arXiv:2305.01471v1 [cs.DS] 2 May 2023

# 带离群点的容量约束／公平聚类的固定参数可处理近似

Rajni Dabas * Neelima Gupta † Tanmay Inamdar ‡

**摘要**

聚类问题，例如 $k$-Median和 $k$-Means，源于位置规划、无监督学习等应用。在许多此类应用中，找到一种点的聚类方式，使其在点的数量上不"偏斜"，十分重要；也就是说，任何簇都不应包含 *too many* 个点。通常通过对簇的大小施加 *capacity constraints* 来对此建模。另一方面，在聚类领域，另一个重要考量是如何处理数据中存在的 *outliers* 。事实上，文献已将上述聚类问题推广为分别处理容量约束和离群点的形式。然而，据我们所知，对于能够同时处理容量约束和离群点的聚类问题，其可近似性研究仍然很少。

我们弥补了这一空白，并开始研究 Capacitated $k$-Median with Outliers(C$k$MO) 问题。在该问题中，我们希望将除 $m$ *outlier points* 之外的所有点划分到至多 $k$ 个簇中，使得：(i) 簇满足容量约束；(ii) 聚类成本最小，其中成本定义为每个 *non-outlier* 点到其所属簇中心的距离之和。

我们为 C$k$MO 设计了首批常数因子近似算法。具体而言，我们的算法针对一般度量空间中的 C$k$MO 问题可返回 $(3+\epsilon)$-近似解；针对常数维欧几里得空间中的该问题可返回 $(1+\epsilon)$-近似解，运行时间为 $f(k, m, \epsilon)\cdot|I_m|^{O(1)}$，其中 $|I_m|$ 表示输入规模。我们还将这些结果扩展到更广泛的一类问题，获得了类似的固定参数可处理近似，包括带离群点的容量约束 $k$-均值/$k$-设施选址问题，以及带离群点的规模均衡公平聚类问题。对于这些问题，我们获得的近似比均达到相应无离群点问题的最佳已知保证。

## 1 引言

聚类问题，例如 $k$-Median， $k$-Means， $k$-Center，以及 Facility Location ，是近似算法文献中研究最为广泛的问题之一。尽管一般而言，这些问题都是NP困难的，但对于其中每一个问题，都已知存在可在多项式时间内实现的常数因子近似算法 [3, 29, 33, 40]。此外，在若干特殊设定下，例如点属于低维欧几里得空间时，还可以获得接近最优的近似比 [3, 6, 19, 21, 23, 28]。

在许多应用中，我们希望为给定的点集找到满足某些附加性质的聚类。因此，人们投入了大量精力研究上述聚类问题的不同约束 版本。值得注意的例子包括处理容量约束、离群点、容错、公平性，以及中心上的背包约束和拟阵约束等聚类问题。通常，处理每种此类约束都需要采用不同的技术，而针对这些问题分别获得良好近似保证的成功程度也各不相同 [4, 5, 9, 14, 15, 17, 37]。然而，自然地，我们希望找到一个能同时满足多项约束的解，这促使我们提出以下问题。

我们能否设计近似算法，以处理聚类问题中的多重正交约束，例如 $k$-Median?

*计算机科学系，德里大学，新德里，印度。电子邮件： rajni@cs.du.ac.in。

†印度新德里德里大学计算机科学系。电子邮箱： ngupta@cs.du.ac.in。

‡挪威卑尔根大学信息学系，挪威卑尔根。电子邮箱： Tanmay.Inamdar@uib.no. 。产生这些研究成果的研究工作获得了欧洲研究委员会（ERC）通过LOPPRE项目提供的资助，项目参考编号为819416。

本文重点研究在 $k$-MEDIAN、 $k$-MEANS 及相关目标函数的背景下，如何处理容量约束（分别为公平性约束）与离群点问题。在讨论本文所研究的问题之前，我们先通过讨论其中几个已在文献中得到广泛研究的特殊情形，逐步引出该问题的定义。

一般形式的 $k$-MEDIAN及其推广如下。给定一个实例 $I = ((P, d)\ C, \mathcal{F}, k)$，其中$(P, d)$是一个度量空间，$C \subseteq P$是由 $n$ 个客户组成的集合，$\mathcal{F} \subseteq P$是候选设施的集合，$k$ 是一个正整数。在 $k$-MEDIAN中，我们要找到一个包含 $k$个设施的集合 $F \subseteq \mathcal{F}$，使目标函数 $\sum_{c \in C} \min_{f \in F} d(c, f)$最小。针对 $k$-MEDIAN，已有一系列近似算法，最终在一般度量空间中实现了多项式时间的2.613近似 [29]；而在低维欧几里得空间、倍增空间以及由无禁式图类诱导的度量空间中，已知有近似方案 [21, 23, 28]。 $k$-MEDIAN WITH OUTLIERS ($k$MO)是 $k$-MEDIAN的一种推广，其中还给定一个非负整数 $m$，表示离群点的数量；这些离群点对应的距离贡献可以从目标函数中排除。在这种情况下，最小化需遍历所有可能的 $m$大小的离群点子集，以及 $k$大小的设施子集。通过迭代舍入，已知一般度量空间中的 $k$MO存在常数近似算法 [32, 37]。

CAPACITATED $k$-MEDIAN(C$k$M) 将容量约束引入了普通的 $k$-MEDIAN 问题。在这里，对于每个候选设施 $f \in \mathcal{F}$，我们都会得到一个表示其容量的非负整数 $u_f$。如果我们决定将 $f$ 纳入大小为 $k$的解中，那么最多可以向 $f$分配 $u_f$ 个客户。目标是找到一个大小为 $k$的解 $F \subseteq \mathcal{F}$ 以及一个满足容量约束的分配，使每个客户到其所分配设施的距离之和最小化。容量约束的存在使问题变得更加困难——事实上，在一般度量空间中为 C$k$M 获得常数因子近似，长期以来一直是近似算法领域中悬而未决的问题。通过设计所谓的 双准则 近似，人们在这一目标上取得了部分进展：这类近似将成本近似到一个常数因子以内，同时还允许“$k$-约束”或容量约束以一个小因子 [11, 12, 13, 18, 26, 31, 39]被违反。

本文旨在研究同时处理容量约束和离群点的聚类问题。据我们所知，只有 $k$-CENTER 和 FACILITY LOCATION 目标函数的相关研究同时处理了这两种约束，参见例如 Cygan 等人 [24] 和 Dabas 和 Gupta [25]。本文研究 $k$MO 和 C$k$M 的一种常见推广。我们称这一问题为 CAPACITATED $k$-MEDIAN WITH OUTLIERS (C$k$MO)，定义如下。

C$k$MO
**输入：** 实例 $I_m = ((P,\ d)\ C,\ \mathcal{F},\ k,\ u,\ m)$，其中 $(P,\ d)\ C,\ \mathcal{F}$的含义如上所述。此外，每个设施 $f \in \mathcal{F}$ 的容量为 $u_f \in \mathbb{N}$，$0 \leq m \leq n$ 表示可以不予分配的客户数量。

**任务：** 求得

- 一个大小至多为 $k$ 的子集 $F \subseteq \mathcal{F}$，
- 一个大小至多为 $m$ 的子集 $C' \subseteq C$ （称为离群点集合），并且
- 一个分配 $\sigma : (C \setminus C') \to F$ 满足容量约束，即对于每个设施 $f \in \mathcal{F}$，都有 $|\sigma^{-1}(f)| \leq u_f$

使得成本 $\sum_{c \in C \setminus C'} d(c, \sigma(c))$ 最小化。

注意，C$k$MO 推广了 C$k$M，而后者本身的多项式时间可近似性仍未解决。该领域近期的一个研究方向是探索多项式时间近似算法之外的方法，以突破常数因子可近似性的障碍，或获得紧的近似因子。

**固定参数可处理近似。** 放宽对多项式运行时间的严格要求后，可以为 $k$-MEDIAN 及其变体获得更优的近似。Cohen-Addad 等人 [20] 设计了一种 $(1+2/e+\epsilon)$ -近似算法，用于 $k$-MEDIAN，其运行时间关于 $k$ 和 $\epsilon$ [1]为固定参数可处理，与复杂性理论下界相匹配

[1]运行时间为 $f(p) \cdot |I|^{O(1)}$ 的算法称为参数 $p$上的 *fixed-parameter tractable*（固定参数可处理），其中$|I|$ 表示输入规模，而 $f$ 是任意函数。

| 问题 | 度量 | 近似 |
|---|---|---|
| $\mathscr{C}$ $k$-Median with Outliers | 任意 | $3+\epsilon$ |
| $\mathscr{C}$ $k$-Median with Outliers | 欧几里得 | $1+\epsilon$ |
| $\mathscr{C}$ $k$-Means with Outliers | 任意 | $9+\epsilon$ |
| $\mathscr{C}$ $k$-Facility Location with Outliers | 任意 | $3+\epsilon$ |

图1：定理 1、定理 2 及其扩展的一些推论。在第一列中，约束 $\mathscr{C}$ 可以替换为 Capacitated 或 $(\alpha,\beta)$-Fair。对于每一行，当 $\mathscr{C}$ =Capacitated时，我们代入 Cohen-Addad 和 Li 提出的相关固定参数可处理近似 [22], ；而当$\mathscr{C}$ $=(\alpha,\beta)$-Fair时，则使用 Bandyapadhyay 等人提出的相关固定参数可处理近似 [7]，从而得到第三列所示的近似比。

最后，通过代入 [22] 针对倍增空间中的 C$k$M 的拟多项式时间近似方案，我们得到了一个 $(1+\epsilon)$-近似算法，用于求解 C$k$MO，其运行时间为 $f(k,\ m,\ \epsilon)\cdot 2^{\mathsf{poly}(\log(|I_m|,1/\epsilon))}$。

界。最近，Agrawal 等人 [2]设计了一种从 $k$MO 到 $k$-Median 的保持近似比的归约，该归约关于 $k$，$m$ 是固定参数可处理的，且 $\epsilon$ （回想一下，$m$ 表示离群点的数量）。作为推论，他们获得了一个 $(1+2/e+\epsilon)$-近似算法，用于求解 $k$MO，该算法关于 $k$，$m$ 是固定参数可处理的，且 $\epsilon$。最后，在 C$k$M 的研究中，Adamczyk 等人 [1] 打破了一般度量空间中常数因子可近似性的瓶颈，首次设计出常数因子近似算法，其运行时间关于 $k$，$\epsilon$ 是固定参数可处理的。随后，Cohen-Addad 和 Li [22]将其改进为：在一般度量空间中，C$k$M 的 $(3+\epsilon)$-近似算法；在低维欧几里得度量空间中，$(1+\epsilon)$-近似算法，运行时间关于 $k$ 和 $\epsilon$是固定参数可处理的。近期也有一些工作致力于为其他相关聚类目标的容量约束版本设计固定参数可处理近似算法，例如半径和 [8, 34]。对于相对容易的 $k$-Center 目标，Goyal 和 Jaiswal [30] 设计了以 $k$ 和 $m$为参数的固定参数可处理近似算法，可处理包括容量约束和离群点在内的不同约束。受这些固定参数可处理近似算法成果的启发，我们继续沿着这条研究路线，研究 C$k$MO 及相关问题。

**近期独立研究。** 在一项近期的独立研究中，Jaiswal 和 Kumar [35] 为带离群点的受约束 $k$-Median/$k$-Means 问题设计了类似的几乎保持近似比的 归约，将其转化为对应的无离群点问题。因此，他们为各种受约束的 $k$-Median/Means with Outliers 问题（包括容量约束、$(\alpha,\ \beta)$-公平性等）获得了类似的近似结果。然而，粗略考察表明，他们的技术与我们的技术截然不同。

