arXiv:2303.12990v1 [cs.IT] 23 Mar 2023

# 关于定重量二进制 $B_2$ 序列

**Jin Sima**、**Yun-Han Li**、**Ilan Shomorony**和**Olgica Milenkovic**，伊利诺伊大学厄巴纳-香槟分校电气与计算机工程系，美国 {jsima,yunhanl2,ilans,milenkov}@illinois.edu

摘要——受基于聚合物的数据存储应用的启发，我们提出了一个新问题：刻画码率并设计定重量二进制 $B_2$ 序列。二进制 $B_2$ 序列是一组长度为$n$ 的二进制字符串，其性质是任意两个不同字符串的实数和各不相同。除了这一基本性质外，定重量二进制 $B_2$序列还满足约束：每个字符串的重量 $\omega$ 固定且相对较小，并随 $n$线性增长。定重量约束确保了低成本合成，并使通过串联质谱仪读取的数据能够得到均匀处理。我们的主要结果包括：将码的大小上界表述为熵优化问题，以及基于 Sidon 序列构造的下界。

## I. 引言

二进制 $B_2$序列由 Lindström 于 [1]年提出，随后在多项后续研究中得到探讨 [2]，[3]，[4]。二进制 $B_2$序列表示一组（码本）固定长度的二进制向量，其中集合内所有码向量两两相加所得的逐位实数和各不相同；因此，给定和便可唯一确定相加的向量。自提出以来，这类序列已在许多领域得到应用，例如搜索算法 [5], [6], 、多址系统设计 [7], [8], 和数据指纹识别 [9], [3]。

二进制 $B_2$-序列的一些较新应用包括基于聚合物的数据存储。基于 DNA、其他天然大分子和合成聚合物的非易失性存储系统，有望实现超高存储密度、长期读取兼容性和出色的耐久性 [10], [11], [12], [13], [14], [15], [16], [17]。合成聚合物是二进制分子存储介质，它们利用质量显著不同的聚合物来表示 0 和 1 [18]。用户定义的二进制字符串通过按所需顺序拼接聚合物符号而生成，并通过测量聚合物字符串的前缀和后缀（或所有子字符串）的质量来读取 [19], [20]。为了仅根据聚合物字符串混合物的前缀和后缀组成，确保能够唯一重建这些字符串，需要遵循一个更复杂的流程，详见 [21], [22]。其中，二进制 $B_h$- 序列（码）[1]，[3]，[4] 用于确保相同长度前缀的质量总和能够唯一确定字符串本身。由于实际中用于表示 1 的聚合物质量远高于用于表示 0 的聚合物，质量差异可能导致严重的碎裂损失，并显著增加化学合成成本，因此宜采用重量相对较小的二进制 $B_2$-序列。这促使我们提出恒定重量 $B_2$-序列设计问题。

我们的主要研究成果包括信息论上界和前缀-后缀拆分上界，它们限制了重量 $\omega$ 随 $n$线性增长的恒定重量二进制 $B_2$-序列的规模。与无约束情形不同，最强上界由一个必须通过数值方法求解的优化问题给出。此外，我们还基于 Sidon 序列给出了构造性的下界。

本文组织如下。第II 节介绍符号、相关概念，以及将 [1] 中的方法推广至定重量二进制字符串情形。第II-A节给出我们的主要结果，即目前已知的定重量二进制$B_2$序列规模的最优上界。第III节给出构造性下界。

## II. 预备知识与熵界

我们用花体大写字母表示集合，用粗体小写字母表示向量。集合的基数用大写字母表示。我们也用 $[n]$ 表示集合 $\{1,2,\ldots,n\}$。除非另有说明，所有对数均以 2 为底。

若一个二进制向量集合 $\mathcal{A}_n \subset \{0,1\}^n$ 满足所有不同字符串对 $\mathbf{c}_1+\mathbf{c}_2$, $\mathbf{c}_1$,$\mathbf{c}_2 \in \mathcal{A}_n$ 的实数值和各不相同，则称其为 $B_2$- 序列集合。若一个 $B_2$-序列集合$\mathcal{A}_n^\omega$ 中的每个向量 $\mathbf{c}=(c_1,\ldots,c_n) \in \mathcal{A}_n^\omega$ 的汉明重量均为 $|\mathbf{c}|, \sum_{i=1}^n c_i=\omega$，则称该集合具有恒定重量 $\omega$ 。

令 $A_n^\omega$ 表示重量为 $\omega$的最大恒定重量 $B_2$-序列集合 $\mathcal{A}_n^\omega$ 的大小。我们关注当 $\bar{\omega}=\frac{\omega}{n}$为常数时， $A_n^\omega$的渐近行为；更准确地说，我们关注渐近码率 $R_{\bar{\omega}}=\limsup_{n\to\infty} \log_n^{A_n^{\bar{\omega}n}}$。[1], [3] 中确立的结果表明，对于不受限制的二进制$B_2$-序列集合，其渐近码率满足 $\leq 0.5753$；这也确立了任意 $\bar{\omega}\in[0,1]$下的 $R_{\bar{\omega}} \leq 0.5753$ 。我们希望改进 $\bar{\omega}\in[0,1/2)$ 时 $R_{\bar{\omega}}$ 的这个上界。

利用信息论论证（在不受限制序列的情形下，这一论证归功于 Katona），可以得到关于 $R_{\bar{\omega}}$ 的一个简单渐近上界：假设所有可能的码向量有序对 $(\boldsymbol{C},\boldsymbol{C}')$ 构成的集合服从均匀概率分布，并利用每一对都会产生唯一和这一事实。注意， $C$ 和 $C'$ 是独立选取的，但允许相等。为了得到关于 $R_{\bar{\omega}}$的上界，我们假设长度为 $n$ 的定重量 $B_2$ 码的大小等于 $A_n^\omega$。于是，所有码向量有序对的熵等于 2 $\log(A_n^\omega)$。令$G=\mathbb{I}(\boldsymbol{C}>\boldsymbol{C}')$为一个指示随机变量，用来表示 $\boldsymbol{c}$ 按字典序排列高于 $\boldsymbol{c}'$这一事件。由于所有无序对的和都各不相同，因此存在一个双射

