \sqcup \ldots \sqcup W_p$, a pairwise disjoint union. Let μ_k, $k \leq n$, denote the Neumann eigenvalues of W and let μ_k^j, $k \leq n_j$, denote the Neumann eigenvalues of W_j (always the eigenvalues are arranged in increasing order). Now arrange the collection $\{\mu_k^j : 1 \leq j \leq p,\ 1 \leq k \leq n_j\}$ in increasing order as $\mu_1^* \leq \ldots \leq \mu_{n^*}^*$ where $n^* = n_1 + \ldots + n_p$). Then

(3) $$\mu_k \geq \mu_k^* \qquad \text{for } 1 \leq k \leq n^*.$$

  Proof: Let $W_0 = W \setminus (W_1 \sqcup \ldots \sqcup W_p)$. Write $\ell^2(W) = \ell^2(W_0) \oplus \ell^2(W_1) \oplus \ldots \oplus \ell^2(W_p)$ in the obvious way. Then μ_i^* are the eigenvalues of $N^* := N_{W_1} \oplus N_{W_1} \oplus \ldots \oplus N_{W_p}$. Further,

[28] One can avoid saying $k \leq n$, by adopting the convention that $\lambda_k^W = \mu_k^W = +\infty$ for $k > n$.

$N_W \geq N_{W_0} \oplus N^* \geq N^*$ in the sense of positive definite order. The second inequality is clear, the first is seen by comparing the quadratic forms:

$$\langle N_W f, f\rangle - \langle (N_{W_0} \oplus N^*)f, f\rangle = \sum_{j\neq j'} \sum_{W_j \ni u \sim v \in W_{j'}} (f(u) - f(v))^2.$$

From $N_W \geq N^*$ and Lemma 9, we get $\mu_k \geq \mu_k^*$. ■

- Let $W \supseteq W_1 \sqcup \ldots \sqcup W_p$, a disjoint union. Further assume that no edge connects vertices in distinct W_js. Let λ_k, $k \leq n$, denote the Dirichlet eigenvalues of W and let λ_k^j, $k \leq n_j$, denote the Dirichlet eigenvalues of W_j (arranged in increasing order). Now arrange the collection $\{\lambda_k^j : 1 \leq j \leq p,\ 1 \leq k \leq n_j\}$ in increasing order as $\lambda_1^* \leq \ldots \leq \lambda_{n^*}^*$ where $n^* = n_1 + \ldots + n_p$. Then

$$\lambda_k \leq \lambda_k^* \qquad \text{for } 1 \leq k \leq n^*. \tag{4}$$

Observe two differences from the previous situation. The inequality is reversed and we have an extra assumption on the W_js.

Proof: Let $W^* = W_1 \sqcup \ldots \sqcup W_p$. The condition that no edge connects two distinct W_js implies that $D_{W^*} = D_{W_1} \oplus \ldots \oplus D_{W_p}$. Hence λ_j^* are the eigenvalues of D_{W^*}. As $W^* \subseteq W$, from (2) we conclude that $\lambda_k \leq \lambda_k^*$. ■

4.1. **Asymptotics of eigenvalues in $\mathbb{Z}^d$.** We know how to compute Dirichlet and Neumann eigenvalues explicitly for axis-parallel rectangles in $\mathbb{Z}^d$. If $W = [0, N_1] \times [0, N_d]$, then we know that the eigenvalues are of the form

$$4\sum_{j=1}^{d} \sin^2(\pi k_j/N_j)$$

for $k_j \geq 1$. If $N_j \sim a_j R$ where $a_j > 0$ and $R \to \infty$, then the above eigenvalue is asymptotic to $\frac{4\pi^2}{R^2}\sum_{j=1}^d \frac{k_j^2}{a_j^2}$. Thus it makes sense to scale all the eigenvalues up by R^2 to get the number $4\pi^2(k_1^2/a_1^2 + \ldots + k_d^2/a_d^2)$ indexed by $k \in \mathbb{N}^d$. How many of these numbers are in $[0, t]$, if t is large? That is the number of lattice points in the ellipse $x_1^2/a_1^2 + \ldots + x_d^2/a_d^2 \leq t/4\pi^2$, which is a scaling of the ellipse $x_1^2/a_1^2 + \ldots + x_d^2/a_d^2 \leq 1$ by $\sqrt{t}/2\pi$. Therefore (this is intuitively clear, but if not it is proved later) the number of lattice points is asymptotic to $(t/4\pi^2)^{\frac{d}{2}}\omega_d a_1 \ldots a_d$, where the ω_d is the volume of the unit ball in $\mathbb{R}^d$ and $\omega_d a_1 \ldots a_d$ is a volume of the ellipse $x_1^2/a_1^2 + \ldots + x_d^2/a_d^2 \leq 1$.