## 1.1 我们的结果与技术

本文的主要结果之一是以下定理。

**定理 1** (非正式表述)。 *存在一种从 CkMO 到 CkM 的保持近似比的归约，其运行时间关于 $k$、$m$ 和 $\epsilon$ 为固定参数可处理，且底层度量空间保持不变。*

我们还可以将这一结果扩展到相关的聚类目标，例如 $k$-Means 和 $k$-Facility Location。回想一下，$k$-Means目标函数涉及距离的平方和，而 $k$-FacilityLocation 是 $k$-Median的推广，其目标函数涉及设施开启成本。对于这两种目标——实际上，对于涉及距离 $z$ 次幂的 (, $z$)-聚类——我们都可以在关于 $k$、$m$ 和 $\epsilon$ 为固定参数可处理的时间内，将带离群点的容量受限聚类问题归约为对应的容量受限聚类变体（不含离群点）。随后，代入已知的最佳无离群点容量受限聚类近似结果，我们便可得到相匹配的近似比。这些推论列于图 1。

**环采样方法及其与相关工作的比较。** 我们算法的一个起点——与许多其他用于 $k$-Median 及其变体（[2, 20, 22]）的固定参数可处理近似一样——是 Chen [16]提出的优雅“环采样”方法。我们首先简要概述这一构造，随后讨论如何在我们的归约框架中使用它来处理容量约束和离群点。

在这一构造中，其理念是进行随机抽样，将给定的 $n$ 客户“压缩”为数量较少的加权客户，使其能够良好地近似原始客户集合。[2] 让我们考虑最简单的 $k$-MEDIAN 设定，其中没有容量约束或离群点，如 [16]。我们从一个针对 $k$-MEDIAN的粗略（常数因子）近似 $F' \subseteq \mathcal{F}$ 开始，并利用它将客户划分为围绕 $F'$中的中心的同心环。在这里，一个环是指这样一个客户子集：其中的客户到某个 $f \in F'$的距离在 $r$和 $2r$之间，并且我们考虑按几何级数增长的半径 $r$。接下来，我们从每个规模足够大的环中抽取客户的均匀样本，并为每个抽取的客户赋予适当的权重。所得加权样本的大小为 $O(k \log n/\epsilon)^2)$。可以证明，该样本中的加权客户以高概率近似保持了相对于任意解的聚类成本。这一点通过一种 Chernoff-Hoeffding 型集中界得到证明，该集中界利用了以下观察结果：尽管样本中每个加权客户的距离是一个随机变量，但由于三角不等式，所有随机变量都落在一个有界区间内。具体而言，对于任意解 $F \subseteq \mathcal{F}$以及任意半径为 $r$的环 $R \subseteq C$，距离 $\{d(c,F)\}_{c\in R}$位于一个长度为 $2r$的区间内。

由于样本规模相对较小，因此在 [20] 中使用该样本对 $k$-MEDIAN 的一部分近似解进行部分枚举，从而实现固定参数可解运行时间。Agrawal 等人 [2] 指出，在 $k$MO 的背景下，通过进一步增大样本规模（此时样本规模也取决于 $m$），集中界足够稳健，能够处理离群点。具体而言，他们证明，对于每个环，即使忽略 距离解最远的至多 $m$ 个点的贡献，集中界仍然成立。

Cohen-Addad 和 Li [22]将环形采样方法扩展到 C$k$M，但由于以下原因，分析更加复杂。考虑一个特定的环 $R$，其中心为 $f$，以及一个大小为 $k$的假设解 $F \subseteq \mathcal{F}$。对于每个 $f_i \in F$，令 $C_{f_i}$ 表示来自 $R$ 且被分配给 $f_i$的客户子集。由于我们从每个环中抽取均匀样本，因此在期望意义下，样本中各个 $|C_{f_i}|$的相对比例得以保持。因此，样本在期望意义下既保持成本，也满足容量约束。然而，我们很可能会从不同的 $C_{f_i}$ 中过度抽取或欠抽取 客户，因此需要界定加权样本的可行分配成本。为此，作者 [22]使用了一个最小费用流（MCF）问题实例。他们利用环中心 $f$ 在过度抽取和欠抽取的不同 $C_{f_i}$ 之间重新路由流，并可利用上述涉及三角不等式论证的一般化形式来界定重新路由的成本。

**我们的抽样与分析**。 由于我们需要同时处理容量约束和离群点，因此在分析中必须更加谨慎。对于 C$k$MO 的一个解，不同于 C$k$M，环 $R$ 中并非所有客户都能分配给设施——其中一些客户可能是离群点。在存在容量约束时，不能像在 $k$MO 中那样，仅忽略距离 $F$ 最远的 $m$ 个客户所作的贡献来处理这一问题。事实上，从$R$中选择最优的加权离群点，取决于在 $R$之外进行抽样的结果。尽管从期望上看，我们保持了 $R$ 中 $|C_{f_i}|$和离群点的相对比例，但我们可能会对不同 $C_{f_i}$ 中的客户进行过度或不足抽样，也可能对离群点集合中的客户进行过度或不足抽样。不过，我们能够将每个 $C_{f_i}$ 中过度/不足抽样的分析，与在 $R$ 中离群点集合过度/不足抽样的分析分离开来。此时，对已分配客户子集进行抽样所产生的误差，可以基本沿用 [22]中的论证来界定。接下来，我们界定从离群点集合中抽样所产生的误差。由于在分析中我们独立处理每个环，因此来自 $R$ 的离群点子集实质上是任意的。为此，我们推广 Agrawal 等人 [2]，的论证，并证明：对于权重至多为$m$的任意 加权客户集合（而不仅仅是距离 $F$ 最远的那些客户），从离群点集合中过度或不足抽样所产生的重新路由成本可以忽略不计。为界定重新路由的成本，我们分析一个名为带离群点的最小费用流（MCFO）的新问题的性质，该问题是 MCF 的推广。本文在多个地方用到了这一问题，而且它本身也具有独立研究价值。

总之，我们证明了环形采样方法可用于构造一个加权样本，其大小为 $W$ $O\left(\left(km \log_{\epsilon^3} n\right)^2\right)$，并且以高概率近似保留任意大小至多为 C$k$MO 的可行解 $F \subseteq \mathcal{F}$的成本 $k$（引理 2）。

接下来，我们讨论如何使用 $W$ 将 C$k$MO 归约为 C$k$M。我们枚举 $W' \subseteq W$ 的所有可能子集，其总权重 $m$ 对应于给定实例的一个最优解（$F^*, \sigma^*$）所对应的离群点集合。利用关于 $|W|$的界，可以很容易地证明，这类猜测的数量上界为 $f(k,m,\epsilon) \cdot |I_m|^{O(1)}$。对于每一个这样的离群点猜测 $W' \subseteq W$，我们通过

[2]这一概念可以形式化，以定义 *coresets*。此处省略形式化定义。

删除 $W'$，即可得到一个实例；再通过为每个加权客户创建多个共址副本，就能轻松将其转换为无权 C$k$M。对于每个这样的 C$k$M 实例，我们使用一个 $\gamma$-近似算法，并返回所有猜测中找到的成本最低的解（要为 C$k$MO 计算最优分配，我们需要能够在多项式时间内求解 MCFO；我们在引理 1 中通过将其归约为最小费用流来证明这一点）。利用引理 2，我们在引理4中证明：在对应于 $W'$ 的"正确"猜测的迭代中，C$k$M 实例的一个 $\gamma$-近似解实际上是原始 C$k$MO 实例的 $(\gamma+\epsilon)$-近似解。

带离群点的平衡公平聚类。$(\alpha,\beta)$-Fair $k$-Median 问题（（$\alpha,\beta$）-F$k$M）最近被提出，并受到了广泛关注。该模型的动机源于差别影响（DI）原则 [17],，该原则指出，不应在决策中使用人口的受保护属性（例如敏感人口统计信息），且决策不应以不成比例的方式影响不同的人口群体。在聚类背景下，通过将客户划分为 $\ell$ 个不同的组来建模人口的受保护属性，目标是找到一种最小成本聚类，确保每个簇中的颜色比例符合指定要求。一个特别值得关注的情形是，我们需要找到这样的簇：其中每个组的"局部比例"大致等于其"全局比例"。最近，Bandyapadhyay 等人 [7] 采用环形采样方法，为该问题及其变体设计了固定参数可处理近似算法。我们研究了该问题的一个推广，称为（$\alpha,\beta$）-Fair$k$-Median with Outliers （（$\alpha,\beta$）-F$k$MO），其中我们希望对客户进行平衡聚类，同时每个组最多排除 $m_j$ 个离群点 $1 \le j \le \ell$。与定理 1 类似，我们给出了从（$\alpha,\beta$）-F$k$MO到（$\alpha,\beta$）-F$k$M的归约，这是本文的第二个主要结果。以下非正式定理对此作出陈述。

**定理 2**（非正式）。 存在一种保持近似比的归约，可将$(\alpha,\beta)$-*FkMO* 转换为$(\alpha,\beta)$- *FkM*，其运行时间关于 $k$、$m=\sum_{j\in[\ell]} m_j$ 和 $\epsilon$ 为固定参数可处理，且底层度量空间保持不变。

与定理 1类似，该定理也可以扩展到相关的 $k$-Means 和 $k$-FacilityLocation 目标函数。该定理的一些推论见图 1。

## 2 预备知识

我们将用 $\mathbb{N}$、 $\mathbb{R}$ 和 $\mathbb{R}_+$ 分别表示非负整数集、实数集和非负实数集。设 $(P, d)$ 为一个度量空间，其中 $P$ 是一个有限点集， $d: P\times P\to\mathbb{R}$ 是满足三角不等式和对称性的距离函数。

**带离群点的最小费用流（MCFO）**。 在带离群点的最小费用流问题（MCFO）中，给定一个流网络 $G=(V, E)$、设施集合 $F\subseteq V$、需求点集合 $C\subseteq V$以及一个整数$m$。每个设施 $i\in F$ 具有一定的供给 $u_i$，每个需求点 $j\in C$ 具有一定的需求 $w(j)$，且每条边 $(u,v)\in E$ 的成本为 $d(u,v)$。目标是利用 $E$ 中的边，将至少 $\sum_{j\in C} w(j) - m$ 的需求从 $C$ 中的需求点输送到 $F$ 中的设施，使得每个设施 $i\in F$ 接收的流量不超过其供给 $u_i$，并使总成本最小化（总成本为用于输送需求的边的成本之和）。MCFO 是经典最小费用流问题（MCF）的推广；当$m=0$时，它与 MCF 相同。MCF 可通过线性规划在多项式时间内求解。

**引理1**。存 在一种多项式时间算法，可为 *MCFO* 问题返回一个整数解。

证明。 为了在多项式时间内求解 MCFO，我们将使用一个最小费用流问题的解。给定一个 MCFO 实例 $I_{\text{MCFO}}$，我们按如下方式构造一个最小费用流实例 $I'$：创建一个供给量为 $u_{d_f}=m$ 的虚拟设施 $d_f$。从 $d_f$ 向其他每个顶点 $v\in V$ 添加一条边，使得 $d(d_f, v)=D$，其中 $D$ 是一个略大于 MCFO 实例中最大距离的值。每个需求点 $j\in C$ 都是一个总需求为 $w(j)$ 的需求点。每个设施 $i\in F\cup\{d_f\}$ 都是一个供给量为 $u_i$ 的供给点。

**论断 1.** 设最优值($I_{MCFO}$) 为 $I_{MCFO}$ 的最优解成本。最优值($I_{MCFO}$) $=$ 最优值($I_{MCF}$) $- X$ 其中$X$ 是给定实例 $I_{MCFO}$ 的固定成本。

*Proof.* Let $m'=\max\{0, \sum_{j\in C} w(j) - \sum_{i\in F} u_i\}$.