它将给定 $G$ 时，数对 $(\boldsymbol{C},\boldsymbol{C}')$ 的条件概率空间映射到由随机变量 $X_i = C_i + C'_i$、 $i \in [n]$组成的概率空间，这些随机变量表示和向量的各个坐标。因此，

$$\begin{aligned} 2\log(A_n^\omega) &= H(\boldsymbol{C},\boldsymbol{C}'|G) + H(G) \\ &= H(X_1, X_2, \ldots, X_n|G) + H(G) \\ &\le H(X_1) + H(X_2) + \ldots + H(X_n) + 1. \end{aligned} \tag{1}$$

假设在该 $B_2$ 序列码的第 $i$ 个坐标（即 $C_i$）观察到 1 的概率为 $p_i, i = 1, \ldots, n$。已知二进制向量的重量为$\omega$，我们有 $\frac{1}{n}(p_1 + p_2 + \ldots + p_n) = \frac{\omega}{n} = \bar{\omega}$。所有可能的 2-和的第 $i$ 个坐标的熵等于 $H(X_i)$，也等于 Binomial$(2, p_i)$ 分布的熵，即概率分别为 0、1 和 2 的分布 $\mathrm{p}^i$ 为

$$p_0^i = (1-p_i)^2,\ p_1^i = 2\,p_i\,(1-p_i),\ p_2^i = p_i^2,\ i \in [n]$$

分别。注意， $H(X_i) = H(\mathrm{p}_i)$ 等于

$$\begin{aligned} &-p_i^2 \log p_i^2 - (1-p_i)^2 \log(1-p_i)^2 - 2p_i(1-p_i)\log p_i \\ &-2p_i(1-p_i)\log(1-p_i) - 2p_i(1-p_i) \\ =&-2p_i \text{对数}\ p_i - 2(1-p_i) \text{对数}(1-p_i) - 2p_i + 2p_i^2 = H_{bin}(p_i)。 \end{aligned}$$

由于 $\frac{d^2 H_{bin}(p_i)}{dp_i^2} = -\frac{1}{\ln(2)}\left(\frac{1}{p_i} + \frac{1}{1-p_i}\right) + 4 < 0$，函数 $H_{bin}(p_i)$ 关于 $p_i$ 是凹函数。因此，我们有

$$\begin{aligned} &H(X_1) + H(X_2) + \ldots + H(X_n) \\ &= H_{bin}(p_1) + H_{bin}(p_2) + \ldots + H_{bin}(p_n) \\ &\le nH_{bin}\left(\frac{\sum_{i=1}^n p_i}{n}\right) = nH_{bin}(\bar{\omega}). \end{aligned} \tag{2}$$

注意， $H_{bin}(\bar{\omega})$表示 Binomial$(2, \frac{\omega}{n})$ 分布的熵，即该分布

$$p_0 = (1-\bar{\omega})^2,\ p_1 = 2\,\bar{\omega}\,(1-\bar{\omega}),\ p_2 = \bar{\omega}^2.$$

由 (1) 和 (2) 可得 $2\log(A_n^\omega) \le nH_{bin}(\bar{\omega}) + 1$ 以及

$$R_{\bar{\omega}} \le \frac{1}{2} H_{bin}(\bar{\omega}). \tag{3}$$

虽然 (3) 为 $R_{\bar{\omega}}$ 提供了一个良好的上界起点，但我们注意到，通过考察码向量之差的各个坐标，而非码向量之和，可以得到另一种熵界。具体而言，我们假设在恒重量 $B_2$ 码中，码向量有序对 $(\boldsymbol{C},\boldsymbol{C}')$ 的集合服从均匀分布，因此有序对的熵仍为 $2\log(A_n^\omega)$。令$Y_i = C_i - C'_i$, $i \in [n]$为表示该有序对中向量之差各坐标取值的随机变量序列。那么，以下两种情况之一成立：(1) 当有序对中的两个向量相等时，对于 $i \in [n]$， $Y_i = 0$；(2) 若该有序对中的两个向量不相等，则 $Y_1, Y_2, \ldots, Y_n$ 与该有序对之间存在一一映射。这是因为，若任意两个不同的有序对 $\mathrm{c}_1, \mathrm{c}_2 \in A_n^\omega$和 $\mathrm{c}_1$,$\mathrm{c}_2 \in A_n^\omega$具有相同的差 $\mathrm{c}_1 - \mathrm{c}_2 = \mathrm{c}_3 - \mathrm{c}_4$,，那么这两个有序对 $\mathrm{c}_1$,$\mathrm{c}_4$ 和 $\mathrm{c}_2, \mathrm{c}_3$就会具有相同的和 $\mathrm{c}_1 + \mathrm{c}_4 = \mathrm{c}_2 + \mathrm{c}_3$,，这违反了 $B_2$约束。接下来，令 $E$为事件 $\{(Y_1, \ldots, Y_n) = 0^n\}$。那么，我们可以将 $2\log(A_n^\omega)$ 上界化为

$$\begin{aligned} &H(C, C'|Y_1, Y_2, \ldots, Y_n) + H(Y_1, Y_2, \ldots, Y_n) \\ =&\left(1 - \frac{1}{A_n^\omega}\right) H(C, C'|Y_1, Y_2, \ldots, Y_n, E^c) \\ &+ \frac{1}{A_n^\omega} H(C, C'|Y_1, Y_2, \ldots, Y_n, E) + H(Y_1, Y_2, \ldots, Y_n) \\ =&\frac{\log A_n^\omega}{A_n^\omega} + H(Y_1, Y_2, \ldots, Y_n) \\ \le&H(Y_1) + H(Y_2) + \ldots + H(Y_n) + \frac{\log A_n^\omega}{A_n^\omega}, \end{aligned}$$

其中 $(1 - \frac{1}{A_n^\omega})$ 等于一对中的两个码向量不相等的概率。假设 $P\{C_i = 1\} = q_i$。那么，与之前一样，我们有 $\frac{1}{n}(q_1 + q_2 + \ldots + q_n) = \bar{\omega}$。 $Y_i$ 的分布 $\mathrm{q}^i$ 为

$$q_0^i = q_i^2 + (1-q_i)^2,\ q_1^i = q_i(1-q_i),\ q_{-1}^i = q_i(1-q_i),\ i \in [n].$$

由于熵函数 $H(\mathrm{q}^i)$ 关于 $\mathrm{q}^i$ 是凹的，

$$\begin{aligned} &H(Y_1) + H(Y_2) + \ldots + H(Y_n) \\ =&H(\mathbf{q}^1) + H(\mathbf{q}^2) + \ldots + H(\mathbf{q}^n) \\ \le&nH\left(\frac{\mathbf{q}^1 + \mathbf{q}^2 + \ldots + \mathbf{q}^n}{n}\right) \\ =&n\Big(-t\log t - (1-t)\log\left(\frac{1-t}{2}\right)\Big), \end{aligned}$$