In summary, if $W_R = \mathbb{Z}^d \cap R\Omega$, where $\Omega = [0, a_1] \times \ldots \times [0, a_d]$ and $R^2\lambda_k^{W_R} \to \lambda_k$ for $k = 1, 2\ldots$ for some numbers $0 \leq \lambda_1 \leq \lambda_2 \leq \ldots$. Further,

$$\#\{k : \lambda_k \leq t\} \sim \kappa_d V(\Omega) t^{\frac{d}{2}}$$

where $\kappa_d = \omega_d(2\pi)^{-d}$ and $V(\Omega)$ is the volume (Lebesgue measure) of Ω. By a similar analysis, $R^2\mu_k^{W_R} \to \mu_k$ for some $0 \leq \mu_1 \leq \mu_2 \leq \ldots$ and

$$\#\{k : \lambda_k \leq t\} \sim \kappa_d V(\Omega) t^{\frac{d}{2}}.$$

We can ask exactly the same question with Ω replaced by another bounded open set in $\mathbb{R}^d$. What happens to the eigenvalues?

Theorem 10 (Weyl's asymptotic law)**.** *Let $\Omega \subseteq \mathbb{R}^d$ be a bounded open set. Let $W_R = \mathbb{Z}^d \cap (R\Omega)$. Then*

(1) $R^2 \lambda_k^R \to \lambda_k(\Omega)$ for each $k \geq 1$, for some numbers $\lambda_1(\Omega) \leq \lambda_2(\Omega) \leq \dots$.

(2) As $t \to \infty$,

$$\#\{k : \lambda_k(\Omega) \leq t\} \sim \kappa_d V(\Omega) t^{-\frac{d}{2}}, \qquad \#\{k : \mu_k(\Omega) \leq t\} \sim \kappa_d V(\Omega) t^{-\frac{d}{2}}.$$

Proof. First consider the case when $\Omega = \Omega_1 \cup \dots \Omega_p$ where each Ω_i is an axis-parallel open rectangle and $\overline{\Omega_i}$ are pairwise disjoint (so Ω is disconnected). Let $W_{i,R} = \mathbb{Z}^d \cap R\Omega_i$. Then for R large enough, $W_{i,R}$ are well-separated from each other and hence $D_{W^R} = D_{W_1^R} \oplus \dots \oplus D_{W_p^R}$. Consequently, the eigenvalues for W^R are just the union of eigenvalues of $W_{i,R}$. From this, both statements follow easily.

Now consider a general bounded open Ω. We can find $\Omega' \subseteq \Omega$ such that Ω' is a union of axis-parallel open rectangles whose closures are disjoint and such that $V(\Omega') \geq (1-\delta)V(\Omega)$. For R large enough,

$$\mu_{k,R}^* \leq \lambda_k^{W_R} \leq \lambda_{k,R}^*$$

where $\mu_{k,R}^*$ (respectively $\lambda_{k,R}^*$) is the increasing enumeration of the union of Neumann eigenvalues (respectively Dirichlet eigenvalues) of $W_{i,R}$, $i \leq p$. ∎

CHAPTER 8

Nevanlinna's value distribution theory and Picard's theorem

Picard's theorem states that in a neighbourhood of an essential singularity, a holomorphic function takes every value among complex numbers with at most one exception. A weaker version is about the range of an entire function in the whole plane. If the function is a non-constant polynomial, it takes every possible value. Otherwise ∞ is an essential singularity and its range omits at most one point (by the stronger statement of Picard, this is true in any neighbourhood of ∞.

There are many proofs of Picard's theorem. Here we present the approach via Nevanlinna's value distribution theory[29].

1. Poisson-Jensen formula

We have stated the Jensen formula before, now we do it in a slightly more general form.

Let f be a meromorphic function on $\mathbb{C}$ and let $a \in \mathbb{C}$. At first, assume that f has neither a pole nor a zero at the origin. Fix $r > 0$ and enumerate the zeros and poles (all non-zero by assumption) of f in $r\mathbb{D}$ as $\zeta_1 \ldots, \zeta_n$ as $\zeta_1, \ldots, \zeta_n$ and $\xi_1, \ldots, \xi_m$, respectively. Then we can write

$$f(z) - a = g(z) \times \prod_{k=1}^{n}(z-\zeta_k)\prod_{\ell=1}^{m}\frac{1}{(z-\xi_\ell)}$$

where g is a meromorphic function with no zeros or poles in $r\mathbb{D}$. Take absolute values, apply logarithm and integrate over $r\mathbb{T}$. By the last thing we shall always mean $\int_{r\mathbb{T}} h = \frac{1}{2\pi}\int_0^{2\pi} h(re^{it})dt$. Use the fact that $\log|g|$ is harmonic in $r\mathbb{D}$ (what if there are zeros or poles on the circle of radius r?) and that $\int_{r\mathbb{T}} \log|\cdot - w| = \log(|w| \vee r)$. Therefore,

$$\begin{aligned}\int_{r\mathbb{T}} \log|f-a| &= \log|g(0)| + \log|c| + \sum_{k=1}^{n}\log r - \sum_{\ell=1}^{m}\log r \\ &= \log|f(0)-a| + \sum_{k=1}^{n}\log\frac{r}{|\zeta_k|} - \sum_{\ell=1}^{m}\log\frac{r}{|\xi_\ell|}.\end{aligned}$$

[29] Good references which we have used and which contain far more material are Hayman's book *Meromorphic functions* (a pre-book version is available online in the TIFR lecture notes series) and Ermenko's lecture notes *Lectures on Nevanlinna theory* (available online). Our presentation follows Eremenko's notes, and streamlined given our limited motivation. Interested readers should read his notes, for the many ideas nicely explained there.

