DNA 覆盖深度问题：对偶性、重量分布与应用

Matteo Bertuzzo[1],Alberto Ravagnani[1],和 Eitan Yaakobi[2]

[1]埃因霍温理工大学，埃因霍温，荷兰[2]以色列理工学院，海法，以色列

2026年3月9日

arXiv:2603.06489v1 [cs.IT] 6 Mar 2026

摘要

DNA 数据存储中的覆盖深度问题，旨在计算恢复所有编码链所需读段数的期望值。给定一个线性码的生成矩阵，该数量等于获得满秩所需随机抽取的列数的期望值。虽然 MDS 码在存在时（即在大有限域上）是最优的，但实际应用中可能需要依赖定义在小有限域上的结构化码族。在本文中，我们基于对偶性论证和扩展重量枚举式的概念，开发了用于求解各种线性码的 DNA 覆盖深度问题的组合工具。利用这些方法，我们推导出单纯形码、汉明码、三元戈莱码、扩展三元戈莱码和一阶里德-穆勒码的闭式公式。本文的核心内容是给出一个一般表达式，用高阶域扩张上的重量分布来表示线性码的覆盖深度。

1 引言

全球产生的数字数据量持续快速增长，存储需求不断攀升，现有供给已无法满足需求，且差距还在持续扩大 [1]。这一趋势凸显了对创新存储技术的迫切需求，这类技术应具备更高的存储密度、更优的效率和更长的使用寿命，突破现有解决方案的局限。在此背景下，DNA存储系统凭借其耐久性、紧凑性和低维护成本，成为一种颇具吸引力的解决方案，尤其适用于长期数据归档 [2], [3]。

DNA数据存储涉及多个步骤的处理流程。First，首先将原始数据从比特字符串编码为基于DNA字母表 A，C，G，T 的序列。随后，这些序列被划分为数据块，并合成为人工DNA分子，称为链。受当前技术限制，合成的链长度最多约为

M.B. 的研究部分由 EuroTech计划 支持。A.R. 的研究由荷兰研究委员会通过 VI.Vidi.203.045、OCENW.KLEIN.539 项资助，并由欧盟委员会提供支持。E.Y. 的研究由欧盟资助（ERC，DNAStorage，101045114）。文中表达的观点和意见仅代表作者本人，并不一定反映欧盟或欧洲研究委员会执行机构的立场。欧盟和资助机构均不对这些观点和意见承担责任。本文部分研究成果曾在 ISIT 2025 上发表。

Email addresses: m.bertuzzo@tue.nl, a.ravagnani@tue.nl, yaakobi@cs.technion.ac.il。

300 个碱基 [4], ，并且每条编码链都会生成多个带噪副本。生成的DNA分子随后被存放在容器中。要取回存储的信息，需要通过测序过程将链转换回DNA序列，并生成多个无序副本，称为读段。最后一步是对读段进行解码，以重建用户的原始数据。这一过程的一个显著特点是，测序以随机方式访问链。

尽管潜力巨大 [5–11], ，与其他存储技术相比，DNA存储目前仍面临实际限制：由于DNA测序仪的效率 [2, 12, 13]，其吞吐量相对较低且成本较高。这些限制与覆盖深度 [14], 这一概念有关，其定义为测序读段数量与设计的DNA链数量之比。该指标在决定测序成本和系统性能方面发挥着关键作用。

在这种情况下，一个核心问题是，平均需要多少读段才能恢复所有编码链。这被称为 *DNA* 覆盖深度问题，其编码理论方面最近在 [15]中首次提出。当数据由线性分组码编码时，覆盖深度问题可以用代数术语表述：给定线性码 $\mathcal{C}$的一个秩为 k的生成矩阵 G，恢复编码链所需的读段数量等于随机抽取列直至得到秩为 k 的矩阵所需抽取列数的期望。近期研究表明，若信息链使用MDS码进行编码，则期望为 $n(H_n - H_{n-k})$，其中 H_i 是第 i个调和数。特别地，MDS 码可达到最优值。然而，这些码仅存在于较大的有限域上。

近年来，人们对覆盖深度问题及相关模型的兴趣日益浓厚。除了 [15], 的完整恢复场景外，已有多项研究探讨了随机访问场景，其目标是高效检索单条信息链 [16–21]。其他扩展还包括将多条链组合起来共同表示一个文件的模型 [22] ，以及对实用编码架构中读写权衡的研究 [23]。此外，研究者还针对组合型基于短聚体的系统 [24–26] 以及采用复合 DNA 字母的存储方案 [27], 提出了进一步的推广。

受 [15], 的研究结果启发，本文研究定义在小有限域上的线性码的覆盖深度问题。我们的主要目标是理解码的结构性质如何决定完整数据恢复所需读段数量的期望值；这一研究方向始于 [28]。为此，我们开发了一套组合与代数工具，核心包括信息集枚举、对偶恒等式和扩展重量枚举式。

这些工具使我们能够为各种经典码族推导出闭式公式。我们展示了如何为单纯形码的期望得到一个简单表达式。随后，利用对偶性论证，我们获得了汉明码、三元戈莱码以及扩展三元戈莱码的显式公式。我们的主要理论贡献，是根据底层码的前几个高阶域扩张的重量分布，给出覆盖深度的一般表达式。这一结果将期望的计算归结为重量枚举。作为应用，我们利用一阶里德-穆勒码已知的扩展重量枚举式，以闭式公式计算其覆盖深度。

本文其余部分安排如下。第 2 节介绍覆盖深度问题及其代数表述。第 3 节计算单纯形码的期望。第 4 节建立一个对偶恒等式，将期望表示为对偶码的信息集。随后，在第 5 节中利用这一结果推导出汉明码和三元戈莱码的闭式公式。第 6 节给出我们的主要结果：用 $\mathcal{C}$ 的高阶扩域码的重量分布表示覆盖深度的一般表达式。最后，第 7 节将该公式应用于一阶里德-穆勒码，得到其显式表达式。

2 问题陈述

在本文中， q 是一个素数幂， $\mathbb{F}_q$ 是含有 q 个元素的有限域。我们令 k 和 n 为满足 $2 \leq k \leq n$ 的正整数。此外，对于正整数 m，我们将 H_m 记为第 m 个调和数：

$$H_m = \sum_{i=1}^{m} \frac{1}{i}.$$

在典型的基于 DNA 的存储系统中，数据被存储为一个长度为 k 的向量，其各项本身是字母表 $\Sigma = \{A, C, G, T\}$ 上长度为 ℓ 的向量（称为链）。为了能够使用编码理论工具，我们将 Σ^ℓ 嵌入 $\mathbb{F}_q$ ，并使用一个 k 维线性分组码 $\mathcal{C} \subseteq \mathbb{F}_q^n$，将信息向量 $(x_1, \ldots, x_k) \in \mathbb{F}_q^k$ 编码为编码向量 $(y_1, \ldots, y_n) \in \mathbb{F}_q^n$。请注意，为了将 Σ^ℓ 与 $\mathbb{F}_q$ 等同，我们需要 $|\Sigma^\ell| = 4^\ell$ 整除 q；然而，本文不考虑对 q 的任何限制，允许它为任意素数幂。

如果用户希望检索所存储的信息，编码链首先经过扩增过程，随后进行测序。这会生成每条链的若干未排序副本，这些副本与原始链相比可能包含错误。它们被称为读段。为简化分析，本文假设在上述任何步骤中均不产生错误，因此该过程的最终输出是一个读段的多重集，且未指定任何顺序。

本文的出发点是一个关于 DNA 数据存储覆盖深度问题的结果 [15], ，其中需要恢复所有信息链。由于这些 k 信息链使用生成矩阵 $G \in \mathbb{F}_q^{k \times n}$进行编码，因此编码链与 G的列之间存在一一对应关系，即第 i条编码链对应于生成矩阵的第 i 列；因此，恢复第 i条信息链等价于恢复第 i个标准基向量，也就是说，该向量必须位于 G中已恢复列的张成空间内，因为我们可以将这些列视为 $\mathbb{F}_q^k$中的向量。受这些结果启发，我们定义本文研究的第一个问题。