我们将分两部分证明这一点：(*i*) 最优值($I_{\text{MCFO}}$) $\leq$最优值($I_{\text{MCF}}$) $-X$ 以及 (*ii*) 最优值($I_{\text{MCFO}}$) $\geq$最优值($I_{\text{MCF}}$) $-X$。

为证明 OPT($I_{\text{MCFO}}$) $\leq$ OPT($I_{\text{MCF}}$) $-X$，令 $S'$ 为值为 OPT($I_{\text{MCF}}$) 的 MCF 最优解。MCFO 的一个解 $S$ 可由 $S'$ 按如下方式得到：丢弃流入虚拟设施 $d_f$ 的全部流量，作为异常值流。注意，该需求至多为 $m$。如果流入 $d_f$ 的流量严格小于$m$（注意，此时它必定为 $m'$），我们将从其他设施中额外选取 $m-m'$需求作为离群点。这是 $m-m'$ 中由 $F$中的设施服务的、成本最高的 $S'$ 需求。令 $Cost(m-m')$ 表示流 $S'$中这部分需求的成本。我们有，OPT($I_{\text{MCFO}}$) $\leq Cost(S) = \text{OPT}(I_{\text{MCF}}) - m' \cdot D - Cost(m-m')$。因此，$X = m' \cdot D + Cost(m-m')$ 的结论成立。

接下来，为证明 OPT($I_{\text{MCFO}}$) $\geq$ OPT($I_{\text{MCF}}$) $-X$ 或 OPT($I_{\text{MCF}}$) $\leq$ OPT($I_{\text{MCF}}$) $+X$，我们将按如下方式，从 $I_{\text{MCFO}}$ 的最优解 $S$ 构造一个 $I_{\text{MCF}}$ 的可行解 $S'$：以 $m' \cdot D$成本将 $m'$ 离群点需求发送至 $d_f$。使用最小费用流将剩余的 $m-m'$ 需求发送至 $F$中的设施。请注意，根据 $m'$的定义，$F$中的设施至少有 $m-m'$ 供给可用。令 $Cost'(m-m')$ 表示为此支付的成本。如果 $Cost'(m-m') \leq Cost(m-m')$，则证明完毕。对于 $X = m' \cdot D + Cost(m-m')$，OPT($I_{\text{MCF}}$) $\leq Cost(S') \leq$ OPT($I_{\text{MCFO}}$) $+X$。否则，存在一个成本为 $<$OPT($I_{\text{MCFO}}$) 的 $I_{\text{MCFO}}$ 解，这与矛盾。 □

□

给定 C$k$MO 的一个实例 $I_m$和一个设施集合 $F \subseteq \mathcal{F}$，我们将 cost$_m(C, F)$ 定义为包含 $m$ 离群点的最优分配成本。如果 $F$中设施的容量约束之和小于 $n-m$，则我们称这样的集合 $F$为不可行，并将 cost$_m(C, F)$ 定义为 $= \infty$。注意，对于给定的可行 $F$，通过求解带异常值的最小费用流问题（ MCFO ），可以轻松确定离群点集合 $C'$以及客户到设施的分配。正如引理1中所述，MCFO 可以在多项式时间内求解，并返回一个整数解。因此，对于给定的 $F \subseteq \mathcal{F}$，可以计算 cost$_m(C, F)$。

在本文中，客户通常带有非负整数权重。本质上，客户的权重可以理解为该客户的多个共址副本。下面给出形式化定义。

**定义 1**。*(WC$k$MO)*。输入为 $I_m = ((P, d)\ W, \mathcal{F}, k, u, m)$，其中 $W \subseteq C \times \mathbb{N}$ 是成对元素的集合$\{(c, w(c)) : c \in C\}$，且 $w : C \to \mathbb{N}$ 是一个权重函数。现在的目标是找到一个大小至多为 $k$ 的子集 $F \subseteq \mathcal{F}$，以及一个分配 $\sigma : C \times F \to \mathbb{N}$，使得：

对于每个 $c \in C$, $\sum_{f \in F} \sigma(c, f) \leq w(c)$,

*2.* 未分配权重总和至多为 $m$，即 $\sum_{c \in C} w(c) - \sum_{c \in C, f \in F} \sigma(c, f) \leq m$,

*3.* 对每个 $f \in F$，分配必须满足容量约束，即 $\sum \sigma(c, f) \leq u_f$，并且$_{c \in C}$

*4.* $\sum_{c \in C, f \in F} \sigma(c, f) d(c, f)$最小化。

与 C$k$MO 类似，给定 WC$k$MO 的一个实例以及一组设施 $F \subseteq \mathcal{F}$，我们将 wcost$_m(W, F)$ 定义为总未分配权重至多为 $m$ 时最优分配的成本。如果 $F$ 中设施的容量约束之和小于 $\sum_{c \in C} w(c) - m$，则称集合 $F$ 不可行，并定义 wcost$_m(C, F) = \infty$。与 C$k$MO 一样，对于给定的可行 $F$，通过求解 MCFO 也可以轻松确定 WC$k$MO 中客户到设施的分配。因此，给定 $F \subseteq \mathcal{F}$ 后即可求得 wcost$_m(W, F)$。

# 3 算法

给定 C$k$MO 的一个实例 $I_m = ((P, d)\ C, \mathcal{F}, k, u, m)$，我们首先构造一个 $(k+m)$-中值问题（无容量约束）实例 $I_0 = ((P, d)\ C, \mathcal{F} \cup C, k+m)$，其中，除了原有设施集合外，我们还在每个客户所在位置设置一个设施。以下结论直接来自一个简单观察：任意解 $I_m$ 都是 $I_0$ 的可行解。

**断言 2**。 $I_0$ 的最优解的值不超过 $I_m$的最优解的值，即最优值$(I_0) \leq$最优值$(I_m)$。

现在，使用任意多项式时间算法（例如 [29]）求解实例 $I_0$，从而得到常数（$\zeta$）因子近似。令 $F_\zeta$ 为算法开启的设施集合。注意，$k_\zeta = |F_\zeta| \le k + m$和

$$\mathsf{cost}_0(C, F_\zeta) \le \zeta \cdot \mathsf{OPT}(I_0) \le \zeta \cdot \mathsf{OPT}(I_m) \tag{1}$$

对于每个设施 $f \in F_\zeta$，令 $C_f \subseteq C$ 表示算法分配给 $f$ 的客户。注意，集合 $C_f$ 两两不相交。令 $\mathsf{Ball}(f, X) \subseteq C_f$ 表示半径为 $X$ 的球，其中包含 $C_f$ 中与 $f$ 的距离不超过 $X$ 的客户。令 $R = \text{成本}_0(C, F_\zeta)\,\zeta n$，并令 $\psi = \lceil \log(\zeta n) \rceil$。

对于设施 $f \in F_\zeta$ 和集合 $C_f$，我们进一步将 $C_f$ 划分为更小的集合，称为环，使得每个环中的点到 $f$ 的距离相近，即

$$C_{f,j} = \begin{cases} \mathsf{Ball}(f, R), & \text{if } j = 0. \\ \mathsf{Ball}(f, 2^j R) \setminus \mathsf{Ball}(f, 2^{j-1} R), & \text{if } 1 \le j \le \psi. \end{cases} \tag{2}$$

令 $s = \frac{a\zeta^2}{\epsilon^3}(m + k \ln n)$，其中 $a$ 是足够大的常数。我们按如下方式从 $C_f$ 中抽取加权点 $W^f$：对于每个集合 $C_{f,j} \subseteq C_f$，如果 $|C_{f,j}| \le s$，则对每个客户 $c \in C_{f,j}$，将 $(c, 1)$ 加入 $W^f$。否则，从 $S_{f,j} \subseteq C_{f,j}$ 中无放回地抽取 $s$ 个随机客户。对于每个客户 $c \in S_{f,j}$，将 $(c, |C_{f,j}|)$ 加入 $W^f$。

对每个簇应用上述过程 $f \in F_\zeta$，我们得到$W = \cup_{f \in F_\zeta} W^f$作为点的加权集合。由于环的数量为 $\log n$，且每个环中的点数至多为 $\frac{a\zeta^2}{\epsilon^3}(m + k \ln n)$，因此加权点的总数为 $|W| = O\left(\frac{(km \log n)^2}{\epsilon^3}\right)$，因为 $a$和 $\zeta$ 是常数。不失一般性，我们假设权重为整数；通过对上述构造稍作修改，$W$ 中的点数至多增加一倍。详情可参见例如 [2, 16]。

我们的关键技术贡献是证明以下引理（引理2），该引理指出，对于所有可行集 $F \subseteq \mathcal{F}$，在剔除总权重为 $m$的离群点后，将加权样本分配给 $F$的成本，与剔除 $m$ 个离群点后，将原始客户集合分配给 $F$的成本相近。这正是 $\epsilon$-核集的定义；这类核集已得到广泛研究，尤其是在（无容量约束的）$k$-MEDIAN 和 $k$-MEANS 的背景下（参见 [23] 及其中的参考文献）。因此，引理2 表明，存在一个规模较小的核集，能够同时处理容量约束和离群点约束；然而，与核集相关的研究不同，本文的重点并非优化核集的规模。为便于理解，该引理的证明安排在第4节和第5 节。

**引理 2.** 对于所有大小为 $k$ 的可行集合 $F \subseteq \mathcal{F}$，$|\text{wcost}_m(W, F) - \text{成本}_m(C, F)| \le \epsilon \text{成本}_m(C, F)$ 的概率至少为 $1 - 1/n$。

假设引理 2 成立，让我们回到算法，看看如何利用它将 C$k$MO 归约为 C$k$M。回顾一下，$W$ 是我们的抽样算法返回的加权客户集合；令 $S$ 为客户集合（即每个对 $(c, w(c)) \in W$ 中的第一个元素组成的集合，其中 $\in W$ 表示一个加权客户）。对于任意子集 $T \subseteq S$，令 $w(T) = \sum_{c_i \in T} w(c_i)$。然后算法按如下方式进行：我们遍历每个大小至多为 $m$ 的猜测 $T \subseteq S$。首先，检查 $w(T) \ge m$；若不成立，则继续检查下一个猜测。现在，假设 $w(T) \ge m$。接着，将 $T$ 中的点排序为 $c_1, c_2, \ldots, c_{m'}$，其中 $m' = |T|$。对于每个 $c_i \in T$，我们猜测一个整数 $0 < z(c_i) \le \min\{m, w(c_i)\}$，使得 $\sum_{c_i \in T} z(c_i) = m$。注意，猜测的数量至多为 $(m')^m \le m^m$。现在，固定这样一个猜测 $\mathbf{z} = (z(c_1)\ z(c_2) \ldots, z(c'_m))$。接下来，我们定义一个新的权重函数 $w_{T,\mathbf{z}} : S \to \mathbb{N}$，如下所示。

$$w_{T,\mathbf{z}}(c) = \begin{cases} w(c) & \text{if } c \notin T \\ w(c_i) - z(c_i) & \text{if } c = c_i \in T \end{cases}$$

令 $W_{T,\mathbf{z}} = \{(c, w_{T,\mathbf{z}}(c)) : c \in S\}$ 表示得到的加权点集，其中点集与 $W$ 中的相同，但权重函数为 $w_{T,\mathbf{z}}$，而非 $w$。现在，算法使用 WC$k$M[3] 实例 $I_0 = ((P, d) W_{T,\mathbf{z}}, \mathcal{F}, k, m_0 = 0)$ 的 $\gamma$-近似算法，求得一个大小至多为 $k$ 的 $F \subseteq \mathcal{F}$，以及一个分配 $\sigma : C \times F \to \mathbb{N}$，使其满足定义 1 中的前三个性质。然后，我们使用 MCFO 计算从 $C$ 到 $F$、含有 $m$ 个离群点的最优分配成本，以及相应的分配 $\sigma$。

[3]现有的用于无权 C$k$M 的 $\gamma$-近似算法可以通过为每个加权客户创建多个共址副本来求解加权实例。

最后，算法在所有猜测（即对 $F^*$、$\sigma^*$以及 $T$和 $\mathbb{z}$的猜测）中，返回成本最小的解（$F^*$, $\sigma^*$）。首先，在引理 3 中，我们分析算法的运行时间；然后在引理 4中，分析其近似保证。

**引理 3.** 令 $T(|I'|, k')$ 表示用于求解 $I'$*of* Capacitated $k'$-Median 的实例的 $\gamma$-近似算法的运行时间。那么，我们的算法的运行时间上界为 $f(k, m, \epsilon) \cdot T(|I_m|, k) \cdot |I_m|^{O(1)}$。

