Pseudorandom Binary Sequences in Number Theory

数论中的伪随机二进制数列

刘华宁 著

科学出版社
www.sciencep.com

数论中的伪随机二进制数列

刘华宁 著

本书得到西北大学研究生创新教育项目资助

科学出版社

北京

内 容 简 介

随着通信与计算机网络的发展，伪随机二进制数列得到了广泛的应用，并已成为密码学的一个基本工具，在构造密码系统中起着重要的作用. 本书介绍了如何基于数论中的 Legendre 符号、Liouville 函数、最大素因子、丢番图逼近、指标、最小非负剩余、Lehmer 问题与 Gallagher 问题等来生成伪随机二进制数列，使用的方法涉及多项式特征和的估计、多项式指数和的估计、Dirichlet L 函数均值、有限域上多项式理论等. 该书是对这一新兴领域十余年来研究工作的一个阶段性总结，其中包含了作者近几年来的研究成果.

本书可供高等院校数学系、计算机系研究生或高年级本科生学习，也可供数论、信息安全与密码学相关专业人员参考.

图书在版编目(CIP)数据

数论中的伪随机二进制数列/刘华宁著. —北京：科学出版社，2008

ISBN 978-7-03-021748-6

Ⅰ. 数… Ⅱ. 刘… Ⅲ. 伪随机码-二进制-数列 Ⅳ. O156.1

中国版本图书馆 CIP 数据核字(2008) 第 057771 号

责任编辑：王丽平 杨 然／责任校对：陈玉凤

责任印制：赵德静／封面设计：王 浩

科 学 出 版 社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

源海印刷有限责任公司 印刷

科学出版社发行 各地新华书店经销

*

2008 年 5 月第 一 版 开本：B5(720×1000)

2008 年 5 月第一次印刷 印张：11

印数：1—3 000 字数：208 000

定价：38.00 元

（如有印装质量问题，我社负责调换〈环伟〉）

前　言

从 20 世纪中叶开始, 随着应用数学与计算机科学的发展, 伪随机数列得到了广泛的应用. 许多论文都是关于这一领域的, 这些论文中提出了大量的思想、方法、工具, 甚至于“伪随机性”这个概念都有着若干的定义, 依赖于具体的应用. 这些论文中大多数是关于 [0,1) 区间的伪随机数列, 而对于伪随机二进制数列研究得比较少, 然而伪随机二进制数列也是非常有用的, 特别是在密码学中.

一个密码系统的安全性可以通过破译该系统的最好算法的计算复杂性来度量, 因而计算复杂性理论已成为现代密码学的基础. 与此同时, 伪随机二进制数列得到了广泛的应用, 并已成为密码学的一个基本工具, 在构造密码系统中起着重要的作用. 具体来说, 基于计算复杂性理论构造的伪随机二进制数列与真随机数列是多项式时间不可区分的, 也是多项式时间不可预测的. 这种类型的伪随机二进制数列具有重要的意义, 用它构造的密码体制具有与用相同长度的真随机数列构造的密码体制同样的安全性.

目前已有的基于计算复杂性理论构造的伪随机二进制数列都是基于大数分解或离散对数等数学难题的, 由于生成速度慢等缺点, 不能完全满足实际的需要. 在实际应用中, 当需要伪随机二进制数列时, 人们通常利用硬件设备或数学方法来获得所需数列. 然而对于得到的数列, 人们往往事先不知道其伪随机性如何, 因此必须进行某些统计测试, 使得伪随机数列满足真随机数列所应具有的某些统计性质或能通过某些统计测试.

基于这些事实, 若能事先构造出一些伪随机二进制数列, 并在理论上研究其伪随机性, 无疑是有意义的. 因此, Mauduit 与 Sárközy 等从 1997 年开始从数论的角度开展了对伪随机二进制数列的研究, 基于数论方法提出了一些新的二进制数列并讨论了其伪随机性.

本书共分 8 章, 介绍了如何基于数论中的 Legendre 符号、Liouville 函数、最大素因子、丢番图逼近、指标、最小非负剩余、Lehmer 问题与 Gallagher 问题等来生成伪随机二进制数列, 使用的方法涉及多项式特征和的估计、多项式指数和的估计、Dirichlet L 函数均值、有限域上多项式理论等. 该书是对这一新兴领域十余年来研究工作的一个阶段性总结, 其中包含了作者近几年来的研究成果. 本书可供高等院校数学系、计算机系研究生或高年级本科生学习, 也可供数论、信息安全与密码学相关专业人员参考.