问题 A （DNA 覆盖深度问题）。 设 $G \in \mathbb{F}_q^{k \times n}$的秩为 k。假设从 G 的列中均匀随机地进行有放回抽取，这意味着每一列都可以被抽取多次。计算所需抽取的列数的期望数量，直到所有标准基向量都属于其 $\mathbb{F}_q$张成空间（或等价地，直到由所抽取列构成的子矩阵的秩为 k）。我们将这一期望记为 $\mathbb{E}[G]$。

一般而言，求出 $\mathbb{E}[G]$ 的闭式表达式颇具挑战。困难源于连续抽取之间复杂的依赖关系：不同于经典的优惠券收集问题 [29–32], ，在这里，新抽取的列所产生的影响取决于所有先前抽取列的$span$ 。具体而言，抽取到此前未抽取过的优惠券并不一定意味着朝着集齐目标取得了进展，因为新列可能不会提高秩。

我们首先证明，与随机访问覆盖深度问题 [18,问题 1], 不同（后者只需恢复一条信息链）， $\mathbb{E}[G]$ 的值仅取决于 G的行空间，也就是说，仅取决于矩阵 G生成的码。

命题 2.1。 设 $G, G' \in \mathbb{F}_q^{k \times n}$具有相同的行空间。则 $\mathbb{E}[G] = \mathbb{E}[G']$。

证明。 由于 G 和 G' 具有相同的行空间，因此存在一个可逆矩阵 $A \in \mathbb{F}_q^{k \times k}$，使得 $G' = AG$。由左乘 A 不会改变列之间线性依赖关系这一事实，可得结论。

□

命题 2.1 表明，对于线性纠错码 $\mathcal{C} \subseteq \mathbb{F}_q^n$，量 $\mathbb{E}[\mathcal{C}]$ 定义良好，即 $\mathbb{E}[G]$，其中 G 是 $\mathcal{C}$ 的任意生成矩阵。因此，我们将交替使用符号 $\mathbb{E}[G]$ 和 $\mathbb{E}[\mathcal{C}]$。现在，我们可以陈述本文将要研究的第二个问题，它与问题 A 密切相关。

问题 B（最优覆盖深度问题）。给定 n、k 和 q，计算值

$$\mathbb{E}_{\mathrm{opt}}[n,k]_q \triangleq \min\{\mathbb{E}[\mathcal{C}] : \mathcal{C} \subseteq \mathbb{F}_q^n \text{ is a code of dimension } k\},$$

并构造一个达到最小值的码 $\mathcal{C}$。

在本文中，$G \in \mathbb{F}_q^{k\times n}$表示一个秩为 k 的矩阵，$\mathcal{C} \subseteq \mathbb{F}_q^n$是由 G 生成的维数为 k 的（线性分组）码，而 $\mathcal{C}^\perp \subseteq \mathbb{F}_q^n$是维数为 $n-k$ 的 $\mathcal{C}$ 的对偶码。

文献 [15] 中得到了若干与问题 A 和问题 B 相关的结果。我们指出其中与本文最相关的一个。

定理 2.2（见 [15,推论 1]）。对于维数为 k 的码 $\mathcal{C} \subseteq \mathbb{F}_q^n$ 的任意生成矩阵 G，有

$$\mathbb{E}[G] \geq \sum_{i=0}^{k-1} \frac{n}{n-i} = n(H_n - H_{n-k}).$$

此外，只有 MDS码的任意生成矩阵才能使该下界取等号。

备注 2.3。定理 2.2 给出了期望的下界，并解决了 MDS 码的问题 A。对于任意满足存在维数为 k 的 MDS 码 $\mathcal{C} \subseteq \mathbb{F}_q^n$ 的参数选择 n, k 和 q，它也解决了问题 B。特别地，假设 MDS 猜想 [33] 成立，我们可以写出

$$\mathbb{E}_{\mathrm{opt}}[n,k]_q = n(H_n - H_{n-k}) \text{ when } q \geq n-1.$$

众所周知，MDS 码只存在于足够大的有限域上。事实上，人们猜想（并已在若干情形下证明），$q \geq n-1$ 是 MDS 码存在的必要条件，只有极少数参数集合例外，它们“只需”满足 $q \geq n-2$；参见 [33–36]。因此，当 q 过小、无法支持 MDS 码存在时，自然有必要研究在此背景下能够得到哪些结果。

3 单纯形码的性能

聚焦于小有限域时，自然会考虑单纯形码。回顾一下，在有限域 $\mathbb{F}_q$ 上的 k 维单纯形码的生成矩阵 G，是将 $\mathbb{F}_q^k$ 的所有非零元素按非零标量倍数等价分类后，作为 G 的列排列而得到的。这个生成矩阵结构简单，因此可以利用标准优惠券收集问题论证的 q 类比，得到 $\mathbb{E}[G]$ 的闭式公式。

定理 3.1。设 $\mathcal{C} \subseteq \mathbb{F}_q^n$ 是维数为 k 的 q 元单纯形码，其中 $n = (q^k - 1)/(q-1)$。我们有

$$\mathbb{E}[\mathcal{C}] = k + \sum_{i=1}^{k} \frac{q^{i-1}-1}{q^k - q^{i-1}}.$$

证明。固定 $\mathcal{C}$ 的任意生成矩阵 G。对于 $i \in \{1,\ldots,k\}$，令 $s_i(\mathcal{C})$ 表示一个随机变量，它决定所选列张成维数为 i 的空间之前所需的抽取次数，

当先前抽取的列张成维数为 $i-1$ 的空间时。注意， $s_1(\mathcal{C})$ 的期望值等于 1，因为 G 的所有列均非零。由于 G 的列是 $\mathbb{F}_q^k$ 中相差非零倍数的元素， $s_i(\mathcal{C})$ 是一个成功概率为

$$\frac{n-\frac{q^{i-1}-1}{q-1}}{n}.$$

因此，根据期望的线性性，我们有

$$\mathbb{E}[\mathcal{C}]=\mathbb{E}\left[\sum_{i=1}^{k}s_i(\mathcal{C})\right]=\sum_{i=1}^{k}\mathbb{E}[s_i(\mathcal{C})]=k+\sum_{i=1}^{k}\frac{q^{i-1}-1}{q^k-q^{i-1}},$$

如所愿。 □

根据实验结果，在参数相同的所有码中，单纯形码的表现最佳。然而，我们仍然没有这一事实的正式证明。

猜想 3.2. 令 $\mathcal{C}\subseteq\mathbb{F}_q^n$ 为维数为 k的 q元单纯形码，其中 $n=(q^k-1)/(q-1)$。则 $\mathcal{C}$ 解决问题 B。

4 一个对偶性结果

一个自然的问题是， $\mathbb{E}[\mathcal{C}]$ 的值与 $\mathbb{E}[\mathcal{C}^\perp]$的值有何关联。尽管一般而言，这两个量并不能相互确定，但在本节中，我们将推导出一个对偶性结果，用对偶码 $\mathcal{C}^\perp$的组合结构来表示 $\mathbb{E}[\mathcal{C}]$ 。在第 5 节中，我们将说明这一结果如何有一项具体应用，即当 $\mathcal{C}$为汉明码时计算 $\mathbb{E}[\mathcal{C}]$ 。

定义 4.1。 设 $\mathcal{C}\subseteq\mathbb{F}_q^n$ 是一个 k 维码，并令 $\pi_S:\mathbb{F}_q^n\to\mathbb{F}_q^{|S|}$ 表示投影到由集合 $S\subseteq\{1,\dots,n\}$ 索引的坐标上。若 $\pi_S(\mathcal{C})$ 的维数为 k，则称 S 是 $\mathcal{C}$ 的**信息集** 。

若 G 是 $\mathcal{C}$ 的生成矩阵，则对于 $0\le s\le n$ ，我们用 g_j 表示 G 的第 j 列，并定义

$$\alpha(\mathcal{C},s)=\left|\left\{S\subseteq\{1,\dots,n\}:|S|=s,\langle g_j:j\in S\rangle=\mathbb{F}_q^k\right\}\right|,$$