证明。注意，构造集合 $W$ 需要多项式时间，因为我们使用 $\zeta$-近似算法求解 $k+m$-Median，并以此作为起点。接下来，我们界定算法尝试的猜测总数。首先，我们猜测一个大小至多为 $m$ 的子集 $T \subseteq W$，这需要 $\binom{|W|}{m} \leq \left(\frac{km}{\epsilon} \log n\right)^{O(m)}$ 的时间；通过对 $m \leq \log n \log\log n$ 是否成立进行标准的分类讨论，可将其上界为 $\left(\frac{km}{\epsilon}\right)^{O(m)} \cdot n^{O(1)}$。接下来，对于每个这样的猜测$T$，算法会猜测向量 $\mathbb{z} = (z(p_1)\ z(p_2) \ldots, z(p_{m'}))$。如前所述，这类猜测的数量上界为 $m^m$。最后，对于每个这样的猜测，我们使用针对 C$k$M 的 $\gamma$-近似算法，其运行时间为 $T(|I_m|, k)$；对于算法返回的解 $F \subseteq \mathcal{F}$，我们使用 MCFO 计算其相对于原始集合 $C$ 的成本。根据引理1，该过程需要多项式时间。因此，所述运行时间界得证。

□

现在，我们分析这个解的成本。

**引理4.** 令 $F^* \subseteq \mathcal{F}$ 为算法返回的至多包含 $k$ 个设施的集合。那么，以至少 $1 - 1/n$的概率，对于任意大小至多为 $k$的可行集合 $F \subseteq \mathcal{F}$，

$$\mathsf{cost}_m(C, F^*) \leq \gamma \cdot \frac{1+\epsilon}{1-\epsilon} \cdot \mathsf{cost}_m(C, F).$$

证明。 引理 2 中的陈述以至少 $1 - 1/n$的概率成立。我们以这一良好事件为条件，并证明当前引理以概率1成立。固定一个如引理陈述所示、大小至多为 $k$的可行集合 $F \subseteq \mathcal{F}$。于是，根据引理 2，我们得到以下不等式:

$$(1-\epsilon) \cdot \mathsf{cost}_m(C, F) \leq \mathsf{wcost}_m(W, F) \leq (1+\epsilon) \cdot \mathsf{cost}_m(C, F). \tag{3}$$

现在，考虑实现 $\mathsf{wcost}_m(W, F)$ 的分配 $\sigma$，并令 $T \subseteq S$为客户 $c$的集合，使得$\sum_{f \in F} \sigma(c, f) < w(c)$。注意，对于每个 $c \in S$, $z(c) := w(c) - \sum_{f \in F} \sigma(c, f)$ 是一个整数。令 $\mathbb{z} = (z(c_1)\ z(c_2) \ldots, z(c'_m))$，其中 $T = \{c_1, c_2, \ldots, c_{m'}\}$的索引任意确定。

容易验证，此处定义的 $(T, \mathbb{z})$ 满足算法中"猜测"的条件。因此，考虑算法中对应于 $T$and $\mathbb{z}$ 的迭代。由此可得 $\mathsf{wcost}_m(W, F) = \mathsf{wcost}_0(W_{T,\mathbb{z}}, F)$。令 $F' \subseteq \mathcal{F}$ 为该次迭代中为 WC$k$M 实例 $I_0 = ((P, d) W_{T,\mathbb{z}}, \mathcal{F}, k, 0)$ 找到的 $\gamma$-近似。由此可得，

$$\mathsf{wcost}_m(W, F') = \mathsf{wcost}_0(W_{T,\mathbb{z}}, F') \leq \gamma \cdot \mathsf{OPT}(I_m) \leq \gamma \cdot \mathsf{wcost}_0(W_{T,\mathbb{z}}, F) \tag{4}$$

这里，$\mathsf{OPT}(I_m)$ 表示 C$k$MO 实例 $I_m$ 的最优解成本；由于 $F$ 是一个至多包含 $k$ 个设施的可行集合，因此可得不等式 $\mathsf{OPT}(I_m) \leq \mathsf{wcost}_0(W_{T,\mathbb{z}}, F)$。

现在，再次由引理 2，我们得到以下不等式

$$(1-\epsilon) \cdot \mathsf{cost}_m(C, F') \leq \mathsf{wcost}_m(W, F') \leq (1+\epsilon) \cdot \mathsf{cost}_m(C, F') \tag{5}$$

最后， $F^* \subseteq \mathcal{F}$ 是所有迭代中找到的最佳解，因此

$$\begin{aligned}
\mathsf{cost}_m(C, F^*) \leq \mathsf{cost}_m(C, F') &\leq \frac{1}{1-\epsilon} \cdot \gamma \cdot \mathsf{wcost}_0(W_{T,\mathbb{z}}, F) && \text{(From (4) and (5)} \\
&= \frac{1}{1-\epsilon} \cdot \gamma \cdot \mathsf{wcost}_m(W, F) && \text{(By definition of } T, \mathbb{z}) \\
&\leq \gamma \cdot \frac{1+\epsilon}{1-\epsilon} \cdot \mathsf{cost}_m(C, F). && \text{(From (3))}
\end{aligned}$$

□

引理 3 和引理 4共同给出了以下定理的证明，该定理是定理 1的正式表述。

**定理3.** 给定一个针对 CAPACITATED $k'$-MEDIAN 的 $\gamma$ 近似算法，其在输入实例 $I'$ 上的运行时间为 $T(|I'|, k')$，我们可以在时间 $f(k, m, \epsilon) \cdot T(|I_m|, k) \cdot |I_m|^{O(1)}$ 内，得到 *CkMO* 实例 $I_m$ 的一个 $(\gamma+\epsilon)$-近似算法。

为了完成定理 3的证明，还需证明引理 2。首先，在第 4 节中，我们在一个特殊情形下证明引理 2。然后，在第 5 节中，我们说明如何将这一分析推广到更一般的情形。

# 4 单环情形

本节将证明引理 2的一个特殊情形。该情形已经展示了稍后证明更一般情形时将用到的大部分技术细节。

考虑簇 $C_f$ 中任意一个环 $C_{f,\,j}$，该簇以设施 $f \in F_\zeta$ 为中心，并且在本节中该设施始终保持固定。我们假设算法只在 $C_{f,\,j}$ 内进行抽样；而所有位于不同于 $C_{f,\,j}$ 的环中的客户，均以权重 1 纳入 $W$。请注意，如果除 $C_{f,\,j}$ 外的所有环所含的点都少于 $s$，则实际算法中可能出现这种情形。设该环的半径为 $R_j = 2^j R$。设环中的客户数量为 $N = |C_j|$。设我们的算法返回的加权客户为 $W = \cup_f W^f$。在本节余下部分，我们将证明以下引理。

**引理 5.** 对于任意大小为 $k$ 的可行 $F \subseteq \mathcal{F}$，以概率 $1 - n^{-(k+\lambda_1)}$ 有 $|\text{加权成本}_m(W, F) - \text{成本}_m(C, F)| \le \epsilon\lambda_2 N R_j$，其中 $\lambda_1$ 和 $\lambda_2$ 是常数。

如引理所述，我们固定一个可行集合 $F \subseteq \mathcal{F}$。在本节余下部分中，我们假设 $N > s = O\left(\frac{km\,\text{对数}\,n}{\epsilon^3}\right)$，否则抽样不会产生任何变化（回想一下，$C_{f,j}$ 之外的所有客户都已属于 $W$，且权重为 1）。在分析中，我们将定义一个随机向量 $\mathbb{X}$ 和一个函数 $g$，使得 $g(\mathbb{X})$ 和加权成本$_m(W, F)$ 同分布。

**定义 $\mathbb{X}$：** 随机向量 $\mathbb{X} \in \mathbb{R}_+^N$ 定义如下：对于每个坐标，以概率 $s/N$ 取值 $N/s$，否则取 0，使得 $\mathbb{E}[\mathbb{X}] = \mathbb{1}$。可以证明，以足够高的概率，该抽样方案恰好选出 $s$ 个点，使用断言 3。

**声明 3** ([22])。 设 $b$ 为一个正整数，并设 $p \in (0,1)$ 使得 $pb$ 为整数。二项分布$(b, p) = pb$ 为 $\Omega(1/\sqrt{b})$。

在声明 $b = N$ 和 $p = s/N$ 中取值后，由声明 3可知，$\mathbb{X}$ 恰有 $N \cdot (s/N) = s$ 个非零项的概率为 $\Omega(1/\sqrt{N})$。给定条件为该事件时，$\mathbb{X}$ 和 $W$ 同分布，即 $\mathbb{X}$表示抽样算法的结果。在本节余下部分，我们分析 $\mathbb{X}$的无条件行为，并证明所需的集中性（如引理 5所述）以高概率成立。随后，利用标准论证可知，即使在给定 5 的高概率事件下，引理 $\mathbb{X}$ 恰有 $s$ 个非零项时，结论仍以高概率成立。

**定义函数 $g$：** 为定义 $g$，我们首先构造一个 MCFO 实例。给定一个大小为 $N$ 的向量 $\mathbb{v}$，其中 $\mathbb{v}$ 中的每个条目都对应于 $C_j$ 中的一个客户，按如下方式创建流实例 FI($\mathbb{v}$)：每个客户 $c \in C_j$有 $\mathbb{v}_c$ 单位需求，$C \setminus C_j$ 中的每个客户有 1 单位需求，簇中心 $f$ 的需求为 $N - \sum_{c \in C_j} \mathbb{v}_c$（可能为负数[4]）。每个设施 $f' \in F$ 有 $u_{f'}$ 单位供给。离群点的数量为 $m$。我们将 $g(\mathbb{v}) : \mathbb{v} \to \mathbb{R}_+$ 定义为 FI($\mathbb{v}$) 的最优流成本。我们定义的流实例是可行的，因为需求总和为 $|C| = n$，离群点数量为 $m$，因此需要满足的需求为 $n - m$；根据对 $F$ 的假设，这是可行的。

注意，$g(\mathbb{E}[\mathbb{X}]) = g(\mathbb{1})$ 恰好是成本$_m(C, F)$。此外，$g(\mathbb{X})$ 和 wcost$_m(W, F)$ 服从相同的分布。现在，我们分两步证明引理 5：$(i)$ $g(\mathbb{X}) \approx \mathbb{E}(g(\mathbb{X}))$ 以高概率成立（由引理 6 证明）；以及 $(ii)$ $\mathbb{E}(g(\mathbb{X})) \approx g(\mathbb{E}[\mathbb{X}])$（由引理 7 证明）。

**引理 6。** $|g(\mathbb{X}) - \mathbb{E}[g(\mathbb{X})]| \le \epsilon N R_j/2$ 的概率为 $\ge 1 - n^{-(k+c)}$。

[4]顶点 $v$ 处的负需求 $d$ 要求必须有 $d$ 单位的流进入 $v$，而正需求则要求指定数量的需求必须流出该顶点。

证明。我们首先证明，对于 $\mathbb{R}_+^N$ 中的 $\ell_1$ 距离， $g(\mathbb{X})$是 $R_j$-利普希茨的，然后应用标准的鞅工具证明 $g(\mathbb{X})$ 集中在其均值附近。为了证明 $g(\mathbb{X})$是 $R_j$-利普希茨的，固定一个客户 $c \in C_j$，并考虑两个向量$\mathbb{v}$、$\mathbb{v}' \in \mathbb{R}_+^N$，其中$\mathbb{v}' = \mathbb{v} + \delta \cdot \mathbb{1}_c$。注意，流实例$\mathbb{v}$与$\mathbb{v}'$ 的流实例相同，只是后者在客户 $c$ 处多了 $\delta$ 需求，并且在簇中心 $f$处少了 $\delta$ 需求。

我们首先根据 $FI(\mathbb{v})$ 的最优流 $\phi$，按如下方式为 $FI(\mathbb{v}')$ 构造一个可行流：创建一个供给量为 $m$ 的虚拟设施 $d_f$，并通过引入成本为 0 的边将其连接到所有需求点。对于每个需求点，将其在 $FI(\mathbb{v})$ 的 $\phi$ 中属于离群点的那部分需求以 0 成本发送到 $d_f$。为 $\mathbb{v}'$ 构造可行流时，从 $c$ 向 $f$ 添加 $\delta$ 单位的流。将流入设施 $d_f$ 的所有需求都视为离群点。不难看出，由此得到的流是 $\mathbb{v}'$ 的可行流，且解的成本最多增加 $\delta R_j$。因此， $g(\mathbb{v}') \leq g(\mathbb{v}) + \delta R_j$。

接下来，我们以类似的方式根据 $FI(\mathbb{v}')$ 的最优流 $\phi$ 为 $FI(\mathbb{v})$ 构造一个可行流：创建一个供给量为 $m$ 的虚拟设施 $d_f$，并通过引入成本为 0 的边将其连接到所有需求点。对于每个需求点，将其在 $FI(\mathbb{v}')$ 的 $\phi$ 中属于离群点的那部分需求以 0 成本发送到 $d_f$。为 $\mathbb{v}$构造可行流时，从 $f$ 向 $c$ 添加 $\delta$ 单位的流。将流入设施 $d_f$ 的所有需求都视为离群点。同样不难看出，这是 $\mathbb{v}$ 的可行流，且解的成本最多增加 $\delta R_j$。因此， $g(\mathbb{v}) \leq g(\mathbb{v}') + \delta R_j$。

证明过程与 Cohen-Addad 和 Li 的做法类似 [22], 也就是说，通过应用针对利普希茨函数的切尔诺夫界（见下述命题1），即可得到所需的界。

命题 *1*（[22]）。设 $x_1$、...、$x_n$ 是独立随机变量，以概率 $p$ 取值 $b$，以概率 $1-p$ 取值 *0*；设 $g: [0,1]^n \to \mathbb{R}$ 是关于 $\ell_1$ 范数的 $L$-利普希茨函数。定义 $X := (x_1, ..., x_n)$ 和 $\mu := \mathbb{E}[g(X)]$。那么，对于 $0 \leq \epsilon \leq 1$， $Pr[|g(X) - \mathbb{E}[g(X)|] \geq \epsilon pnbL] \leq 2e^{-\epsilon^2 pn/3}$.