If f has a zero or pole at the origin, write $f(z) - a = c_f(a)z^p + O(z^{p+1})$ as $z \to 0$, apply the above formula to $f_1(z) = (f(z) - a)z^{-p}$ and get

$$\int_{r\mathbb{T}} \log|f-a| = \log|c_f(a)| + p\log r + \sum_{k=1}^{n} \log\frac{r}{|\zeta_k|} - \sum_{\ell=1}^{m} \log\frac{r}{|\xi_\ell|}.$$

We rewrite it in terms of certain counting functions. Let $n_f(r,a)$ denote the number of solutions to $f(z) = a$ for $z \in r\mathbb{D}$, counted with multiplicity. We write $n_f^+(r,a) = n_f(r,a) - n_f(0,a)$. For example, in the above situation, if $p > 0$ then $n_f(r,a) = n + p$ and $n_f(r,\infty) = m$ while if $p < 0$ then $n_f(r,a) = n$ and $n_f(r,\infty) = m + p$. It turns out (in fact as the Poisson-Jensen formula above shows) that it is easier to access a different "counting function", known as the *Nevanlinna counting function*, defined as

$$N_f(r,a) := \sum_{z \in f^{-1}\{a\} \cap r\mathbb{D}} \log\frac{r}{|z|} \quad = \quad n_f(0,a)\log r + \int_0^r \frac{n_f^+(t,a)}{t}dt.$$

The last line is got by writing $\log\frac{r}{|z|}$ as $\int_0^r \frac{\mathbf{1}_{|z|<t<r}}{t}dt$. With this, we arrive at the version of Poisson-Jensen that we want.

$$\int_{r\mathbb{T}} \log|f-a| = \log|c_f(a)| + N_f(r,a) - N_f(r,\infty). \tag{5}$$

It is convenient and easier for understanding to think of the case when $f - a$ has no zeros or poles at the origin. In that case, we can simply write $c_f(a) = f(0) - a$ and also have the less clumsy looking expression $N_f(r,a) = \int_0^r \frac{n_f(t,a)}{t}dt$. The reason why we could not define $N_f(r,a)$ like this in general is that the integral does not converge unless $n_f(0,a) = 0$. For a given f, if 0 is not a pole, then there is at most one $a \in \mathbb{C}$ (namely $a = f(0)$) for which we need the general expression (5), in all other cases we may write $f(0) - a$ in place of $c_f(a)$.

Here is an exercise to show how one may extract information about n_f from N_f, at least for large r.

Exercise 11. Show that $N_f(r,a) - N_f(1,a) \le (n_f(r,a) - n_f(1,a))\log r$ and $n_f(r,a)\log 2 \le N_f(2r,a)$.

Remark 12. In many books, the notation $m(r,f)$ and $N(r,f)$ are used for what we have denoted $m_f(r,\infty)$ and $N_f(r,\infty)$. This is quite reasonable, as $m_f(r,a)$ and $N_f(r,a)$ can then be simply written as $m(r,\frac{1}{f-a})$ and $N(r,\frac{1}{f-a})$.

2. The first fundamental theorem of Nevanlinna

We shall use quite often the positive part of the logarithm, $\log_+ t = \max\{\log t, 0\}$ for $t > 0$. The following elementary properties will be useful.

Exercise 1. For any $z, w \in \mathbb{C}$, show that

(1) $\log_+|z+w| \le \log_+|w| + \log_+|z| + \log 2$.

(2) $\log_+|zw| \le \log_+|z| + \log_+|w|$.

Let f be a meromorphic function and let $a \in \mathbb{C} \cup \{\infty\}$. Introduce the *proximity function*

$$m_f(r,a) := \begin{cases} \int_{r\mathbb{T}} \log_+ |f| & \text{if } a = \infty, \\ \int_{r\mathbb{T}} \log_+ \frac{1}{|f-a|} & \text{if } a \neq \infty. \end{cases}$$

It measures how close f is to a on average over the circle $r\mathbb{T}$. In contrast $n_f(r,a)$ (or $N_f(r,a)$, if loosely interpreted), measure how often f is actually equal to a in the disk $r\mathbb{D}$. The first fundamental theorem of Nevanlinna makes the paradoxical sounding assertion that the only way for f to avoid taking the value a (or to keep $N_f(r,a)$ small) is to come close to a (i.e., increase $m_f(r,a)$) quite often!