该量与 G 的选取无关，并统计 $\mathcal{C}$ 中基数为 s 的信息集的数量。

注意，对于 $0\le s\le k-1$，有 $\alpha(\mathcal{C},s)=0$。按照命题 2.1 中相同的推理，可以验证 $\alpha(\mathcal{C},s)$ 仅取决于 G 所生成的码 $\mathcal{C}$ 。因此，我们将符号 $\alpha(\mathcal{C},s)$ 和 $\alpha(\mathcal{C},s)$ 交替使用。

我们首先用刚才引入的值 $\alpha(\mathcal{C},s)$ 表示 $\mathbb{E}[\mathcal{C}]$ 。证明与 [18, 引理 1] 的证明类似，因此本文省略。

命题 4.2。 对于任意 k 维码 $\mathcal{C}\subseteq\mathbb{F}_q^n$ ，我们有

$$\mathbb{E}[\mathcal{C}]=nH_n-\sum_{s=k}^{n-1}\frac{\alpha(\mathcal{C},s)}{\binom{n-1}{s}}.$$

在继续之前，我们说明如何利用命题 4.2 轻松计算 MDS 码的期望。

例4.3. 设 $\mathcal{C} \subseteq \mathbb{F}_q^n$ 是一个维数为 k 的 MDS 码，并设 G 是 $\mathcal{C}$ 的生成矩阵。由于 G 是一个 MDS矩阵， G 的任意 k 列都线性无关。因此，我们有

$$\alpha(\mathcal{C}, s) = \binom{n}{s} \quad \text{for } k \le s \le n.$$

将这些值代入命题 4.2 的公式，我们得到

$$\mathbb{E}[\mathcal{C}] = nH_n - \sum_{s=k}^{n-1} \frac{\binom{n}{s}}{\binom{n-1}{s}},$$

经过简单计算，可化简为 $n(H_n - H_{n-k})$。

现在我们转向对偶性结果。我们将 $\alpha(\mathcal{C}, s)$ 的值与对偶码 $\mathcal{C}^\perp$ 的结构联系起来。为此，引入一些辅助量会很方便。我们将向量 $x \in \mathbb{F}_q^n$ 的汉明支撑记为 $\sigma(x) = \{i : x_i \neq 0\}$。对于码 $\mathcal{C} \subseteq \mathbb{F}_q^n$ 和子集 $S \subseteq \{1, \dots, n\}$，我们令 $\mathcal{C}(S) = \{x \in \mathcal{C} : \sigma(x) \subseteq S\}$。集合 S 的补集记为 $S^{\mathsf{c}} = \{1, \dots, n\} \setminus S$。

记号4.4。对于码 $\mathcal{C} \subseteq \mathbb{F}_q^n$、 $1 \le \ell \le k$ 和 $0 \le s \le n$，令

$$\beta_\ell(\mathcal{C}, s) = |\{S \subseteq \{1, \dots, n\} : |S| = s, \dim(\mathcal{C}(S^{\mathsf{c}})) = \ell\}|.$$

本节的主要工具是以下结果。

引理 4.5. 设 $\mathcal{C} \subseteq \mathbb{F}_q^n$ 是一个维数为 k 的码。对于所有满足 $1 \le \ell \le k$ 和 $0 \le s \le n$ 的 ℓ, s，我们有

$$\beta_\ell(\mathcal{C}, s) = \beta_{\ell+s-k}(\mathcal{C}^\perp, n-s). \tag{1}$$

特别地，

$$\alpha(\mathcal{C}, s) = \beta_{s-k}(\mathcal{C}^\perp, n-s).$$

证明. 取一个任意的、基数为 s 的集合 $S \subseteq \{1, \dots, n\}$。利用秩-零化度定理，我们得到

$$\dim(\pi_S(\mathcal{C})) + \dim(\ker(\pi_S)) = k, \tag{2}$$

我们可以将其改写为

$$\dim(\pi_S(\mathcal{C})) + \dim(\mathcal{C}(S^{\mathsf{c}})) = k. \tag{3}$$

此外，由 [37, 定理 24]，我们有

$$|\mathcal{C}(S)| = \frac{|\mathcal{C}|}{q^{n-s}} |\mathcal{C}^\perp(S^{\mathsf{c}})|,$$

也就是说，

$$\dim(\mathcal{C}(S)) = k - n + s + \dim(\mathcal{C}^\perp(S^{\mathsf{c}})). \tag{4}$$

因此，由 (2) 可知，$\dim(\pi_S(\mathcal{C})) = t$ 当且仅当 $\dim(\mathcal{C}(S^{\mathsf{c}})) = k - t$；而 (4) 告诉我们，后一个等式等价于 $\dim(\mathcal{C}^\perp(S)) = s - t$。这一切表明存在一个双射

$$\{S \subseteq \{1, \dots, n\} : |S| = s, \dim(\mathcal{C}(S^{\mathsf{c}})) = k - t\} \to \{S \subseteq \{1, \dots, n\} : |S| = n - s, \dim(\mathcal{C}^\perp(S^{\mathsf{c}})) = s - t\},$$

由此我们得到引理的第一部分。对于第二部分，只需利用以下事实：

$$\alpha(\mathcal{C}, s) = \beta_0(\mathcal{C}, s), \tag{5}$$

这显然可由定义推出。将此等式与 (1) 结合，我们便得到该引理的第二部分。

□

下一个推论可直接由命题 4.2 和引理 4.5 得出。我们将在下一节中使用它。

推论 4.6. 设 $\mathcal{C} \subseteq \mathbb{F}_q^n$ 是一个维数为 k的码。我们有

$$\mathbb{E}[\mathcal{C}] = nH_n - \sum_{s=k}^{n-1} \frac{\beta_{s-k}(\mathcal{C}^\perp, n-s)}{\binom{n-1}{s}}. \tag{6}$$

5 汉明码与三元戈莱码的性能

我们应用第 4 节的结果，计算 $\mathbb{E}[\mathcal{C}]$ 的值，其中 $\mathcal{C}$ 分别为汉明码、三元戈莱码和扩展三元戈莱码。

下面的结果是引理 4.5 的直接应用，它使我们能够用汉明码对应的对偶码（即单纯形码）来表示汉明码的期望值。

定理 5.1. 设 $\mathcal{C} \subseteq \mathbb{F}_q^n$ 为冗余度为 r 的 q 元汉明码，其中 $n =(q^r-1)/(q-1)$。我们有

$$\mathbb{E}[\mathcal{C}] = nH_n - \sum_{\ell=1}^{r} \frac{1}{\binom{n-1}{n-\ell}} \frac{\prod_{i=0}^{\ell-1} \frac{q^r-q^i}{q-1}}{\ell!}.$$

证明. 我们有 $\dim(\mathcal{C}) = n-r$，因此可以将 (6) 改写为

$$\mathbb{E}[\mathcal{C}] = nH_n - \sum_{\ell=1}^{r} \frac{\beta_{r-\ell}(\mathcal{C}^\perp, \ell)}{\binom{n-1}{n-\ell}},$$

其中，对偶码 $\mathcal{C}^\perp$ 是 $[n,r]_q$ 单纯形码。将 (3) 应用于记号4.4，可得

$$\beta_{r-\ell}(\mathcal{C}^\perp, \ell) = |\{S \subseteq \{1,\dots,n\} : |S| = \ell, \dim(\pi_S(\mathcal{C}^\perp)) = \ell\}|.$$

剩下的工作是计算基数为 ℓ 且对应列线性无关的子集数量。为此，我们再次利用以下事实：单纯形码的任意生成矩阵的列，都是 $\mathbb{F}_q^r$ 中所有非零向量在非零标量倍数意义下的代表。因此，我们有

$$\beta_{r-\ell}(\mathcal{C}^\perp, \ell) = \frac{\prod_{i=0}^{\ell-1} \left(\frac{q^r-1}{q-1} - \frac{q^i-1}{q-1} \right)}{\ell!} = \frac{\prod_{i=0}^{\ell-1} \frac{q^r-q^i}{q-1}}{\ell!},$$

