arXiv:2603.09846v1 [cs.CG] 10 Mar 2026

低维欧几里得空间中聚类的近乎最优上界与下界

Vincent Cohen-Addad
Google Research
cohenaddad@google.com

Karthik C. S.*罗格斯大学
karthik.cs@rutgers.edu

David Saulpic 巴黎西岱大学，法国国家科学研究中心
david.saulpic@irif.fr

Chris Schwiegelshohn†
奥胡斯大学
schwiegelshohn@cs.au.dk

摘要

k-中位数和 k-均值聚类目标是用于度量空间中聚类建模的经典目标。给定度量空间中的一组点，k-中位数（分别为 k-均值）问题的目标是找到 k 个代表点，使每个点到其最近代表点的距离之和（分别为距离平方和）最小。Cohen-Addad、Feldmann 和 Saulpic [在 JACM'21] 中证明，对于低维欧几里得度量下的 k-中位数和 k-均值问题，均可在近线性时间内获得 $(1+\varepsilon)$ 因子近似，其中 d 是维度，n是输入点的数量。

我们将这一运行时间改进至 $2^{\widetilde{O}(1/\varepsilon)^{d-1}} \cdot n \cdot \mathrm{polylog}(n)$，并证明了几乎匹配的下界：根据针对 3-SAT 的间隙指数时间假设，不存在能够为 k-means 实现 $(1+\varepsilon)$-近似的 $2^{o(1/\varepsilon^{d-1})} n^{O(1)}$ 算法。

*本研究得到美国国家科学基金会资助，资助项目编号为 CCF-2313372 和 CCF-2443697；同时获得西蒙斯基金会资助（资助编号 825876，获奖者 Thu D. Nguyen），并部分得到保加利亚教育与科学部对 INSAIT、索非亚大学"圣克莱门特·奥赫里德斯基"的支持。该支持属于保加利亚国家科研基础设施路线图的一部分。

†本研究部分获丹麦独立研究基金（DFF）Sapere Aude 研究领军人才资助（编号 1051-00106B）以及 Google Research 奖的支持。

1 引言

k-median 和 k-means 问题是用于建模聚类的经典目标优化任务，广泛应用于数据挖掘和机器学习领域。给定一个点集 $P \subset X$和度量空间 中的候选中心 $S \subset X$ (X, Δ)， k-median 问题[1] 的目标是找到一组 k 点 $S \subset S$，称为中心，使每个点到其最近中心的距离（由 Δ给出）之和最小，其中每个点属于 P，其最近中心属于 S。（在 k-means 问题中，目标则是最小化每个点到其最近中心的平方距离之和，其中每个点属于 P，其最近中心属于 S。）

关于 k-中位数的算法研究始于 20 世纪 60 年代初的运筹学领域 [37,47, 30], ， 而 k-均值的研究则通常被认为始于 Lloyd 的开创性工作 [40], ， 该工作将聚类应用于模拟信号的量化。此后，这两个问题都受到了极大的关注。显然，问题的复杂度会随承载输入点的底层度量空间而变化。从否定的角度来看，已知这两个问题在欧几里得空间中均为 NP 难问题，即使这些点位于欧几里得平面中（且 k 很大） [45, 41], ，或者 $k = 2$ [23, 2] （且维度很大）。当 k 和 d 均为任意值时，多个研究团队已证明这两个问题具有近似困难性 [6, 38, 17, 19]。如果输入点位于任意度量空间中，则还存在更强的困难性结果（参见 Guha 和 Kuller [29] 以及 [18]）。

精细复杂性。 尽管存在这些困难性结果，欧几里得输入在统计学和机器学习应用中的重要性促使研究人员开发各种方法和算法，尤其是通过研究输入位于欧几里得空间时该问题的参数化复杂性。输入的维度 d 和目标聚类数量 k 都曾作为参数进行研究。对于精确算法，Imai、Katoh 和 I mai 的开创性工作 [32] 表明，可以在 $O(n^{kd+1})$ –时间内计算 k-中位数和 k-均值问题的精确解；相比之下，~朴素的暴力枚举需要 $O(k^n)$ 时间。另一方面，Cohen-Addad、de Mesmay、Rotenberg 和 Roytman [11] 后来证明，对于（离散）k-中位数或 k-均值问题，不存在 $n^{o(k)}$ 精确算法，即使维度为 4 也是如此，从而部分证明了 $O(n^{kd+1})$算法的最优性。

为规避这一困难，研究人员转而研究近似算法。当维度 d 和精度 ε 作为参数时，一系列工作 [5, 36, 28, 20, 10, 13] 将近似方案的运行时间从准多项式时间 $2^{(\log(n)/\varepsilon)^{d-1}}$改进到了 $2^{(1/\varepsilon)^{O(d^2)}}n$。参数 d^2 中的双重指数依赖，与其他类似结果形成了鲜明对比，例如旅行商问题（TSP）。对于这一问题，Kisfaludi-Bak、Nederlof 和 Wegrzycki [35] 近期确定了关于 d的精确依赖关系，证明存在一个运行时间基本为 $2^{O\left(1/\varepsilon^{d-1}\right)}n$的算法，并在 Gap-ETH 假设下证明了 $2^{\Omega\left((1/\varepsilon)^{d-1}\right)}$ 这一下界 [26, 43]（见定义 4.1）。鉴于这一结果，我们考虑以下问题：

问题 1.1。 能否为低维输入中的 k-中值和 k-均值问题设计一种 $2^{O\left(1/\varepsilon^{d-1}\right)}n^{O(1)}$ 近似方案？能否为这两个问题设计一种速度更快的 $2^{o\left(1/\varepsilon^{d-1}\right)}n^{O(1)}$ 近似方案？

[1]这种变体有时也称为离散 k-中位数问题。在连续的 k-中位数问题中，我们可以从 X中的任意位置选取 k 中心 S。

1.1 我们的结果

本文在回答问题 1.1方面取得了重大进展。我们的首要贡献是设计了一种改进的算法：

定理 1.2。 对于任意 $\varepsilon > 0$ 和维度 d， $\mathbb{R}^d$ 中的 k-中位数和 k-均值问题都可以在时间 $2^{\widetilde{O}(1/\varepsilon^{d-1})} n \cdot \mathrm{polylog}(n)$内近似到 $(1+\varepsilon)$ 因子。[2]

上述算法是我们的主要贡献，它改进了 [13]中的技术。它也很容易扩展到连续的 k-中位数和 k-均值问题（参见备注 2.1）。

在技术层面上，该算法加深了我们对聚类四叉树 分解的理解。基于这种分解的一系列研究 [5, 36, 28, 20, 10, 13] 将聚类近似方案（即针对任意足够小的常数 $(1+\varepsilon)$的 因子近似算法ε）的运行时间从准多项式时间 $2^{(\log(n)/\varepsilon)^{d-1}}$改进到了 $2^{(1/\varepsilon)^{O(d^2)}} n$ [13]。这些工作的主要概念性洞见是，限制分解过程中所需的门户 数量，从而使得强制路径经过门户所引起的失真至多为 $(1+\varepsilon)$因子。 [13] 的主要贡献在于，处理完输入中极小的一部分后，在四叉树分解的每一层级上只需常数个门户。然而，这个常数仍远大于其他自然问题所需的数量。

事实上，深入分析四叉树剖分的性能以及实现 $(1+\varepsilon)$因子近似所需的门户数量，一直是设计欧几里得空间近似方案的一条富有成果的研究方向。Arora 最初的工作 [3, 4] 展示了旅行商问题（TSP）和其他路径规划问题的近似方案，起初运行时间为准多项式时间（当 $d > 2$时），随后运行时间达到 $n \log(n)^{O(1/\varepsilon)^{d-1}}$。对于 TSP，运行时间随后得到显著改进：首先，Rao 和 Smith [46] 实现了 $(1/\varepsilon)^{1/\varepsilon^{d-1}} n \log n$；接着，Bartal 和 Gottlieb [7] 将其降至 $2^{1/\varepsilon^{O(d)}} n$，关于 n呈线性，但关于 d并非最优。最后， [35] 确定了关于 d的精确依赖关系，证明存在一种算法，其运行时间本质上为 $2^{O(1/\varepsilon^{d-1})} n$，并给出了$2^{\Omega((1/\varepsilon)^{d-1})}\mathrm{poly}(n)$ 的下界（基于 Gap-ETH 假设） [26, 43] （见定义 4.1）。

鉴于上述文献，我们的贡献因此是对聚类问题的四叉树剖分所做的改进分析，使我们能够（几乎）匹配研究充分的 TSP 所需的门户数量。这需要与 TSP 不同的分析和技术，尤其是因为当距离被平方时它们会失效——正如在 k-均值问题中所发生的那样。我们注意到，由于其简单性，四叉树的应用远远超出了低维空间中的近似算法。对于聚类，四叉树例如被用于必须保持较低内存的流式场景[8], 也被用于设计高维空间中的线性时间算法 [21] 或用于构造差分隐私算法 [12]。因此，更好地理解四叉树的性能，即使在最简单的场景中，也可能产生深远的影响。

我们以一个条件下界补充定理 1.2 ， 该下界表明我们的算法在 Gap-ETH 假设下近乎最优：

定理 1.3。 假设 *Gap-ETH* 假设成立，对于每个整数 $d \geq 2$，都存在一个常数 $c > 0$，使得不存在近似方案能够在时间 $2^{c(1/\varepsilon)^{d-1}}\mathrm{poly}(N)$ 内，给定欧几里得空间 $\mathbb{R}^d$ 中的离散 k-均值（或离散 k-中位数）实例（包含N 个点）和参数 $\varepsilon > 0$，计算出 $(1+\varepsilon)$ 因子近似。

[2]该 $\widetilde{O}$ 符号隐藏了对 d的指数依赖，以及对 $\log(1/\varepsilon)$的多项式依赖。

这个下界是通过融合 [38, 11, 17] 以及 de Berg、Bodlaender、Kisfaludi-Bak、Marx 和 van der Zanden 的框架 [24]中的技术思想得到的。将顶点覆盖问题归约为聚类问题的思路，在 [38, 11, 17], 中得到了详细阐述； [24] 则建立了一个框架，用于证明低维空间中一系列几何问题的精细化下界。我们将这些方法结合起来，得到了定理 1.3。

扩展。 在 [13] 中提出的框架 k-均值（以及 k-中位数）的许多变体，例如奖品收集 k-均值、带离群点的 k-均值以及设施选址。由于我们的结构定理是对其结构定理的直接改进，因此我们也改进了这些问题的运行时间。我们在附录B中概述了论证。我们的分析还改进了倍增度量（一种欧几里得空间的推广）下的结果，不过所得复杂度 $2^{\widetilde{O}(1/\varepsilon^d)}$并非紧确。因此，本文聚焦于欧几里得情形，在这种情形下，我们的分析几乎是紧确的。

1.2 我们的技术

为便于说明，下面我们详细介绍欧几里得平面中 k-均值问题的上界（即，满足 $d = 2$）。输入是一组点 $P \subset \mathbb{R}^2$、一组候选中心 $C \subset \mathbb{R}^2$，目标是计算一组中心 $S \in C^k$ （C^k 是 C 的 k-路划分乘积），使点 P 到其在 S 中最近中心的平方距离之和最小。

上界。 我们的起点是配备门户的四叉树剖分。我们将在接下来的几段中简要回顾其构造及主要性质。

首先，找到一个包含所有输入点的轴对齐矩形。然后，随机将该矩形分割成 4 个边长大约缩小一半的轴对齐矩形。继续递归地进行这种分割，直到每个矩形只包含一个点：这就定义了输入的递归分解。我们称两个点在层级 i 被切割，是指同时包含这两个点的最小矩形位于该分解的第 i^{th} 层级。假设分解的叶子直径为 1，则第 i^{th} 层级的矩形直径大约为 2^i。

现在，沿每个矩形的边界规则地放置一组门户。我们不再用直线将点连接到中心，而只考虑遵循门户的路径，其定义如下。要连接在层级 i 被切割的两个点 p 和 q，一条遵循门户的路径由依次连接 p、p_1、……、p_i、q_i、……、q_1、q 的一系列线段组成，其中 p_j（分别地， q_j）是层级 j 上包含 p（分别地， q）的矩形的一个门户。只考虑这些路径会强制产生一些绕行，但一个简单的动态规划可以计算出最佳的遵循门户的解，其复杂度自然（以指数方式）依赖于门户的数量。此外，如果沿每条边界放置 $1/\rho$ 个门户，可以算出在层级 i 被切割的两个点 p、q 之间的遵循门户的路径长度（本质上）为 $\|p - q\| + \rho 2^i$。事实上，由于各矩形边界的尺寸按几何级数递减，经过层级 i 上某个门户所产生的绕行主导了所有其他绕行；又因为该层级矩形的直径为 2^i，通过规则地放置 $1/\rho$ 个门户，我们可以保证存在一个门户，它与直线路径穿越边界之处的距离至多为 $\rho 2^i$。因此，分析的目标是证明：对于少量门户，存在一个具有几乎最优代价的遵循门户的解。

标准论证如下。四叉树分解的主要性质是，对于任意一对点 p、$q \in \mathbb{R}^2$，它们在层级 i 被切割的概率为 $\|p - q\|/2^i$。使用 $1/\rho$ 个门户，如果它们在层级 i被切割，那么通过门户连接 p 和 q 所需的绕行距离为 $\rho 2^i$。因此，期望

层级 i的绕行距离为 $\|p-q\|/2^i \cdot \rho 2^i = \rho\|p-q\|$。由于分解中有 $\log n$ 个层级，期望绕行距离为 $\rho\|p-q\| \cdot \log n$，因此必须取 $\rho = \varepsilon/\log n$ 才能使期望绕行距离足够小。

