中国科学技术大学

博士学位论文

集合系及其在存储系统中的应用

作者姓名: 余文俊

学科专业: 应用数学

导师姓名: 张先得 特任教授 葛根年 教授

完成时间: 二〇二一年十月二十九日

University of Science and Technology of China

A dissertation for doctor's degree

Set systems and their applications in storage systems

Author: Wenjun Yu

Speciality: Applied Mathematics

Supervisors: Prof. Xiande Zhang, Prof. Gennian Ge

Finished time: October 29, 2021

中国科学技术大学学位论文原创性声明

本人声明所呈交的学位论文，是本人在导师指导下进行研究工作所取得的成果。除已特别加以标注和致谢的地方外，论文中不包含任何他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的贡献均已在论文中作了明确的说明。

作者签名：余文俊　　签字日期：2021.10.29

中国科学技术大学学位论文授权使用声明

作为申请学位的条件之一，学位论文著作权拥有者授权中国科学技术大学拥有学位论文的部分使用权，即：学校有权按有关规定向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅，可以将学位论文编入《中国学位论文全文数据库》等有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。本人提交的电子文档的内容和纸质论文的内容相一致。

保密的学位论文在解密后也遵守此规定。

☑ 公开　☐ 保密（____ 年）

作者签名：余文俊　　导师签名：张先得

签字日期：2021.10.29　　签字日期：2021.10.29

摘　要

近年来，极值组合的蓬勃发展不仅推动组合理论的快速发展，而且还促进存储系统的广泛应用。在极值组合中，集合系是备受瞩目的研究对象。常见的集合系有图、超图、相交集族，这些集合系也常被用于刻画存储系统中的编码问题。本学位论文一方面对集合系的极值问题进行理论上的探索，另一方面利用极值组合方法推动解决存储系统中的相关应用问题。具体而言，本文利用集合系中的外代数与图论知识对房屋分配、分布式存储以及 DNA 编码等问题进行研究并取得一定的进展。

第一章将介绍集合系中相关的概念以及房屋分配和存储系统所研究问题的背景，并概述本人对这些问题的主要贡献。

第二章从理论角度出发研究 Bollobás 型集合系，并利用该理论解决源自房屋分配的相关问题。房屋分配是组合优化中的经典问题，它被用于解决各类现实问题。其中，Pareto 最优分配是房屋分配问题中最令人关心的分配之一。本章着重研究 Pareto 最优分配在房屋分配中的数目。该问题可以部分地转化成如下 Bollobás 型集族猜想：

令 a, b 是两个正整数，$\{A_i, B_i\}_{i=1}^m$ 是一组集合对，其中对任意 $i \in [m]$ 均有 $|A_i| = a$, $|B_i| = b$，并且要求 $A_i \cap B_j = \varnothing$ 当且仅当 $i = j$，同时对任意 $1 \leqslant i, j \leqslant m$ 均有 $A_i \cap A_j \neq \varnothing$。则猜测集合对的数目 $m \leqslant \binom{a+b-1}{a-1}$。

此猜想可以看成一个带有相交性质的 Bollobás 型集合系问题。借助于经典的外代数方法，本章证明了该猜想以及相关稳定性结果，并将这些结论推广到有限维实空间上。

第三章研究集合系在分布式存储系统中的应用。本章关注的集合系是图与超图，运用这些集合系分析分布式存储系统中的编码问题。为存储海量的数据，一种基于分布式存储方案的 DRESS 码受到了广泛地关注，它主要由一个外层 MDS 码和一个内层部分重复码嵌套组成。考虑到原始数据带有 “热度” 这一特性，本章研究同时具有最优容量的部分重复码和热度均衡的分布式存储系统。为获得高性能的分布式存储系统，本章提出了一个新型的访问均衡模型，并将该问题转化成极值组合中的一类图标号问题。本章通过研究图标号问题给出几类达到访问均衡的最优部分重复码，改进了 Dau 等人提出的 MaxMinSum 模型。

第四章研究集合系在 DNA 存储系统中的应用。本章研究的集合系是受限相交族，并从该角度对 DNA 存储系统中的编码问题进行探究。DNA 作为一类优异的存储介质，相比较于传统的存储介质具有高密度存储、长使用寿命以及易复制等诸多优势。DNA 上的编码可以帮助解决如今数据极速增长和长期存储的需

求。本章依据现在的生物技术以及 DNA 自身的生物特性研究了 DNA 存储系统中的平衡 (t,k,v) 集合码，它为 DNA 存储系统提供了一类优异的编码。显而易见的是，平衡 (t,k,v) 集合码等价于组合设计中带标号的 t 填充设计。本章首先把平衡集合码转换成图或者超图上的极值结构并改进平衡集合码的上界，然后再利用一些图论、组合设计中的技巧寻找极值结构，最后给出所有参数为 $(2,3,v)$，$(2,4,v)$，$(3,4,v)$ 的最优平衡集合码，这完整地构造出所有 $k \leqslant 4$ 的最优平衡集合码。此外，对于所有 $k \geqslant 4$ 的偶数和 $t \geqslant 3$ 的正整数，本章还分别构造出渐近最优平衡 $(2,k,v)$ 集合码和渐近最优平衡 $(t,t+1,v)$ 集合码。相比较于以前的工作，本章提供了更简单的最优平衡 $(2,3,v)$ 集合码和更多参数的最优平衡 (t,k,v) 集合码的构造。

第五章对其他极值组合问题和个人成果做简要地介绍。

关键词：极值组合；集合系；图论；组合设计；外代数；存储系统；部分重复码；集合码

ABSTRACT

Extremal combinatorics has been rapidly developed in recent years, which not only immensely enhances the development of the theory of Combinatorics, but also promotes the widespread application of storage systems. Set systems are attractive research objects in extremal combinatorics. There are some common set systems, such as graphs, hypergraphs and intersecting set families, which can be used to characterize the coding problems in storage systems. This thesis mainly explores some extremal problems related to set systems and use this theory to obtain several good results in storage systems. More precisely, by applying certain exterior algebra method and graph theory, we investigate several problems on house allocation, distributed storage systems and DNA data storage systems.

In Chapter 1, we will introduce some relative concepts of set systems and several backgrounds of problems arisen from house allocation and storage systems. Meanwhile, we summarize our main contributions to those problems.

In Chapter 2, we mainly study the Bollobás type set system from the theoretical perspective, and utilize it to solve some related problems from house allocation. House allocation is a classical problem of combinatorial optimization, which has been used to solve various practical problems. Especially, we focus on the number of Pareto optimal matchings in house allocation, which is one of the most concerned matchings in house allocation. This problem can be partially converted to the following conjecture of Bollobás-type set system.

Let $\{A_i, B_i\}_{i=1}^m$ be a collection of set pairs such that $A_i \cap B_j = \emptyset$ if and only if $i = j$ and $A_i \cap A_j \neq \emptyset$ for any $1 \leqslant i, j \leqslant m$. Let $|A_i| = a$, $|B_i| = b$ for any $i \in [m]$, where a, b are positive integers, then it is conjectured that $m \leqslant \binom{a+b-1}{a-1}$.

This conjecture can be considered as a problem of Bollobás-type set system with intersecting property. We prove this conjecture and its stability results by the classical exterior algebra argument. In addition, we generalize these results to the setting of finite dimensional real spaces.

In Chapter 3, we study the application of set systems in distributed storage systems. The set systems we are concerned with are graphs and hypergraphs, which are used to analyse the coding problems in distributed storage systems. In order to storage massive data, one DRESS code based on distributed storage scheme has received wide attention. This code consists of the concatenation of an outer MDS code and an inner fractional

repetition code. As the raw data has its own "popularity", we consider one distributed storage system with optimal fractional repetition codes and balanced popularities. To design a distributed storage system with high performance, we propose a novel access-balancing model and turn this problem into the problem of graph labelings, which is a research topic of extremal combinatorics. By studying the graph labeling problems, we provide several optimal fractional repetition codes with access-balancing property and refine the MaxMinSum model introduced by Dau *et al*.

In Chapter 4, we study the application of set systems in the DNA storage systems. The set systems we study are restricted intersecting families. From this point of view, we explore the coding problems in DNA storage systems. As one excellent storage media, DNA, compared to conventional media, has many advantages, including high data density, longevity and ease of copying information. DNA codes can help to overcome challenges from the increasing amount of data and the need for long-term data storage. According to the current biotechnology and DNA's own biological features, we mainly discuss the balanced (t, k, v) set codes, which provide a kind of good DNA codes. A balanced (t, k, v) set code is equivalent to a labeled t-packing in combinatorial designs. After turning balanced set codes into some extremal structures on graphs or hypergraphs, we improve the upper bound of balanced set codes. Furthermore, by couples of techniques from graph theory and combinatorial designs, we exhibit constructions of optimal balanced $(2, 3, v), (2, 4, v), (3, 4, v)$ set codes, which contains optimal balanced (t, k, v) set codes for all $k \leqslant 4$. Besides, for even number $k \geqslant 4$ and integer $t \geqslant 3$, we construct asymptotically optimal balanced $(2, k, v)$ set codes and asymptotically optimal balanced $(t, t + 1, v)$ set codes respectively. Comparing with former works, we provide a simpler construction of optimal balanced $(2, 3, v)$ set codes and more constructions of optimal balanced (t, k, v) set codes with different parameters.

In Chapter 5, we will shortly introduce other extremal combinatorial problems and our contributions.

Key Words: Extremal combinatorics; Set system; Graph theory; Combinatorial design; Exterior algebra; Storage system; Fractional repetition code; Set code

目 录

插 图 清 单

表 格 清 单

第 1 章　绪　　论

极值组合的蓬勃发展促进了存储系统以及其他应用数学领域的进步。反之，应用领域源源不断产生出来的问题也推动极值组合在新方向上的前进。特别地，在如今大数据时代下，海量的数据急需存储至相应的系统中。人们为达到数据在不同场景下的存储目标，对存储系统也提出了各种各样的需求，例如：高安全、低成本、强隐私等。这些大都可以转换成满足特定条件的极值组合问题。简而言之，极值组合主要研究组合数学中的极值问题，同时也为存储系统中的编码问题提供最优的解决方案。这使得极值组合在存储系统领域发挥着强大的作用。

通常来说，存储系统中的研究对象基本都能用极值组合中的集合系来表示。作为极值组合研究最多的对象之一，集合系包含了耳熟能详的图、超图、相交集族等。这促使组合数学和应用数学相关领域学者同时对集合系展开大量研究。本文首先从理论角度探索集合系中的极值问题，证明了集合系中 Bollobás 型猜想并利用该理论结果改进房屋分配中的相关结论。紧接着本文在应用层面上，从集合系的角度刻画两类存储系统中的编码问题：部分重复码上的访问均衡以及 DNA 存储中的集合码。本文借助于极值组合的若干方法成功地获得一些最优的解决方案。下面将概述各子课题的研究背景以及本文对各子课题的贡献。

1.1　房屋分配中的组合问题

组合优化是应用数学的热门方向之一，大量的困难问题实质上都对应于某种优化问题。例如：网络拥塞、交通调度、背包问题、最短路径问题等。其中房屋分配作为当中的经典问题受到学者们广泛地关注。房屋分配主要考虑如下问题：在给定买家集 $\boldsymbol{A}$、房屋集 $\boldsymbol{B}$ 以及买家对房屋的喜好排序表之后，如何寻找一个从买家集 $\boldsymbol{A}$ 到房屋集 $\boldsymbol{B}$ 之间的单射 σ，称为分配，使得买家都尽可能分配到各自喜欢的房子。特别地，有如下标准判断两个房屋分配 σ 与 σ' 的优异程度：如果对于任意的买家 $i \in \boldsymbol{A}$，在买家 i 的房屋排序表中，σ 分配的房屋 $\sigma(i)$ 都不低于 σ' 分配的房屋 $\sigma'(i)$，则分配 σ 是好于分配 σ'。人们依据此标准主要考虑一类最优的房屋分配，帕累托最优 (Pareto optimal) 分配：不存在其他分配好于该分配。

帕累托最优分配的数目是房屋分配中的重要指标，它可以通过房屋集 $\boldsymbol{B}$ 中可达集 $\boldsymbol{E}$ 的数目来间接刻画。其中可达集 $\boldsymbol{E} \subseteq \boldsymbol{B}$ 是某个帕累托最优分配 σ 的像，也即 $\boldsymbol{E} = \{\sigma(a) : a \in \boldsymbol{A}\}$。当 $|\boldsymbol{E}| = m$ 的时候，$f(m)$ 表示房屋集 $\boldsymbol{B}$ 中可达集的数目。Gerbner 等人在文献[1] 探索了该问题，并给出如下极值集合论中的猜想和

$f(m)$ 的估计。

猜想 1.1 [1] 令 $AK(n,k,t)$ 表示 k 一致 t 相交集族 $\mathcal{F} \subseteq \binom{[n]}{k}$ 的最大值，$\{(A_i,B_i)\}_{i=1}^m$ 为集合对，其中对任意 $i \in [m]$ 均有 $|A_i| = a \leqslant b = |B_i|$。假设存在某个整数 $t \geqslant 0$，满足

- 对任意的 $1 \leqslant i,j \leqslant m$，$|A_i \cap A_j| \geqslant t$，
- 对任意的 $1 \leqslant i \leqslant m$，$|A_i \cap B_i| = 0$，
- 对任意的 $1 \leqslant i \neq j \leqslant m$，$|A_i \cap B_j| > 0$。

则

$$m \leqslant AK(a+b,a,t)。 \tag{1.1}$$

定理 1.2 [1] 假设猜想 1.1成立，则 $f(m) \leqslant AK(m,m,\lceil \frac{m}{2} \rceil)$。

特别地，猜想 1.1是极值集合论中重要基石之一的 Bollobás 定理某种推广形式。下面简要提及关于交叉相交集合对的著名 Bollobás 定理。

定理 1.3 (Bollobás 定理) [2] 设 $(A_1,B_1),\cdots,(A_m,B_m)$ 是集合对，其中对任意 $1 \leqslant i \leqslant m$ 均有 $|A_i| = a$，$|B_i| = b$。假设

- 对任意的 $1 \leqslant i \leqslant m$，$A_i \cap B_i = \varnothing$，
- 对任意的 $i \neq j$，$A_i \cap B_j \neq \varnothing$。

则

$$m \leqslant \binom{a+b}{a}。 \tag{1.2}$$

进一步地，等号成立当且仅当存在某个大小为 $a+b$ 的集合 X，使得对每个 i，A_i 是 X 的 a 元子集且 $B_i = X \setminus A_i$。

本文的主要工作在于利用 Lovász[3] 优雅的外代数 (或者楔乘积) 方法以及 Scott 和 Wilmer[4] 关于外代数与超图的新对应证明了有限维实向量空间上的新型带权重 Bollobás 定理以及相关的推论和稳定性定理。以上的工作可以用下面三个定理概括。

定理 1.4 (向量空间上的新型 Bollobás 定理)　设 $\{(A_i,B_i)\}_{i=1}^m$ 是有限维实向量空间中的子空间对，其中对任意 $1 \leqslant i \leqslant m$ 均有 $\dim(A_i) = a_i$，$\dim(B_i) = b_i$ 及 $a_i \leqslant b_i$。假设存在某个整数 $t \geqslant 0$，满足

- 对任意的 $1 \leqslant i,j \leqslant m$，$\dim(A_i \cap A_j) > t$，
- 对任意的 $1 \leqslant i \leqslant m$，$\dim(A_i \cap B_i) \leqslant t$，
- 对任意的 $1 \leqslant i < j \leqslant m$，$\dim(A_i \cap B_j) > t$，
- 对任意 $1 \leqslant i \leqslant m$ 均有 $a_i + b_i = N$ 和 $a_1 \leqslant a_2 \leqslant \cdots \leqslant a_m$，其中 N 是某个给定的正整数。

则

$$\sum_{i=1}^m \binom{N-(2t+1)}{a_i-(t+1)}^{-1} \leqslant 1。 \tag{1.3}$$

并且当对任意 $1 \leqslant i \leqslant m$ 均有 $a_i < b_i$ 的时候，等号成立仅当 $a_1 = a_2 = \cdots = a_m$ 且 $b_1 = b_2 = \cdots = b_m$。

作为定理 1.4 的直接推论，我们有如下关于子集合对的带权重 “半捆” 型 Bollobás 定理。

定理 1.5 (集合上的新型 Bollobás 定理)　设 $\{(A_i, B_i)\}_{i=1}^m$ 是集合对，其中对任意 $1 \leqslant i \leqslant m$ 均有 $|A_i| = a_i \leqslant |B_i| = b_i$。假设存在某个整数 $t \geqslant 0$，满足

- 对任意的 $1 \leqslant i, j \leqslant m$，$|A_i \cap A_j| > t$，
- 对任意的 $1 \leqslant i \leqslant m$，$|A_i \cap B_i| \leqslant t$，
- 对任意的 $1 \leqslant i < j \leqslant m$，$|A_i \cap B_j| > t$，
- 对任意 $1 \leqslant i \leqslant m$ 均有 $a_i + b_i = N$ 和 $a_1 \leqslant a_2 \leqslant \cdots \leqslant a_m$，其中 N 是某个给定的正整数。

则

$$\sum_{i=1}^{m} \binom{N-(2t+1)}{a_i-(t+1)}^{-1} \leqslant 1。 \tag{1.4}$$

特别地，定理1.5 可以完美地解决猜想 1.1中 $t = 1$ 的情形，从而给出可达集数目的上界。

定理 1.6 (稳定性定理)　设 $\{(A_i, B_i)\}_{i=1}^m$ 是集合对，其中对任意 $1 \leqslant i \leqslant m$ 均有 $|A_i| = a < |B_i| = b$。假设

- 对任意的 $1 \leqslant i, j \leqslant m$，$|A_i \cap A_j| > 0$，
- $|A_i \cap B_j| = 0$ 当且仅当 $i = j$。

则

$$m \leqslant \binom{a+b-1}{a-1}。$$

特别地，定理中等号成立当且仅当集合对的结构如下：在大小为 $a+b$ 的背景集 $X = \bigcup_{i=1}^m (A_i \cup B_i)$ 中，$\{A_i\}_{i=1}^m$ 是 X 中所有包含某个固定元素的 a 元子集，且对任意的 i，$B_i = X \setminus A_i$。

1.2　部分重复码上的访问均衡问题

伴随着现代互联网的快速发展，海量的数据需要使用更加安全、经济的方式存储。相较于较传统的整体存储以及重复存储的方案，近年来提出的分布式存储系统则具有更便捷的操作、更低的运营成本以及更强的性能等优势。对于需要长期存储且带有 “热度” 的数据而言，访问均衡是衡量存储效率的关键性指标。在结合分布式存储系统的结构和数据带有热度的特点之后，设计出分布式存储系统中合理的数据排布方案显得尤为重要。为此，本文提出一个新的数据排布模型并找到几类达到访问均衡的最优部分重复码。

在本文所考虑的分布式存储系统中，每一个文件被分割成若干个大小相同的子文件。首先这些子文件通过外层 MDS 码被编码成一些数据块。然后每个数据块被复制固定次数之后，再通过内层部分重复码将它们分布在存储系统中多个存储节点中[5-7]。这样一个外层 MDS 码与内层部分重复码的嵌套方案保障分布式存储系统具有低冗余和可修复的性能。例如著名的最小带宽生成 (MBR) 码[8] 就属于该类分布式存储系统。人们在研究如何均衡每个存储节点内部热度和的时候，通常会利用大量的组合设计构型来排布数据，例如按照斯坦纳三元系的排布方案[9-11]。在现实生活中，Hadoop 分布式文件系统和 Google 文件系统都采取了这样的策略[12]。

访问均衡问题是通过综合每个数据块热度之后再来均衡每个存储节点的访问请求[13]。更具体地说，首先数据块热度将依据其访问频率进行标号，然后要求每个存储节点内的所有数据块标号之和尽可能地均衡。Dau 和 Milenkovic[14] 针对均衡性给出了一些相关标准，例如 MaxMinSum (或者 MinMaxSum) 标准。他们成功地发现所有达到 MaxMinSum 标准的斯坦纳三元系和对偶斯坦纳三元系。进一步关于达到 MaxMinSum 标准的这类问题，文献[15] 研究了柯克曼系，文献[16] 研究了部分斯坦纳系。

分布式存储系统中的存储效率，也就是分布式存储系统中可存储文件的最大数量，它唯一取决于系统中的内层部分重复码。存储效率达到最大的部分重复码是人们关注的研究对象，也被称为最优部分重复码。尽管组合设计构型被广泛地使用在存储方案中的底层构架，但是通常来说它们不是最优部分重复码。为此，Silberstein 和 Etzion[17] 研究出一些基于图和设计的最优部分重复码。其中基于 Turán 图和大围长的图是两类重复次数为 2 的最优部分重复码，基于横截设计和广义多边形则是重复次数更大的最优部分重复码。

本文的工作主要聚焦于最优重复码的访问均衡，观察到 Dau 和 Milenkovic 提出的 MaxMinSum 模型只关注到所有存储节点内部热度和的最小值，对于均衡性的刻画过于局部化。因此本文将引入一个考虑所有存储节点内部热度和方差最小的模型，称之为 MinVar 模型。该新模型为分布式存储系统中热度数据的长期存储提供了一个全局均衡标准。本文利用极值组合中的技巧将 MinVar 模型问题等价于集合系上的区组标号问题。具体描述如下：

问题 1.7 (MinVar 模型问题)　给定一个 n 阶 ρ 一致 α 正则的集合系 $S = (V, E)$，其中 $V = \{v_1, v_2, \cdots, v_n\}$，$E = \{e_1, e_2, \cdots, e_\theta\}$ 且参数满足关系 $\theta = n\alpha/\rho$。那么构造一个由 S 生成且达到 *MinVar* 的 (n, α, ρ) 部分重复码，这样的问题等价于寻找一个 E 上的区组标号，也即，从集合 E 到集合 $[\theta]$ 之间的双射 σ，使得访问方差

$$Var(S_\sigma) := \| I(S)(\sigma(e_1), \cdots, \sigma(e_\theta))^T - (\bar{a}, \cdots, \bar{a})^T \|_{\mathbb{L}^2}$$

最小化。其中 $\bar{a}$ 等于平均热度 $\frac{\rho\theta(\theta+1)}{2n}=\frac{\alpha(\theta+1)}{2}$ 且对任意向量 $\mathbf{x}=(x_1,x_2,\cdots,x_n)^T$，符号 $\|\mathbf{x}\|_{\mathbb{L}^2}:=\sum_{i=1}^{n}x_i^2$。

特别地，MinVar 模型达到方差为零的时候恰好对应为组合学的图幻标问题，由此我们发现了如下访问请求达到均衡的最优部分重复码。

定理 1.8　令 $r\geqslant 2$，r 整除 n 且 $\alpha=(r-1)\frac{n}{r}$。当下列任意条件之一成立的时候，最优 $(n,\alpha,2)$ 部分重复码达到访问均衡。

(1) $n=r=2$ 或者 $n=r\geqslant 6$ 且 $n\not\equiv 0 \pmod 4$;

(2) $n=2r\geqslant 6$;

(3) $n\geqslant 3r$，除了 $r\equiv 0 \pmod 4$ 且 $\frac{n}{r}$ 是奇数的情形。

我们接着对线性集合系上的 MinVar 模型提出一个对偶等价的 MinPS 问题，它是图的顶点标号问题。具体描述如下：

问题 1.9 (MinPS 问题)　给定一个 θ 阶 d 正则图 G，其中图的顶点为 $v_1,v_2,\cdots,v_\theta$。哪个重量函数 $f:V(G)\to[\theta]$ (双射) 使得 $M(f):=\sum_{v_i\sim v_j}f(v_i)f(v_j)$ 最小化?

本文利用重量转移、归纳等方法解决了多个完全图的并、多个 Turán 图的并以及圈上的 MinPS 问题，随之解决了对应的 MinVar 问题。此外，本文提供了这些访问均衡的部分重复码的显示构造。

1.3　DNA 存储中的集合码

互联网的蓬勃发展随之产生了海量数据，这些数据主要利用光介质和磁介质媒体记录。现有存储系统保证数据能够有效接受、检索和复制[18]，其主要的特性包括：随机访问、高精度检索、低成本和实时操作。但是面对数据的极速增长和长期存储需求，未来可行的 DNA 存储系统和多聚合物存储系统[19-22] 对比于现有存储系统具有诸多显著的优势。比如，DNA 存储具有存储密度高、使用寿命长和复制简易等优势[19]。利用如今的 DNA 测序、合成以及编辑技术优势[23-24]，DNA 存储系统有潜力突破传统存储系统遇到的瓶颈并发挥巨大的存储效益。

近年来人们提出一些 DNA 上的编码技术模型[25-29]。在这些模型中，数据将被排布在 DNA 双链的不同切割点中。为了防止由切割点而导致的 DNA 断裂以及纠正 DNA 上的数据读取错误，数据被要求尽可能地均匀分布在 DNA 的两条链中且数据被存储在少量重叠的切割点中。Gabrys 等人[30] 发现该问题等价于一个具有较小差异值和较小相交数的 (t,k,v) 集合码，通常来说 (t,k,v) 集合码是一个 $\mathcal{S}\subseteq\binom{[v]}{k}$ 的集族，其中 $[v]$ 的任意 t 元集最多包含在一个 $S\in\mathcal{S}$ 中。同时

Gabrys 等人[30] 也为具有最小差异值的 (t,k,v) 集合码提供了理论上界：

$$A(t,k,v)\leqslant \frac{\binom{\lceil v/2\rceil}{\lfloor t/2\rfloor}\binom{\lfloor v/2\rfloor}{\lceil t/2\rceil}}{\binom{\lceil k/2\rceil}{\lfloor t/2\rfloor}\binom{\lfloor k/2\rfloor}{\lceil t/2\rceil}}。\tag{1.5}$$

该集合码与组合设计的 t 填充理论[31] 密切相关，但是构造出具有较小差异值和较小相交数的最大集合码却是一个困难的挑战。

本文的主要工作是构造出拥有最小差异值和固定相交数的最大 (t,k,v) 集合码，也称为最优平衡 (t,k,v) 集合码。对于 $k\leqslant 4$ 的时候，我们利用图的语言将最优平衡集合码等价于图上边标号问题或者特殊 “平衡表”，并改进了平衡 (t,k,v) 集合码的上界。本文借助于一些组合设计以及图论结果，例如，正交阵列、Howell 设计、边染色理论等，构造出一些最优平衡集合码。这些工作分别通过以下三个定理表示。

定理 1.10　平衡 $(2,3,v)$ 集合码的上界为：

$$A(2,3,v)\leqslant\begin{cases}2m^2, & \text{如果 } v=4m,\ m\geqslant 2,\\ 2m^2+m, & \text{如果 } v=4m+1,\ m\geqslant 2,\\ 2m^2+2m, & \text{如果 } v=4m+2,\ m\geqslant 1,\\ (2m+1)(m+1), & \text{如果 } v=4m+3,\ m\geqslant 0。\end{cases}$$

并且对任意 $v\geqslant 6$，达到该上界的平衡 $(2,3,v)$ 集合码都有显示构造。

在文献[30] 中，虽然 Gabrys 等人利用一种新型的拉丁矩阵构造出了达到上界的平衡 $(2,3,v)$ 集合码，但是我们的构造与其本质不同并且更加的简单。

定理 1.11　平衡 $(2,4,v)$ 集合码的上界为：

$$A(2,4,v)\leqslant\begin{cases}m^2, & \text{如果 } v=4m \text{ 或者 } 4m+1,\ m\geqslant 1,\\ m^2+m, & \text{如果 } v=4m+2 \text{ 或者 } 4m+3,\ m\geqslant 1。\end{cases}$$

并且对任意 $v\geqslant 11$，达到该上界的平衡 $(2,4,v)$ 集合码都有显示构造。

定理 1.12　平衡 $(3,4,v)$ 集合码的上界为：

$$A(3,4,v)\leqslant\begin{cases}2m^3-m^2, & \text{如果 } v=4m,4m+1,\ m\geqslant 1,\\ 2m^3+m^2, & \text{如果 } v=4m+2,\ m\geqslant 1,\\ 2m^3+3m^2+m, & \text{如果 } v=4m+3,\ m\geqslant 1。\end{cases}$$

并且对于任意 $v\geqslant 5$，达到该上界的平衡 $(3,4,v)$ 集合码都有显示构造。

对于 $k\geqslant 4$ 的情形，我们分别利用广义 Reed-Solomon 码和 Rödl Nibble 方法，构造出所有偶数 $k\geqslant 4$ 的渐近最优平衡 $(2,k,v)$ 集合码和所有正整数 $t\geqslant 3$ 的渐近最优平衡 $(t,t+1,v)$ 集合码。相比较于 Gabrys 等人[30] 的工作，本文给出更加简单的最优平衡 $(2,3,v)$ 集合码构造，同时提供几类新参数的最优平衡集合码的构造。