本书是在作者的博士学位论文部分内容基础上加工而成的, 作者谨以此书向导

师张文鹏教授表示由衷的感谢. 感谢新加坡的 Harald Niederreiter 教授、澳大利亚的 Igor Shparlinski 教授、法国的 Christian Mauduit 教授对作者研究工作的关注, 作者对此深表感激. 当然还要感谢美国的李文卿教授, 由于她, 作者才得以了解到伪随机二进制数列这一新兴方向, 并得以进入一个新的天地. 衷心感谢本人的博士后合作导师王小云教授, 与王老师的多次讨论使作者得以深刻理解伪随机二进制数列在密码学中的意义.

感谢西北大学研究生处、"211" 办公室、数学系, 以及山东大学数学与系统科学学院、博士后管理办公室对作者写作本书的大力支持. 该书的写作和出版得到了国家重大基础研究计划 973 课题 (编号: 2007CB807902, 2007CB807903)、国家自然科学基金 (编号: 10671155)、陕西省自然科学基金 (编号: 2006A04)、中国博士后科学基金 (编号: 20070421084)、山东省博士后创新项目 (编号: 200702036) 的资助. 此外, 本书得到西北大学研究生创新教育项目资助, 该项目是西北大学"211 工程"建设公共服务体系项目的子项目之一.

限于作者水平, 书中疏漏之处在所难免, 欢迎批评指正.

刘华宁

2008 年 4 月

目　录

第 1 章　伪随机二进制数列的测度

当前人类已经进入了一个崭新的时代, 传统的商务活动、事务处理以及政府服务等越来越多地通过开放的计算机和通信网络来实施和提供. 只有在开放网络能提供安全通信的情况下, 上述活动才能顺利实现, 而各种形式的密码则是解决这一问题的基本理论和方法.

一个密码系统的安全性可以通过破译该系统的最好算法的计算复杂性来度量, 因而计算复杂性理论已成为现代密码学的基础. 与此同时, 伪随机二进制数列得到了广泛的应用, 并已成为密码学的一个基本工具, 在构造密码系统中起着重要的作用. 具体来说, 基于计算复杂性理论构造的伪随机二进制数列与真随机数列是多项式时间不可区分的, 也是多项式时间不可预测的. 这种类型的伪随机二进制数列具有重要的意义, 用它构造的密码体制具有与用相同长度的真随机数列构造的密码体制同样的安全性.

目前已有的基于计算复杂性理论构造的伪随机二进制数列都是基于大数分解或离散对数等数学难题的, 由于生成速度慢等缺点, 不能完全满足实际的需要. 在实际应用中, 当需要伪随机二进制数列时, 人们通常利用硬件设备或数学方法来获得所需数列. 然而对于得到的数列, 人们往往事先不知道其伪随机性如何, 因此必须进行某些统计测试, 使得伪随机数列满足真随机数列所应具有的某些统计性质或能通过某些统计测试.

基于这些事实, 若能事先构造出一些伪随机二进制数列, 并从数学理论上研究其伪随机性, 无疑是有意义的. Mauduit 与 Sárközy 等从 1997 年开始从数论的角度开展了对伪随机二进制数列的研究, 基于数论方法提出了一些新的二进制数列并讨论了其伪随机性.

本章我们介绍 Mauduit 与 Sárközy 等引入的用于研究伪随机二进制数列的一些测度, 研究这些测度的背景、性质, 讨论测度之间的关系以及与线性复杂度、Gowers 范数等性质的联系, 从而为后面第 3~8 章研究伪随机二进制数列提供了基本的工具.

§1.1　伪随机测度

研究伪随机二进制数列, 首先需要引入伪随机测度, 以衡量二进制数列的伪随机性.

怎样定义伪随机测度? 一般来说, 人们会不自觉地提出关于伪随机性的越来越多的标准, 期望能借此选出越来越好的数列. 然而, 当这些标准越来越多时, 人们处理起来就会越来越困难, 甚至很难找到符合要求的数列.

例如, Calabro 与 Wolf[3] 于 1968 年引入了最佳二进阵列的定义.