The proof is merely a rephrasing of the Poisson-Jensen formula. Write $\log|w| = \log_+|w| - \log_+ \frac{1}{|w|}$ to see that

$$\begin{aligned}\int_{r\mathbb{T}} \log|f-a| &= \int_{r\mathbb{T}} \log_+|f-a| - \int_{r\mathbb{T}} \log_+ \frac{1}{|f-a|} \\ &= m_f(r,\infty) - m_f(r,a) + \left[\int_{r\mathbb{T}} \log_+|f-a| - \log_+|f|\right].\end{aligned}$$

Plug this into (5) and rearrange terms to get

$$m_f(r,\infty) + N_f(r,\infty) = m_f(r,a) + N_f(r,a) + \log|c_f(a)| + \int_{r\mathbb{T}} \log_+|f-a| - \log_+|f|.$$

By the exercise, the integrand in the last summand is at most $\log_+|a| + \log 2$, while the integral is actually an average (we integrate against $dt/2\pi$ over $[0,2\pi)$), hence the same bound holds for the integral. Thus, we arrive at

$$m_f(r,\infty) + N_f(r,\infty) = m_f(r,a) + N_f(r,a) + O(1)$$

where the $O(1)$ term is at most $|\log|c_f(a)|| + \log_+|a| + \log 2$ in absolute value. It remains bounded as $r \to \infty$.

The quantity $T_f(r,a) := m_f(r,a) + N_f(r,a)$ is called the *Nevanlinna characteristic function*. The conclusion above can be rewritten as follows.

Theorem 2 (Nevanlinna's first fundamental theorem)**.** *Let f be a meromorphic function. Then $T_f(r,a) = T_f(r,\infty) + O(1)$ as $r \to \infty$ for any $a \in \mathbb{C}$ (the constants implicit in $O(1)$ do depend on a and on f).*

In summary, we started out wanting to know about $n_f(r,a)$, the number of times the value a is taken by f inside $r\mathbb{D}$. Following where the equations lead, we shifted our goal post to understanding $N_f(r,a)$. Examples show that $N_f(r,a)$ can be wildly different for different values of a. The key insight is Nevanlinna's introduction of the proximity function m_f, which restores balance by capturing whatever goes missing in N_f as a varies. In other words, their sum, T_f is a "conserved quantity" (changes little as a varies).

Example 3. Let f be a polynomial of degree d. Then for large r, we have $|f(z)| = cr^d + O(r^{d-1})$ and hence $m_f(r,\infty) \approx d\log r$, $N_f(r,\infty) = 0$. But for any finite value of a, we have $N_f(r,a) \approx d\log r$

(once all the roots of $f-a$ are inside the disk, their contributions keep growing like $\log r$) while $m_f(r,a)=0$ (since $|f-a|>1$ everywhere on $r\mathbb{T}$ for large r). Thus, we see that as a varies over $\mathbb{C}$, both m_f and N_f are individually conserved, while as a becomes ∞, their contributions switch!

Example 4. Let $f(z)=e^z$. If $a=0$ or $a=\infty$, then $N_f(r,a)=0$ while $m_f(r,0)=\frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} r\cos t dt$ and $m_f(r,\infty)=\frac{1}{2\pi}\int_{\pi/2}^{3\pi/2} r\cos t\, dt$ are both equal to $\frac{r}{\pi}$. Overall $T_f(r,0)=T_f(r,\infty)=\frac{r}{\pi}$. Both 0 and ∞ are special as they are not in the range of f.

Now suppose $a=1$. Then $f-1$ has simple zeros at $2\pi i n$, $n\in\mathbb{Z}$. Hence

$$\begin{aligned} N_f(r,1) &= \sum_{n=-\lfloor r/2\pi\rfloor}^{\lfloor r/2\pi\rfloor} \log\frac{r}{2\pi n} \\ &= \frac{r}{\pi} - \log r + O(1) \end{aligned}$$

by Stirlings' formula (exercise!). To calculate $m_f(r,1)$, we must find $z\in r\mathbb{T}$ for which $|e^z-1|<1$. This happens only in arcs near $\pm ir$ (since $|e^{ir}|=1$). Need to complete and get $m_f(r,1)\approx\log r$.

3. The sphere as $\mathbb{C}\cup\{\infty\}$

The stereographic projection is the correspondence between $\mathbb{C}_\infty := \mathbb{C}\cup\{\infty\}$ and S^2 (here regarded as the sphere of radius $\frac12$ centered at $(0,0,\frac12)$ in $\mathbb{R}^2$) given by

$$z=x+iy\mapsto \hat{z}=\left(\frac{x}{x^2+y^2+1},\frac{y}{x^2+y^2+1},\frac{x^2+y^2}{x^2+y^2+1}\right) \quad\text{and}\quad \infty\mapsto(0,0,1).$$

Its inverse is given by

$$(u,v,w)\mapsto\frac{u+iv}{1-w}.$$

By the chordal distance on the sphere, we mean the Euclidean distance in $\mathbb{R}^3$. We shall write $d(z,w)=\|\hat{z}-\hat{w}\|_{\mathbb{R}^3}$ for $z,w\in\mathbb{C}_\infty$.