第 2 章　房屋分配中的组合问题

这一章将讨论 Bollobás 型集合系中的理论问题并解决房屋分配中的相关问题：研究房屋分配中可达集的数目。Gerbner 等人在文献[1] 中研究了该问题，同时提出如下集合系上的猜想并利用该猜想改进对可达集数目的估计。

令 $\{(A_i, B_i)\}_{i=1}^m$ 为集合对，其中对任意 $i \in [m]$ 均有 $|A_i| = a \leqslant |B_i| = b$。假设存在某个整数 $t \geqslant 0$，满足

- 对任意的 $1 \leqslant i, j \leqslant m$，$|A_i \cap A_j| \geqslant t$，
- 对任意的 $1 \leqslant i \leqslant m$，$|A_i \cap B_i| = 0$，
- 对任意的 $1 \leqslant i \neq j \leqslant m$，$|A_i \cap B_j| > 0$。

则猜测

$$m \leqslant AK(a+b, a, t),$$

其中 $AK(n,k,t)$ 表示 k 一致 t 相交集族 $F \subseteq \binom{[n]}{k}$ 的最大值。

外代数方法和集合系相关结论将被用来证明集合系上的 Bollobás 型不等式和相关的稳定性的结果，其中的推论证明了上述猜想在 $t = 1$ 的时候成立。从而改进可达集数目的上界估计。

2.1　介绍

在房屋分配问题中，主要研究的是在给定的买家集 A，房屋集 B 以及每个买家各自心中的房屋喜好排序列表之后。如何制定一个房屋分配方案使得每个买家都分配到了各自的房子。其中房屋分配 σ 是从买家集 A 到房屋集 B 之间的单射。如果两个不同的房屋分配 σ 与 σ' 满足：对于任意的买家 i，在买家 i 的房屋排序表中，房屋 $\sigma(i)$ 都不低于房屋 $\sigma'(i)$，那称分配 σ 好于分配 σ'。其中最令人关注的帕累托最优 (Pareto optimal) 分配是指，不存在其他的分配好于该分配。

房屋集 B 中的子集 E 称为是 可达的，如果存在一个帕累托最优分配 σ，使得 $\{\sigma(a) : a \in A\} = E$。令 $|E| = m$，$f(m)$ 表示 B 中可达集的数目。那么 $f(m)$ 可以来度量分配中的帕累托最优分配的数量。

Gernber 等人在文献[1] 中给出了如下的猜想和 $f(m)$ 上界的估计。

猜想 2.1 [1] 令 $AK(n,k,t)$ 表示 k 一致 t 相交集族 $F \subseteq \binom{[n]}{k}$ 的最大值，$\{(A_i, B_i)\}_{i=1}^m$ 为集合对，使得对任意 $i \in [m]$ 均有 $|A_i| = a \leqslant |B_i| = b$。假设存在某个整数 $t \geqslant 0$，满足

- 对任意的 $1 \leqslant i, j \leqslant m$，$|A_i \cap A_j| \geqslant t$，
- 对任意的 $1 \leqslant i \leqslant m$，$|A_i \cap B_i| = 0$，

- 对任意的 $1 \leqslant i \neq j \leqslant m$，$|A_i \cap B_j| > 0$。

则

$$m \leqslant AK(a+b, a, t)。$$

定理 2.2 [1] 假设猜想 2.1成立，则 $f(m) \leqslant AK(m, m, \lceil \frac{m}{2} \rceil)$。

下面我们主要研究猜想 2.1 在 $t = 1$ 的情形以及相关的定理。

1965 年 Bollobás 证明了如下关于交叉相交集合对的定理。它是极值集合论中的重要基石之一。

定理 2.3 (Bollobás 定理) [2] 令 $(A_1, B_1), \cdots, (A_m, B_m)$ 是一组集合，其中对任意 $1 \leqslant i \leqslant m$ 均有 $|A_i| = a$ 及 $|B_i| = b$。假设

- 对任意的 $1 \leqslant i \leqslant m$，$A_i \cap B_i = \varnothing$，
- 对任意的 $i \neq j$，$A_i \cap B_j \neq \varnothing$。

则

$$m \leqslant \binom{a+b}{a}。 \tag{2.1}$$

进一步地，等号成立当且仅当存在某个大小为 $a+b$ 的集合 X，使得对任意的 i，A_i 是 X 的 a 元子集且 $B_i = X \setminus A_i$。

经过多年的发展，关于 Bollobás 定理的诸多证明方法和各种各样的拓展发表在大量文献中 (见[4,32-49])。在这诸多证明中，Lovász 利用外代数 (或者楔乘积) 的证明[3] 被视为令人惊艳的优雅证明，并且在处理这种约束类型的集合对或者子空间对问题中提供了全新的方向。同时 Lovász 将这个定理推广到向量空间。

Füredi[32] 利用 Lovász 的方法在 1984 年证明了如下向量空间中的 Bollobás 定理。

定理 2.4 [32] 令 $(A_1, B_1), \cdots, (A_m, B_m)$ 是有限维实向量空间中的非平凡子空间对，其中对任意 $1 \leqslant i \leqslant m$ 均有 $\dim(A_i) = a$，$\dim(B_i) = b$。假设对某个整数 $t \geqslant 0$，满足

- 对任意的 $1 \leqslant i \leqslant m$，$\dim(A_i \cap B_i) \leqslant t$，
- 对任意的 $1 \leqslant i < j \leqslant m$，$\dim(A_i \cap B_j) \geqslant t+1$。

则

$$m \leqslant \binom{a+b-2t}{a-t}。$$

最近 Scott 和 Wilmer[4] 沿着 Lovász 和 Füredi 的道路建立起外代数和超图的新对应。该方法被证明是处理具有交叉相交性质的 Bollobás 型集合对的高效方法，并证明了如下有限维实向量空间中的带权重的 Bollobás 定理。

定理 2.5 [4] 令 $(A_1, B_1), \cdots, (A_m, B_m)$ 是有限维实向量空间的非平凡子空间对，其中对任意的 $1 \leqslant i \leqslant m$，记 $a_i = \dim(A_i)$，$b_i = \dim(B_i)$，假设

- 对任意的 $1 \leqslant i \leqslant m$，$\dim(A_i \cap B_i) = 0$，

- 对任意的 $1 \leqslant i \neq j \leqslant m$，$\dim(A_i \cap B_j) > 0$，
- $a_1 \leqslant a_2 \leqslant \cdots \leqslant a_m$ 及 $b_1 \geqslant b_2 \geqslant \cdots \geqslant b_m$。

则

$$\sum_{i=1}^{m} \binom{a_i + b_i}{a_i}^{-1} \leqslant 1。$$

本章利用外代数方法和文献[4] 中外代数与超图的新对应证明了实向量空间中的新型带权重 Bollobás 定理。与定理 2.5不同，本章将原始的限制条件推广到：对任意的 $1 \leqslant i \leqslant m$ 均有 $\dim(A_i \cap B_i) \leqslant t$ 以及对任意的 $1 \leqslant i < j \leqslant m$ 和某个整数 $t \geqslant 0$ 均有 $\dim(A_i \cap B_j) > t$。此外，本章额外要求子空间集 $\{A_i\}_{i=1}^m$ 是 $(t+1)$ 相交的，从而称该定理为带权重 “半捆” 型 Bollobás 定理。它的正式描述如下。

定理 2.6　令 $\{(A_i, B_i)\}_{i=1}^m$ 是有限维实向量空间的子空间对，其中对任意 $1 \leqslant i \leqslant m$ 均有 $\dim(A_i) = a_i$，$\dim(B_i) = b_i$ 及 $a_i \leqslant b_i$。假设对某个整数 $t \geqslant 0$，满足

- 对任意的 $1 \leqslant i, j \leqslant m$，$\dim(A_i \cap A_j) > t$，
- 对任意的 $1 \leqslant i \leqslant m$，$\dim(A_i \cap B_i) \leqslant t$，
- 对任意的 $1 \leqslant i < j \leqslant m$，$\dim(A_i \cap B_j) > t$，
- 对任意 $1 \leqslant i \leqslant m$ 均有 $a_i + b_i = N$ 和 $a_1 \leqslant a_2 \leqslant \cdots \leqslant a_m$，其中 N 是某个给定的正整数。

则

$$\sum_{i=1}^{m} \binom{N - (2t+1)}{a_i - (t+1)}^{-1} \leqslant 1。$$

当对任意 $1 \leqslant i \leqslant m$ 均有 $a_i < b_i$ 的时候，等号成立仅当 $a_1 = a_2 = \cdots = a_m$ 及 $b_1 = b_2 = \cdots = b_m$。

作为定理 2.6 的直接推论，我们有如下关于子集合对的带权重 “半捆” 型 Bollobás 定理。

定理 2.7　令 $\{(A_i, B_i)\}_{i=1}^m$ 是集合对，其中对任意的 $1 \leqslant i \leqslant m$ 均有 $|A_i| = a_i \leqslant |B_i| = b_i$。假设对某个整数 $t \geqslant 0$，满足

- 对任意的 $1 \leqslant i, j \leqslant m$，$|A_i \cap A_j| > t$，
- 对任意的 $1 \leqslant i \leqslant m$，$|A_i \cap B_i| \leqslant t$，
- 对任意的 $1 \leqslant i < j \leqslant m$，$|A_i \cap B_j| > t$，
- 对任意 $1 \leqslant i \leqslant m$ 均有 $a_i + b_i = N$ 和 $a_1 \leqslant a_2 \leqslant \cdots \leqslant a_m$，其中 N 是某个给定的正整数。

则

$$\sum_{i=1}^{m} \binom{N - (2t+1)}{a_i - (t+1)}^{-1} \leqslant 1。 \tag{2.2}$$

当 $t = 0$，且对任意 $1 \leqslant i \leqslant m$ 均有 $a_i = a, b_i = b$ 的时候，定理 2.7 证实了 Gerbner 等人在房屋分配中的猜想。

Bollobás 在定理 2.3的证明中说明式子 (2.1) 中等号成立当且仅当背景集 X 的大小为 $a+b$，$\{A_1,\cdots,A_m\}=\binom{X}{a}$ 以及 $B_i=X\setminus A_i$。同样地，在定理 2.7 中等式等号成立且 $t=0$ 及 $a<b$ 时，本章也确定出 $\{(A_i,B_i)\}_{i=1}^m$ 的唯一结构。

定理 2.8　令 $\{(A_i,B_i)\}_{i=1}^m$ 是集合对，其中对任意 $1\leqslant i\leqslant m$ 均有 $|A_i|=a<|B_i|=b$。假设

- 对任意的 $1\leqslant i,j\leqslant m$，$|A_i\cap A_j|>0$，
- $|A_i\cap B_j|=0$ 当且仅当 $i=j$。

则

$$m\leqslant\binom{a+b-1}{a-1}。$$

特别地，定理中等号成立当且仅当集合对的结构如下：在大小为 $a+b$ 的背景集 $X=\bigcup_{i=1}^m(A_i\cup B_i)$ 中，$\{A_i\}_{i=1}^m$ 是 X 中所有包含某个固定元素的 a 元子集，且对任意的 i，$B_i=X\setminus A_i$。

注解　当 $a=b$ 的时候，背景集合 X 的大小为 $2a$。此时集合系 $\{A_i\}_{i=1}^m$ 的极值情形各种各样，所以上述集合系 $\{A_i\}_{i=1}^m$ 和 $\{B_i\}_{i=1}^m$ 的极值结构同样也不是唯一的。

2.2　准备工作

本小节将回顾由 Scott 和 Wilmer 在文献[4] 提出的外代数与超图之间的联系以及一些与定理证明相关的已知结果。

2.2.1　外代数和超图之间的关系

给定整数 n 和 r，其中 $0\leqslant r\leqslant n$。令 $[n]=\{1,\cdots,n\}$，$\binom{[n]}{r}=\{A\subseteq[n]:|A|=r\}$。如果背景集为 $[n]$ 的超图 $\mathcal{A}\subseteq\binom{[n]}{r}$，则称其为 r 一致的。对任意的有限整数集合 $S\subseteq[n]$，本章规定其中的元素 $S=\{n_1,n_2,\cdots,n_{|S|}\}$ 总是按照 $n_1<n_2<\cdots<n_{|S|}$ 排列。

令 $V=\mathbb{R}^n$ 是一个 n 维实向量空间，其标准基为 $E=\{e_1,\cdots,e_n\}$。令

$$\bigwedge V=\bigoplus_{r=0}^{n}\bigwedge{}^{r}V$$

为 V 的标准分级外代数，其中 $\bigwedge^r V$ 是 V 的 r 阶外代数。它是由形如 $e_{i_1}\wedge e_{i_2}\wedge\cdots\wedge e_{i_r}$ 的元素所生成的空间。对于一个可逆线性变换 $F\in GL_n(\mathbb{R})$，定义 F 的 E 矩阵为线性变换 F 在标准基 E 下的表示矩阵。在不引起歧义下，F 同时表示可逆线性变换和 F 的 E 矩阵且矩阵 F 的项与列记作

$$F=(f_1|\cdots|f_n)=(f_{ij})_{n\times n}。$$

对于子集 $A \in \binom{[n]}{r}$，

$$f_A := \wedge_{a\in A} f_a \in \bigwedge^r V。$$

因此，

$$f_A \wedge f_B = \begin{cases} \mathbf{0} & A \cap B \neq \varnothing; \\ (-1)^{\rho(A,B)} f_{A\cup B} & A \cap B = \varnothing, \end{cases} \tag{2.3}$$

其中 $\rho(A, B)$ 定义如下：

$$\rho(A, B) = |\{(a, b) \in A \times B : a > b\}|。$$

通过楔乘积的线性性质，集合 $\mathcal{W}(F, n, r) = \left\{ f_A : A \in \binom{[n]}{r} \right\}$ 是向量空间 $\bigwedge^r V$ 的一组基，且 $\dim(\bigwedge^r V) = \binom{n}{r}$。

对于 r 一致超图 $\mathcal{A} \subseteq \binom{[n]}{r}$，

$$W(F, \mathcal{A}) := \operatorname{span}\{f_A : A \in \mathcal{A}\}$$

为 $\bigwedge^r V$ 中的线性子空间。注意到 $\dim(W(F, \mathcal{A})) = |\mathcal{A}|$ 且 f_A 和 $W(F, \mathcal{A})$ 都与 F 的选取相关。另一方面如果对某个超图 $\mathcal{A} \subseteq \binom{[n]}{r}$，存在子空间 $W = W(F, \mathcal{A})$，则称子空间 $W \subseteq \bigwedge^r V$ 为关于 F 的*单空间*。给定非零向量 $w \in \bigwedge^r V$，令 w 在 $\mathcal{W}(F, n, r)$ 的基下的展开式为 $w = \sum_{A\in\binom{[n]}{r}} m_A f_A$，则 w 关于 F 的*初始集*，记作 $\operatorname{ins}_F(w)$，定义如下：

$$\operatorname{ins}(w) = \max\left\{ A \in \binom{[n]}{r} : m_A \neq 0 \right\} \in \binom{[n]}{r},$$

其中最大值取自 $\binom{[n]}{r}$ 上的*反协字典序*：对于集合 $A, B \in \binom{[n]}{r}$，如果 A 和 B 的对称差的最大元落在集合 B 中，则称 $A > B$。例如文献[50] 章节 5 中的例子。注意到 $w \mapsto \operatorname{ins}(w)$ 形成了从 $\bigwedge^r V$ 到 $\binom{[n]}{r}$ 的满射。给定子空间 $W \subseteq \bigwedge^r V$，$W$ 上关于 F 的*初始超图*定义为：

$$\mathcal{H}_F(W) = \{\operatorname{ins}(w) : w \in W, w \neq \mathbf{0}\} \subseteq \binom{[n]}{r}。$$

Scott 和 Wilmer[4] 基于上面超图与子空间的对应证明了如下基本结论。它揭示了 $\mathcal{A} \mapsto W(F, \mathcal{A})$ 是 $[n]$ 上 r 一致超图和 $\bigwedge^r V$ 中关于 F 的单空间之间的双射。

引理 2.9　[4] 令 $V = \mathbb{R}^n$，$F \in GL_n(\mathbb{R})$ 以及 $0 \leqslant r \leqslant n$。则

(i) 对任意的子空间 $W \subseteq \bigwedge^r V$，$\dim(W) = |\mathcal{H}_F(W)|$，

(ii) 对任意 F 上的单空间 W，$W(F, \mathcal{H}_F(W)) = W$，

(iii) 对任意的 $\mathcal{A} \subseteq \binom{[n]}{r}$，$\mathcal{H}_F(W(F, \mathcal{A})) = \mathcal{A}$。

相交超图 $\mathcal{A}$ 是指对任意 $A, B \in \mathcal{A}$ 均有 $A \cap B \neq \varnothing$。当 $\mathcal{A} \subseteq \binom{[n]}{r}$ 及 $r \leqslant n/2$ 的时候，经典的 Erdős-Ko-Rado 定理[51] 得到 $|\mathcal{A}| \leqslant \binom{n-1}{r-1}$，且当 $r < n/2$ 的时候，等号成立当且仅当 $\mathcal{A}$ 是一个 完整 1 星，即包含某个固定元的所有 r 子集。Scott 和 Wilmer 在文献[4] 中介绍了 自消子空间 的定义：如果对任意 $v, w \in W$ 均有 $v \wedge w = \mathbf{0}$，那称子空间 $W \subseteq \bigwedge V$ 是 自消的。他们证明了如下自消空间的结果。

定理 2.10 [4] 令 $V = \mathbb{R}^n$，W 是 $\bigwedge^r V$ 上的自消子空间以及 $r \leqslant n/2$。则 $\mathcal{H}_F(W) \subseteq \binom{[n]}{r}$ 是一个相交超图。因此，

$$\dim(W) = |\mathcal{H}_F(W)| \leqslant \binom{n-1}{r-1}。$$

类似于完整 1 星的向量空间 $\{v \wedge z : z \in \bigwedge^{r-1} V\}$ 显然是 $\bigwedge^r V$ 上的自消子空间且维数为 $\binom{n-1}{r-1}$。然而当 $r < n/2$ 的时候，找出所有达到极值的自消空间仍然是一个公开困难问题。

2.2.2　局部 LYM 不等式

作为极值集合论的基本结果，局部 LYM 不等式与定理 2.3 和反链的 Sperner 定理紧密相连。给定一致超图 $\mathcal{A} \subseteq \binom{[n]}{a}$ 和非负整数 b，其中 $a+b \leqslant n$。定义 $\partial^b \mathcal{A} := \left\{ B \in \binom{[n]}{a+b} : \text{存在某个} A \in \mathcal{A}, \text{使得 } A \subseteq B \right\}$ 为 $\mathcal{A}$ 的 b 上影；对满足条件 $b' \leqslant a$ 的非负整数 b'，定义 $\delta_{b'}(\mathcal{A}) := \left\{ B \in \binom{[n]}{b'} : \text{存在某个} A \in \mathcal{A}, \text{使得 } B \subseteq A \right\}$ 为 $\mathcal{A}$ 的 b 下影。局部 LYM 不等式[50] 可以如下表述：

$$\frac{|\partial^b \mathcal{A}|}{\binom{n}{a+b}} \geqslant \frac{|\mathcal{A}|}{\binom{n}{a}}。$$

Erdős-Ko-Rado 定理和 Kruskal-Katona 定理能直接推导出如下相交集族的局部 LYM 不等式 (也见文献[52] 中的不等式 (3))。

引理 2.11　令 n, a, b 是非负整数，其中 $a + b \leqslant n$ 且 $2a \leqslant n$。假设 $\mathcal{A} \subseteq \binom{[n]}{a}$ 是一个相交超图，则

$$\frac{|\partial^b \mathcal{A}|}{\binom{n-1}{a+b-1}} \geqslant \frac{|\mathcal{A}|}{\binom{n-1}{a-1}}。$$

当 $2a < n$ 的时候，等号成立当且仅当 $\mathcal{A}$ 是完整 1 星或者 $b = 0$。特别地，当 $\mathcal{A}$ 仅由一个大小为 a 的集合组成的时候，等号成立当且仅当 $b = 0$。

下面将详细地证明该引理。

证明　令 $\overline{\mathcal{A}} = \{[n] \setminus A : A \in \mathcal{A}\}$。因此 $\overline{\mathcal{A}} \subseteq \binom{[n]}{n-a}$ 且 $|\overline{\mathcal{A}}| = |\mathcal{A}|$ 同时 $\overline{\partial^b \mathcal{A}} = \delta_{n-a-b}(\overline{\mathcal{A}})$。令 x 是满足关系 $n - a \leqslant x \leqslant n$ 和 $|\overline{\mathcal{A}}| = \binom{x}{n-a}$ 的实数。由于 $\mathcal{A}$

是相交集族，Erdős-Ko-Rado 定理推导出 $|\mathcal{A}| \leqslant \binom{n-1}{a-1}$。这意味着 $x \leqslant n-1$。此外，Kruskal-Katona 定理[53] 推导出

$$|\partial^b \mathcal{A}| = |\delta_{n-a-b}(\overline{\mathcal{A}})| \geqslant \binom{x}{n-a-b},$$

且

$$\frac{|\partial^b \mathcal{A}|}{|\mathcal{A}|} \geqslant \frac{\binom{x}{n-a-b}}{\binom{x}{n-a}} \geqslant \frac{\binom{n-1}{n-a-b}}{\binom{n-1}{n-a}} = \frac{\binom{n-1}{a+b-1}}{\binom{n-1}{a-1}}。$$

上述式子中的第二个不等式是依据 $\frac{\binom{x}{n-a-b}}{\binom{x}{n-a}}$ 是减函数和 $x \leqslant n-1$ 所得。从而等号成立当且仅当 $|\mathcal{A}| = \binom{n-1}{a-1}$ 或者 $b = 0$。这等价于 $\mathcal{A}$ 是完整 1 星或者 $b = 0$。

当 $\mathcal{A}$ 是仅由一个大小为 a 的集合时，$|\partial^b \mathcal{A}| = \binom{n-a}{b}$ 且等号成立当且仅当 $b = 0$。引理得证。 ∎

下面将基于外代数和超图的对应证明实自消子空间的局部 LYM 不等式。

对于两个子空间 $U, W \subseteq \bigwedge V$，定义

$$U \wedge W := \operatorname{span}\{u \wedge w : u \in U, w \in W\}。$$

给定矩阵 $F \in GL_n(\mathbb{R})$ 以及超图 $\mathcal{A} \subseteq \binom{[n]}{r}$ 和矩阵 F 上的单子空间 $W(F, \mathcal{A}) \subseteq \bigwedge^r V$，定义

$$W(F, \mathcal{A}) \wedge \bigwedge{}^c V = \operatorname{span}\left\{ f_A \wedge f_J : A \in \mathcal{A}, J \in \binom{[n]}{c} \right\}。$$

根据等式 (2.3)，$W(F, \mathcal{A}) \wedge \bigwedge^c V = W(F, \partial^c \mathcal{A})$。对于一般的子空间 $W \subseteq \bigwedge^r V$，有如下包含关系：

$$\mathcal{H}_F\left(W \wedge \bigwedge{}^c V\right) \supseteq \left\{ A \cup J : A \in \mathcal{H}_F(W), J \in \binom{[n] \setminus A}{c} \right\} = \partial^c(\mathcal{H}_F(W))。 \tag{2.4}$$

这些结论已经能够充分证明 Scott 和 Wilmer[4] 证明子空间上的局部 LYM 不等式

$$\frac{\dim(W \wedge \bigwedge^c V)}{\binom{n}{r+c}} \geqslant \frac{\dim(W)}{\binom{n}{r}}。$$

现在本章将上述的不等式拓展到自消子空间。

定理 2.12　令 $V = \mathbb{R}^n$ 及 $W \subseteq \bigwedge^r V$ 是自消子空间，其中 $0 < 2r \leqslant n$。则对任意 $0 \leqslant c \leqslant n-r$，有

$$\frac{\dim(W \wedge \bigwedge^c V)}{\binom{n-1}{r+c-1}} \geqslant \frac{\dim(W)}{\binom{n-1}{r-1}}。 \tag{2.5}$$

当 $2r < n$ 的时候，等号成立当且仅当存在可逆矩阵 $F \in GL_n(\mathbb{R})$ 使得 $\mathcal{H}_F(W)$ 是完整 1 星或者 $c = 0$。特别地当 $\dim(W) = 1$ 的时候，等号成立仅当 $c = 0$。

证明　给定 $F \in GL_n(\mathbb{R})$。等式 (2.4) 和引理 2.11 可以得到

$$\frac{\dim(W \wedge \bigwedge^c V)}{\binom{n-1}{r+c-1}} = \frac{|\mathcal{H}_F(W \wedge \bigwedge^c V)|}{\binom{n-1}{r+c-1}} \geqslant \frac{|\partial^c(\mathcal{H}_F(W))|}{\binom{n-1}{r+c-1}} \geqslant \frac{|\mathcal{H}_F(W)|}{\binom{n-1}{r-1}} = \frac{\dim(W)}{\binom{n-1}{r-1}}。$$

注意到式子 (2.5) 中的等号成立当且仅当上式的两个不等号同时取等号。因此引理 2.11 可以推出式子 (2.5) 中的等号成立仅当存在可逆矩阵 $F \in GL_n(\mathbb{R})$ 使得 $\mathcal{H}_F(W)$ 是完整 1 星或者 $c = 0$。当 $\dim(W) = 1$ 的时候，等号成立仅当 $c = 0$。定理得证。 ∎

2.3　主定理的证明

本节将证明定理 2.6、定理 2.7 和定理 2.8。

2.3.1　半捆型 Bollobás 定理

证明定理 2.6 的主要思想来源于文献[4] 中定理 4.5 的证明方法。我们首先通过递归的方式构造出一系列自消子空间 $Z_i \subseteq \bigwedge^{a_i} V$，它们可以把 A_i 之间的相交性质和 A_i 与 B_j 之间的相交性质同时在 Z_i 中体现出来，接着利用自消子空间上的局部 LYM 不等式来控制 Z_i 的维数，最后利用定理 2.10 推导出主定理。

Füredi 一般位置方法[32] 被应用在集合对的定理证明中，这里我们同样使用该方法证明主定理。

引理 2.13　[32] 令 $V = \mathbb{R}^n$，$U_1, \cdots, U_m$ 是 V 的真子空间。则存在 k 维的子空间 V'，使得对任意的 $1 \leqslant i \leqslant m$，均有

$$\dim(U_i \cap V') = \max\{\dim(U_i) + k - n, 0\}。$$

下面将给出定理 2.6的证明过程，为了方便起见，定理 2.6 如下重新叙述。

定理 2.14　令 $\{(A_i, B_i)\}_{i=1}^m$ 是有限维实向量空间的子空间对，其中对任意 $1 \leqslant i \leqslant m$ 均有 $\dim(A_i) = a_i$，$\dim(B_i) = b_i$ 及 $a_i < b_i$。假设存在某个正整数 $t \geqslant 0$，满足

(i) 对任意的 $1 \leqslant i, j \leqslant m$，$\dim(A_i \cap A_j) > t$，

(ii) 对任意的 $1 \leqslant i \leqslant m$，$\dim(A_i \cap B_i) \leqslant t$，

(iii) 对任意的 $1 \leqslant i < j \leqslant m$，$\dim(A_i \cap B_j) > t$。

进一步地，如果 $a_1 \leqslant a_2 \leqslant \cdots \leqslant a_m$，且存在某个正整数 N，使得对任意 $1 \leqslant i \leqslant m$ 均有 $a_i + b_i = N$。则

$$\sum_{i=1}^m \binom{N-(2t+1)}{a_i-(t+1)}^{-1} \leqslant 1。 \tag{2.6}$$

上式等号成立仅当 $a_1 = a_2 = \cdots = a_m$ 且 $b_1 = b_2 = \cdots = b_m$。

证明　令 $V=\mathbb{R}^n$。我们首先证明定理在 $t=0$ 且 $n=N$ 这一特殊情形时成立，而其他一般的情形将被归约到该特殊情形。

现在令 $t=0$ 且对任意 $1\leqslant i\leqslant m$ 均有 $n=N=a_i+b_i>2a_i$。同时选取一个可逆矩阵 $F=(f_{ij})=(f_1|f_2|\cdots|f_n)\in GL_n(\mathbb{R})$ 使得 $\{f_1,f_2,\cdots,f_n\}$ 是 V 的一组基。沿用文献[54] 当中的记号，对于 k 维子空间 $T\subseteq V$，通过选取 T 中的任意一组基 $\{v_1,\cdots,v_k\}$ 来定义 $\wedge T\in\bigwedge^k V$，并令

$$\wedge T=v_1\wedge v_2\wedge\cdots\wedge v_k。$$

除去 $\wedge T$ 的系数之后，它是唯一确定的且 $\mathrm{span}\{\wedge T\}$ 是良好定义的一维子空间。

对任意的 $i\in[m]$，记

$$\tilde{A}_i=\wedge A_i\in\bigwedge^{a_i}V\ \text{ 且 }\tilde{B}_i=\wedge B_i\in\bigwedge^{b_i}V。$$