我们将命题1应用于函数 $g$，其中 $X := \mathbb{X}$， $p := s/N$， $n := N$， $b := 1/p$，以及 $L := R_j$，从而得到以下结果：

$$\begin{aligned}
\Pr[|g(\mathbb{X}) - \mathbb{E}[g(\mathbb{X})]| \geq (\epsilon/2)NR_j] &= \Pr[|g(\mathbb{X}) - \mathbb{E}[g(\mathbb{X})|] \geq (\epsilon/2)pnbL] \\
&\leq 2 \cdot \exp\left(\frac{-(\epsilon/2)^2 pn}{3}\right) \\
&= 2 \cdot \exp\left(\frac{-(\epsilon/2)^2 (s/N) N}{3}\right) \\
&= \exp\left(-\Theta\left(\epsilon^2 s\right)\right) \\
&= \exp\left(-\Theta\left(\epsilon^2 \cdot \frac{m + k\log n}{\epsilon^2}\right)\right) \\
&\leq n^{-(k+\lambda_1)}
\end{aligned} \tag{6}$$

其中，最后一个等式根据 $s$ 的定义得出，而 $\lambda_1$ 是最后一个不等式中的常数。至此，引理 6的证明结束。 □

**引理 7。** $|\mathbb{E}[g(\mathbb{X})] - g(\mathbb{E}[\mathbb{X}])| \leq \epsilon\lambda_2 NR_j$ 其中 $\lambda_2$ 是一个常数。

证明。我们分两部分证明（$i$） $g$（$\mathbb{E}[\mathbb{X}]$） $\leq \mathbb{E}[g$（$\mathbb{X}$）]（证明见第 4.1 节），以及（$ii$） $\mathbb{E}[g$（$\mathbb{X}$）] $\leq g$（$\mathbb{E}[\mathbb{X}]$） $+\epsilon\lambda_2 NR_j$（证明见第 4.2 节）。 □

引理 5 可由将引理 6 和引理 7 相加并修改 $\lambda_2 \leftarrow \lambda_2 + 1/2$。

## 4.1 引理的证明 7，第 $(i)$ $g(\mathbb{E}[\mathbb{X}]) \leq \mathbb{E}[g(\mathbb{X})]$

为了证明 $g(\mathbb{E}[\mathbb{X}]) \leq \mathbb{E}[g(\mathbb{X})]$, ，我们为 $FI(\mathbb{E}[\mathbb{X}])$ 构造一个成本不超过 $\mathbb{E}[g(\mathbb{X})]$ 的可行解。由于最小费用流的成本只会更低，因此我们得到所需结果。

设向量 $\mathbb{X}$ 的取值分别为$\mathbb{v}_1$、$\mathbb{v}_2$、.....，对应的概率为 $p_1$、$p_2$、.....。我们有， $\mathbb{E}[g(\mathbb{X})] = \sum_i p_i g(\mathbb{v}_i)$。设 $\phi_i$ 为针对 FI($\mathbb{v}_i$) 得到的流。现在，考虑将 $i$ 乘以 $p_i$ 后求和得到的流 $\phi$。注意， $\phi$ 的成本至多为 $\sum_i p_i g(\mathbb{v}_i)$，即 $= \mathbb{E}[g(\mathbb{X})]$。

接下来，我们将证明 $\phi$ 是 FI($\sum_i p_i \mathbb{v}_i$) = FI($\mathbb{E}(\mathbb{X})$) 的可行流。对于客户 $c$，设 $y_{c,i}$ 为 FI($\mathbb{v}_i$) 中的需求，设 $o_{c,i}$ 为作为离群点保留的需求，即在 $\phi_i$ 中满足了 $(y_{c,i} - o_{c,i})$ 的需求。因此，在 $\phi$ 中，总需求 $\sum_{c\in C}\sum_i p_i \cdot (y_{c,i} - o_{c,i})$ 得到了满足。并且，$\sum_{c\in C}\sum_i p_i \cdot (y_{c,i} - o_{c,i}) = \sum_i p_i \cdot \sum_{c\in C}(y_{c,i} - o_{c,i}) = \sum_{c\in C}(y_{c,i} - o_{c,i}) = |C| - m_o$ 倒数第二个等式成立，是因为概率之和为 1；最后一个等式成立，是因为 $\phi_i$ 是 FI($\mathbb{v}_i$) 的可行流。

接下来我们证明，流 $\phi$ 中每个设施都满足容量约束。对于任意设施 $j \in F$，设 $j_i$ 为流 $\phi_i$ 中流入 $j$ 的总流量。因此，流 $\phi$ 中流入 $j$ 的总流量为 $\sum_i j_i p_i \le (\sum_i p_i) \cdot \max_i j_i = \max_i j_i = j_{i^*}$ （假设如此），由于 $\phi^*$ 是 FI($\mathbb{v}_i$) 的可行流，该值不超过 $j$ 的容量。

## 4.2 引理 7的证明，第 $(ii)$ $\mathbb{E}[g(\mathbb{X})] \le g(\mathbb{E}[\mathbb{X}]) + \epsilon\lambda_2 N R_j$

为证明这一点，我们首先证明断言 4 ，然后证明引理 8。

**断言 4.** 以概率 *1*， $g(\mathbb{X}) \le g(\mathbb{E}[\mathbb{X}]) + nNR_j$。

证明。 $g(\mathbb{X})$ 的值落在一个长度为 $N/s \cdot N \cdot R_j \le N^2 R_j$ 的区间内，因为 $\mathbb{X} \in [0, N/s]^N$ 且函数$g$是 $R_j$-利普希茨函数。 $g(\mathbb{E}[\mathbb{X}])$ 也落在同一区间内，因为 $\mathbb{E}[\mathbb{X}] = \mathbb{1} \in [0, N/s]^N$。因此，断言得证。 □

现在我们证明以下引理，这是我们分析中的关键技术步骤。

**引理 8.** 以 $1 - n^{-10}$ 的概率， $g(\mathbb{X}) \le g(\mathbb{E}[\mathbb{X}]) + \epsilon\lambda_2 N R_j$，其中 $\lambda_2$ 是一个常数。

证明. 为证明这一点，我们从 $FI(\mathbb{1})$ 的带离群点的最小费用流中构造一个适用于 $FI(\mathbb{X})$ 的可行流 $\phi$，使得流 $\phi$ 的成本不超过 $g(\mathbb{E}[\mathbb{X}]) + \epsilon\lambda_2 N R_j$。

考虑 FI($\mathbb{1}$) 的最优带离群点最小费用流 $\phi'$。设 $C^o$ 为该流中被设为离群点的客户集合。注意，$|C^o| \le m_o$。创建一个虚拟设施 $d_f$，并以 0 成本将其连接到 $C^o$ 中的所有客户。对于每个客户 $c \in C^o$，在 $\phi'$ 中添加从 $c$ 到 $d_f$ 的 1 单位流。注意， $g(\mathbb{1})$ 不变。此外，将从环中心 $f'$ 到 $d_f$ 的成本设为 0。

对于所有不在 $C_j$ 中的客户，我们按照 $\phi'$ 中的相同方式路由其 1 单位需求。剩下 $C_j$ 中的需求以及环中心 $f'$ 处的额外 $N - \sum_{c\in C_j} \mathbb{v}_c$ 需求。

对于每个设施 $f \in \mathcal{F} \cup d_f$，令 $C_j^f \subseteq C_j$ 为 $\phi'$ 中由 $f$ 服务的客户集合，并令 $S^f \subseteq C_j^f$ 为 $C_j^f$ 中的抽样客户。如果 $S^f$ 欠采样，即 $S^f \cdot \frac{N}{s} \le C_j^f$，那么在 $\phi$ 中我们路由：

1. 从每个 $c \in S^f$ 向设施 $f$ 输送 $\frac{N}{s}$ 单位的流，并且
2. 将 $|C_j^f| - |S^f| \cdot \frac{N}{s}$ 个单位的流从环中心 $f'$ 发送到设施 $f$。

而如果 $S^f$ 是过采样的，即 $|S^f| \cdot \frac{N}{s} > |C_j^f|$，那么我们首先从 $S^f$ 中随机抽取一个子样本，例如大小为 $\lfloor |C_j^f| \cdot \frac{s}{N} \rfloor$ 的 $S_s^f$。现在在 $\phi$ 中，我们进行如下路由：

1. 从每个 $c \in S_s^f$ 向设施 $f$发送 $\frac{N}{s}$ 单位的流，
2. 从环中心 $f'$ 向设施 $f$ 发送 $|C_j^f| - |S_s^f| \cdot \frac{N}{s}$ 单位的流，并且，
3. 从每个 $c \in S^f \setminus S_s^f$ 向环中心 $f'$发送 $\frac{N}{s}$ 单位的流。

注意，流入 $d_f$ 的流总量为 $|C^o| \le m_o$。将所有流入的需求都分配给 $d_f$ 在 $\phi$中的离群点。可以很容易验证，所得的流 $\phi$ 是实例 FI($\mathbb{X}$) 的可行流。

位于 $C_j \setminus C^o$ 的客户流成本与 Cohen-Addad 和 Li [22] 中的相同，如下述引理所述：

**引理 9** ([22])。 以至少 $1 - n^{-10}$ 的概率， $C_j \setminus C^o$ 中客户在流 $\phi$ 中的成本有界于 $\sum_{f\in F\setminus d_f} \sum_{c\in C_j^f \setminus C^o} d(c, f) + 0.48\epsilon N R_j$。

现在，对于 $C^o$ 中的客户，从客户到 $d_f$ 或从环中心到 $d_f$ 所支付的成本为 0。因此，我们额外支付的唯一成本是从客户 $c \in C^o$ 到环中心 $f'$ 的成本，该成本至多为 $R_j$。由于 $|C^o| \leq m$，额外成本总计至多为 $m \cdot R_j \leq \frac{\epsilon}{\lambda_2} N \cdot R_j$，其中 $\lambda_2$ 是一个常数。将此结果与引理 9 相加，可得总成本至多为：

$$\sum_{f \in F} \sum_{c \in C_j} d(c, f) + 0.48\epsilon N R_j + \epsilon \lambda_2 N R_j. \tag{7}$$

