

DAVENPORT · 圆法 · 高中详解版

# Waring 问题的历史

*Waring's problem: history*

**本章要解决什么 / 读完能掌握什么** 本章是全书的“历史与背景”引子，没有任何困难的证明，但\*\*埋下了后面所有章节要用到的核心记号与思想\*\*。读完你应当能回答这些问题：

1. 什么是 **Waring 问题**？它在问“每个正整数能不能写成  $s$  个  $k$  次方之和”。
2. 两个核心函数  $g(k)$  与  $G(k)$  各是什么意思，区别在哪。
3. Hardy 与 Littlewood 的**圆法**（circle method）大致是怎么把“数表法个数”这件纯算术的事，变成一个**积分**去算的——也就是公式 (2.7)：
$$r(N) = \int_0^1 (T(\alpha))^s e(-N\alpha) d\alpha$$
是怎么来的。
4. 什么是**渐近公式** (2.2)、什么是**奇异级数**  $\mathfrak{S}(N)$ 、为什么它们能证明“大数一定可表示”。

本页把译文按段落顺序逐段讲，凡是原书写得简略、跳步、或默认你“早就会”的地方，全部从高中知识从零补全。公式里出现的  $\sum$ （求和号）、 $\int_0^1$ （定积分）、 $e^{2\pi i\alpha}$ （复指数）、 $O(\cdot)$ 、 $\Gamma$ （伽马函数）、同余等，第一次出现时都会单独讲清。

阅读方式：黑色正文是**译文原句**（忠实于 Davenport 原书）；带颜色的框（目标／符号／分步推演／图）是为**高中生加的详解**，逐步推演、举例、配图，不用比喻、不省略。

## 第1段 Waring 的原始陈述与问题 (2.1)

Edward Waring 在他的《代数沉思录》(*Meditationes algebraicae*, 1770) 中陈述道：每个数都可表示为 4 个平方数之和，或 9 个立方数之和，或 19 个四次方数之和，“依此类推”。由最后这句话可推知，他想要断言的是：对于每个  $k \geq 2$ ，都存在某个  $s$ ，使得每个正整数  $N$  可表示为

$$N = x_1^k + x_2^k + \cdots + x_s^k, \quad (2.1)$$

其中  $x_i \geq 0$ 。

**这段在讲什么** 本段交代问题的来历：250 多年前一位英国数学家 Waring 凭直觉写下一句话，后人把它精确化为一个数学命题 (2.1)。本段不证明任何东西，只是把“要研究的问题”说清楚。

先把 Waring 那三句具体话翻成现代语言，逐条看：

- “每个数都可表示为 4 个平方数之和”：任何正整数  $N$ ，都能写成  $N = x_1^2 + x_2^2 + x_3^2 + x_4^2$ ，其中  $x_i$  是非负整数（即  $0, 1, 2, 3, \dots$ ）。这里  $k = 2$ （平方就是 2 次方）， $s = 4$ 。例如  $7 = 2^2 + 1^2 + 1^2 + 1^2$ ， $6 = 2^2 + 1^2 + 1^2 + 0^2$ 。这条后来叫**四平方和定理**（Lagrange 1770 年证明）。
- “或 9 个立方数之和”： $k = 3$ （立方=3 次方）， $s = 9$ 。任何  $N$  都能写成至多 9 个非负整数的立方之和。例如  $23 = 2^3 + 2^3 + 1^3 + 1^3 + 1^3 + 1^3 + 1^3 + 1^3 + 1^3$ （一个 8、一个 8，再加 7 个 1，共 9 项）。
- “或 19 个四次方数之和”： $k = 4$ （“biquadrate” 就是四次方）， $s = 19$ 。

**符号  $N = x_1^k + x_2^k + \cdots + x_s^k$  怎么读**

- $N$ ：要被表示的那个正整数（被表示的“目标数”）。

- $k$ : 幂次 (指数), 是固定的,  $k \geq 2$ 。  $k = 2$  是平方,  $k = 3$  是立方, 等等。  $x^k$  表示  $x$  自乘  $k$  次。
- $s$ : 用到的项数, 也就是“用几个  $k$  次方加起来”。
- $x_1, x_2, \dots, x_s$ : 这  $s$  个底数, 每个都是非负整数  $x_i \geq 0$ 。下标  $i$  只是给它们编号 (第 1 个、第 2 个……第  $s$  个)。
- 省略号  $\dots$ : 表示中间一样的项都照写, 一共  $s$  项。

注意允许  $x_i = 0$ : 因为  $0^k = 0$ , 加上一个 0 不改变和, 这等价于说“至多  $s$  个非零的  $k$  次方”。后文有时要求  $x_i \geq 1$  (不许取 0), 那是另一种计数口径, 下文会专门区分。

为什么“依此类推”这句话这么关键? Waring 只写了  $k = 2, 3, 4$  三个具体例子 (对应 4, 9, 19), 但他加了一句“and so on (依此类推)”。Davenport 指出: 正是这句话, 让后人推断 Waring 想断言的是一个对所有  $k \geq 2$  都成立的普遍命题

**Waring 断言 (精确版)** 对每个整数  $k \geq 2$ , 都存在一个整数  $s$  (这个  $s$  可以依赖  $k$ , 不同的  $k$  用不同的  $s$ ), 使得每一个正整数  $N$  都能写成  $s$  个非负整数的  $k$  次方之和, 即形如 (2.1)。

**“存在  $s$ ”与“每个  $N$ ”的顺序不能颠倒** 这句断言的逻辑结构是: 先固定  $k$ , 再找一个  $s$  (一个数), 让这个  $s$  对所有  $N$  都管用。关键在于  $s$  只依赖  $k$ , 不依赖  $N$ ——同一个  $s$  要同时对  $N = 1, 2, 3, \dots$  一直到无穷大都成立。如果允许  $s$  随  $N$  变大而变大, 那这命题就毫无内容了 (任何  $N$  都能写成  $N$  个  $1^k$  之和, 取  $s = N$  即可)。难点正在于“一个固定的  $s$  通吃所有  $N$ ”。这就是为什么这是个深刻问题。

$$N = x_1^k + x_2^k + \cdots + x_s^k$$

共  $s$  项，每个  $x_i \geq 0$ ； $k$  固定， $s$  只跟  $k$  有关，对所有  $N$  通用

例 ( $k=2, s=4$ ):

$$7 = 2^2 + 1^2 + 1^2 + 1^2$$

$$6 = 2^2 + 1^2 + 1^2 + 0^2$$

Waring 问题的核心对象：把任意正整数  $N$  拆成固定个数  $s$  个  $k$  次方之和。

## 第 2 段 Hilbert 的证明、Hurwitz 的铺垫，与 $g(k)$

这个断言首先由 Hilbert 于 1909 年证明。Hilbert 的证明是一项非常伟大的成就，尽管其中部分功劳也应归于 Hurwitz，他的工作提供了出发点。Hurwitz 此前已经证明：若该断言对某个指数  $k$  成立，则它对  $2k$  也成立。通常用  $g(k)$  表示使得每个  $N$  都可表示的最小  $s$  值。如今  $g(k)$  的确切值对所有  $k$  值都已知晓。我在此不讨论 Hilbert 的证明方法；关于这一点，可参阅 Stridsberg [79]、Schmidt [72] 或 Rieger [71] 的论文。

**这段在讲什么** 讲历史进展：从 1770 年 Waring 提出，到 1909 年 Hilbert 证明它“真的成立”，中间有 139 年没人能证。这段还引入了第一个核心记号  $g(k)$ 。

**历史脉络（为什么这是大事）。** Waring 1770 年只是“断言”，没有证明。这个断言看起来简单，证起来极难：你要找一个固定的  $s$ ，保证**无穷多个**  $N$  全部能被表示，不能有任何漏网之鱼。整整 139 年，数学界都没能对一般的  $k$  证明它（只有零星几个特殊的  $k$  被攻下，例如  $k=2$  的四平方和 1770 年由 Lagrange 解决）。直到 1909 年，David Hilbert（当时世界顶尖数学家）才证明：对每个  $k \geq 2$ ，这样的  $s$  确实存在。这就是为什么 Davenport 称之为“非常伟大的成就”。

**Hurwitz 的归纳一步：“ $k$  成立  $\Rightarrow 2k$  成立”是什么意思** Hurwitz 证明了一条**递推性质**：假设我们已经知道 Waring 断言对幂次  $k$  成立（即存在某个  $s$  能把所有数表成  $s$  个  $k$  次方之和），那么它对幂次  $2k$  也自动成立。这在逻辑上是“**把指数翻倍**”的一步。它的用处：如果你能从某个起点（比如  $k = 2$  已知成立）出发，靠翻倍得到  $k = 4, 8, 16, \dots$ 。但光靠翻倍并不能覆盖所有  $k$ （得不到  $k = 3, 5, 6, 7, \dots$ ），所以它只是“提供了出发点 / 铺垫”，真正把所有  $k$  一网打尽的是 Hilbert。这正是 Davenport 说“部分功劳归 Hurwitz，但主要成就是 Hilbert”的原因。

**核心记号  $g(k)$ ：最小的可行项数** 对固定的  $k$ ，能让“每个正整数  $N$  都可表示成  $s$  个  $k$  次方之和”的  $s$  有很多（ $s$  越大越容易满足，因为多出来的项可以全取 0）。我们关心其中**最小的那个**，记作  $g(k)$ ：

$$g(k) = \min\{s : \text{每个正整数 } N \text{ 都可写成 } s \text{ 个 } k \text{ 次方之和}\}.$$

读作“ $g$  在  $k$  处的值”。Waring 的三个例子其实就是在猜测：

$$g(2) = 4, \quad g(3) = 9, \quad g(4) = 19.$$

（这三个值后来都被严格证明是对的。）Hilbert 定理保证  $g(k)$  是一个**有限**的数（而不是无穷大），这正是断言成立的含义。译文末句说“如今  $g(k)$  的确切值对所有  $k$  已知”——也就是这个最难的常数问题，现在基本算清楚了。