因为 $n=a_i+b_i=\dim(A_i)+\dim(B_i)$，所以

$$\tilde{A}_i\wedge\tilde{B}_j\begin{cases}\neq\mathbf{0}, & \text{如果 } i=j;\\ =\mathbf{0}, & \text{如果 } i<j。\end{cases}$$

我们将从 $Z_0=\{\mathbf{0}\}$ 出发递归地构造一系列的子空间 $Z_i\subseteq\bigwedge^{a_i}V$。递归方法为：对任意的 $0\leqslant i\leqslant m-1$，其递归关系为

$$Z_{i+1}=\mathrm{span}\left\{Z_i\wedge\bigwedge^{a_{i+1}-a_i}V,\tilde{A}_{i+1}\right\}。$$

因为对任意 $1\leqslant i,j\leqslant m$ 均有 $\dim(A_i\cap A_j)>0$，所以对任意 $0\leqslant i\leqslant m-1$ 均有 Z_{i+1} 是自消子空间。记 Y_i 是由 Z_i 和 $\bigwedge^{a_{i+1}-a_i}V$ 楔乘积而来的，也即

对任意 $0\leqslant i\leqslant m-1$ 均有

$$Y_i=Z_i\wedge\bigwedge^{a_{i+1}-a_i}V\subseteq\bigwedge^{a_{i+1}}V。$$

断言：对任意 $0\leqslant i\leqslant m-1$，均有如下关系

$$\dim(Z_{i+1})=\dim(Y_i)+1,\tag{2.7}$$

及

$$\frac{\dim(Y_i)}{\binom{N-1}{a_{i+1}-1}}\geqslant\frac{\dim(Z_i)}{\binom{N-1}{a_i-1}}。\tag{2.8}$$

根据 Z_{i+1} 的定义知道

$$Z_{i+1}=\mathrm{span}\{\tilde{A}_{i+1},Y_i\}。$$

一方面 $\tilde{\boldsymbol{A}}_{i+1} \wedge \tilde{\boldsymbol{B}}_{i+1} \neq \boldsymbol{0}$。另一方面，由于 $\boldsymbol{Z}_i \wedge \bigwedge^{a_{i+1}-a_i} V = \operatorname{span}\{\tilde{\boldsymbol{A}}_h \wedge \bigwedge^{a_{i+1}-a_h} V \ : \ h \leqslant i\}$ 且对任意 $h \leqslant i$ 均有 $\tilde{\boldsymbol{A}}_h \wedge \tilde{\boldsymbol{B}}_{i+1} = \boldsymbol{0}$，所以对任意 $\boldsymbol{y} \in \boldsymbol{Y}_i = \boldsymbol{Z}_i \wedge \bigwedge^{a_{i+1}-a_i} V$ 均有 $\boldsymbol{y} \wedge \tilde{\boldsymbol{B}}_{i+1} = \boldsymbol{0}$。

因此 $\tilde{\boldsymbol{A}}_{i+1} \notin \boldsymbol{Y}_i$ 且式子 (2.7) 成立。最后根据定理 2.12 得到式子 (2.8) 也成立。并且

$$\frac{\dim(\boldsymbol{Y}_i)}{\binom{N-1}{a_{i+1}-1}} = \frac{\dim(\boldsymbol{Z}_i \wedge \bigwedge^{a_{i+1}-a_i} V)}{\binom{N-1}{a_{i+1}-1}} \geqslant \frac{\dim(\boldsymbol{Z}_i)}{\binom{N-1}{a_i-1}},$$

其中等号成立仅当 $a_{i+1} = a_i$ 或者 $\mathcal{H}_F(\boldsymbol{Z}_i)$ 是完整 1 星。

在自消空间 $\boldsymbol{Z}_m$ 中，定理 2.10 以及断言中的关系式子 (2.7) 和式子 (2.8) 推导出

$$1 \geqslant \frac{\dim(\boldsymbol{Z}_m)}{\binom{N-1}{a_m-1}} = \frac{1+\dim(\boldsymbol{Y}_{m-1})}{\binom{N-1}{a_m-1}} \geqslant \frac{1}{\binom{N-1}{a_m-1}} + \frac{\boldsymbol{Z}_{m-1}}{\binom{N-1}{a_{m-1}-1}} = \cdots \geqslant \sum_{i=1}^{m} \frac{1}{\binom{N-1}{a_i-1}}。 \tag{2.9}$$

这证明式子 (2.6) 在 $t = 0$ 和 $n = N$ 的时候成立。

现在考虑 $\{(\boldsymbol{A}_i, \boldsymbol{B}_i)\}_{1 \leqslant i \leqslant m}$ 在式子 (2.9) 中等号成立时结构是怎样的。在 $a_1 = 1$ 时，$m = 1$。结论自然成立。现在假设 $a_1 > 1$，由于 $\boldsymbol{Z}_1 = \operatorname{span}\{\tilde{\boldsymbol{A}}_1\}$，便有 $\dim(\boldsymbol{Z}_1) = 1$。然后依据定理 2.12 中 $i = 1$ 的情形知道式子 (2.8) 中的等号成立仅当 $a_2 = a_1$。现在假设存在某个 $1 < s \leqslant m-1$ 使得 $a = a_1 = \cdots = a_s < a_{s+1}$，那么 $\mathcal{H}_F(\boldsymbol{Z}_s)$ 是大小为 $\binom{N-1}{a-1}$ 的完整 1 星。一方面，这意味着

$$1 \geqslant \frac{\dim(\boldsymbol{Z}_m)}{\binom{N-1}{a_m-1}} \geqslant \sum_{i=s+1}^{m} \frac{1}{\binom{N-1}{a_i-1}} + \frac{s}{\binom{N-1}{a-1}}。$$

另一方面，对任意 $1 \leqslant i \leqslant s$，$\boldsymbol{Z}_s = \operatorname{span}\{\tilde{\boldsymbol{A}}_1, \cdots, \tilde{\boldsymbol{A}}_s\}$。根据式子 (2.7) 得到 $\dim(\boldsymbol{Z}_s) = \dim(\boldsymbol{Z}_1) + s - 1 = s$ 成立。这意味着 $s = \dim(\boldsymbol{Z}_s) = |\mathcal{H}_F(\boldsymbol{Z}_s)| = \binom{N-1}{a-1}$，于是 $s = m$。因此，式子 (2.9) 中的等号成立仅当 $a_1 = a_2 = \cdots = a_m$ 且 $b_1 = b_2 = \cdots = b_m$。

上面完整地证明了定理在 $t = 0$ 且 $n = N$ 这一特殊情形下成立。

现在假设 $t \geqslant 0$ 且 $n \geqslant N$。令 $n' = n - t$，依据引理 2.13 找到一个 V 中的 n' 维子空间 V'，使得

- 对任意的 $1 \leqslant i \leqslant m$，$\dim(\boldsymbol{A}_i \cap V') = a_i - t$，$\dim(\boldsymbol{B}_i \cap V') = b_i - t$，
- 对任意的 $1 \leqslant i, j \leqslant m$，$\dim(\boldsymbol{A}_i \cap \boldsymbol{A}_j \cap V') > 0$，
- 对任意的 $1 \leqslant i \leqslant m$，$\dim(\boldsymbol{A}_i \cap \boldsymbol{B}_i \cap V') = 0$，
- 对任意的 $1 \leqslant i < j \leqslant m$，$\dim(\boldsymbol{A}_i \cap \boldsymbol{B}_j \cap V') > 0$。

记 $a'_i = a_i - t$，$b'_i = b_i - t$ 和 $n'' = n' - a'_i - b'_i = n' - (N - 2t)$。我们再一次地运用引理 2.13 找到一个 V' 中的 n'' 维子空间 V''，使得 $\dim((\boldsymbol{A}_i \cap V') \cap V'') = 0$，且对任意 $i \in [m]$ 均有 $\dim((\boldsymbol{B}_i \cap V') \cap V'') = 0$。令 $\boldsymbol{Q}$ 是 V'' 在空间 V' 中的正交补

空间，也即 $V' = V'' \oplus Q$。定义线性映射 $\phi : V' \to Q$ 如下

$$\text{对任意的 } (a,b) \in V'' \times Q,\ \phi(a+b) = b,$$

上面的符号 $\times$ 代表笛卡尔积。这里笛卡尔积 $V'' \times Q$ 与直和 $V'' \oplus Q$ 使用不同的符号。因为对任意 $1 \leqslant i \leqslant m$ 均有 $A_i' = \phi(A_i \cap V')$, $B_i' = \phi(A_i \cap V')$，所以下面的关系成立

- 对任意的 $1 \leqslant i \leqslant m$，$\dim(A_i') = a_i'$，$\dim(B_i') = b_i'$，
- 对任意的 $1 \leqslant i, j \leqslant m$，$\dim(A_i' \cap A_j') > 0$，
- 对任意的 $1 \leqslant i \leqslant m$，$\dim(A_i' \cap B_i') = 0$，
- 对任意的 $1 \leqslant i < j \leqslant m$，$\dim(A_i' \cap B_j') > 0$。

从而空间 Q 上有一系列的新子空间对 $\{(A_i', B_i')\}_{i=1}^m$ 且它们满足定理的所有条件，其中 $\dim(Q) = \dim(A_i') + \dim(B_i') = N - 2t$。依据前面关于 $t = 0$ 且 $n = N - 2t = a_i' + b_i'$ 的特殊情形证明知道

$$\sum_{i=1}^{m} \frac{1}{\binom{(N-2t)-1}{a_i'-1}} = \sum_{i=1}^{m} \frac{1}{\binom{N-(2t+1)}{a_i-(t+1)}} \leqslant 1。$$

上面等号成立仅当 $a_1 = \cdots = a_m$ 且 $b_1 = \cdots = b_m$。定理得证。 ∎

注意到在不考虑等号成立的情形下，定理 2.6 的条件 $a_i < b_i$ 可以放松到 $a_i \leqslant b_i$。最后文献[32] 的方法给出定理 2.7 的证明。

定理 2.7的证明　令 $X = \bigcup_{i=1}^m (A_i \cup B_i)$ 是大小为 n 的背景集，且对任意的 $x \in X$ 对应唯一的向量 $v(x) \in \mathbb{R}^n$，同时 $\{v(x) : x \in X\}$ 形成 $\mathbb{R}^n$ 的一组基。

现在对 $1 \leqslant i \leqslant m$，定义 $\bar{A}_i$ 为空间 $\mathbb{R}^n$ 中由 $\{v(x) : x \in A_i\}$ 所生成的 a_i 维空间，$\bar{B}_i$ 为 $\mathbb{R}^n$ 中由 $\{v(x) : x \in B_i\}$ 所生成的 b_i 维空间。因此 $\{A_i\}_{i=1}^m$ 之间的相交条件和 $(A_i, B_j)_{1 \leqslant i \leqslant j \leqslant m}$ 之间的相交条件意味着 $\{(\bar{A}_i, \bar{B}_i)\}_{i=1}^m$ 是满足定理 2.6 的一组子空间。从而

$$\sum_{i=1}^{m} \frac{1}{\binom{N-(2t+1)}{a_i-(t+1)}} \leqslant 1。$$

∎

2.3.2　稳定性结果

这一小节将利用外代数和超图之间的关系和著名 Erdős-Ko-Rado 定理[51] 中的稳定性结果来证明定理 2.8。

定理 2.8的证明　令 $X=\bigcup_{i=1}^{m}(A_i\cup B_i)$ 是大小为 n 的背景集。对任意的 $x\in X$ 对应唯一的向量 $v(x)\in V=\mathbb{R}^{a+b}$，且 $\{v(x):x\in X\}$ 落在一般位置，也即，其中任何 $a+b$ 个向量都是线性独立的。类似地对任意的 $1\leqslant i\leqslant m$，定义 $\bar{A}_i$ 和 $\bar{B}_i$ 是分别由 $\{v(x):x\in A_i\}$ 和 $\{v(x):x\in B_i\}$ 所生成的子空间。

令 $\{e_1,e_2,\cdots,e_{a+b}\}$ 是 $\mathbb{R}^{a+b}$ 的一组标准基。不失一般性，我们假设 $X=[n]$，$A_1=\{b+1,b+2,\cdots,a+b\}$，$B_1=\{1,2,\cdots,b\}$ 以及上述分配满足对任意的 $x\in[a+b]$，$v(x)=e_x$。因此有 $\bar{A}_1=\mathrm{span}\{e_{b+1},e_{b+2},\cdots,e_{a+b}\}$ 且 $\bar{B}_1=\mathrm{span}\{e_1,e_2,\cdots,e_b\}$。

因为 $\{A_i\}_{i=1}^{m}$ 是相交集族，所以

$$W=\mathrm{span}\{\wedge\bar{A}_1,\wedge\bar{A}_2,\cdots,\wedge\bar{A}_m\}$$

是 $\bigwedge^a V$ 中的自消子空间，从而

$$(\wedge\bar{A}_i)\wedge(\wedge\bar{B}_j)\begin{cases}\neq\mathbf{0},\ \text{如果 } i=j;\\ =\mathbf{0},\ \text{如果 } i<j。\end{cases}$$

根据第 3 版的三角准则 (见文献[54] 中的命题 2.9) 推导出 $\wedge\bar{A}_1,\wedge\bar{A}_2,\cdots,\wedge\bar{A}_m$ 在 $\bigwedge^a V$ 中线性独立。因此 $\dim(W)=m$。令 $F=(e_1|e_2|\cdots|e_{a+b})\in GL_{a+b}(\mathbb{R})$。则根据引理 2.9得到

$$|\mathcal{H}_F(W)|=\dim(W)=\binom{a+b-1}{a-1}。$$

根据定理 2.10，$\mathcal{H}_F(W)\subseteq\binom{[a+b]}{a}$ 是相交集族。因为 $a<b$，所以利用 Erdős-Ko-Rado 定理的极值情形得到 $\mathcal{H}_F(W)$ 必须是完整 1 星。

上面的假设保证了 $f_{A_1}=\wedge_{i=b+1}^{a+b}e_i=\wedge\bar{A}_1\in W$ 且 $\mathrm{ins}(f_{A_1})=A_1\in\mathcal{H}_F(W)$。因为 $\mathcal{H}_F(W)$ 是 $[a+b]$ 中的完整 1 星，所以存在 $x\in[a]$ 使得对任意 $w\in W$ 均有 $b+x\in\mathrm{ins}(w)$。

断言：对任意的 $2\leqslant i\leqslant m$，均有 $b+x\in A_i$。

假设断言不正确，那么存在 $2\leqslant i_0\leqslant m$ 使得 $b+x\notin A_{i_0}$。假设 $A_1\cap A_{i_0}=\{b+i_1,\cdots,b+i_k\}$，其中 $\varnothing\neq\{i_1,\cdots,i_k\}\subsetneq[a]$。那么 $x\notin\{i_1,\cdots,i_k\}$。根据 $\bar{A}_{i_0}$ 的定义，我们假设 $\bar{A}_{i_0}=\mathrm{span}\{e_{b+i_1},\cdots,e_{b+i_k},v_{k+1}\cdots,v_a\}$。其中对任意的 $j=k+1,\cdots,a$，向量 $v_j\in\mathbb{R}^{a+b}\setminus\{e_{b+1},e_{b+2},\cdots,e_{a+b}\}$。从而对于 $\wedge\bar{A}_{i_0}\in W$，它的展开为

$$\wedge\bar{A}_{i_0}=(\wedge_{j=1}^{k}e_{b+i_j})\wedge(\wedge_{j=k+1}^{a}v_j)=\sum_{C\in\binom{[a+b]}{a}}m_C f_C。\tag{2.10}$$

因此根据 f_C 的定义，式子 (2.10) 中的每个 C 都满足 $m_C\neq0$ 且 $\{b+i_1,\cdots,b+i_k\}\subseteq C$。另外由于 $b+x\in\mathrm{ins}(\wedge\bar{A}_{i_0})$ 和反协字典序的定义，式子 (2.10) 中满足 $m_C\neq0$ 的 C 同样满足关系

$$C\cap(\{b+x,b+x+1,\cdots,a+b\}\setminus\{b+i_1,\cdots,b+i_k\})\neq\varnothing。$$

否则令 $D(x)=\{b+x,b+x+1,\cdots,a+b\}\setminus\{b+i_1,\cdots,b+i_k\}$ 并假设存在子集 $C_0\in\binom{[a+b]}{a}$ 使得 $C_0\cap D(x)=\varnothing$。则对任意的 $C\cap D(x)\neq\varnothing$，在反协字典序下 $C_0>C$。这与猜测 $b+x\in\mathrm{ins}(\wedge\bar{A}_{i_0})$ 相矛盾。因此

$$(\wedge_{i\in[a]\setminus\{i_1,\cdots,i_k\}}e_{b+i})\wedge(\wedge\bar{A}_{i_0})=(\wedge_{i\in[a]\setminus\{i_1,\cdots,i_k\}}e_{b+i})\wedge(\sum_{C\in\binom{[a+b]}{a}}m_Cf_C)=\mathbf{0}。$$

这表明这 $2a-k$ 个向量 $e_{b+1},\cdots,e_{a+b},v_{k+1},\cdots,v_a$ 在 $\mathbb{R}^{a+b}$ 中是线性相关的，这与 $\{v(x):x\in X\}$ 中任何 $a+b$ 个向量线性独立矛盾。因此对每个 $2\leqslant i\leqslant m$，$b+x\in A_i$。

最后让我们考虑集合对 $\{(A_i\setminus\{b+x\},B_i)\}_{i=1}^m$。因为 $A_i\cap B_i=\varnothing$，所以对任意的 $i\in[m]$，$b+x\notin B_i$。从而 $\{(A_i\setminus\{b+x\},B_i)\}_{i=1}^m$ 继承了 $\{(A_i,B_i)\}_{i=1}^m$ 交叉相交的性质。根据定理 2.3 推出 $m=\binom{a+b-1}{a-1}$ 当且仅当存在大小为 $a+b$ 背景集 X，使得对任意的 i，A_i 是集合 X 的 a 元子集且 $B_i=X\setminus A_i$。∎

2.4　小结

本章利用外代数的方法证明了在有限维实向量空间上带权重的半捆型 Bollobás 定理。该定理的一个应用是解决了 Gerbner 等人在文献[1] 的猜想。进一步地，本章改进了房屋分配中的可达集数目的上界。此外在极值集合论的理论层面上，我们确定了有限集合对上的半捆型 Bollobás 定理中唯一的极值结构。

Gerbner 等人[1] 进一步地研究中提出了本章开头的猜想 2.1。

猜想 2.1 与定理 2.7 当中 t 交叉相交的限制不同，它只要求 $\{A_i\}_{i=1}^m$ 和 $\{B_j\}_{j=1}^m$ 是交叉相交的。因此定理 2.7 和猜想 2.1 之间还是存在差距。特别地，Scott 和 Wilmer 在参考文献[4] 的第二个版本中利用内乘积方法给出了猜想 2.1 的证明。

定理 2.6 要求 $a_i+b_i=N$ 是固定的。但是在定理 2.5 中是没有这样的要求限制的。然而根据 Adam Wagner 提供的反例说明这样的限制是不可以移除的：

例 2.1　令

$$A_1=\{0,3\},\ A_2=\{3,5\},\ A_3=\{3,4\}\text{ 且 }A_4=\{0,4,5\};$$
$$B_1=\{4,5\},\ B_2=\{0,4\},\ B_3=\{0,5\}\text{ 且 }B_4=\{1,2,3\}。$$

则对任意 $1\leqslant i\leqslant 3$ 均有 $a_i=b_i=2$ 及 $a_4=b_4=3$。这意味着 $\sum_{i=1}^4\binom{a_i+b_i-1}{a_i-1}^{-1}=1.1>1$。

稳定性是极值集合问题的自然研究方向，从稳定性角度出发得到的定理同样值得进一步地深刻研究。因为定理 2.7 分别证明了子空间对和集合对的上界，那么下面的平凡结构是否是唯一达到本章提供的上界的结构?

$\{A_i\}_{i=1}^m$ 是 $\mathbb{R}^{a+b}$ 中的一个 t 星 且 $B_i=A_i^{\perp}\in\mathbb{R}^{a+b}$。

如果结构不唯一，那么其他所有达到极值的结构又是如何的呢? 这个问题或许不是一个简单的问题。根据我们的了解，定理 2.4 中的唯一结构现在仍然是未解决的。因此我们或许需要通过其他的方法来解决这种类型问题。

第 3 章　部分重复码上的访问均衡问题

本章的目标是构造出达到访问均衡的分布式存储系统并以此来提高整个存储系统的稳定性。为此，本章在部分重复码的基础上提出了一类新型的组合模型，称之为 MinVar 模型。由于部分重复码通常是用图或者集合系的语言描绘，因此 MinVar 模型将被描绘成图或者集合系上的优化问题：寻找集合系上的区组标号，使得所有顶点标号的方差最小化，其中顶点标号为所有与其关联的区组标号之和。该模型与 Dau 和 Milenkovic 的 MaxMinSum 模型[14] 有着显著的差异并且具有独特的现实意义。本章解决了许多与 MinVar 问题对偶等价的图顶点标号问题 (MinPS 问题)，其中包含了多个完全图的并、多个 Turán 图的并以及圈上的顶点标号，进而获得了一系列达到访问均衡的部分重复码。

3.1　介绍

Dau 和 Milenkovic[14] 从分布式存储系统中的访问均衡问题出发[8] 提出了一类基于组合设计的点标号问题。在该问题框架中，文件首先被分割成若干个大小相同的子文件并通过外层 MDS 码的方式把它们编码成一些数据块。然后每个数据块被复制相同次数，再按照内层部分重复 (FR) 码的方式分布在多个存储节点当中[5-7]。这样的一个外层 MDS 码与内层部分重复码的嵌套方案保障了存储系统的低冗余性和可修复性。在传输过程中，具有恰当修复特性的最小带宽生成 (MBR) 码[8] 就属于该分布式存储系统。在研究该套系统的过程中，数据块的复制次数以及内层部分重复码中节点的数目通常是固定的。因此大量的组合设计构型被使用到该系统中，例如利用斯坦纳三元系来排布数据[9-11]。熟知的 Hadoop 分布式文件系统和 Google 文件系统都运用了这样的策略[12]。

通常来说，访问均衡问题旨在利用数据块的热度信息来均衡每个存储节点的访问请求[13]。在 Dau 和 Milenkovic 的模型中[14]，他们用数据块的访问热度对其标号并要求每个存储节点内的数据块标号和尽可能地均衡。这等价于在集合系中寻找一个合适的点标号，使得集合系中的每个区组内的标号和尽可能地相等。特别地，他们定义了一些标准来衡量均衡性。例如 MaxMinSum (或者 MinMaxSum) 标准：要求区组内部标号之和的最小值最大化 (或者区组内部标号之和的最大值最小化)，并成功地发现所有达到 MaxMinSum 标准的斯坦纳三元系以及对偶斯坦纳三元系。进一步地，文献[15] 和文献[16] 分别研究达到 MaxMinSum 标准的柯克曼系和部分斯坦纳系。

尽管组合设计构型广泛地应用在存储方案的底层构架中，但是通常来说它

们不是最好的部分重复码，针对 DRESS 码[7,55] 中所能存储的最大文件数目，内层部分重复码是决定最大存储文件数目的唯一因素。因此达到最大存储文件数目的部分重复码被称为最优的。Silberstein 和 Etzion[17] 研究了基于图和设计上的最优部分重复码。他们构造了两类重复次数是 2 的最优部分重复码，其中一类部分重复码基于 Turán 图，另一类基于大围长的图。对于重复次数更大的情形，他们构造出由横截设计和广义多边形生成的最优部分重复码。

本章主要聚焦于最优重复码的访问均衡问题。注意到，MaxMinSum 标准模型只关注到了所有存储节点内部标号和的最小值，对于均衡性的刻画过于局部化。因此我们引入一个考虑所有存储节点内部标号和方差最小的模型。该新模型旨在寻找所有存储节点热度和方差最小的方案，称之为 MinVar 模型。而该模型恰好是一个集合系的区组标号问题。于是，寻找到一个特殊的区组标号，使得集合系中所有顶点标号方差最小，显得尤为重要。这里的顶点标号为所有与该顶点关联的区组标号之和。特别地，当最小方差等于零的时候，该问题恰好等价于图幻标问题。由此本章发现了诸多访问请求达到完美平衡的最优部分重复码。第二个贡献在于对线性集合系的 MinVar 问题给出了对偶等价的 MinPS 问题，该等价问题是一个图顶点标号问题。本章通过极值组合的方法解决了特定图的顶点标号问题，从而获得一些达到 MinVar 标准的部分重复码。

3.2　准备工作

这一小节将给出一些在本章非常有用的编码、图和集合系的概念以及它们之间的关系。

3.2.1　部分重复码

El Rouayheb 和 Ramchandran[7] 提出了 DRESS (基于简单存储的分布式重复) 码的概念，它是一个外层 MDS 码和一个内层部分重复码的嵌套组合。

令 $[\theta] := \{1, 2, \cdots, \theta\}$，$n, \alpha, \theta, \rho$ 是满足等式 $n\alpha = \theta\rho$ 的正整数。则 (n, α, ρ) 部分重复码 $\mathcal{C}$ 是 $[\theta]$ 中的 n 个子集 $N_1, N_2, \cdots, N_n$，其中每个子集的大小为 α 且 $[\theta]$ 中的每个符号都恰好出现在 $\mathcal{C}$ 中 ρ 个子集中。

$[(\theta, M), k, (n, \alpha, \rho)]$ DRESS 码是由外层 (θ, M) MDS 码和内层 (n, α, ρ) 部分重复码 $\mathcal{C}$ 嵌套组成。首先文件 $\mathbf{f} = (x_1, x_2, \cdots, x_M) \in \mathbb{F}_q^M$ 通过外层 MDS 码被编码变成码字 $y_{\mathbf{f}} = (y_1, y_2, \cdots, y_\theta)$。然后将 $y_{\mathbf{f}}$ 中每个符号通过内层部分重复码 $\mathcal{C}$ 的方式排布在 n 存储节点中: 如果在 $\mathcal{C}$ 中，$i \in N_j$，那么将符号 y_i 排布在第 j 个节点中。根据部分重复码的定义可以发现每个节点恰好存储 α 个符号同时每个符号恰好排布在 ρ 个节点中。

有效的DRESS码需要包含下面两个基本性质:(1) 当某个节点 j 损坏后,系统可以通过寻找其他 $d=\alpha$ 个节点使得每个节点恰好传输一个符号来修复出节点 j。这个修复带宽 d 与 MBR 码的修复带宽是一样的。(2) 所存储的文件可以通过调取任意 k 个节点来重建。根据 MDS 的性质知道参数满足 $\min_{|I|=k}|\cup_{i\in I}N_i|\geqslant M$。因此有效的 DRESS 码总是假设 $M=M(k)=\min_{|I|=k}|\cup_{i\in I}N_i|$。

在能够保障重建和修复的功能后,存储文件的数目最大化成为大家的关注重点。我们根据文献[5] 的研究总是要求对所有的 $i\neq j$, $|N_i\cap N_j|\leqslant 1$。令 $A(n,k,\alpha,\rho)$ 为所有 $[(\theta,M),k,(n,\alpha,\rho)]$ DRESS 码中最大的 $M(k)$，该参数的值是由内层部分重复码唯一决定的。文献[7] 中给出了 $A(n,k,\alpha,\rho)$ 的两个上界。

$$A(n,k,\alpha,\rho)\leqslant\left\lfloor\frac{n\alpha}{\rho}\left(1-\frac{\binom{n-\rho}{k}}{\binom{n}{k}}\right)\right\rfloor, \tag{3.1}$$

$$\begin{aligned}&A(n,k,\alpha,\rho)\leqslant\varphi(k),\ \text{其中}\\&\varphi(1)=\alpha,\varphi(k+1)=\varphi(k)+\alpha-\left\lceil\frac{\rho\varphi(k)-k\alpha}{n-k}\right\rceil。\end{aligned} \tag{3.2}$$

假设对任何给定的 k，部分重复码满足 $\min_{|I|=k}|\cup_{i\in I}N_i|=A(n,k,\alpha,\rho)$，则称其为 k-*最优的*。假设对所有的 $k\leqslant\alpha$，分布式重复码都是 k-最优的，则称其为*最优的*。

Silberstein 和 Etzion[17] 在 $\rho=2$ 的时候构建了两类最优部分重复码，其中一个是基于 Turán 图，另一个是基于大围长的图。当 $\rho>2$ 的时候，他们构造出由横截设计和广义多边形生成的最优部分重复码。