然而，拥有 $\log n/\varepsilon$ 许多门户并不能带来高效的算法。更重要的是，这种分析无法推广到 k-均值问题：尽管 p 与 q 之间的期望距离很小，但期望平方距离却并不小！事实上，层级 i 处的期望平方绕行为 $\|p-q\|/2^i \cdot (\rho 2^i)^2 = \rho^2\|p-q\|2^i$，而项 2^i 难以得到有效控制。为了解决这一问题，[13] 对输入进行了细致的预处理，以确保没有点会在高于 $\|p-q\|/\varepsilon$的层级处被切割。在这种情况下，采用$1/\varepsilon^2$ 个门户就足以保证较小的绕行，同时也能保证较小的平方绕行，因为这并非平均情形下的论证（不同于前面所见的论证）。这种预处理基本如下：从一个常数因子近似解 $\mathcal{A}$ 开始（它是一组 k 个中心），对于每个点 p，令$\mathcal{A}_p$ 表示它到 $\mathcal{A}$中最近中心的距离。如果球 $B(p, \mathcal{A}_p/\varepsilon)$ 在高于 $\log(\mathcal{A}_p/\varepsilon^2)$的层级处被分解切割，则将 p 替换为其最近中心在 $\mathcal{A}$中的一个副本。他们证明，每个点被替换的概率至多为 ε，因此期望代价增加为 $\varepsilon\mathrm{cost}(P, \mathcal{A}) = O(\varepsilon)\mathrm{OPT}$。此外，[13] 证明，放置 $1/\varepsilon^3$ 个门户可以确保：对于 $\mathcal{A}$中的任意中心，要么可以通过门户以较小的绕行将其连接到最优解，要么可以假设该中心属于最优解。因此，预处理之后，每个客户点都可以通过门户连接到最优解，而动态规划则可以计算出一个近优解。

然而，从某种意义上说，这种分析可视为“最坏情况”分析，因为它以相同的方式处理所有在低于 $\log(\mathcal{A}_p/\varepsilon^2)$ 的层级上被切割的球所对应的客户点。我们的贡献是将平均情况分析与 [13] 的技术结合起来，从而减少所需的门户数量。我们根据每个点从近似解 $\mathcal{A}$ 中被切割的层级，为其谨慎地定义预算，这与 [13], 中的做法相同，但现在还考虑该点从最优解中被切割的层级。我们首先证明，以常数概率，该预算非常低廉，即仅为最优代价的 ε 倍。随后我们证明，这一预算足以支付输入的预处理代价，以及为使一个解遵守门户而产生的绕行代价。算法与 [13] 中的相同：我们的贡献在于进行了更为精细的分析。

这项分析与 [13] 不同，因为他们容许的绕行与最优解无关：利用这一额外信息，我们能够进行更精确的分析，并将所需门户数量减少到 $(\log(1/\varepsilon)/\varepsilon)^{d-1}$，而非 $1/\varepsilon^{O(d)}$。尽管整体图景仍然相似，但要在以下两者之间找到恰当的平衡并非易事：（1）预算足够小，不会导致代价大幅增加；（2）预算又足以支付使最优解遵守门户的代价。这是对他们证明的非平凡扩展。事实上，这需要我们对如何连接一个被严重切割的点 p 有新的认识；也就是说，该点在高于 $\log(\|p-s\|/\varepsilon)$ 的层级上被其最优中心 s 切割。为此，我们证明，我们的预算允许以遵守门户的方式，将该点连接到距离 $\mathcal{A}(p)$ 最近的中心 $\tilde{s}$；其中，$\mathcal{A}(p)$ 是常数因子近似解 $\mathcal{A}$ 中距离 p 最近的中心。

下界。 我们的下界构造建立在 de Berg 等人 [25]的技术之上。给定一个 3-SAT 公式，他们展示了如何构造一个嵌入到 $\mathbb{R}^d$ 中的图，使得在该嵌入图上求解顶点覆盖等价于判定该公式是否具有有效赋值。该图有 $|E| = n^{d/(d-1)}$ 条边。我们利用（如今已属标准，参见例如 [11]）k-means 与顶点覆盖之间的关系，并构造一个 k-means 实例，使得计算其最优代价即可判定该公式是否可满足。

为应对近似问题，我们注意到，只需依赖 Gap-ETH [26, 43] 而非 ETH，[25]的框架便可轻松扩展，以处理完备性和可靠性情形中的常数间隙

（与 [35]的工作很相似）。因此，我们证明，任何针对 k-均值聚类的 $(1+\varepsilon)$因子近似，都能导出一个顶点覆盖，它覆盖了几乎所有的边，仅留下 $O(\varepsilon|E|)$ 数量的边未被覆盖。

因此，从一个可嵌入到 $\mathbb{R}^d$ 中网格上的顶点覆盖问题近似困难性实例出发，其中每条边都可以看作一条线段，我们将这些线段的中点视为待聚类的点，并将对应于顶点的点视为候选中心集合。该聚类实例嵌入的一个关键性质是，靠近某条边中点的候选中心只有与该边相接的两个顶点所对应的点。在完备性情形下，选取与顶点覆盖中顶点对应的点，便可保证每个点都靠近某个 k 个中心之一（特别是覆盖该边的顶点）；而在可靠性情形下，给定任何聚类代价较低的 k 个中心，都可以恢复出原图的一个顶点覆盖，并覆盖大多数边。

1.3 更多相关工作

我们此前介绍了 k-均值和 k-中值问题的困难性结果；就上界而言，对于一般度量空间，[15] 最近取得突破，为 k-中值问题实现了 $(2+\varepsilon)$ 近似，并为 k-均值问题实现了 5.6 近似 [9]。这两项工作改进了一系列长期研究成果，这些成果基于原始-对偶技术以及设施选址中所谓的拉格朗日乘子保持算法，参见例如 [34, 33, 39]。

为了提升这些近似比，可以采用以聚类数为参数的算法 k：对于 k-中位数问题和 k-均值问题，分别可以得到近似比达到下界的固定参数可处理算法，即 $1+2/e$ [33]，$1+8/e$ [16]。

过去几十年里，人们开发了其他专用于欧几里得空间的技术。降维 [42] 和核心集 [22] 可以将输入的维度降至 $O(\log k/\varepsilon^{-2})$，并将不同点的数量降至 $\widetilde{O}(k\varepsilon^{-O(1)})$，从而得到基于朴素枚举的简单固定参数可处理算法。近期的论文改进了这一简单方法，降低了对 k 的依赖（在指数中）[1]。

在引言中，我们介绍了大部分与低维聚类相关的文献——这类研究关注以 d 为参数的问题。要在关于 n 和 k 的多项式时间内计算一个 $(1+\varepsilon)$-近似，任何算法的运行时间都至少必须关于 d 呈双重指数级，因为该问题在维度 $\Omega(\log n)$ 下是 APX- 困难的。这些算法中最优的是 [13], ，其运行时间为近线性级别 $f(\varepsilon,d)\cdot n\operatorname{polylog} n$。

最后，研究者也从不同参数角度探讨了聚类的可处理性，例如代价。Fomin、Golovach 和 Simonov [27], 提出了一种用于 k-median 的 $D^D\mathrm{poly}(nd)$ 精确算法，其中 D 为最优代价。

1.4 论文结构

本文重点介绍 k-means，因为使用四叉树处理平方距离无疑更为困难。第 2 节介绍证明定理 1.2 所需的概念和定义。第 3 节分析我们提出的多项式时间近似方案，并由此得出定理 1.2；随后在第 4 节给出下界结果，从而得到定理 1.3。

我们在附录 A 中介绍如何调整论证以适用于 k-中位数问题。正文包含结构定理的证明，该定理允许减少四叉树所需的门户数量。我们在附录B中将算法结果推广到其他目标函数。

2 预备知识

2.1 定义

我们考虑欧几里得空间 $\mathbb{R}^d$，其度量为： $\mathrm{dist}(p,q) := \sqrt{\sum_{i=1}^{d}(p_i - q_i)^2}$。~~对于任意点~~ p和 $r \geq 0$，以 p 为中心、半径为 r 的球是 $B(p,r) := \{x \in \mathbb{R}^d : \mathrm{dist}(p,x) \leq r\}$。（离散）$k$-中位数和 k-均值问题定义如下：给定一组客户点 P 以及一组候选中心 C。任意大小为 k的 C 子集都是一个解 。目标是计算使$\mathrm{cost}(P,S) := \sum_{p\in P}\mathrm{dist}(p,S)^z$最小的解 S，其中对于 k-中位数问题， $z = 1$；对于 k-均值问题， $z = 2$。若一个解的代价至多为最小代价的 α 倍，则称该解为α-近似解。

注 2.1。 本文研究 离散欧几里得聚类。在连续版本中，中心可以任意放置在空间中。一项经典结果展示了如何通过计算一组包含 $1/\varepsilon^d \cdot |P|$ 候选中心的集合，将连续问题转化为离散问题，并得到 $(1+\varepsilon)$近似 *[44]*。

对于某个点集 S 和任意点 $p \in \mathbb{R}^d$，我们定义 $S(p) = \arg\min_{s\in S}\mathrm{dist}(p,s)$ 为 S 中距离 p 最近的点；并令 $S_p := \mathrm{dist}(p,S(p))$ 为 p 到其最近点 S 的距离。

我们假设两点之间的最小距离为 1，并将 $\mathrm{diam}(P)$ 记为 P 中两点之间的最大距离。我们还假设 $P \subset B(0,\mathrm{diam}(P))$。根据聚类问题的标准论证，我们可以假设 $\mathrm{diam}(P) = \mathrm{poly}(n)$ [14]。

层次分解 $\mathcal{D}$是一个序列 $\mathcal{B}_1$、 $\mathcal{B}_2$...，满足对于任意 i， $\mathcal{B}_i$是对 P的划分，并且$\mathcal{B}_i$ 是 $\mathcal{B}_{i+1}$的加细：也就是说， $\mathcal{B}_i$ 的每个部分都完全包含于 $\mathcal{B}_{i+1}$ 的某个部分中（注意，在此定义中，部分会随着 i递增）。

2.2 层次分解及其性质

引理 2.2。 对于任意点集 $P \subset \mathbb{R}^d$ 和任意 $\rho > 0$，存在一个随机分层分解$\mathcal{D}$，使得部分 $B \in \mathcal{B}_i$ 的直径至多为 2^{i+1}， $|\mathcal{D}| \leq \lceil\log(\mathrm{diam}(P))\rceil$，并且层级 $i-1$上的每个部分 $B \in \mathcal{B}_i$ 至多在 $2^{O(d)}$ 个部分中被加细，且：

1. 对于任意 $p \in \mathbb{R}^d$、半径 r和层级 i，有

$$\Pr[\mathcal{D}\ cuts\ B(p,r)\ at\ a\ level\ i] \leq dr/2^i.$$

2. 对于分解中的每个部分 $B \in \mathcal{B}_i$，都存在一个至多包含 $1/\rho^{d-1}$ 个门户的集合，使得对于任意点 $p \in B$， $q \notin B$，都存在一个靠近连接 p 与 q 的直线的门户 $x \in \mathcal{P}_B$，即 $\|p-x\|+\|x-q\| \leq \|p-q\| + \rho 2^{i+1}$。此外，位于 B 中的任何层级为 $i+1$ 的门户，也是 B 的门户。

该分解可以在 $(1/\rho)^{O(d)} n\log(\mathrm{diam}(P))$ 时间内找到。

在欧几里得空间中，该分解正是四叉树剖分。前述陈述是对 [13] 中引理 2.3 在欧几里得空间中的改写。改进后的性质1 例如见 [31]中的引理 11.3。由于 $\mathrm{diam}(P) = \mathrm{poly}(n)$，该分解可以在近线性时间内计算出来。

我们用 $\tau(\varepsilon,d) = \log(d) + \log(1/\varepsilon)$ 表示一个参数，该参数在本文中经常使用。

连接两个点 p 和 q 且在层级 i 处切割的遵守门户 路径，由一系列线段组成，这些线段连接 p、p_1、……、 p_i、 q_i、……、 q_1、 q，其中 p_j（分别地， q_j）是包含 p的层级 j矩形的一个门户

（分别地， q）。推广而言，若连接客户点与中心的所有路径都遵守门户，则一个-means 解遵守门户。回顾一下， $\mathcal{A}_p$ 是从 p 到解中最近点的距离 $\mathcal{A}$。我们将使用的一个重要定义是严重切割，它是对 [13] 略作修改后的版本：

定义 2.3. 设 $P \subset \mathbb{R}^d$, $\mathcal{D}$ 是 P 的一个层次分解，并且 $\varepsilon > 0$。

对于任意球 $B(x,r)$，我们将球 $B(x,r)$ 被 $\mathcal{D}$ 切割的层级定义为 $\mathrm{lvl}(x,r)$。

如果 $\mathrm{lvl}(x,r) \geq \log(3r) + \tau(\varepsilon, d)$，则称球 $B(x,r)$ 相对于 $\mathcal{D}$ 被严重切割。如果点 p 相对于 $\mathcal{D}$ 和 $\mathcal{A}$ 被严重切割，则称该点被严重切割；也就是说， $B(p, 3\mathcal{A}_p)$ 相对于 $\mathcal{D}$ 被严重切割（即 $\mathrm{lvl}(p, 3\mathcal{A}_p) \geq \log(3\mathcal{A}_p) + \tau(\varepsilon, d)$）。

下文中，我们将考察相对于两个不同集合被严重切割的点： $\mathcal{A}$ 是算法已知的常数因子近似，而 $\mathcal{S}$ 是算法未知的最优解的轻微修改，稍后我们将具体说明；这两个解都独立于分解 $\mathcal{D}$。对于 P 中的输入点，我们将考察它们是否相对于 $\mathcal{D}$ 和 $\mathcal{A}$ 被严重切割。对于 $\mathcal{A}$ 的中心，则考察其是否相对于 $\mathcal{D}$ 和 $\mathcal{S}$ 被严重切割。此外， $\mathcal{D}$ 始终相同，因此为简便起见，我们称其为相对于 $\mathcal{A}$（或相对于 $\mathcal{S}$）被严重切割。

下面的引理界定了点被严重切割的概率，它是性质 1 的直接推论。

引理 2.4。 设 P， $S \subset \mathbb{R}^d$ 以及 $\mathcal{D}$ 是由引理 2.2给出的随机层次分解，且 $\rho = O(\varepsilon / \log(1/\varepsilon))$。对于任意点 $p \in \mathbb{R}^d$， p 相对于 S 被严重切割的概率至多为 ε。