由此即可得出该结论。 □

现在我们转而讨论三元戈莱码。首先，我们对命题 4.2 中的公式进行改进，该公式涉及码的最小距离，因此需要计算的未知值更少。

Corollary 5.2. 设 $\mathcal{C} \subseteq \mathbb{F}_q^n$ 是维数为 k 、最小距离为 d的码。我们有

$$\mathbb{E}[\mathcal{C}] = n(H_n - H_{d-1}) - \sum_{s=k}^{n-d} \frac{\alpha(\mathcal{C}, s)}{\binom{n-1}{s}}.$$

证明. 由(5)可知 $\alpha(\mathcal{C},s)=|\{S\subseteq[n]:\ |S|=s,\dim(\mathcal{C}(S^c))=0\}|$。这意味着，若 $n-s\le d-1$，则有$\dim(\mathcal{C}(S^c))=0$，从而 $\alpha(\mathcal{C},s)=\binom{n}{s}$。因此，

$$\begin{aligned}\mathbb{E}[\mathcal{C}] &= nH_n-\sum_{s=k}^{n-1}\frac{\alpha(\mathcal{C},s)}{\binom{n-1}{s}}\\ &= nH_n-\left(\sum_{s=k}^{n-d}\frac{\alpha(\mathcal{C},s)}{\binom{n-1}{s}}+\sum_{s=n-d+1}^{n-1}\frac{n}{n-s}\right)\\ &= n(H_n-H_{d-1})-\sum_{s=k}^{n-d}\frac{\alpha(\mathcal{C},s)}{\binom{n-1}{s}}.\end{aligned}$$

□

当码 $\mathcal{C}$ 为三元戈莱码或扩展三元戈莱码时，推论 5.2 给出一个公式，只需计算 $\alpha(\mathcal{C},s)$ 诸值中的一个，从而使问题 A 对这两个码更容易求解。我们将利用码的重量枚举式；它出现在本节，并在第 6 节中更为突出。

定义 5.3. 设 $\mathcal{C}\subseteq\mathbb{F}_q^n$ 为一个码。则 **齐次重量枚举式**（关于 $\mathcal{C}$ ）是多项式

$$W_{\mathcal{C}}(X,Y)=\sum_{i=0}^{n}W_i(\mathcal{C})X^{n-i}Y^i,\ \text{where}\ W_i(\mathcal{C})=|\{c\in\mathcal{C}:\omega^H(c)=i\}|.$$

该多项式也可以表示为一元形式 $W_{\mathcal{C}}(Z)$，称为**重量枚举式**，它与 $W_{\mathcal{C}}(X,Y)$ 的联系如下：

$$W_{\mathcal{C}}(X,Y)=X^nW_{\mathcal{C}}(X^{-1}Y),\quad W_{\mathcal{C}}(Z)=W_{\mathcal{C}}(1,Z).$$

数 $\{W_i(\mathcal{C}):\ 0\le i\le n\}$ 称为 $\mathcal{C}$ 的**重量分布** 。

定理 5.4. 设 $\mathcal{C}$ 为长度 $n=11$、维数 $k=6$、最小距离 $d=5$ 的三元戈莱码。我们有

$$\mathbb{E}[\mathcal{C}]=n(H_n-H_{d-1})-\frac{\binom{n}{k}-\frac{W_k(\mathcal{C}^\perp)}{2}}{\binom{n-1}{k}}\approx 8.416.$$

证明。 根据推论 5.2，当 $s\ge n-d+1=7$ 时，我们有 $\alpha(\mathcal{C},\ s)=\binom{n}{s}$。因此，我们只需计算 $\alpha(\mathcal{C},\ k)$。注意，

$$\begin{aligned}\alpha(\mathcal{C},k) &= |\{S\subseteq\{1,\dots,n\}:\ S\text{ is a minimal information set of }\mathcal{C}\}|\\ &= |\{S\subseteq\{1,\dots,n\}:\ |S|=k\,,\nexists\, x\in\mathcal{C}^\perp\setminus\{0\}\ \text{with}\ \sigma(x)\subseteq S\}|\\ &= \binom{n}{k}-|\{S\subseteq\{1,\dots,n\}:\ |S|=k\,,\exists\, x\in\mathcal{C}^\perp\setminus\{0\}\ \text{with}\ \sigma(x)\subseteq S\}|.\end{aligned}$$

三元戈莱码的齐次重量枚举式为

$$W_{\mathcal{C}}(X,Y)=X^{11}+132X^6Y^5+132X^5Y^6+330X^3Y^8+110X^2Y^9+24Y^{11};\qquad(7)$$

参见 [38, 例 7.6.2]。现在我们将（7）与麦克威廉恒等式结合起来：

$$W_{\mathcal{C}^\perp}(X,Y)=\frac{1}{|\mathcal{C}|}W_{\mathcal{C}}(X+(q-1)Y,X-Y);$$

例如参见 [34, 第 5 章，定理 13]。我们得到，对偶码 $\mathcal{C}^\perp$ 的齐次重量枚举式为

$$W_{\mathcal{C}^\perp}(X,Y) = X^{11} + 132X^5Y^6 + 110X^2Y^9.$$

因此，

$$\begin{aligned}\alpha(\mathcal{C},k) &= \binom{n}{k} - |\{S \subseteq \{1,\dots,n\} : |S| = k\,, \exists\, x \in \mathcal{C}^\perp \setminus \{0\} \text{ with } \sigma(x) \subseteq S\}| \\ &= \binom{n}{k} - \frac{W_k(\mathcal{C}^\perp)}{2},\end{aligned}$$

其中，后一个等式成立，是因为如果非零码字 $x \in \mathcal{C}^\perp$ 的支撑包含于一个 k 子集 $S \subseteq \{1,\dots,n\}$ 中，那么 $2x$ 是 $\mathcal{C}^\perp$ 中唯一另一个支撑包含于 S 的非零码字。这是由对偶码 $\mathcal{C}^\perp$ 的最小距离为 $d(\mathcal{C}^\perp) = 6$ 这一事实得出的。

□

对扩展三元戈莱码的类似结果也可以通过类似推理得到，因此省略证明。

定理 5.5. 设 $\mathcal{C}$ 是长度为 $n = 12$、维数为 $k = 6$、最小距离为 $d = 6$ 的扩展三元戈莱码。那么

$$\mathbb{E}[\mathcal{C}] = n(H_n - H_{d-1}) - \frac{\binom{n}{k} - \frac{W_k(\mathcal{C})}{2}}{\binom{n-1}{k}} \approx 8.124.$$

6 期望与重量分布

在上一节中，我们利用对其重量枚举式的了解，计算了三元戈莱码和扩展三元戈莱码的 $\mathbb{E}[\mathcal{C}]$ 值。自然会问，这是否是普遍现象，以及期望 $\mathbb{E}[\mathcal{C}]$ 是否总能用 $\mathcal{C}$ （或等价地用 $\mathcal{C}^\perp$）的重量枚举式表示。下面的例子给出了否定的答案。

例 6.1. 设 $\mathcal{C}_1$ 和 $\mathcal{C}_2$ 分别是由以下矩阵生成的长度为 $n = 12$ 、维数为 $k = 3$的二元码

$$G_1 = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

and

$$G_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix},$$

。可以验证， $\mathcal{C}_1$ 和 $\mathcal{C}_2$ 具有相同的重量枚举式，即

$$W_{\mathcal{C}_1}(Z) = 1 + Z^3 + Z^4 + Z^5 + Z^7 + Z^8 + Z^9 + Z^{12} = W_{\mathcal{C}_2}(Z).$$

另一方面，我们有

$$\mathbb{E}[\mathcal{C}_1] = \frac{1229}{210} \approx 5.852 \quad \text{and} \quad \mathbb{E}[\mathcal{C}_2] = \frac{2633}{462} \approx 5.699.$$