3.2.2　集合系和图

对于有限点集 V，令 $\binom{V}{r}$ 表示 V 中所有的 r 元子集。如果二元组 $S=(V,E)$ 满足 $E\subseteq\binom{V}{r}$,则称为 r *一致集合系*,其中 E 中的元素称为 *区组*。S 的 *阶*为点集的数目 $|V|$, S 的 *大小*为区组的数目 $|E|$。当 $r=2$ 的时候,这样的二元组 (V,E) 就是广为熟知的图,其中 V 和 E 分别被称为顶点集和边集。当 $r>2$ 的时候, S 为通常熟知的 r *一致超图* (或者 r 图),其中 E 被称为超边集。如果集合系中任意两个不同的区组相交至多一个点,称为是 *线性的*。$S=(V,E)$ 的 2 *影子*记作 $\partial_2S=(V',E')$ 是一个 2 一致集合系，其中 $V'=V,E'=\{\{a,b\}:\{a,b\}\subseteq B,B\in E\}$。显然地，如果 S 是 2 一致集合系，则 $\partial_2S=S$。

对任意两个点 $x,y\in V$，如果存在一个区组 $e\in E$ 使得 $\{x,y\}\subseteq e$，则称 x,y 在 S 中是 *相邻的*。一个点 x 与一个区组 $e\in E$ 如果满足 $x\in e$，则称它们是 *关联的*。点 x 的 *度* 指的是与 x 相关联的区组数目，记作 $d(x)$。对一个给定的正整数 d，如果集合系满足对任意的 $x\in V$，都有 $d(x)=d$，则称集合系是 d *正则的*。

集合系 $S=(V,E)$ 的 关联矩阵 $I(S)$ 是一个 $|V|\times|E|$ 的二元矩阵，其中行和列分别用 V 和 E 表示，使得 $I(S)_{i,e}=1$ 当且仅当 $i\in e$。集合系 $S=(V,E)$ 的 线图 $L(S)$ 是一个多重图 (V',E')，其中 $V'=E$ 且任意两个不同的区组 $e,e'\in E$ 之间由 $|e\cap e'|$ 条边连接。若 S 是线性集合系，那它的线图 $L(S)$ 就是一个简单图。S 的 对偶集合系 S^* 为满足条件 $I(S^*)=I(S)^T$ 的集合系。

现在我们列举一些图论中常用的图。当 $|V|=n$ 时，图 $\left(V,\binom{V}{2}\right)$ 称为 完全图，记作 K_n。一个 r 部图是指它的顶点可以划分成 r 个部分且任意两个顶点只有它们分属于不同部分的时候才相连。进一步地，如果任意两个属于不同部的顶点都相连，则称该图为 完全 r 部图。如果在完全 r 部图中，对任意的 $i\in[r]$，第 i 部的顶点数为 m_i，则记作 $K_{m_1,m_2,\cdots,m_r}$。*Turán* 图 $T(n,r)$ 是一个 n 个顶点的完全 r 部图，其中每部的顶点数是 $\lceil\frac{n}{r}\rceil$ 或者 $\lfloor\frac{n}{r}\rfloor$。

图 $G=(V,E)$ 的 邻接矩阵 $A(G)$ 是一个 $|V|\times|V|$ 的矩阵，它的行和列都分别用 V 来表示，其中当 $\{i,j\}\in E$ 的时候 $A(G)_{i,j}=1$ 其余的时候为 0。顶点 x 的 邻居 是由所有与 x 相邻的顶点组成的，记作 $N(x)$。圈 G 是一个连通的 2 正则图，记作 $C_n=(v_1,v_2,\cdots,v_n)$，其中边集为 $v_i\sim v_{i+1},i\in[n-1]$ 和 $v_n\sim v_1$。图的 围长 为图中最短圈的长度。$G=(V,E)$ 的 独立集 是一组两两不相邻的顶点集。图 G 的 完美匹配 是覆盖所有顶点的互不相邻的边。如果图 G 的边集可以划分成一些完美匹配，则称 G 是 可 1-因子分解的。

3.2.3　基于集合系的部分重复码

给定一个 (n,α,ρ) 部分重复码 C，它的 关联矩阵 $I(C)$ 是一个满足 $n\alpha=\theta\rho$ 的 $n\times\theta$ 二元矩阵，其中矩阵的行由码的节点表示，矩阵的列由外层 MDS 码所生成的码字符号表示，矩阵的项 $I(C)_{i,j}$ 按照如下定义：

$$I(C)_{i,j}=\begin{cases}1 & \text{如果节点 } i \text{ 包含符号 } j,\\ 0 & \text{否则。}\end{cases}$$

注意到矩阵 $I(C)$ 的每行恰有 α 个 '1'，每列恰有 ρ 个 '1'。显然 n 阶 ρ 一致 α 正则的集合系 S 可以生成满足关系 $I(C)=I(S)$ 的 (n,α,ρ) 部分重复码 C。由于矩阵 $I(S)$ 的转置可以视为 S 的对偶集合系的关联矩阵，那么它是一个 $n\alpha/\rho$ 阶 α 一致 ρ 正则的集合系。由此生成了满足关系 $I(C')=I(S)^T$ 的 $(n\alpha/\rho,\rho,\alpha)$ 部分重复码 C'。

3.3　部分重复码中新的访问均衡模型

对任意正则一致集合系，它都能按照小节 3.2.3 的方式生成部分重复码。外层 MDS 码编码来的码字 $(y_1,y_2,\cdots,y_\theta)$ 中的符号将被视为区组的标号，那么标记

为 x 的存储节点拥有所有包含点 x 的区组标号，其中 $y_1, y_2, \cdots, y_\theta$ 的标号直接取自它们的热度 (也就是它们的访问频率)。实验表明数据块的访问频率服从 Zipf 法则[56]：对于某个常数 $\beta > 0$，数据块中热度第 i 高的访问频率为 $1/i^\beta$。为简化模型，我们假设 $y_1, y_2, \cdots, y_\theta$ 的标号为连续的整数 $[\theta]$。同时本章的小结部分将讨论在 Zipf 法则下的访问均衡问题。一个节点的总热度为该节点所存储的数据块标号总和。文献[14] 的作者为满足每个节点的访问请求尽可能的均衡提出了数据块的排布策略，称为 MaxMinSum 方案。MaxMinSum 问题本质上是一个对偶设计的点标号问题，它旨在找到所有点标号中区组内部点标号和的最小值最大化。他们同时解决在斯坦纳三元系和对偶斯坦纳三元系上的 MaxMinSum 问题[14]。在此我们重述一下部分重复码上的 MaxMinSum 问题，也即，一个区组标号问题。

问题 3.1　[14] 给定一个 n 阶 ρ 一致 α 正则的集合系 $S = (V, E)$，其中 $V = \{v_1, v_2, \cdots, v_n\}$，$E = \{e_1, e_2, \cdots, e_\theta\}$ 并且参数满足关系 $\theta = n\alpha/\rho$。那么构造一个由 S 生成并达到 *MaxMinSum* 的 (n, α, ρ) 部分重复码，这样的问题等价于寻找一个集合系 E 上的区组标号，也即，从集合 E 到集合 $[\theta]$ 的双射 σ，使得最小访问和

$$MinSum(S_\sigma) := \|I(S)(\sigma(e_1), \sigma(e_2), \cdots, \sigma(e_\theta))^T\|_{\mathbb{L}^{min}}$$

最大化。其中对任意向量 $\mathbf{x} = (x_1, x_2, \cdots, x_n)^T$，符号 $\|\mathbf{x}\|_{\mathbb{L}^{min}} := \min\{x_i\}$。

文献[14] 的作者同时提出了 MinMaxSum 模型，它追寻 *最大访问和*

$$MaxSum(S_\sigma) := \|I(S)(\sigma(e_1), \sigma(e_2), \cdots, \sigma(e_\theta))^T\|_{\mathbb{L}^{max}},$$

最小化。其中对任意向量 $\mathbf{x} = (x_1, x_2, \cdots, x_n)^T$，符号 $\|\mathbf{x}\|_{\mathbb{L}^{max}} := \max\{x_i\}$。然而部分重复码通常是不能够在这两个模型下同时达到最优。比如下面的例子。

例 3.1　令 C 是基于完全图 K_4 的部分重复码，其中包含 $\{1, 2, 3, 4\}$ 四个节点和六个符号。考虑如下 C 上的两个标号 σ_1 和 σ_2：

σ_1：

$$\begin{aligned}
&\sigma_1(12) = 3, \sigma_1(13) = 1, \sigma_1(14) = 6,\\
&\sigma_1(23) = 5, \sigma_1(24) = 2, \sigma_1(34) = 4,\\
&MinSum(S_{\sigma_1}) = \|(10, 10, 10, 12)^T\|_{\mathbb{L}^{min}} = 10,\\
&MaxSum(S_{\sigma_1}) = \|(10, 10, 10, 12)^T\|_{\mathbb{L}^{max}} = 12。
\end{aligned}$$

σ_2：

$$\begin{aligned}
&\sigma_2(12) = 3, \sigma_2(13) = 1, \sigma_2(14) = 5,\\
&\sigma_2(23) = 6, \sigma_2(24) = 2, \sigma_2(34) = 4,\\
&MinSum(S_{\sigma_2}) = \|(9, 11, 11, 11)^T\|_{\mathbb{L}^{min}} = 9,\\
&MaxSum(S_{\sigma_2}) = \|(9, 11, 11, 11)^T\|_{\mathbb{L}^{max}} = 11。
\end{aligned}$$

简单地计算可以知道标号 σ_1 的最小访问和达到最大，而标号 σ_2 最大访问和达到最小并且标号 σ_2 在 MaxMinSum 模型下不是最优标号。现在我们观察所有节点热度的方差，也就是数值 $\sum_{i=1}^{n}(p_i-\bar{p})^2$，其中 p_i 是第 i 个节点的热度，$\bar{p}$ 是所有节点的热度平均值。易知 $I(S)(\sigma(e_1),\sigma(e_2),\cdots,\sigma(e_\theta))^T=(p_1,p_2,\cdots,p_n)^T$，则上面例子中的标号 σ_1 和标号 σ_2 具有相同的热度方差。那么 MaxMinSum 模型不能获取所有访问均衡的标号。

即使两个标号在 MaxMinSum 模型中表现相同，但是他们在 MinVar 模型下仍有可能表现不一样。比如下面例子：

例 3.2　我们考虑一个基于完全图 K_8 的部分重复码。通过计算机的搜索找到两个在 MaxMinSum 模型下最优的标号 σ_1 和 σ_2，但是在考虑热度方差的时候标号 σ_2 是比标号 σ_1 更加优异的。

σ_1：

$\sigma_1(12)=1,\sigma_1(13)=2,\sigma_1(14)=3,\sigma_1(15)=14,$
$\sigma_1(16)=26,\sigma_1(17)=27,\sigma_1(18)=28,\sigma_1(23)=4,$
$\sigma_1(24)=10,\sigma_1(25)=17,\sigma_1(26)=21,\sigma_1(27)=23,$
$\sigma_1(28)=25,\sigma_1(34)=24,\sigma_1(35)=22,\sigma_1(36)=15,$
$\sigma_1(37)=16,\sigma_1(38)=18,\sigma_1(45)=20,\sigma_1(46)=19,$
$\sigma_1(47)=12,\sigma_1(48)=13,\sigma_1(56)=8,\sigma_1(57)=11,$
$\sigma_1(58)=9,\sigma_1(67)=7,\sigma_1(68)=5,\sigma_1(78)=6,$

$MinSum(S_{\sigma_1})=\|(101,101,101,101,101,101,102,104)^T\|_{\mathbb{L}^{min}}=101$。

σ_2：

$\sigma_2(12)=1,\sigma_2(13)=2,\sigma_2(14)=3,\sigma_2(15)=14,$
$\sigma_2(16)=26,\sigma_2(17)=27,\sigma_2(18)=28,\sigma_2(23)=4,$
$\sigma_2(24)=10,\sigma_2(25)=17,\sigma_2(26)=21,\sigma_2(27)=23,$
$\sigma_2(28)=25,\sigma_2(34)=24,\sigma_2(35)=22,\sigma_2(36)=15,$
$\sigma_2(37)=16,\sigma_2(38)=18,\sigma_2(45)=20,\sigma_2(46)=19,$
$\sigma_2(47)=12,\sigma_2(48)=13,\sigma_2(56)=9,\sigma_2(57)=11,$
$\sigma_2(58)=8,\sigma_2(67)=7,\sigma_2(68)=5,\sigma_2(78)=6,$

$MinSum(S_{\sigma_2})=\|(101,101,101,101,101,102,102,103)^T\|_{\mathbb{L}^{min}}=101$。

在考虑访问均衡问题时候，节点的热度方差绝对是一个重要的指标，因此下一节将介绍这个新模型。

3.3.1　最小方差模型

本小节提出了一个新的数据块排布策略，称之为 *最小方差 (MinVar)* 排布方案。在重新对数据块标号之后，最小方差排布方案旨在最小化部分重复码中节点

热度的方差。该问题用如下公式陈述。

问题 3.2　给定一个 n 阶 ρ 一致 α 正则的集合系 $S=(V,E)$，其中 $V=\{v_1,v_2,\cdots,v_n\}$，$E=\{e_1,e_2,\cdots,e_\theta\}$ 并且参数满足关系 $\theta=n\alpha/\rho$。那么构造一个由 S 生成并且达到 *MinVar* 的 (n,α,ρ) 部分重复码，这样的问题等价于寻找一个 E 上的区组标号，也即，从集合 E 到集合 $[\theta]$ 的双射 σ，使得访问方差

$$Var(S_\sigma) := \|I(S)(\sigma(e_1),\cdots,\sigma(e_\theta))^T-(\bar{a},\cdots,\bar{a})^T\|_{\mathbb{L}^2}$$

最小化。其中 $\bar{a}$ 等于平均热度 $\frac{\rho\theta(\theta+1)}{2n}=\frac{\alpha(\theta+1)}{2}$ 且对任意向量 $\mathbf{x}=(x_1,x_2,\cdots,x_n)^T$，符号 $\|\mathbf{x}\|_{\mathbb{L}^2}:=\sum_{i=1}^n x_i^2$。

令 $p_\sigma=(p_1,\cdots,p_n):=I(S)(\sigma(e_1),\cdots,\sigma(e_\theta))^T$。由 $I(S)$ 的定义，p_i 是节点 i 的热度且等于 $\sum_{j:v_i\in e_j}\sigma(e_j)$。$Var(S_\sigma)/n$ 是 n 个节点热度分布的方差，它度量了每个节点的热度与平均热度的差异大小。由于参数 n 是固定的，最小化 $Var(S_\sigma)$ 的值可以得到访问尽可能均衡的部分重复码。

记

$$\mathsf{MinVar}(S)=\min_\sigma Var(S_\sigma)。$$

则一个基于 S 的部分重复码且拥有满足关系 $Var(S_\sigma)=\mathsf{MinVar}(S)$ 的标号 σ 称为 *MinVar 部分重复码*。如果 S 是一个正则图，则 $\mathsf{MinVar}(S)=0$ 当且仅当该图是超幻的。下一节将回顾一些图幻标结果。

3.3.2　幻标

1963 年，Sedláček 在考虑数论中的幻方概念的时候提出了图论中的幻标概念[57]。此后 Stewart 在文献[58] 和[59] 中研究了图中边标号的各种各样的问题。

给定连通图 $G=(V,E)$ 和从 E 到正整数集的单射 σ，令

$$\sigma^*(v):=\sum_{e\in E:v\in e}\sigma(e)。$$

如果对所有的 $v\in V$，都有 $\sigma^*(v)=\lambda$，那称标号 σ 是图 G 上指数为 λ 的 *幻标*。进一步地，当 $\{\sigma(e):e\in E\}$ 是由连续正整数组成的时候，称标号 σ 是 *超幻的*。如果图 G 存在一个超幻标 (幻标)，则称图 G 是超幻的 (幻的)。

迄今为止，人们已经发表了大量关于幻图和超幻图的文章。例如文献[60-66]。当图的标号集合是 $[|E|]$，且对任意顶点 $v\in V$，$\sigma^*(v)=\deg(v)(1+|E|)/2$，则称该图是 *度幻的*[67-68]。注意到如果图 G 是正则图，那么 G 是超幻的当且仅当 G 是度幻的[68]。我们推荐读者在综合文献[69] 中进一步地了解相关知识。

如果图 G 是超幻的 (或者度幻的) 正则图，则根据 $Var(G_\sigma)$ 和度幻标的定义知道超幻标 σ 满足 $Var(G_\sigma)=0$。也就是说超幻正则图能构造出达到 MinVar 的部分重复码且访问方差为零。

Ivančo 在文献[65] 中给出了所有超幻完全多部正则图的刻画，其中完全多部正则图就是大家熟知的 Turán 图。在此，我们总结一下相关结论。

定理 3.3　[65] 当 r 整除 n 且 $r \geqslant 2$ 的时候，Turán 图 $T(n,r)$ 是超幻的当且仅当下列条件之一满足：

(1) $n = r$，也即完全图 K_n，其中 $n = 2$ 或者 $n \geqslant 6$ 且 $n \not\equiv 0 \pmod 4$；

(2) $n = 2r \geqslant 6$，也即 Turán 图 $T(2r,r)$，其中 $r \geqslant 3$；

(3) $n \geqslant 3r$，除了 $r \equiv 0 \pmod 4$ 且 $\frac{n}{r}$ 为奇数的情形。

定理 3.4　[17] r 整除 n 且 $r \geqslant 2$ 的 Turán 图 $T(n,r)$ 能生成最优 $(n,\alpha,2)$ 部分重复码。其中 $\alpha = (r-1)\frac{n}{r}$。

定理 3.3 和定理 3.4 可以立即推导出以下结果。

推论 3.5　令 $r \geqslant 2$，r 整除 n 且 $\alpha = (r-1)\frac{n}{r}$。当下列任意条件之一成立的时候，最优 $(n,\alpha,2)$ 部分重复码的访问方差为零。

(1) $n = r = 2$ 或者 $n = r \geqslant 6$ 且 $n \not\equiv 0 \pmod 4$；

(2) $n = 2r \geqslant 6$；

(3) $n \geqslant 3r$，除了 $r \equiv 0 \pmod 4$ 且 $\frac{n}{r}$ 是奇数的情形。

问题 3.6　当 $n = r \equiv 0 \pmod 4$ 时，也即，顶点数被 4 整除的完全图，或者 $r \equiv 0 \pmod 4$ 且 $\frac{n}{r}$ 是奇数的 Turán 图时，那么 Turán 图 $S = T(n,r)$ 的最小方差 $\mathsf{MinVar}(S)$ 是多少?

文献[17] 的作者指出大围长的图也可以生成最优部分重复码。

定理 3.7　[17] 如果存在顶点数为 n 同时围长为 g 的 α 正则图，那么存在参数为 $(n,\alpha,2)$ 的 k-最优的部分重复码。进一步如果 $g \geqslant \alpha + 1$，则存在最优部分重复码。

问题 3.8　对于大围长的正则图，它的最小访问方差是多少?

注意到问题 3.2 可以看作幻标问题的拓展。当图不是超级幻的时候，它将给出一个标号与超幻标的差异参考。

3.3.3　等价问题

问题 3.2 是为了寻找集合系上的区组标号使得访问方差最小。下面将对偶地提出一个关于线图上的顶点标号问题。

为了方便起见，我们总是假设 $S = (V,E)$ 是一个 n 阶 ρ 一致 α 正则的集合系，其中 $V = \{v_1, v_2, \cdots, v_n\}$，$E = \{e_1, e_2, \cdots, e_\theta\}$ 且满足关系 $\theta = n\alpha/\rho$。令 $I(S)$ 是行和列分别由 V 和 E 标记的关联矩阵，$\mathbf{1}_n$ 为 n 长的全 1 列向量。则

$$I(S)\cdot\mathbf{1}_\theta = \alpha\cdot\mathbf{1}_n \text{ 且 } \mathbf{1}_n^T\cdot I(S) = \rho\cdot\mathbf{1}_\theta^T。$$

记 $M(S) := I(S)^T \cdot I(S)$，则

$$M(S)_{(i,j)} = \begin{cases} \rho & \text{如果 } i = j; \\ |e_i \cap e_j| & \text{如果 } i \neq j。 \end{cases}$$

如果 S 是线性集合系，则 $M(S) = \rho I_\theta + A(L(S))$，其中 I_θ 是大小为 θ 的单位矩阵。$A(L(S))$ 是线图 $L(S)$ 的邻接矩阵。显然 $L(S)$ 是一个有 θ 个点的 d 正则图，其中 $d = \rho(\alpha - 1)$，也即，$A(L(S))$ 中每行每列都恰好有 d 个 '1'。如果 S 不是线性集合系，则 $A(L(S))$ 是多重图 $L(S)$ 的邻接矩阵，其中多重图的顶点 e_i 与顶点 e_j 恰好由 $|e_i \cap e_j|$ 条平行边连接。

引理 3.9　给定一个 n 阶 ρ 一致 α 正则的线性集合系 $S = (V, E)$ 和一个边标号 σ 满足 $\sigma(e_j) = i_j$，则访问方差

$$Var(S_\sigma) = (i_1, i_2, \cdots, i_\theta) A(L(S)) (i_1, i_2, \cdots, i_\theta)^T + c,$$

其中 $c = c(\theta, \rho, \alpha)$ 是一个常数。

证明　由于 $\sigma(e_j) = i_j$ 是一个双射，

$$\sum_{j=1}^{\theta} i_j = \sum_{j=1}^{\theta} j = \frac{\theta^2 + \theta}{2},$$

$$\sum_{j=1}^{\theta} {i_j}^2 = \sum_{j=1}^{\theta} j^2 = \frac{\theta(\theta+1)(2\theta+1)}{6}。$$

等式 $I(S) \cdot \mathbf{1}_\theta = \alpha \cdot \mathbf{1}_n$ 和等式 ${\mathbf{1}_\theta}^T \cdot A(L(S)) = \rho(\alpha - 1)\mathbf{1}_\theta$ 可以计算出访问方差为：

$$\begin{aligned} Var(S_\sigma) =& \| I(S)(i_1 - \frac{\bar{a}}{\alpha}, i_2 - \frac{\bar{a}}{\alpha}, \cdots, i_\theta - \frac{\bar{a}}{\alpha})^T \|_{\mathbb{L}^2} \\ =& (i_1 - \frac{\bar{a}}{\alpha}, i_2 - \frac{\bar{a}}{\alpha}, \cdots, i_\theta - \frac{\bar{a}}{\alpha}) \cdot M(S) \cdot (i_1 - \frac{\bar{a}}{\alpha}, i_2 - \frac{\bar{a}}{\alpha}, \cdots, i_\theta - \frac{\bar{a}}{\alpha})^T \\ \triangleq& (i_1, i_2, \cdots, i_\theta) M(S) (i_1, i_2, \cdots, i_\theta)^T + c_1 \\ \triangleq& (i_1, i_2, \cdots, i_\theta) A(L(S)) (i_1, i_2, \cdots, i_\theta)^T + c_1 + c_2 \\ \triangleq& \sum_{e_k \cap e_l \neq \varnothing} i_k i_l + c, \end{aligned}$$

其中

$$\begin{aligned} c_1 &= (-\frac{\bar{a}}{\alpha}\mathbf{1}_\theta^T) M(S) (-\frac{\bar{a}}{\alpha}\mathbf{1}_\theta) - 2(i_1, i_2, \cdots, i_\theta) M(S) (\frac{\bar{a}}{\alpha}\mathbf{1}_\theta) \\ &= \frac{\bar{a}^2}{\alpha^2}\mathbf{1}_\theta^T \rho\alpha \mathbf{1}_\theta - 2\frac{\bar{a}}{\alpha}(i_1, i_2, \cdots, i_\theta)\rho\alpha\mathbf{1}_\theta \\ &\qquad (\text{因为 } M(S)\mathbf{1}_\theta = \rho\alpha\mathbf{1}_\theta) \\ &= \frac{\bar{a}^2}{\alpha}\rho\theta - 2\bar{a}\rho \sum_{j=1}^{\theta} j = \bar{a}\rho\theta\left(\frac{\bar{a}}{\alpha} - (\theta+1)\right) \\ &= -\frac{\bar{a}\rho\theta(\theta+1)}{2}, \quad \left(\text{Since } \frac{\bar{a}}{\alpha} = \frac{\theta+1}{2}\right) \end{aligned}$$

及 $c_2 = \rho \sum_{i=1}^{\theta} i^2$。则 $c = c_1 + c_2 = \frac{\rho\theta(\theta+1)(2\theta+1)}{6} - \frac{\rho\alpha\theta(\theta+1)^2}{4}$。 ∎

我们从引理 3.9中知道问题 3.2 只需要考虑 S 的线图。因此如下关于图顶点标号优化问题十分重要。当 $G = L(S)$ 的时候，下面问题等价于问题 3.2。

问题 3.10　给定一个 θ 阶 d 正则图 G，其中图的顶点为 $v_1, v_2, \cdots, v_\theta$。哪个重量函数 $f : V(G) \to [\theta]$ (双射) 使得 $M(f) := \sum_{v_i \sim v_j} f(v_i)f(v_j)$ 最小化? 其中每条边在 $M(f)$ 的求和项中只计算一次。记 $M(G) = \min_f M(f)$ 并且称其为图 G 的 MinPS (它是最小乘积和的缩写)。

研究问题 3.10 中多个完全图的并、多个 Turán 图的并及圈的原因是：正如问题 3.6 和问题 3.8，集合系 S 是 Turán 图或者大围长图的 MinVar(S) 值得大家关注，但是它们能等价地转为研究其线图 $G = L(S)$ 的 MinPS。例如，集合系 S 是圈 (大围长的图)，它的线图 $L(S)$ 仍然是其本身。它的 MinPS 会在后面章节 3.4.3 中确定。然而对于一般的 Turán 图 $T(n, r)$，线图 $L(S)$ 的 MinPS 是很难确定的，甚至对于 $n \equiv 0 \pmod 4$ 的 Turán 图 $S = T(n, n)$，也即完全图 K_n 也是很难确定的。在这样情况下，完全图 K_n 可以分解成一个 Turán 图 $T(n, n/4)$ (超级幻图) 和 $n/4$ 个不交的 K_4 (它的线图为 $n/4$ 个不交的 Turán 图 $T(6, 3)$)。正如引理 3.14所述，当考虑多个 Turán 图的并的 MinPS 时，多个完全图的并的 MinPS 需要被确定下来。此外，上述图的结果可以解决一些部分重复码的最小方差问题 (问题 3.2)，后续小节将给出相关的例子。

下面给出图 G 的 MinPS 的上界。

引理 3.11　给定一个 θ 阶 d 正则图 G，$M(G) \leqslant \frac{d(3\theta+2)\theta(\theta+1)}{24}$。

证明　通过计算 $M(f)$ 的平均值：

$$\begin{aligned}
\sum_{\sigma \in S_\theta} M(f_\sigma) &= \sum_{\sigma \in S_\theta} \sum_{v_i \sim v_j} f_\sigma(v_i) f_\sigma(v_j) \\
&= \sum_{v_i \sim v_j} \sum_{\sigma \in S_\theta} f_\sigma(v_i) f_\sigma(v_j) \\
&= \sum_{v_i \sim v_j} (\theta - 2)! \sum_{1 \leqslant a \neq b \leqslant \theta} ab \\
&= \sum_{v_i \sim v_j} (\theta - 2)! \left(\left(\sum_{i=1}^{\theta} i \right)^2 - \sum_{i=1}^{\theta} i^2 \right) \\
&= \frac{d(3\theta + 2)\theta^2(\theta^2 - 1)}{24} (\theta - 2)!。
\end{aligned}$$

最后 $M(G) \leqslant \frac{d(3\theta+2)\theta(\theta+1)}{24}$。 ∎

3.4　确定问题 3.10中的 MinPS 值

这一节聚焦解决问题 3.10 中的几类图，例如多个完全图的并、多个 Turán 图的并以及圈。为了方便起见，对于任意的整数 $a \leqslant b$，$[a,b]$ 表示整数集合 $\{a, a+1, \cdots, b\}$。

在解决问题 3.10之前，图 3.1可以用来阐释 MinVar 问题和 MinPS 问题在线性集合系和对偶集合系之间的关系。

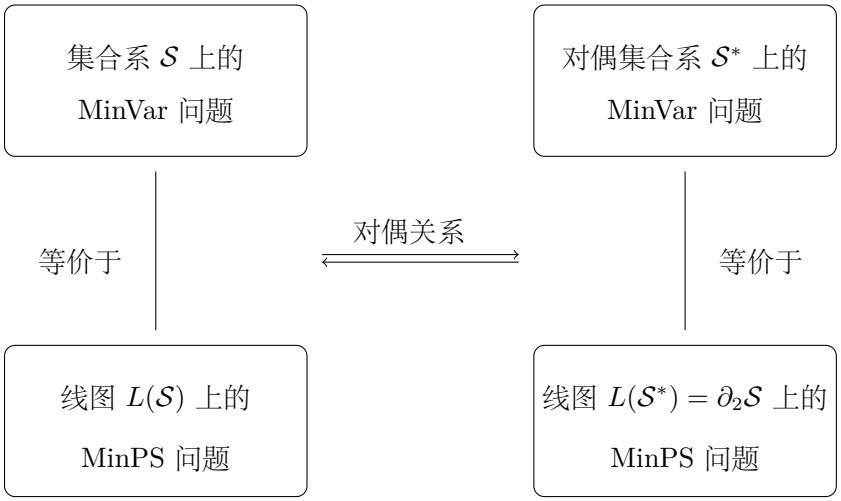


图 3.1　线性集合系与其对偶集合系在 MinVar 问题和 MinPS 问题之间的关系。因为 S^* 中任意两个区组相交当且仅当在 S 中对应的点在同一个区组里面，所以 $L(S^*) = \partial_2 S$