给定一个随机分解后，会为每个点分配一个 预算。在下面的定义中，可以将 $\mathcal{A}$ 视为算法已知的一个 $O(1)$-近似，而将 $\mathcal{S}$ 视为最优解。我们将在引理 3.4 中证明，该预算很小，并且可以计算出一个代价至多为最优代价加上该预算的解。

定义 2.5。 令 $P \subset \mathbb{R}^d$, $\mathcal{D}$ 为 P, $\varepsilon > 0$的层次分解，且 $\mathcal{A}$ 和 $\mathcal{S}$ 为 P上k-均值问题的两个解。

对于球 $B(x,r)$，相对于 B和 $\mathcal{D}$ 的绕行 为 $\det_{\mathcal{D}}(x,r) = \varepsilon 2^{\mathrm{lvl}(x,r)} r + \varepsilon^2 2^{2\,\mathrm{lvl}(x,r)}$。
点 $p \in P$ 相对于 $\mathcal{A}$ 和 $\mathcal{S}$ $b(p,\varepsilon) = b_1(p,\varepsilon) + b_2(p,\varepsilon) + b_3(p,\varepsilon)$具有预算，其中

$$b_1(p,\varepsilon) = \begin{cases} \det_{\mathcal{D}}(p, 3\mathcal{A}_p) \text{ if } \mathrm{lvl}(p, 3\mathcal{A}_p) \leq \log(3\mathcal{A}_p) + \tau(\varepsilon, d) \\ 0 \text{ otherwise} \end{cases}$$

$$b_2(p,\varepsilon) = \begin{cases} \det_{\mathcal{D}}\left(p, 3(\mathcal{A}_p + \mathcal{S}_p)\right) \text{ if } \mathrm{lvl}(p, 3(\mathcal{A}_p + \mathcal{S}_p)) \leq \log(3(\mathcal{A}_p + \mathcal{S}_p)) + \tau(\varepsilon, d) \\ 36d\mathcal{A}_p^2 + 16d\mathcal{S}_p^2 \text{ otherwise} \end{cases}$$

$$b_3(p,\varepsilon) = \begin{cases} \det_{\mathcal{D}}(\mathcal{A}(p), 3\mathcal{S}_{\mathcal{A}(p)}) \text{ if } \mathrm{lvl}(\mathcal{A}(p), 3\mathcal{S}_{\mathcal{A}(p)}) \leq \log(3\mathcal{S}_{\mathcal{A}(p)}) + \tau(\varepsilon, d) \\ 0 \text{ otherwise} \end{cases}$$

3 结构定理与近似算法

3.1 路线图

现在所有定义都已就绪，我们将更精确地概述定理1.2的证明思路。我们的目标是证明，引理 2.2 中的分解 $\rho = \varepsilon / \log(1/\varepsilon)$ 足够精确，使得动态规划能够计算出一个遵循门户的解，并且该解是 $(1+\varepsilon)$-近似。

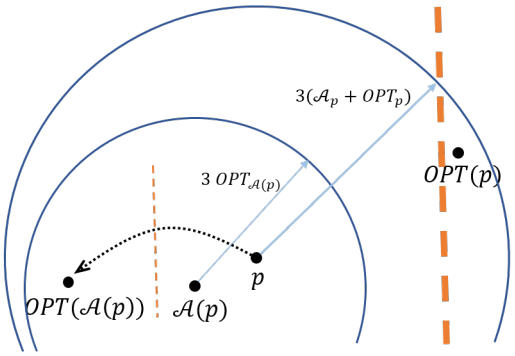

图1：对一种情况区别的示意图。 $B(p, 3(\mathcal{A}_p + \text{最优解}_p))$ 被橙色粗虚线严重切割，因此 p 无法通过门户连接到最优解(p)。然而， p 和 $\mathcal{A}(p)$并未被严重切割，因此 p可以改为连接到最优解$(\mathcal{A}(p))$，并沿着橙色细虚线绕行。这种重新分配的代价计入 $b_2(p)$。

为此，我们证明可以找到一个解 S^*，使其满足：(1) 它是 $(1+\varepsilon)$-近似；(2) 从每个点到 S^* 的绕行距离不超过该点的预算。最后，只需给出预算的上界，这只需应用引理 2.2 和一阶矩分析即可。为了找到这样的解 S^*，我们从最优解出发（它满足条件 (1)，但不满足条件 (2)），并对其进行转换，使每个客户点都能经由门户连接。

首先，考虑一个简单的情况：客户点 p 没有被严重切割，且 $B(p, 3(\mathcal{A}_p + \text{最优解}_p))$ 没有被严重切割。此时，从 p 绕行到最优解$_p$ 所产生的代价可以计入 $b_2(p, \varepsilon)$，因为距离 p 最近的最优中心位于后一个球内。当这个球被严重切割时，我们不再用其最近中心最优解(p)服务 p，而改用最优解$(\mathcal{A}(p))$服务它。我们证明，这笔代价总能计入 p的预算。事实上，在 $\mathcal{A}(p)$本身没有被严重切割的子情形中，我们可以证明， p 与最优解$(\mathcal{A}(p))$ 之间的绕行代价相对于二者间的距离是可承受的；而这段距离可以计入预算 $b_2(p, \varepsilon)$。图1展示了这一子情形。然而，在 $\mathcal{A}(p)$也被严重切割的子情形中，我们别无选择，只能将该中心加入最优解。这会导致解中的中心略多——从期望来看$(1+\varepsilon)k$ 如此，因为每个中心被严重切割的概率为 ε ——因此我们需要移除一些最优解中的中心，同时避免代价大幅增加。这一步已在 [13], 中介绍，我们将在第3.2节中回顾。

这基本上结束了 p 未被严重切割的情形。当它被严重切割时，分解$\mathcal{D}$ 对其距离的近似效果很差；不过这种情况发生的概率极小，我们可以将 p 移动到 $\mathcal{A}(p)$ （将这一移动计入 $b_1(p, \varepsilon)$）。随后，与前述类似的分类讨论表明， $\mathcal{A}(p)$ 到最优解的绕行可以计入 $b_2(p, \varepsilon)$ 和 $b_3(p, \varepsilon)$。

在本节的剩余部分，我们说明如何从最优解中移除一些中心（以便随后加入 $\mathcal{A}$的严重切割的中心），然后证明结构定理，该定理表明存在满足(1)和(2)的解。我们在第3.4节阐述依赖于该结构定理并能计算出良好近似的动态规划，从而完成定理1.2的证明。

3.2 准备实例

我们考虑将最优解的中心到 $\mathcal{A}$ 的映射，定义如下：对最优解中的任意 $f \in$ ，令 $\mathcal{A}(f)$ 表示 $\mathcal{A}$ 中距离 f最近的中心。回顾一下，对于客户点 c， $\mathcal{A}(c)$ 是在 $\mathcal{A}$中服务 c 的中心。

对于任意中心 ℓ $\mathcal{A}$的，定义 $\psi(\ell)$ 为映射到 ℓ的最优解的中心集合，即 ，

$\psi(\ell) = \{f \in \text{最优解} \mid \mathcal{A}(f) = \ell\}$。定义 $\mathcal{A}^1$ 为满足以下条件的中心 ℓ 集合：对于 $\mathcal{A}$，存在唯一的$f \in$ 最优解，使得 $\mathcal{A}(f) = \ell$，即 $\mathcal{A}^1 = \{\ell \in \mathcal{A} \mid |\psi(\ell)| = 1\}$。令 $\mathcal{A}^0 = \{\ell \in \mathcal{A} \mid |\psi(\ell)| = 0\}$，且 $\mathcal{A}^{\geq 2} = \mathcal{A} \setminus (\mathcal{A}^1 \cup \mathcal{A}^0)$。类似地，定义 $\text{OPT}^1 = \{f \in \text{OPT} \mid \mathcal{A}(f) \in \mathcal{A}^1\}$ 和 $\text{OPT}^{\geq 2} = \{f \in \text{OPT} \mid \mathcal{A}(f) \in \mathcal{A}^{\geq 2}\}$。注意，由于 $|\text{OPT}^1| = |\mathcal{A}^1|$ 且不妨设 $|\text{OPT}| = |\mathcal{A}| = k$，因此 $|\text{OPT}^{\geq 2}| = |\mathcal{A}^0| + |\mathcal{A}^{\geq 2}|$。

如 [13], 所示，可以从最优解中移除部分中心$^{\geq 2}$，从而得到以下结果：

引理 3.1 （参见 [13]中的步骤 1 和引理 4.1）。 存在一个集合 $\mathcal{H} \subseteq$ 最优解$^{\geq 2}$，使得：

- $|\mathcal{H}| = \lfloor \varepsilon \cdot |\text{最优解}^{\geq 2}|/2 \rfloor$,
- 对于任意 $\ell \in \mathcal{A}$， $\psi(\ell)$ 中距离 ℓ最近的中心不在 $\mathcal{H}$中，
- 令最优解′ = 最优解 $\setminus \mathcal{H}$：最优解′ 的代价至多为 $(1 + O(\varepsilon))\text{cost}(\text{最优解}) + O(\varepsilon)\text{cost}(\mathcal{A})$。

此外，我们对输入集合 P 进行变换，使其具有更多结构，如下所述。给定一个随机化的~层次分解 $\mathcal{D}$ 和一个解 $\mathcal{A}$，我们按如下方式定义集合 P。对于每个点 p,

$$\widetilde{p} = \begin{cases} \mathcal{A}(p) \text{ if } p \text{ is badly cut w.r.t. } \mathcal{D} \text{ and } \mathcal{A} \\ p \text{ otherwise.} \end{cases}$$

We n注意， $\widetilde{P} := \{\widetilde{p} : p \in P\}$是一个随机变量，取决于 $\mathcal{D}$的随机性。

3.3 结构化解的构造

利用引理 3.1，现在可以证明近似最优的门户遵循解的存在性。令 $\mathcal{A}$为一个固定解，OPT′ 为引理 3.1 针对 P. 令 $\mathcal{D}$ 为一个随机分层分解，并令 B_D 为 $\mathcal{A}$ 中相对于 OPT 被严重切割的中心集合′，也就是说， $c \in B_D$ 当球$B(c, 3\text{OPT}'_c)$ 在高于 log(3OPT的某个层级被切割时。

定义 3.2。 对于 $\varepsilon > 0$，如果 $\mathcal{D}$ 具有 ε-小失真，则称

1. 相对于 $\mathcal{A}$ 和 OPT' 的总预算有界： $\sum_{p \in P} b(p, \varepsilon) \leq \varepsilon(\text{cost}(P, OPT) + \text{cost}(P, \mathcal{A}))$。*2.* 存在一个解 S^*，其中包含 B_D，且满足 $\sum_{p \in P} \left(\text{dist}(p, \widetilde{p}) + \text{dist}(\widetilde{p}, S^*)\right)^2 \leq (1+\varepsilon)\text{cost}(P, OPT) + \varepsilon\text{cost}(P, \mathcal{A})$; *3.* 对于每个点 $p \in P$，其到 S^* 的绕行距离至多为 $b(p, \varepsilon)$，即存在一个中心 $s \in S^*$，它在层级 i从 $\widetilde{p}$处被切割，且满足 $(\text{dist}(p, \widetilde{p}) + \text{dist}(\widetilde{p}, s) + \varepsilon 2^i)^2 \leq \left(\text{dist}(p, \widetilde{p}) + \text{dist}(\widetilde{p}, S^*)\right)^2 + b(p, \varepsilon)$。

如引言中所述，我们将按以下方式利用这些性质：标准动态规划将计算使遵守门户的解达到最小值 $\widetilde{\text{cost}}(P, S) := \sum_p (\text{dist}(p, \widetilde{p}) + \text{dist}(\widetilde{p}, S))^2$。使解遵守门户会产生误差 $\varepsilon 2^i$，对于在层级 i被切割的两个点：性质3 因此将此误差限制在 $b(p, \varepsilon)$以内。定义3.2的性质1确保总预算有界，而性质2 则将 $\widetilde{\text{cost}}$ 与 cost联系起来。

低维空间中 k-均值问题的主要结构定理如下：

定理3.3 (结 构定理)。 对于任意 $\varepsilon > 0$， $\mathcal{D}$ 以至少 $2/$ 的概率具有 ε-小失真 3.