其中 $t = \frac{\sum_{i=1}^n [q_i^2 + (1-q_i)^2]}{n}$。注意，根据詹森不等式，$t \ge \bar{\omega}^2 + (1-\bar{\omega})^2 \ge \frac{1}{2}$，并且对于函数 $-t\log t - (1-t)\log(\frac{1-t}{2})$，其值随 $t \ge \frac{1}{2}$ 的增大而减小。因此，由前一个不等式可得

$$H(Y_1) + H(Y_2) + \ldots + H(Y_n) \le nH(\mathbf{q}),$$

其中，分布 $\mathrm{q}$ 定义为

$$q_0 = \bar{\omega}^2 + (1-\bar{\omega})^2,\ q_1 = \bar{\omega}(1-\bar{\omega}),\ q_{-1} = \bar{\omega}(1-\bar{\omega}).$$

因此， $2\log(A_n^\omega) \le nH(\mathrm{q}) + \log_{A_n^\omega} A_n^\omega$，并且

$$R_{\bar{\omega}} \le \frac{1}{2} H(\mathbf{q}). \tag{4}$$

容易证明，上界 (3) 严格优于 (4)。然而，正如我们稍后将看到的，与直接基于信息论论证得到的上界相比，界 (4) 更适用于推导上界。

表I中列出了本节推导的信息论界的数值，以及后文将介绍的改进上界的数值，适用于不同的权重 $\omega$，其随 $n$ 线性变化。可以看出，随着 $\omega$ 减小，所有上界的数值都越来越接近。表中还列出了其他结果，包括下界，将在最后一节中进一步讨论。

表 I 二元恒定重量 $B_2$序列大小的上界和下界

| $\bar{\omega}$ 的值 | 0.5 | 0.4 | 0.345 | 0.2 | 0.1 | 0.05 | 0.02 |
|---|---|---|---|---|---|---|---|
| 熵界 (3) | 0.75 | 0.731 | 0.704 | 0.562 | 0.379 | 0.239 | 0.122 |
| 熵界 (4) | 0.75 | 0.739 | 0.723 | 0.612 | 0.43 | 0.274 | 0.139 |
| 上界（5） | 0.6 | 0.6 | 0.594 | 0.515 | 0.365 | 0.235 | 0.121 |
| 上界（27） | 0.6 | 0.59 | 0.575 | 0.487 | 0.349 | 0.228 | 0.12 |
| 下界 | 0.25 | 0.259 | 0.263 | 0.232 | 0.166 | 0.108 | 0.056 |

A. 改进的上界

通过调整并推广 $R_{\bar{\omega}}$ 近期针对无约束 $B_2$码向量提出的一种方法，我们可以为受约束的[4] 码本设计得到 $B_2$ 的改进上界。其证明结合了熵界与前缀-后缀分解方法，该方法最早见于 [2]。为完整起见，我们首先介绍如何将前缀-后缀分解方法扩展到恒定重量 $B_2$ 码本。随后，我们通过将两种界——信息论界和前缀-后缀界——结合起来加以改进[4]。针对一般码本设计的方法与我们的方法之间的主要差异 [4], [2] 在于，我们采用了更复杂的熵界，根据前缀的重量对码向量进行分组$\mathbf{c} \in \mathcal{A}_n^{\omega}$，并运用专门的计数技巧。值得注意的是， [4] 中的方法并未改进仅通过前缀-后缀分解所能得到的结果 [2], ，而我们的方案则改进了恒定重量码本的熵方法和前缀-后缀方法。证明见附录A。

引理1. 将每个码向量拆分为 $\mathbf{c} = \mathbf{ab} \in \mathcal{A}_n^{\omega}$，其中 $\mathbf{a} \in \{0,1\}^e$ 且 $\mathbf{b} \in \{0,1\}^{n-e}$。令 $e = \bar{e}n$，其中 $\bar{e}$是 $[0,1]$中的常数。则，$R_{\bar{\omega}}$ 可以上界为

$$\min_{\bar{e}\in[0,1]} \max_{\bar{\omega}'\in[\max\{0,\bar{\omega}-1+\bar{e}\},\min\{\bar{e},\bar{\omega}\}]} \max\left\{ H(\frac{\bar{\omega}'}{\bar{e}})\cdot\bar{e}, \frac{1}{2}\left[H(\frac{\bar{\omega}'}{\bar{e}})\cdot\bar{e} + H(\frac{2\bar{\omega}''}{1-\bar{e}})\cdot(1-\bar{e}) + 2\bar{\omega}''\right]\right\}, \tag{5}$$

其中 $\bar{\omega}'' = \min\{\bar{\omega} - \bar{\omega}', \frac{1-\bar{e}}{4}\}$。通过对 $\bar{e}$进行优化，可以得到不同 $\omega$'s 对应的上界数值。部分结果见表 I。

基于上述结果及其证明，接下来我们将阐述主要结果，它给出了关于恒定重量二进制 $B_2$序列的一个更紧的渐近上界。令

$$\mathcal{B}_n^{\omega'} = \{\mathbf{c} : \mathbf{c} = \mathbf{ab} \in \mathcal{A}_n^{\omega},\ \mathbf{a} \in \{0,1\}^e,\ |\mathbf{a}| = \omega'\}, \tag{6}$$

其中 $e = \bar{e}n$ 和 $\omega' = \bar{\omega}'n$ 为常数，满足 $\bar{e}, \bar{\omega}' \in [0,1]$；令 $\mathcal{A}_n^{\omega}$ 中前缀长度为 $e$ 且重量为 $\omega' \in [\max\{0, \omega - n + e\}, \min\{e, \omega\}]$ 的码向量集合。为方便记号，我们也用 $f = n - e$ 表示后缀的长度。注意，

$$A_n^{\omega} = \sum_{\omega'\in[\max\{0,\omega-f\},\min\{e,\omega\}]} |\mathcal{B}_n^{\omega'}|. \tag{7}$$

因此，我们需要为任意$n$, $e$ 以及 $\omega' \in [\max\{0, \omega - f\}, \min\{e, \omega\}]$，确定 $|\mathcal{B}_n^{\omega'}|$ 的上界。