注意到 $\mathcal{C}_1$ 和 $\mathcal{C}_2$ 是不等价码，否则必然有$\mathbb{E}[\mathcal{C}_1] = \mathbb{E}[\mathcal{C}_2]$。

此时，人们可能会想知道， $\mathbb{E}[\mathcal{C}]$ 的值能否用码 $\mathcal{C}$的某种组合不变量来表示。虽然我们已提供证据表明，重量分布并非这样的不变量，但本节将证明，更精细的不变量足以做到这一点，即 $\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^m}$ 的扩张码对 $1\le m\le n$的重量分布。

定义 6.2. 设 $\mathcal{C}\subseteq\mathbb{F}_q^n$ 是一个码， $m\ge 1$ 是一个整数。 $\mathcal{C}$ 在 $\mathbb{F}_{q^m}$ 上的**扩张码**记作 $\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^m}$，是由 $\mathcal{C}$ 的码字构成的 $\mathbb{F}_{q^m}$-线性组合的集合。

本节余下部分将用于证明以下主要结果，该结果将 $\mathbb{E}[\mathcal{C}]$ 与扩张码的重量分布联系起来。

The定理 6.3. 设 $\mathcal{C}\subseteq\mathbb{F}_q^n$ 是一个维数为 k 且最小距离为 d 的码。我们有

$$\mathbb{E}[\mathcal{C}]=n(H_n-H_{d-1})-\sum_{r=k}^{n-d}\frac{1}{\binom{n-1}{r}}\sum_{\ell=0}^{n-r}\binom{n-\ell}{r}\sum_{m=0}^{n}(-1)^m W_\ell(\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^m})\gamma(q,m,n),$$

其中

$$\gamma(q,m,n)=\sum_{j=m}^{n}\left(\prod_{\nu=0}^{j-1}\frac{1}{q^j-q^\nu}\right)q^{\binom{j}{2}+\binom{j-m}{2}}\begin{bmatrix}j\\m\end{bmatrix}_q,$$

并约定当 $m=0$ 时， $\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^m}$ 是零码。

我们认为前述结果本身就很有趣。不过，在第 7 节中，我们还将说明如何用它计算 $\mathbb{E}[\mathcal{C}]$，其中 $\mathcal{C}$ 是任意一个一阶 q 元 Reed-Muller 码。

定理 6.3 的证明依赖于四个我们分别建立的引理。在本节余下部分中，对于任意 $m\ge 1$，我们固定 $\mathbb{F}_{q^m}$ 在 $\mathbb{F}_q$ 上的一组基 $\Omega_m=\{\omega_m^1,\dots,\omega_m^m\}$，并记 $\mathbb{F}_{q^m}\to\mathbb{F}_q^m$: φ_m 为关于基 Ω_m 的展开映射，它是一个 $\mathbb{F}_q$-同构。

对于一个 $\mathbb{F}_q$-子空间 $V\subseteq\mathbb{F}_q^n$，我们令 $\mathbb{F}_q^{m\times n}[V]$ 表示由行向量来自 V 的 $m\times n$ 矩阵构成的 $\mathbb{F}_q$-空间。当 $S\subseteq\{1,\dots,n\}$ 时，我们令 $\mathbb{F}_q^{m\times n}[S]$ 表示这样一个矩阵集合：其任意 $i\notin S$ 所索引的列均恒为零。

映射 φ_m 诱导出如下映射 $\widehat{\varphi}_m$： $\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^m}\to\mathbb{F}_q^{m\times n}[\mathcal{C}]$。对于 $x\in\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^m}$，将 $x=($ x_1、……、 $x_n)$ 写成标准坐标，并构造矩阵 $M_m(x)$，其列为向量 $\varphi_m(x_1)^\top$、……、 $\varphi_m(x_n)^\top$，其中上标“$\top$”表示转置。容易验证， $M_m(x)\in\mathbb{F}_q^{m\times n}[\mathcal{C}]$。我们定义 $\widehat{\varphi}_m$： $x\mapsto M_m(x)$。

引理 6.4 （参见 $[39,$ 命题 $6]$）。 设 $x\in\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^m}$ 和 $S\subseteq\{1,\dots,n\}$。以下条件等价：

1. $\sigma(x)\subseteq S$，2. $\widehat{\varphi}_m$ $(x)\in\mathbb{F}_q^{m\times n}[S]$。

下一个结果回顾一个著名的组合恒等式。通过按秩对元素属于 $\mathbb{F}_q$ 的 $j\times r$ 矩阵进行计数即可得到该恒等式。

引理 6.5 （参见 $[40,$ 引理 $26]$）。 对于所有 j， $r\in\mathbb{N}$ 我们有

$$q^{jr}=\sum_{i=0}^{r}\begin{bmatrix}j\\i\end{bmatrix}_q\begin{bmatrix}r\\i\end{bmatrix}_q\prod_{s=0}^{i-1}(q^i-q^s).$$

最后两个引理建立了我们证明主要结果所需的特设反演公式。

引理 6.6. 设 $\{x_i\}_{i=0}^{\infty}$ 是一个整数序列。设

$$y_m = \sum_{i=0}^{m} \begin{bmatrix} m \\ i \end{bmatrix}_q x_i \quad \text{for all } m \geq 0.$$

Then

$$x_i = \sum_{m=0}^{i} (-1)^{i-m} q^{\binom{i-m}{2}} \begin{bmatrix} i \\ m \end{bmatrix}_q y_m \quad \text{for all } i \geq 0.$$

证明。 固定 $i \geq 0$。我们有

$$\begin{aligned}
\sum_{m=0}^{i} (-1)^{i-m} q^{\binom{i-m}{2}} \begin{bmatrix} i \\ m \end{bmatrix}_q y_m &= \sum_{m=0}^{i} (-1)^{i-m} q^{\binom{i-m}{2}} \begin{bmatrix} i \\ m \end{bmatrix}_q \sum_{j=0}^{m} \begin{bmatrix} m \\ j \end{bmatrix}_q x_j \\
&= \sum_{j=0}^{i} x_j \sum_{m=j}^{i} (-1)^{i-m} q^{\binom{i-m}{2}} \begin{bmatrix} i \\ m \end{bmatrix}_q \begin{bmatrix} m \\ j \end{bmatrix}_q \\
&= \sum_{j=0}^{i} x_j \sum_{m=j}^{i} (-1)^{i-m} q^{\binom{i-m}{2}} \begin{bmatrix} i \\ j \end{bmatrix}_q \begin{bmatrix} i-j \\ i-m \end{bmatrix}_q \\
&= \sum_{j=0}^{i} \begin{bmatrix} i \\ j \end{bmatrix}_q x_j \sum_{m=j}^{i} (-1)^{i-m} q^{\binom{i-m}{2}} \begin{bmatrix} i-j \\ i-m \end{bmatrix}_q \\
&= \sum_{j=0}^{i} \begin{bmatrix} i \\ j \end{bmatrix}_q x_j \sum_{m=0}^{i-j} (-1)^{m} q^{\binom{m}{2}} \begin{bmatrix} i-j \\ m \end{bmatrix}_q = x_i.
\end{aligned}$$

后一个等式成立，因为

$$\sum_{m=0}^{i-j} (-1)^m q^{\binom{m}{2}} \begin{bmatrix} i-j \\ m \end{bmatrix}_q = 0$$

除非 $i = j$ （此时其值为 1），根据 q-二项式定理。 □

下一个反演公式的证明与前一个类似，因此略去。

引理 6.7。 设 $n \geq 1$，且 $x_0, x_1, \ldots, x_n$ 为整数。设

$$y_i = \sum_{j=i}^{n} \begin{bmatrix} j \\ i \end{bmatrix}_q x_j \quad \text{for } 0 \leq i \leq n.$$

Then

$$x_j = \sum_{i=j}^{n} (-1)^{i-j} q^{\binom{i-j}{2}} \begin{bmatrix} i \\ j \end{bmatrix}_q y_i \quad \text{for all } 0 \leq j \leq n.$$