**为什么  $g(2) = 4$  不能再小（即 3 个平方不够）** 要说明“最小是 4”，需要两件事：(i) 4 个平方够用（Lagrange 定理）；(ii) 3 个平方**不够**——必须找到一个数它写不成 3 个平方和。取  $N = 7$ 。三个平方数（含 0）只能取自  $\{0, 1, 4\}$ （因为  $3^2 = 9 > 7$ ）。把任意三个这种数相加，逐一验证最大也凑不出 7：

- 1 用到的平方块只有 0, 1, 4。三块之和的所有可能： $0+0+0 = 0$ ,  
 $\dots$ ,  $4+1+1 = 6$ ,  $4+4+0 = 8$ （超了）,  $4+1+0 = 5$ ,

$$1+1+1 = 3, \quad 4+4+4 = 12 \cdots \cdots$$

2 逐个验证发现：能凑出的  $\leq 7$  的值是  $0, 1, 2, 3, 4, 5, 6, 8, \dots$ ，唯独凑不出 7。

3 所以 7 需要第 4 个平方： $7 = 2^2 + 1^2 + 1^2 + 1^2$ 。这证明  $s = 3$  不够，故  $g(2) \geq 4$ ；配合 Lagrange 的  $g(2) \leq 4$ ，得  $g(2) = 4$ 。



这也顺便说明了  $g(k)$  的“脆弱”：它的大小往往被个别特别难表示的小数（如这里的 7）顶上去。这一点正是下文要引入  $G(k)$  的动机。

### 第 3 段 Hardy–Littlewood 的工作与渐近公式 (2.2)

Hardy 与 Littlewood 的工作出现在一系列题为“关于数的分拆”（“On Partitio Numerorum”，简记 P.N.）的若干篇论文中，该系列的其余论文主要关注 Goldbach 的三素数问题。在 P.N. I [37] 中，他们得到了关于  $r(N)$ （即  $N$  表示为形如 (2.1) 且  $x_i \geq 1$  的表法数目）的一个渐近公式，只要  $s \geq s_0(k)$ （ $k$  的某个确切函数）即成立。该渐近公式具有如下形式：

$$r(N) = C_{k,s} N^{s/k-1} \mathfrak{S}(N) + O(N^{s/k-1-\delta}), \quad (2.2)$$

其中  $\delta > 0$  且

$$C_{k,s} = \frac{\Gamma(1 + 1/k)^s}{\Gamma(s/k)} > 0.$$

**这段在讲什么** 从“能不能表示”（是非题）升级到“有多少种表示法”（计数题）。Hardy 与 Littlewood 不满足于知道  $N$  能表示，他们要算出表示法的

个数  $r(N)$  大约是多少——这就是渐近公式 (2.2)。这一段是全章技术上的核心预告。

**背景：P.N. 系列与 Goldbach 问题** “Partitio Numerorum” 是拉丁文“数的分拆”。Hardy 与 Littlewood (20 世纪初英国剑桥的两位大数学家，长期合作) 发表了一系列以此命名的论文，编号 P.N. I, II, IV, VI, VIII……方括号里的 [37][38] 等是原书参考文献的编号。这套方法 (后来叫圆法, circle method) 不仅用于 Waring 问题，也用于 **Goldbach 三素数问题** (每个充分大的奇数是三个素数之和)。换句话说，本章学的圆法是一套**通用武器**，Waring 只是它的一个应用战场。

**核心记号  $r(N)$ ：表示法的个数**  $r(N)$  表示：把  $N$  写成形如 (2.1) 的方式，且要求每个  $x_i \geq 1$  (注意这里不许取 0)，一共有多少种有序写法。“有序”指：  $(x_1, x_2, \dots, x_s)$  看成一个有顺序的列，调换顺序算作不同的表示。例如  $k=2, s=2, N=5$ :  $5 = 1^2 + 2^2 = 2^2 + 1^2$ ，这两种被算作  $r(5) = 2$  (如果还允许更多就再加)。为什么从“是非”改成“数个数”？因为**只要能证明  $r(N) > 0$ ，就说明  $N$  至少有一种表示，从而  $N$  可表示**。所以“算出  $r(N)$  大约多大、并说明它是正的”这件强得多的事，自动顺带解决了原来的可表示性问题。这就是 Hardy-Littlewood 的思路：用更强的**定量**结果去碾压原来的**定性**问题。

**“渐近公式”是什么意思 (高中视角从零讲)** “渐近”指  $N$  很大时的近似行为。一个**渐近公式**就是：把一个难算的量 (这里是  $r(N)$ ) 写成“一个主项 (main term, 能写成漂亮的初等表达式) + 一个误差项 (error term, 比主项小得多)”。(2.2) 正是这种形式：

$$\underbrace{r(N)}_{\text{真实个数}} = \underbrace{C_{k,s} N^{s/k-1} \mathfrak{S}(N)}_{\text{主项: 能写出来的近似值}} + \underbrace{O\left(N^{s/k-1-\delta}\right)}_{\text{误差: 保证比主项小}}.$$

只要误差项确实比主项小（这由那个  $-\delta$  保证，见下方  $O$  记号说明），主项就“统治”了  $r(N)$  的大小，于是我们就近似地知道了  $r(N)$ 。

**$O(\cdot)$  (大  $O$ ) 记号：从零讲起**  $O(f(N))$  读作“大  $O$  的  $f(N)$ ”，意思是“某个量的绝对值不超过常数倍的  $f(N)$ ”。准确地说，写

$$A(N) = O(f(N))$$

表示：存在一个与  $N$  无关的正常数  $C$ ，使得当  $N$  充分大时  $|A(N)| \leq C f(N)$ 。它是用来控制误差大小的简写，不关心误差的精确值，只关心它“最多有多大”。本章还会用到与之配套的  $\ll$  记号（Vinogradov 记号）： $A \ll B$  与  $A = O(B)$  完全同义，即存在常数  $C$  使  $|A| \leq C B$ 。例如“ $x^2 \ll x^3$ （当  $x \geq 1$ ）”是对的。

为什么误差项  $N^{s/k-1-\delta}$  确实比主项  $N^{s/k-1}$  小 主项的“个头”是  $N$  的  $\frac{s}{k} - 1$  次方；误差项是  $N$  的  $\frac{s}{k} - 1 - \delta$  次方。两者相除：

$$\frac{\text{误差}}{\text{主项}} \asymp \frac{N^{s/k-1-\delta}}{N^{s/k-1}} = N^{-\delta} = \frac{1}{N^{\delta}}.$$

因为  $\delta > 0$ ，当  $N \rightarrow \infty$  时  $N^{-\delta} \rightarrow 0$ ，所以这个比值  $N^{-\delta} \rightarrow 0$ 。也就是说误差相对主项可以忽略，主项越来越准。这个正的小数  $\delta$ （读作 delta）就是为了“拉开主项与误差的次方差距”而设的，它有多大不重要，只要  $> 0$  即可。

**主项里的各个因子** 主项  $C_{k,s} N^{s/k-1} \mathfrak{S}(N)$  由三部分相乘：

$N^{s/k-1}$ （个头因子）

这是主项随  $N$  增大的增长速度。指数  $\frac{s}{k} - 1$  来自一个直观的“维数 / 体积”估计：每个  $x_i$  大约能取到  $0 \sim N^{1/k}$ （因为  $x_i^k \leq N$ ）， $s$  个变量给出大约  $(N^{1/k})^s = N^{s/k}$  种组合，再被“和恰好等于  $N$ ”这一个约束“除掉一维”，于是  $N^{s/k}/N = N^{s/k-1}$ 。后面的章节会把这个直觉变成严格推导。

$C_{k,s}$ （一个正的常数因子）

只依赖  $k$  与  $s$ ，与  $N$  无关，下面单独讲它的公式。

### $\mathfrak{S}(N)$ (奇异级数)

一个反映“同余条件”的算术因子，随  $N$  而变，下一段专门讲。

常数  $C_{k,s} = \frac{\Gamma(1 + 1/k)^s}{\Gamma(s/k)}$ ：先认识  $\Gamma$  (伽马函数) 高中学过阶乘

$n! = 1 \cdot 2 \cdots n$ ，但它只对正整数  $n$  有定义。伽马函数  $\Gamma$  (读“gamma”) 是阶乘对所有正实数的“连续推广”，满足

$$\Gamma(n) = (n-1)! \quad (n \text{ 为正整数}), \quad \Gamma(x+1) = x\Gamma(x).$$

比如  $\Gamma(1) = 0! = 1$ ,  $\Gamma(2) = 1! = 1$ ,  $\Gamma(4) = 3! = 6$ 。它还能取小数自变量，比如  $\Gamma(1 + \frac{1}{k})$  里  $1 + \frac{1}{k}$  一般不是整数，这正需要  $\Gamma$  的连续推广才有意义。这里你不必会算  $\Gamma$  的具体值，只需知道两点：(1) 它是阶乘的推广；(2) 对正自变量

它取正值，所以  $C_{k,s} = \frac{\Gamma(1 + 1/k)^s}{\Gamma(s/k)} > 0$  (分子分母都是正数)。这个

$C_{k,s} > 0$  很关键——它保证主项是正的，为下文“ $r(N) > 0 \Rightarrow$ 可表示”埋下伏笔。

## 第 4 段 奇异级数 $\mathfrak{S}(N)$ 与它的正下界 (2.3)

在上述公式中， $\mathfrak{S}(N)$  是一个纯算术性质的无穷级数，Hardy 与 Littlewood 称之为奇异级数 (singular series)。他们进一步证明了

$$\mathfrak{S}(N) \geq \gamma > 0, \quad (2.3)$$

其中  $\gamma$  与  $N$  无关，只要  $s \geq s_1(k)$ 。然而他们在该阶段并未给出  $s_1(k)$  的任何确切值。于是该公式蕴含着

$$r(N) \sim C_{k,s} N^{s/k-1} \mathfrak{S}(N) \quad (2.4)$$

当  $N \rightarrow \infty$ , 只要  $s \geq \max(s_0(k), s_1(k))$ , 从而独立地给出了 Hilbert 定理的一个证明。