定义 1.1　设 $S=[S(x_1,\cdots,x_n)]$ 是一个 n 维 $N_1\times N_2\times\cdots\times N_n$ 阶的矩阵, 其中 $0\leqslant x_i\leqslant N_i-1(1\leqslant i\leqslant n)$. 如果高维矩阵 S 满足

(I) 元素为 ±1, 即 $S(x_1,\cdots,x_n)=\pm1$,

(II) 异相自相关函数为 0, 即

$$\sum_{x_1 \bmod N_1}\cdots\sum_{x_n \bmod N_n}S(x_1,\cdots,x_n)S(x_1+r_1,\cdots,x_n+r_n)$$
$$=\begin{cases}E, & 当\ (r_1,\cdots,r_n)=(0,\cdots,0),\\ 0, & 当\ (r_1,\cdots,r_n)\neq(0,\cdots,0),\end{cases}$$

那么就称 S 是体积为 $E=N_1\times\cdots\times N_n$ 的 n 维最佳二进数列, 其中 x_i+r_i 是在模 N_i 上运算.

在 20 世纪 60 年代末期最佳二进阵列刚刚诞生时, 由于良好的相关特性, 它受到了通信与电子领域内众多学者的广泛重视. 但是由于当时缺乏计算机的帮助, 人们很快发现: 最佳二进阵列虽然性能特别好但却难以寻找 (在 20 世纪 60 年代仅找到两个二维 2×2 阶和 4×4 阶最佳二进阵列). 所以, 人们只好忍痛割爱去寻找其他性能较好的伪噪声阵列来代替最佳二进阵列. 从 20 世纪 70 年代初到 80 年代末, 人们再也没有去认真研究过最佳二进阵列, 几乎将它遗忘了. 直到 80 年代末, 由于计算机的飞速发展, 最佳二进阵列才成为学术界的一个研究热点.

因此, 我们提出的伪随机测度, 既要能很好地衡量二进制数列的性质, 还要具有一定的可操作性.

首先, 我们希望伪随机测度能反映出在实际应用中最为重要以及研究得最深入的一些随机性质, 例如: 正则性、一致分布、相关性. 其次, 伪随机测度应该是定义在二进制数列上的实值函数, 这样相同长度的两个数列就可以比较伪随机性. 最后, 伪随机测度应该具有不同的层次.

对于第三条要求, 我们以一个例子来说明. 在文献 [32] 的第 162 页有这样一个伪随机测度的定义.

定义 1.2　设

$$E_N=(e_1,\cdots,e_n)\in\{-1,+1\}^N,\quad X=(x_1,\cdots,x_k)\in\{-1,+1\}^k,$$

$M\in\mathbb{N}$ 且 $M\leqslant N+1-k$,

$$T(E_N,M,X)=|\{n:0\leqslant n<M,(e_1,\cdots,e_{n+k})=X\}|,$$

如果对任意自然数 $k \leqslant \dfrac{\log N}{\log 2}$ 以及任意 $X \in \{-1,+1\}^k$ 都有

$$\left|T(E_N, N+1-k, X) - \frac{N+1-k}{2^k}\right| \leqslant \frac{1}{\sqrt{N}},$$

则称 E_N 是伪随机的.

根据定义 1.2, 如果某数列 E_N 满足

$$\left|T(E_N, N+1-k, X) - \frac{N+1-k}{2^k}\right| \leqslant \frac{1.1}{\sqrt{N}},$$

那么 E_N 是伪随机的吗? 更进一步, 如果此数列在其他方面还具有非常好的性质, 那么怎么看待这个数列呢?

因此我们提出的伪随机测度应该灵活一点, 要有层次.

设 $E_N = (e_1, \cdots, e_n) \in \{-1,+1\}^N$, Mauduit 与 Sárközy[35] 定义了下面的测度.

定义 1.3 k 阶正则测度:

$$N_k(E_N) = \max_{X \in \{-1,+1\}^k} \max_{0<M \leqslant N+1-k} \left|T(E_N, M, X) - \frac{M}{2^k}\right|,$$

一致分布测度 (Well-distribution measure):

$$W(E_N) = \max_{a,b,t} \left|\sum_{j=0}^{t-1} e_{a+jb}\right|, \quad 1 \leqslant a \leqslant a+(t-1)b \leqslant N,$$

k 阶相关测度 (Correlation measure of order k):

$$C_k(E_N) = \max_{M,D} \left|\sum_{n=1}^{M} e_{n+d_1} e_{n+d_2} \cdots e_{n+d_k}\right|, \quad 0 \leqslant d_1 < \cdots < d_k \leqslant N-M,$$

其中 $D = (d_1, \cdots, d_k)$.