定理 *6.3* 的证明。 在整个证明中，对于 $0 \leq j \leq k$ 和 $0 \leq r \leq n$，我们令

$$\widehat{\beta}_j(\mathcal{C}, r) = |\{S \subseteq \{1, \ldots, n\} : |S| = r, \dim(\mathcal{C}(S)) = j\}|.$$

我们对集合的基数进行双重计数

$$\mathcal{A}_m = \{(M, S) : M \in \mathbb{F}_q^{m \times n}[\mathcal{C}],\ S \subseteq \{1, \ldots, n\},\ |S| = r,\ \sigma(\mathrm{rowsp}(M)) \subseteq S\}$$

对于 $m \geq 0$，其中 $\mathbb{F}_q^{0\times n} = \{\emptyset\}$ and rowsp$(\emptyset) = \{0\}$ 按照惯例。另一方面，

$$\begin{aligned}|\mathcal{A}_m| &= \sum_{\substack{S\subseteq\{1,\ldots,n\}\\|S|=r}} |\{M \in \mathbb{F}_q^{m\times n}[\mathcal{C}] \,:\, \sigma(\mathrm{rowsp}(M)) \subseteq S\}| \\ &= \sum_{\substack{S\subseteq\{1,\ldots,n\}\\|S|=r}} |\mathcal{C}(S)|^m = \sum_{j=0}^{n} \sum_{\substack{S\subseteq\{1,\ldots,n\}\\|S|=r\\ \dim(\mathcal{C}(S))=j}} q^{jm} = \sum_{j=0}^{n} q^{jm}\widehat{\beta}_j(\mathcal{C},r).\end{aligned}$$

另一方面，

$$\begin{aligned}|\mathcal{A}_m| &= \sum_{M\in\mathbb{F}_q^{m\times n}[\mathcal{C}]} |\{S \subseteq \{1,\ldots,n\} \,:\, |S| = r,\, \sigma(\mathrm{rowsp}(M)) \subseteq S\}| \\ &= \sum_{x\in\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^m}} |\{S \subseteq \{1,\ldots,n\} \,:\, |S| = r,\, \sigma(x) \subseteq S\}| \\ &= \sum_{\ell=0}^{n} \sum_{\substack{x\in\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^m}\\ \omega^H(x)=\ell}} \binom{n-\ell}{r-\ell} = \sum_{\ell=0}^{n} \binom{n-\ell}{r-\ell} W_\ell(\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^m}),\end{aligned}$$

其中倒数第二个等号由引理 6.4 得出。因此

$$\sum_{j=0}^{n} q^{jm}\widehat{\beta}_j(\mathcal{C},r) = \sum_{\ell=0}^{n} \binom{n-\ell}{r-\ell} W_\ell(\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^m}) \tag{8}$$

对所有 $m \geq 0$ 均成立。将 (8) 与引理 6.5 结合可得

$$\sum_{i=0}^{m} \begin{bmatrix} m \\ i \end{bmatrix}_q \left(\prod_{s=0}^{i-1}(q^i - q^s)\right) \sum_{j=0}^{n} \begin{bmatrix} j \\ i \end{bmatrix}_q \widehat{\beta}_j(\mathcal{C},r) = \sum_{\ell=0}^{n} \binom{n-\ell}{r-\ell} W_\ell(\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^m}).$$

注意，前一个恒等式具有如下形式

$$\sum_{i=0}^{m} \begin{bmatrix} m \\ i \end{bmatrix}_q x_i = y_m,$$

这使我们能够应用引理 6.6，得到

$$\sum_{j=m}^{n} \begin{bmatrix} j \\ m \end{bmatrix}_q \widehat{\beta}_j(\mathcal{C},r) = \left(\prod_{s=0}^{m-1} \frac{1}{q^m - q^s}\right) \sum_{i=0}^{m} (-1)^{m-i} q^{\binom{m-i}{2}} \begin{bmatrix} m \\ i \end{bmatrix}_q \sum_{\ell=0}^{n} \binom{n-\ell}{r-\ell} W_\ell(\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^i})$$

对所有 $m \geq 0$ 均成立。因此，我们可以应用引理 6.7，得到

$$\widehat{\beta}_m(\mathcal{C},r) = \sum_{j=m}^{n} (-1)^{j-m} q^{\binom{j-m}{2}} \begin{bmatrix} j \\ m \end{bmatrix}_q \left(\prod_{s=0}^{j-1} \frac{1}{q^j - q^s}\right) \sum_{i=0}^{j} (-1)^{j-i} q^{\binom{j-i}{2}} \begin{bmatrix} j \\ i \end{bmatrix}_q \sum_{\ell=0}^{n} \binom{n-\ell}{r-\ell} W_\ell(\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^i})$$

对所有 $m \geq 0$。

综上，由于 $\beta_0(\mathcal{C},\ r) = \alpha(\mathcal{C},\ n-r)$，我们得到

$$\begin{aligned}\alpha(\mathcal{C},n-r) &= \sum_{j=0}^{n} (-1)^j q^{\binom{j}{2}} \left(\prod_{s=0}^{j-1} \frac{1}{q^j - q^s}\right) \sum_{i=0}^{j} (-1)^{j-i} q^{\binom{j-i}{2}} \begin{bmatrix} j \\ i \end{bmatrix}_q \sum_{\ell=0}^{n} \binom{n-\ell}{r-\ell} W_\ell(\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^i}) \\ &= \sum_{\ell=0}^{n} \binom{n-\ell}{r-\ell} \sum_{i=0}^{n} (-1)^i W_\ell(\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^i}) \sum_{j=i}^{n} \left(\prod_{s=0}^{j-1} \frac{1}{q^j - q^s}\right) q^{\binom{j}{2}+\binom{j-i}{2}} \begin{bmatrix} j \\ i \end{bmatrix}_q,\end{aligned}$$

或者，等价地，

$$\alpha(\mathcal{C},r)=\sum_{\ell=0}^{n-r}\binom{n-\ell}{r}\sum_{i=0}^{n}(-1)^i W_\ell(\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^i})\sum_{j=i}^{n}\left(\prod_{s=0}^{j-1}\frac{1}{q^j-q^s}\right)q^{\binom{j}{2}+\binom{j-i}{2}}\begin{bmatrix}j\\ i\end{bmatrix}_q.$$

Co将后者与推论 5.2 中的一般公式结合起来，我们得到所需的结果 lt. □

注意，定理 6.3 表明，对于任意码 $\mathcal{C}\subseteq\mathbb{F}_q^n$，已知扩域码 $\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^i}$（$0\le i\le n$）的所有重量分布，就足以确定期望 $\mathbb{E}[\mathcal{C}]$。

7 一阶 Reed-Muller 码的性能

本节展示定理 6.3 的一个直接应用，为 $\mathbb{E}[\mathcal{C}]$ 建立闭式公式，其中 $\mathcal{C}$ 是任意一阶 q 元 Reed-Muller 码。回顾一下，对于整数 $s\ge 2$，一阶 q 元 Reed-Muller 码 $\mathcal{C}\subseteq\mathbb{F}_q^n$ 的长度为 $n=q^{s-1}$，维数为 $k=s$，最小距离为 $d=(q-1)q^{s-2}$；参见 [41, 定义 15]。我们首先介绍扩展重量枚举式，这是证明主要结果所需的工具。

定义 7.1。设 $\mathcal{C}\subseteq\mathbb{F}_q^n$是一个码，且 $J\subseteq\{1,\dots,n\}$。设

$$B_J(U)=U^{\dim(\mathcal{C}(J^{\mathrm{c}}))}-1,\qquad B_t(U)=\sum_{\substack{J\subseteq\{1,\dots,n\}\\ |J|=t}}B_J(U).$$