**这段在讲什么** 引入主项里最神秘的因子——**奇异级数**  $\mathfrak{S}(N)$ , 并说明: 只要它有一个**不随  $N$  跑掉的正下界**, 整个渐近公式就能**独立地**重新证明 Hilbert 定理。这是本段的高潮。

**什么是“级数”（无穷级数），先从零讲** 高中学过有限项相加。**无穷级数**就是“无穷多项相加”，写成  $\sum_{q=1}^{\infty} a_q = a_1 + a_2 + a_3 + \cdots$ , 它的值定义为前  $n$  项部分和  $a_1 + \cdots + a_n$  当  $n \rightarrow \infty$  的极限（若极限存在，就说级数**收敛**）。奇异级数  $\mathfrak{S}(N)$  就是这样一个无穷和，它的每一项都是由**同余信息**算出来的纯算术量（所以叫“纯算术性质”）。本章只需把它当作主项里一个“与同余有关的正因子”，它的精确定义与求值放在后面的专门章节。

**为什么需要引入奇异级数？它解决了什么困难** 单凭“个头”  $N^{s/k-1}$  还不足以正确预测  $r(N)$ , 因为表示法个数会受**整除 / 同余**的影响。举个直觉例子：如果某个  $N$  模某个数  $q$  的余数“天生”就难凑成  $s$  个  $k$  次方之和（局部障碍），那它的表示法就会偏少甚至为 0。奇异级数  $\mathfrak{S}(N)$  正是把**所有模  $q$  的局部信息打包成一个因子**：

- 若某个  $N$  存在“局部障碍”（某个模  $q$  根本无解），则  $\mathfrak{S}(N) = 0$ , 主项消失，确实表示不了；
- 若处处无障碍，则  $\mathfrak{S}(N)$  是一个正数，主项为正。

这就是它叫“奇异（singular，本义为‘特殊 / 主导’）级数”的来历：它精确地编码了表示法个数对同余条件的依赖。它解决的困难是：**把全局的“能否表示、有几种”问题，化约为一族容易的“模  $q$  局部问题”**。

**同余 / 模运算，从零讲（理解奇异级数所需）** “ $a$  与  $b$  模  $q$  同余”，记作  $a \equiv b \pmod{q}$ , 意思是  $a - b$  能被  $q$  整除，即  $a, b$  除以  $q$  余数相同。例如

$17 \equiv 2 \pmod{5}$  (都余 2)。研究“ $N$  能否写成  $s$  个  $k$  次方之和”时，常先看“在模  $q$  意义下能否凑出  $N$  的余数”——这就是“局部问题”，奇异级数把每个  $q$  的局部解的多少汇总起来。

**不等式  $\mathfrak{S}(N) \geq \gamma > 0$  (编号 2.3) 在说什么**  $\gamma$  (读“gamma”，这里是个小正数，与上文  $\Gamma$  函数无关，只是同名希腊字母) 是一个不依赖  $N$  的正的下界。这句话保证：只要项数  $s$  足够大 ( $s \geq s_1(k)$ )，奇异级数永远~~不会~~贴着 0，而是被一个固定的正数  $\gamma$  从下方顶住。为什么这点至关重要？因为如果  $\mathfrak{S}(N)$  可以随  $N$  任意接近 0，主项  $C_{k,s} N^{s/k-1} \mathfrak{S}(N)$  就可能被误差项淹没，从而无法断定  $r(N) > 0$ 。有了统一正下界  $\gamma$ ，主项至少有  $C_{k,s} N^{s/k-1} \gamma$  这么大，稳稳压过误差，于是  $r(N) > 0$  对所有大  $N$  成立。

**两个门槛  $s_0(k)$  与  $s_1(k)$ ，以及  $\max$  的意思**

$s_0(k)$

使渐近公式 (2.2) 本身成立所需的项数门槛 (误差能压住所需)。

$s_1(k)$

使奇异级数正下界 (2.3) 成立所需的项数门槛。

$\max(s_0(k), s_1(k))$

两者中的较大者。要同时用到 (2.2) 和 (2.3)，就得让  $s$  同时跨过两个门槛，所以取较大的那个。

“ $s_1(k)$  在该阶段没有确切值”是说：他们只证明了“存在这样的门槛”，但当时没算出它具体多大——这是后续工作要改进的地方。

**渐近等价号  $\sim$  (编号 2.4) 从零讲**  $A(N) \sim B(N)$  (读作“ $A$  渐近等于  $B$ ”) 的定义是

$$\lim_{N \rightarrow \infty} \frac{A(N)}{B(N)} = 1.$$

意思是  $N$  很大时两者**比值趋于 1**，即相对误差趋于 0（但绝对差可以很大）。

(2.4) 就是把 (2.2) 丢掉误差项后的“主项主宰”版本：

$$r(N) \sim C_{k,s} N^{s/k-1} \mathfrak{S}(N)。$$

推演：为什么 (2.2)+(2.3) 就能独立证明 Hilbert 定理 原文说“从而独立地给出了 Hilbert 定理的一个证明”，这一步原书没展开，我们补全：

1 取  $s \geq \max(s_0(k), s_1(k))$ ，则 (2.2) 与 (2.3) 同时可用。

2 由 (2.3)，主项满足  $C_{k,s} N^{s/k-1} \mathfrak{S}(N) \geq C_{k,s} N^{s/k-1} \gamma$ ，这是一个正的、随  $N$  增大而增大的量（因为  $C_{k,s} > 0$ ， $\gamma > 0$ ，且当  $s > k$  时指数  $\frac{s}{k} - 1 > 0$ ）。

3 由 (2.2)， $r(N) = \text{主项} + O(N^{s/k-1-\delta})$ 。误差项相对主项是  $N^{-\delta} \rightarrow 0$ （前面已算），所以存在  $N_0$ ，当  $N \geq N_0$  时误差的绝对值小于主项的一半。

4 于是当  $N \geq N_0$ ：  
$$r(N) \geq \text{主项} - |\text{误差}| > \text{主项} - \frac{1}{2} \text{主项} = \frac{1}{2} \text{主项} > 0。$$

5  $r(N) > 0$  意味着  $N$  至少有一种  $s$  个  $k$  次方的表示，故每个  $N \geq N_0$  都可表示。再加上  $\setminus(N$  ◆

“独立地”一词的含义：这条证明完全走**解析（积分 / 级数）**路线，跟 Hilbert 当年的纯代数证明思路毫无关系，是对同一个定理的**另起炉灶**的新证明。

## 第 5 段 记号 $G(k)$ ，以及对它的上界

Hardy 与 Littlewood 引入记号  $G(k)$  表示使得每个充分大的  $N$  都可表示为形如 (2.1) 的最小  $s$  值；这个函数实际上比  $g(k)$  更具意义，因为后者受到表示某一两个特殊数  $N$  的难度的影响。在 P.N. II [38] 和 P.N. IV [39] 中，Hardy 与 Littlewood 证明了渐近公式以及  $\mathfrak{S}(N)$  的下界两者都对  $s \geq (k-2)2^{k-1} + 5$  成立，这蕴含着

$$G(k) \leq (k-2)2^{k-1} + 5.$$

**这段在讲什么** 引入第二个、也是圆法真正关心的核心记号  $G(k)$ ，并解释它为什么比  $g(k)$  更“干净”、更有意义；最后给出 Hardy-Littlewood 得到的第一个上界。

**核心记号  $G(k)$ ：只管“充分大的  $N$ ”**

$$G(k) = \min\{s : \text{每个充分大的 } N \text{ 都可写成 } s \text{ 个 } k \text{ 次方之和}\}.$$

“充分大的  $N$ ”指：存在某个界  $N_0$ ，对所有  $N \geq N_0$  都可表示（允许有限多个小的  $N$  例外）。这和  $g(k)$  的唯一区别就在这里：

$g(k)$

要求**每一个**正整数（含所有小数）都可表示，一个都不能漏。

$G(k)$

只要求**所有足够大**的正整数可表示，少数顽固的小数可以网开一面。

因此恒有  $G(k) \leq g(k)$ （要求更宽松，需要的项数不会更多）。

**为什么  $G(k)$  “比  $g(k)$  更有意义”** 回看前面  $g(2) = 4$  的讨论： $g(k)$  的大小常常被个别又小又难表示的数顶高，这些“钉子户”是局部的、偶然的现象，跟  $k$  次方和的本质规律关系不大。最有名的例子是  $k = 3$ ：实际上只有 23 和 239 这两个数需要用满 9 个立方，其余所有数 8 个甚至更少就够。正因为 23, 239 这两个“特例”， $g(3) = 9$  被顶得很高；但一旦放过这有限几个小数，所需项数立刻下降——这正是  $G(3)$  想刻画的。所以  $G(k)$  过滤掉了这些偶然的小数障碍，反

映的是  $k$  次方和的渐近本质，这就是 Davenport 说它“更具意义”的原因，也是圆法（专攻大  $N$  的渐近行为）真正能发力的对象。

$g(k)$ : 覆盖所有  $N$  (含所有小数, 一个不漏)



$G(k)$ : 只管充分大的  $N$  (放过左边有限个小数)



$g(k)$  被左边个别难表示的小数顶高； $G(k)$  放过有限个小数，只看大  $N$  的渐近规律，因此  $G(k) \leq g(k)$  且更本质。