定理 1.1 对任意 N, E_N 以及 $k < N$, 有 $N_k(E_N) \leqslant \max\limits_{1 \leqslant t \leqslant k} |C_t(E_N)|$.

证 对任意 $k, N \in \mathbb{N}$, $X = (x_1, \cdots, x_k) \in \{-1,+1\}^k$ 以及 $1 \leqslant M \leqslant N+1-k$ 可得

$$\begin{aligned} &\left|T(E_N, M, X) - \frac{M}{2^k}\right| \\ =&\left|\left|\{n : 0 \leqslant n < M, (e_{n+1}, \cdots, e_{n+k}) = X\}\right| - \frac{M}{2^k}\right| \end{aligned}$$

$$
\begin{aligned}
&= \left| \sum_{n=0}^{M-1} \frac{x_1 \cdots x_k}{2^k} \prod_{j=1}^{k} (e_{n+j} + x_j) - \frac{M}{2^k} \right| \\
&= \left| \frac{x_1 \cdots x_k}{2^k} \sum_{1 \leqslant d_1 < \cdots < d_t \leqslant k} \left(\prod_{j \in \{1, \cdots, k\} \setminus \{d_1, \cdots, d_t\}} x_j \right) \sum_{n=0}^{M-1} e_{n+d_1} \cdots e_{n+d_t} \right| \\
&\leqslant \frac{1}{2^k} \sum_{\substack{D \subset \{1,2,\cdots,k\} \\ D \neq \varnothing}} |V(E_N, M, D)| \leqslant \frac{1}{2^k} \sum_{t=1}^{k} \binom{k}{t} C_t(E_N) \\
&\leqslant \max_{1 \leqslant t \leqslant k} |C_t(E_N)| .
\end{aligned}
$$

由此可证定理. ■

定理 1.1 说明 k 阶正则测度可由相关测度决定. 一致分布测度与相关测度之间似乎没有什么关系. 由文献 [32] 中第 168 页的几个例子我们知道, 当 $C_k(E_N)$ 很小时, $W(E_N)$ 有可能比较大; 另一方面, 文献 [35] 中也有例子说明了当 $W(E_N)$ 很小时, $C_k(E_N)$ 也可能很大. 因此, 对于二进制数列 E_N, 如果它的一致分布测度 $W(E_N)$ 与 k 阶相关测度 $C_k(E_N)$ (至少对于较小的 k) 的阶相对于 N 来说都比较小的话, 就认为是一个"好"的伪随机二进制数列. 有时候为了方便, 也把一致分布测度 $W(E_N)$ 与 k 阶相关测度 $C_k(E_N)$ 组合起来.

定义 1.4　k 阶联合测度 (Combined PR-measure of order k):

$$
Q_k(E_N) = \max_{a,b,t,D} \left| \sum_{j=0}^{t} e_{a+jb+d_1} e_{a+jb+d_2} \cdots e_{a+jb+d_k} \right|,
$$
$$
1 \leqslant a + jb + d_i \leqslant N (j = 0, 1, \cdots, t; i = 1, 2, \cdots, k),
$$

其中 $D = (d_1, \cdots, d_k)$.

§1.2　测度之间的关系[39]

我们进一步深入讨论一致分布测度 $W(E_N)$ 与 k 阶相关测度 $C_k(E_N)$ 之间的数量关系.

定理 1.2　对任意 $N \in \mathbb{N}$ 以及 $E_N = (e_1, \cdots, e_N) \in \{-1, +1\}^N$, 有

$$
W(E_N) \leqslant 3 \left(N C_2(E_N) \right)^{1/2}. \tag{1.1}
$$

证　设

$$
a, b, t \in \mathbb{N}, \qquad 1 \leqslant a \leqslant a + (t-1)b \leqslant N, \tag{1.2}
$$