引理 2. 对于任意 $\omega' \in [\max\{0, \omega - f\}, \min\{e, \omega\}]$，有

$$\log|\mathcal{B}_n^{\omega'}| \le \max\left\{ e\,H\left(\frac{\omega'}{e}\right) + \log n, \frac{1}{2}\left[e\,H\left(\frac{\omega'}{e}\right) + f H(p_0, p_1, p_{-1})\right] + 1\right\}, \tag{8}$$

其中 $p_0 = \frac{\omega''^2 + f - \omega''^2}{f^2}$ $p_1 = p_{-1} = \frac{1-p_0}{2}$ $\omega'' = ()()$、$\omega - \omega'$ $H\ x = -x$ $x - 1 - x$ $1 -$ 和 。函数 ( ) log ( ) log($x$) 表示二元香农熵函数，而函数 $H(p_0, p_1, p_{-1}) = -p_0 \log p_0 - p_1 \log p_1 - p_{-1} \log p_{-1}$表示上述分布为 ($p_0$, $p_1, p_{-1}$) 的三元随机变量的熵。

证明。固定 $\omega'$并令 $B_n^{\omega'} = |\mathcal{B}_n^{\omega'}|$。令 $\{\mathbf{a}_1、\ldots、\mathbf{a}_r\} = \{\mathbf{a}: \mathbf{ab} \in \mathcal{B}_n^{\omega'}$（其中某个 b}）为 $\mathcal{B}_n^{\omega'}$中所有码向量的可能前缀集合。注意，每个 $\mathbf{a}_i$的重量为 $\omega'$，且此类向量至多有 $r$ 个不同取值，其中 $r \le \binom{e}{\omega'}$。令 $\mathcal{S}_i = \{\mathbf{b}: \mathbf{a}_i\mathbf{b} \in \mathcal{B}_n^{\omega'}\}$, $i \in [r]$为 $\mathcal{B}_n^{\omega'}$ 中以 $\mathbf{a}_i$ 为前缀的码向量的后缀集合（可能为空）。于是，

$$B_n^{\omega'} = \sum_{i=1}^{r} |\mathcal{S}_i|. \tag{9}$$

现在考虑属于同一组 $\mathcal{S}_i$ （对于某个 $i \in [r]$）的所有后缀对的集合，记作

$$\mathcal{D} = \cup_{i=1}^{r} \{(\mathbf{b}_1, \mathbf{b}_2) : \mathbf{b}_1, \mathbf{b}_2 \in \mathcal{S}_i\}. \tag{10}$$

注意，$\mathbf{b}_1$ 和 $\mathbf{b}_2$ 可以相同；并且，只要 $\mathbf{b}_1$,$\mathbf{b}_2 \in \mathcal{S}_i$, $i \in [r]$, 互不相同，($\mathbf{b}_1$,$\mathbf{b}_2$) 和 ($\mathbf{b}_2$,$\mathbf{b}_1$) 就被视为两个不同的对。因此， $\mathcal{D}$ 是一个多重集。于是

$$|\mathcal{D}| = \sum_{i=1}^{r} |\mathcal{S}_i|^2 \ge \frac{(\sum_i^r |\mathcal{S}_i|)^2}{r} = \frac{(B_n^{\omega'})^2}{r}, \tag{11}$$

其中，该界由柯西-施瓦茨不等式得出。此外，考虑 $\mathcal{D}$ 中所有数对之间的差，

$$\mathcal{Z} = \{\mathbf{b}_1 - \mathbf{b}_2 : (\mathbf{b}_1, \mathbf{b}_2) \in \mathcal{D}\}, \tag{12}$$

其中 $\mathcal{Z}$是一个多重集。 $\mathcal{Z}$中 $0^f$（长度为 $f$ 的全零向量）的重数恰好为 $B_n^{\omega'}$。此外， $\mathcal{Z}$ 中任意非零元素的重数恰好为一。为说明这一点，反设存在不同的不等元素对 ($\mathbf{b}_1$,$\mathbf{b}_2$)、($\mathbf{b}_3$,$\mathbf{b}_4$) $\in \mathcal{D}$，满足 $\mathbf{b}_1 - \mathbf{b}_2 = \mathbf{b}_3 - \mathbf{b}_4$。根据 $\mathcal{D}$ 的定义，对于某个 $i \in [r]$，有 $\mathbf{b}_1$,$\mathbf{b}_2 \in \mathcal{S}_i$；对于某个 $j \in [1, r]$，有 $\mathbf{b}_3$,$\mathbf{b}_4 \in \mathcal{S}_j$。这意味着 $\mathbf{a}_i\mathbf{b}_1, \mathbf{a}_i\mathbf{b}_2, \mathbf{a}_j\mathbf{b}_3, \mathbf{a}_j\mathbf{b}_4$ 是 $\mathcal{B}_n^{\omega'}$中的码向量。于是，

$$\mathbf{a}_i\mathbf{b}_1 - \mathbf{a}_i\mathbf{b}_2 = \mathbf{a}_j\mathbf{b}_3 - \mathbf{a}_j\mathbf{b}_4, \tag{13}$$

这与 $\mathcal{B}_n^{\omega'}$ 是一个二进制 $B_2$-序列这一事实相矛盾。

接下来，我们将 [4] 中的信息论论证推导加以推广。从 $\mathcal{D}$ 中均匀随机地选取一对，并将结果记为一对随机变量

$(\boldsymbol{X},\boldsymbol{Y})$。那么，差 $\boldsymbol{X}-\boldsymbol{Y}$ 在 $\mathcal{Z}\backslash\{0^f\}$ 上均匀分布（即，在给定 $\boldsymbol{X}-\boldsymbol{Y}\neq 0^f$ 的条件下）。令 $E$ 为事件 $\{\boldsymbol{X}-\boldsymbol{Y}\neq 0^f\}$。那么，$H(\boldsymbol{X},\boldsymbol{Y})$ 等于

$$\begin{aligned}&H(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{X}-\boldsymbol{Y})=H(\boldsymbol{X}-\boldsymbol{Y})+H(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{X}-\boldsymbol{Y})\\&=H(\boldsymbol{X}-\boldsymbol{Y})+\Pr(E)H(\boldsymbol{X},Y|\boldsymbol{X}-\boldsymbol{Y},E)\\&+\Pr(E^c)H(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{X}-\boldsymbol{Y},E^c).\end{aligned}\tag{14}$$

显然，$H(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{X}-\boldsymbol{Y},E^c)=0$，因为 $\mathcal{Z}$ 中不同的非零元素的重数均为 1。此外，

$$H(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{X}-\boldsymbol{Y},E)=\log B_n^{\omega'},\ \Pr(E)=\frac{B_n^{\omega'}}{|\mathcal{D}|}\le\frac{r}{B_n^{\omega'}},$$