推演：为什么“渐近公式 + 正下界对  $s$  成立”就给出  $G(k) \leq s$  原书直接从“两个结论对  $s \geq (k-2)2^{k-1} + 5$  成立”跳到  $G(k) \leq (k-2)2^{k-1} + 5$ ，这一步逻辑补全：

- 1 设某个具体的  $s$  (这里  $s = (k-2)2^{k-1} + 5$ ) 同时满足渐近公式 (2.2) 与正下界 (2.3)。
- 2 按上一段“独立证明 Hilbert”的推演，对这个  $s$ ，存在  $N_0$  使每个  $N \geq N_0$  都有  $r(N) > 0$ ，即都可表示成  $s$  个  $k$  次方之和。
- 3 “每个充分大的  $N$  可表示成  $s$  个  $k$  次方之和”正是  $G(k) \leq s$  的定义 (因为  $G(k)$  是最小的能做到这件事的项数，它当然  $\leq$  任何一个已被验证能做到的  $s$ )。
- 4 代入  $s = (k-2)2^{k-1} + 5$  即得  $G(k) \leq (k-2)2^{k-1} + 5$ 。 ◆

这个上界长什么样（代入小  $k$  感受一下） $(k-2)2^{k-1}+5$  随  $k$  指数级增长（含  $2^{k-1}$  因子），增长很快：

1  $k=2$ :  $(2-2) \cdot 2^1 + 5 = 0 + 5 = 5$ ，即  $G(2) \leq 5$ 。（真值其实是  $G(2) = 4$ ，可见此上界不紧。）

2  $k=3$ :  $(3-2) \cdot 2^2 + 5 = 4 + 5 = 9$ ，即  $G(3) \leq 9$ 。

3  $k=4$ :  $(4-2) \cdot 2^3 + 5 = 16 + 5 = 21$ ，即  $G(4) \leq 21$ 。（下一段会改进到 19。）

4  $k=5$ :  $(5-2) \cdot 2^4 + 5 = 48 + 5 = 53$ 。可见随  $k$  增大上界迅速变大，这正是后人努力改进的动机。

## 第 6 段 更好的上界 $G(4) \leq 19$ ，与 P.N. VIII

在 P.N. VI [40] 中，他们对  $G(k)$  找到了一个更好的上界（尽管对渐近公式的成立性并非如此），特别地，他们证明了  $G(4) \leq 19$ 。该系列的最后一篇论文 P.N. VIII [41] 则完全关注奇异级数以及由它引出的同余问题。

**这段在讲什么** 说明一条重要的方法论：求  $G(k)$  的上界，不一定非得证明完整的渐近公式，可以用更省力的办法压低上界（这一思想第 9 段会展开）。并补充 P.N. VIII 的主题。

**“更好的上界，但对渐近公式并非如此”怎么理解** 这句话里藏着一个关键区分，是后面圆法的灵魂：

证明渐近公式（强）

要算出  $r(N)$  (全部表示法的精确个数) 的近似值, 技术要求高, 需要的  $s$  较大。

证明  $G(k) \leq s$  (弱)

只需说明  $r(N) > 0$  (哪怕只有一种表示也行), 不必精确计数, 技术要求低, 能容许更小的  $s$ 。

所以 P.N. VI 用了某种不依赖完整渐近公式的技巧, 把  $G(4)$  的上界从上一段的 21 压到了 19——但这个更小的  $s$  已经不足以撑起完整的渐近公式了。一句话: “证明能表示”比“数清表示法个数”要求低, 所以能用更少的项数。

**为什么  $G(4)$  特别有名 (四次方的局部障碍)** 四次方 ( $k = 4$ ) 有一个出名的同余障碍: 任何整数的四次方模 16 只能是 0 或 1。

1 偶数  $x = 2m$ :  $x^4 = 16m^4 \equiv 0 \pmod{16}$ 。

2 奇数  $x$ : 可验证  $x^4 \equiv 1 \pmod{16}$  (例如  $1^4 = 1, 3^4 = 81 = 16 \cdot 5 + 1, 5^4 = 625 = 16 \cdot 39 + 1$ )。

3 所以  $s$  个四次方之和模 16 的余数, 就等于其中“奇四次方”的个数模 16。形如  $16n + 15$  的数模 16 余 15, 要让奇四次方的个数  $\equiv 15 \pmod{16}$ , 至少需要 15 个“奇四次方”, 于是这类数至少要 15 项——这已说明  $G(4) \geq 15$ 。

4 要再逼到 16, 需要更细一类数  $16^t \cdot 31$  (它们可任意大, 恰对应“充分大的  $N$ ”, 且可证必须用满 16 个四次方), 由此  $G(4) \geq 16$ 。这类同余分析也解释了为什么  $G(4)$  的值卡在十几这个量级。

(如今已知精确值  $G(4) = 16$ , 但那是更晚的成果; Hardy-Littlewood 当年得到的是上界 19。)这个例子让你具体看到“同余问题”如何直接限制

$G(k)$ ，也就呼应了下一句。

**P.N. VIII 为什么“完全关注奇异级数与同余”** 前面说过，奇异级数  $\mathfrak{S}(N)$  把所有模  $q$  的局部信息打包；而像上面  $G(4)$  的 mod 16 障碍这种事，正是“同余问题”。Hardy–Littlewood 在系列最后一篇专门研究  $\mathfrak{S}(N)$  何时为正、何时为 0（即何时存在局部障碍），这等价于研究同余方程

$x_1^k + \cdots + x_s^k \equiv N \pmod{q}$  是否有解、有多少解。这是奇异级数的算术内核，本章只点到为止，细节留待后续章节。

## 第 7 段 出发点：母函数（生成函数）

Hardy 与 Littlewood 以  $r(N)$  的母函数（生成函数）作为他们的出发点，也就是幂级数

$$\sum_{N=0}^{\infty} r(N) z^N = \left( \sum_{n=0}^{\infty} z^{n^k} \right)^s.$$

**这段在讲什么** 揭示圆法的第一块基石——生成函数。这是把“数论计数问题”翻译成“函数 / 级数问题”的关键一步：把要算的所有  $r(N)$  一次性装进一个幂级数里。

**求和号  $\sum$  从零讲  $\sum$** （希腊字母大写 Sigma，读“西格玛”）是“连加”的简写。例如

$$\sum_{n=0}^{\infty} a_n = a_0 + a_1 + a_2 + \cdots, \quad \sum_{x=1}^P b_x = b_1 + b_2 + \cdots + b_P.$$

下标  $n = 0$ （或  $x = 1$ ）告诉你从哪开始加，上标  $\infty$ （或  $P$ ）告诉你加到哪结束， $\sum$  右边是“通项”（第  $n$  项长什么样）。

**什么是“生成函数 / 母函数”？为什么要引入它** 我们有无穷多个要研究的数  $r(0), r(1), r(2), \dots$ 。生成函数的想法是：给每个  $r(N)$  配一个“标签”  $z^N$  ( $z$  是一个形式变量)，把它们全部加成一个幂级数：

$$F(z) = \sum_{N=0}^{\infty} r(N)z^N = r(0) + r(1)z + r(2)z^2 + \dots$$

这样，“ $z^N$  前的系数”就是  $r(N)$ 。一个函数  $F(z)$  就把无穷多个数论量整体打包带走了。引入它的动机：单看  $r(N)$  一个个地算很难；但打包成函数后，我们可以用函数的运算（乘法、积分）来一次性操控所有  $r(N)$ ，把组合 / 数论问题转化为分析问题。这正是“解析方法”的精髓。

推演：为什么  $\sum r(N)z^N = \left(\sum_{n \geq 0} z^{n^k}\right)^s$  (原书直接给出，这里补全)

1 先看单个因子

$g(z) = \sum_{n=0}^{\infty} z^{n^k} = z^{0^k} + z^{1^k} + z^{2^k} + \dots = 1 + z + z^{2^k} + z^{3^k} + \dots$ 。  
它的特点：每个“ $k$  次方指数”  $n^k$  恰好贡献一项  $z^{n^k}$ ，系数都是 1。

2 把  $s$  个这样的因子相乘： $g(z)^s = \underbrace{g(z) \cdot g(z) \cdots g(z)}_{s \uparrow}$ 。展开时，要从

每个因子中各挑一项相乘。第  $i$  个因子挑出  $z^{x_i^k}$  (对应某个非负整数  $x_i$ )， $s$  个相乘得

$$z^{x_1^k} \cdot z^{x_2^k} \cdots z^{x_s^k} = z^{x_1^k + x_2^k + \cdots + x_s^k}.$$

这里用了同底数幂相乘指数相加： $z^a \cdot z^b = z^{a+b}$ 。

3 于是  $g(z)^s$  展开后，每一种挑法  $(x_1, \dots, x_s)$  贡献一个  $z^{x_1^k + \cdots + x_s^k}$ 。把指数恰好等于  $N$  的项归并到一起：指数为  $N$  的项有多少个？正是满足  $x_1^k + \cdots + x_s^k = N$  的有序非负整数解  $(x_1, \dots, x_s)$  的个数。

- 4 按定义，这个解的个数就是  $z^N$  前的系数，记作  $r(N)$ （这里口径是  $x_i \geq 0$  的表法数）。所以

$$g(z)^s = \sum_{N=0}^{\infty} r(N) z^N.$$

合并指数、按  $N$  归类——这就是生成函数“自动完成计数”的魔力：  
多项式 / 幂级数相乘，等于把各项指数相加，于是系数自动统计出“和  
等于  $N$  的方案数”。◆

每个因子里挑一项：  $z^{x_1^k} \cdot z^{x_2^k} \cdot \dots \cdot z^{x_s^k}$

指数相加  $\rightarrow z^{x_1^k + x_2^k + \dots + x_s^k}$

把“指数恰好 =  $N$ ”的项收集起来：

$x_1^k + \dots + x_s^k = N$  的方案

$\rightarrow$  共  $r(N)$  个

$z^N$  前的系数 =  $r(N)$

结论：  $(\sum z^{(n^k)})^s = \sum r(N) z^N$  —— 幂级数相乘自动完成“数表法个数”