并规定当 $n>N$ 时 $e_n=0$. 不妨设 $t\geqslant 2$, 则由式 (1.2) 可得

$$b<N,\qquad \text{以及}\qquad t-1\leqslant (t-1)b\leqslant N-a\leqslant N-1,$$

从而

$$t\leqslant N. \tag{1.3}$$

则有

$$\begin{aligned}\sum_{i=a}^{a+b-1}\left(\sum_{j=0}^{t-1}e_{i+jb}\right)^2&=\sum_{i=a}^{a+b-1}\left(\sum_{j=0}^{t-1}(e_{i+jb})^2+2\sum_{0\leqslant j_1<j_2\leqslant t-1}e_{i+j_1b}e_{i+j_2b}\right)\\&=\sum_{i=a}^{a+b-1}\left(t+2\sum_{d=1}^{t-1}\sum_{j_1=0}^{t-1-d}e_{i+j_1b}e_{i+j_1b+db}\right)\\&=tb+2\sum_{d=1}^{t-1}\sum_{j_1=0}^{t-1-d}\sum_{i=a}^{a+b-1}e_{i+j_1b}e_{i+j_1b+db}\\&=(t-1)b+b+2\sum_{d=1}^{t-1}\sum_{n=a}^{a+(t-d)b-1}e_ne_{n+db}\\&<N+N+2\sum_{d=1}^{t-1}\left|\sum_{n=a}^{a+(t-d)b-1}e_ne_{n+db}\right|.\end{aligned}\tag{1.4}$$

容易证明

$$\left|\sum_{n=a}^{a+(t-d)b-1}e_ne_{n+db}\right|\leqslant 2C_2(E_N). \tag{1.5}$$

则由式 (1.3), 式 (1.4) 和式 (1.5) 可得

$$\begin{aligned}\left(\sum_{j=0}^{t-1}e_{a+jb}\right)^2&\leqslant\sum_{i=a}^{a+b-1}\left(\sum_{j=0}^{t-1}e_{i+jb}\right)^2<2N+2\sum_{d=1}^{t-1}2C_2(E_N)\\&=2N+4(t-1)C_2(E_N)\leqslant 6NC_2(E_N).\end{aligned}\tag{1.6}$$

这就证明了定理. ■

式 (1.1) 中的估计式在不考虑系数的情况下是最佳的, 我们将通过证明下面的定理来说明这一点.

定理 1.3 设 $k,N\in\mathbb{N}$, $N>N_0$, 且

$$N^{3/4}\leqslant k\leqslant N, \tag{1.7}$$

则存在数列 $E_N \in \{-1,+1\}^N$ 满足

$$W(E_N) \geqslant k \tag{1.8}$$

以及

$$C_2(E_N) \leqslant 120 \max\left\{\frac{k^2}{N}, (N\log N)^{1/2}\right\}. \tag{1.9}$$

此外由式 (1.9) 可得

$$(NC_2(E_N))^{1/2} < 11\max\{k, N^{3/4}(\log N)^{1/4}\}, \tag{1.10}$$

则当 $k \geqslant N^{3/4}(\log N)^{1/4}$ 时, 由式 (1.1), 式 (1.8) 和式 (1.10) 可知, 对于该数列 E_N 有

$$k \leqslant W(E_N) \leqslant 3\,(NC_2(E_N))^{1/2} < 33k.$$

在证明定理 1.3 之前, 先证明两个引理.

引理 1.1　设 $\mathcal{A}$ 为正整数的有限集合, $d \in \mathbb{N}$, 并定义 $f(\mathcal{A}, d)$ 为满足方程

$$a - a' = d, \qquad a \in \mathcal{A}, \qquad a' \in \mathcal{A} \tag{1.11}$$

的解的个数. 假定正整数 N 充分大, k 满足 $N^{3/4} \leqslant k \leqslant \dfrac{N}{10}$, 则存在集合 $\mathcal{A} \subset \{1,2,\cdots,N\}$ 满足

$$|\mathcal{A}| = k \tag{1.12}$$

和

$$f(\mathcal{A}, d) < 30\frac{k^2}{N}, \quad \text{对任意 } 1 \leqslant d \leqslant N. \tag{1.13}$$

证　记 $M = 30\dfrac{k^2}{N}$, $\mathcal{F} = \{\mathcal{A} : \mathcal{A} \subset \{1,2,\cdots,N\}, |\mathcal{A}| = k\}$, 以及 $\mathcal{F}_d = \{\mathcal{A} : \mathcal{A} \in \mathcal{F}, f(\mathcal{A}, d) \geqslant M\}$. 显然集合

$$\mathcal{F} \setminus \bigcup_{d=1}^{N-1} \mathcal{F}_d \tag{1.14}$$

中的任意子集 $\mathcal{A}$ 都满足式 (1.12) 和式 (1.13), 现在只需证明式 (1.14) 中的集合非空. 我们将给出 $|\mathcal{F}_d|$ 的上界估计, 使得

$$\sum_{d=1}^{N-1} |\mathcal{F}_d| < |\mathcal{F}| = \binom{N}{k} \tag{1.15}$$

成立, 从而引理得证.