论文其余部分将用于证明该定理。

3.3.1 预算界定

引理 3.4. 固定两个解 $\mathcal{A}$，并令 $b(p,\varepsilon)$ 为 p 关于解 $\mathcal{A}$ 和 S 的预算。以 $2/3$ 的概率（关于 D 的随机性），$\sum_p b(p,\varepsilon) = O(d^2\log(1/\varepsilon)\cdot\varepsilon)(\mathrm{cost}(P,\mathcal{A})+\mathrm{cost}(P,S))$

证明。我们先界定预算的期望，然后使用马尔可夫不等式。回顾一下，任意球$B(x,r)$ 在层级 i 被切割的概率为 $dr/2^i$，被严重切割的概率为 ε。此外，当$B(x,r)$在层级 i被切割时，则有 $\det_D(x,r) = \varepsilon r 2^i + \varepsilon^2 2^{2i}$。

为简化分析，我们先分析变量的期望

$$\beta(x,r) := \begin{cases} \det_D(x,r) \text{ when } \mathrm{lvl}(x,r) \le \log(r)+\tau(\varepsilon,d) \\ 0 \text{ otherwise.} \end{cases}$$

为此，我们区分以下几种情况：$\mathrm{lvl}(x,r)$ 至多为 $\log(r)$、介于 $\log(r)$ 和 $\log(r)+\tau(\varepsilon,d)$之间，以及至少为 $\log(r)+\tau(\varepsilon,d)$。由此得到以下上界：

$$\begin{aligned}
\mathbb{E}[\beta(x,r))] &\le \sum_{i\le\log(r)} \det_D(x,r) + \sum_{i=\log(r)}^{\log(r)+\tau(\varepsilon,d)} \det_D(x,r)\cdot\Pr[D \text{ cuts } B(p,r) \text{ at a level } i] \\
&\le \sum_{i\le\log(r)} \left(\varepsilon 2^i r + \varepsilon^2 2^{2i}\right) + \sum_{i=\log(r)}^{\log(r)+\tau(\varepsilon,d)} \left(\varepsilon 2^i r + \varepsilon^2 2^{2i}\right)\cdot\frac{dr}{2^i} \\
&\le 3\varepsilon r^2 + \sum_{i=\log(r)}^{\log(r)+\tau(\varepsilon,d)} \left(\varepsilon d r^2 + \varepsilon^2 2^i\cdot dr\right) \\
&\le 3\varepsilon r^2 + \tau(\varepsilon,d)\cdot\varepsilon d r^2 + \varepsilon^2\cdot dr\cdot 2^{\log(r)+\tau(\varepsilon,d)} \\
&= O(d^2\log(1/\varepsilon))\varepsilon r^2,
\end{aligned}$$

其中最后一行用到了 $2^{\tau(\varepsilon,d)} = 2^{\log(d)+\log(1/\varepsilon)} = O(d/\varepsilon)$。

由于 $b_1(p,\varepsilon) = \beta(p,3\mathcal{A}_p)$和 $b_3(p,\varepsilon) = \beta(\mathcal{A}(p),3S_{\mathcal{A}(p)})$，我们直接得到 $\mathbb{E}[b_1(p,\varepsilon)+b_3(p,\varepsilon)] \le O(d^2\log(1/\varepsilon)\varepsilon)(\mathcal{A}_p^2+S_p^2)$。界定 $b_2(p,\varepsilon)$稍微复杂一些：$b_2(p,\varepsilon) = \beta(p,\mathcal{A}_p+S_p)+X(p,\varepsilon)$，其中当 $\mathrm{lvl}(p,3(\mathcal{A}_p+S_p)) > \log(3(\mathcal{A}_p+S_p))+\tau(\varepsilon,d)$时，$X(p,\varepsilon) := 36d\mathcal{A}_p^2+16dS_p^2$。因此，只需计算 $\mathbb{E}[X(p,\varepsilon)]$。

引理中分解的性质 2.2 保证 $\mathrm{lvl}(p,3(\mathcal{A}_p+S_p)) > \log(3(\mathcal{A}_p+S_p))+\tau(\varepsilon,d)$的概率至多为 ε，因此 $\mathbb{E}[X(p,\varepsilon)] \le \varepsilon\cdot(36d\mathcal{A}_p^2+16dS_p^2)) = O(\varepsilon d(\mathcal{A}_p^2+S_p^2))$。

该 ₀refore, $\mathbb{E}[b_2(p,\varepsilon)] = \mathbb{E}[\beta(p,3(\mathcal{A}_p+S_p)] + \mathbb{E}[X(p,\varepsilon)] \le O(d^2\log(1/\varepsilon)\cdot\varepsilon)(\mathcal{A}_p^2+S_p^2)$

将这些不等式分别对 b_1、b_2、b_3 求和，并对所有 $p\in P$ 求和，可得 $\mathbb{E}[\sum_{p\in P} b(p,\varepsilon)] \le O(d^2\log(1/\varepsilon)\cdot\epsilon)(\mathrm{cost}(P,\mathcal{A})+\mathrm{cost}(P,S))$。由马尔可夫不等式可得该引理。 □

3.3.2 S^*的构造

为了构造 S^*，我们从引理3.1 得到的最优解 OPT 出发，并执行以下两个步骤。

对于任意 $\ell\in\mathcal{A}$，定义 f_ℓ 为其在 $\psi(\ell)$中的最近中心，平局时任意打破。我们从S^* = 最优解′ 开始，该最优解由引理 3.1得到。注意，对于每个 $\ell\in\mathcal{A}^1\cup\mathcal{A}^{\ge 2}$，引理 3.1的第二条性质保证 f_ℓ 是最优解的一个中心′。 S^* 随后按如下方式变换：

- **步骤 1.** 对于每个中心 $\ell \in B_D \setminus \mathcal{A}^0$（即，$\ell$ 相对于 $\mathcal{D}$ 和最优解′被严重切割，且 $\psi(\ell) \neq \emptyset$），在 S^*中将 f_ℓ替换为 ℓ。
- **步骤 2.** 将所有相对于 $\mathcal{D}$ 和最优解′ 被 $\mathcal{A}^0$ 严重切割的中心添加到 S^*。

该过程的两个关键性质如下：

事实 3.5 (命题 *4.3* 和 *4.4*)。 以概率 $4/5$，$|S^*| \leq k$，因此 S^* 是一个有效的解，并且 $\text{cost}(P, S^*) \leq (1+\varepsilon)\text{cost}(P, \text{最优解}) + \varepsilon\text{cost}(P, \mathcal{A})$。此外，对于任意点 p，有 $\text{dist}(p, S^*) \leq 2\,\text{dist}(p, \text{最优解}') + \text{dist}(p, \mathcal{A})$。

此外，S^* 的构造确保：

事实 3.6. 对于任意点 p，

$$\begin{aligned}\text{dist}(p, \widetilde{p}) + \text{dist}(\widetilde{p}, S^*) &\leq 3\mathcal{A}_p + 2OPT'_p \\ \text{dist}(p, S^*(\mathcal{A}(p))) &\leq 3\mathcal{A}_p + 2OPT'_p.\end{aligned}$$

此外，对于任意中心 $\ell \in \mathcal{A}$，有 $\text{dist}(\ell, S^*) \leq 2\,\text{dist}(\ell, \text{最优解}')$。

证明。 第一个不等式由以下事实得出：$\text{dist}(\widetilde{p}, S^*) \leq \text{dist}(\widetilde{p}, S^*(p)) \leq \text{dist}(p, \widetilde{p}) + \text{dist}(p, S^*(p))$以及事实 3.5。第二个不等式的推导类似：$\text{dist}(p, S^*(\mathcal{A}(p))) \leq \mathcal{A}_p + \text{dist}(\mathcal{A}(p), S^*(\mathcal{A}(p)) \leq \mathcal{A}_p + \text{dist}(\mathcal{A}(p), S^*(p)) \leq 2\mathcal{A}_p + \text{dist}(p, S^*) \leq 3\mathcal{A}_p + 2\text{最优解}'_p$。

对于第三个不等式，设 ℓ 是 $\mathcal{A}$ 的一个中心，$\text{OPT}'(\ell)$ 是其在 OPT' 中的最近中心。接下来，要么 $\text{OPT}'(\ell) \in S^*$，此时由于 $\text{dist}(\ell, S^*) = \text{dist}(\ell, \text{OPT}')$，结论成立；要么 $\text{OPT}'(\ell)$ 在步骤 1 中被移除。在这种情况下，它被替换为中心 ℓ'，该中心是 $\mathcal{A}$ 中距离 $\text{OPT}'(\ell)$ 最近的中心：特别地，$\text{dist}(\ell', \text{OPT}'(\ell)) \leq \text{dist}(\ell, \text{OPT}'(\ell))$。因此，有：

$$\begin{aligned}\text{dist}(\ell, S^*) &\leq \text{dist}(\ell, \ell') \leq \text{dist}(\ell, \text{OPT}'(\ell)) + \text{dist}(\ell', \text{OPT}'(\ell)) \\ &\leq 2\,\text{dist}(\ell, \text{OPT}')\end{aligned}$$

□

现在我们可以利用这个解 S^* 来证明 $\mathcal{D}$ 具有 ε-小失真。

3.3.3 界定 S^*的代价

引理 3.7. S^* 包含 B_D，且以至少 $4/5$ 的概率，有

$$\sum_{p \in P} \left(\text{dist}(p, \widetilde{p}) + \text{dist}(\widetilde{p}, S^*)\right)^2 \leq (1+\varepsilon)\text{cost}(P, OPT) + \varepsilon\text{cost}(P, \mathcal{A})$$

证明。 根据构造，S^* 确实包含 B_D。为简化方程，我们记 p b.c.，当 p 相对于 $\mathcal{A}$ 被严重切割时（此时 $\widetilde{p} = \mathcal{A}(p)$）。有：

$$\sum_{p \in P} \left(\text{dist}(p, \widetilde{p}) + \text{dist}(\widetilde{p}, S^*)\right)^2 \leq \sum_{p \in P} \text{dist}(p, S^*)^2 + \sum_{p \text{ b.c.}} \left(\text{dist}(p, \widetilde{p}) + \text{dist}(\widetilde{p}, S^*)\right)^2. \tag{1}$$

根据事实3.5，第一项至多为 $(1+\varepsilon)\mathrm{cost}(P, \text{最优解}) + \varepsilon\mathrm{cost}(P, \mathcal{A})$：因此，我们只关注第二项。为此，我们使用事实3.6：

$$\sum_{p \text{ b.c.}} \left(\mathrm{dist}(p, \widetilde{p}) + \mathrm{dist}(\widetilde{p}, S^*)\right)^2 \le \sum_{p \text{ b.c.}} \left(3\mathcal{A}_p + 2\mathrm{OPT}'_p\right)^2 \le 18 \sum_{p \text{ b.c.}} \mathcal{A}_p^2 + \mathrm{OPT}'^2_p$$

每个点 p 被严重切割的概率为 ε：因此，期望为

$$\mathbb{E}[18 \sum_{p \text{ b.c.}} \mathcal{A}_p^2 + \mathrm{OPT}'^2_p] \le 18\varepsilon \sum_{p \in P} \mathcal{A}_p^2 + \mathrm{OPT}'^2_p \le 18\varepsilon(\mathrm{cost}(P, \mathcal{A}) + \mathrm{cost}(P, \mathrm{OPT}'))$$

根据马尔可夫不等式，以概率 $4/5$ 有

$$\sum_{p \text{ b.c.}} \left(\mathrm{dist}(p, \widetilde{p}) + \mathrm{dist}(\widetilde{p}, \mathrm{OPT}')\right)^2 \le 18 \cdot 5 \cdot \varepsilon(\mathrm{cost}(P, \mathcal{A}) + \mathrm{cost}(P, \mathrm{OPT}')).$$

将此界代入式（1），并利用引理 3.1 对最优解的代价进行界定′，可得：

$$\begin{aligned}\sum_{p \in P} \left(\mathrm{dist}(p, \widetilde{p}) + \mathrm{dist}(\widetilde{p}, S^*)\right)^2 &\le (1+\varepsilon)\mathrm{cost}(P, \mathrm{OPT}) + \varepsilon\mathrm{cost}(P, \mathcal{A}) + 18 \cdot 5 \cdot \varepsilon(\mathrm{cost}(P, \mathcal{A}) + \mathrm{cost}(P, \mathrm{OPT}')) \\ &= (1 + O(\varepsilon))\mathrm{cost}(P, \mathrm{OPT}) + O(\varepsilon)\mathrm{cost}(P, \mathcal{A}).\end{aligned}$$

□

3.3.4 界定绕行

~现在我们证明，通过分解的门户连接 P 与 S^* 所产生的绕行 $\mathcal{D}$ 在定义2.5中规定的预算范围内。

引理 3.8。 设 $P \subset \mathbb{R}^d$, $\mathcal{A}$ 是针对 k-均值问题的一个解，其中 P，且 $\mathcal{D}$ 是一个随机分解。设 S^* 为引理3.7给出的解。

那么，对于任意 $p \in P$，都存在一个中心 $s \in S^*$，它在某个层级 i从 $\widetilde{p}$切割而来，使得 $(\mathrm{dist}(p, \widetilde{p}) + \mathrm{dist}(\widetilde{p}, s) + \varepsilon 2^i)^2 \le (\mathrm{dist}(p, \widetilde{p}) + \mathrm{dist}(\widetilde{p}, S^*))^2 + b(p, \varepsilon)$。

证明。 回顾一下，当球 $B(x, 3\,\mathrm{dist}(x, S))$ 在高于 $\log(9\,\mathrm{dist}(x, S)) + \tau(\varepsilon, d)$的层级被~切割时，点 $x \in \mathbb{R}^d$ 相对于 $\mathcal{D}$和集合 S被严重切割。再回顾一下我们对 P和 S^*的构造：当点 p相对于 $\mathcal{D}$ 和 $\mathcal{A}$被严重切割时，$\widetilde{p} = \mathcal{A}(p)$；当中心 $\ell \in \mathcal{A}$相对于 $\mathcal{D}$ 和最优解′被严重切割时，$\ell \in S^*$。

为证明该引理，我们进行分类讨论，首先依据 p 相对于 $\mathcal{D}$和 $\mathcal{A}$是否被严重切割。在其被严重切割的情形中，成立 $\widetilde{p} = \mathcal{A}(p)$。随后，我们有如下情形。

1. 若 $\mathcal{A}(p)$ 相对于 $\mathcal{D}$ 和最优解′被严重切割，则 $\mathcal{A}(p) \in S^*$：因此，$\widetilde{p} = \mathcal{A}(p) = S^*(p)$，且 $\widetilde{p}$ 和 $S^*(p)$未被切割，例如 $i = -\infty$。因此，$s = \mathcal{A}(p)$满足该引理，其中 $(\mathrm{dist}(p, \widetilde{p}) + \mathrm{dist}(\widetilde{p}, s) + \varepsilon 2^i)^2 = \mathrm{dist}(p, \widetilde{p})^2$。

2. $\mathcal{A}(p)$ 相对于 $\mathcal{D}$ 和最优解′并未被严重切割。设 $s = S^*(\mathcal{A}(p))$，且 i 为 $\mathcal{A}(p)$ 与 s被切割时所处的层级。我们记 $(\mathrm{dist}(p, \widetilde{p}) + \mathrm{dist}(\widetilde{p}, s) + \varepsilon 2^i)^2 = (\mathcal{A}_p + \mathrm{dist}(\widetilde{p}, S^*))^2 + 2\varepsilon 2^i \cdot \mathcal{A}_p + 2\varepsilon 2^i \cdot S^*_{\mathcal{A}(p)} + \varepsilon^2 2^{2i}$。