其扩展重量枚举式为 $\mathcal{C}$ 是

$$W_{\mathcal{C}}(X,Y,U)=X^n+\sum_{t=0}^{n}B_t(U)(X-Y)^tY^{n-t}.$$

事实证明，First 阶 q 元 Reed-Muller 码的扩展重量枚举式是已知的，并且已按如下形式明确计算出来。

引理 7.2（见 [41, 定理 18]）。设 $\mathcal{C}\subseteq\mathbb{F}_q^n$为维数为 s的 First 阶 q 元 Reed-Muller 码，其中 $n=q^{s-1}$。其扩展重量枚举式等于

$$\begin{aligned}W_{\mathcal{C}}(X,Y,U)=&\sum_{t=1}^{s}\left(\prod_{j=0}^{t-1}(U-q^j)\right)\begin{bmatrix}s-1\\ t-1\end{bmatrix}_q Y^{q^{s-1}}\\&+\sum_{t=0}^{s-1}\left(\prod_{j=0}^{t-1}(U-q^j)\right)q^t\begin{bmatrix}s-1\\ t\end{bmatrix}_q X^{q^{s-1}-1-t}Y^{q^{s-1}-q^{s-1-t}}.\end{aligned}\tag{9}$$

现在，我们已经具备了求解 First 阶 q 元 Reed-Muller 码的 DNA 覆盖深度问题所需的全部要素。

定理 7.3. 设 $\mathcal{C}\subseteq\mathbb{F}_q^n$为维数为 s 的一阶 q 元 Reed-Muller 码，其中 $n=q^{s-1}$ 且 $d=(q-1)q^{s-2}$。我们有

$$\begin{aligned}\mathbb{E}[\mathcal{C}]=n(H_n-H_{d-1})-\sum_{r=s}^{q^{s-2}}\frac{1}{\binom{q^{s-1}-1}{r}}\sum_{t=0}^{\lfloor s-1-\log_q(r)\rfloor}\binom{q^{s-1-t}}{r}\\ \cdot\sum_{i=0}^{q^{s-1}}(-1)^i\left(\prod_{j=0}^{t-1}(q^i-q^j)\right)q^t\begin{bmatrix}s-1\\ t\end{bmatrix}_q\gamma(q,i,q^{s-1}),\end{aligned}$$

其中

$$\gamma(q,i,q^{s-1})=\sum_{j=i}^{q^{s-1}}\left(\prod_{\nu=0}^{j-1}\frac{1}{q^j-q^\nu}\right)q^{\binom{j}{2}+\binom{j-i}{2}}\begin{bmatrix}j\\ i\end{bmatrix}_q.$$

证明. 我们首先将扩展重量枚举式与扩域码联系起来。更确切地说，码 $\mathcal{C}$ 在 $U=q^m$ 上的扩展重量枚举式与扩张码 $\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^m}$的重量枚举式一致:

$$W_{\mathcal{C}}(X,Y,q^m)=W_{\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^m}}(X,Y)\quad\text{for all } m\geq 0; \tag{10}$$

参见 [39, 命题 5]。因此，将(10)与(9)结合，我们得到扩张码 $\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^m}$ 的重量枚举式，该码对应于一阶 q元Reed-Muller码 $\mathcal{C}$ ，其维数为 s:

$$\begin{aligned}W_{\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^m}}(X,Y)=&\sum_{t=1}^{s}\left(\prod_{j=0}^{t-1}(q^m-q^j)\right)\begin{bmatrix}s-1\\ t-1\end{bmatrix}_q Y^{q^{s-1}}\\&+\sum_{t=0}^{s-1}\left(\prod_{j=0}^{t-1}(q^m-q^j)\right)q^t\begin{bmatrix}s-1\\ t\end{bmatrix}_q X^{q^{s-1-t}}Y^{q^{s-1}-q^{s-1-t}}\end{aligned}$$

对所有 $m\geq 0$成立。就本证明而言，我们将需要系数

$$W_{q^{s-1}-q^{s-1-t}}(\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^m})=\left(\prod_{j=0}^{t-1}(q^m-q^j)\right)q^t\begin{bmatrix}s-1\\ t\end{bmatrix}_q,\quad\text{for } 0\leq t\leq s-1. \tag{11}$$

由定理 6.3可知

$$\sum_{r=k}^{n-d}\frac{1}{\binom{n-1}{r}}\sum_{\ell=0}^{n-r}\binom{n-\ell}{r}\sum_{i=0}^{n}(-1)^i W_\ell(\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^i})\gamma(q,i,n)=n(H_n-H_{d-1})-\mathbb{E}[\mathcal{C}]. \tag{12}$$

由(11)，(12)的左端为

$$\begin{aligned}&\sum_{r=k}^{n-d}\frac{1}{\binom{n-1}{r}}\sum_{\ell=0}^{n-r}\binom{n-\ell}{r}\sum_{i=0}^{n}(-1)^i W_\ell(\mathcal{C}\otimes_{\mathbb{F}_q}\mathbb{F}_{q^i})\gamma(q,i,n)\\&\quad=\sum_{r=s}^{q^{s-2}}\frac{1}{\binom{q^{s-1}-1}{r}}\sum_{\substack{0\leq\ell\leq q^{s-1}-r,\\ \ell=q^{s-1}-q^{s-1-t},\\ 0\leq t\leq s-1}}\binom{q^{s-1}-\ell}{r}\sum_{i=0}^{q^{s-1}}(-1)^i\left(\prod_{j=0}^{t-1}(q^i-q^j)\right)q^t\begin{bmatrix}s-1\\ t\end{bmatrix}_q\gamma(q,i,q^{s-1}),\end{aligned}$$

它可以改写为

$$\sum_{r=s}^{q^{s-2}}\frac{1}{\binom{q^{s-1}-1}{r}}\sum_{t=0}^{\lfloor s-1-\log_q(r)\rfloor}\binom{q^{s-1-t}}{r}\sum_{i=0}^{q^{s-1}}(-1)^i\left(\prod_{j=0}^{t-1}(q^i-q^j)\right)q^t\begin{bmatrix}s-1\\ t\end{bmatrix}_q\gamma(q,i,q^{s-1})$$

因为对于任意 $0\leq t\leq s-1$，条件 $0\leq q^{s-1}-q^{s-1-t}\leq q^{s-1}-r$ 等价于$0\leq t\leq\lfloor s-1-\log_q(r)\rfloor$。结合 (12) 即可得到所需结果。 □

8 结论与未来工作

本文从编码理论的角度研究了 DNA 数据存储中的覆盖深度问题。我们运用从组合学到对偶理论的各种方法，并探索重量分布和扩域码之间的联系，推导出了多种码族的期望闭式表达式。除了本文所呈现的结果之外，还有若干值得探索的研究方向。以下是其中两个:

1. 我们猜想，对于单纯形码存在的参数，单纯形码能够解决问题 B。一个有趣的挑战是，刻画在既不存在 MDS 码也不存在单纯形码的情况下性能最佳的码。2. 尽管我们给出了多种期望的闭式表达式，但一般而言，推导这类公式并不容易。因此，建立通用下界或开发近似技术将十分有价值。