考虑集合

$$\mathcal{A} \in \mathcal{F}_d, \tag{1.16}$$

并记 $t = [M/3]$. 我们将通过下面的递推关系定义集合:

$$(\mathcal{A} =)\mathcal{A}_0 \supset \mathcal{A}_1 \supset \cdots \supset \mathcal{A}_t$$

与整数

$$a_1 \in \mathcal{A}_0, \quad a_2 \in \mathcal{A}_1, \quad \cdots, \quad a_t \in \mathcal{A}_{t-1}.$$

首先规定

$$\mathcal{A}_0 = \mathcal{A}. \tag{1.17}$$

假设对于 $1 \leqslant i \leqslant t$ 集合 $\mathcal{A}_0, \mathcal{A}_1, \cdots, \mathcal{A}_{i-1}$ 以及整数 $a_1, a_2, \cdots, a_{i-1}$ 已经定义, 并满足

$$(\mathcal{A} =)\mathcal{A}_0 \supset \mathcal{A}_1 \supset \cdots \supset \mathcal{A}_{i-1}, \tag{1.18}$$

$$|\mathcal{A}_j| = |\mathcal{A}| - 2j, \qquad 0 \leqslant j \leqslant i-1, \tag{1.19}$$

$$f(\mathcal{A}_j, d) \geqslant M - 3j, \qquad 0 \leqslant j \leqslant i-1, \tag{1.20}$$

$$a_j \in \mathcal{A}_{j-1}, a_j + d \in \mathcal{A}_{j-1}, a_j \notin \mathcal{A}_j, a_j + d \notin \mathcal{A}_j, \quad 0 < j \leqslant i-1. \tag{1.21}$$

由式 (1.20) 可得 $f(\mathcal{A}_{i-1}, d) \geqslant M - 3(i-1) > M - 3t \geqslant 0$, 从而方程

$$a - a' = d, \quad a \in \mathcal{A}_{i-1}, \quad a' \in \mathcal{A}_{i-1} \tag{1.22}$$

至少有一个解. 设 $a_i \in \mathcal{A}_{i-1}$ 是满足上式的最小整数, 由此定义 $\mathcal{A}_i$ 为

$$\mathcal{A}_i = \mathcal{A}_{i-1} \setminus \{a_i, a_i + d\}.$$

显然 $\mathcal{A}_i$ 也满足式 (1.18) 到式 (1.21) 中的条件, 从而得到集合

$$(\mathcal{A} =)\mathcal{A}_0 \supset \mathcal{A}_1 \supset \cdots \supset \mathcal{A}_t$$

与整数

$$a_1 \in \mathcal{A}_0, \quad a_2 \in \mathcal{A}_1, \quad \cdots, \quad a_t \in \mathcal{A}_{t-1},$$

此外可知 $\mathcal{A}$ 是集合 $\{a_1, a_1 + d\}, \cdots, \{a_t, a_t + d\}$, $\mathcal{A}_t$ 的不相交并集.

在 $\{1, 2, \cdots, N\}$ 中选取 $a_1, \cdots, a_t$, 最多有 $\binom{N}{t}$ 种选取方式, 这些数唯一决定

了 $a_1+d,\cdots,a_t+d$. 则由式 (1.12) 与式 (1.19), $\mathcal{A}_t$ 中的元素可在剩下的 $N-2t$ 个整数中选取, 最多有 $\binom{N-2t}{|\mathcal{A}_t|}=\binom{N-2t}{k-2t}$ 种选取方式. 则有

$$\begin{aligned}|\mathcal{F}_d|&\leqslant\binom{N}{t}\binom{N-2t}{k-2t}=\frac{N!(N-2t)!}{t!(N-t)!(k-2t)!(N-k)!}\\&=\frac{k!(N-2t)!}{t!(N-t)!(k-2t)!}\binom{N}{k}\\&=\frac{(k-2t+1)(k-2t+2)\cdots k}{t!(N-2t+1)(N-2t+2)\cdots(N-t)}\binom{N}{k}\\&\leqslant\frac{k^{2t}}{t!(N-2t)^t}\binom{N}{k}.\end{aligned}$$