生成函数的关键机制：幂级数相乘 = 指数相加，于是  $z^N$  的系数自动等于“和为  $N$  的表示法个数”  $r(N)$ 。

## 第 8 段 Cauchy 的系数公式与 Vinogradov 的简化

他们借助 Cauchy 给出幂级数系数的公式，利用沿圆周  $|z| = \rho$ （其中  $\rho$  略小于 1）取的围道积分，把  $r(N)$  用此函数表示出来。Vinogradov 在 1928 年作出了一项有益的技术简化；它在于用一个有限的指数和来代替幂级数，其效果是消除了 Hardy 与 Littlewood 原始论述中出现的若干不重要的复杂之处。

**这段在讲什么** 讲圆法名字的由来（“沿圆周积分”），以及 Vinogradov 把无穷幂级数换成有限指数和的关键简化——正是这个简化，让下文 (2.5)–(2.7)

那套更干净的写法成为可能。本章后续就用 Vinogradov 的版本。

**怎样从生成函数  $F(z)$  反取出系数  $r(N)$ ？这就是 Cauchy 公式** 有了

$F(z) = \sum_N r(N)z^N$ ，怎么单独把第  $N$  个系数  $r(N)$  抠出来？复变函数论里的 Cauchy 系数公式说：

$$r(N) = \frac{1}{2\pi i} \oint_{|z|=\rho} \frac{F(z)}{z^{N+1}} dz,$$

其中  $\oint_{|z|=\rho}$  表示让  $z$  沿复平面上半径为  $\rho$  的圆周绕一圈做积分（这叫围道积分 contour integral）。这一步你不需要会复变积分，只需理解它的作用：它是

“从打包好的函数里挑出某一个系数”的提取器。“圆法”（circle method）的名字正是来自这条‘沿圆周积分’。

**为什么取  $|z| = \rho$  且  $\rho$  略小于 1，而不是恰好 = 1？** 这是一个“取舍”的细节，原书一笔带过，补全理由：幂级数  $\sum r(N)z^N$  当  $|z|$  太大时可能发散（加不出有限值），所以不能随便取大的半径；它在  $|z| < 1$  内部收敛。取  $\rho$  稍微小于 1，是为了让积分路径落在“级数收敛、函数有意义”的安全区内，同时又尽量贴近  $|z| = 1$ ——因为决定  $r(N)$  大小的关键信息都集中在单位圆周  $|z| = 1$  附近（尤其是  $z$  接近某些特殊根处）。所以“略小于 1”是“既要收敛安全、又要贴近主战场”的折中。

**Vinogradov 1928 的简化：把无穷幂级数换成有限指数和** Hardy-Littlewood 原版用无穷幂级数  $\sum_{n=0}^{\infty} z^{n^k}$ ，加上  $\rho \rightarrow 1$  的极限处理，技术上很啰嗦。

Vinogradov 的改良是：

- 不让  $n$  跑到无穷，而是截断成有限项：只取  $x = 1, 2, \dots, P$ （ $P$  是一个有限的正整数）；
- 不在  $|z| = \rho < 1$  的圆上积分，而是直接令  $z = e^{2\pi i \alpha}$  落在单位圆  $|z| = 1$  上，让  $\alpha$  在  $[0, 1]$  上跑。

这样无穷级数变成**有限和**  $T(\alpha) = \sum_{x=1}^P e(\alpha x^k)$  (就是下一段的 (2.5)), 围道积分变成  $[0, 1]$  上的普通定积分 (2.7)。好处：有限和处处收敛、不用担心发散，也不必处理  $\rho \rightarrow 1$  的极限，从而“消除了若干不重要的复杂之处”。代价仅是要选一个合适的截断长度  $P$  (下段会看到  $P = [N^{1/k}]$  即可)，这是非常划算的简化，所以本章及全书都采用 Vinogradov 的版本。

## 第 9 段 记号 $e(t)$ 与生成函数 $T(\alpha)$ (公式 2.5)

记  $e(t) = e^{2\pi i t}$ 。对于实变量  $\alpha$ ，我们定义  $T(\alpha)$  为

$$T(\alpha) = \sum_{x=1}^P e(\alpha x^k), \quad (2.5)$$

其中  $P$  是一个正整数。

**这段在讲什么** 正式登场两个贯穿全书的记号：复指数简写  $e(t)$ ，以及 Vinogradov 版的有限指数和  $T(\alpha)$  (公式 2.5)。这是圆法的“主角函数”。

**复数与“复数的模长”，从零讲（看懂  $e^{2\pi i t}$  的前置知识）** 高中接触过虚数单位  $i$ ，满足  $i^2 = -1$ 。一个复数写成  $z = a + bi$  ( $a, b$  实数)，可看成平面上的点  $(a, b)$ 。它的模长（绝对值）是它到原点的距离

$$|z| = \sqrt{a^2 + b^2}.$$

模长衡量复数的“大小”。后面会反复用到它来估计  $T(\alpha)$  的大小。

**复指数  $e(t) = e^{2\pi i t}$  从零讲：单位圆上转圈的点** 这是全书最重要的记号，务必吃透。由欧拉公式  $e^{i\theta} = \cos \theta + i \sin \theta$ ，令  $\theta = 2\pi t$ ：

$$e(t) := e^{2\pi i t} = \cos(2\pi t) + i \sin(2\pi t).$$

几何意义： $e(t)$  是复平面单位圆（半径 1 的圆）上的一个点，对应转过的角度为  $2\pi t$  弧度。关键性质：

### 模长恒为 1

$|e(t)| = \sqrt{\cos^2(2\pi t) + \sin^2(2\pi t)} = 1$ 。所以它永远在单位圆上跑，不会变大变小。

### 周期为 1

$t$  每增加 1，角度增加  $2\pi$ （整整一圈），点回到原处： $e(t+1) = e(t)$ 。这就是为什么后文积分区间取长度恰为 1 的  $[0, 1]$ ——一个周期。

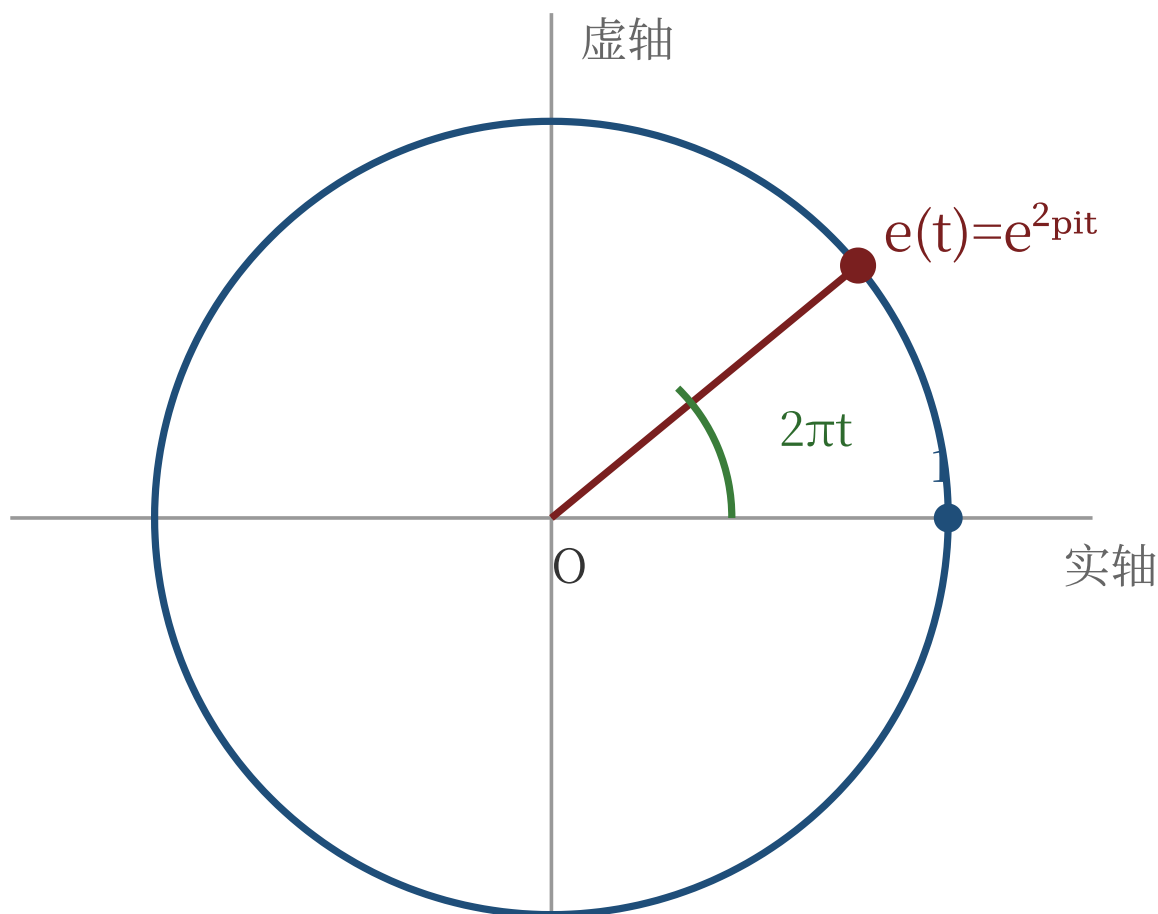
### 相乘 = 角度相加

$e(s)e(t) = e^{2\pi i s}e^{2\pi i t} = e^{2\pi i(s+t)} = e(s+t)$ 。这条让“指数相加”变得极方便，下段计算  $(T(\alpha))^s$  全靠它。

$$e(0) = 1, e(\text{整数}) = 1$$

当  $t$  是整数  $m$  时， $e(m) = \cos(2\pi m) + i \sin(2\pi m) = 1$ 。这条是下一段“积分挑系数”的命门。