参考文献

[1] J. Rydning, 《Worldwide IDC Global DataSphere Forecast, 2022–2026: Enterprise Or- ganizations Driving Most of the Data Growth》, 国际数据公司（*IDC*）, 2022 年。[2] DNA 数据存储联盟, 《保存我们的数字遗产：DNA 数据存储简介》, 2021 年。[3] D. Markowitz, 《生物学就是你所需要的一切》, 高性能计算、网络、存储与分析国际会议，2023 年。[4] E. M. LeProust、B. J. Peck、K. Spirin、H. B. McCuen、B. Moore、E. Namsaraev 和 M. H. Caruthers, 《采用新型脱嘌呤控制工艺合成高质量长链（150 聚体）寡核苷酸文库》, 《核酸研究》, 第 38 卷，第 2522–2540 页，2010 年。[5] S. Tabatabaei Yazdi、Y. Yuan、J. Ma、H. Zhao 和 O. Milenkovic, 《一种可重写、随机访问的 DNA 存储系统》, 《科学报告》, 第 5 卷，第 1 期，第 1–10 页，2015 年。[6] M. Blawat、K. Gaedke、I. Huetter、X.-M. Chen、B. Turczyk、S. Inverso、B. W. Pruitt 和 G. M. Church, 《DNA 数据存储的前向纠错》, 《计算机科学程序汇刊》, 第 80 卷，第 1011–1022 页，2016 年。[7] J. Bornholt、R. Lopez、D. M. Carmean、L. Ceze、G. Seelig 和 K. Strauss, 《一种基于 DNA 的归档存储系统》, 载于第二十一届编程语言与操作系统架构支持国际会议论文集，第 637–649 页，2016 年。[8] S. H. T. Yazdi、R. Gabrys 和 O. Milenkovic, 《便携式无差错 DNA 数据存储》, 《科学报告》, 第 7 卷，第 1 期，文章编号 5011，2017 年。[9] L. Organick、S. D. Ang、Y.-J. Chen、R. Lopez、S. Yekhanin、K. Makarychev、M. Z. Racz、G. Kamath、P. Gopalan、B. Nguyen 等, 《大规模 DNA 数据存储中的随机访问》, 《自然·生物技术》, 第 36 卷，第 3 期，第 242–248 页，2018 年。[10] L. Anavy、I. Vaknin、O. Atar、R. Amit 和 Z. Yakhini, 《利用复合 DNA 字母减少合成循环次数的 DNA 数据存储》, 《自然·生物技术》, 第 37 卷，第 10 期，第 1229–1236 页，2019 年。[11] D. Bar-Lev、I. Orr、O. Sabary、T. Etzion 和 E. Yaakobi, 《深度 DNA 存储：通过编码理论与深度学习实现可扩展、稳健的 DNA 存储》, 《自然·机器智能》, 第 7 卷，第 639–649 页，2025 年。

[12] S. H. T. Yazdi、H. M. Kiah、E. Garcia-Ruiz、J. Ma、H. Zhao 和 O. Milenkovic，《基于 DNA 的存储：趋势与方法》，*IEEE* 分子、生物与多尺度通信汇刊，第 1 卷，第 3 期，第 230–248 页，2015 年。[13] I. Shomorony、R. Heckel 等，《DNA 数据存储的信息论基础》，《基础与趋势》® 通信与信息论，第 19 卷，第 1 期，第 1–106 页，2022 年。[14] R. Heckel、G. Mikutis 和 R. N. Grass，《DNA 数据存储信道的特征描述》，《科学报告》，第 9 卷，第 1 期，文章编号 9663，2019 年。[15] D. Bar-Lev、O. Sabary、R. Gabrys 和 E. Yaakobi，《覆盖所有碱基：如何最大限度地降低 DNA 存储系统的测序覆盖度》，*IEEE* 信息论汇刊，第 71 卷，第 1 期，第 192–218 页，2024 年。[16] A. Boruchovsky、O. Elishco、R. Gabrys、A. Gruica、I. Tamo 和 E. Yaakobi，《勇闯 First：DNA 存储中的随机访问问题》，*arXiv* 预印本：*2501.12274*，2025 年。[17] R. Cao 和 X. Chen，《优化 DNA 存储中的测序覆盖深度：来自 DNA 存储数据的见解》，载于*2025* 年 *IEEE* 信息论国际研讨会（*ISIT*），第 1–6 页，IEEE，2025 年。[18] A. Gruica、D. Bar-Lev、A. Ravagnani 和 E. Yaakobi，《DNA 存储随机访问效率的组合视角》，*IEEE* 信息论汇刊，第 71 卷，第 12 期，第 9395–9412 页，2025 年。[19] A. Gruica、M. Montanucci 和 F. Zullo，《DNA 存储随机访问码的几何》，*arXiv* 预印本：*2411.08924*，2024 年。[20] S.. Bodur、S. Lia、H. H. L´opez、R. Ludhani、A. Ravagnani 和 L. Seccia，《DNA 覆盖深度问题中的随机变量》，*arXiv* 预印本：*2507.20645*，2025 年。[21] C. Wang 和 E. Yaakobi，《DNA 存储中的随机访问：算法、构造与界》，*arXiv* 预印本：*2601.07053*，2026 年。[22] H. Abraham、R. Gabrys 和 E. Yaakobi，《覆盖所有碱基：DNA 测序效率的新一局》，载于*2024* 年 *IEEE* 信息论国际研讨会（*ISIT*），第 464–469 页，IEEE，2024 年。[23] S. Chandak、K. Tatwawadi、B. Lau、J. Mardia、M. Kubit、J. Neu、P. Griffin、M. Woot- ters、T. Weissman 和 H. Ji，《利用低密度奇偶校验码改进 DNA 存储的读写成本权衡》，载于*2019* 年第 *57* 届 *Allerton* 通信、控制与计算年会（*Allerton*），第 147–156 页，IEEE，2019 年。[24] I. Preuss、B. Galili、Z. Yakhini 和 L. Anavy，《组合式 DNA 存储系统的测序覆盖度分析》，*IEEE* 分子、生物与多尺度通信汇刊，2024 年。[25] R. Sokolovskii、P. Agarwal、L. A. Croquevielle、Z. Zhou 和 T. Heinis，《面向基于组合基序的 DNA 存储的优惠券收集信道编码》，*arXiv* 预印本：*2406.04141*，2024 年。[26] I. Preuss、M. Rosenberg、Z. Yakhini 和 L. Anavy，《利用短序列组合编码实现高效 DNA 数据存储》，$bioRxiv$，第 2021–08 页，2021 年。

[27] T. Cohen 和 E. Yaakobi，《优化复合 DNA 的解码概率与覆盖率》，载于*2024* 年 *IEEE* 信息论国际研讨会 *(ISIT)*，第 1949–1954 页，IEEE，2024 年。[28] M. Bertuzzo、A. Ravagnani 和 E. Yaakobi，《小字母表 DNA 存储中的覆盖深度问题》，载于*2025* 年 *IEEE* 信息论国际研讨会 *(ISIT)*，第 1-6 页，IEEE，2025 年。[29] D. J. Newman，《双 Dixie 杯问题》，《美国数学月刊》，第 67 卷，第 1 期，第 58–61 页，1960 年。[30] P. Erd¨os 和 A. R´enyi，《概率论的一个经典问题》，*Magyar Tud.Akad. Mat. Kutat´o Int. K¨ozl*，第 6 卷，第 1 期，第 215–220 页，1961 年。[31] W. Feller，《概率论及其应用导论》，第 1 卷，第 3 版。Wiley，纽约，1968 年。[32] P. Flajolet、D. Gardy 和 L.Thimonier，《生日悖论、集邮者问题、缓存算法与自组织搜索》，《离散应用数学》，第 39 卷，第 3 期，第 207–229 页，1992 年。[33] B. Segre，《有限空间中的正规有理曲线与 k-弧》，*Ann. Mat. Pura Appl.*，第 39 卷，第 357–379 页，1955 年。[34] F. J. MacWilliams 和 N. J. A. Sloane，《纠错码理论》，第 16 卷。爱思唯尔，1977 年。[35] S. Ball，《有限向量空间中每个基大小子集都是基的向量集合》，*J. Eur. Math. Soc.*，第 14 卷，第 3 期，第 733–748 页，2012 年。[36] S. Ball 和 J. De Beule，《有限向量空间中每个基大小子集都是基的向量集合 II》，*Des. Codes Cryptogr.*，第 65 卷，第 5–14 页，2012 年。[37] A. Ravagnani，《正则格上支撑码的对偶性及其在枚举组合学中的应用》，*Des. Codes Cryptogr.*，第 86 卷，第 2035–2063 页，2018 年。[38] W. C. Huffman 和 V. Pless，《纠错码基础》，剑桥大学出版社，2003 年。[39] R. P. M. J. Jurrius 和 R. Pellikaan，《扩展权重枚举式与广义权重枚举式》，国际编码与密码学研讨会论文集 *WCC-2009*，第 76–91 页，2009 年。.[40]G. Landsberg，《关于一个计数问题及与之相关的级数》，*J. Reine Angew. Math.*，第 11 卷，第 87-88 页，1893 年。[41] R. P. M. J. Jurrius，《有限空间中的码的权重枚举》，*Des. CodesCryptogr.*，第 63 卷，第 3 期，第 321–330 页，2012 年。