Exercise 1. Show that $d(z,w)=\frac{|z-w|}{\sqrt{1+|z|^2}\sqrt{1+|w|^2}}$. By continuity, the right hand side is interpreted as $\frac{1}{\sqrt{1+|z|^2}}$ if $w=\infty$ and $z\neq\infty$.

Exercise 2. Show that rotations of the sphere correspond to linear transformations $z\mapsto\frac{az+b}{-\bar{b}z+\bar{a}}$ on $\mathbb{C}_\infty$, where $a,b\in\mathbb{C}$ with $|a|^2+|b|^2=1$.

Exercise 3. Show that the uniform measure on S^2 (normalized to have total mass 1) corresponds to the measure $d\mu(a)=\frac{dm(a)}{\pi(1+|a|^2)^2}$ on $\mathbb{C}_\infty$ where $dm(a)$ is the Lebesgue measure on $\mathbb{C}$ (omitting a point is irrelevant when considering absolutely continuous measures).

The following lemma will be useful. Let μ denote the Cauchy-like measure in the exercise above.

Lemma 4. $\int_{\mathbb{C}}\log|w-a|d\mu(a)=\frac12\log(1+|w|^2)$.

It is possible to prove this in a pedestrian way by integrating in polar co-ordinates and using the formula for $\int_{r\mathbb{T}} \log|\cdot - a|$. Here is a proof avoiding calculations.

Proof. We claim that for all $w \in \mathbb{C}$,

$$(1) \qquad \int_{\mathbb{C}} \log\left[\frac{|w-a|}{\sqrt{1+|w|^2}\sqrt{1+|a|^2}}\right] d\mu(a) = \int_{\mathbb{C}} \log\frac{1}{\sqrt{1+|a|^2}} d\mu(a).$$

The integrand is the logarithm of the chordal distance, and the measure is the uniform measure on S^2, both invariant under rotations of the sphere. Hence the integral on the left has the same value at $\varphi(w)$ as at w for any conformal automorphism φ of S^2, which implies that the left hand side of (1) is independent of w. Now let $w \to \infty$ and see that the integrand becomes $\log\frac{1}{\sqrt{1+|a|^2}}$, giving the right hand side of (1).

Now, the left hand side of (1) is

$$\int_{\mathbb{C}} \log|w-a| d\mu(a) - \frac{1}{2}\log(1+|w|^2)\int_{\mathbb{C}} d\mu(a) + \int_{\mathbb{C}} \frac{1}{\sqrt{1+|a|^2}} d\mu(a).$$

The last term is the same as right hand side, hence the first two terms must cancel each other. ■

Remark 5. The mathematically correct analogue of Electrostatic potential in two dimensions, induced by a unit charge at a location w is $\log|\cdot - w|$. Its similarity to the inverse distance potential $1/\|\cdot - w\|$ in three dimensions (both in gravitation and electrostatics, up to a question of sign) is that the Laplacian of the potential is (a multiple) of δ_w (interpreted appropriately in distributional sense). In view of this, if we think of a measure μ as a charge distribution, then the potential induced by it is $w \mapsto \int \log|z-w| d\mu(z)$. This is called the Riesz potential of the measure μ and the above lemma computes the Riesz potential of a particular measure. Another example where exact computation can be done is the uniform distribution on the circle $r\mathbb{T}$, whose potential is $\log(|z| \vee r)$, a fact we have used for instance in the proof of Jensen's formula. Riesz potential is a fundamental object in *potential theory*.

4. An exact conservation formula

We now present an observation of Ahlfors and Shimizu that one can modify the proximity function so as to get an exact conservation of its sum with the Nevanlinna counting function. The modification is quite geometric in flavour.

Let $d(z,w) = \frac{|z-w|}{\sqrt{1+|z|^2}\sqrt{1+|w|^2}}$ denote the chordal distance on $\mathbb{C}_\infty$ got by identifying it with S^2 via stereographic projection. We define the modified proximity function

$$\hat{m}_f(r,a) = \int_{r\mathbb{T}} \log\frac{1}{d(f(\cdot),a)}.$$

The chordal distance is always between 0 and 1, hence there is no need to write $\log_+$ here, and $\hat{m}$ is necessarily positive. By the lemma, we know that $\log\frac{1}{d(w,\infty)} = \int_{\mathbb{C}} \log|w-a| d\mu(a)$. Thus we may write

$$\begin{aligned}\hat{m}_f(r,\infty) &= \int_0^{2\pi}\left[\int_{\mathbb{C}}\log|f(re^{i\theta})-a|d\mu(a)\right]\frac{d\theta}{2\pi}\\ &= \int_{\mathbb{C}}\left[\int_{r\mathbb{T}}\log|f(re^{i\theta})-a|\,\frac{d\theta}{2\pi}\right]d\mu(a) \quad \text{(interchange justified?)}.\end{aligned}$$