为什么要用  $e(t)$  这个简写？因为  $e^{2\pi i t}$  写起来长，而且  $2\pi$  这个因子在圆法里到处都是；用  $e(t)$  能让公式干净得多。



模长  $|e(t)| = 1$  (始终在圆上)  
 $t$  增加 1  $\rightarrow$  转一整圈回到原处 (周期 1)

$e(t) = e^{2\pi i t}$  是单位圆上的点，角度  $2\pi t$ ，模长恒为 1，周期为 1。整数  $t$  时落在 1 处。

**生成函数  $T(\alpha) = \sum_{x=1}^P e(\alpha x^k)$  (公式 2.5) 逐符号拆解**

$\alpha$

读“阿尔法”，一个**实变量**，把它想成在  $[0, 1]$  上滑动的旋钮。 $T$  是  $\alpha$  的函数。

$x$

求和指标，依次取  $1, 2, 3, \dots, P$ 。

$x^k$

底数  $x$  的  $k$  次方 ( $k$  是固定的幂次)。

$$e(\alpha x^k) = e^{2\pi i \alpha x^k}$$

单位圆上角度为  $2\pi\alpha x^k$  的点。每个  $x$  给一个这样的点。

$$\sum_{x=1}^P$$

把  $x = 1$  到  $P$  的这  $P$  个单位圆上的点（复数）相加。

$P$

截断长度（有限的正整数），即只用到底数不超过  $P$  的  $k$  次方。下段说明取  $P = [N^{1/k}]$  就够。

所以  $T(\alpha)$  是把  $P$  个“单位圆上的点”  $e(\alpha \cdot 1^k), e(\alpha \cdot 2^k), \dots, e(\alpha \cdot P^k)$  相加得到的一个复数，它随  $\alpha$  变化。它就是 Vinogradov 版的生成函数——对应上一段那个无穷幂级数因子  $\sum z^{n^k}$ ，只是把  $z$  换成  $e(\alpha)$ 、并截断到  $P$  项。

## 第 10 段 展开 $(T(\alpha))^s$ （公式 2.6）

则

$$(T(\alpha))^s = \sum_m r'(m) e(m\alpha), \quad (2.6)$$

其中  $r'(m)$  表示  $m$  表示为

$$x_1^k + \dots + x_s^k, \quad (1 \leq x_i \leq P)$$

的表法数目。

**这段在讲什么** 把主角函数  $T(\alpha)$  自乘  $s$  次，展开后自动统计出表示法个数。这是第 7 段“幂级数相乘 = 计数”的 Vinogradov 翻版，但现在是有有限和、用  $e(m\alpha)$  代替  $z^m$ 。

推演：为什么  $(T(\alpha))^s = \sum_m r'(m) e(m\alpha)$

（原书只给结果，这里完整展开）

- 1 写开  $s$  次幂为  $s$  个相同因子相乘，每个因子都用不同的求和指标 ( $x_1, \dots, x_s$ ) 以示区别：

$$(T(\alpha))^s = \left( \sum_{x_1=1}^P e(\alpha x_1^k) \right) \left( \sum_{x_2=1}^P e(\alpha x_2^k) \right) \cdots \left( \sum_{x_s=1}^P e(\alpha x_s^k) \right).$$

- 2 乘开括号：从每个因子里各挑一项相乘，所有挑法求和。即对所有有序组  $(x_1, \dots, x_s)$  (每个分量从 1 到  $P$ ) 求和：

$$(T(\alpha))^s = \sum_{x_1=1}^P \sum_{x_2=1}^P \cdots \sum_{x_s=1}^P e(\alpha x_1^k) e(\alpha x_2^k) \cdots e(\alpha x_s^k).$$

- 3 用上一段“相乘 = 角度相加”的性质  $e(a)e(b) = e(a+b)$ ，把  $s$  个  $e$  合并：

$$e(\alpha x_1^k) \cdots e(\alpha x_s^k) = e(\alpha x_1^k + \cdots + \alpha x_s^k) = e(\alpha (x_1^k + \cdots + x_s^k))$$

这里把公因子  $\alpha$  提了出来。

- 4 记每一组的“和”为  $m = x_1^k + \cdots + x_s^k$ ，则该项就是  $e(m\alpha)$ 。于是整个和按  $m$  的值归类：指数（这里指  $e(\cdot\alpha)$  中的  $m$ ）等于同一个  $m$  的项，应当合并。

- 5 合并后， $e(m\alpha)$  前面的系数 = “有多少组  $(x_1, \dots, x_s)$  满足  $x_1^k + \cdots + x_s^k = m$  且每个  $1 \leq x_i \leq P$ ”。把这个个数记作  $r'(m)$ 。于是

$$(T(\alpha))^s = \sum_m r'(m) e(m\alpha).$$

求和指标  $m$  跑遍所有可能出现的和（从最小  $s \cdot 1^k = s$  到最大  $s \cdot P^k$ ）。



$r'(m)$  与前面  $r(N)$  的区别：那一撇是什么意思  $r'(m)$ （读“r prime m”）和  $r(N)$  都是“表法个数”，但限制不同：

$r(N)$

表示成  $s$  个  $k$  次方之和、 $x_i \geq 1$ ，底数不设上限。

$r'(m)$

同样要  $x_i \geq 1$ ，但底数被截断在  $1 \leq x_i \leq P$ （因为  $T(\alpha)$  只加到  $P$ ）。

那一撇正是提醒你：这是“带截断  $P$ ”的计数。下一段要做的，就是说明只要  $P$  取得够大，截断不丢解， $r'(N)$  就等于  $r(N)$ 。

$$(T(\alpha))^s = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_s} e(\alpha x_1^k) \cdot e(\alpha x_2^k) \cdots e(\alpha x_s^k)$$

相乘=角度相加  $\rightarrow e(\alpha \cdot (x_1^k + x_2^k + \cdots + x_s^k))$

令  $m = x_1^k + \cdots + x_s^k \rightarrow$  每组贡献一个  $e(m\alpha)$

按  $m$  归类： $e(m\alpha)$  的系数 = 满足和为  $m$  ( $1 \leq x_i \leq P$ ) 的组数 =  $r'(m)$

$$(T(\alpha))^s = \sum_m r'(m) e(m\alpha)$$

与生成函数同理： $(T(\alpha))^s$  展开后， $e(m\alpha)$  的系数自动等于“和为  $m$  的（截断）表示法个数”  $r'(m)$ 。

**第 11 段 选  $P$  使  $r'(N) = r(N)$ ，并用积分挑出  $r(N)$ （公式 2.7）**

若  $P \geq [N^{1/k}]$ （其中  $[\lambda]$  表示实数  $\lambda$  的整数部分），则  $r'(N)$  就是  $N$  表示为形如 (2.1) 且  $x_i \geq 1$  的表法总数。因此  $r'(N) = r(N)$ 。若我们

将 (2.6) 的两边乘以  $e(-N\alpha)$  并在单位区间  $[0, 1]$  上（或任何长度为 1 的区间上）积分，便得到

$$r(N) = \int_0^1 (T(\alpha))^s e(-N\alpha) d\alpha. \quad (2.7)$$

**这段在讲什么** 这是全章的技术高潮：两件事。(1) 选对截断  $P$ ，让截断计数  $r'(N)$  等于真计数  $r(N)$ ；(2) 用一个“积分提取系数”的技巧（正交关系），把  $r(N)$  写成漂亮的定积分 (2.7)——这就是后面所有章节真正动手计算的出发点。

**取整记号  $[\lambda]$  与条件  $P \geq [N^{1/k}]$**   $[\lambda]$ （读“ $\lambda$  的整数部分”， $\lambda$  读“lambda”）表示不超过  $\lambda$  的最大整数（向下取整）。例如  $[3.7] = 3$ ,  $[5] = 5$ ,  $[2.0] = 2$ 。为什么取  $P \geq [N^{1/k}]$  就能保证不丢解？关键观察：在  $x_1^k + \cdots + x_s^k = N$  且各  $x_i \geq 1$  的表示里，每个底数  $x_i$  都不会太大。

1 因为其余各项  $\geq 0$ （实际  $\geq 1$ ），单看一项有

$$x_i^k \leq x_1^k + \cdots + x_s^k = N。$$

2 两边开  $k$  次方： $x_i \leq N^{1/k}$ 。

3 而  $x_i$  是整数，整数且  $\leq N^{1/k}$ ，就一定  $\leq [N^{1/k}]$ （不超过  $N^{1/k}$  的最大整数）。

4 所以只要截断上限  $P \geq [N^{1/k}]$ ，每个真正的表示里的底数都落在  $1 \leq x_i \leq P$  内，一个解都不会被截断丢掉。反过来， $1 \leq x_i \leq P$  的解本来就是合法表示。于是截断计数与真计数完全一致：

$$r'(N) = r(N)。◆$$

这就是  $P$  的取法理由：取得恰好够大 ( $\geq [N^{1/k}]$ ) 以不丢解，又不必更大（更大只是多算些和必然  $> N$  的项，对  $r'(N)$  无影响），干净利落。

**定积分  $\int_0^1$  从零讲（看懂 (2.7) 所需）**  $\int_0^1 f(\alpha) d\alpha$  读作“ $f$  从 0 到 1 的定积分”。直观上它表示函数  $f$  在区间  $[0, 1]$  上的“累积量 / 面积”（对复值函数则是实部、虚部分别累积）。本段只需用到它一条关键性质（线性性）：积分可以逐项进行，且常数可以提到积分号外——

$$\int_0^1 \left( \sum_m c_m g_m(\alpha) \right) d\alpha = \sum_m c_m \int_0^1 g_m(\alpha) d\alpha.$$