根据事实3.6， $s \in B(\mathcal{A}(p), 2\text{最优解}'_{\mathcal{A}(p)})$，因此 $i \le i' := \mathrm{lvl}\left(\mathcal{A}(p), 3\text{最优解}'_{\mathcal{A}(p)}\right)$。结合 $\mathcal{A}(p)$ 并未被严重切割这一事实，因此有 $2\varepsilon 2^i \cdot \mathcal{S}^*_{\mathcal{A}(p)} + \varepsilon^2 2^{2i} \le 2\varepsilon 2^{i'} \cdot 2\text{最优解}'_{\mathcal{A}(p)} + \varepsilon^2 2^{2i'} \le b_3(p)$。于是，只需对 $2\varepsilon 2^i \cdot \mathcal{A}_p$ 给出上界。为此，我们细化分类讨论，依据球 $B(p, 3(\mathcal{A}_p + \text{最优解}'_p))$ 是否被严重切割，并使用来自 $b_2(p)$ 的预算。

（a）如果该球没有被严重切割，那么我们注意到事实 3.6 保证 s 和 $\mathcal{A}(p)$ 都属于 $B(p, 3(\mathcal{A}_p + \text{最优解}'_p))$，因此 $i \le i'' := \mathrm{lvl}(p, 3(\mathcal{A}_p + \text{最优解}'_p))$。因此， $2\varepsilon 2^i \cdot \mathcal{A}_p \le \varepsilon 2^{i''} 3(\mathcal{A}_p + \text{最优解}'_p) \le b_2(p)$。（b）如果该球被严重切割，那么除了最优解之外，我们还使用 $i \le i' \le \log(3\text{最优解}'_{\mathcal{A}(p)}) + \tau(\varepsilon, d)$，如下所示：$'_{\mathcal{A}(p)} \le \mathcal{A}_p + \text{最优解}'_p$ 如下： $2\varepsilon 2^i \cdot \mathcal{A}_p \le 6\varepsilon \cdot \text{最优解}'_{\mathcal{A}(p)} \cdot 2^{\tau(\varepsilon,d)} \mathcal{A}_p \le 6d(\mathcal{A}_p + \text{最优解}'_p)\mathcal{A}_p \le b_2(p)$。

综合这些情况可得，当 $\mathcal{A}(p)$ 未被严重切割时， $(\mathrm{dist}(p, \widetilde{p}) + \mathrm{dist}(\widetilde{p}, s) + \varepsilon 2^i)^2 \le (\mathcal{A}_p + \mathrm{dist}(\widetilde{p}, \mathcal{S}^*))^2 + b_2(p) + b_3(p)$。

现在我们转向 p 未被严重切割的情况。在这种情况下， $\widetilde{p} = p$：因此我们只需给出 $\left(\mathrm{dist}(p, \mathcal{S}^*) + \varepsilon 2^i\right)^2$ 的上界。此外，球 $B(p, 3\mathcal{A}_p)$ 在层级 $\mathrm{lvl}(p, 3\mathcal{A}_p) \le \log(3\mathcal{A}_p) + \tau(\varepsilon, d)$ 处被切割。我们根据球 $B(p, 3(\mathcal{A}_p + \text{最优解}'_p))$ 是否被严重切割进行分类讨论。注意，根据事实 3.5， $\mathcal{S}^*(p) \in B(p, 3(\mathcal{A}_p + \text{最优解}'_p))$。

1. 如果 $B(p, 3(\mathcal{A}_p + \text{最优解}'_p))$ 没有被严重切割，那么 p 和 $s = \mathcal{S}^*(p)$ 在不高于 $i = \mathrm{lvl}(p, 3(\mathcal{A}_p + \text{最优解}'_p))$ 的层级上被切割，并且 s 满足引理： $(\mathrm{dist}(p, s) + \varepsilon 2^i)^2 \le \mathrm{dist}(p, \mathcal{S}^*)^2 + b_2(p)$。
2. 如果 $B(p, 3(\mathcal{A}_p + \text{最优解}'_p))$ 被严重切割，我们证明 $s = \mathcal{S}^*(\mathcal{A}(p))$ 满足引理。具体而言，

(a) 如果 $\mathcal{A}(p)$ 被严重切割，那么 $s = \mathcal{A}(p)$；在这种情况下，由于 p 没有被严重切割， p 和 s 在不高于 $i = \mathrm{lvl}(p, 3\mathcal{A}_p)$ 和 $(\mathrm{dist}(p, s) + \varepsilon 2^i)^2 \le \mathrm{dist}(p, \mathcal{S}^*)^2 + b_1(p) + b_2(p)$ 的层级上被切割。(b) 如果 $\mathcal{A}(p)$ 没有被严重切割，那么可以将 p 分配给 $s = \mathcal{S}^*(\mathcal{A}(p))$。这里的关键观察是，它们在不高于 $i = \max\left(\mathrm{lvl}(p, \mathcal{A}_p), \mathrm{lvl}(\mathcal{A}(p), 3\text{最优解}'_{\mathcal{A}(p)})\right)$ 的层级上被切割：事实上，根据事实 3.6 s 位于球 $B(\mathcal{A}(p), 3\text{最优解}'_{\mathcal{A}(p)})$ 中，因此 s 和 $\mathcal{A}(p)$ 在层级 i 属于同一个单元；类似地， p 和 $\mathcal{A}(p)$ 在层级 i 属于同一个单元，因此， p 和 s 在不高于 i 的层级上被切割。此外，根据事实 3.6， p 和 s 满足 $\mathrm{dist}(p, s) \le 3\mathcal{A}_p + 2\text{最优解}'_p$。因此，利用对任意 a、 b、 c、 $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ 均有：

$$\begin{aligned}(\mathrm{dist}(\widetilde{p}, s) + \varepsilon 2^i)^2 &\le (3\mathcal{A}_p + 2\mathrm{OPT}'_p + \varepsilon 2^i)^2 \\ &\le 27\mathcal{A}_p^2 + 12\mathrm{OPT}'^2_p + 3\varepsilon^2 2^{2i}.\end{aligned}$$

现在，由于 $B(p, 3(\mathcal{A}_p + \text{最优解}'_p))$ 遭到严重切割， $27\mathcal{A}_p^2 + 12\text{最优解}'^2_p \le b_2(p)$。此外，根据 i 的上界，我们有 $3\varepsilon^2 2^{2i} \le b_1(p) + b_3(p)$，因为 p 和 $\mathcal{A}(p)$ 都未遭到严重切割。因此， $(\mathrm{dist}(\widetilde{p}, s) + \varepsilon 2^i)^2 \le b_1(p) + b_2(p) + b_3(p)$，至此该情形及引理得证。

□

3.4 算法：如何使用结构定理

假设 $\mathcal{D}$ 具有 ρ-小失真，对于 $\rho = O(\varepsilon / \log(1/\varepsilon))$ （其中 O 隐藏了 d 中的依赖关系），并令 $\mathcal{S}^*$ 为定义3.2所给出的解。如前所述，我们的算法将采用标准的动态规划，按如下方式运行。分解 $\mathcal{D}$ 将增添一组

门户（定义见引理 2.2），其中 $\rho = \varepsilon/\log(1/\varepsilon)$。动态规划（稍后将详细描述）计算出最优的遵循门户的解，即每条连接客户点与中心的路径仅在门户处穿过 $\mathcal{D}$ 的边界的解。更准确地说，它计算出具有（近似）最小 $\widetilde{\text{cost}}(P,S) := \sum_p (\text{dist}(p,\widetilde{p}) + \text{dist}(\widetilde{p},S))^2$

定理 3.9。 存在一个运行时间为 $(2/\varepsilon)^{O\left(\log(1/\varepsilon)/\varepsilon\right)^{d-1}} \cdot n\log^3 n$ 的算法，可以计算出一个遵循门户的解 S，使得 $\widetilde{\text{cost}}(P,S) \le (1+\varepsilon)\min_{S'} \widetilde{\text{cost}}(P,S')$。

门户集合的精确性和嵌套性确保，当两个点 x 和 y 在层级 i 被切割时，x 和 y 之间遵循门户的路径长度为 $\|x-y\| + \rho 2^{i+1}$。此外，定理 3.9 中的算法计算出的解至少与 S^* 一样好：根据定义 3.2 的性质 3，S^* 的遵循门户 $\widetilde{\text{cost}}$ 至多为 $\sum_p (\text{dist}(p,\widetilde{p}) + \text{dist}(\widetilde{p},S^*))^2 + b(p,\rho)$。性质 1 确保总预算至多为 $\rho(\text{cost}(P,\text{OPT}) + \text{cost}(P,\mathcal{A})) = O(\rho)\text{cost}(P,\text{OPT})$，因为 $\mathcal{A}$ 是一个 $O(1)$-近似。结合定义 3.2 的性质 2，这表明算法计算出的解的遵循门户 $\widetilde{\text{cost}}$ 至多为 $(1+O(\varepsilon))\text{cost}(P,\text{OPT})$。

最后，我们指出，满足门户约束的 $\widetilde{\text{cost}}$ 是代价的上界：沿直线行进而非经过门户只会降低代价；并且对于任意点 p, $\text{dist}(p,S) \le \text{dist}(p,\widetilde{p}) + \text{dist}(\widetilde{p},S)$。因此，动态规划计算出的解是$k$-means 的 $(1+O(\varepsilon))$-近似。

动态规划：定理的证明 3.9 用于证明定理的算法3.9 是一个相当标准的动态规划（DP），详细描述见 [13]。这里我们简要概述一下。对于动态规划，我们将把分解 $\mathcal{D}$ 表示为一棵二叉树，而不是 2^d叉树。[3]

一个配置由以下参数组成：

- 四叉树的一个节点；设 D 为其直径，
- 对于该节点的每个门户 p_i，有两个值 ℓ_i 和 s_i，它们是 εD 的倍数，且取值范围为 $[0, D/\varepsilon + 1/\varepsilon]$，
- 一个布尔标志，表示单元的最近中心距离是否至少为 D/ε，
- 目标成本 c_0，其幂为 $(1+\varepsilon/\log n)$，取值范围为 $[\text{cost}(\mathcal{A})/n, (1+\varepsilon)\text{cost}(\mathcal{A})]$。

值 ℓ_i 编码了从门户 p_i 到部分内部最近中心的距离，以及 s_i 到部分外部最近中心的距离。

对于每种配置，动态规划会计算在单元中放置多少个中心才能匹配门户参数，并使解的 $\widetilde{\text{cost}}$ 至多为 c_0。

动态规划的第一个基本情形是节点为叶节点时：那么，如 [13], 所述，单元中有来自 $\widetilde{P}$ 的一个点，两种选择分别是在单元中设置一个中心，使得 $\widetilde{\text{cost}} = 0$，或者通过遵循门户的路径连接到外部的中心，其距离由以下值给出：s_i。

动态规划的另一个基本情形是：当布尔标志表明最近的中心距离超过 D/ε 时，可以假设单元中的每个点都位于该单元的中心，就像其中只有一个点一样。随后，按照叶节点的方式计算该配置的值。

[3]严格来说，在 $\mathcal{D}$ 中，每个单元都被分解为 2^d 个子节点，这些子节点是通过用 d 个超平面 H_1, ..., H_d对其进行切割得到的。我们不采用这种表示方式，而是先沿着 H_1进行切割，得到两个子节点，再分别沿着 H_2对它们进行切割，使每个子节点各自再产生两个子节点，依此类推。

在非基例中，递归过程也很标准：对于给定的配置，动态规划会枚举节点两个子节点的所有可能兼容配置，即：当门户相同时， ℓ_i和 s_i 相同；否则满足三角不等式；并且两个估计成本之和等于 c_0 （允许舍入误差）。

正确性的证明可在 [13]的第 4.1 节中找到。由于门户更少，复杂度得以降低：根据引理 2.2 及 $\rho = \varepsilon/\log(1/\varepsilon)$，共有 $(\log(1/\varepsilon)/\varepsilon)^{d-1}$ 个门户。对于每个门户， $1/\varepsilon$ 是 ℓ_i 和 s_i的取值选择数。此外， $O(\log^2 n)$ 是 c_0的可能取值数。最后，~树中有 $O(nd\log n)$ 个节点，因为每个叶节点至少包含一个来自 P的点，并且树的深度为 $O(d\log n)$，因为 $|\mathcal{D}| \le \log(\mathrm{diam}(P))$， $\mathrm{diam}(P) = \mathrm{poly}(n)$，而二叉树的层级数是分解 d的 $\mathcal{D}$倍。因此，动态规划的复杂度为 $(2/\varepsilon)^{O(\log(1/\varepsilon)/\varepsilon)^{d-1}}dn\log^3 n$。

4 细粒度　　ned低维空间中的近似困难性

Th本节的目标是证明定理1.3，为方便起见，我们在此重述　　：

定理 1.3. 在 *Gap-ETH* 假设下，对于每一个整数 $d \ge 2$，存在常数 $c > 0$，使得不存在近似方案能够在给定欧几里得空间 $\mathbb{R}^d$ 中含有N 个点的离散 k-均值（或离散 k-中位数）实例以及参数 $\varepsilon > 0$时，在时间$2^{c(1/\varepsilon)^{d-1}}\mathrm{poly}(N)$内计算出 $(1+\varepsilon)$-因子近似。

我们的困难性结果基于 Gap-ETH 假设：

定义 4.1 （间隙指数时间假设（Gap-ETH） [26, 43]）。 存在常数 θ， $\delta > 0$，使得任何以 *3-CNF* 公式 φ 为输入（该公式含有 n个变量和 $O(n)$ 个子句）且能够区分 $\mathrm{SAT}(\varphi) = 1$ 与 $\mathrm{SAT}(\varphi) < 1-\delta$的算法，其运行时间必须至少为 $2^{\theta n}$。

为了在我们的归约中保持间隙，我们需要确保每个变量只出现在少数子句中。为此，3-CNF 公式是不够的，但下面的引理表明，可以假设每个变量至多出现在 3 个子句中，且每个子句的大小至多为 3（这称为 $(3,3)$-CNF）：