其中不等式由（11）得出。因此，

$$\Pr(E)H(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{X}-\boldsymbol{Y},E)\le\frac{r}{B_n^{\omega'}}\log B_n^{\omega'}.\tag{15}$$

将（14）、（15）与 $H(\boldsymbol{X},\boldsymbol{Y})=\log|\mathcal{D}|\ge\frac{(B_n^{\omega'})^2}{r}$ 结合起来，我们得到

$$\log(\frac{(B_n^{\omega'})^2}{r})\le H(X-\boldsymbol{Y})+\frac{r}{B_n^{\omega'}}\log B_n^{\omega'}.\tag{16}$$

为了得到 $B_n^{\omega'}$ 的上界，我们需要针对常重向量对 $H(\boldsymbol{X}-\boldsymbol{Y})$ 的上界进行专门推导。

令 $n_{ij},\ i\in[f],\ j\in[r]$ 分别表示 $\mathcal{S}_j$ 中第 $i$ 个坐标为 1 的后缀数量。根据熵函数的次可加性，我们有

$$H(\boldsymbol{X}-\boldsymbol{Y})\le\sum_{i=1}^{f}H(X_i-Y_i)\le fH(\frac{\sum_{i=1}^{f}\mathbf{p}^i}{f}),\tag{17}$$

其中，$\mathbf{p}^i=(p_0^i,p_1^i,p_{-1}^i)$ 是 $X_i-Y_i$ 的分布，

$$\begin{aligned}&p_0^i=\frac{\sum_{j=1}^{r}[n_{ij}^2+(|\mathcal{S}_j|-n_{ij})^2]}{\sum_{j=1}^{r}|\mathcal{S}_j|^2},\ p_1^i=\frac{\sum_{j=1}^{r}n_{ij}(|\mathcal{S}_j|-n_{ij})}{\sum_{j=1}^{r}|\mathcal{S}_j|^2},\\&p_{-1}^i=\frac{\sum_{j=1}^{r}n_{ij}(|\mathcal{S}_j|-n_{ij})}{\sum_{j=1}^{r}|\mathcal{S}_j|^2}.\end{aligned}\tag{18}$$

我们接下来证明，平均分布记作

$$\mathbf{p}^*=\frac{\sum_{i=1}^{f}\mathbf{p}^i}{f},\tag{19}$$

满足 $p_0^*\ge\frac{(\omega')^2+(f-\omega'')^2}{f^2}$，其中 $\omega''=\omega-\omega'$ 是 $\mathcal{B}_n^{\omega'}$ 中所有码向量后缀的重量，即 $\mathcal{S}_j$中向量的重量，$j\in[r]$。由式（18）可得

$$\begin{aligned}p_0^*=&\frac{\sum_{i=1}^{f}p_0^i}{f}=\frac{\sum_{j=1}^{r}\left(\sum_{i=1}^{f}[n_{ij}^2+(|\mathcal{S}_j|-n_{ij})^2]\right)}{f(\sum_{j=1}^{r}|\mathcal{S}_j|^2)}\\\ge&\frac{\sum_{j=1}^{r}\left(\frac{(\sum_{i=1}^{f}n_{ij})^2}{f}+\frac{(\sum_{i=1}^{f}(|\mathcal{S}_j|-n_{ij})^2}{f}\right)}{f(\sum_{j=1}^{r}|\mathcal{S}_j|^2)}\\\overset{(a)}{=}&\frac{\sum_{j=1}^{r}\left(|\mathcal{S}_j|^2(\omega'')^2+|\mathcal{S}_j|^2(f-\omega'')^2\right)}{f^2(\sum_{j=1}^{r}|\mathcal{S}_j|^2)}\\=&\frac{(\omega'')^2+(f-\omega'')^2}{f^2},\end{aligned}\tag{20}$$

其中 $(a)$ 由以下事实可知：$\mathcal{S}_j$ 中的向量重量固定且等于 $\omega''$。此外，我们有 $p_1^*=p_{-1}^*$，因为 $p_1^i=p_{-1}^i$ 由式（18）得出。注意，$p_0^*\ge\frac{1}{2}$，且熵函数 $H(p_0^*,\frac{1-p_0^*}{2},\frac{1-p_0^*}{2})$ 关于 $p_0^*$ 在 $p_0^*\ge\frac{1}{2}$ 时递减。因此，结合式（20）和式（17），可得

$$H(\boldsymbol{X}-\boldsymbol{Y})\le fH(p_1,p_0,p_{-1}),\tag{21}$$

其中 $p_0=\frac{(\omega'')^2+(f-\omega'')^2}{f^2}$ 且 $p_1=p_{-1}=\frac{1-p_0}{2}$。结合 (16) 和 (21)，得到

$$\log B_n^{\omega'}\le\frac{1}{2}(\log r+fH(p_0,p_1,p_{-1}))+\frac{r}{2B_n^{\omega'}}\log B_n^{\omega'}.\tag{22}$$

最后，为证明 (8)，反设

$$\begin{aligned}&\log B_n^{\omega'}>eH\left(\frac{\omega'}{e}\right)+\log n,\ \text{and}\\&\log B_n^{\omega'}>\frac{1}{2}\left[eH\left(\frac{\omega'}{e}\right)+fH(p_0,p_1,p_{-1})\right]+1,\end{aligned}\tag{23}$$

由 (22)、(23)，以及 $r\le 2^{eH(\frac{\omega'}{e})}$ 这一事实，可得

$$1<\frac{r}{2B_n^{\omega'}}\log B_n^{\omega'},\tag{24}$$

以及下面这个与 (24) 矛盾的不等式：

$$B_n^{\omega'}>n2^{eH\left(\frac{\omega'}{e}\right)}\ge nr\ge r\log B_n^{w'}.\tag{25}$$

□

结合引理 2 和 (7)，我们得出

$$\begin{aligned}\log A_n^{\omega}\le&\max_{\omega'\in[\max\{0,\omega-f\},\min\{e,\omega\}]}\max\Big\{eH\left(\frac{\omega'}{e}\right),\\&\frac{1}{2}\left[eH\left(\frac{\omega'}{e}\right)+fH(p_0,p_1,p_{-1})\right]\Big\}+\log n,\end{aligned}\tag{26}$$