也就是“和的积分 = 积分的和”“常数倍照搬”。

核心引理：正交关系  $\int_0^1 e(m\alpha) d\alpha = \begin{cases} 1, & m = 0 \\ 0, & m \neq 0 \end{cases}$  ( $m$  为整数) 这是“积分挑系数”的全部秘密，原书默认你会，这里从零证：

1 情形  $m = 0$ ：  $e(0 \cdot \alpha) = e(0) = 1$ （常函数）。所以  $\int_0^1 1 d\alpha = 1 - 0 = 1$ 。

2 情形  $m \neq 0$ （整数）：此时  $e(m\alpha) = e^{2\pi i m \alpha}$  是个指数函数，对  $\alpha$  求原函数：

$$\int e^{2\pi i m \alpha} d\alpha = \frac{1}{2\pi i m} e^{2\pi i m \alpha} + C,$$

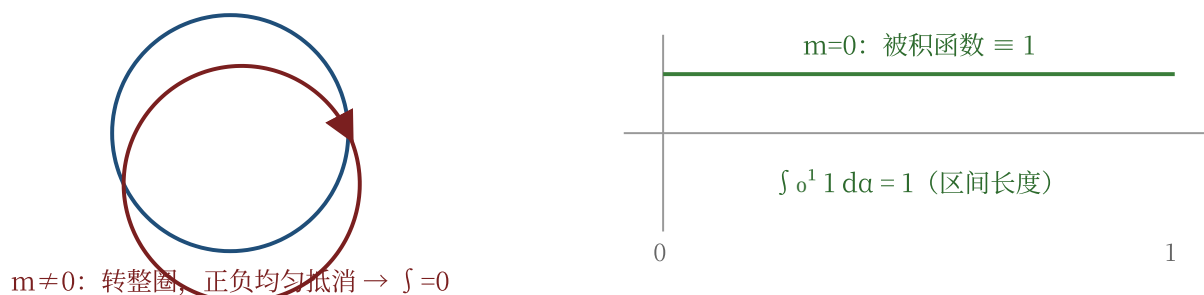
（验证：把右边对  $\alpha$  求导，链式法则得  $\frac{1}{2\pi i m} \cdot 2\pi i m e^{2\pi i m \alpha} = e^{2\pi i m \alpha}$ ，正确。）

3 代入上下限 0 到 1：

$$\int_0^1 e^{2\pi i m \alpha} d\alpha = \frac{1}{2\pi i m} \left( e^{2\pi i m \cdot 1} - e^{2\pi i m \cdot 0} \right) = \frac{1}{2\pi i m} (e(m) - e(0))$$

4 因为  $m$  是整数，前面讲过  $e(m) = 1$ ，又  $e(0) = 1$ ，所以括号  $= 1 - 1 = 0$ 。故整个积分  $= 0$ 。

5 合并两种情形即得正交关系。直观解释： $m \neq 0$  时  $e(m\alpha)$  在  $[0, 1]$  上绕单位圆整整转  $|m|$  圈，正负均匀抵消，累积为 0； $m = 0$  时它恒等于 1，累积为 1。 ◆



正交关系： $m \neq 0$  时积分为 0（转整圈抵消）， $m = 0$  时为 1。这把“积分”变成一台只挑出  $m = 0$  那一项的“筛子”。

推演：把 (2.6) 乘  $e(-N\alpha)$  再积分，得到 (2.7) 现在用上面的正交关系把 (2.7) 推出来，每一步都说清。

1 从 (2.6) 出发： $(T(\alpha))^s = \sum_m r'(m) e(m\alpha)$ 。两边乘以  $e(-N\alpha)$ ：

$$(T(\alpha))^s e(-N\alpha) = \sum_m r'(m) e(m\alpha) e(-N\alpha).$$

2 用  $e(a)e(b) = e(a+b)$ ： $e(m\alpha)e(-N\alpha) = e((m-N)\alpha)$ 。于是右边  $= \sum_m r'(m) e((m-N)\alpha)$ 。

- 3 两边对  $\alpha$  在  $[0, 1]$  上积分，并用积分的线性性（和的积分=积分的和，常数  $r'(m)$  提出）：

$$\int_0^1 (T(\alpha))^s e(-N\alpha) d\alpha = \sum_m r'(m) \int_0^1 e((m - N)\alpha) d\alpha.$$

- 4 对每个内层积分用正交关系：指数里的整数是  $m - N$ 。
- 当  $m \neq N$  时  $m - N \neq 0$ ，积分 = 0，整项消失；
  - 当  $m = N$  时  $m - N = 0$ ，积分 = 1，该项保留为  $r'(N) \cdot 1$ 。
- 所以右边整个求和只剩  $m = N$  这一项：

$$\int_0^1 (T(\alpha))^s e(-N\alpha) d\alpha = r'(N).$$

- 5 由本段开头取  $P \geq [N^{1/k}]$  已证  $r'(N) = r(N)$ ，代入即得目标公式

$$r(N) = \int_0^1 (T(\alpha))^s e(-N\alpha) d\alpha \quad (2.7)$$

一句话总结这套机制：乘上  $e(-N\alpha)$  再积分，就像一把“筛子”，把  $(T(\alpha))^s$  这堆  $e(m\alpha)$  里恰好  $m = N$  的那一项的系数  $r(N)$  单独筛出来，其余全部因“转整圈抵消”而归零。这正是圆法把数论计数  $r(N)$  变成一个定积分的核心一招。

**第 12 段 (2.7) 是出发点：与 Hardy–Littlewood 围道积分的对应**

这就是我们研究 Waring 问题工作的出发点。它对应于 Hardy 与 Littlewood 所用的  $r(N)$  的围道积分，只是把  $z$  替换成了  $e^{2\pi i\alpha}$ 。

**这段在讲什么** 承上启下：宣布 (2.7) 就是后面所有章节的**起跑线**，并点明它和第 8 段 Cauchy 围道积分其实是同一件事的两种写法。

**两种写法的对应关系（为什么说“只是把  $z$  换成  $e^{2\pi i\alpha}$ ”）**

**Hardy-Littlewood 原版**

$$r(N) = \frac{1}{2\pi i} \oint_{|z|=\rho} \frac{F(z)}{z^{N+1}} dz, \text{ 沿复平面圆周 } |z| = \rho \text{ 积分。}$$

**Vinogradov / 本书版**

$$r(N) = \int_0^1 (T(\alpha))^s e(-N\alpha) d\alpha, \text{ 沿实区间 } [0, 1] \text{ 积分。}$$

做一个变量替换  $z = e^{2\pi i\alpha}$  就能看出二者一致：当  $\alpha$  从 0 跑到 1， $z = e^{2\pi i\alpha}$  恰好沿**单位圆**  $|z| = 1$  **绕一圈**（因为  $e(t)$  周期为 1、模长为 1）。所以“沿  $\alpha \in [0, 1]$  的实积分”就是“沿单位圆周的围道积分”，而  $1/z^{N+1}$  中的因子化成了  $e(-N\alpha)$ 。两种语言、同一件事——只不过 Vinogradov 版省去了复变积分的包袱，只剩高中也能逐步看懂的实定积分。这正是本书选它作出发点的原因。

## 第 13 段 本书的首要目标与条件 $s \geq 2^k + 1$ （华罗庚不等式）

我们的首要目标是建立渐近公式 (2.2) 对  $r(N)$  当  $N \rightarrow \infty$  时的成立性，所受的条件为  $s \geq 2^k + 1$ 。借助华罗庚于 1938 年发现的一个不等式（下文引理 3.2），有可能以一种相对简单的方式做到这一点。值得注意的是：就渐近公式本身而言，迄今尚未对小  $k$  值的条件  $s \geq 2^k + 1$

作出任何改进。对于大的  $k$ ，Vinogradov 已经证明形如  $s > Ck^2 \log k$  的条件即足够。

**这段在讲什么** 给出本书接下来要达到的**具体目标**：在条件  $s \geq 2^k + 1$  下证明渐近公式 (2.2)。并交代两个关键事实：靠华罗庚 1938 年的不等式能相对简单地做到；以及对大  $k$  Vinogradov 有更省项的条件。

**条件  $s \geq 2^k + 1$  怎么读、有多大**  $2^k$  是 2 的  $k$  次方，随  $k$  指数增长。条件  $s \geq 2^k + 1$  说：要让本书这套相对简单的方法跑通渐近公式，项数  $s$  至少要  $2^k + 1$ 。代入感受：

1  $k = 2: 2^2 + 1 = 5。$

2  $k = 3: 2^3 + 1 = 9。$

3  $k = 4: 2^4 + 1 = 17。$

4  $k = 5: 2^5 + 1 = 33。$

这是一个“够用但不一定最优”的门槛——好处是**换来证明的简单**。译文特别提醒：**对小  $k$  而言，这个  $2^k + 1$  至今没被改进过**（就渐近公式本身而言），可见它相当本质，不是随便估的。

**华罗庚不等式 (Hua's inequality, 1938)：为什么需要它、它解决了什么** 华罗庚 (Loo-Keng Hua, 中国现代数学奠基人之一) 1938 年发现的一个关于指数和  $T(\alpha)$  的**积分不等式**（本书后文的引理 3.2）。**动机 / 它解决的困难**：要从积分 (2.7) 估计  $r(N)$ ，必须先**控制住  $T(\alpha)$  的大小**——特别是要估计形如  $\int_0^1 |T(\alpha)|^{2^k} d\alpha$  这类“高次幂的平均大小”。这本身很难。华罗庚不等式恰好给出了这种积分的一个干净上界，从而能把 (2.7) 中“次弧”部分的贡献压成误差

项。正因为有了它，本书才能“以相对简单的方式”在  $s \geq 2^k + 1$  下完成渐近公式。它是后面技术推导的**关键工具**，所以这里先郑重预告。

**大  $k$  的 Vinogradov 条件  $s > Ck^2 \log k$ ，以及里面的记号** 对很大的  $k$ ， $2^k + 1$  大得惊人（指数增长），太浪费。Vinogradov 用更深的方法证明：只要  $s > Ck^2 \log k$  就够。