引理 4.2 （推论 6.1，见 [35]）。 存在常数 δ， $\gamma > 0$，使得不存在运行时间为 $2^{\gamma n}$ 的算法，在给定具有 n 个变量和 m 个子句的 *(3,3)-CNF* 公式 ϕ 时，能够区分以下情形：*(i)* ϕ 可满足，或 *(ii)* 每个赋值至少违反 δm 个子句，除非 *Gap-ETH* 不成立。

接下来，我们使用的来自 [24] 的结果如下。对于任意具有 n 个变量和 m 个子句的 (3, 3)-CNF 公式 ϕ，存在一个图 $G = (V, E)$ 和一个嵌入 $\Pi : V \to \mathbb{R}^d$，具有如下性质：

- $|E| = \Theta(n^{d/(d-1)})$,
- E 可以划分为 E_1 和 E_2，其中，对于每条边 $(u,v) \in E_1$， $\|\Pi(u) - \Pi(v)\| = 2$；对于任意边 $(u,v) \in E_2$， $\|\Pi(u) - \Pi(v)\| = 2\sqrt{2}$。
- 对于每条边 $(u,v) \in E$，中点 $(\Pi(u) + \Pi(v))/2$ 到 $\Pi(V)$ 中任意其他不与该边关联的嵌入顶点的欧几里得距离平方至少为 4。

嵌入性质证明概要，来自 *[24]*。 基于 de Berg 等人的框架，我们按如下方式构造图 G 及其嵌入。首先，将公式 ϕ 的关联图画在一个网格中，并利用立方体布线定理（ [24]中的定理3.3），将其布线到一个 d 维网格，其边

长为 $O(n^{\frac{1}{d-1}})$。该诱导网格图的顶点总数为 $O(n^{\frac{d}{d-1}})$，即边数 $|E|$ 以 $\Theta(n^{\frac{d}{d-1}})$为上界。

该构造将顶点和边替换为特定的组件：

- 每个变量 v_i 都替换为一个变量构件，它是一个长度为八的环。
- 每个包含三个文字的子句都替换为一个长度为九的环。
- 每个包含两个文字的子句都替换为一条边。
- 连接这些构件的导线用奇数长度的简单路径表示。

该图嵌入到扩展的 d 维网格中。该网格由标准网格以及添加到二维面上的对角边组成。为精确确定空间距离，嵌入采用仿射变换对空间进行倾斜，使标准网格边（E_2）的长度恰为 $2\sqrt{2}$（平方距离为 8），而所使用的对角边（E_1）的长度恰为 2（平方距离为 4）。此外，该变换保证任意一条边的中点与任意一个非关联顶点之间的平方距离至少为 4。

□

此外，图 G 通过其局部顶点覆盖与公式 ϕ 相关联。令 W 为 G中的全局导线集合，并令 部分指任意变量构件、子句构件或导线。

- 对于任意长度为 $2t_w + 1$的导线 $w \in W$，有效的最小局部顶点覆盖必须恰好包含其中 t_w 个内部顶点。
- 对于任意子句 C，令 G_C 为由 C的构件、其变量的构件以及连接导线 $W_C \subseteq W$所诱导的子图。
- 我们可以从 G_C 的任意有效最小局部顶点覆盖中解码出一个满足赋值；该覆盖在子句环上恰好使用 5 个顶点，在每个变量环上使用 4 个顶点，并且在每条连接导线上使用 t_w 个顶点。
- 类似地，对于含 2 个文字的子句，我们可以从一个最小局部顶点覆盖中解码出一个满足赋值；该覆盖在子句线段上恰好使用 1 个顶点，在每个变量环上使用 4 个顶点，并且在每条导线上使用 t_w 个顶点。
- 总体而言，如果 ϕ 可满足（包含 n_v 个变量、 n_2 个大小为二的子句以及 n_3 个大小为三的子句），那么 G 的最小顶点覆盖大小恰好为 $k := 4n_v + 5n_3 + n_2 + \sum_{w \in W} t_w$。

注意，在图 G中，任何与满足赋值对应的最小顶点覆盖都会紧密地恰好包含每个变量环中的 4 个顶点、每个三文字子句环中的 5 个顶点，以及每个二文字子句线段中的 1 个顶点。因此，根据 G 上大小为 k 的全局最小顶点覆盖，可以高效地恢复一个满足赋值。

基于这样的图 G，我们构造一个离散聚类实例，其中客户点集合为 P，候选中心集合为 C，具体如下： $C = \{\Pi(v) : v \in V\}$ 是所有顶点的嵌入，而 P 由 E所有边的中点构成的集合并集组成；也就是说，对于任意一条边 $(u, v) \in E$，其中点 $p_{uv} := (\Pi(u) + \Pi(v))/2$都属于 P。

嵌入的性质保证了以下事实。对于任意边 $e \in E_1$，距离 p_e 最近的候选中心的距离为 1（平方距离为 1）；对于任意边 $e \in E_2$，距离最近的候选中心

到 p_e 的距离为 $\sqrt{2}$ （平方距离为 2）。因此，如果 G 有一个大小为 k 的顶点覆盖，那么最优离散 k-均值代价恰好为 $|E_1|+2|E_2|$，最优离散 k-中位数代价恰好为 $|E_1|+\sqrt{2}|E_2|$。

我们的主要结果如下，它将离散聚类解的代价与公式ϕ联系起来：

引理 4.3。 设 S 是 $G=(V,E)$ 上代价至多为 $(1+\varepsilon)\cdot(|E_1|+2|E_2|)$ （对于离散 k-均值聚类）或 $(1+\varepsilon)\cdot(|E_1|+\sqrt{2}|E_2|)$（对于离散 k-中位数聚类）的解，其中$k=4n_v+5n_3+n_2+\sum_{w\in W} t_w$。那么，我们可以在关于 $|V|$ 的多项式时间内构造一个变量赋值，使 ϕ中至少有 $m-c'\varepsilon|E|$ 个子句得到满足，其中 $c'>0$为某个常数。

证明。 我们将 S 的中心看作覆盖图 G中的边： S 的中心对应于该图中的顶点，而客户点对应于边中点。当 S 在某个部分上不诱导出有效的局部最小顶点覆盖时，我们称该部分被 S 摧毁。

我们可以将被破坏部分的数量与未覆盖边的数量联系起来，从而把聚类解的代价与有效赋值联系起来。我们将被破坏部分分为三种类型：

- **类型（A）：** 因 S 在这些部分中放置的顶点过少而被破坏的部分（不足） .
- **类型（B）：** 因 S 在这些部分中放置的顶点过多而被破坏的部分（过剩）。
- **类型（C）：** 接收的顶点数量恰好正确，但至少有 1 条内部边未被覆盖的部分。

设 D 为所有 A 型部分中的顶点总不足量， S 为 B 型部分中的顶点总过剩量。由于全局恰好使用了 k 个中心，总不足量必须等于总过剩量，即 $D=S$。设 k_A、 k_B和 k_C 分别为每种类型的部分数量，因此被破坏部分的总数为 $\delta_{parts}=k_A+k_B+k_C$。注意， $D\geq k_A$ 且 $S\geq k_B$。

我们构造中的每个部分 P都是一个子图，其中每个顶点的内部度至多为 2。设 E_P 为该部分的内部边数，k_P 为其最优局部顶点覆盖大小。如果一个类型（A）部分缺少 $d\geq 1$ 个顶点，那么它包含 k_P-d 个顶点，最多可以覆盖 $2(k_P-d)$ 条边。因此，完全未覆盖的内部边至少有 $E_P-2(k_P-d)=2d+(E_P-2k_P)$条。

我们可以通过检查每种构件类型对应的 E_P和 k_P来计算这个下界。对于变量构件，它是一个长度为八的环，其 $E_P=8$，最优覆盖大小为 $k_P=4$，因此有 $2d+(8-8)=2d\geq d$ 条未覆盖边。对于包含三个文字的子句，该构件是一个长度为九的环，最优覆盖大小为 $k_P=5$，因此有 $2d+(9-10)=2d-1$ 条未覆盖边。由于类型（A）部分按定义存在缺失，我们知道 $d\geq 1$，这保证了 $2d-1\geq d$。同样的计算也适用于其余部分：连线构件是奇数长度的简单路径（其中 $E_P=2t_w-1$ 且 $k_P=t_w$），而恰好包含两个文字的子句使用一条边作为子句构件（其中 $E_P=1$ 且 $k_P=1$）。在这两种情况下，公式计算结果仍为 $2d-1\geq d$。

因此，缺少 d 个顶点的 A 型部分总会使至少 d 条内部边完全未被覆盖。从全局求和来看，A 型部分至少贡献 $D\geq k_A$ 条未覆盖的内部边。此外，根据定义，每个 C 型部分都不会诱导出有效的局部顶点覆盖，因此它必须至少留下 1 条未被覆盖的内部边，总体上至少贡献 k_C 条未覆盖边。

由于不同部分的内部边两两不相交，我们可以将它们相加而不会重复计数。未覆盖边的总数 U 有如下界：

$$U\geq D+k_C$$

由于 $D = S$，我们知道 $D = \frac{D+S}{2} \geq \frac{k_A+k_B}{2}$。代入后得到：

$$U \geq \frac{k_A + k_B}{2} + k_C = \frac{k_A + k_B + 2k_C}{2} \geq \frac{\delta_{parts}}{2}$$

因此， G中至少有 $\delta_{parts}/2$ 条未覆盖边。

如果一个子句被破坏，或者该子句中的某个变量被破坏，或者其某个变量/导线被破坏，或者它们之间存在不匹配（从而使一条连接边未被覆盖），我们就称该子句为坏 。

设 β 为坏子句的总数。我们可以将这些子句划分为包含至少一个被破坏部分的 β_{dest} 个子句，以及没有被破坏部分但包含不匹配的 β_{mis} 个子句（因此 $\beta = \beta_{\text{dest}} + \beta_{\text{mis}}$）。由于单个被破坏部分至多对应3个坏子句，这 β_{dest} 个子句对应于至少 $\beta_{\text{dest}}/3$ 个被破坏部分。如前所述，这产生至少 $\frac{1}{2}(\beta_{\text{dest}}/3) = \beta_{\text{dest}}/6$ 条未覆盖的内部边。这 β_{mis} 个不匹配的子句具有完全有效的局部部分，但每个子句必须至少有一条未覆盖的连接边（从而产生 $\geq \beta_{\text{mis}}$ 条未覆盖的连接边）。由于内部边与连接边互不相交，未覆盖边的总数 U 至少为 $\frac{\beta_{\text{dest}}}{6} + \beta_{\text{mis}} \geq \frac{\beta_{\text{dest}}+\beta_{\text{mis}}}{6} = \frac{\beta}{6}$。

根据嵌入性质，一条未覆盖边的客户点（即边中点）与任意选定中心的平方距离至少为 4。对于离散 k-均值问题，若该边被局部覆盖，其代价将为 1（对于 E_1）或 2（对于 E_2）。由于局部覆盖边的最大标准代价为 2，一条未覆盖边会引入至少为 $4 - 2 = 2$ 的严格加性惩罚费用。因此，离散 k-均值解的总代价下界为 $|E_1| + 2|E_2| + 2\left(\frac{\beta}{6}\right) = |E_1| + 2|E_2| + \frac{\beta}{3}$。

对于离散 k-中位数问题（其最小化标准的、未平方距离之和），分析完全类似。未覆盖边的边中点与任意选定中心的距离至少为 $\sqrt{4} = 2$，而局部覆盖边的代价为 1（对于 E_1）或 $\sqrt{2}$（对于 E_2）。由于局部覆盖边的最大标准代价为 $\sqrt{2}$，一条未覆盖边会引入至少为 $2 - \sqrt{2} > 0.58$ 的严格加性惩罚费用。因此，离散 k-中位数解的总代价下界为 $|E_1| + \sqrt{2}|E_2| + 0.58\left(\frac{\beta}{6}\right) \geq |E_1| + \sqrt{2}|E_2| + \frac{\beta}{11}$。

反过来，给定一个代价至多为 $(1+\varepsilon)\cdot$最优解的解 S。对于 k-均值聚类，这意味着额外惩罚费用有上界 $\frac{\beta}{3} \leq \varepsilon(2|E_1| + 2|E_2|) = 2\varepsilon|E|$，化简后得到 $\beta \leq 6\varepsilon|E|$。对于 k-中位数聚类，惩罚费用有上界 $\frac{\beta}{11} \leq \varepsilon(\sqrt{2}|E_1| + \sqrt{2}|E_2|) = \sqrt{2}\varepsilon|E|$，化简后得到 $\beta \leq 11\sqrt{2}\varepsilon|E|$。在这两种情况下，我们都可以将坏子句的数量界定为 $c'\varepsilon|E|$，其中 $c' > 0$ 为常数（例如，对于 k-均值聚类， $c' = 6$；对于 k-中位数聚类， $c' = 16$）。

最后，我们可以高效地为所有非坏子句计算满足赋值。为此，我们依赖 [24], 的结果；他们展示了如何在给定 G的有效顶点覆盖时，为 ϕ 解码出一个有效赋值。仅考虑非坏子句时，这些子图构成一个有效的局部顶点覆盖，从而为所有 $m - c'\varepsilon|E|$ 非坏子句提供有效赋值。

□

利用引理 4.3，我们可以得出定理 1.3的证明。

证明。 设 δ， γ为引理 4.2中的常数。注意，对于一个 (3,3)-CNF 公式，子句的数量m 与变量的数量呈线性关系 n ($m = \Theta(n)$)。

反设存在一个近似方案，以 $\varepsilon > 0$ 作为输入，并在时间 $2^{c(1/\varepsilon)^{d-1}}\text{poly}(N)$ 内计算离散 k-means（或 k-median）的 $(1+\varepsilon)$-近似解，