3.4.1　多个完全图的并

令 mK_r 为 m 个完全图 K_r 的不交并。本小节考虑 mK_r 的 MinPS 并将其用于引理 3.14 中来确定多个 Turán 图的并 $mT(n,r)$ 的 MinPS。

当 $m = 1$ 时，它就是一个完全图 K_r。对于任意的标号 f，$M(f)$ 显然是一个常数且 $M(K_r) = \sum_{1 \leqslant i < j \leqslant r} ij$。下例的部分重复码 S 的线图 $L(S)$ 是完全图，例如 S 是 2-$(q^2+q+1, q+1, 1)$ 设计 (也即，对称设计[31])，其中任何两个不同的区组恰好相交一个点。

引理 3.12　当 $m \geqslant 2$ 时，$M(mK_r)$ 的 MinPS 能确定，并能找到每个点的标号和尽可能均匀的标号。

证明　设 f 是图 mK_r 的点标号，令 V_i 是第 i 个完全图 K_r 上的标号集合，则

$|V_i| = r$ 且 $\bigcup_{i=1}^{m} V_i = [mr]$。同时

$$\begin{aligned} M(f) &= \sum_{i=1}^{m} \sum_{u<v\in V_i} uv = \sum_{i=1}^{m} \frac{1}{2}\left(\left(\sum_{u\in V_i} u\right)^2 - \sum_{u\in V_i} u^2\right) \\ &= \frac{1}{2}\sum_{i=1}^{m}\left(\sum_{u\in V_i} u\right)^2 - \frac{1}{2}\sum_{i=1}^{mr} i^2 \\ &\geqslant \frac{m}{2}\left(\frac{\sum_{i=1}^{mr} i}{m}\right)^2 - \frac{mr(mr+1)(2mr+1)}{12} \\ &= \frac{mr^2(mr+1)^2}{8} - \frac{mr(mr+1)(2mr+1)}{12}。 \end{aligned}$$

上面的等式成立当且仅当对任意的 $i \in [m]$，$\sum_{u\in V_i} u$ 的值都相同。除了 r 是奇数且 m 是偶数的情形 (这个情形下，$\frac{\sum_{i=1}^{mr} i}{m}$ 不是整数) 其余情形都可以取等号，而在这个例外的情形下，对任意的 $i \in [m]$，$\sum_{u\in V_i} u$ 两两最多相差 1 的时候取到最小值。 ∎

注解　引理 3.12 中的 $M(mK_r)$ 能通过下面的构造达到。

(1) 当 r 是偶数的时候，对任意的 $i \in [m]$，令

$$V_i = \left[\tfrac{(i-1)r}{2}+1, \tfrac{ir}{2}\right] \bigcup \left[mr+1-\tfrac{ir}{2}, mr-\tfrac{(i-1)r}{2}\right] = V_i^{(r)}。$$

则对任意的 i，$\sum_{u\in V_i} u = \frac{(rm+1)r}{2}$ 且 $M(mK_r) = \frac{mr^2(mr+1)^2}{8} - \frac{mr(mr+1)(2mr+1)}{12}$。

(2) 当 m 和 $r > 1$ 都是奇数的时候，令

$$V_i = V_i^{(r-3)} \bigcup \left\{(r-3)m+i, \tfrac{(2r-3)m-1}{2}+i, rm+2-2i\right\}, i \in \left[\tfrac{m+1}{2}\right]$$

并且对任意的 $i \in \left[\frac{m+3}{2}, m\right]$，

$$V_i = V_i^{(r-3)} \bigcup \left\{(r-3)m+i, \tfrac{(2r-5)m-1}{2}+i, (r+1)m+2-2i\right\}。$$

则对任意的 i，$\sum_{u\in V_i} u = \frac{(rm+1)r}{2}$ 且 $M(mK_r) = \frac{mr^2(mr+1)^2}{8} - \frac{mr(mr+1)(2mr+1)}{12}$。

(3) 当 $r > 1$ 是奇数且 m 是偶数的时候，令

$$V_i = V_i^{(r-3)} \bigcup \left\{(r-3)m+i, \tfrac{(2r-3)m}{2}+i, rm+2-2i\right\}, i \in \left[\tfrac{m}{2}\right]$$

并且对任意的 $i \in \left[\frac{m}{2}+1, m\right]$，

$$V_i = V_i^{(r-3)} \bigcup \left\{(r-3)m+i, \tfrac{(2r-5)m}{2}+i, (r+1)m+1-2i\right\}。$$

则

$$\sum_{u\in V_i} u = \begin{cases} \frac{(rm+1)r+1}{2}, & i \in \left[\frac{m}{2}\right]; \\ \frac{(rm+1)r-1}{2}, & i \in \left[\frac{m}{2}+1, m\right]。 \end{cases}$$

因此，$M(mK_r) = \frac{mr^2(mr+1)^2+m}{8} - \frac{mr(mr+1)(2mr+1)}{12}$。

3.4.2　多个 Turán 图的并

接下来，当 r 整除 n 的时候，Turán 图 $T(n,r)$ 的 MinPS 将被确定。特别地，对偶横截设计 TD$(r, n/r)$[31] 的线图就是 Turán 图 $T(n,r)$。首先，下面简单但是非常有用的注解将频繁地被应用到以后的证明中。

注解　假设 f 是一个达到图 G MinPS 值的重量函数。对每个顶点 v，记 $f(N(v)) := \sum_{u\in N(v)} f(u)$。如果 v_i 和 v_j 不相邻且 $f(v_i) < f(v_j)$，则 $f(N(v_i)) \geqslant f(N(v_j))$。

事实上，如果 $f(N(v_i)) < f(N(v_j))$，那么通过交换一下 v_i 和 v_j 的重量可以获得新的重量函数 f'，其中 $f'(N(v_i)) = f(N(v_i)), f'(N(v_j)) = f(N(v_j))$ 且 $f'(v_i) = f(v_j), f'(v_j) = f(v_i)$。但是

$$\begin{aligned} M(f') - M(f) =& f'(v_i)f'(N(v_i)) + f'(v_j)f'(N(v_j)) \\ & - f(v_i)f(N(v_i)) - f(v_j)f(N(v_j)) \\ =& f(v_i)(f(N(v_j)) - f(N(v_i))) + f(v_j)(f(N(v_i)) - f(N(v_j))) \\ =& (f(N(v_i)) - f(N(v_j)))(f(v_j) - f(v_i)) < 0, \end{aligned}$$

这与 $M(f) = M(G)$ 这样的事实矛盾！

引理 3.13　令 r 整除 n 且 $r \geqslant 2$，则 Turán 图 $T(n,r)$ 的 MinPS 为

$$M(T(n,r)) = \sum_{1\leqslant j<j'\leqslant r} \left(\sum_{k=(j-1)n/r+1}^{jn/r} k \right) \left(\sum_{l=(j'-1)n/r+1}^{j'n/r} l \right)。$$

证明　令 $m = n/r \geqslant 2$ (当 $m = 1$ 时，该图为完全图)，假设 $G = (V, E)$ 是 Turán 图 $T(n,r)$，其顶点集划分为 $V = V_1 \cup V_2 \cup \cdots \cup V_r$ 且每部的大小为 $|V_i| = m$。对于给定的重量函数 f，记 $f(V_i) = \sum_{v\in V_i} f(v)$ 和 $f(V) = \sum_{v\in V} f(v)$，则 $M(f) = \sum_{1\leqslant i<j\leqslant r} f(V_i)f(V_j)$。不失一般性，我们假设 $f(V_1) \leqslant f(V_2) \leqslant \cdots \leqslant f(V_r)$。

断言：使得 $M(f)$ 最小的重量函数满足以下性质：对任意 $i \in [r]$ 均有 $f(V_i) = \sum_{j=(i-1)m+1}^{im} j$，也即，点集 V_i 上的标号为 $(i-1)m+1, (i-1)m+2, \cdots, im$。

反证法证明断言：假设有两个指标 $i < i'$，存在点集 V_i 中标号 x 和点集 $V_{i'}$

中的标号 y 满足 $x > y$。通过互换标号 x 和 y 的值得到新重量函数 f'，且满足

$$\begin{aligned} M(f') - M(f) =& \sum_{1\leqslant i<j\leqslant r} f'(V_i)f'(V_j) - \sum_{1\leqslant i<j\leqslant r} f(V_i)f(V_j) \\ =&(f(V_i) - x + y)(f(V_{i'}) - y + x) - f(V_i)f(V_{i'}) \\ =&(f(V_i) - f(V_{i'}))(x - y) - (x - y)^2。\end{aligned}$$

不等式 $f(V_{i'}) \geqslant f(V_i)$ 和 $x > y$ 推出 $M(f') < M(f)$，矛盾！此时引理结论成立。 ∎

现在我们考虑多个 Turán 图的并 $mT(n,r)$，其中 r 整除 n。$mT(n,r)$ 的 MinPS 将会在引理 3.19中用来估计 MinVar(K_{4r})。

引理 3.14　当 r 整除 n 时，$mT(n,r)$ 的 MinPS 可以由 mK_r 的 MinPS 确定。

证明　令 $l = n/r \geqslant 2$，对任意的 $i \in [r]$，$j \in [m]$。令 V_{ij} 是第 j 个图 $T(n,r)$ 中第 i 个部分的所有标号集合。则 $|V_{ij}| = l$ 及 $\bigcup_{i\in[r],j\in[m]} V_{ij} = [mn]$。假设重量函数 f 的 $M(f)$ 值是最小的。令 $V_j = \bigcup_{i\in[r]} V_{ij}$。引理 3.13的证明知道对任意的 j，V_{1j} 是 V_j 中最小的 l 个整数，V_{2j} 是 $V_j \setminus V_{1j}$ 中最小的 l 个整数，其他的以此类推。为了方便起见，记该性质为 P。

断言：在 V_{ij} 中任何两个不同的集合 V_{ij} 和 $V_{i'j'}$，$V_{i'j'}$ 中的元素都比 V_{ij} 中的元素小，或者 $V_{i'j'}$ 中的元素都比 V_{ij} 中的元素大。也就是说，V_{ij} 必定是某个集合 $U_t = [l(t-1)+1, lt]$，其中 $t \in [mr]$。

反证法证明此断言：假设存在两个集合 V_{ij} 与 $V_{i'j'}$ 和整数 $x, y \in V_{ij}$ 与整数 $z, w \in V_{i'j'}$ 使得 $x > z$ 和 $y < w$。根据性质 P 得到 $j \neq j'$。令 $F_j = \sum_{s\neq i} V_{sj}$ 和 $F_{j'} = \sum_{a\neq i'} V_{aj'}$。假设 $F_j \geqslant F_{j'}$。令 f' 为互换 x 和 z 之后的重量函数。则

$$M(f') - M(f) = (zF_j + xF_{j'}) - (xF_j + zF_{j'}) = (z - x)(F_j - F_{j'}) \leqslant 0。$$

因此互换 x 和 z 之后，$M(f)$ 的值是不增的。我们持续这样的操作，最终可以得到 V_{ij} 的元素比 $V_{i'j'}$ 的元素都小。继续对形如 V_{ij} 和 $V_{i'j'}$ 这样的集合对进行上述步骤，最后每个集合 V_{ij} 将是某个集合 U_t。

上述的断言可以保证每个 V_{ij} 是某个集合 U_t。现在我们计算 $M(f)$，在接下

来的等式中假设 $V_{ij}=U_t$ 和 $V_{i'j}=U_{t'}$。

$$\begin{aligned}M(f)&=\sum_{j\in[m]}\sum_{1\leqslant i<i'\leqslant r}\left(\sum_{u\in V_{ij}}u\right)\left(\sum_{u\in V_{i'j}}u\right)\\&=\sum_{j\in[m]}\sum_{1\leqslant i<i'\leqslant r}\frac{l(2lt-l+1)}{2}\times\frac{l(2lt'-l+1)}{2}\\&=\binom{r}{2}\frac{ml^2(1-l)^2}{4}+l^3(1-l)\sum_{j\in[m]}\sum_{1\leqslant i<i'\leqslant r}\frac{t+t'}{2}+l^4\sum_{j\in[m]}\sum_{1\leqslant i<i'\leqslant r}tt'\\&=\binom{r}{2}\frac{ml^2(1-l)^2}{4}+\frac{l^3-l^4}{4}rm(rm+1)(r-1)+l^4M(\bar{f}),\end{aligned}$$

其中 $\bar{f}$ 是图 mK_r 诱导的点标号。注解 3.4.1 能帮助找到最优的标号 $\bar{f}$ 使得 $M(\bar{f})=M(mK_r)$，从而推导出 $mT(n,r)$ 的最优标号 f。∎

3.4.3　圈

长圈是大围长的图，并且它的线图就是自身。现在我们考虑 C_θ 的 MinPS，记 $M_\theta=M(C_\theta)$。

引理 3.15　对任意给定的 $\theta\geqslant 3$，$M_{\theta+2}\geqslant M_\theta+\theta^2+4\theta+5$。

证明　反证法证明：假设存在一个圈 $C_{\theta+2}$ 上的标号 f 使得 $M(f)$ 最小并且满足 $M(f)<M_\theta+\theta^2+4\theta+5$。从标号 f 出发，如果能推导出一个圈 C_θ 的新标号 f' 满足 $M(f')<M_\theta$，那引理证明完毕。下面将证明分成三种情况进行构造新标号 f'。

情形一: 如果在圈 $C_{\theta+2}$ 中，有这样的标号片段 $\cdots x\ \theta+2\ 1\ y\cdots$。则该圈将进行如下两步操作。

(S1) 删除标号为 1 和标号为 $\theta+2$ 的两个顶点，并将标号为 x 和标号为 y 的两个顶点连接起来。$M(f)$ 将变成 $M=M(f)-x(\theta+2)-(\theta+2)-y+xy$。

(S2) 将剩下长度为 θ 的圈中每个标号 l 换做 $l-1$。则获得一个圈 C_θ 的标号 f'。

现在我们计算 $M(f')$。在步骤 (S1) 中，$M=\sum_{u\sim v}uv$，其中每个 $u\in[2,\theta+1]$ 在求和项中出现两次。由于 $uv=(u-1)(v-1)+u+v-1$，便有

$$M=\sum_{u\sim v}(u-1)(v-1)+2\sum_{i=2}^{\theta+1}i-\theta=M(f')+\theta^2+2\theta。$$

因此在步骤 (S2) 之后，

$$\begin{aligned}M(f')=&M(f)-x(\theta+2)-(\theta+2)-y+xy-\theta^2-2\theta\\<&M_\theta+\theta^2+4\theta+5-x(\theta+2)-(\theta+2)-y+xy-\theta^2-2\theta\\=&M_\theta+(x-1)(y-\theta-2)+1。\end{aligned}$$

由于 $y < \theta + 2$ 和 $x \geqslant 2$，可知 $M(f') < M_\theta$，这与 $M(f)$ 最小矛盾。结论得证。

情形二: 假设在圈 $C_{\theta+2}$ 中，有这样的标号片段 $\cdots x\ \theta+2\ z\ 1\ y \cdots$。删除标号为 1 和标号为 $\theta + 2$ 的两个顶点，并连接 xz 和 yz。则 $M(f)$ 变为

$$M = M(f) - (x + z)(\theta + 2) - (y + z) + xz + yz。$$

再进行情形一中的步骤 (S2)，获得了圈 C_θ 的一个标号 f'。与情形一类似，有

$$\begin{aligned} M(f') =& M - \theta^2 - 2\theta \\ =& M(f) - (x + z)(\theta + 2) - (y + z) + xz + yz - \theta^2 - 2\theta \\ <& M_\theta + \theta^2 + 4\theta + 5 - (x + z)(\theta + 2) - (y + z) + xz + yz - \theta^2 - 2\theta \\ =& M_\theta + (x - 2)(z - \theta - 2) + (z - 1)(y - 1) - z\theta。 \end{aligned}$$

不等式 $y, z < \theta + 2$ 和 $x \geqslant 2$ 可以推出 $M(f') < M_\theta$，矛盾。结论得证。

情形三: 假设在圈 $C_{\theta+2}$ 中，有这样的标号片段 $\cdots x\ \theta+2\ z \cdots w\ 1\ y \cdots$。

断言： z 和 w 的顶点不相邻。

否则 $f(N(z)) = \theta + 2 + w$ 且 $f(N(1)) = y + w$，因此 $f(N(1)) < f(N(z))$，这与注解 3.4.2矛盾。相似地 x 和 y 也不相邻。假设 $w > y$ ($w < y$ 的情况类似)，有一个标记 a 在标记 w 的左边，也即片段为 $\cdots a\ w\ 1\ y \cdots$。则该圈进行如下两步操作。

(T1) 删除标号为 1 和标号为 w 的两个顶点，并连接标号为 a 和标号为 y 的两个顶点。则 $M(f)$ 的值变为 $M = M(f) - aw - w - y + ay$。

(T2) 把剩下长度为 θ 的圈中所有的标号 $l > w$ 换做 $l - 2$ 同时把所有的标号 $l < w$ 换做 $l - 1$。则获得一个圈 C_θ 的标号 f'。

由于 $(u - 2)(v - 2) = uv - 2(u + v) + 4$, $(u - 1)(v - 2) = uv - 2u - v + 2$ 和 $(u - 1)(v - 1) = uv - (u + v) + 1$,

$$\begin{aligned} M(f') =& M - \sum_{u\sim v:u,v>w} (2u + 2v - 4) - \sum_{u\sim v:u<w<v} (2u + v - 2) - \sum_{u\sim v:u,v<w} (u + v - 1) \\ =& M - 4\sum_{i=w+1}^{\theta+2} i - 2\sum_{i=2}^{w-1} i + \sum_{u\sim v:u<w<v} (v - u + 2) + \sum_{u\sim v:u,v>w} 4 + \sum_{u\sim v:u,v<w} 1 \\ \triangleq& M - 2(\theta + w + 3)(\theta - w + 2) - (w + 1)(w - 2) + S, \end{aligned}$$

其中 S 是最后三项的和。那么 $M(f') - M_\theta$ 小于

$$\begin{aligned} T(w) :=& \theta^2 + 4\theta + 5 - aw - w - y + ay - 2(\theta + w + 3)\times \\ & (\theta - w + 2) - (w + 1)(w - 2) + S。 \end{aligned}$$

断言： 对任何的 w，$T(w) < 0$。

如果断言成立，则 $M(f') < M_\theta$ 与假设矛盾。结论得证。为了证明断言，S 的上

界有如下的估计。当 $u \sim v$ 且 $u < w < v$ 的时候，$v - u + 2 \geqslant 4$。因此 $[2, w-1]$ 和 $[w+1, \theta+2]$ 之间的交叉边越多，S 越大。依据 w 的取值，下面分两种情形进行考虑。

当 $w \leqslant \frac{\theta}{2} + 2$ 的时候，则 $|[2, w-1]| \leqslant |[w+1, \theta+2]|$。因此

$$\begin{aligned} S \leqslant & 2 \sum_{i=\theta-w+5}^{\theta+2} i - 2 \sum_{i=2}^{w-1} i + 4(w-2) + 4(\theta - 2w + 4) \\ = & (2\theta - w + 11)(w-2) - (w+1)(w-2) + 4(\theta - 2w + 4) \\ = & -2w^2 + (2\theta + 6)w - 4, \end{aligned}$$

取 $w = \frac{\theta}{2} + 2$ 的时候，上面的取值最大。从而

$$\begin{aligned} T(w) &\leqslant -w^2 + (2\theta + 8)w - (\theta + 3)^2 + a(y - w) - y \\ &= a(y - w) - y - \frac{\theta^2}{4} + 3 < 0。 \end{aligned}$$

当 $w \geqslant \frac{\theta}{2} + 2$ 的时候，则 $|[2, w-1]| \geqslant |[w+1, \theta+2]|$。因此

$$\begin{aligned} S \leqslant & 2 \sum_{i=w+1}^{\theta+2} i - 2 \sum_{i=2}^{\theta-w+3} i + 4(\theta - w + 2) + (2w - \theta - 4) \\ = & (\theta + w + 7)(\theta - w + 2) - (\theta - w + 5)(\theta - w + 2) + (2w - \theta - 4) \\ = & -2w^2 + (2\theta + 4)w + \theta, \end{aligned}$$

取 $w = \theta + 1$ 的时候，上面的取值最大。从而

$$\begin{aligned} T(w) &\leqslant -w^2 + (2\theta + 5)w - \theta^2 - 5\theta - 5 + (a-1)(y - w) \\ &= (a-1)(y - w) - 1 < 0。 \end{aligned}$$

∎

接下来证明引理 3.15中圈 C_θ 的 MinPS 能够达到。

引理 3.16　对任何给定的 $\theta \geqslant 3$，$M_{\theta+2} = M_\theta + \theta^2 + 4\theta + 5$。因此 $M_{2k+1} = (4k^3 + 12k^2 + 14k + 3)/3$，并对任意 $k \geqslant 1$，有 $M_{2k+2} = (4k^3 + 18k^2 + 29k + 12)/3$。

证明　我们对任意的 $\theta \geqslant 3$ 进行归纳证明更强的结果：圈 C_θ 中存在一个达到 M_θ 的标号 f 且标号 1 和标号 θ 的顶点相邻。

对于 $\theta = 3$ 时，$M(f)$ 是一个常数且 $M_3 = 11$，同时标号 1 和标号 3 相邻。对于 $\theta = 4$ 时，由引理 3.13，$M_4 = 21$ 且标号 1 和标号 4 相邻。假设对所有圈长不超过 θ 的圈，结论都正确。当 $\theta + 2$ 的时候，下面证明结论也是正确的。根据假设在圈 C_θ 中存在一个达到 $M(f') = M_\theta$ 的标号 f' 并且标号 1 和标号 θ 相邻。

现在将 f' 中的每个标号都加 1，则标号 2 和标号 $\theta+1$ 相邻。通过在标号 2 和标号 $\theta+1$ 之间插入两个标号为 $1,\theta+2$ 的顶点。片段形如 $\cdots 2\ \theta+2\ 1\ \theta+1\cdots$ 并且得到了圈 $C_{\theta+2}$ 的标号 f。容易验证

$$\begin{aligned} M(f) =& M_\theta + 2\sum_{i=1}^{\theta} i + \theta - 2(\theta+1) + 2(\theta+2) + (\theta+2) + (\theta+1) \\ =& M_\theta + \theta^2 + 4\theta + 5。 \end{aligned}$$

因此，$M_{\theta+2} \leqslant M_\theta + \theta^2 + 4\theta + 5$。从而引理 3.15 证明了结论。根据 M_3 和 M_4 的值和递归方法能精确地计算出 M_θ，也就是引理中的结果。这里省略了具体计算。 ∎

进一步地，引理 3.16 中圈 C_θ 的标号同时是问题 3.1中最小访问和达到最大值的标号。

定理 3.17　圈 C_θ 中存在一个标号同时是问题 3.1和问题 3.2中的最优标号。

证明　令集合系 $S_\theta = (V, E)$ 的线图是圈 C_θ，则 S_θ 同样是长度为 θ 的圈。引理 3.16中圈 C_θ 的顶点标号 f 自然地诱导出集合系 S_θ 的边标号 σ，它最小化 MinVar 模型中 S_θ 的访问方差。

断言：标号 σ 同时最大化 MaxMinSum 模型中的最小访问和。

我们知道 S_θ 的最小访问和最大值至多为 θ。当 $\theta = 3,4$ 的时候，定理显然成立。假设对所有圈长不超过 θ 时，断言都成立，现在我们考虑 $\theta+2$ 的情形。根据假设，引理 3.16 的重量函数 f' 诱导出 C_θ 的边标号 σ'，且该边标号最大化 S_θ 的最小访问和，也即，对于圈 C_θ 的任意边 $v_i \sim v_j$，$f'(v_i) + f'(v_j) \geqslant \theta$。根据引理 3.16构造出圈 $C_{\theta+2}$ 上的标号 f 形如 $\cdots 2\ \theta+2\ 1\ \theta+1\cdots$，它是从标号 f' 通过每个标号加 1，然后删除边 $2 \sim \theta+1$ 并添加 $2 \sim \theta+2, \theta+2 \sim 1, 1 \sim \theta+1$ 三条边得到的。显然圈 $C_{\theta+2}$ 的任意边 $v_i \sim v_j$，$f(v_i) + f(v_j) \geqslant \theta+2$。因此，诱导的边标号 σ 最大化了 $S_{\theta+2}$ 的最小访问和。 ∎

3.5　估计问题 3.2中的 MinVar(S)

本节将估计某些集合系 S 的 MinVar(S)。

引理 3.18　令 $H_1 = (V, E_1)$ 和 $H_2 = (V, E_2)$ 是两个顶点集相同的图。如果 $E_1 \cap E_2 = \varnothing$ 和 H_1 是超幻的，则图 $G = (V, E_1 \cup E_2)$ 的最小访问方差满足

$$\text{MinVar}(G) \leqslant \text{MinVar}(H_2)。$$

证明　定义图 G 的边标号 σ 如下。对图 H_1 用 $[|E_2|+1, |E_1|+|E_2|]$ 来标记边，并采用 H_1 的超幻标，使得对任意的 $v \in V$，$\sigma^*(v)$ 是一个常数。对图

H_2 用 $[1,|E_2|]$ 来标记边，并使得 σ 在 H_2 上的访问方差等于 $\mathsf{MinVar}(H_2)$。因此，$Var(G_\sigma)=\mathsf{MinVar}(H_2)$。引理得证。 ∎

接下来考虑问题 3.6 中的公开情形：K_{4r} 的访问方差，其中基于完全图的部分重复码产生了第一类具有恰当修复特性的 MBR 码[5-6]。为了估计 $\mathsf{MinVar}(K_{4r})$，K_{4r} 可以被视为一个多个完全图的并 rK_4 和 Turán 图 $T(4r,r)$ 的组合。定理 3.3 显示 $T(4r,r)$ 是超幻的，再依据引理 3.18，$\mathsf{MinVar}(K_{4r})\leqslant\mathsf{MinVar}(rK_4)$。特别地，$rK_4$ 的线图是 $rT(6,3)$。$M(rT(6,3))$、引理 3.14确定的 MinPS 和引理 3.9确定的 $\mathsf{MinVar}(rK_4)$ 一起得出 $\mathsf{MinVar}(K_{4r})$ 的上界。

引理 3.19　对任意的正整数 r，

$$\mathsf{MinVar}(K_{4r})\leqslant\begin{cases}3r, & r\text{ 是奇数时},\\ 7r, & r\text{ 是偶数时}。\end{cases}$$

证明　将引理 3.9中的参数设置为 $\rho=2,\alpha=3,\theta=6r$，并将引理 3.14中的参数设置为 $m=r,n=6,r=3$。则根据引理 3.9 和引理 3.12 得到 $\mathsf{MinVar}(K_{4r})$ 的上界为 $32M-72r^2-18r+c$，其中 $c=-r(6r+1)(30r+7)$ 且

$$M=\begin{cases}\frac{45r^3+36r^2+7r}{8}, & r\text{是奇数时},\\ \frac{45r^3+36r^2+8r}{8}, & r\text{是偶数时}。\end{cases}$$

最终引理得证。 ∎

结合引理 3.9和引理 3.11中的平均值上界，$\mathsf{MinVar}(K_{4r})=O(r^7)$。而引理 3.19得到的结果 $O(r)$ 远远改进了平均值上界。另一方面，接下来给出 $\mathsf{MinVar}(K_{4r})$ 的下界并得到 $\mathsf{MinVar}(K_{4r})=\Theta(r)$。事实上，$K_{4r}$ 节点热度平均值等于 $(4r-1)(8r^2-2r+1)/2$，但是它的小数部分为 0.5。又因为每个节点的热度是整数，所以有 $\mathsf{MinVar}(K_{4r})\geqslant 4r\times(0.5)^2=r$。因此，我们提出以下问题。

问题 3.20　引理 3.19中的上界是否是紧的?