$C$

某个**绝对常数**（一个固定的正数，不依赖  $k$ ），具体多大不重要。

$\log k$

$k$  的对数。这里增长很慢。

$k^2 \log k$

相比  $2^k$ ，它是**多项式**  $\times$  **对数级**，增长慢得多。

对比一下大  $k$  时两者的差距： $2^k$  是指数爆炸，而  $k^2 \log k$  温和得多。所以对大  $k$ ，Vinogradov 的条件远优于  $2^k + 1$ ；但它的方法复杂，本书在入门阶段先用  $2^k + 1$  的简单路线。这就是“取舍”：小  $k$  用简单方法（ $2^k + 1$ ），大  $k$  才值得上 Vinogradov 的重武器。

## 第 14 段 证 $G(k) \leq s$ 比证渐近公式更省：Vinogradov 与 Davenport

若我们对某个特定的  $s$  值（比如  $s = s_1$ ）证明了渐近公式成立，那么就可推出每个大数都可表示为  $s_1$  个  $k$  次方之和，从而  $G(k) \leq s_1$ 。但为了证明这一点，并不必证明关于表法总数的渐近公式；只需对某种特殊类型的表法证明它即可，即作为  $s_1$  个  $k$  次方之和。这使得有可能得到比由渐近公式的成立性所能得到的更好的  $G(k)$  估计。1934 年 Vinogradov 证明了对大的  $k$  有  $G(k) < Ck \log k$ ，我们将在第 9 章给

出一个证明。对小  $k$  已知的最佳结果由 Davenport 于 1939–41 年间找到 [19]。

**这段在讲什么** 把第 6 段那条方法论说透：要压低  $G(k)$  的上界，不必证完整渐近公式，从而能用更少的项数  $s$ 。并报告  $G(k)$  上界的两项里程碑成果。

推演：渐近公式成立  $\implies G(k) \leq s_1$ （与第 5 段同理，复述要点）

- 1 设对  $s = s_1$ ，渐近公式 (2.2) 成立且主项为正（奇异级数有正下界）。
- 2 则对充分大的  $N$ ， $r(N) = \text{正主项} + \text{更小误差} > 0$ 。
- 3  $r(N) > 0 \implies N$  可表示成  $s_1$  个  $k$  次方之和。对一切大  $N$  成立即  $G(k) \leq s_1$ 。 ◆

**关键洞察：**“只要有一种表示”比“数清所有表示”要求低得多 证渐近公式是去算  $r(N)$  的精确个数——这要求对积分 (2.7) 做非常精细的估计，主项误差都要拿捏，因此  $s$  必须够大。但要得到  $G(k) \leq s_1$ ，逻辑上只需  $r(N) > 0$ ，即哪怕只存在一种表示就够了。这给了我们偷懒的空间：

- 可以只对“某种特殊类型的表示”去计数（比如限定某些  $x_i$  落在特定范围、或带特殊结构的解），只要能证明这种特殊表示的个数  $> 0$ ；
- 对“特殊类型”计数，往往比对“全部表示”计数容易控制误差，于是能容许更小的  $s$ 。

结论：“求  $G(k)$  上界”能用比“求渐近公式”更少的项数，这就是为什么 P.N. VI 能把  $G(4)$  压到比渐近公式所需更小的值（第 6 段）。这是圆法里反复使用的省力策略。

两项里程碑结果，以及记号  $<$ 、 $\log$

### Vinogradov 1934: $O(G(k))$

注意比第 13 段渐近公式的大  $k$  条件  $s > Ck^2 \log k$  还要小一个  $k$  的因子（这里是  $k \log k$  而非  $k^2 \log k$ ）——正是上面“只需存在性”省下来的。本书第 9 章会给证明。

### Davenport 1939–41: 小 $k$ 的最佳结果 [19]

Davenport（本书作者本人）当年得到了对小  $k$ （如  $k = 4, 5, 6$  等）当时已知最好的  $G(k)$  上界。这也解释了作者为何对此题如此熟稔——他是这段历史的亲历者与推进者。

$\log k$  是  $k$  的对数，增长极慢； $C$  仍是绝对常数。 $Ck \log k$  随  $k$  增长比  $k^2 \log k$  更慢，是更强（更小）的上界。

## 第 15 段 Linnik 的“初等”证明与华罗庚不等式的影响

关于 Hilbert 定理的一个新的“初等”证明由 Linnik 于 1943 年给出 [58]，并被 Khintchine 选入他的“三颗明珠” [53] 之中。这个证明的根本思想无疑是受到了 Hardy–Littlewood 方法某些特征的启发，特别是受到了华罗庚不等式的启发。

**这段在讲什么** 收尾：交代 Hilbert 定理还有第三条路线——Linnik 1943 年的“初等”证明，并点明它的灵感来自圆法（尤其华罗庚不等式）。这说明本章讲的解析思想影响深远，连“不用解析工具”的证明也是受它启发。

**“初等”证明是什么意思（不是“简单”的意思）** 数学里的“初等”

（elementary）有特定含义：指不使用复变函数、积分、无穷级数等分析工具，只用整数、不等式、组合计数等较“基本”的手段。它不等于“容易”——Linnik 的初等证明其实相当精巧。

- Hilbert 1909: 代数路线（第一条证明）。

- Hardy–Littlewood / 圆法：解析路线（用积分 (2.7)、奇异级数），本章主角。
- Linnik 1943：初等路线（避开解析工具），但其思想内核仍源自圆法，特别是华罗庚不等式所体现的“控制高次和的平均大小”这一想法。

## 人物与出处背景

### Linnik

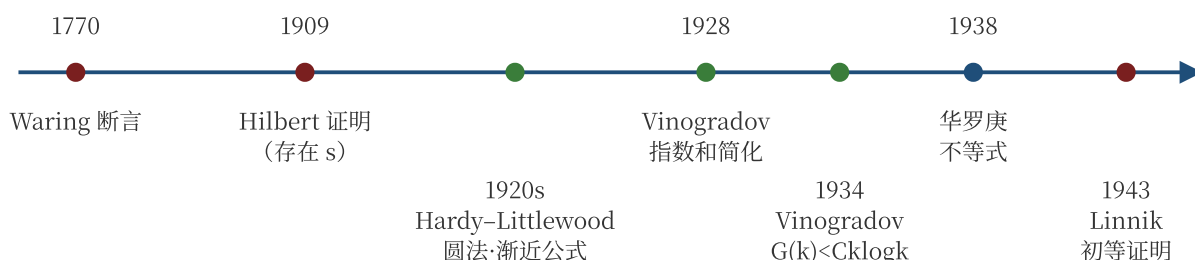
苏联数论学家，1943 年给出 Hilbert 定理的初等新证。

### Khintchine（辛钦）

苏联著名数学家，他写过一本小册子《三颗明珠》(Three Pearls of Number Theory)，精选三个优美的数论证明向大众讲解，Linnik 的这个证明被收入其中——说明它在数学审美上被高度认可。

### 华罗庚不等式的影响

它在本章已两次出现（第 13 段作为本书证渐近公式的工具，本段作为 Linnik 初等证明的灵感来源）。这条 1938 年的不等式因此成为贯穿 Waring 问题各条证明路线的一根**共同主线**，也呼应了全章开头“圆法是一套影响深远的通用思想”的定位。



Waring 问题的历史主线：从 1770 年 Waring 的断言，到 1909 年 Hilbert 证明其存在性，再到 1920 年代 Hardy–Littlewood 的圆法、Vinogradov 的简化与突破、华罗庚不等

## 全章小结：这条逻辑链请记牢

### 把整章串成一条线

1 问题：每个正整数  $N$  能否写成固定个数  $s$  个  $k$  次方之和？  
(Waring 断言，Hilbert 1909 证明可行。)

2 两个常数：  $g(k)$  (覆盖所有  $N$  的最小项数) 与  $G(k)$  (覆盖所有大  $N$  的最小项数，更本质，  $G \leq g$ )。

3 升级为计数：定义表法个数  $r(N)$ ；只要证  $r(N) > 0$  就说明可表示，所以追求更强的渐近公式 (2.2)：  
$$r(N) = C_{k,s} N^{s/k-1} \mathfrak{S}(N) + O(N^{s/k-1-\delta}).$$

4 奇异级数  $\mathfrak{S}(N)$ ：主项里编码同余信息的算术因子；它有正下界  
(2.3)  $\Rightarrow$  主项压过误差  $\Rightarrow r(N) > 0 \Rightarrow$  Hilbert 定理与  $G(k)$  上界。

5 圆法机器：生成函数  $\rightarrow$  截断成指数和  $T(\alpha) = \sum_{x=1}^P e(\alpha x^k)$  (2.5)  
 $\rightarrow$  自乘得  $(T(\alpha))^s = \sum_m r'(m) e(m\alpha)$  (2.6)  $\rightarrow$  取  $P \geq [N^{1/k}]$   
使  $r'(N) = r(N) \rightarrow$  用正交关系“积分筛系数”得出发点

$$r(N) = \int_0^1 (T(\alpha))^s e(-N\alpha) d\alpha. \quad (2.7)$$

6 本书目标：在  $s \geq 2^k + 1$  下用华罗庚不等式证 (2.2)；并知道“求  $G(k)$  上界”比“证渐近公式”省项，由此有 Vinogradov、Davenport 的更好结果。

后面每一章，本质上都是在**动手估计这个积分 (2.7)**：把区间  $[0, 1]$  切成“主弧”（ $\alpha$  接近简单分数处，贡献主项）和“次弧”（其余，贡献误差），分别处理。这就是“圆法”的全部戏剧将要上演的舞台。

---

返回 [全书目录](#)