修改常数 $\lambda_2 = 0.48 + \lambda_2$ 即可得到引理8。 □

根据断言4 和引理8，我们有

$$\begin{aligned} \mathbb{E}[g(\mathbb{X})] &\leq n^{-10} \cdot (g(\mathbb{E}[\mathbb{X}]) + nNR_j) + (1 - n^{-10})(g(\mathbb{E}[\mathbb{X}]) + \lambda_2 \epsilon N R_j) \\ &= g(\mathbb{E}[\mathbb{X}]) + (n^{-10} \cdot n + (1 - n^{-10}) \cdot \lambda_2) N R_j \\ &\leq g(\mathbb{E}[\mathbb{X}]) + \epsilon \lambda_2 N R_j, \end{aligned} \tag{8}$$

由此完成了对 $\mathbb{E}[g(\mathbb{X})] \leq g(\mathbb{E}[\mathbb{X}]) + \epsilon \lambda_2 N R_j$ 的证明。

# 5 多环情形的分析

在第 4 节中，我们展示了如何界定单环情形下的误差。现在考虑多环情形，并对所有环应用并集界，以得到所需的结论。

按任意顺序考虑这些环 $\sigma$。对于任意两个环 $C_{f,j}$ 和 $C_{f',j'}$，若 $C_{f',j'}$ 在 $\sigma$ 中排在 $C_{f,j}$ 之后，我们称 $(f, j) < (f', j')$。固定一个环 $C_{f,j}$。现在，我们仿照第 4 节中定义的函数，为每个环定义一个函数 $g_{f,j}$。如第 4 节所做，为了定义 $g_{f,j}$，我们首先创建一个与大小为 $n$ 的向量 $\mathbb{Y}$ 对应的 MCFO 实例 FI($\mathbb{Y}$)。为创建该实例，我们定义一个随机向量 $\mathbb{X} \in \mathbb{R}^{|C_f|}$，$_{,j}|$，其中每个坐标以概率 $|C_f$ 取值 $\overset{s}{,j}|$，$|C_f$，$_{,j}|$，否则取 0。在 FI($\mathbb{Y}$) 中，每个客户$c \in C_{f,j}$的需求为 $\mathbb{X}_c$，簇中心 $f$ 的需求为 $|C_{f,j}| - \sum_{c \in C_f, j} \mathbb{X}_c$。但是，由于我们也考虑来自其他环的样本，$g_{f,j}$ 也将取决于这些样本。假设我们为每个环$C_{f',j'}$（$f', j' \neq (f, j)$）固定样本 $S_{f',j'}$。在 FI($\mathbb{Y}$) 中，将每个客户 $c \in S_{f',j'}$ 的需求设为 $|C_{\underset{s}{f'}, \ _{j'}|}$。$g_{f,j}(\mathbb{Y})$ 现在是 FI($\mathbb{Y}$) 的最优成本。

令 $\mathbb{E}_{S_{f',j'}:(f',j') > (f,j)}[g_{f,j}(\mathbb{Y}) | S_{f',j'}: (f', j') < (f, j)]$，或简写为 $\mathbb{E}_{>(f,j)}[g_{f,j}(\mathbb{Y})]$，表示在所有样本 $S_{f',j'}$ 上，给定固定样本 $S_{f',j'}$ 时，$(f', j') > (f, j)$ 的 $g_{f,j}$ 的期望，其中所有 $(f', j') < (f, j)$ 均成立。类似地，定义 $\mathbb{E}_{S_{f',j'}:(f',j') \geq (f,j)}[g_{f,j}(\mathbb{Y}) | S_{f',j'}: (f', j') < (f, j)]$，或简写为 $\mathbb{E}_{\geq(f,j)}[g_{f,j}(\mathbb{Y})]$，表示在所有样本 $S_{f',j'}$ 上，给定固定样本 $S_{f',j'}$ 时，$(f', j') \geq (f, j)$ 的 $g_{f,j}$ 的期望，其中所有 $(f', j') < (f, j)$ 均成立。

设 $(f_1, j_1)$ 和 $(f_l, j_l)$ 分别是按此顺序排列的第一个和最后一个环的索引。那么，$\mathbb{E}_{\geq(f_1,j_1)}[g_{f,j}(\mathbb{Y})] = \text{成本}_m(C, F)$，且 $\mathbb{E}_{>(f_l,j_l)}[g_{f,j}(\mathbb{Y})] = \text{wcost}_m(W, F)$。

根据这些定义，我们得到如下引理，它类似于单环情形中的引理6和引理7（合并后）。

**引理 10**。以至少 $1 - n^{-(k+\lambda_1)}$ 的概率，对于任意环 $C_{f,\ j}$，$|\mathbb{E}_{>(f,j)}[g_{f,j}(\mathbb{Y})] - \mathbb{E}_{\geq(f,j)}[g_{f,j}(\mathbb{Y})]| \leq \epsilon \lambda_2 |C_{f,\ j}| R_j$ 其中 $R_j$ 是环 $C_{f,\ j}$ 的半径，而 $\lambda_1$、$\lambda_2$ 是常数。

对所有环求和，并使用 $\mathbb{E}_{>(f,j)}[\text{wcost}_m(W, F)]$ 和 $\mathbb{E}_{\geq(f,j)}[\text{wcost}_m(W, F)]$,各自的定义，我们得到以下引理：

**引理 11**。对于任意可行的 $F \subseteq \mathcal{F}$，$|\mathbb{E}_{>(f,j)}[\text{加权成本}_m(W, F)] - \mathbb{E}_{\geq(f,j)}[\text{加权成本}_m(W, F)]| \leq \epsilon \lambda_2 |C_{f,\ j}| R_j$ 以概率 $1 - n^{-(k+\lambda_1)}$。

我们对所有可能的可行解集合取并集界，由此可知，引理 11中的不等式以概率 $\leq n^{-c}$ 失效，因此该引理以高概率成立。现在考虑按照 $\sigma$遍历所有环 $C_{f,\ j}$ 的过程。在给定 $S_{f',\ j'}$ 的选择后，对所有 $O((k + m) \log n) \leq n^2$ 个环应用引理 11，其中 $f'$，$j'$ 为 $<(f,\ j)$。以高概率可得：

$$
\begin{aligned}
|\mathsf{wcost}_m(W,F) - \mathbb{E}[\mathsf{wcost}_m(W,F)]| &\le \sum_{(f,j)} \epsilon\lambda_2 |C_{f,j}| R_j \\
&= 2\epsilon\lambda_2 \cdot \sum_{(f,j)} |C_{f,j}| \cdot R_j/2 \\
&\le 2\epsilon\lambda_2 \cdot \mathsf{cost}_0(C, F_\zeta) \\
&\le 2\epsilon\lambda_2 \cdot \zeta \cdot \mathsf{OPT}(I) \\
&\le O(\epsilon)\mathsf{cost}_m(C,F),
\end{aligned} \tag{9}
$$

其中，倒数第二个不等式由公式（1）推出。注意， $\mathbb{E}[\mathsf{w成本}_m(W,F)] = \mathsf{成本}_m(C,F)$。因此，将 $\epsilon$ 缩小一个常数因子，以便从 $O(\epsilon)\mathsf{成本}_m(C,F)$得到 $\epsilon\cdot\mathsf{成本}_m(C,F)$，即可得到引理 2。

# 6 其他目标函数的扩展与$(\alpha,\beta)$-公平性

## 6.1 带离群点的容量约束 $k$-均值/$k$-设施选址

首先，我们考虑 Capacitated $k$-Facility Location with Outliers (C$k$FLO)，它是带设施开启成本的C$k$MO 的推广。在这里，每个 $f \in \mathcal{F}$ 都有开启成本 $o_f \ge 0$ （开启成本可能各不相同）。C$k$FLO 的定义与 C$k$MO 完全相同，唯一的区别是目标成本现在还包括设施开启成本。更正式地说，我们要找到满足其他条件的集合 $C' \subseteq C$ 和 $F \subseteq \mathcal{F}$，并使 $\sum_{f\in F} o_f + \sum_{c\in C\setminus C'} d(c,\sigma(s))$ 最小化。我们首先注意到， [22] 中的 $3+\epsilon$- 近似可以轻松调整用于 C$k$FLO—— 在枚举过程中，还需要猜测每个 “ 领导点 ” 最近设施的开启成本，误差不超过因子 $1+\epsilon$。接着，我们从 C$k$MO 到 C$k$M 的归约在存在开启成本时也同样适用——事实上，加权样本 $W$ 以高概率近似保留了关于所有集合 $F \subseteq \mathcal{F}$ 的距离，而设施的开启成本不受抽样过程影响。因此，给出从 C$k$MO 到 C$k$M 归约的定理 1，实际上可推广为给出从 C$k$FLO 到 C$k$FL 的固定参数可处理归约。特别地，我们在一般度量空间中为 C$k$FLO 获得了 $(3+\epsilon)$-近似。

Capacitated $(k,z)$-Clustering with Outliers 的定义与 C$k$MO 类似，但其目标函数现在涉及距离的 $z$ 次幂，其中 $z \ge 1$ 为固定值。特别地， $z=1$ 对应于 C$k$MO，而 $z=2$ 对应于 Capacitated $k$-Means with Outliers。已知环形采样方法可以推广到处理目标函数中距离的 $z$ 次幂，但分析会更加繁琐，因为距离的 $z$ 次幂仅满足三角不等式的“较弱版本”。不过，有几处步骤需要谨慎处理。例如，在重新路由连接到不同设施的客户所产生的流时，以及处理离群点时，引理 8 中的保证要求对每个环额外计入一项 $O_z(\epsilon)\cdot$ 最优值，其中大 O 记号中的常数隐藏了关于 $z$ 的指数级因子（这是可以接受的，因为 $z$ 是固定值）。将所有环上的误差相加后，现在会多出一项 $O_z((k+m)\log n\cdot\epsilon)\cdot$ 最优值。因此，通过对 $\epsilon$ 进行适当缩放，我们可以得到所需的保证。技术细节留待未来版本说明。

## 6.2 ($\alpha$, $\beta$)-带离群点的公平聚类

这里，我们简要概述如何将 ($\alpha$, $\beta$)-Fair $k$-Median with Outliers (($\alpha$, $\beta$)-BF$k$MO) 归约为其无离群点对应问题。下面给出 ($\alpha$, $\beta$)-F$k$MO 的形式化定义。

**定义 2** (($\alpha,\beta$)-F$k$MO)。 给定有限点集、 $C$ 和 $\mathcal{F}$ 以及一个整数 $k$，我们还给定了将 $C$ 划分为 $\ell$ 个组的一个划分，记为 $\{C_1,\dots,C_\ell\}$，两个公平性向量 $\alpha$、$\beta \in [0,1]^\ell$ 以及一个离群点向量$\mathrm{m} \in \mathbb{N}^\ell$。目标是找到一个大小至多为 $k$ 的子集 $F \subseteq \mathcal{F}$、一个离群点集合 $C'$ 以及一个分配$\sigma$： $C\setminus C' \to F$，使得：

*1.* 每种颜色的客户数量 $g \in [\ell]$在每个簇中最多为 $\alpha_g$ 且至少为 $\beta_g$ 该簇客户总数的比例。更正式地说，对于每个 $g \in [\ell]$以及每个 $f \in F$，以下条件成立： $\beta_g \le \frac{|\sigma^{-1}(f)\cap C_g|}{|\sigma^{-1}(f)|} \le \alpha_g$.