其中 $p_0=\frac{\omega''^2+f-\omega''^2}{f^2}$ $p_1=p_{-1}=\frac{1-p_0}{2}$ $\omega''=()()$，且，$\omega-\omega'$ $e$ 对于任意满足 $e+f=n$ 的和 $f$ 选择。因此，$R_{\bar{\omega}}$ 的一个上界为

$$\begin{aligned}R_{\bar{\omega}}\le&\max_{\bar{e}\in[0,1]}\max_{\bar{\omega}'\in[\max\{0,\bar{\omega}-1+\bar{e}\},\min\{\bar{e},\bar{\omega}\}]}\max\Big\{\bar{e}H\left(\frac{\bar{\omega}'}{\bar{e}}\right),\\&\frac{1}{2}\left[\bar{e}H\left(\frac{\bar{\omega}'}{\bar{e}}\right)+(1-\bar{e})H(p_0,p_1,p_{-1})\right]\Big\},\end{aligned}\tag{27}$$

其中 $p_0=\frac{}{1-\bar{e}^2}$ $p_1=p_{-1}=\frac{1-p_0}{2}$ $\bar{\omega}''=$ $\bar{\omega}-\bar{\omega}'$ $(\bar{\omega}'')^2+(1-\bar{e}-\bar{\omega}'')^2()$，，且。注意，上界 ( ) 小于 [2] 当 $\omega<\frac{n}{2}$ 时的上界，且小于已知的最佳上界 0。5753 对于无约束二元 $B_2$-序列，该上界由 [3] 在 $\omega\le 0.345n$时报告。关于上界实际值的更多详情，请参见表 I 。

27 0.6

### III. 下界

接下来，我们介绍一种用于构造恒定重量二元$B_2$码$\mathcal{A}_n^\omega$的构造，其大小为 $(\frac{n}{\omega})^{\frac{\omega}{2}+o(\omega)}$并且

$$(\left\lfloor\frac{n}{\omega}\right\rfloor)^{\frac{\omega\lceil\frac{n}{\omega}\rceil-n}{2(\lceil\frac{n}{\omega}\rceil-\lfloor\frac{n}{\omega}\rfloor)}}(\left\lceil\frac{n}{\omega}\right\rceil)^{\frac{n-\omega\lfloor\frac{n}{\omega}\rfloor}{2(\lceil\frac{n}{\omega}\rceil-\lfloor\frac{n}{\omega}\rfloor)}}2^{o(\omega)},$$

分别适用于 $\frac{n}{\omega}$为整数和非整数实数的情况。该构造表明，当 $\frac{1}{\omega}$为整数时， $R_{\bar{\omega}} \geq \frac{\bar{\omega}}{2}\log(\frac{1}{\omega})$，而

$$R_{\bar{\omega}} \geq \frac{\bar{\omega}\lceil\frac{1}{\bar{\omega}}\rceil-1}{2(\lceil\frac{1}{\bar{\omega}}\rceil-\lfloor\frac{1}{\bar{\omega}}\rfloor)}\log(\left\lfloor\frac{1}{\bar{\omega}}\right\rfloor)+\frac{1-\bar{\omega}\lfloor\frac{1}{\bar{\omega}}\rfloor}{2(\lceil\frac{1}{\bar{\omega}}\rceil-\lfloor\frac{1}{\bar{\omega}}\rfloor)}\log(\left\lceil\frac{1}{\bar{\omega}}\right\rceil)$$

否则。一个重要的观察是，我们的构造虽然在概念上简单，但得到的码率至少达到无约束恒定重量码最大可能码率$\binom{n}{\omega}$的 $1\ 4_{\circ}$

接下来，为简便起见，我们假设 $\frac{n}{\omega}$ 为整数，并且与前文一样，令 $\omega \leq \frac{n}{2}$ 。我们的思路是找到一个满射线性映射 $F$: $\{0,1,2\}^n \to [0,(\frac{n}{\omega})^\omega-1]$，将字母表 $\{0,1,2\}$ 上任意长度为 $n$ 的向量映射为 $[0,(\frac{n}{\omega})^\omega-1]$ 中的一个整数。更准确地说，映射$F$需要满足以下两个性质：

(A) 对于任意整数 $i \in [0,(\frac{n}{\omega})^\omega-1]$，存在一个重量为 $\omega$的向量 c $\in\{0,1\}^n$，使得 $F$(c) $=i_{\circ}$ (B) 对于任意向量 c、c′ $\in\{0,1\}^n$，有 $F$(c) $+F$(c′) $=F$(c $+$c′)。注意，c $+$c′ $\in\{0,1,2\}^n{}_{\circ}$

给定映射 $F$，我们从集合 $[0,(\frac{n}{\omega})^\omega-1]$构造一个整数 Sidon 集[23]。根据 Bose-Chawla 构造，存在一个由整数构成的集合 $\{i_1,\ldots,i_{(\frac{n}{\omega}\ \frac{\omega}{2}+o\ \omega}\}\subset[0,\ \frac{n}{\omega}\ ^\omega-)\ (\ )\ (\ ),$ $1]\qquad\frac{n}{\omega}\ \frac{\omega}{2}+o\ \omega$ 其大小为 ( ) ( )，且集合中任意两个整数之和都不相同。于是，由映射$F$的性质 (A) 可知，对于每个 $i_j,\ j\in[1,(\frac{n}{\omega})^{\frac{\omega}{2}+o(\omega)],$，都存在一个重量为 $\omega$ 的向量 c$_j \in\{0,1\}^n$，使得 $F$(c$_j$) $=i_{j\circ}$ 最后，根据映射 $F$ 的性质 (B) 以及集合$\{i_1,\ldots,i_{(\frac{n}{\omega})}\ ^{\frac{\omega}{2}+o(\omega)}\}$的定义，集合 $\{$c1, . . . , c$_{(\frac{n}{\omega})}\ ^{\frac{\omega}{2}+o(\omega)}\}$是一个重量为 $\omega$ 的二元$B_2$码本。

对于任意整数 $k\in[0,n-1]$，令 $k=a_k\omega+b_k$，其中 $a_k=\lfloor\frac{k}{\omega}\rfloor$且 $b_k=k$模 $\omega$。对于任意 c $\in\{0,1,2\}^n$，定义

$$F(\mathbf{c})\triangleq\sum_{i=1}^{n}a_{i-1}\left(\frac{n}{\omega}\right)^{b_{i-1}}c_i. \tag{28}$$