特别地，最近 Colbourn[70] 通过直接构造 K_{4r} 的标号证明了 $\mathsf{MinVar}(K_{4r})=r$，由此解决了问题 3.20。

对于 Turán 图的情形，令 $G=L(T(n,r))$ 是满足 r 整除 n 及 $r\geqslant 2$ 的 Turán 图线图。问题 3.6 显示 $r\equiv 0\pmod 4$ 且 $\frac{n}{r}$ 为奇数的 Turán 图是唯一未解决的情形。Chetwynd 和 Hilton 关于正则图有以下的性质。

定理 3.21　[71] 令图 G 是 $2n$ 阶 d 正则图且 $d\geqslant\frac{12}{7}n$。则图 G 是可 1 因子分解的。

根据定理 3.21，当 $r\geqslant 7$，n 是偶数且 r 整除 n 的时候，$T(n,r)$ 是可 1-因子分解的。由于 $T(n,r)$ 中的完美匹配是线图 $L(T(n,r))$ 中的独立集，令 $m=\frac{(r-1)n}{r}$ 和 $d=$

$2(m-1)$, 则 $G = L(T(n,r)) = (V,E)$ 是 m 部 d 正则图, 其中 $V = V_1 \cup V_2 \cup \cdots \cup V_m$, 且对任意 $i \in [m]$ 均有 $|V_i| = \frac{n}{2}$。进一步地, 对每个 $i \neq j \in [m]$, 由点集 $V_i \cup V_j$ 诱导的子图是 2 正则图。

问题 3.22　对任意的 $r \geqslant 7$ 和被 r 整除的偶数 n, 如何确定 $G = L(T(n,r))$ 的 MinPS?

3.6　小结

我们从 DRESS 码和分布式存储系统中的访问均衡问题问题出发提出了新组合模型 (MinVar 模型): 集合系中满足问题 3.2 访问方差最小的区组标号问题。当考虑的图不是超幻的时候, 该问题被视为广义的幻标问题。进一步地, 线性集合系的 MinVar 问题 (问题 3.2) 和图的点标号问题 (问题 3.10) 之间具有等价关系。我们通过解决这两类问题构造出几类基于特殊图的最优部分重复码且拥有最小的访问方差。此外, 根据访问均衡问题在分布式存储系统中的应用, 问题 3.2和问题 3.10本身也值得进一步地研究。特别地, 问题 3.6 更值得进一步研究。

问题 3.6: 当 $r \equiv 0 \pmod 4$ 且 $\frac{n}{r}$ 是奇数的时候, Turán 图 $S = T(n,r)$ 的 MinVar(S) 是多少?

我们接下来简单讨论模型中数据块的访问频率服从 Zipf 法则的情形。运用相同的记号, 问题 3.10 将改写成如下:

问题 3.23　给定一个 θ 阶 d 正则图 G, 其顶点集为 $v_1, v_2, \cdots, v_\theta$ 以及某个实数 $\beta > 0$。如何寻找一个双射 $\ell : V(G) \to \{1, 1/2^\beta, \cdots, 1/\theta^\beta\}$, 使得 $\overline{M}(\ell) = \sum_{v_i \sim v_j} \ell(v_i)\ell(v_j)$ 最小化? 记 $\overline{M}(G) = \min_\ell \overline{M}(\ell)$ 为图 G 的 MinPS。

我们利用同样的方法得到问题 3.23中一些平行的结论。首先, 当 $G = K_\theta$ 时, 对任意的顶点标号, $\overline{M}(\ell)$ 是一个常数。对于 Turán 图, 当第 k 个部分的顶点标号为 $\left\{\frac{1}{((k-1)r+1)^\beta}, \frac{1}{((k-1)r+2)^\beta}, \cdots, \frac{1}{(kr)^\beta}\right\}$ 的时候, 标号可以达到 $\overline{M}(T(n,r))$。然而对于对个完全图的并或者多个 Turán 图的并, 如下集合划分问题的答案将帮助计算 $\overline{M}(mK_r)$ 和 $\overline{M}(mT(n,r))$。

问题 3.24　如何在多项式时间内寻找集合 $\left\{1, \frac{1}{2^\beta}, \cdots, \frac{1}{(mr)^\beta}\right\}$ 的 m 等份划分 $S_1, S_2, \cdots, S_m$, 使得 $\sum_{i=1}^{m} \left(\sum_{j \in S_i} j\right)^2$ 最小?

最后我们强调即使要求有零访问方差的性质, 等式 (3.1) 和等式 (3.2) 中关于 $[(\theta, M), k, (n, \alpha, \rho)]$DRESS 码的最大文件数 $A(n,k,\alpha,\rho)$ 的上界仍然是紧的, 其中推论 3.5 给出了一些最优部分重复码具有零访问方差的性质。对于带有访问方差限制的部分重复码, 改进等式 (3.1) 和等式 (3.2) 中的界是十分有意思的工作。但是运用文献[7] 中的原始证明很难解决考虑访问方差限制的问题。因此我们认为这个问题值得进一步地研究。

第 4 章　DNA 存储中的集合码

本章研究一类基于 DNA 存储的编码问题：构造出参数为 (t,k,v) 的最优平衡集合码。该编码本质上是一类具有固定相交数和最小差异值的集合系。本章首先将达到最优的平衡集合码等价于图中边标号问题或者特殊 “平衡表”，然后利用图论和组合设计相关知识求解问题，最后构造出所有 $k \leqslant 4$ 的最优平衡集合码，也即参数为 $(2,3,v),(2,4,v)$ 和 $(3,4,v)$ 的最优平衡集合码。并对任意的偶数 $k \geqslant 4$，本章利用广义 Reed-Solomon 码构造出渐近最优的 $(2,k,v)$ 集合码。对任意的正整数 $t \geqslant 2$，本章利用超图中的 Rödl Nibble 方法证明了参数为 $(t,t+1,v)$ 渐近最优集合码的存在性。

4.1　介绍

现代的数据存储系统主要依靠光学介质和磁介质媒体来记录海量数据，且存储系统保证数据能够有效地接受，检索和复制[18]。现有存储系统主要特性包括：随机访问、高精度检索、低成本和实时操作。随着如今数据的极速增长和长期的存储需求，传统的存储系统面临着 DNA 存储系统和多聚合物存储系统[19-22]的竞争。利用现有的 DNA 测序、合成及编辑技术优势[23-24]，DNA 存储系统有潜力突破传统存储系统遇到的瓶颈。此外 DNA 比常见的存储介质具有高密度存储、使用寿命长和易复制等诸多优势[19]。

近来许多学者[25-29] 提出了一些 DNA 上的编码技术。数据在这些技术模型中被排布在 DNA 双链的不同切割点中。由于 DNA 是双链结构 (一条链和一条反链)，数据需要尽可能均匀地分布在两条链当中以防止由切割点导致的断裂。同时为纠正数据读取错误，信息将被存储在一些只有少量重叠的切割点中。因此该问题等价于寻找一个具有较小差异值和较小相交数的集合码，该集合码问题主要关注如何构造出最大的子集族，使得该集族有较小的差异值和相交值。更确切地说，在给集族中背景集的每个元素标记 $+1$ 或者 -1 之后，每个子集内的标号和绝对值较小，且任何两个不同的子集具有较小的相交数。迄今为止已有许多关于集合差异值理论或者双染色理论的工作，例如文献[72-74]。此外该理论在各个领域有着广泛地应用。例如伪随机性与独立置换生成[75]、ϵ 近似[76]、箱嵌入、格近似以及图谱理论[77-78]。尽管有许多关于集合相交数的研究工作，例如 t 填充设计理论[31]，但是构造具有较小差异值和较小相交数的最大集合码仍然面临许多的挑战。一个自然的想法是在斯坦纳系上寻找标号，使得该系统有较小的差异值。然而 Gabrys 等人[30] 指出在斯坦纳系上面没有差异值为零的标号，并给出

拥有最小差异值和有限相交数的集合码的上界。这意味着大家需要同时关注集合码的差异值和相交数。

本章主要构造拥有最小差异值的集合码 (称为平衡集合码) 使其达到 Gabrys 的理论上界[30]，同时证明该上界在一些参数下是渐近最优的。对于 $k \leqslant 4$ 情形，最优平衡 (t,k,v) 集合码的构造等价于寻找图上的边标号或者是寻找特殊 “平衡表”。本章造借助于一些特殊的组合设计构型，例如正交阵列、Howell 设计，拉丁方等，构造出所有 $k \leqslant 4$ 的最优平衡集合码。第二个贡献在于找到了两类渐近最优平衡集合码：渐近最优 $(2,k,v)$ 平衡集合码，其中 k 是大于等于 4 的偶数；渐近最优 $(t,t+1,v)$ 平衡集合码，其中 t 是大于等于 3 的正整数。

4.2　准备工作

本小节将介绍一些必要的符号、集合码上的结果以及后面构造中所需的图和组合设计知识。

4.2.1　集合码

令 $t<k<v$ 为正整数，(t,k,v) 集合码是一个集族 $\mathcal{S} \subseteq \binom{[v]}{k}$ 使得 $[v]$ 中任何的 t 元集至多包含在一个 $S \in \mathcal{S}$ 当中。注意到，集合码的概念恰好与组合设计理论[31] 中的 t 填充一致。令 $L:[v] \to \{-1,+1\}$ 是 $[v]$ 上的标号且对每个 $S \in \mathcal{S}$，记 $L(S)=\sum_{x\in S} L(x)$。则 $\mathcal{S}$ 的 *差异值*为满足对任意的 $S \in \mathcal{S}, |L(S)| \leqslant d$ 的最小整数 d 。集合码 $\mathcal{S}$ 称为 *平衡的*[30] 是指对于偶数 k，集合码 $\mathcal{S}$ 的差异值为 0，对于奇数 k，集合码 $\mathcal{S}$ 的差异值为 1。

令集合 $[1,6]$ 中的顶点标记都为 -1 同时集合 $[7,12]$ 中的顶点标记都为 $+1$，则下面的集族 $\mathcal{S}_4$ 是平衡 $(2,4,12)$ 集合码，集族 $\mathcal{S}_3$ 是平衡 $(2,3,12)$ 集合码。

$$\begin{aligned}\mathcal{S}_4 = \{&\{1,4,7,10\},\{2,5,7,11\},\{3,6,7,12\},\\ &\{2,6,8,10\},\{3,4,8,11\},\{1,5,8,12\},\\ &\{3,5,9,10\},\{1,6,9,11\},\{2,4,9,12\}\}。\end{aligned} \tag{4.1}$$

$$\begin{aligned}\mathcal{S}_3 = \{&\{1,2,7\},\{3,7,9\},\{4,7,10\},\{5,7,11\},\{6,7,12\},\{2,3,8\},\\ &\{4,8,9\},\{5,8,10\},\{6,8,11\},\{1,8,12\},\{5,6,9\},\{1,9,11\},\\ &\{2,9,12\},\{1,6,10\},\{2,10,11\},\{3,10,12\},\{3,4,11\},\{4,5,12\}\}。\end{aligned} \tag{4.2}$$

令 $A(t,k,v)$ 为平衡 (t,k,v) 集合码的最大值，其中达到最大值的集合码称为 *最优的*。平衡集合码的核心问题为：给定 t,k 和 v 之后，如何确定出 $A(t,k,v)$ 的值? 文献[30] 给出了如下 $A(t,k,v)$ 的上界。

定理 4.1　[30] 给定正整数 $0 < t < k < v$，

$$A(t,k,v) \leqslant \frac{\binom{\lceil v/2\rceil}{\lfloor t/2\rfloor}\binom{\lfloor v/2\rfloor}{\lceil t/2\rceil}}{\binom{\lceil k/2\rceil}{\lfloor t/2\rfloor}\binom{\lfloor k/2\rfloor}{\lceil t/2\rceil}}。 \tag{4.3}$$

作者在文献[30] 中利用一种新型的拉丁矩阵构造出所有 $t = 2, k = 3$ 和 $v \geqslant 8$ 的最优平衡集合码。当 $t = 3$ 且 $k = 4$ 的时候，他们依据横截设计构造出 v 是 8 的倍数的最优平衡集合码。此外，他们借助限制和的构造证明了对所有参数为 $t \geqslant 2, k = t + 1$ 和偶数 v 的 (t,k,v) 集合码是渐近达到定理 4.1 中的上界。也即

$$\lim_{v\to\infty} \frac{A(t,t+1,v)\binom{\lceil (t+1)/2\rceil}{\lfloor t/2\rfloor}\binom{\lfloor (t+1)/2\rfloor}{\lceil t/2\rceil}}{\binom{\lceil v/2\rceil}{\lfloor t/2\rfloor}\binom{\lfloor v/2\rfloor}{\lceil t/2\rceil}} = 1。$$

这激发我们研究最大平衡集合码的渐近行为。

4.2.2　图与超图

在本章中，图的概念与第三章中的图概念稍微不同，这里会使用更加广义的图。与前面章节不同，本章所提及的超图更加强调其结构概念。

图 G 是一组集合对 $(V(G), E(G))$，其中点集 $V(G)$ 是一个非空集合同时边集 $E(G)$ 是一组无序对 (x,y) 其中 $x, y \in V(G)$。如果存在边 e，使得 $e = \{x,y\}$，则称顶点 $x, y \in V(G)$ 在 G 中 相邻 并记作 $x \sim y$。对两条边 $e_1, e_2 \in E(G)$，如果 $e_1 \cap e_2 \neq \varnothing$，则它们也称作在 G 中 相邻。环 $\ell = \{x\}$ 是一条特殊的边满足 $x \sim x$。重边 $e \in E(G)$ 是一条边满足 G 中有另外一条 e' 使得 $e = e'$。简单图是一个无环无重边的图。环图是一个包含环的图。对于正整数 c，简单图 G 的 c 边染色为映射 $f : E(G) \to [c]$ 使得 G 中任意相邻的两条边 e_1, e_2 都满足 $f(e_1) \neq f(e_2)$，其中使得 G 有 c 边染色的最小整数 c 称为 G 的 边染色数，记为 $\chi'(G)$。

令 G 为简单图。对于顶点 $v \in V(G)$，顶点 v 的 邻居为 $N(v) = \{u \in V(G) : u \sim v\}$ 同时顶点 v 的 点度为 $d(v) = |N(v)|$。令 H 是一个超图，对任意两个不同的顶点 $x, y \in V(H)$，x, y 的 共度为 $d(x,y) = |\{e \in E(H) : x, y \in e\}|$。图 (超图) 的 匹配 指的是一组互不相交的边 (超边)，进一步地，完美匹配 是包含所有顶点的匹配。

4.2.3　组合设计

n 阶 拉丁方 L 是一个 $n \times n$ 阵列，其中每项都是 n 元集中的一个元素且集合中的每个元素恰好出现在每行和每列各一次。假设 L_1 和 L_2 是两个 n 阶的拉丁方，其中的项分别属于集合 X 和 Y。如果对任意的 $x \in X, y \in Y$ 都存在唯一的格子 (i,j) 使得 $L_1(i,j) = x$ 且 $L_2(i,j) = y$，那称 L_1 和 L_2 是 正交的。

定理 4.2 [79] 对任意的正整数 $n \neq 2, 6$，都存在一对正交的 n 阶拉丁方。

阶为 $2n$ 且尺寸为 s 的 Howell 设计，记为 $H(s, 2n)$ 是一个 $s \times s$ 表格，其中的每个格子是空的或者是集合 X ($|X| = 2n$) 中一个无序对，同时满足

(a) 每行和每列具有拉丁性质 (也即，X 中的每个元素恰好出现在每行和每列的一个格子中),

(b) X 中的无序对出现在至多一个格子中。

下面提及一些关于 Howell 设计存在性的结果。

定理 4.3 [80] 对于正整数 $n+1$，$H(n+1, 2n)$ Howell 设计存在当且仅当 $n \notin \{1, 2, 4\}$。

令 L_1 和 L_2 是一对 n 阶正交拉丁方，$H(n, 2n)$ Howell 设计可以通过对任意的 $1 \leqslant i, j \leqslant n$，在 (i, j) 位置的格子中放置 $\{L_1(i, j), L_2(i, j) + n\}$ 来构造。因此定理 4.2 带来了一些 $H(n, 2n)$ Howell 设计的存在结果。

(k, n) **正交阵列** A，记作 $OA(k, n)$，是一个 $n^2 \times k$ 阵列，其中每项取自于集合 $[n]$，使得在 A 中的任意两列，任何点对 $(x, y) \in [n] \times [n]$ 都恰好出现在一行中。

4.3　小 k 下的最优平衡集合码

当 $k \leqslant 4$ 的时候，本小节对所有可能的 t 和 v 都给出 $A(t, k, v)$ 的确切值。特别地，对 $(t, k) \in \{(2, 3), (2, 4), (3, 4)\}$ 和所有的 v，最优平衡 (t, k, v) 集合码有精确的构造。此外，当 $(t, k) = (2, 3)$ 的时候，虽然文献[30] 中给出了最优平衡码构造，但是我们的构造与其本质不同并且更加的简单。

以下简述本章构造的主要思想。假设 S 是在背景集 $[v]$ 上的平衡 (t, k, v) 集合码。令 P_+ 是标号为 $+1$ 且大小为 p_+ 的点集，P_- 是标号为 -1 且大小为 p_- 的点集。则 S 的每个区组 B 都可以划分成两部分 $B = B_+ \cup B_-$，其中 $B_+ = B \cap P_+$ 和 $B_- = B \cap P_-$。根据平衡性的要求，B_+ 和 B_- 是几乎相等的。当 $k = 3, 4$ 的时候，它们的大小为 1 或者 2。现在定义一个顶点集为 P_+ 的图 G_+。对任意的 $B \in S$，集合 $B_+ \subseteq P_+$ 定义为图 G_+ 的边。注意到，当存在不同的区组 B 和 B' 满足 $B_+ = B'_+$ 的时候，图 G_+ 是包含重边的。当某个 B_+ 的大小为 1 的时候，图 G_+ 是包含环的。同时图 G_+ 的每条边 B_+ 用 B_- 来标记。上述方法可以从平衡集合码构造出标号图 G_+ (可能含重边或者环)，其中每条边 B_+ 加上其标号 $L(B_+) = B_-$ 组成集合码的一个区组。也就是说，图 G_+ 的边数等于集合码 S 的大小。类似地，平衡集合码可以构造出标号图 G_-，其中每条边 B_- 的标号为 $L(B_-) = B_+$。

例如，在式子 (4.1) 中的平衡 $(2, 4, 12)$ 集合码 S_4 形成如下的标号图 G_4:

标号图 G_+ 可以构造一个大小为 $p_+ \times p_+$ 的 "平衡表" T_+。它的行和列都分别由 $V(G) = P_+$ 标记。当 $\{i, j\}$ 是图中的一条边时，(i, j) 位置的格子填充所有 $\{i, j\}$

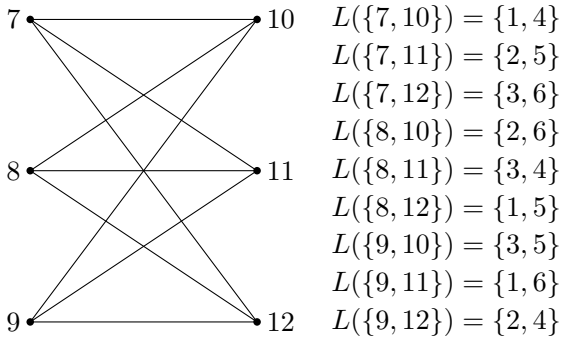


图 4.1　集合码 S_4 对应的标号图 $G_4 = K_{3,3}$ 及其标号 L

上的标号。否则格子是空的。$\boldsymbol{T}_+$ 自然是对称的且表格右上角区域 (包含主对角区域) 的标号数目等于集合码的区组数目，从而我们定义该数目为表格的大小。例如在表 4.1 中基于标号图 $\boldsymbol{G}_4$ 上的 “平衡表” $\boldsymbol{T}_4$ 的大小为 9。

表 4.1　基于标号图 G_4 上的 “平衡表” T_4，其中表的第一行和第一列代表图 G_4 的顶点，并无实际作用

	7	8	9	10	11	12
7				{1,4}	{2,5}	{3,6}
8				{2,6}	{3,4}	{1,5}
9				{3,5}	{1,6}	{2,4}
10	{1,4}	{2,6}	{3,5}			
11	{2,5}	{3,4}	{1,6}			
12	{3,6}	{1,5}	{2,4}			

本章的方法是通过相反方向构造出平衡集合码，首先刻画出标号图或者平衡表上的充分条件，使得对应的码是平衡 (t,k,v) 集合码，然后再构造出最大的表格。特别地，该表格与组合设计理论中著名 Howell 设计紧密相关。

4.3.1　参数为 $(t,k) = (2,4)$ 的最优平衡集合码

本小节给出平衡 $(2,4,v)$ 集合码的构造。根据定义知道图中的每条边与其相应的标号都是 2 元集且在标号图中没有重边。因此平衡 $(2,4,v)$ 集合码等价于顶点集为 $\boldsymbol{P}_+$ 标号集落在 $\binom{P_-}{2}$ 上的简单图 $\boldsymbol{G}_+$ 同时满足如下条件：

(C1)　如果任意两条边 e_1, e_2 相邻，则标号 $\boldsymbol{L}(e_1)$ 和 $\boldsymbol{L}(e_2)$ 互不相交；

(C2)　如果任意两条边 e_1, e_2 不相邻，则标号 $\boldsymbol{L}(e_1)$ 和 $\boldsymbol{L}(e_2)$ 是两个不同的 2 元集。

同样集合码与标号图 G_- 有相似的等价关系。这些等价关系可以稍微改进式子 (4.3) 中的上界 $A(2,4,v) \leqslant \frac{\lfloor v/2 \rfloor \lceil v/2 \rceil}{4}$。

命题 4.4

$$A(2,4,v) \leqslant \begin{cases} m^2, & \text{如果 } v = 4m \text{ 或者 } 4m+1,\ m \geqslant 1, \\ m^2+m, & \text{如果 } v = 4m+2 \text{ 或者 } 4m+3,\ m \geqslant 1。 \end{cases} \tag{4.4}$$

证明　我们依据命题的描述只需要证明 $v = 4m+1$ 和 $v = 4m+3$ 的情形。假设根据平衡 $(2,4,v)$ 集合码 $\mathcal{S}$ 获得的标号图为 G_+ 和 G_-，则根据性质 (C1) 得到所有相邻至同一个顶点的边集上的标号是互不相交的，因此图 G_+ 的每个顶点的点度至多为 $\lfloor p_-/2 \rfloor$。类似地图 G_- 的每个顶点的点度至多为 $\lfloor p_+/2 \rfloor$。因此 $e(G_+) \leqslant \frac{p_+ \lfloor p_-/2 \rfloor}{2}$ 且 $e(G_-) \leqslant \frac{p_- \lfloor p_+/2 \rfloor}{2}$。由于图 G_+ 和图 G_- 的边数都等于集合码的区组数目，便有 $|\mathcal{S}| \leqslant \min\{\frac{p_+ \lfloor p_-/2 \rfloor}{2}, \frac{p_- \lfloor p_+/2 \rfloor}{2}\}$。因此 $A(2,4,v) \leqslant \max_{p_+,p_-} \min\{\frac{p_+ \lfloor p_-/2 \rfloor}{2}, \frac{p_- \lfloor p_+/2 \rfloor}{2}\}$。取 $v = p_+ + p_- = 4m+1$ 或者 $4m+3$ 的时候，命题中的结果便成立。这里省略了简单地计算。 ∎

构造出满足性质 (C1)-(C2) 的简单标号图 G_+ 等价于构造出一个对称表 T_+ 其中每个格子是空的或者是一个 P_- 上的 2 元集，同时满足如下性质：

(T1) 所有对角线上的格子都是空的，且主对角以上的非空格子中元素是互不相同的；

(T2) 每行或者每列中所有非空格子构成了一个 P_- 上的匹配。

接下来将构造出满足上面条件的平衡表，且该平衡表对应的平衡集合码大小达到式子 (4.4) 中的上界。当 $v \in \{6,8,9,10\}$ 的时候，计算机的搜索发现式子 (4.4) 中的上界是达不到的。与之对应的最优平衡 $(2,4,v)$ 集合码将列举在本章末尾的附录 4.6。

定理 4.5　对于除了 $v \in \{6,8,9,10\}$ 的其他任意 $v \geqslant 4$，式子 (4.4) 中的上界是紧的。

证明　我们根据命题 4.4 把 v 按照模 4 划分成四种情形。当 v 是偶数的时候，达到式子 (4.4) 上界的最优平衡 $(2,4,v)$ 集合码同时也是最优平衡 $(2,4,v+1)$ 集合码。因此我们只需要考虑 $v = 4m$ 和 $v = 4m+2$ 情形下的定理。为了方便起见，令 O_m 表示 $m \times m$ 的空表，对所有的表 W，令 W^T 表示表 W 的转置。

当 $v = 4m$ 的时候，目标为构造一个基于标号图 $G_+ = K_{m,m}$ 的平衡表 T_+，其中标号图的顶点集为 $P_+ = [2m+1, 4m]$ 且每条边的标号为一个 $P_- = [2m]$ 上的 2 元集。根据性质 (T1) 和 (T2)，该表 T_+ 是一个大小为 m^2 的 $2m \times 2m$ 表，且形如 $\begin{bmatrix} O_m & W; & W^T & O_m \end{bmatrix}$，其中 W 是用 $[2m]$ 中不同二元对填充的 $m \times m$ 表，使得每行的所有格子构成了 $[2m]$ 中的完美匹配。这恰好是一个 $H(m, 2m)$ Howell 设计。当 $m \neq 2, 6$ 的时候，定理 4.2 保证了该 Howell 设计的存在性。由此提供了所有

形如 $v=4m$ 或者 $4m+1$ 但是除了 $v\in\{8,9,24,25\}$ 情形的最优平衡 $(2,4,v)$ 集合码。

当 $v=4m+2$ 的时候，目标为构造一个大小为 m^2+m 平衡表 $\boldsymbol{T}_+$。它是基于如下的图 $\boldsymbol{G}_+=\boldsymbol{K}_{m+1,m+1}-\boldsymbol{M}$，其中顶点集为 $\boldsymbol{P}_+=[2m+1,4m+2]$ 且每条边的标号是一个 $\boldsymbol{P}_-=[2m]$ 上的 2 元集，$\boldsymbol{M}$ 是完全图 $\boldsymbol{K}_{m+1,m+1}$ 中的一个完美匹配。则该表 $\boldsymbol{T}_+$ 是一个 $(2m+2)\times(2m+2)$ 表，其形如 $=\begin{bmatrix}\boldsymbol{O}_{m+1} & \boldsymbol{W};\ \boldsymbol{W}^T & \boldsymbol{O}_{m+1}\end{bmatrix}$，其中 $\boldsymbol{W}$ 是一个 $(m+1)\times(m+1)$ 的表，使得每行恰有一个空的格子且其他的 m 个格子形成 $[2m]$ 中的完美匹配。进一步地，主对角线以上的非空格子的元素都是互不相同。这意味着表 $\boldsymbol{W}$ 的确是一个 $\boldsymbol{H}(m+1,2m)$ Howell 设计。当 $m\notin\{1,2,4\}$ 的时候，定理 4.3 提供了该 Howell 设计的存在性。由此构造出所有形如 $v=4m+2$ 或者 $4m+3$ 但是除了 $v\in\{6,7,10,11,18,19\}$ 情形的最优平衡 $(2,4,v)$ 集合码。

现在我们解决其他额外的情形，当 $v\in\{7,18,24\}$ 的时候，附录 4.6 直接给出了达到式子 (4.4) 上界的最优平衡 $(2,4,v)$ 集合码，并根据上述的性质同时解决了 $v=19$ 和 25 的情形。最后剩下的唯一情形为 $v=11$，在任何的最优平衡 $(2,4,12)$ 集合码中删去点集 $[12]$ 中的任意一个顶点和所有包含该点的区组，剩下的集合码则是最优平衡 $(2,4,11)$ 集合码。事实上对所有的偶数 v，我们都可以从最优平衡 $(2,4,v)$ 集合码中删去任意一个顶点和所有包含该点的区组来获得最优平衡 $(2,4,v-1)$ 集合码。 ∎

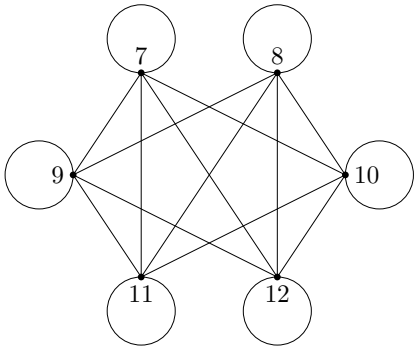
4.3.2　参数为 $(t,k)=(2,3)$ 的最优平衡集合码

本小节给出最优平衡 $(2,3,v)$ 集合码的新构造，它与文献[30] 的方法本质不同。我们将构造出所有最优平衡 $(2,3,v)$ 集合码都会达到式子 (4.3) 的上界 $A(2,3,v)\leqslant\frac{\lfloor v/2\rfloor\lceil v/2\rceil}{2}$。更确切地说：

$$A(2,3,v)\leqslant\begin{cases}2m^2, & \text{如果 } v=4m,\ m\geqslant 2,\\ 2m^2+m, & \text{如果 } v=4m+1,\ m\geqslant 2,\\ 2m^2+2m, & \text{如果 } v=4m+2,\ m\geqslant 1,\\ (2m+1)(m+1), & \text{如果 } v=4m+3,\ m\geqslant 0。\end{cases}\tag{4.5}$$