*2.* 对于每种颜色 $g \in [\ell]$，离群点的数量至多为 $m_g$，即 $\forall g \in [\ell], |c:\ c \in C' \cap C_g| \leq m_g$，并且
*3.* 最小化以下目标函数：成本$(\sigma) = \sum_{c \in C \setminus C'} d(c, \sigma(c))$。

我们简要概述为在 $(\alpha, \beta)$-F$k$MO 的背景下，将我们的方法调整为 C$k$MO 所需的修改。令 $I_m = ((P, d)$ $C, \mathcal{F}, k,$ m $\alpha, \beta)$ 表示给定的 $(\alpha, \beta)$-F$k$MO 实例。令 $m := \sum_{g \in [\ell]} m_g$ 表示各颜色的离群点总数。首先，我们通过忽略公平性约束，并在每个客户处添加一个共址设施，定义一个合适的 $(k+m)$-Median 实例 $I_0$。容易得出，OPT($I_0$) ≤OPT($I_m$)。随后，我们为 $I_0$求得一个常数近似解，并用它进行环采样。环的定义与之前相同；不过，对于每个环 $C_{f,j}$ 以及每种颜色 $g \in [\ell]$，我们需要抽取 $O((m + k \log n)/\epsilon^3)$ 个点。注意，样本中的每个加权客户都保留其原有颜色。因此，样本 $W$的总规模至多为 $\left(^{km\ell}\log_{\epsilon}{}^{n}\right)^{O(1)}$。在枚举阶段，我们尝试猜测加权离群点的确切分布；对于每一种猜测，我们通过删除相应的猜测离群点，得到一个 $(\alpha, \beta)$-F$k$M 实例。我们对每个这样的实例使用一个 a$\gamma$-近似算法，并且可以论证，最优解是 $I_m$ 的一个 $(\gamma + \epsilon)$-近似解。

现在我们概述一下分析过程。首先，给定一个可行集合 $F \subseteq \mathcal{F}$，我们需要计算一个满足（$\alpha$，$\beta$）-公平性约束的可行分配。为此，我们需要构造一个类似于 MCFO 的算法，我们在下一小节中将其称为 带离群点的加权公平分配 （WFAO）。正如 [7], 中所指出的，即使没有离群点，该问题也是 NP-困难的；但他们为该问题设计了一种 FPT 算法。我们将他们的思路推广到能够处理离群点的情形。在算法中，我们使用 WFAO 来寻找可行分配；在分析中，我们也使用它来论证各种重新分配的成本始终有界。与之前一样，我们分析中的一个关键思想是，证明每个环中离群点集合的过采样/欠采样成本可以忽略不计，因为样本大小相对于 $m$（即离群点总数）而言足够大。在第6.3 节中，我们给出 WFAO 的形式化定义，并为该问题设计一种 FPT 算法。其余分析与 C$k$MO 的分析类似，因此推迟到本文的后续完整版中给出。我们还指出，与上一小节一样，（$\alpha$，$\beta$）-F$k$MO 的论证可以直接推广到 $k$-Means 和 $k$-Facility Location 目标函数。此外，整个方法还可以推广到客户组不一定彼此不相交的情形。我们省略这些细节。

## 6.3 带离群点的加权公平分配

对于固定的设施集合 $k$ $F = \{f_1, f_2, \ldots, f_k\}$，我们将 wcost$_m(W, M, F)$ 定义为所有满足上述性质 1–3 的 $M \in \mathcal{M}$ 中，最优分配的成本。如果不存在这样的分配，我们就称这样的 $F$ 和 $M$ 不可行，并定义 wcost$_m(W, M, F)$ $= \infty$。注意，给定 $F$ 和 $M$ 后，wcost$_m(W, M, F)$ 可以通过求解一个 MCFO 得到。但是，由于 $\mathcal{M}$ 的规模可能非常大，我们无法尝试所有可能的 $M$。为此，对于固定且可行的 $F$，我们修改 [7] 中提出的加权公平分配算法，以求得满足公平性和离群点约束的 $W$ 到 $F$ 的最优分配。我们将这个问题称为带离群点的加权公平分配问题（WFAO）。$(\alpha, \beta)$-公平聚类的分配问题，以及由此得到的带离群点的 $(\alpha, \beta)$-公平聚类分配问题，都是 NP 难的 [10]。因此，我们给出一个固定参数可处理算法，其运行时间关于 $k$ 和 $\ell$ 为固定参数可处理。

**引理 12.** 存在一种算法，可返回 *WFAO* 的整数解。该算法的运行时间为 $(k\ell)^{O(k\ell)} n^{O(1)} L$，其中 $L$ 是问题实例中距离和权重的编码总位数 $I_m$。

证明。 我们将 WFAO 表述为一个混合整数线性规划。对于每个客户 $c \in C$ 和每个设施 $f \in F$，令 $x_{cf}$ 表示客户 $c$ 分配给设施 $f$ 的权重。对于每个设施 $f \in F$ 和每个组 $g \in \{1, \ldots, \ell\}$，令 $y_{fg}$ 表示组 $g$ 中的客户分配给设施 $f$ 的权重。对于每个客户 $c \in C$，令 $z_c$ 表示客户 $c$ 作为离群点的权重。现在，WFAO 可表述为如下混合整数线性规划：

$$\text{Minimize } \mathsf{cost}(x, y, z)$$

$$
\begin{aligned}
\text{subject to,} \quad & \textstyle\sum_{f\in F} x_{cf} + z_c = w(c) && \forall c \in C && (10)\\
& \textstyle\sum_{c:c\in C_g} z_c \le m_g && \forall g \in \{1,\dots,\ell\} && (11)\\
& \textstyle\sum_{c:c\in C_g} x_{cf} = y_{fg} && \forall f \in F, \forall g \in \{1,\dots,\ell\} && (12)\\
& \textstyle\sum_{c:c\in C_g} x_{cf} \le \alpha_g \sum_{c\in C} x_{cf} && \forall f \in F, \forall g \in \{1,\dots,\ell\} && (13)\\
& \textstyle\sum_{c:c\in C_g} x_{cf} \ge \beta_g \sum_{c\in C} x_{cf} && \forall f \in F, \forall g \in \{1,\dots,\ell\} && (14)\\
& x_{cf}, z_c \ge 0 && \forall c \in C, \forall f \in F && (15)\\
& y_{fg} \in \mathbb{N} && \forall c \in C, \forall g \in [\ell] && (16)
\end{aligned}
$$

其中，在 $k$-中位数问题的情形下，成本$(x, y, z) = \sum_{c\in C}\sum_{f\in\mathcal{F}} d(c, f)x_{cf}$，而在 $k$-均值问题的情形下，成本$(x, y, z) = \sum_{c\in C}\sum_{f\in\mathcal{F}} d(c, f)^2 x_{cf}$。约束 10 确保对于任意客户 $c$，分配给 $F$ 中设施的权重与作为离群点的权重之和恰好等于 $c$ 的权重。约束 11 确保每个组 $g \in [\ell]$ 的离群点数量满足上界。约束 13 和 14 是公平性约束。

我们使用以下命题求解上述混合整数线性规划。

**命题 2** ([27, 36, 38])**.** 存在一种算法，可在 $O(a^{2.5a+o(p)}b^4L)$ 时间内求解混合整数线性规划，且空间复杂度关于 $L$呈多项式，其中 $a$ 是整数变量的数量，$b$ 是非整数变量的数量，而 $L$ 是给定实例的位长。

注意，在我们的混合整数线性规划中，$a = k\ell$ and $b = nm$。因此，我们可以在时间 $(k\ell)^{O(k\ell)}(nm)^{O(1)}L$内求得最优解 $\{x_{cf}\}, \{y_{fg}\}, \{z_c\}$。

$\{y_{fg}\}$ 是整数，但 $\{x_{cf}\}$ 和 $\{z_c\}$ 可能是分数。接下来我们证明，可以利用 $\{y_{fg}\}$的整数性找到另一个最优解，使得 $\{x_{cf}\}$ 和 $\{z_c\}$ 也为整数。对于每个组 $g \in [\ell]$，考虑如下网络：对于每个客户 $c \in C_g$，设置一个供给节点，其供给量为 $w(c)$；对于每个设施 $f \in F$，设置一个需求节点，其需求量为 $y_{fg}$。在每个客户 $c \in C_g$与每个设施 $f \in F$之间添加一条容量无限的边，其成本为 $d(c, f)$。由于所有需求量和供给量均为整数，因此存在一个成本最小的最大流（记为 $\{x'_{cf}\}$），且该流为整数流；该流可以通过最小费用流在多项式时间内计算得到。用 $\{x'_{cf}\}$替换 $\{x_{cf}\}$，即可得到另一个最优解，其中 $\{x_{cf}\}$和 $\{y_{fg}\}$为整数。注意，一旦我们令 $\{x_{cf}\}$为整数，$\{z_c\}$也为整数，因为 $w(c)$是整数，且对于所有 $c \in C$，$z_c = w(c) - \sum_{f\in F} x_{cf}$。对每个组 $g \in [\ell]$重复上述步骤后，该解便完全为整数。 □

# 7 结论

本文进一步展示了环形采样方法的强大之处，证明它足够灵活，能够处理聚类问题中的多种约束。具体而言，我们为 C$k$MO 设计了固定参数可处理近似算法。我们还将相关论证推广到处理离群点，同时保持 ($\alpha$, $\beta$)-公平性。我们的工作留下了几个尚待解决的问题：我们还能同时处理哪些其他同时正交约束？能否将这一方法推广到其他类型的聚类目标，例如半径和聚类？或许，与本文结果更为相关的另一个开放问题是，能否改进 C$k$MO 的近似保证（即使是在容量约束均匀这一特殊情形下）？一种途径是改进 C$k$M 的近似算法。

# 参考文献


[1] Marek Adamczyk、Jaroslaw Byrka、Jan Marcinkowski、Syed Mohammad Meesum 和 Michal Wlo- darczyk。容量约束 k-中位数的常数因子固定参数可处理近似。载于 Michael A. Bender、Ola Svensson 和 Grzegorz Herman 编，第*27*届欧洲算法年度研讨会，*ESA 2019*，*2019*年*9*月*9*日至*11*日，德国慕尼黑/加兴，第144卷*LIPIcs*，第1:1–1:14页。Schloss Dagstuhl - 莱布尼茨信息学中心¨，2019年。 数字对象标识符:10.4230/LIPIcs.ESA.2019.1。[2] Akanksha Agrawal、Tanmay Inamdar、Saket Saurabh 和 Jie Xue。聚类重要因素：带离群点聚类的最优近似，2023年。 arXiv:2212.00696。

[3] Sara Ahmadian、Ashkan Norouzi-Fard、Ola Svensson 和 Justin Ward。通过原始-对偶算法，为 k-均值和欧几里得 k-中位数提供更好的保证。SIAM 计算杂志，49(4)，2020年。数字对象标识符:10.1137/ 18M1171321.[4] Hyung-Chan An、Aditya Bhaskara、Chandra Chekuri、Shalmoli Gupta、Vivek Madan 和 Ola Svensson。树在容量约束 k-中心中的核心作用。数学规划，154(1-2):29–53，2015年。数字对象标识符:10.1007/ s10107-014-0857-y.[5] Hyung-Chan An、Mohit Singh 和 Ola Svensson。基于 Lp 的容量约束设施选址算法。载于 FOCS，2014年，第256–265页，2014年。[6] Sanjeev Arora、Prabhakar Raghavan 和 Satish Rao。欧几里得 k-中位数及相关问题的近似方案。载于第三十届 ACM 计算理论年度研讨会论文集，STOC '98，第106–113页，美国纽约，1998年。ACM。网址:http://doi.acm.org/10.1145/ 276698.276718，数字对象标识符:10.1145/276698.276718.[7] Sayan Bandyapadhyay、Fedor V. Fomin 和 Kirill Simonov。度量空间和欧几里得空间中公平聚类的核集及其应用。载于 Nikhil Bansal、Emanuela Merelli 和 James Worrell 编，《第48届自动机、语言与编程国际研讨会》，ICALP 2021，2021年7月12日至16日，苏格兰格拉斯哥（线上会议），LIPIcs 第198卷，第23:1–23:15页。Schloss Dagstuhl - 莱布尼茨信息学中心，2021年。数字对象标识符:10.4230/LIPIcs.ICALP.2021.23.[8] Sayan Bandyapadhyay、William Lochet 和 Saket Saurabh。用于最小化簇半径总和的容量约束聚类 FPT 常数近似算法。CoRR，abs/2303.07923，2023年。arXiv:2303.07923，数字对象标识符:10.48550/arXiv.2303.07923.[9] Manisha Bansal、Naveen Garg 和 Neelima Gupta。容量约束设施选址的5-近似算法。载于 ESA，第33–144页，2012年。[10]Suman Bera、Deeparnab Chakrabarty、Nicolas Flores 和 Maryam Negahbani。公平聚类算法。载于 H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alch´e-Buc、E. Fox 和 R. Garnett 编，《神经信息处理系统进展》第32卷。Curran Associates, Inc.，2019年。网址：https://proceedings.neurips.cc/paper_files/paper/2019/file/ fc192b0c0d270dbf41870a63a8c76c2f-Paper.pdf.[11] Jaroslaw Byrka、Krzysztof Fleszar、Bartosz Rybicki 和 Joachim Spoerhase。硬容量约束 k-中位数问题的双因子近似算法。载于 ACM-SIAM 离散算法研讨会，2013年。[12] Jaroslaw Byrka、Bartosz Rybicki 和 Sumedha Uniyal。适用于均匀容量约束 k-中位数问题且允许容量违反的近似算法。载于 Quentin Louveaux 和 Martin Skutella 编，《整数规划与组合优化——第18届国际会议 IPCO 2016，比利时列日，2016年6月1日至3日，论文集》，计算机科学讲义第9682卷，第262–274页。施普林格，2016年。数字对象标识符:10.1007/978-3-319-33461-5_22.[13] Moses Charikar、Sudipto Guha、 ´Eva Tardos 和 David B. Shmoys。k-中位数问题的常数因子近似算法（扩展摘要）。载于 STOC，1999年，第1–10页。数字对象标识符:10.1145/ 301250.301257.[14] Moses Charikar、Samir Khuller、David M. Mount 和 Giri Narasimhan。含离群点的设施选址问题算法。载于 S. Rao Kosaraju 编，《第十二届离散算法年度研讨会论文集》，2001年1月7日至9日，美国华盛顿哥伦比亚特区，第642–651页。ACM/SIAM，2001年。网址:http://dl.acm.org/citation.cfm?id=365411.365555.[15] Danny Z. Chen、Jian Li、Hongyu Liang 和 Haitao Wang。拟阵中心问题与背包中心问题。Algorithmica，75(1):27–52，2016年5月。数字对象标识符:10.1007/s00453-015-0010-1.[16] Ke Chen。度量空间和欧几里得空间中 k-中位数与 k-均值聚类的核集及其应用。SIAM计算杂志，39(3):923–947，2009年。