显然， $F$ 满足性质 (B)。为证明 $F$满足 (A)，注意任意整数 $m\in[0,(\frac{n}{\omega})^\omega-1]$都有一个$\frac{n}{\omega}$进制表示 $m=\sum_{i=0}^{\omega-1}m_i\left(\frac{n}{\omega}\right)^i$，其中 $m_i\in[0,\frac{n}{\omega}-1]$。令 c$_m$为 $\{0,1\}^n$中的一个向量，其值为 1 的位的索引由 $\{m_i\omega+i$ 给出： $i\in[0,\omega-1]\}$。那么，c$_m$的重量为 $\omega$，且 $F$(c$_m$) $=\sum_{i=0}^{\omega-1}a_{m_i\omega+i}(\frac{n}{\omega})^{b_{m_i\omega+i}}=m_{\circ}$ 因此$F$ 满足性质 (B)。

### 参考文献


[1] B. Lindström，“从和确定两个向量”，组合理论杂志，第 6 卷，第 4 期，第 402–407 页，1969 年。[2] ——，“关于向量的 b2 序列”，数论杂志，第 4 卷，第 3 期，第 261–265 页，1972 年。

[3] G. Cohen、S. Litsyn 和 G. Zémor，“二进制 b2 序列：一个新的上界”，组合理论杂志 A 辑，第 94 卷，第 1 期，第 152–155 页，2001 年。[4] S. Della Fiore 和 M. Dalai，“关于 2-可分码和 b2 码的注记”，离散数学，第 345 卷，第 3 期，第 112751 页，2022 年。[5] A. G. D'yachkov，“筛查实验设计讲义”，arXiv 预印本 arXiv:1401.7505，2014 年。[6] M. Bouvel、V. Grebinski 和 G. Kucherov，“受生物信息学应用启发的图上组合搜索：简要综述”，载于国际计算机科学图论概念研讨会。施普林格，2005 年，第 16–27 页。[7] V. Gritsenko、G. Kabatiansky、V. Lebedev 和 A. Maevskiy，“噪声多址加法信道的签名码”，设计、码与密码学，第 82 卷，第 1 期，第 293–299 页，2017 年。[8] Y. Polyanskiy，“大规模多址接入的信息论视角”，短期课程（幻灯片），Skoltech 技术学院，俄罗斯莫斯科，2018 年。[9] D. Boneh 和 J. Shaw，“数字数据的抗合谋指纹识别”，IEEE 信息论汇刊，第 44 卷，第 5 期，第 1897–1905 页，1998 年。[10] N. Goldman、P. Bertone、S. Chen、C. Dessimoz、E. M. LeProust、B. Sipos 和 E. Birney，“迈向合成 DNA 中实用、高容量、低维护的信息存储”，《自然》，第 494 卷，第 7435 期，第 77 页，2013 年。[11] G. M. Church、Y. Gao 和 S. Kosuri，“DNA 中的下一代数字信息存储”，《科学》，第 337 卷，第 6102 期，第 1628–1628 页，2012 年。[在线]。可访问：http://science.sciencemag.org/content/

337 6102 1628 / / R. N. Grass、R. [12] Heckel、M. Puddu、D. Paunescu 和 W. J. Stark，“利用纠错码在二氧化硅中的 DNA 上稳健地化学保存数字信息”，《德国应用化学国际版》，第 54 卷，第 8 期，第 2552–2555 页，2015 年。[13] S. H. T. Yazdi、Y. Yuan、J. Ma、H. Zhao 和 O. Milenkovic，“一种可重写、随机访问的 DNA 数据存储系统”，《科学报告》，第 5 卷，第 1 期，第 1–10 页，2015 年。[14] S. H. T. Yazdi、R. Gabrys 和 O. Milenkovic，“便携且无差错的 DNA 数据存储”，《科学报告》，第 7 卷，第 1 期，第 1–6 页，2017 年。[15] S. K. Tabatabaei、B. Wang、N. B. M. Athreya、B. Enghiad、A. G. Hernandez、C. J. Fields、J.-P. Leburton、D. Soloveichik、H. Zhao 和 O. Milenkovic，“通过酶促切口在天然 DNA 序列上存储数据的 DNA 打孔卡”，《自然·通讯》，第 11 卷，第 1 期，第 1–10 页，2020 年。[16] C. Pan、K. Tabatabaei、S. H. T. Yazdi、A. G. Hernandez、C. M. Schroeder 和 O. Milenkovic，“利用机器学习重建的可重写二维 DNA 数据存储”，bioRxiv，2021 年。[17] S. K. Tabatabaei、B. Pham、C. Pan、J. Liu、S. Chandak、S. A. Shorkey、A. G. Hernandez、A. Aksimentiev、M. Chen、C. M. Schroeder 等，“通过神经网络纳米孔读出处理扩展 DNA 数据存储系统的分子字母表”，《纳米快报》，第 22 卷，第 5 期，第 1905–1914 页，2022 年。[18] C. Laure、D. Karamessini、O. Milenkovic、L. Charles 和 J.-F. Lutz，“二维编码：利用有意分散性提升序列编码聚合物条形码的信息容量”，《德国应用化学》，第 128 卷，第 36 期，第 10 880–10 883 页，2016 年。 [在线]。可访问：https://onlinelibrary.wiley.com/doi/abs/

10 1002 201605279 。/ange. J. [19] Acharya、H. Das、O. Milenkovic、A. Orlitsky 和 S. Pan，“从子串组成重建字符串”，SIAM 离散数学杂志，第 29 卷，第 3 期，第 1340–1371 页，2015 年。[20] S. Pattabiraman、R. Gabrys 和 O. Milenkovic，“基于聚合物的数据存储的重建与纠错码”，载于 2019 年 IEEE 信息论研讨会（ITW）。IEEE，2019 年，第 1–5 页。[21] R. Gabrys、S. Pattabiraman 和 O. Milenkovic，“从前缀和后缀组成重建编码字符串的混合物”，载于 2020 年 IEEE 信息论研讨会（ITW）。IEEE，2021 年，第 1–5 页。[22] ——，“从前缀/后缀组成重建字符串集合”，IEEE 通信汇刊，2022 年。[23] I. Z. Ruzsa，“一个无限 Sidon 序列”，数论杂志，第 68 卷，第 1 期，第 63–71 页，1998 年。