其中 $N = |V| = \Theta(n^{d/(d-1)})$ 是实例的规模，而 $c > 0$ 是一个足够小、待确定的常数，以便导出矛盾。

固定一个包含 n 个变量和 m 个子句的 (3,3)-CNF 公式 ϕ，并按照描述构造图 G 及相应的聚类实例。我们设置近似参数 ε，使坏子句的数量 $c'\varepsilon|E|$ 恰好为 $\frac{\delta}{2}m$。由于 $|E| = \Theta(n^{d/(d-1)})$ 和 $m = \Theta(n)$，因此必须设置为：

$$\varepsilon = \Theta\left(\frac{n}{n^{d/(d-1)}}\right) = \Theta(n^{-1/(d-1)})$$

注意，这意味着 $(1/\varepsilon)^{d-1} = \Theta(n)$。

根据引理4.3，如果我们使用这个特定的 ε运行假设的近似方案，就能在多项式时间内解码出一个满足所有非坏子句的赋值。因此，该嵌入确保：如果 ϕ可满足，则最优离散 k-均值（或 k-中值）代价恰好为最优解 OPT。如果该方案返回的解的代价至多为 $(1+\varepsilon)\cdot$OPT，那么我们就能恢复一个满足除$\frac{\delta}{2}m < \delta m$ 个子句之外所有子句的赋值。

该方案使用此 ε 运行所需的时间为：

$$2^{c(1/\varepsilon)^{d-1}}\text{poly}(N) = 2^{c\cdot\Theta(n)}\text{poly}(n).$$

通过将算法运行时间中的常数 c 选得足够小，总运行时间将严格小于 $2^{\gamma n}$，这与 Gap-ETH 矛盾。□

参考文献

[1] Fateme Abbasi、Sandip Banerjee、Jaroslaw Byrka、Parinya Chalermsook、Ameet Gadekar、Kamyar Khodamoradi、D´aniel Marx、Roohani Sharma 和 Joachim Spoerhase。《一般范数目标下聚类的参数化近似方案》。载于 第*64*届 *IEEE* 计算机科学基础年会，*FOCS 2023*，圣克鲁兹，加利福尼亚州，美国，*2023*年*11*月*6*日至*9*日，第 1377–1399 页。IEEE，2023年。 doi:10.1109/FOCS57990.2023.00085。

[2] Enver Aman、Karthik C. S. 和 Sharath Punna。《k-着色与欧几里得 k-均值之间的联系》。载于 Timothy M. Chan、Johannes Fischer、John Iacono 和 Grzegorz Herman 编，第*32*届欧洲算法年会，*ESA 2024*，*2024*年*9*月*2*日至*4*日，英国伦敦皇家霍洛威学院，第 308 卷，*LIPIcs*，第 9:1–9:18 页。Schloss Dagstuhl - Leibniz 信息学中心，2024年。网址：https://doi.org/10.4230/LIPIcs.ESA.2024.9, doi:10.4230/LIPICS.ESA.2024.9。

[3] Sanjeev Arora。欧几里得旅行商问题及其他几何问题的多项式时间近似方案。载于第*37*届计算机科学基础年会，*FOCS '96*，伯灵顿，佛蒙特州，美国，*1996*年*10*月*14-16*日，第2–11页。IEEE计算机学会，1996年。 doi:10.1109/SFCS.1996.548458。

[4] Sanjeev Arora。欧几里得旅行商问题及其他几何问题的多项式时间近似方案。《*ACM*杂志》，45(5):753–782，1998年。 doi:10.1145/290179.290180。

[5] 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Jeffrey Scott Vitter主编，第三十届年度*ACM*论文集

计算理论研讨会，达拉斯，得克萨斯州，美国，*1998*年*5*月*23*日至*26*日，第106–113页。ACM，1998年。 doi:10.1145/276698.276718。

[6] Pranjal Awasthi、Moses Charikar、Ravishankar Krishnaswamy 和 Ali Kemal Sinop。欧几里得 k 均值的近似困难性。载于 Lars Arge 和 J´anos Pach 编，《第*31*届国际计算几何研讨会，*SoCG 2015*，*2015*年*6*月*22*日至*25*日，荷兰埃因霍温》，LIPIcs 第34卷，第754–767页。Schloss Dagstuhl - Leibniz-Zentrum f¨ur Informatik，2015年。 doi:10.4230/LIPIcs.SOCG.2015.754。

[7] Yair Bartal 和 Lee-Ad Gottlieb。欧几里得旅行商问题的线性时间近似方案。载于《第*54*届 *IEEE* 计算机科学基础年会，*FOCS 2013*，*2013*年*10*月*26*日至*29*日，美国加利福尼亚州伯克利》，第698–706页。IEEE计算机学会，2013年。 doi:10.1109/FOCS.2013.80。

[8] Vladimir Braverman、Gereon Frahling、Harry Lang、Christian Sohler 和 Lin F. Yang。《高维动态数据流中的聚类》。载于 Doina Precup 和 Yee Whye Teh 编，第*34*届国际机器学习会议，*ICML 2017*，澳大利亚新南威尔士州悉尼，*2017*年*8*月*6*日至*11*日，第70卷，机器学习研究论文集，第576–585页。PMLR，2017。网址： http://proceedings.mlr.press/v70/braverman17a.html。

[9] Moses Charikar、Vincent Cohen-Addad、Ruiquan Gao、Fabrizio Grandoni、Euiwoong Lee 和 Ernest van Wijland。《改进的（度量）k-均值贪心近似》。载于第*66*届 *IEEE* 计算机科学基础年会，*FOCS 2025*，澳大利亚悉尼，*2025*年*12*月*14*日至*17*日，第233–240页。IEEE，2025。 doi:10.1109/FOCS63196.2025.00016。

[10] Vincent Cohen-Addad。低维*k*均值聚类的快速近似方案。载于 Artur Czumaj（编），第二十九届 *ACM-SIAM* 离散算法研讨会论文集（*SODA 2018*），美国路易斯安那州新奥尔良，*2018*年*1*月*7*日至*10*日，第 430–440 页。SIAM，2018。 doi:10.1137/1.9781611975031.29。

[11] Vincent Cohen-Addad、Arnaud de Mesmay、Eva Rotenberg 和 Alan Roytman。低维聚类的祸根。载于*2018*年 *ACM-SIAM* 离散算法研讨会（*SODA*）论文集，2018。网址： https://epubs.siam.org/doi/abs/10.1137/1.9781611975031.30， arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9781611975031.30， doi:10.1137/1.9781611975031.30。

[12] Vincent Cohen-Addad、Alessandro Epasto、Silvio Lattanzi、Vahab Mirrokni、Andres Mu˜noz Medina、David Saulpic、Chris Schwiegelshohn 和 Sergei Vassilvitskii。通过分层分离树实现可扩展的差分隐私聚类。载于 Aidong Zhang 和 Huzefa Rangwala 编，*KDD '22*：第 *28* 届 *ACM* 知识发现与数据挖掘会议，美国华盛顿特区，*2022* 年 *8* 月 *14* 日至 *18* 日，第 221–230 页。ACM，2022 年。 doi:10.1145/3534678.3539409。

[13] Vincent Cohen-Addad、Andreas Emil Feldmann 和 David Saulpic。倍增度量中聚类的近线性时间近似方案。《*ACM*杂志》，68(6):44:1–44:34，2021 年。 doi:10.1145/3477541。

[14] Vincent Cohen-Addad、Andreas Emil Feldmann 和 David Saulpic。倍增度量中聚类的近线性时间近似方案。发表于《*ACM*杂志》，第 68 卷，2021 年。

[15] Vincent Cohen-Addad、Fabrizio Grandoni、Euiwoong Lee、Chris Schwiegelshohn 和 Ola Svensson。《度量 k-中位数的 $(2+\epsilon)$-近似算法》。载于 Michal Kouck´y 和 Nikhil Bansal 编，

第*57*届*ACM*年度计算理论研讨会论文集，*STOC 2025*，捷克布拉格，*2025*年*6*月*23*日至*27*日，第615–624页。ACM，2025年。 doi:10.1145/3717823.3718299。

[16] Vincent Cohen-Addad、Anupam Gupta、Amit Kumar、Euiwoong Lee 和 Jason Li。针对 k-中位数和 k-均值的紧确 FPT 近似算法。载于 Christel Baier、Ioannis Chatzigiannakis、Paola Flocchini 和 Stefano Leonardi（编），第*46*届国际自动机、语言与编程研讨会，*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。Schloss Dagstuhl - Leibniz-Zentrum f¨ur Informatik，2019。 doi:10.4230/LIPIcs.ICALP.2019.42。

[17] Vincent Cohen-Addad 和 Karthik C. S.。lp度量下聚类的不可近似性。载于 David Zuckerman（编），第*60*届 *IEEE*计算机科学基础年会，*FOCS 2019*，美国马里兰州巴尔的摩，*2019*年*11*月*9*日至*12*日，第519–539页。IEEE计算机学会，2019。 doi:10.1109/FOCS.2019.00040。

[18] Vincent Cohen-Addad、Karthik C. S. 和 Euiwoong Lee。关于不含候选中心的聚类问题的可近似性。载于 D´aniel Marx 编，*2021*年 *ACM-SIAM*离散算法研讨会论文集，*SODA 2021*，虚拟会议，*2021*年*1*月*10*日至*13*日，第2635–2648页。SIAM，2021年。 doi:10.1137/1.9781611976465.156。

[19] Vincent Cohen-Addad、Karthik C. S. 和 Euiwoong Lee。Johnson覆盖假设：ℓ_p-度量空间中 k-均值和 k-中位数的不可近似性。载于 Joseph (Seffi) Naor 和 Niv Buchbinder 编，*2022*年 *ACM-SIAM*离散算法研讨会论文集，*SODA 2022*，虚拟会议 / 亚历山大，弗吉尼亚州，美国，*2022*年*1*月*9*日至*12*日，第1493–1530页。SIAM，2022年。 doi:10.1137/1.9781611977073.63。

[20] Vincent Cohen-Addad、Philip N. Klein 和 Claire Mathieu。局部搜索为欧几里得度量和无小子图度量下的 k-均值与 k-中位数提供近似方案。*SIAM*计算杂志，48(2):644– 667，2019。 doi:10.1137/17M112717X。

[21] Vincent Cohen-Addad、Silvio Lattanzi、Ashkan Norouzi-Fard、Christian Sohler 和 Ola Svensson。通过拒绝采样实现快速且准确的 k-均值++。载于 Hugo Larochelle、Marc'Aurelio Ranzato、Raia Hadsell、Maria-Florina Balcan 和 Hsuan-Tien Lin 编，《神经信息处理系统进展 *33*：神经信息处理系统年会 *2020*，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上举行》，2020。网址：https://proceedings.neurips.cc/paper/2020/hash/babcff88f8be8c4795bd6f0f8cccca61-Abstract.html。

[22] Vincent Cohen-Addad、David Saulpic 和 Chris Schwiegelshohn。《用于聚类的新核心集框架》。载于 Samir Khuller 和 Virginia Vassilevska Williams 编，*STOC '21*：第*53*届*ACM SIGACT*年度计算理论研讨会，线上活动，意大利，*2021*年*6*月*21*日至*25*日，第169–182页。ACM，2021年。 doi:10.1145/3406325.3451022。

[23] Sanjoy Dasgupta 和 Yoav Freund。《用于向量量化的随机投影树》。*IEEE* 信息论汇刊，55(7):3229–3242，2009年。

[24] Mark de Berg、Hans L. Bodlaender、S´andor Kisfaludi-Bak、D´aniel Marx 和 Tom C. van der Zan- den。几何交图中指数时间假设紧算法与下界的一个框架。*SIAM*计算杂志，49(6):1291–1331，2020。 doi:10.1137/20M1320870。

[25] Mark de Berg、Hans L Bodlaender、S´andor Kisfaludi-Bak、D´aniel Marx 和 Tom C van der Zanden。几何交图中 ETH 紧算法与下界的一个框架。载于第50届年度 ACM SIGACT计算理论研讨会论文集，第574–586页，2018年。[26] Irit Dinur。从间隙3sat到多项式间隙标签覆盖的轻度指数归约。计算复杂性电子讨论会，TR16-128，2016年。网址：https://eccc.weizmann.ac.il/report/2016/128，arXiv:TR16-128。[27] Fedor V. Fomin、Petr A. Golovach 和 Kirill Simonov。参数化k聚类：易处理性孤岛。《计算机与系统科学杂志》，117:50–74，2021年。网址：https://doi.org/10.1016/j.jcss.2020.10.005，doi:10.1016/J.JCSS.2020.10.005。[28] Zachary Friggstad、Mohsen Rezapour 和 Mohammad R. Salavatipour。局部搜索为倍增度量中的k-均值提供多项式时间近似方案。SIAM计算杂志，48(2):452–480，2019年。doi:10.1137/17M1127181。[29] Sudipto Guha 和 Samir Khuller。贪心算法卷土重来：改进的设施选址算法。《算法杂志》，31(1):228–248，1999年。网址：http://dx.doi.org/10.1006/jagm.1998.0993，doi:10.1006/jagm.1998.0993。[30] S Louis Hakimi。通信网络中交换中心的最优布局及一些相关图论问题。《运筹学》，13(3):462–475，1965年。[31] Sariel Har-Peled。几何近似算法。第173卷。美国数学学会，2011年。[32] Mary Inaba、Naoki Katoh 和 Hiroshi Imai。加权Voronoi图与随机化在基于方差的k聚类中的应用（扩展摘要）。载于第十届年度计算几何研讨会论文集，石溪，纽约州，美国，1994年6月6日至8日，第332–339页，1994年。doi:10.1145/177424.178042。[33] Kamal Jain、Mohammad Mahdian 和 Amin Saberi。设施选址问题的一种新贪心方法。载于第34届年度 ACM计算理论研讨会论文集，2002年5月19日至21日，蒙特利尔，魁北克省，加拿大，第731–740页，2002年。网址：http://doi.acm.org/10.1145/509907.510012，doi:10.1145/509907.510012。[34] Kamal Jain 和 Vijay V. Vazirani。使用原始-对偶框架和拉格朗日松弛的度量设施选址与k-中位数问题近似算法。《ACM杂志》，48(2):274–296，2001年。doi:10.1145/375827.375845。[35] S´andor Kisfaludi-Bak、Jesper Nederlof 和 Karol Wegrzycki。欧几里得旅行商问题的 Gap-ETH 紧近似方案。《ACM期刊》，72(6):1–48，2025年。[36] Stavros G. Kolliopoulos 和 Satish Rao。欧氏k中位数问题的近线性时间近似方案。SIAM计算杂志，37(3):757–782，2007年。doi:10.1137/S0097539702404055。[37] Alfred A Kuehn 和 Michael J Hamburger。用于仓库选址的启发式程序。《管理科学》，9(4):643–666，1963年。[38] Euiwoong Lee、Melanie Schmidt 和 John Wright。改进并简化k-均值的不可近似性结果。《信息处理快报》，120:40–43，2017年。doi:10.1016/j.ipl.2016.11.009。