我们可以轻易地验证 $A(2,3,4)=1$ 和 $A(2,3,5)=2$。证明中的构造方法与小节 4.3.1 相似，但是由于现在的区组大小是 3，点集为 $\boldsymbol{P}_+$ 的标号图 $\boldsymbol{G}_+$ 在 $|\boldsymbol{B}_+|=1$ 的时候包含环，在 $|\boldsymbol{B}_+|=2$ 的时候包含通常的边。且这些边对应的标号分别为 $\boldsymbol{P}_-$ 上的 2 元集和 1 元集。进一步地，图 $\boldsymbol{G}_+$ 中允许包含重环。那么平衡 $(2,3,v)$ 集合码就等价于环图 $\boldsymbol{G}_+$ 同时满足：(1) 如果两条边 (或者两个环，或者一条边一个环) 相邻，则它们的标号互不相交；(2) 所有大小为 2 的标号都是互不相同的。

例如，式子 (4.2) 中的平衡 $(2,3,12)$ 集合码 $\boldsymbol{S}_3$ 等价于如下的标号图 $\boldsymbol{G}_3$。



$$
\begin{array}{ll}
L(\{7\})=\{1,2\} & L(\{7,12\})=\{6\} \\
L(\{8\})=\{2,3\} & L(\{8,9\})=\{4\} \\
L(\{9\})=\{5,6\} & L(\{8,10\})=\{5\} \\
L(\{10\})=\{1,6\} & L(\{8,11\})=\{6\} \\
L(\{11\})=\{3,4\} & L(\{8,12\})=\{1\} \\
L(\{12\})=\{4,5\} & L(\{9,11\})=\{1\} \\
L(\{7,9\})=\{3\} & L(\{9,12\})=\{2\} \\
L(\{7,10\})=\{4\} & L(\{10,11\})=\{2\} \\
L(\{7,11\})=\{5\} & L(\{10,12\})=\{3\}
\end{array}
$$

图 4.2　集合码 S_3 对应的标号图 G_3 及其标号 L

那么该等价的对称平衡表 $\boldsymbol{T}_+$ 有如下更简单的描述：

(S1) 对角线格子上是空的或者是一组 2 元集，每个非对角格子是空的或者是一个 1 元集；

(S2) 对角线格子上的所有 2 元集都是不同的；

(S3) $\boldsymbol{P}_-$ 的每个点在每行或者每列上最多出现一次。

例如，基于标号图 $\boldsymbol{G}_3$ 的平衡表 $\boldsymbol{T}_3$ 如下。表 $\boldsymbol{T}_3$ 可以视为从循环表中移动某些元素至对角格子获得而来的。我们将该想法拓展至接下来的构造中。

表 4.2　基于标号图 G_3 的平衡表 T_3，其中表的第一行和第一列是表格的行列标号，并无实际作用

	7	8	9	10	11	12
7	{1,2}		{3}	{4}	{5}	{6}
8		{2,3}	{4}	{5}	{6}	{1}
9	{3}	{4}	{5,6}		{1}	{2}
10	{4}	{5}		{1,6}	{2}	{3}
11	{5}	{6}	{1}	{2}	{3,4}	
12	{6}	{1}	{2}	{3}		{4,5}

定理 4.6　对于除了 $v \in \{4,5\}$ 的其他任意 $v \geqslant 3$，平衡 $(2,3,v)$ 集合码的上界 (4.5) 是紧的。

证明　我们将根据 v 模 4 划分成四种情形。当 $v = 8,9$ 的时候，文献[30] 给出了相关的结果。

当 $v = 4m$ 且 $m \geqslant 3$ 的时候，目标是寻找一个平衡表 $\boldsymbol{T}_+$，该表基于一个顶点集为 $\boldsymbol{P}_+ = [2m+1, 4m]$ 的标号图 $\boldsymbol{G}_+$。这里图 $\boldsymbol{G}_+$ 是 $\boldsymbol{K}_{2m} - \boldsymbol{M}$ 与所有 $\boldsymbol{P}_+$ 上环的并，其中 $\boldsymbol{M}$ 是一个完美匹配。因此图 $\boldsymbol{G}_+$ 总共有 $2m^2$ 条边。令 $\boldsymbol{T}_0$ 是一个首行为 $(1,2,\cdots,2m)$ 且每次向左平移一个位置而来的 $2m \times 2m$ 循环表。因此，$\boldsymbol{T}_0$ 是一个

在 $[2m]$ 上的拉丁方，其中 $T_0(i,j) \equiv i+j-1 \pmod{2m}$。注意到 T_0 的项都限制在集合 $[2m]$ 中。下面将修改表 T_0 来获得满足 (S1)-(S3) 性质的表 T_+。这个修改过程将分为两种情形。

情形一：m 为奇数时。对所有 $i \in [2m]$ 的奇数，项 $T_0(i,i+1)$ 被移动到 (i,i) 格子中。对所有 $i \in [2m]$ 的偶数，项 $T_0(i,i-1)$ 被移动到 (i,i) 格子中。由此得到期望的表 T_+。例如当 $m=3$ 的时候，表 4.2 就是一个例子。表 T_+ 显然还是对称的，且每个对角格子中包含一个 2 元集，所有非对角格子是空的或者是一个 1 元集同时 $[2m]$ 的每个元素恰好出现在每行一次。进一步地，表 T 的大小为 $2m+\frac{2m(2m-2)}{2}=2m^2$。因此定理只要充分地证明 T_+ 所有对角格子的项互不相同，那么就构造出最优平衡集合码。其中奇数行中对角格子的项为 $\mathcal{O}=\{\{4i-3,4i-2\} \pmod{2m} : 1 \leqslant i \leqslant m\}$，偶数行中对角格子的项为 $\mathcal{E}=\{\{4j-2,4j-1\} \pmod{2m} : 1 \leqslant j \leqslant m\}$。因为 m 是奇数，所以接下来的关系成立：如果 $4i-2=4j-2 \pmod{2m}$，则 $i=j$。因此这些 2 元集互不相同。定理得证。

情形二：m 为偶数时。对所有的 $1 \leqslant i \leqslant m-1$，项 $T_0(i,i+m-1)$ 被移动至 (i,i) 格子中。对所有的 $m \leqslant i \leqslant 2m-2$，项 $T_0(i,i-m+1)$ 被移动至 (i,i) 格子中。最后项 $T_0(2m-1,2m)$ 被移动至 $(2m-1,2m-1)$ 格子中，项 $T_0(2m,2m-1)$ 被移动至 $(2m,2m)$ 格子中。通过上面的方法得到期望的表 T_+。与情形一类似，定理只需要证明表 T_+ 中对角格子的项互不相同。其中前 $m-1$ 行中对角格子的项为 $A=\{\{2i-1,2i+m-2\} \pmod{2m} : 1 \leqslant i \leqslant m-1\}$，紧接后 $m-1$ 行中对角格子的项为 $B=\{\{2j-1,2j-m\} \pmod{2m} : m \leqslant j \leqslant 2m-2\}$ 以及最后两行对角格子项为 $C=\{\{2m-3,2m-2\},\{2m-2,2m-1\}\}$。显然 A 或者 B 中内部两个 2 元集互不相同。因为 C 的 2 元集的差异都为 1，但是 A 或者 B 中的 2 元集的差异为 $m-1 \geqslant 2$，所以集合 C 与集合 A 和集合 B 都不相交。最后，假设 $A \cap B \neq \varnothing$。那么存在某个 $1 \leqslant i \leqslant m-1$ 和某个 $m \leqslant j \leqslant 2m-2$ 使得 $2i-1 \equiv 2j-1 \pmod{2m}$ 和 $2i+m-2 \equiv 2j-m \pmod{2m}$。这是不能同时成立的，因此该假设不成立。定理得证。

综上两种情形，我们就找到 $v=4m$ 且 $m \geqslant 3$ 的最优平衡 $(2,3,v)$ 集合码。

当 $v=4m+1$ 且 $m \geqslant 3$ 的时候，目标是寻找一个平衡表 T_+，该表基于一个顶点集为 $P_+=[2m,4m+1]$ 的图 G_+，其中图 G_+ 是完全图 K_{2m} 和所有环的并。因此 G_+ 总共有 $2m^2+m$ 条边。平衡表 T_+ 则是通过在 $v=4m$ 的平衡表中添加符号 $2m+1$ 到每个空格子而成的。平衡表 T_+ 对应的码显然是最优平衡 $(2,3,v)$ 集合码。

当 $v=4m+3$ 的时候，目标是在寻找一个 $(2m+1)\times(2m+1)$ 的平衡表 T_+，该表基于一个顶点集为 $P_+=[2m+3,4m+3]$ 的图 G。该图只包含重环不包含通常的边。令 $P_-=[2m+2]$ 且 $M_1,M_2,\cdots,M_{2m+1}$ 是 P_- 的 1 因子。则表 T_+ 中第

(i,i) 位置的对角格子填充完美匹配 M_i 中的所有 2 元集且其他非对角格子都是空的。表 T_+ 显然满足性质 (S1)-(S3)。这就找到了大小为 $(2m+1)(m+1)$ 的最优平衡 $(2,3,v)$ 集合码。

当 $v=4m+2$ 的时候，从最优平衡 $(2,3,v+1)$ 集合码中删去一个 P_+ 的点和所有包含该点的区组，该方法可以找到最优平衡 $(2,3,v)$ 集合码。 ∎

4.3.3　参数为 $(t,k)=(3,4)$ 的最优平衡集合码

本小节构造最优平衡 $(3,4,v)$ 集合码。根据定义知道标号图的每条边和其对应的标号都是一个 2 元集，但是图中可能包含重边。为了方便起见，我们将相同位置的重边视为简单图中的边 e，并定义边 e 上的标号为重图中 e 上的所有标号并。也就是说，$L(e):=\{B_-:e\cup B_-\in\mathcal{S}\}\subseteq\binom{P_-}{2}$。则平衡 $(2,4,v)$ 集合码等价于一个顶点集为 P_+ 的简单图 G_+。同时满足如下的性质：

(Q1) 对任意的边 e，对应的标号 $L(e)$ 是一个 P_- 中的匹配；

(Q2) 如果两条边 e_1 和 e_2 相邻，则 $L(e_1)\cap L(e_2)=\varnothing$。

基于此标号，$|\mathcal{S}|=\sum_{e\in E(G_+)}|L(e)|$。

如下的命题稍微地改进了式子 (4.3) 中的上界。

命题 4.7

$$A(3,4,v)\leqslant\begin{cases}2m^3-m^2, & \text{如果 } v=4m,4m+1,\ m\geqslant 1,\\ 2m^3+m^2, & \text{如果 } v=4m+2,\ m\geqslant 1,\\ 2m^3+3m^2+m, & \text{如果 } v=4m+3,\ m\geqslant 1。\end{cases}\tag{4.6}$$

证明　由于 $v=4m+3$ 的情形和式子 (4.3) 的值是一致的，因此命题只需要检验 $v=4m+1$ 和 $4m+2$ 的情形。

假设标号图 G_+ 和 G_- 是从平衡 $(3,4,v)$ 集合码 $\mathcal{S}$ 获得而来的，那么根据性质 (Q1)，对任意的 $e\in E(G_+)$, $|L(e)|\leqslant\lfloor p_-/2\rfloor$。相似地，对任意的 $e\in E(G_-)$, $|L(e)|\leqslant\lfloor p_+/2\rfloor$。则 $|\mathcal{S}|\leqslant\min\left\{\binom{p_+}{2}\lfloor p_-/2\rfloor,\binom{p_-}{2}\lfloor p_+/2\rfloor\right\}$。因此 $A(3,4,v)\leqslant\max_{p_+,p_-}\min\left\{\binom{p_+}{2}\lfloor p_-/2\rfloor,\binom{p_-}{2}\lfloor p_+/2\rfloor\right\}$。通过选取 $v=p_++p_-=4m+1$ 或者 $4m+2$，我们便得到命题结论。这里省略了简单地计算。 ∎

定理 4.8　对所有 $v\geqslant 5$ 的情形，平衡 $(3,4,v)$ 集合码的上界 (4.6) 都是紧的。

证明　当 $v=4m$ 的时候，目标是构造一个顶点集为 $P_+=[2m+1,4m]$ 的图 $G_+=K_{2m}$，其中图中每条边的标号是一个 $P_-=[2m]$ 中的完美匹配，且对任意两条相邻的边 e_1,e_2, $L(e_1)\cap L(e_2)=\varnothing$。该标号图可以得到一个大小为 $m\binom{2m}{2}=2m^3-m^2$ 的集合码。由于 $P_-=[2m]$ 有 $2m-1$ 个不交的完美匹配。假设其中的每个完美匹配视为一种颜色，那么上面的要求等价于寻找图 G_+ 的 $(2m-1)$

边染色。著名的 Vizing 定理可以找到该边染色，这同时也找到了 $v = 4m + 1$ 的最优平衡集合码。

当 $v = 4m+2$ 和 $4m+3$ 的时候，构造方式类似于 $v = 4m$ 的方法。当 $v = 4m+2$ 的时候，我们把 $G_+ = K_{2m+1}$ 的每条边用 $P_- = [2m + 1]$ 中的一个大小为 m 的匹配标记，且该标记满足性质 (Q2)。这就找到了大小为 $m\binom{2m+1}{2} = 2m^3 + m^2$ 的平衡集合码。因为 $[2m + 1]$ 有 $2m + 1$ 个大小为 m 的匹配并且这些匹配是边不交的。上面的要求等价于寻找完全图 K_{2m+1} 的 $(2m + 1)$ 边染色，同样利用 Vizing 定理可以保证找到该边染色。当 $v = 4m + 3$ 的时候，对应的图为 $G_+ = K_{2m+1}$ 且标号集 $P_- = [2m + 2]$。图 G_+ 的每条边用 P_- 的一个完美匹配标记。因为 $[2m + 2]$ 有 $2m + 1$ 个边不交的完美匹配，我们只要寻找图 K_{2m+1} 的 $(2m + 1)$ 边染色。Vizing 定理同样保证了该边染色存在。这就找到了大小为 $(m+1)\binom{2m+1}{2} = 2m^3+3m^2+m$ 的平衡集合码。 ∎

4.4　大 k 下的最优和渐近最优平衡集合码

本小节处理大 k 下的平衡 (t, k, v) 集合码，并提供了两类渐近最优的平衡集合码。当 $k \geqslant 4$ 为偶数的时候，我们可以从正交阵列中构建一些最优平衡 $(2, k, v)$ 集合码。进一步地，该构造给出了渐近最优的平衡 $(2, k, v)$ 集合码。此外，对所有的 $t \geqslant 2$，Rödl Nibble 方法保障了渐近最优平衡 $(t, t + 1, v)$ 集合码的存在性。

4.4.1　参数为 $t = 2$ 以及偶数 k 的渐近最优平衡集合码

对任意的偶数 $k \geqslant 4$，本小节通过稍微修改正交阵列 $OA(k, m)$ 的每行以及选取一个合适的标号得到了最优的平衡 $(2, k, km)$ 集合码。根据正交阵列的构造得到一组渐近最优平衡 $(2, k, v)$ 集合码。

定理 4.9　令 $k \geqslant 4$ 为偶数。假设存在一个正交阵列 $OA(k, m)$，则存在一个最优平衡 $(2, k, km)$ 集合码。

证明　根据上界 (4.3) 得到 $A(2, k, km) \leqslant m^2$。令 $A = (a_{ij})_{m^2\times k}$ 是一个正交阵列 $OA(k, m)$ 其中对任意 $i \in [m^2]$ 均有 $a_{ij} \in [m]$。正交阵列 A 的第 i 行可以用来定义一个区组 $B_i = \{a_{i1}, a_{i2} + m, \cdots, a_{ik} + (k - 1)m\} \subseteq [km]$。根据正交阵列的定义知道对任意的 $i \neq j$，B_i 与 B_j 至多相交一个元素。从而区组集合 $\{B_i : i \in [m^2]\}$ 形成 $[km]$ 上的 $(2, k, km)$ 集合码。进一步地，集合 $[km/2]$ 中的每个元素标记为 $+1$ 同时集合 $[km/2 + 1, km]$ 中的每个元素标记为 -1。因此该方法找到了最优平衡 $(2, k, km)$ 集合码。 ∎

我们知道 $(k, 2, k - 1)_m$ MDS 码是正交 $OA(k, m)$。此外对任意的整数 $2 \leqslant k$ 和 $k \leqslant m$ 的素幂 m，广义 Reed–Solomon 码[81] 是 $(k, 2, k - 1)_m$ MDS 码，因此我们

有如下的渐近结果。

定理 4.10　对任意的整数 $2 \leqslant k$ 和 $k \leqslant m$ 的素幂 m，存在一个最优平衡 $(2, k, km)$ 集合码。同时平衡 $(2, k, v)$ 集合码的上界 (4.3) 是渐近紧的。

证明　素数稠密性保证了对任意的实数 $\frac{v}{k}$ 都存在一个素数 m，使得 $(1 - o(1))\frac{v}{k} \leqslant m \leqslant \frac{v}{k}$。定理 4.9 知道存在一个大小为 m^2 的平衡 $(2, k, km)$ 集合码。同时 $A(2, k, v) \geqslant m^2 \geqslant (1 - o(1))(\frac{v}{k})^2$，也即 $\lim\limits_{v\to\infty} A(2, k, v) \geqslant (\frac{v}{k})^2$。这意味着上界 (4.3) 是渐近紧的。 ∎

4.4.2　参数为 $k = t + 1$ 的渐近最优平衡集合码

本小节主要关注平衡 $(t, t + 1, v)$ 集合码。正如本章预备知识章节中所述，一个 (t, k, v) 集合码恰是一个 t 填充。我们在构造平衡集合码之前先介绍 t 填充设计理论中最著名的工作：Rödl Nibble 方法。我们为寻找平衡集合码首先将平衡集合码和辅助超图的匹配建立了等价关系，然后利用著名的概率方法：Rödl Nibble 方法 (特殊的一致超图中存在一个近似的完美匹配) 找到渐近最优平衡 $(t, t + 1, v)$ 集合码。

对整数 $0 < t < k < v$，令 $m(t, k, v)$ 表示 (t, k, v) 集合码的最大区组数。Erdős 和 Hanani 猜测对任意 $t < k$ 均有 $\lim\limits_{v\to\infty} \frac{m(t,k,v)}{\binom{v}{t}/\binom{k}{t}} = 1$。不久后，Rödl 创造的 Nibble 方法[82] 肯定了此猜想。

现在简单介绍杰出的 Rödl Nibble 方法。

定理 4.11　[83] 对所有的 $r \geqslant 2$ 和实数 $k \geqslant 1$ 和 $a > 0$，存在实数 $\gamma = \gamma(r, k, a) > 0$ 和整数 $d_0 = d_0(r, k, a)$ 使得对任意的 $n \geqslant D \geqslant d_0$，下面的结论成立。

在任意 V 上 r 一致超图 $\mathcal{H}$ 中，$|V| = n$ 且超图每个顶点点度都是正整数，同时满足下面的条件：

(1) 除了至多 γn 个点其余所有顶点 x，$d(x) = (1 \pm \gamma)D$；

(2) 对所有的 $x \in V, d(x) < kD$；

(3) 对任意两个不同的 $x, y \in V, d(x, y) < \gamma D$，

则超图包含一个大小至多为 $(1 + a)n/r$ 的覆盖。

特别地，定理中的 $\gamma \leqslant \epsilon^{r-1}\delta$，其中 ϵ 和 $\delta < 1/10$ 是满足如下条件的正实数：

$$\begin{cases} \frac{\epsilon}{1-e^{-\epsilon}} + r\epsilon < 1 + a, \\ (1 + 4\delta)\frac{\epsilon}{1-e^{-\epsilon}} + r\epsilon < 1 + a。 \end{cases} \tag{4.7}$$

我们通过选取一些特殊参数获得如下结论。该结论意味着对任何共度较小、顶点数充分大的近似正则一致超图，超图都有一个近似的完美匹配。

引理 4.12　对任意的整数 $r \geqslant 2$ 和满足条件 $\gamma^{1/r} < (1-e^{-\gamma^{1/r}})(1+\gamma^{1/r})$ 的正实数 $\gamma < 10^{-r}$，令 $\mathcal{H}$ 是一个顶点集大小为 $|V(\mathcal{H})| = n$ 的 r 一致超图 $V(\mathcal{H})$，同时存在正实数 D_0 以及 D，满足对任意的 $x \in V(\mathcal{H})$，均有 $(1-\gamma)D \leqslant d(x) \leqslant (1+\gamma)D$，同时对任意不同的顶点 $x, y \in V(\mathcal{H})$，均有 $d(x,y) \leqslant \gamma D$。则对任意的 $n \geqslant D \geqslant D_0$，超图 $\mathcal{H}$ 包含一个大小至少为 $n/r - 2an$ 的匹配，其中 $a = (r+8)\gamma^{1/r}$。

证明　令 ϵ 和 $\delta < 1/10$ 是两个满足关系 $\epsilon = \delta = \gamma^{1/r}$ 的正实数，也即 $\gamma = \epsilon^{r-1}\delta$ 且 $\frac{\epsilon}{1-e^{-\epsilon}} \leqslant 1+\epsilon$。这意味着参数 ϵ, δ 和 a 满足式子 (4.7) 的要求。根据定理 4.11，超图 $\mathcal{H}$ 有一个大小至多为 $(1+a)n/r$ 的覆盖 C。

对任意的顶点 $x \in V(\mathcal{H})$，令 $v(x) := |\{e \in C : x \in e\}|$ 以及 $X_1 := \{x \in V(\mathcal{H}) : v(x) = 1\}$。于是 $\sum_{x \in V(\mathcal{H})} v(x) \leqslant (1+a)n$ 且

$$\begin{aligned}\sum_{x \in V(\mathcal{H})} v(x) &= \sum_{x \in V(\mathcal{H}), v(x)=1} v(x) + \sum_{x \in V(\mathcal{H}), v(x) \geqslant 2}, v(x) \\ &\geqslant |X_1| + 2(n - |X_1|) = 2n - |X_1|。\end{aligned}$$

因此 $|X_1| \geqslant (1-a)n$ 且 $\sum_{x \in V(\mathcal{H}), v(x) \geqslant 2} v(x) \leqslant 2an$。通过删去覆盖 C 中所有 $v(x) \geqslant 2$ 的顶点 x，超图中便有一个至少 $(1+a)n/r - 2an > n/r - 2an$ 条边的匹配。 ∎

定理 4.13　令 $t \geqslant 2$ 为正整数，下面渐近关系成立

$$\lim_{v \to \infty} \frac{A(t, t+1, v)\binom{\lceil (t+1)/2 \rceil}{\lfloor t/2 \rfloor}\binom{\lfloor (t+1)/2 \rfloor}{\lceil t/2 \rceil}}{\binom{\lceil v/2 \rceil}{\lfloor t/2 \rfloor}\binom{\lfloor v/2 \rfloor}{\lceil t/2 \rceil}} = 1。$$

证明　首先，我们提出一个特殊一致超图 $\mathcal{H}$，并证明平衡集合码和 $\mathcal{H}$ 的匹配是等价的。

令 P_+, P_- 为点集，我们构造一个顶点集 $V = \bigcup_{i \in \mathcal{I}} \binom{P_+}{i} \times \binom{P_-}{t-i}$ 的辅助超图 $\mathcal{H} = (V, E)$，其中 $\mathcal{I}$ 将在后面确定出来。令 $\mathcal{B}$ 是一组 k 元集，使得对任意的 $B \in \mathcal{B}$，$B = B_+ \cup B_-$，其中 $B_+ = B \cap P_+, B_- = B \cap P_-$ 同时 $|B_+| = \lceil k/2 \rceil, |B_-| = \lfloor k/2 \rfloor$。对任意的 $B \in \mathcal{B}$，它对应于一条超边 $e_B = \{\binom{B_+}{i} \times \binom{B_-}{t-i} : i \in \mathcal{I}\}$。这里的指标集 $\mathcal{I} = \{0 \leqslant i \leqslant t : \binom{|B_+|}{i} \times \binom{|B_-|}{t-i} \neq 0\}$。

断言： $\mathcal{H}$ 的匹配 $\mathcal{M}$ 对应大小为 $|\mathcal{M}|$ 的平衡 (t,k,v) 集合码。

假设 $\mathcal{M} = \{e_{B_1}, e_{B_2}, \cdots, e_{B_m}\}$，令 $S = \{B_1, B_2, \cdots, B_m\}$。显然 S 中的集合是平衡的。如果我们能证明 S 中的任意两个不同的区组相交至多 $t-1$ 个点，则 S 是一个 (t,k,v) 集合码。现在假设存在两个不同的区组 B_i 和 B_j 使得 $|B_i \cap B_j| \geqslant t$，则存在一个 t 元集 $T = T_+ \cup T_- \subseteq B_i \cap B_j$ 其中 $T_+ = T \cap P_+, T_- = T \cap P_-$。然而 $T_+ \times T_-$ 是超图 $\mathcal{H}$ 中的一个顶点，但是 $T_+ \times T_- \in e_{B_i}, T_+ \times T_- \in e_{B_j}$。这与 $\mathcal{M}$ 是匹配矛盾。因此断言成立。

对于平衡 $(t, t+1, v)$ 集合码，这里选取参数 $k = t+1$，$|P_+| = \lceil v/2 \rceil, |P_-| = \lfloor v/2 \rfloor$ 以及指标集 $\mathcal{I} = \{\lceil \frac{t+1}{2} \rceil - 1, \lceil \frac{t+1}{2} \rceil\}$。则辅助超图 $\mathcal{H} = (V, E)$ 的顶点集

$V=\bigcup_{i\in\mathcal{I}}\binom{P_+}{i}\times\binom{P_-}{t-i}$ 边集 $E=\{e_B:B\in\mathcal{B}\}$。则超图 $\mathcal{H}$ 是 n 个顶点的 r 一致超图，其中

$$n=\binom{\lceil v/2\rceil}{\lceil (t+1)/2\rceil-1}\binom{\lfloor v/2\rfloor}{\lfloor (t+1)/2\rfloor}+\binom{\lceil v/2\rceil}{\lceil (t+1)/2\rceil}\binom{\lfloor v/2\rfloor}{\lfloor (t+1)/2\rfloor-1},$$

$$r=\binom{\lceil k/2\rceil}{\lceil (t+1)/2\rceil-1}\binom{\lfloor k/2\rfloor}{\lfloor (t+1)/2\rfloor}+\binom{\lceil k/2\rceil}{\lceil (t+1)/2\rceil}\binom{\lfloor k/2\rfloor}{\lfloor (t+1)/2\rfloor-1}=t+1。$$

令 $X=X_+\cup X_-$ 是超图 $\mathcal{H}$ 中的一个顶点，则点度 $d(X)=|\{e_B\in E:X\in e_B\}|=|\{B\in\mathcal{B}:X\subseteq B\}|=\binom{|P_+|-|X_+|}{\lceil k/2\rceil-|X_+|}\binom{|P_-|-|X_-|}{\lfloor k/2\rfloor-|X_-|}$，更确切说

$$d(X)=\begin{cases}\lceil v/2\rceil-\lceil (t+1)/2\rceil+1, & \text{如果 } |X_+|=\lceil\frac{t+1}{2}\rceil-1,\\ \lfloor v/2\rfloor-\lfloor (t+1)/2\rfloor+1, & \text{如果 } |X_+|=\lceil\frac{t+1}{2}\rceil。\end{cases}\tag{4.8}$$

令 $X,Y\in V$ 是超图中两个不同的顶点，则 $d(X,Y)=|\{e_B\in E:X,Y\in e_B\}|=|\{B\in\mathcal{B}:S_X\cup S_Y\subseteq B\}|$，其中 S_X,S_Y 分别对应于顶点 X 和 Y 所对应的 t 元集。因为 $|S_X\cup S_Y|\geqslant t+1=k$ 所以 $d(X,Y)=0$ 或者 1。

当顶点数 v 充分大的时候，这里选取一个充分小的 $\epsilon>0$，整数 $D=\lceil v/2\rceil-\lceil (t+1)/2\rceil+1$ 以及 $\gamma=4/D$。则 $\gamma<10^{-t-1}$ 且 $\gamma^{1/(t+1)}<(1-e^{-\gamma^{1/(t+1)}})(1+\gamma^{1/(t+1)})$。根据引理 4.12，超图 $\mathcal{H}$ 有一个大小至少为 $n/r-2an$ 的匹配 $\mathcal{M}$，其中 v 趋向无穷大的时候，我们有一个大小为 $(1-o(1))n/r$ 的平衡集合码 S。

当 t 是偶数时，不妨设 $t=2s$，此时平衡集合码的上界 (4.3) 为 $\frac{(v/2)^{2s}}{s!(s+1)!}+o(v^{2s})$。同时构造出的平衡集合码 S 大小为 $(1-o(1))\frac{n}{r}$ 其中 $n=\frac{2s+1}{s!(s+1)!}(v/2)^{2s}+o(v^{2s})$ 和 $r=2s+1$，也即找到一个大小为 $\frac{(v/2)^{2s}}{s!(s+1)!}+o(v^{2s})$ 的平衡集合码。