## 附录

我们将每个向量 $\mathbf{c} = \mathbf{ab} \in \mathcal{A}_n^{\omega}$ 拆分为 $\mathbf{a} \in \{0,1\}^e$ 和 $\mathbf{b} \in \{0,1\}^f$。接下来，我们根据 $\mathbf{a}$ 的重量对向量 $\mathbf{c} \in \mathcal{A}_n^{\omega}$ 分组，并使用定义 $\mathcal{B}^{\omega'} = \{\mathbf{c} : \mathbf{c} = \mathbf{ab} \in \mathcal{A}_n^{\omega}、\ \mathbf{a} \in \{0,1\}^e、\ |\mathbf{a}| = \omega'\}$。

引理 3. 对任意 $\omega' \in [\max\{0, \omega - f\}, \min\{e, \omega\}]$，我们有

$$\log |\mathcal{B}_n^{\omega'}| \leq \max \Bigg\{ H(\frac{\omega'}{e}) \cdot e + \log(n+1) + 1, \\ \frac{1}{2}\Big[ H(\frac{\omega'}{e}) \cdot e + \log(n+1) + H(\frac{2\omega''}{f}) \cdot f \\ + 2\omega'' + \log(n(n+1)) \Big] + 1 \Bigg\}, \tag{29}$$

其中 $\omega'' = \min\{\frac{f}{4}, \omega - \omega'\}$ and $H(x) = -x \log_2 x - (1-x) \log_2 (1-x)$ 是熵函数。

证明。证明的第一部分与引理 2 的证明思路相同。令 $\{\mathbf{a} \in \{0,1\}^e \in \mathcal{B}_n^{\omega'}\}$ $\mathcal{S}_i = \{ \quad _i \in \mathcal{B}_n^{\omega'} \}$ $,, \mathbf{a}_r\} = \{\mathbf{a} : \mathbf{a}, \mathbf{ab}$ 对某个 $\mathbf{b}, \mathbf{b} : \mathbf{ab}$，并且 $\mathcal{D} = \cup_{i=1}^{r} \{(\mathbf{b}_1, \mathbf{b}_2) : \mathbf{b}_1, \mathbf{b}_2 \in \mathcal{S}_i\}$。由此得到 (11)。证明的其余部分与引理 2 的证明不同，因为我们使用组合论论证 [2]。

令 $\{\mathbf{v}_1, \ldots, \mathbf{v}_{|\mathcal{D}|}\} = \{\mathbf{b}_1 - \mathbf{b}_2 : (\mathbf{b}_1, \mathbf{b}_2) \in \mathcal{D}\}$ 为 $\mathcal{D}$ 中各对后缀之差构成的向量集合，并计入重数。于是，$0^f$ 在 $\{\mathbf{v}_1, \ldots, \mathbf{v}_{|\mathcal{D}|}\}$ 中的重数为 $B_n^{\omega'}$，而 $\{\mathbf{v}_1, \ldots, \mathbf{v}_{|\mathcal{D}|}\}$ 中每个非零向量的重数均为 1。对于 $i \in [|\mathcal{D}|]$, $j \in [1, f]$，令 $h_{ij} = 1$：若 $\mathbf{v}_i$ 的第 $j$ 位为 0，则取该值；否则取 $h_{ij} = -1$。于是，对于 $j \in [1, f]$ 和 $\sum_{i=1}^{|\mathcal{D}|} \sum_{j=1}^{f} h_{ij} \geq 0$，有 $\sum_{i=1}^{|\mathcal{D}|} h_{ij} \geq 0$。

由于对于每个具有 $k$ 个非零项的 $\mathbf{v}_i$，都有 $\sum_{j=1}^{f} h_{ij} = f - 2k$，因此，对于任何至少有 $\frac{f}{2}$ 个非零项的 $\mathbf{v}_i$，都有 $\sum_{j=1}^{f} h_{ij} \leq 0$。此外，具有 $k$ 个非零项的差向量至多有 $\binom{f}{k} 2^k$ 个。令 $|\mathbf{v}_i|$ 表示 $\mathbf{v}_i$, $i \in [|\mathcal{D}|]$ 中的非零项数。那么，

$$\begin{aligned} 0 \leq \sum_{i=1}^{|\mathcal{D}|} \sum_{j=1}^{f} h_{ij} &\overset{(a)}{\leq} f B_n^{\omega'} + \sum_{i:|\mathbf{v}_i| \geq 1} (\sum_{j=1}^{f} h_{ij}) \\ &\overset{(b)}{\leq} f B_n^{\omega'} + \sum_{i:1 \leq |\mathbf{v}_i| \leq \min\{\frac{f}{2}, 2(\omega - \omega')\}} (\sum_{j=1}^{f} h_{ij}) \\ &- \sum_{i:|\mathbf{v}_i| > \frac{f}{2}} 1 \overset{(c)}{\leq} f B_n^{\omega'} + 2f\omega'' \binom{f}{2\omega''} 2^{2\omega''} - \sum_{i:|\mathbf{v}_i| > \frac{f}{2}} 1, \end{aligned} \tag{30}$$

其中 $\omega'' = \min\{\omega - \omega', \frac{f}{4}\}$，$(a)$ 源于 $0^f$ 在 $\{\mathbf{v}_1, \ldots, \mathbf{v}_s\}$ 中的重数为 $B_n^{\omega'}$；$(b)$ 源于当 $|\mathbf{v}_i| > \frac{f}{2}$ 时 $\sum_{j=1}^{f} h_{ij} < 0$，以及 $|\mathbf{v}_i| \leq 2(\omega - \omega')$；而 $(c)$ 源于满足 $|\mathbf{v}_i| = k$ 的 $\mathbf{v}_i$ 的数量至多为 $\binom{f}{k} 2^k$，以及这样一个事实：数量随着 $k$ 当 $k \leq$ 最小值 $\{\frac{f}{2}, 2(\omega - \omega')\}$。式 (30) 蕴含

$$\begin{aligned} |\mathcal{D}| &\leq (f+1)|\mathcal{B}_n^{\omega'}| + 2(f+1)\omega'' \binom{f}{2\omega''} 2^{2\omega''} \\ &\leq (f+1)|\mathcal{B}_n^{\omega'}| + 2^{fH(\frac{2\omega''}{f}) + 2\omega'' + \log(n(n+1))} \end{aligned} \tag{31}$$

结合式（11）和式（31），可得

$$|\mathcal{B}_n^{\omega'}|^2 \leq (f+1)|\mathcal{B}_n^{\omega'}| r + 2^{fH(\frac{2\omega''}{f}) + 2\omega'' + \log(n(n+1))} r.$$

由于 $r \leq 2^{eH(\frac{\omega'}{e})}$，于是有式（29）。根据引理 3和式（7），可得式（5）。 □