The inner integral is equal to $\log|f(0)-a| - N_f(r,\infty) + N_f(r,a)$ by Jensen's formula. Integrating over a (w.r.t. μ) gives

$$\begin{aligned}\hat{m}_f(r,a) &= \int_{\mathbb{C}}\log|f(0)-a|d\mu(a) - N_f(r,\infty) + \int_{\mathbb{C}} N_f(r,a)d\mu(a)\\ &= \frac{1}{2}\log(1+|f(0)|^2) - N_f(r,\infty) + \int_{\mathbb{C}} N_f(r,a)d\mu(a)\end{aligned}$$

where we used the lemma again to compute the first term. The first term can be written as $\hat{m}_f(0,\infty)$. The third term is a global average of the Nevanlinna counting function with respect to μ and hence depends only on r. We denote it by $\hat{T}_f(r)$ and call it the modified Nevanlinna characteristic. Then rearranging terms we have

$$\hat{m}_f(r,\infty) + N_f(r,\infty) = \hat{T}_f(r) + \hat{m}_f(0,\infty).$$

Now take any $a \in \mathbb{C}$ and fix φ, a conformal automorphism of S^2 that maps a to ∞. Then if $g := \varphi \circ f$, clearly $m_g(r,\infty) = m_f(r,a)$ and $N_g(r,\infty) = N_g(r,a)$ and $\hat{T}_g(r) = \hat{T}_f(r)$. Consequently, applying the above formula to g, we arrive at

$$\hat{m}_f(r,a) + N_f(r,a) = \hat{T}_f(r) + \hat{m}_f(0,a)$$

In other words, $[\hat{m}_f(r,a) - \hat{m}_f(0,a)] + N_f(r,a) = \hat{T}_f(r)$, a quantity that is independent of a. This is an exact conservation law in contrast to $m_f(r,a) + N_f(r,a)$, which was only conserved up to some additive bounded error.

5. Second fundamental theorem of Nevanlinna

Let $d\nu(z) = \rho(z)d\mu(z)$ be any probability measure on $\mathbb{C}\cup\{\infty\}$ with density ρ with respect to the uniform measure on S^2. Set $\lambda(r) := \int_{r\mathbb{T}} \frac{|f'(\cdot)|^2}{(1+|f(\cdot)|^2)^2}\rho(\cdot)$.

Lower bound for λ: By Jensen's inequality for convex functions,

$$\begin{aligned}\log\lambda(r) &\ge \int_{r\mathbb{T}}\log\left[\frac{|f'(\cdot)|^2}{(1+|f(\cdot)|^2)^2}\rho(\cdot)\right]\\ &= 2\int_{r\mathbb{T}}\log|f'| - 4\int_{r\mathbb{T}}\log\sqrt{1+|f|^2} + \int_{r\mathbb{T}}\log\rho.\end{aligned}$$

By the Poisson-Jensen formula, the first integral is $\log|f'(0)| + N_{f'}(r,0) - N_{f'}(r,\infty)$ while the second integral is equal to $\hat{m}_f(r,\infty)$, which we write as $\hat{T}_f(r) - N_f(r,\infty)$. Therefore,

$$\frac{1}{2}\log\lambda(r) \ge N_{1,f}(r) - 2\hat{T}_f(r) + \frac{1}{2}\int_0^{2\pi}\log\rho(f(re^{i\theta}))\frac{d\theta}{2\pi}$$

where $N_{1,f}(r) = N_{f'}(r,0) + 2N_f(r,\infty) - N_{f'}(r,\infty)$. We claim that $N_{1,f}(r) \geq 0$, which implies that

$$\frac{1}{2}\log\lambda(r) \geq -2\hat{T}_f(r) + \frac{1}{2}\int_0^{2\pi} \log\rho(f(re^{i\theta}))\frac{d\theta}{2\pi}. \tag{2}$$

The non-negativity of $N_{1,f}$ is best seen from the fact that it is a "counting function". Consider the critical points of f and let $n_{1,f}(r)$ denote the number of critical points of f in $r\mathbb{D}$ (counted with multiplicty, as always). Critical points mean zeros of the derivative, but here we should be careful that there can be critical points at the poles of f too, as we shall explain below. We wish to show that $n_{1,f}(r) = n_{f'}(r,0) + 2n_f(r,\infty) - n_{f'}(r,\infty)$.