注意到 $N^{3/4}\leqslant k\leqslant\dfrac{N}{10}$, $t=[M/3]$, 则由 Stirling 公式可得

$$\begin{aligned}|\mathcal{F}_d|&\leqslant\frac{k^{2t}}{(t/3)^t(N/3)^t}\binom{N}{k}=\left(\frac{9k^2}{tN}\right)^t\binom{N}{k}=\left(\frac{9M}{30t}\right)^t\binom{N}{k}\\&\leqslant\left(\frac{10}{11}\right)^t\binom{N}{k}<\left(\frac{11}{12}\right)^{M/3}\binom{N}{k}=\left(\frac{11}{12}\right)^{10k^2/N}\binom{N}{k}\\&\leqslant\left(\frac{11}{12}\right)^{10N^{1/2}}\binom{N}{k}<\frac{1}{N}\binom{N}{k}.\end{aligned}\tag{1.23}$$

由式 (1.23) 可得式 (1.15), 从而完成引理的证明. ■

引理 1.2　*存在正整数 N_0 使得对 $N>N_0$, $1\leqslant V\leqslant N$ 有*

$$\frac{1}{2^V}\sum_{|r-V/2|>15(N\log N)^{1/2}}\binom{V}{r}<\frac{1}{N^4}.$$

证　该引理可通过简单的直接计算, 或者利用二项式分布的初等估计来证明.■

现在证明定理 1.3. 当 $k>\dfrac{N}{10}$ 时, 由式 (1.8) 立即可得式 (1.9), 因此下面不妨假设 $k\leqslant\dfrac{N}{10}$. 记

$$\Delta=30\max\left\{\frac{k^2}{N},(N\log N)^{1/2}\right\},$$

则式 (1.9) 可写成

$$C_2(E_N)\leqslant 4\Delta.\tag{1.24}$$

设集合 $\mathcal{A}\subset\{1,2,\cdots,N\}$ 满足引理 1.1 中的条件 (1.12) 和 (1.13), 并定义 $\mathcal{E}$ 为满足 $e_n=+1, n\in\mathcal{A}$ 的二进制数列 $E_N\in\{-1,+1\}^N$ 的集合, 从而有 $|\mathcal{E}|=2^{N-|\mathcal{A}|}=$

2^{N-k}. 在 $\mathcal{E}$ 中以 $1/2^{N-k}$ 的概率随机选取一个 $E_N \in \mathcal{E}$, 换句话说, 这个二进制数列 $E_N = (e_1, \cdots, e_N)$, 当 $n \in \mathcal{A}$ 时 $e_n = +1$, 而当 $n \notin \mathcal{A}$ 时则有

$$P(e_n = +1) = P(e_n = -1) = \frac{1}{2}, \qquad \text{对于 } n \notin \mathcal{A}.$$

我们将在下面证明, 这样的 $E_N \in \mathcal{E}$ 以大于 1/3 的概率满足式 (1.8) 和式 (1.9), 这就是说存在数列 E_N 符合定理 1.3 的要求, 从而完成定理 1.3 的证明.

显然有

$$W(E_N) \geqslant \left|\sum_{n=1}^{N} e_n\right| = \left|\sum_{\substack{n\leqslant N\\ n\in\mathcal{A}}} e_n + \sum_{\substack{n\leqslant N\\ n\notin\mathcal{A}}} e_n\right| = \left|k + \sum_{\substack{n\leqslant N\\ n\notin\mathcal{A}}} e_n\right|. \tag{1.25}$$

此外由对称性, 可得

$$\begin{aligned} P\left(\sum_{\substack{n\leqslant N\\ n\notin\mathcal{A}}} e_n \geqslant 0\right) &= \frac{1}{2}P\left(\sum_{\substack{n\leqslant N\\ n\notin\mathcal{A}}} e_n \neq 0\right) + P\left(\sum_{\substack{n\leqslant N\\ n\notin\mathcal{A}}} e_n = 0\right) \\ &\geqslant \frac{1}{2}\left(P\left(\sum_{\substack{n\leqslant N\\ n\notin\mathcal{A}}} e_n \neq 0\right) + P\left(\sum_{\substack{n\leqslant N\\ n\notin\mathcal{A}}} e_n = 0\right)\right) = \frac{1}{2}. \end{aligned} \tag{1.26}$$

则由式 (1.25) 和式 (1.26) 有

$$P\left(W(E_N) \geqslant k\right) \geqslant P\left(\left|k + \sum_{\substack{n\leqslant N\\ n\notin\mathcal{A}}} e_n\right| \geqslant k\right) \geqslant P\left(\sum_{\substack{n\leqslant N\\ n\notin\mathcal{A}}} e_n \geqslant 0\right) \geqslant \frac{1}{2}. \tag{1.27}$$

接下来我们给出 $P(C_2(E_N) > 4\Delta)$ 的上界估计.