[17] Flavio Chierichetti、Ravi Kumar、Silvio Lattanzi 和 Sergei Vassilvitskii。通过公平小簇实现公平聚类。载于 Isabelle Guyon、Ulrike von Luxburg、Samy Bengio、Hanna M. Wallach、Rob Fergus、S. V. N. Vishwanathan 和 Roman Garnett（编），《2017年神经信息处理系统大会论文集》第30卷：神经信息处理系统年会，2017年12月4日至9日，美国加利福尼亚州长滩，第5029–5037页，2017年。网址：https://proceedings.neurips.cc/paper/2017/hash/978fce5bcc4eccc88ad48ce3914124a2-Abstract.html。[18] Julia Chuzhoy 和 Yuval Rabani。非均匀容量约束下 k-中位数问题的近似。载于《第十六届 ACM-SIAM 离散算法年度研讨会论文集》，SODA 2005，加拿大不列颠哥伦比亚省温哥华，2005年1月23日至25日，第952–958页。SIAM，2005年。网址：http: //dl.acm.org/citation.cfm?id=1070432.1070569。[19] Vincent Cohen-Addad。倍增度量中容量受限聚类的近似方案。载于 Shuchi Chawla（编），《2020年 ACM-SIAM 离散算法研讨会论文集》，SODA 2020，美国犹他州盐湖城，2020年1月5日至8日，第2241–2259页。SIAM，2020年。数字对象标识符：10.1137/1. 9781611975994.138。[20] Vincent Cohen-Addad、Anupam Gupta、Amit Kumar、Euiwoong Lee 和 Jason Li。针对 k-中位数和 k-均值的紧致固定参数可处理近似。载于 Christel Baier、Ioannis Chatzigiannakis、Paola Flocchini 和 Stefano Leonardi（编），《第46届自动机、语言与编程国际研讨会》，ICALP 2019，2019年7月9日至12日，希腊帕特雷，第132卷 LIPIcs，第42:1–42:14页。达格施图尔信息学莱布尼茨国际论文集中心，2019年。数字对象标识符：10.4230/LIPIcs.ICALP.2019.42。[21] Vincent Cohen-Addad、Philip N. Klein 和 Claire Mathieu。局部搜索为欧几里得度量和无小图度量中的 k-均值与 k-中位数问题带来近似方案。SIAM 计算杂志，48(2):644–667，2019年。数字对象标识符：10.1137/17M112717X。[22] Vincent Cohen-Addad 和 Jason Li。论容量受限聚类的固定参数可处理性。载于 Christel Baier、Ioannis Chatzigiannakis、Paola Flocchini 和 Stefano Leonardi（编），《第46届自动机、语言与编程国际研讨会》，ICALP 2019，2019年7月9日至12日，希腊帕特雷，第132卷 LIPIcs，第41:1–41:14页。达格施图尔信息学莱布尼茨国际论文集中心，2019年。数字对象标识符：10.4230/LIPIcs.ICALP.2019.41。[23] Vincent Cohen-Addad、David Saulpic 和 Chris Schwiegelshohn。新的聚类核心集框架。载于 Samir Khuller 和 Virginia Vassilevska Williams（编），STOC '21：《第53届 ACM SIGACT 计算理论年度研讨会》，线上活动，意大利，2021年6月21日至25日，第169–182页。ACM，2021年。数字对象标识符：10.1145/3406325.3451022。[24] Marek Cygan 和 Tomasz Kociumaka。带离群点的容量约束 k-中心问题的常数因子近似。载于理论计算机科学算法与系统研讨会，第251–262页，2014年。[25] Rajni Dabas 和 Neelima Gupta。带离群点 / 惩罚项的容量约束设施选址。载于 Yong Zhang 、 Dongjing Miao 和 Rolf H. M¨ohring（编），《计算与组合学——第28届国际会议》，COCOON 2022，中国深圳，2022年10月22日至24日，论文集，第13595卷《计算机科学讲义》，第549–560页。施普林格，2022年。数字对象标识符：10.1007/978-3-031-22105-7_49。[26] G¨okalp Demirci 和 Shi Li。允许（1+epsilon）容量违反的容量约束 k-中位数问题的常数近似。载于 Ioannis Chatzigiannakis、Michael Mitzenmacher、Yuval Rabani 和 Da- vide Sangiorgi（编），《第43届自动机、语言与编程国际研讨会》（ICALP 2016），第55卷《信息学莱布尼茨国际论文集》（LIPIcs），第73:1–73:14页，德国达格施图尔，2016年。达格施图尔信息学莱布尼茨国际论文集中心。网址：http: //drops.dagstuhl.de/opus/volltexte/2016/6211，数字对象标识符：10.4230/LIPIcs.ICALP.2016.73。[27] Andr´as Frank 和 ´Eva Tardos。同时丢番图逼近在组合优化中的应用。《组合数学》，7(1):49–65，1987年3月。数字对象标识符：10.1007/BF02579200。[28] Zachary Friggstad、Mohsen Rezapour 和 Mohammad R. Salavatipour。局部搜索为倍增度量中的 k- 均值问题带来多项式时间近似方案。SIAM 计算杂志，48(2):452–480，2019 年。数字对象标识符：10.1137/17M1127181。

[29] Kishen N. Gowda、Thomas W. Pensyl、Aravind Srinivasan 和 Khoa Trinh。针对 k-中位数问题的改进双点舍入算法与黄金障碍。载于 Nikhil Bansal 和 Viswanath Nagarajan（编），《2023年 ACM-SIAM 离散算法研讨会论文集》，SODA 2023，意大利佛罗伦萨，2023年1月22日至25日，第987–1011页。SIAM，2023年。数字对象标识符：10.1137/1.9781611977554.ch38。[30] Dishant Goyal 和 Ragesh Jaiswal。受约束 k-中心和 k-供应商问题的紧致固定参数可处理近似。《理论计算机科学》，940（部分卷）:190–208，2023年。数字对象标识符：10.1016/j.tcs.2022.11.001。[31] Sapna Grover、Neelima Gupta、Samir Khuller 和 Aditya Pancholi。均匀硬容量约束背包中位数问题的常数因子近似算法。载于 Sumit Ganguly 和 Paritosh K. Pandya（编），《第38届IARCS软件技术与理论计算机科学基础年会》，FSTTCS 2018，2018年12月11日至13日，印度艾哈迈达巴德，第122卷 LIPIcs，第23:1–23:22页。达格施图尔信息学莱布尼茨国际论文集中心，2018年。数字对象标识符：10.4230/LIPIcs. FSTTCS.2018.23。[32] Anupam Gupta、Benjamin Moseley 和 Rudy Zhou。广义 k-中位数问题的结构化迭代舍入。载于 Nikhil Bansal、Emanuela Merelli 和 James Worrell（编），《第48届自动机、语言与编程国际研讨会》，ICALP 2021，2021年7月12日至16日，苏格兰格拉斯哥（线上会议），第198卷 LIPIcs，第77:1–77:18页。达格施图尔信息学莱布尼茨国际论文集中心，2021年。数字对象标识符：10.4230/LIPIcs.ICALP.2021.77。[33] Dorit S Hochbaum。集合覆盖问题和顶点覆盖问题的近似算法。SIAM计算杂志，11(3):555–556，1982年。[34] Tanmay Inamdar 和 Kasturi R. Varadarajan。容量约束半径和聚类：一种固定参数可处理近似。载于 Fabrizio Grandoni、Grzegorz Herman 和 Peter Sanders（编），《第28届欧洲算法研讨会》，ESA 2020，2020年9月7日至9日，意大利比萨（线上会议），第173卷 LIPIcs，第62:1–62:17页。达格施图尔信息学莱布尼茨国际论文集中心，2020年。数字对象标识符：10.4230/LIPIcs.ESA.2020.62。[35] Ragesh Jaiswal 和 Amit Kumar。受约束设定下聚类重要因素，2023年。arXiv：2305.00175。[36]Ravi Kannan。闵可夫斯基凸体定理与整数规划。《数学运筹学》，12(3):415–440，1987年8月。[37] Ravishankar Krishnaswamy、Shi Li 和 Sai Sandeep。通过迭代舍入实现带离群点的 k-中位数和 k-均值问题的常数近似。载于 Ilias Diakonikolas、David Kempe 和 Monika Henzinger（编），《第50届 ACM SIGACT 计算理论年会论文集》，STOC 2018，美国加利福尼亚州洛杉矶，2018年6月25日至29日，第646–659页。ACM，2018年。数字对象标识符：10.1145/3188745.3188882。[38] H. W. Lenstra。固定数量变量的整数规划。《数学运筹学》，8(4):538–548，1983年11月。数字对象标识符：10.1287/moor.8.4.538。[39]Shanfei Li。均匀硬容量约束 k-中位数问题的改进近似算法。载于 Klaus Jansen、Jos´e D. P. Rolim、Nikhil R. Devanur 和 Cristopher Moore（编），《近似、随机化与组合优化：算法与技术》（APPROX/RANDOM 2014），第28卷《信息学莱布尼茨国际论文集》（LIPIcs），第325–338页，德国达格施图尔，2014年。达格施图尔信息学莱布尼茨国际论文集中心。网址：http:// drops.dagstuhl.de/opus/volltexte/2014/4706，数字对象标识符：10.4230/LIPIcs.APPROX-RANDOM.2014.325。[40] Shi Li。无容量限制设施选址问题的 1.488 近似算法。载于《第38届自动机、语言与编程国际会议论文集——第二卷》，ICALP'11，第77–88页。施普林格出版社，2011年。网址：http://dl.acm.org/citation.cfm?id=2027223.2027230。[41] Clemens R¨osner 和 Melanie Schmidt。带约束的隐私保护聚类。载于 Ioannis Chatzigiannakis、Christos Kaklamanis、D´aniel Marx 和 Donald Sannella（编），《第45届国际

自动机、语言与编程研讨会，*ICALP2018*，*2018*年*7*月*9*日至*13*日，捷克共和国布拉格，第107卷，*LIPIcs*，第96:1–96:14页。达格施图尔城堡莱布尼茨信息学中心¨，2018年。 数字对象标识符:10.4230/LIPIcs.ICALP.2018.96。