To see this, observe that if f behaves like z^p near 0 for some $p \geq 1$, then f' behaves like z^{p-1} near zero, hence the order of the critical point is $p-1$. But that is exactly the contribution of the origin to $n_{f'}(0,\infty)$ (while its contribution to $n_f(r,\infty)$ and $nf'(r,\infty)$ is nil). If f behaves like z^{-p} for some $p \geq 1$, we make a change of co-ordinates $g(z) = 1/f(z)$ to see that $g(z)$ looks like z^p near 0, hence we have a critical point of order $p-1$, which can also be written as $2p-(p+1)$. Here p is the order of the pole of f at 0 and $p+1$ is the order of the pole of f' at 0 (while $n_f(r,0)$ gets no contribution from the origin). Taking the two cases together, we see that $n_{1,f}(r) = n_{f'}(r,0) + 2n_f(r,\infty) - n_{f'}(r,\infty)$. In short, $n_{1,f}(r) \geq 0$, as it counts something. Multiply by $1/r$ and integrate (or better, take care of the possibility that the origin is also a critical point) to deduce that $N_{1,f}(r) = \int_0^r \frac{1}{s}n_{1,f}(s)ds$ is also non-negative.

Upper bound for λ: In the case when $\rho = 1$, we have seen that

$$\int_0^s \lambda(u)udu = \frac{1}{2}\int_{s\mathbb{D}} \frac{|f'(z)|^2}{\pi(1+|f(z)|^2)^2}dm(z) = \frac{1}{2}\int_{\mathbb{C}} n_f(s,a)d\mu(a)$$

whence $\int_0^r \frac{1}{s}\int_0^s \lambda(u)udu = \frac{1}{2}\int_{\mathbb{C}} N_f(s,a)d\mu(a)$ which is just $\hat{T}_f(s)$. This motivates us to consider the same quantity for general ρ. Let $G(r) = \int_0^r \frac{1}{s}\left[\int_0^s \lambda(u)udu\right]ds$. Then,

$$\begin{aligned} G(r) &= \int_0^r \frac{1}{s}\int_{s\mathbb{D}} n_f(s,a)\rho(a)\,d\mu(a) \\ &= \int_{r\mathbb{D}} N_f(r,a)\rho(a)\,d\mu(a). \end{aligned}$$

Since $N_f(r,a) \leq \hat{T}_f(r)$ and $\rho(a)d\mu(a)$ has total mass 1, we see that $G(r) \leq \hat{T}_f(r)$ for all r.

We must go from the bound on G to a bound on λ. It is in general not possible, since the derivative can be arbitrarily large on a very short interval without affecting the function much. But that is what we shall show, that the derivative can bounded using the function, except on a small set.

Lemma 6. *If $g : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is an increasing function such that $g(x) \to \infty$ as $x \to \infty$, then the set $E = \{x \in \mathbb{R}_+ : g'(x) > g(x)(\log g(x))^2\}$ has finite Lebesgue measure.*

For any increasing function $\varphi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ such that $\frac{1}{\varphi(t)}dt$ is integrable near infinity, the proof below show that $\{x : g'(x) > \varphi(g(x))\}$ has finite Lebesgue measure. We shall only use the case $\varphi(t) = t(\log t)^2$.

Proof. By the well-known idea of Chebyshev, we write

$$m(E) \le \int_E \frac{g'(x)}{g(x)(\log g(x))^2}dx \;\le\; \int_0^\infty \frac{g'(x)}{g(x)(\log g(x))^2}dx \;=\; \int_{g(0)}^\infty \frac{1}{t(\log t)^2}dt$$

which is finite (no need to worry about the case $g(0) = 0$, since we can omit a bounded interval and apply the above reasoning on $[a, \infty)$ where $g(a) > 0$). Therefore $m(E)$ is finite. ■

Applying the Lemma to G, we see that the set $E = \{r : G'(r) \ge \hat{T}_f(r)(\log \hat{T}_f(r))^2\}$ has finite Lebesgue measure. But $G'(r) = \frac{1}{r}\int_0^r \lambda(u)udu$. Apply the Lemma again, this time to $rG'(r)$ (which is clearly increasing in r) to see that $F = \{r : r\lambda(r) \ge rG'(r)(\log(rG'(r)))^2\}$ has finite Lebesgue measure. Clearly, for all $r \notin E \cup F$, we have

$$\lambda(r) \le \hat{T}_f(r)(\log \hat{T}_f(r))^2(\log r + \log \hat{T}_f(r) + 2\log\log \hat{T}_f(r))^2$$

In summary, $\lambda(r) \le 10\hat{T}_f(r)(\log r + \log \hat{T}_f(r))^3$ outside a set of r of finite Lebesgue measure.

A specific choice of ρ: Putting the upper and lower bounds on $\lambda(r)$, we see that

$$(3) \qquad -2\hat{T}_f(r) + \frac{1}{2}\int_0^{2\pi} \log \rho(f(re^{i\theta}))\frac{d\theta}{2\pi} \le C(\log \hat{T}_f(r) + \log r) \quad \text{for } r \notin E'$$

where E' has finite Lebesgue measure. Now we make a specific choice of ρ. The goal is to get a control on $\sum_{k=1}^q m_f(r, a_k)$ where $a_1, \dots, a_q$ are fixed distinct points in $\mathbb{C} \cup \{\infty\}$.