当 t 是奇数时，不妨设 $t=2s+1$，此时平衡集合码的上界 (4.3) 为 $\frac{(v/2)^{2s+1}}{(s+1)!(s+1)!}+o(v^{2s+1})$。同时构造的平衡集合码 S 大小为 $(1-o(1))\frac{n}{r}$ 其中 $n=2\frac{(v/2)^{2s+1}}{s!(s+1)!}+o(v^{2s+1})$ 和 $r=2(s+1)$，也即找到一个大小为 $\frac{(v/2)^{2s+1}}{(s+1)!(s+1)!}+o(v^{2s+1})$ 的平衡集合码。

因此，$A(t,t+1,v)$ 是渐近紧的。 ∎

4.5　小结

当 $t=2,k=3,4$ 的时候，我们为最优平衡 (t,k,v) 集合码和图标号问题建立起了等价关系并简练地给出达到上界 (4.3) 的平衡集合码确切构造。事实上，当 $t=2$ 的时候，它等价对应于一个超图标号问题。这里我们对奇数 k 描述该超图标号问题。

问题 4.14　正点集为 P_+ 负点集为 P_- 的平衡 $S(2,k,v)$ 集合码等价于有 $|S(2,k,v)|$ 条边的线性超图 $G=(V(G),E(G))$，使得 $V(G)=P_+$ 且所有的超边大小为 $\lceil k/2\rceil$ 或者 $\lfloor k/2\rfloor$ 以及超边标号 $L:E(G)\to\binom{[P_-]}{\lceil k/2\rceil}\cup\binom{[P_-]}{\lfloor k/2\rfloor}$ 满足：

(1) 如果两条超边 e_1, e_2 相交一个顶点，则标号 $L(e_1)$ 和 $L(e_2)$ 互不相交；

(2) 如果两条超边 e_1, e_2 不相交，则标号 $L(e_1)$ 和 $L(e_2)$ 相交至多一个元素。

找到满足上述标号的最大超图将会是一件十分有意义的事情。此外，我们利用边染色理论找到了最优平衡 $(3,4,v)$ 集合码，并利用正交阵列找到渐近最优平衡 $(2,k,v)$ 集合码以及利用概率方法找到渐近最优平衡 $(t,t+1,v)$ 集合码。

4.6　附录

下面列举出 $v \in \{6,7,8,9,10,18,23,24,25\}$ 的最优平衡 $(2,4,v)$ 集合码。

表 4.3　例外情形下的最优平衡 $(2,4,v)$ 集合码，其中 $[1,\lfloor v/2\rfloor]$ 总是标记为 -1 同时 $[\lfloor v/2\rfloor+1,v]$ 总是标记为 $+1$

v	最优平衡 $(2,4,v)$ 集合码例子				
6	{1,2,4,5}				
7	{1,2,4,5}	{1,3,6,7}			
8	{1,2,5,6}	{1,3,7,8}			
9	{1,2,5,6}	{2,3,7,8}	{3,4,5,9}		
10	{1,3,8,9}	{2,4,8,10}	{3,5,6,10}	{4,5,7,9}	{1,2,6,7}
18	{6,8,9,16}	{2,5,9,13}	{3,8,11,13}	{5,6,11,14}	{1,5,10,15}
	{2,6,10,12}	{1,6,13,18}	{4,6,15,17}	{3,5,12,16}	{3,4,9,18}
	{1,7,9,14}	{4,7,13,16}	{4,8,10,14}	{1,4,11,12}	{1,2,16,17}
	{2,3,14,15}	{2,7,11,18}	{5,8,17,18}	{7,8,12,15}	{3,7,10,17}
23	{7,11,17,23}	{3,6,17,19}	{4,6,13,15}	{4,9,12,23}	{5,6,12,14}
	{5,8,20,22}	{6,9,18,22}	{4,5,16,21}	{7,9,13,14}	{8,10,13,17}
	{8,11,14,18}	{7,8,12,15}	{3,7,18,21}	{7,10,16,20}	{6,8,16,23}
	{5,11,13,19}	{1,4,17,20}	{1,11,15,16}	{1,2,13,23}	{1,10,12,18}
	{3,10,15,23}	{3,4,14,22}	{2,6,20,21}	{10,11,21,22}	{2,9,16,17}
	{2,5,15,18}	{2,10,14,19}	{3,11,12,20}	{8,9,19,21}	{1,7,19,22}
24	{5,11,13,15}	{4,5,16,19}	{1,5,18,20}	{5,8,14,24}	{2,6,16,20}
	{2,9,17,23}	{8,11,20,22}	{6,7,13,23}	{8,10,13,18}	{10,11,16,21}
	{1,2,15,21}	{4,8,21,23}	{7,9,21,22}	{1,3,23,24}	{5,10,22,23}
	{4,9,13,24}	{2,4,14,18}	{1,6,14,22}	{6,12,19,21}	{3,5,17,21}
	{6,11,17,18}	{2,11,19,24}	{3,4,15,20}	{4,12,17,22}	{1,12,13,16}
	{7,8,15,16}	{3,9,14,16}	{9,10,19,20}	{7,12,20,24}	{2,3,13,22}
	{1,8,17,19}	{6,10,15,24}	{7,10,14,17}	{9,12,15,18}	{11,12,14,23}
	{3,7,18,19}				
25	{10,12,14,23}	{9,10,17,19}	{6,9,14,21}	{2,13,15,24}	{5,12,20,25}
	{2,5,17,23}	{3,6,16,17}	{6,8,18,24}	{1,2,16,22}	{4,11,18,21}
	{2,10,18,25}	{5,11,14,19}	{7,11,15,25}	{2,7,14,20}	{4,10,15,16}
	{3,5,21,24}	{4,12,19,24}	{1,10,20,24}	{1,13,17,18}	{4,13,20,23}
	{7,13,16,19}	{1,8,21,25}	{3,7,18,22}	{3,9,15,20}	{3,13,14,25}
	{4,8,14,17}	{3,8,19,23}	{4,9,22,25}	{7,12,17,21}	{8,12,15,22}
	{1,6,15,19}	{5,9,16,18}	{6,11,22,23}	{8,11,16,20}	{10,13,21,22}
	{7,9,23,24}				

第 5 章　其他在研问题

这一章将罗列一些在攻读博士学位期间所研究的其他极值组合方面的问题。这些课题虽然不在集中考虑存储系统中的问题，但是它们仍然属于极值组合以及相关应用中的问题。在此，我们简要地介绍这些研究课题。这些课题主要包括：有限域上的 Erdős-Falconer 距离问题，Hamilton 幂圈分解以及多访问的编码缓存方案。

5.1　有限域上的 Erdős-Falconer 距离问题

极值组合中著名的 Erdős-Falconer 距离问题旨在确定当一个集合多大的时候，该集合包含许多不同的距离。例如，如下经典的 Erdős-Falconer 猜想：

猜想 5.1　对任意的正整数 d，假设 $\mathbb{R}^d$ 中的集合 S 的 Hausdorff 维数超过 $\frac{d}{2}$，则 S 中包含所有不同距离的集合其 Lebesgue 测度是严格大于 0。

Falconer 首先证明当 $\mathbb{R}^d$ 的紧集 A 的 Hausdorff 维数超过 $\frac{d+1}{2}$ 时，A 中不同欧式距离的集合具有正的 Lebesgue 测度。近年来大量的研究工作极力解决该猜想。与经典的 Erdős-Falconer 问题不同，我们研究有限域上 Erdős-Falconer 型的汉明距离问题，在该研究中：考虑的对象为有限域上的集合，测度则采用汉明距离以及被度量的集合用距离图替换两点之间的距离。该问题为经典 Erdős-Falconer 的推广问题，这里采用图论的语言简述如下：

问题 5.2　对于给定的图 H，当有限域 $\mathbb{F}_q^n$ 中的集合 A 多大的时候可以保证 A 中包含正比例数目的等距 H? 其中 A 中的等距 H 指的是一个嵌入 $\phi : V(H) \to A$ 使得对图 H 中任意的边 uv，$\phi(u)$ 和 $\phi(v)$ 的汉明距离等于某个数 d_{uv}。

通过结合拓展的模版本 Delsarte 不等式和极值组合中著名的 dependent random choice 方法，我们解决了二部图 H 上的 Erdős-Falconer 型的汉明距离问题。主要的结论如下：

定理 5.3　对任意的素幂 q 和充分大的 n，令 H 是 turán 数为 $ex(n, H) = O(n)$ 的二部图，假设集合 $A \subseteq \mathbb{F}_q^n$ 满足对某个实数 $c_1 = c_1(q, H) > 0$，$|A| > q^{(1-c_1)n}$。则 A 包含 $\Omega(n)$ 个不同的等距 H，其中的每条边的汉明距离都是某个给定的自然数。

定理 5.4　对任意的素幂 q 和充分大的 n，令 H 是一个二部图，假设集合 $A \subseteq \mathbb{F}_q^n$ 满足对某个实数 $c_2 = c_2(q, H) > 0$，$|A| > q^{(1-c_2)n}$。则 A 包含 $\Omega(n)$ 个不同的等距 H。

5.2　Hamilton 幂圈分解

图分解的研究由来已久，大量的图论学者都致力解决该问题。具体来说，给定一个图 H，如果图 G 可以分解成若干个边不交的 H，则称图 G 有 H 分解。对于不同的 G 和 H 的图分解问题，学者们提供了许多著名的定理。例如：组合设计中的 Kirkman 定理，Wilson 定理，Keevash 关于 t 设计存在性定理。

针对完全图上的 Hamilton 圈分解问题，Walecki 率先给出如下结论。

定理 5.5　对任意的奇数 $n \geqslant 3$，完全图 K_n 都有 Hamilton 圈分解。

我们利用分解集研究完全图上的 Hamilton 幂圈分解问题，并获得如下 Hamilton 2 幂圈分解和更一般的 Hamilton k 幂圈分解结果。

定理 5.6　令 $k \geqslant 2$ 为正整数，$n = 2km + 1$ 为素数，假设存在一个完美 $B[-k, k](n)$ 分解集，则完全图 K_n 有 Hamilton k 幂圈分解。

推论 5.7　当素数 n 是满足 $n \equiv 1 \pmod 4$ 及 4 整除 2 在模 n 下的阶，则完全图 K_n 有 Hamilton 2 幂圈分解。

目前该课题仍处于研究阶段。

5.3　多访问编码缓存方案

作为 5G 通信中关键技术之一的缓存技术，缓存通过提前下载部分数据来减少网络高峰期的数据传输，从而大大地缓解网络拥塞。为此设计出各种高效的缓存方案变得尤为重要。在著名的 Maddah-Ali-Niesen 方案之后，大量的学者提出了许多的缓存方案。我们着眼考虑其中的多访问编码缓存方案，也即，多个缓存可以被同一用户访问且同一缓存可以被多个用户访问的方案。我们特别考虑 Rajan 等人提出的基于交叉可分解设计 (Cross Resolvable Design) 的多访问编码缓存方案。在同等数目和同种型号的缓存下，该方案相比较于传统的 Maddah-Ali-Niesen 方案具有更多服务用户数和更小分包数的优势。

对于该方案的核心构型：交叉可分解设计 (CRD)，我们的贡献是发现构型 CRD 等价于具有一定正则性质的 r 部 r 一致超图或者是固定强度的正交阵列。我们依据一些已有的超图构造以及正交阵列的构造，例如：完全 r 部 r 一致均衡超图，MDS 码等，获得了许多基于 CRD 的多访问编码缓存方案。目前该课题尚处于研究阶段。

参 考 文 献

[1] GERBNER D, KESZEGH B, METHUKU A, et al. Set systems related to a house allocation problem[J]. Discrete Mathematics, 2020, 343(7): 111886.

[2] BOLLOBÁS B. On generalized graphs[J]. Acta Mathematica Hungarica, 1965, 16(3-4): 447-452.

[3] LOVÁSZ L. Topological and algebraic methods in graph theory[C]//Graph theory and related topics (Proc. Conf., Univ. Waterloo, Waterloo, Ont., 1977). 1979: 1-14.

[4] SCOTT A, WILMER E. Combinatorics in the exterior algebra and the Bollobás two families theorem[J]. Journal of the London Mathematical Society, online, 2021.

[5] RASHMI K V, SHAH N B, KUMAR P V, et al. Explicit construction of optimal exact regenerating codes for distributed storage[C]//2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton). 2009: 1243-1249.

[6] SHAH N B, RASHMI K V, KUMAR P V, et al. Distributed storage codes with repair-by-transfer and nonachievability of interior points on the storage-bandwidth tradeoff[J]. IEEE Transactions on Information Theory, 2011, 58(3): 1837-1852.

[7] EL ROUAYHEB S, RAMCHANDRAN K. Fractional repetition codes for repair in distributed storage systems[C]//2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton). 2010: 1510-1517.

[8] DIMAKIS A G, GODFREY P B, WU Y, et al. Network coding for distributed storage systems [J]. IEEE transactions on information theory, 2010, 56(9): 4539-4551.

[9] OLMEZ O, RAMAMOORTHY A. Fractional repetition codes with flexible repair from combinatorial designs[J]. IEEE Transactions on Information Theory, 2016, 62(4): 1565-1591.

[10] OLMEZ O, RAMAMOORTHY A. Repairable replication-based storage systems using resolvable designs[C]//2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton). IEEE, 2012: 1174-1181.

[11] ZHU B, SHUM K W, LI H, et al. General fractional repetition codes for distributed storage systems[J]. IEEE Communications Letters, 2014, 18(4): 660-663.

[12] CIDON A, RUMBLE S, STUTSMAN R, et al. Copysets: Reducing the frequency of data loss in cloud storage[C]//Proceedings of 2013 USENIX Annual Technical Conference (ATC 2013). USENIX. 2013: 37-48.

[13] CHERKASOVA L, GUPTA M. Analysis of enterprise media server workloads: access patterns, locality, content evolution, and rates of change[J]. IEEE/ACM Transactions on networking, 2004, 12(5): 781-794.

[14] DAU H, MILENKOVIC O. MaxMinSum steiner systems for access balancing in distributed storage[J]. SIAM Journal on Discrete Mathematics, 2018, 32(3): 1644-1671.

[15] BRUMMOND W M. Kirkman systems that attain the upper bound on the minimum block sum, for access balancing in distributed storage[J]. arXiv:1906.02157, 2019.

[16] CHEE Y M, COLBOURN C J, DAU H, et al. Access balancing in storage systems by labeling partial Steiner systems[J]. Designs, Codes and Cryptography, 2020, 88(11): 2361-2376.

[17] SILBERSTEIN N, ETZION T. Optimal fractional repetition codes based on graphs and designs [J]. IEEE Transactions on Information Theory, 2015, 61(8): 4164-4180.

[18] GODA K, KITSUREGAWA M. The history of storage systems[J]. Proceedings of the IEEE, 2012, 100(Special Centennial Issue): 1433-1440.

[19] CHURCH G M, GAO Y, KOSURI S. Next-generation digital information storage in DNA[J]. Science, 2012, 337(6102): 1628-1628.

[20] GOLDMAN N, BERTONE P, CHEN S, et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA[J]. Nature, 2013, 494(7435): 77-80.

[21] GRASS R N, HECKEL R, PUDDU M, et al. Robust chemical preservation of digital information on DNA in silica with error-correcting codes[J]. Angewandte Chemie International Edition, 2015, 54(8): 2552-2555.

[22] ZHIRNOV V, ZADEGAN R M, SANDHU G S, et al. Nucleic acid memory[J]. Nature materials, 2016, 15(4): 366-370.

[23] YAZDI S H T, KIAH H M, GARCIA-RUIZ E, et al. DNA-based storage: Trends and methods [J]. IEEE Transactions on Molecular, Biological and Multi-Scale Communications, 2015, 1 (3): 230-248.

[24] TANG Y, YEHEZKEALLY Y, SCHWARTZ M, et al. Single-error detection and correction for duplication and substitution channels[J]. IEEE Transactions on Information Theory, 2020, 66(11): 6908-6919.

[25] GABRYS R, KIAH H M, MILENKOVIC O. Asymmetric Lee distance codes for DNA-based storage[J]. IEEE Transactions on Information Theory, 2017, 63(8): 4982-4995.

[26] KIAH H M, PULEO G J, MILENKOVIC O. Codes for DNA sequence profiles[J]. IEEE Transactions on Information Theory, 2016, 62(6): 3125-3146.

[27] YAZDI S H T, KIAH H M, GABRYS R, et al. Mutually uncorrelated primers for DNA-based data storage[J]. IEEE Transactions on Information Theory, 2018, 64(9): 6283-6296.

[28] YAZDI S H T, GABRYS R, MILENKOVIC O. Portable and error-free DNA-based data storage[J]. Scientific reports, 2017, 7(1): 1-6.

[29] YAZDI S H T, YUAN Y, MA J, et al. A rewritable, random-access DNA-based storage system [J]. Scientific reports, 2015, 5(1): 1-10.

[30] GABRYS R, DAU H, COLBOURN C J, et al. Set-codes with small intersections and small discrepancies[J]. SIAM Journal on Discrete Mathematics, 2020, 34(2): 1148-1171.

[31] COLBOURN C J. The crc handbook of combinatorial designs[M]. CRC press, 2010.

[32] FÜREDI Z. Geometrical Solution of an Intersection Problem for Two Hypergraphs[J]. European Journal of Combinatorics, 1984, 5(2): 133-136.

[33] ALON N. An extremal problem for sets with applications to graph theory[J]. Journal of Combinatorial Theory, Series A, 1985, 40(1): 82-89.

[34] ALON N, KALAI G. A simple proof of the upper bound theorem[J]. European Journal of Combinatorics, 1985, 6(3): 211-214.

[35] BLOKHUIS A. Solution of an extremal problem for sets using resultants of polynomials[J]. Combinatorica, 1990, 10(4): 393-396.

[36] GRIGGS J R, STAHL J, TROTTER W T. A Sperner theorem on unrelated chains of subsets [J]. Journal of Combinatorial Theory, Series A, 1984, 36(1): 124-127.

[37] KALAI G. Weakly saturated graphs are rigid[M]//North-Holland Mathematics Studies: volume 87. Elsevier, 1984: 189-190.

[38] KALAI G. Intersection patterns of convex sets[J]. Israel Journal of Mathematics, 1984, 48 (2-3): 161-174.

[39] KATONA G. Solution of a problem of A. Ehrenfeucht and J. Mycielski[J]. Journal of Combinatorial Theory Series A, 1974, 17: 265-266.

[40] LOVÁSZ L. Flats in matroids and geometric graphs[C]//Combinatorial Surveys (Proc. 6th British Combinatorial Conference. 1977: 45-86.

[41] TARJÁN T G. Complexity of lattice-configurations[J]. Studia Sci. Math. Hungar., 1975, 10 (1-2).

[42] FRANKL P. An extremal problem for two families of sets[J]. European Journal of Combinatorics, 1982, 3(2): 125-127.

[43] KANG D Y, KIM J, KIM Y. On the Erdős-Ko-Rado theorem and the Bollobás theorem for t-intersecting families[J]. European Journal of Combinatorics, 2015, 47: 68-74.

[44] KIRÁLY Z, NAGY Z L, PÁLVÖLGYI D, et al. On families of weakly cross-intersecting set-pairs[J]. Fundamenta Informaticae, 2012, 117(1-4): 189-198.

[45] PATKÓS B. On the general position problem on Kneser graphs[J]. Ars Math. Contemp., 2019, 18: 273-280.

[46] TALBOT J. A new Bollobás-type inequality and applications to t-intersecting families of sets [J]. Discrete mathematics, 2004, 285(1-3): 349-353.

[47] TUZA Z. Inequalities for two set systems with prescribed intersections[J]. Graphs and Combinatorics, 1987, 3(1): 75-80.

[48] TUZA Z. Applications of the set-pair method in extremal hypergraph theory[J]. Extremal problems for finite sets (Visegr'ad, 1991), Bolyai Soc. Math. Stud., 1994, 3: 479-514.

[49] TUZA Z. Applications of the se-pair method in extremal problems, II[J]. Combinatorics, Paul Erdős is eighty, Vol. 2 (Keszthely, 1993), Bolyai Soc. Math. Stud., 1996, 2: 459-490.

[50] BOLLOBÁS B. Combinatorics: set systems, hypergraphs, families of vectors, and combinatorial probability[M]. Cambridge University Press, 1986.

[51] KATONA G. Intersection theorems for systems of finite sets[J]. Acta Mathematica Academiae Scientiarum Hungaricae, 1964, 15(3-4): 329-337.

[52] WANG J. Intersecting antichains and shadows in linear lattices[J]. Journal of Combinatorial Theory, Series A, 2011, 118(7): 2092-2101.

[53] LOVÁSZ L. Combinatorial problems and exercises: volume 361[M]. American Mathematical Soc., 2007.

[54] BABAI L, FRANKL P. Linear algebra methods in combinatorics[M]. University of Chicago, 1988.

[55] PAWAR S, NOORSHAMS N, EL ROUAYHEB S, et al. Dress codes for the storage cloud: Simple randomized constructions[C]//2011 IEEE International Symposium on Information Theory Proceedings. 2011: 2338-2342.

[56] BRESLAU L, CAO P, FAN L, et al. Web caching and Zipf-like distributions: Evidence and implications[C]//Proceedings of the IEEE International Conference on Computer Communications: volume 1. 1999: 126-134.

[57] SEDLÁČEK J. Problem 27. Theory of graphs and its applications[C]//Proc. Symp. Smolenice. Praha. 1963: 163-164.

[58] STEWART B M. Magic graphs[J]. Canadian Journal of Mathematics, 1966, 18: 1031-1059.

[59] STEWART B M. Supermagic complete graphs[J]. Canadian Journal of Mathematics, 1967, 19: 427-438.

[60] DOOB M. Generalizations of magic graphs[J]. Journal of Combinatorial Theory, Series B, 1974, 17(3): 205-217.

[61] DOOB M. Characterizations of regular magic graphs[J]. Journal of Combinatorial Theory, Series B, 1978, 25(1): 94-104.

[62] JEURISSEN R H. Magic graphs, a characterization[J]. European Journal of Combinatorics, 1988, 9(4): 363-368.

[63] SEDLÁČEK J. On magic graphs[J]. Mathematica slovaca, 1976, 26(4): 329-335.

[64] SHIU W C, LAM P C B, LEE S M. On a construction of supermagic graphs[J]. Journal of Combinatorial Mathematics and Combinatorial Computing, 2002, 42: 147-160.

[65] IVANČO J. On supermagic regular graphs[J]. Mathematica Bohemica, 2000, 125(1): 99-114.

[66] SUN G C, GUAN J, LEE S M. A labeling algorithm for magic graph[J]. Congressus Numerantium, 1994: 129-138.

[67] BEZEGOVÁ L, IVANČO J. A characterization of complete tripartite degree-magic graphs[J]. Discussiones Mathematicae Graph Theory, 2012, 32(2): 243-253.

[68] BEZEGOVÁ L, IVANČO J. An extension of regular supermagic graphs[J]. Discrete mathematics, 2010, 310(24): 3571-3578.

[69] GALLIAN J A. A dynamic survey of graph labeling[J]. Electronic Journal of combinatorics, 2018, 1(DynamicSurveys): DS6.

[70] COLBOURN C J. Egalitarian edge orderings of complete graphs[J]. Graphs and Combinatorics, 2021: 1-9.

[71] CHETWYND A G, HILTON A J W. Regular graphs of high degree are 1-factorizable[J]. Proceedings of the London Mathematical Society, 1985, 3(2): 193-206.

[72] DOERR B, SRIVASTAV A. Multicolour discrepancies[J]. Combinatorics, Probability and Computing, 2003, 12(4): 365-399.

[73] LOVÁSZ L, SPENCER J, VESZTERGOMBI K. Discrepancy of set-systems and matrices[J]. European Journal of Combinatorics, 1986, 7(2): 151-160.

[74] MUTHUKRISHNAN S, NIKOLOV A. Optimal private halfspace counting via discrepancy [C]//Proceedings of the forty-fourth annual ACM symposium on Theory of computing. 2012: 1285-1292.

[75] ARMONI R, SAKS M, WIGDERSON A, et al. Discrepancy sets and pseudorandom generators for combinatorial rectangles[C]//Proceedings of 37th Conference on Foundations of Computer Science. 1996: 412-421.

[76] MATOUŠEK J, WELZL E, WERNISCH L. Discrepancy and approximations for bounded VC-dimension[J]. Combinatorica, 1993, 13(4): 455-466.

[77] DOERR B. Lattice approximation and linear discrepency of totally unimodular matrices[C]// Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms. 2001: 119-125.

[78] SOLYMOSI J. Incidences and the spectra of graphs[M]//Combinatorial number theory and additive group theory. Springer, 2009: 299-314.

[79] STINSON D R. Combinatorial Designs: Construction and Analysis[M]. Springer Science & Business Media, 2004.

[80] SCHELLENBERG P J, STINSON D R, VANSTONE S A, et al. The existence of Howell designs of side $n+1$ and order $2n$[J]. Combinatorica, 1981, 1(3): 289-301.

[81] HUFFMAN W C, PLESS V. Fundamentals of error-correcting codes[M]. Cambridge university press, 2010.

[82] RÖDL V. On a packing and covering problem[J]. European Journal of Combinatorics, 1985, 6(1): 69-78.

[83] ALON N, SPENCER J H. The probabilistic method[M]. John Wiley & Sons, 2004.

致　谢

自 2016 年起，从中国科学技术大学一路走进葛根年教授创办的组合数学与信息交叉科学研究团队过程中，我的博士阶段也即将结束。在这几年过程中，我的成长离不开亲爱的老师、家人、朋友和同学的帮助与支持。至此让我向他们表示真诚的感谢！

首先我要特别地感谢科大导师张先得特任教授和北京团队的葛根年教授。在科大的学习阶段，张老师在学习和生活上给予了极大的帮助。她引领着我进入组合数学和编码理论的领域，让我踏上这条有趣学术道路，积极推荐我参加各种学术会议和暑期研讨班，丰富了我的专业知识，提升了我的学术眼见。在北京团队的学习阶段，葛根年教授渊博的知识、开阔的眼界使我接触到了学术的前沿。您创新的思维、严谨的学术作风提升了我的学术素养。面对新的学术挑战鼓励我勇往直前。这些宝贵的经历令我终生受益。

感谢在科大的一起学习的同学：陈婷婷、石飞、杨倩倩学姐、邱瑜师兄、魏歆、王琛、李言智、谢天颖等，是你们让度过了快乐的科大时光。感谢北京团队优秀的毕业师兄们：魏恒嘉师兄、胡思煌师兄、李抒行师兄、张一炜师兄、上官冲师兄、汪馨师兄、张韬师兄、丁报昆师兄、马景学师兄等，你们优异的表现激励我进步，也同时感谢你们在学习生活上对我的照顾。感谢北京团队的同门：孔祥梁师兄、钱昺辰师兄、叶左、奚元霄、徐子翔、韩雪娇、兰昭君、谢城飞、徐民、李好阳、孙钰博、刘欣、刘雨等。尤其感谢孔祥梁师兄、钱昺辰师兄、奚元霄、徐子翔对我学术研究的帮助和照顾。在这段学习生活时光，我们留下了美好的回忆。

感谢我的朋友们：远在德国的石磊 (Daniel)，北京的彭峰、杨鑫、田昌昊、陈筱静，以及家乡的何杨、占施宇、邹润、黄娇磊、匡阳等。你们给我带来的欢快和经验。

感谢亲爱的袁博君一直对我的理解、支持、鼓励和陪伴。为我前进提供不断的动力。感谢博君父母和袁林姐的支持与鼓励。最后要感谢一直鼓励，默默支持我的父亲、母亲。你们是我坚强的后盾，温暖的港湾。

感谢所有关心爱护我的人！谢谢你们！

余文俊

2021 年 10 月

在读期间发表的学术论文与取得的研究成果

已发表论文

1. **Wenjun Yu**, Xiande Zhang, and Gennian Ge, Optimal Fraction Repetition Codes for Access-Balancing in Distributed Storage, *IEEE Transactions on Information Theory*, vol. 67, no. 3, pp. 1630–1640, March 2021, doi: 10.1109/TIT.2020.3039901.
2. **Wenjun Yu**, Xiangliang Kong, Yuanxiao Xi, Xiande Zhang, and Gennian Ge, Bollobás-type theorems for hemi-bundled two families, *European Journal of Combinatorics*, accepted, 2021.

待发表论文

1. **Wenjun Yu**, Yuanxiao Xi, Xin Wei, and Gennian Ge, Balanced set codes with small intersections, in preparation.
2. Zixiang Xu, **Wenjun Yu**, and Gennian Ge, Embedding bipartite distance graphs under Hamming metric in finite fields, submitted to *Journal of Combinatorial Theory, Series A*.