由 $C_2(E_N)$ 的定义易得

$$\begin{aligned} P(C_2(E_N) > 4\Delta) &= P\left(\max_{M,d_1,d_2}\left|\sum_{n=1}^{M} e_{n+d_1}e_{n+d_2}\right| > 4\Delta\right) \\ &\leqslant \sum_{M,d_1,d_2} P\left(\left|\sum_{n=1}^{M} e_{n+d_1}e_{n+d_2}\right| > 4\Delta\right). \end{aligned} \tag{1.28}$$

对任意 $E_N \in \mathcal{E}$ 有

$$\begin{aligned}\sum_{n=1}^{M} e_{n+d_1}e_{n+d_2} &= \sum_{\substack{n\leqslant M\\ n+d_1\in\mathcal{A},n+d_2\in\mathcal{A}}} e_{n+d_1}e_{n+d_2} + \sum_{\substack{n\leqslant M\\ n+d_1\in\mathcal{A},n+d_2\notin\mathcal{A}}} e_{n+d_1}e_{n+d_2}\\ &\quad + \sum_{\substack{n\leqslant M\\ n+d_1\notin\mathcal{A},n+d_2\in\mathcal{A}}} e_{n+d_1}e_{n+d_2} + \sum_{\substack{n\leqslant M\\ n+d_1\notin\mathcal{A},n+d_2\notin\mathcal{A}}} e_{n+d_1}e_{n+d_2}\\ &= \Sigma_1+\Sigma_2+\Sigma_3+\Sigma_4. \end{aligned} \tag{1.29}$$

由 Δ 与 $\mathcal{A}$ 的定义有

$$\begin{aligned}\Sigma_1 &\leqslant \sum_{\substack{n\leqslant M\\ n+d_1\in\mathcal{A},n+d_2\in\mathcal{A}}} 1 \leqslant \sum_{\substack{a,a'\in\mathcal{A}\\ a-a'=d_2-d_1}} 1\\ &= f(\mathcal{A}, d_2-d_1) < 30\frac{k^2}{N} \leqslant \Delta. \end{aligned} \tag{1.30}$$

则由式 (1.29) 与式 (1.30) 可得

$$\left\{E_N : \left|\sum_{n=1}^{M} e_{n+d_1}e_{n+d_2}\right| > 4\Delta\right\} \subset \bigcup_{i=2}^{4}\{E_N : |\Sigma_i| > \Delta\}. \tag{1.31}$$

接下来只需给出 $P(|\Sigma_i| > \Delta)$ (对任意 M, d_1, d_2 以及 $2 \leqslant i \leqslant 4$) 的上界估计.

首先考虑 $i = 2$ 的情形. 根据 $\mathcal{E}$ 的定义, 对任意 E_N 有

$$\Sigma_2 = \sum_{\substack{n\leqslant M\\ n+d_1\in\mathcal{A},n+d_2\notin\mathcal{A}}} e_{n+d_2},$$

这是 V 个独立随机变量 e_{n+d_2} (e_{n+d_2} 取 $-1, +1$ 的概率各为 $1/2$) 之和, 其中 $V = |\{n : n \leqslant M, n+d_1 \in \mathcal{A}, n+d_2 \notin \mathcal{A}\}|$. 对给定的 E_N, 记

$$r = |\{n : n \leqslant M, n+d_1 \in \mathcal{A}, n+d_2 \notin \mathcal{A}, e_{n+d_2} = +1\}|$$

以及

$$s = |\{n : n \leqslant M, n+d_1 \in \mathcal{A}, n+d_2 \notin \mathcal{A}, e_{n+d_2} = -1\}|,$$

则有

$$|\Sigma_2| = |r-s| = |2r-V| = 2\left|r-\frac{V}{2}\right|,$$

从而 $|\Sigma_2| > \Delta$ 当且仅当 $\left|r-\dfrac{V}{2}\right| > \dfrac{\Delta}{2}$.