Apparently the first to do this was F. Nevanlinna (not R. Nevanlinna whose theory we are discussing but his brother!) and his idea was to take ρ to be the (density of) hyperbolic measure on $\mathbb{C} \setminus \{a_1, \dots, a_q\}$. What this means is that ρ must satisfy (here Δ is the Laplacian)

$$\Delta\rho(z) = -\rho(z) \qquad \text{or is it } \Delta\left[\frac{\rho(z)}{(1+|z|^2)^2}\right] = -\frac{\rho(z)}{(1+|z|^2)^2}??$$

for $z \in \mathbb{C} \setminus \{a_1, \dots, a_q\}$. However, the solution is not given by an explicit formula, and one must approach it indirectly. Ahlfors simplified the approach by using an explicit density that is explicit and easy to analyse, but retains the essential features of the Hyperbolic measure, mainly its behaviour near the points $a_1, \dots, a_q$ (it is not necessary here to know that it has these properties, the proof is self-contained).

Fix $\beta > 0$ and let $\rho(z) = C_\beta \prod_{k=1}^{q} \frac{1}{d(z,a_k)^\beta}$, where the constant C_β is chosen so that $\int \rho d\mu = 1$. To make this normalization possible, we assume that $\beta < 2$ (note that $\frac{1}{|z|^\beta}$ is integrable in a disk around the origin in $\mathbb{C}$ if and only if $\beta < 2$). Then

$$\int_0^{2\pi} \log \rho(f(re^{i\theta}))\frac{d\theta}{2\pi} = \beta\sum_{k=1}^n \hat{m}_f(r, a_k) + \log C_\beta.$$

Plugging this into (3), and rearranging, we arrive at

$$\sum_{k=1}^{q} \hat{m}_f(r, a_k) \leq \frac{4}{\beta}\hat{T}_f(r) + O(\log \hat{T}_f(r) + \log r) \tag{4}$$

except for a set of finite measure of r. Since β can be taken arbitrarily close to 2, we deduce that

$$\sum_{k=1}^{q} \hat{m}_f(r, a_k) \leq 2\hat{T}_f(r)(1 + o(1)) \quad \text{for } r \notin E'. \tag{5}$$

The $o(1)$ term is unspecified here. If one is more careful, one can get the error term as before. We state this as a theorem.

Theorem 7 (Second fundamental theorem)**.** *Let f be a meromorphic function on $\mathbb{C}$ and let $a_1, \ldots, a_q$ be distinct points in $\mathbb{C} \cup \{\infty\}$. Then, $\sum_{k=1}^{q} \hat{m}_f(r, a_k) \leq 2\hat{T}_f(r) + S_f(r)$ where $S_f(r) = O(\log \hat{T}_f(r) + \log r)$ for r outside a subset of finite Lebesgue measure.*

For the applications below, our less precise statement (5) is sufficient. But first let us remark how the more precise form can be obtained. If we were allowed to set $\beta = 2$, then (4) would immediately give the second fundamental theorem. We cannot set $\beta = 2$ because ρ would then be not integrable. However, we can introduce logarithmic corrections to make it integrable. More precisely, set

$$\rho(z) = C \prod_{k=1}^{q} \frac{1}{d(z, a_k)^2 \, (\log d(z, a_k))^2}.$$

The $\log d(z, a_k)$ terms ensure integrability, but are mild enough that their contributions to the term $\int_0^{2\pi} \log \rho(f(re^{i\theta}))d\theta$ can be absorbed into the error terms. We do not give these details here.

6. Picard's theorem

Let f be a meromorphic function whose range misses three points $a_1, \ldots, a_q$ in $\mathbb{C} \cup \{\infty\}$. Then $N_f(r, a_k) = 0$ for all $r > 0$ and $k \leq q$. Therefore, $\hat{m}_f(r, a_k) = \hat{T}_f(r)$. But then the second fundamental theorem forces that $q\hat{T}_f(r) \leq 2\hat{T}_f(r)(1 + o(1))$ (except for a finite Lebesgue measure set of r). Since $\hat{T}_f(r) \to \infty$ as $r \to \infty$, this forces $q \leq 2$. In particular, the range of an entire function (which already misses ∞) can miss at most one point in the complex plane. This is Picard's theorem!

More generally, define the defect of a point a as $\delta_f(a) := \liminf\limits_{r \to \infty} \frac{\hat{m}_f(r, a_k)}{\hat{T}_f(r, a_k)}$. Then for any meromorphic function

$$\sum_{a \in \mathbb{C}_\infty} \delta_f(a) \leq 2.$$

This means that for most values of a, we must have that $N_f(r, a)$ is almost the whole of $\hat{T}_f(r)$.

CHAPTER 9

Discrete harmonic functions

Bounded and positive harmonic functions on $\mathbb{Z}^d$ and other graphs. Buhovsky-Sodin-Logunov-Malinnikova result.