[39] Shi Li 和 Ola Svensson。通过伪近似逼近 k-中位数。SIAM计算杂志，45(2):530–547，2016。 doi:10.1137/130938645.[40] Stuart Lloyd。脉冲编码调制中的最小二乘量化。IEEE信息论汇刊，28(2):129–137，1982。[41] Meena Mahajan、Prajakta Nimbhorkar 和 Kasturi R. Varadarajan。平面k-均值问题是 NP难。Theor. Comput. Sci.，442:13–21，2012。[42] Konstantin Makarychev、Yury Makarychev 和 Ilya P. Razenshteyn。Johnson-Lindenstrauss 变换在 k-均值和 k-中位数聚类中的性能。载于第三十届ACM年会论文集，即第51届ACM SIGACT计算理论研讨会（STOC 2019），美国亚利桑那州凤凰城，2019年6月23日至26日，第1027–1038页，2019。 doi:10.1145/3313276.3316350.[43] Pasin Manurangsi 和 Prasad Raghavendra。生日重复定理与稠密 CSP 近似的复杂性。载于 Ioannis Chatzigiannakis、Piotr Indyk、Fabian Kuhn 和 Anca Muscholl（编），第44届国际自动机、语言与编程研讨会（ICALP 2017），2017年7月10日至14日，波兰华沙，LIPIcs 第80卷，第78:1–78:15页。Schloss Dagstuhl - Leibniz-Zentrum für Informatik，2017。 doi:10.4230/LIPIcs.ICALP.2017.78.[44] Jiŕı Matousek。关于近似几何k-聚类。Discret. Comput. Geom.，24(1):61–84，2000。doi:10.1007/s004540010019.[45] Nimrod Megiddo 和 Kenneth J Supowit。若干常见几何选址问题的复杂性。SIAM 计算杂志，13(1):182–196，1984。[46] Satish B Rao 和 Warren D Smith。通过“伸展器”和“榕树”逼近几何图。载于第三十届ACM计算理论研讨会年会论文集，第540–550页，1998。[47] John Fred Stollsteimer。技术变革和产量扩张对加利福尼亚梨产区梨营销设施的最佳数量、规模及位置的影响。加利福尼亚大学伯克利分校，1961。

低维情形下 k-中位数的算法

对于 k-中位数，由于距离没有平方，我们需要略微调整预算和论证。给定这一修改后的预算，算法的结构几乎相同；因此，我们在此仅给出 k-中位数的结构定理的证明。

定义 A.1. 设 $P \subset \mathbb{R}^d$、 $\mathcal{D}$ 是 P、 $\varepsilon > 0$、 $\mathcal{A}$ 的层次分解，且 $\mathcal{S}$ 是 P上k-中位数的两个解。

对于球 $B(x,r)$，其相对于 B和 $\mathcal{D}$ 的绕行 为 $\det_{\mathcal{D}}(x,r) = \varepsilon 2^{\mathrm{lvl}(x,r)}$。

点 $p \in P$相对于 $\mathcal{A}$ 和 $\mathcal{S}$具有 k-中位数预算 $b(p,\varepsilon) = b_1(p,\varepsilon) + b_2(p,\varepsilon) +$ $b_3(p,\varepsilon)$,

with

$$b_1(p,\varepsilon) = \begin{cases} \det_{\mathcal{D}}(p, 3\mathcal{A}_p) \text{ if } \mathrm{lvl}(p, 3\mathcal{A}_p) \le \log(3\mathcal{A}_p) + \tau(\varepsilon, d) \\ 0 \text{ otherwise} \end{cases}$$

$$b_2(p,\varepsilon) = \begin{cases} \det_{\mathcal{D}}\left(p, 3(\mathcal{A}_p + \mathcal{S}_p)\right) \text{ if } \mathrm{lvl}(p, 3(\mathcal{A}_p + \mathcal{S}_p)) \le \log(3(\mathcal{A}_p + \mathcal{S}_p)) + \tau(\varepsilon, d) \\ 3\mathcal{A}_p + 2\mathcal{S}_p \text{ otherwise} \end{cases}$$

$$b_3(p,\varepsilon) = \begin{cases} \det_{\mathcal{D}}(\mathcal{A}(p), 3\mathcal{S}_{\mathcal{A}(p)}) \text{ if } \mathrm{lvl}(\mathcal{A}(p), 3\mathcal{S}_{\mathcal{A}(p)}) \le \log(3\mathcal{S}_{\mathcal{A}(p)}) + \tau(\varepsilon, d) \\ 0 \text{ otherwise} \end{cases}$$

我们的目标是证明，与定理3.3类似， $\mathcal{D}$ 对于 k-中位数问题具有较小的失真，其定义如下：

定义 A.2. 对于 $\varepsilon > 0$，我们称 $\mathcal{D}$具有 ε-小失真的 k-中位数，当且仅当

1. 关于 $\mathcal{A}$ 和最优解′的总 k-中位数预算有界： $\sum_{p\in P} b(p,\varepsilon) \le \varepsilon(\mathrm{cost}(P, \text{最优解}) + \mathrm{cost}(P, \mathcal{A}))$。
2. 存在一个解 S^*，其中包含 $B_{\mathcal{D}}$，且满足 $\sum_{p\in P} \mathrm{dist}(p, \widetilde{p}) + \mathrm{dist}(\widetilde{p}, S^*) \le (1+\varepsilon)\mathrm{cost}(P, \text{最优解}) + \varepsilon\mathrm{cost}(P, \mathcal{A})$；*3.* 对于每个点 $p \in P$，其到 S^* 的绕行至多为 $b(p,\varepsilon)$，也就是说，存在一个中心 $s \in S^*$，它在层级 i 处从 $\widetilde{p}$切割出来，满足
$\mathrm{dist}(p, \widetilde{p}) + \mathrm{dist}(\widetilde{p}, s) + \varepsilon 2^i \le \mathrm{dist}(p, \widetilde{p}) + \mathrm{dist}(\widetilde{p}, S^*) + b(p,\varepsilon)$。

该解 S^* 的构造与第 3.3 节完全相同，唯一的区别是， $\mathcal{A}$和 OPT′ 现在是关于 k-中位数而非 k-均值的解。

预算界定。 对于 k-均值，我们首先证明预算是有界的：

引理 A.3。 固定两个解 $\mathcal{A}$ 和 S，令 $b(p,\varepsilon)$ 为相对于解 $\mathcal{A}$ 和 S的 p 的预算。以 $2/3$ 的概率（关于 $\mathcal{D}$的随机性）， $\sum_p b(p,\varepsilon) = O(d\log(1/\varepsilon)\cdot\varepsilon)(\mathrm{cost}(P,\mathcal{A}) + \mathrm{cost}(P,S))$。

证明。 我们先界定预算的期望，然后使用马尔可夫不等式。回顾可知，任意球$B(x,r)$ 在层级 i 被切割的概率为 $\sqrt{d}r/2^i$，被严重切割的概率为 ε。 此外，当$B(x,r)$ 在层级 i被切割时，则有 $\det_{\mathcal{D}}(x,r) = \varepsilon 2^i$。

为了稍微简化分析，我们先分析该变量的期望

$$\beta(x,r) := \begin{cases} \det_{\mathcal{D}}(x,r) \text{ when } \mathrm{lvl}(x,r) \le \log(r) + \tau(\varepsilon, d) \\ 0 \text{ otherwise.} \end{cases}$$

为此，我们区分 区分 $\mathrm{lvl}(x,r)$ 至多为 $\log(r)$、介于 $\log(r)$ 和 $\log(r)+$之间的情况

$\tau(\varepsilon, d)$，以及至少为 $\log(r) + \tau(\varepsilon, d)$ 的情况。这给出了以下上界：

$$\begin{aligned}
\mathbb{E}[\beta(x, r))] &\le \sum_{i \le \log(r)} \det_D(x, r) + \sum_{i=\log(r)}^{\log(r)+\tau(\varepsilon,d)} \det_D(x, r) \cdot \Pr[\boldsymbol{D} \text{ cuts } B(p, r) \text{ at a level } i] \\
&\le \sum_{i \le \log(r)} \varepsilon 2^i + \sum_{i=\log(r)}^{\log(r)+\tau(\varepsilon,d)} \varepsilon 2^i \cdot \frac{\sqrt{d} r}{2^i} \\
&\le 2\varepsilon r + \sum_{i=\log(r)}^{\log(r)+\tau(\varepsilon,d)} \varepsilon \sqrt{d} r \\
&\le 2\varepsilon r + \tau(\varepsilon, d) \cdot \varepsilon \sqrt{d} r \\
&= O(\sqrt{d} \log(d/\varepsilon))\varepsilon r,
\end{aligned}$$

其中最后一步使用了 $\tau(\varepsilon, d) = \log(d)/2 + \log(1/\varepsilon)$。

由于 $b_1(p, \varepsilon) = \beta(p, 3\mathcal{A}_p)$ 和 $b_3(p, \varepsilon) = \beta(\mathcal{A}(p), 3\mathcal{S}_{\mathcal{A}(p)})$，我们直接得到 $\mathbb{E}[b_1(p, \varepsilon) + b_3(p, \varepsilon)] \le O(\sqrt{d} \log(d/\varepsilon))\varepsilon(\mathcal{A}_p + \mathcal{S}_p)$。

对 $b_2(p, \varepsilon)$ 进行界定只稍微复杂一些：$b_2(p, \varepsilon) = \beta(p, \mathcal{A}_p + \mathcal{S}_p) + X(p, \varepsilon)$，其中 $X(p, \varepsilon) := 3\mathcal{A}_p + 2\mathcal{S}_p$ 当 $\mathrm{lvl}(p, 3(\mathcal{A}_p + \mathcal{S}_p)) > \log(3(\mathcal{A}_p + \mathcal{S}_p)) + \tau(\varepsilon, d)$ 时。因此，只需计算 $\mathbb{E}[X(p, \varepsilon)]$。引理 2.2 中分解的性质保证了 $\mathrm{lvl}(p, 3(\mathcal{A}_p + \mathcal{S}_p)) > \log(3(\mathcal{A}_p + \mathcal{S}_p)) + \tau(\varepsilon, d)$ 的概率至多为 ε，因此 $\mathbb{E}[X(p, \varepsilon)] \le \varepsilon \cdot (3\mathcal{A}_p + 2\mathcal{S}_p) = O(\varepsilon)(\mathcal{A}_p + \mathcal{S}_p)$。

Therefore, $\mathbb{E}[b_2(p, \varepsilon)] = \mathbb{E}[\beta(p, 3(\mathcal{A}_p + \mathcal{S}_p)] + \mathbb{E}[X(p, \varepsilon)] \le O(\varepsilon)(\mathcal{A}_p + \mathcal{S}_p)$.

将这些不等式对 b_1、b_2、b_3 以及所有 $p \in P$ 求和，我们得到 $\mathbb{E}[\sum_{p \in P} b(p, \varepsilon)] \le O(\sqrt{d} \log(d/\varepsilon)\varepsilon)(\mathrm{cost}(P, \mathcal{A}) + \mathrm{cost}(P, \mathcal{S}))$。由马尔可夫不等式即得该引理。 □

代价界定 $\mathcal{S}^*$。现在，与 k-均值完全一样，解 $\mathcal{S}^*$ 是近最优的：我们可以用它证明如下引理，类似于引理3.7。为完备性起见，我们重复该证明。

引理 A.4。 $\mathcal{S}^*$ 包含 B_D，并且以至少 4/5 的概率成立

$$\sum_{p \in P} \mathrm{dist}(p, \widetilde{p}) + \mathrm{dist}(\widetilde{p}, \mathcal{S}^*) \le (1 + \varepsilon)\mathrm{cost}(P, OPT) + \varepsilon \mathrm{cost}(P, \mathcal{A})$$

证明。根据构造，$\mathcal{S}^*$ 确实包含 B_D。为简化方程，当 p 相对于 $\mathcal{A}$ 被严重切割时（此时 $\widetilde{p} = \mathcal{A}(p)$），我们将 p 记作b.c.。于是有

$$\sum_{p \in P} \mathrm{dist}(p, \widetilde{p}) + \mathrm{dist}(\widetilde{p}, \mathcal{S}^*) \le \sum_{p \in P} \mathrm{dist}(p, \mathcal{S}^*) + \sum_{p \text{ b.c.}} \mathrm{dist}(p, \widetilde{p}) + \mathrm{dist}(\widetilde{p}, \mathcal{S}^*). \tag{2}$$

根据 $\mathcal{S}^*$ 的性质，第一项至多为 $(1 + \varepsilon)\mathrm{cost}(P, \text{最优解}) + \varepsilon \mathrm{cost}(P, \mathcal{A})$：因此，我们只关注第二项。为此，我们使用事实 3.6：

$$\sum_{p \text{ b.c.}} \mathrm{dist}(p, \widetilde{p}) + \mathrm{dist}(\widetilde{p}, \mathcal{S}^*) \le \sum_{p \text{ b.c.}} 3\mathcal{A}_p + 2\mathrm{OPT}'_p$$

每个点 p 被严重切割的概率为 ε：因此，其期望满足

$$\mathbb{E}[3\sum_{p\text{ b.c.}}\mathcal{A}_p+\mathrm{OPT}'_p]\le 3\varepsilon\sum_{p\in P}\mathcal{A}_p+\mathrm{OPT}'_p\le 3\varepsilon(\mathrm{cost}(P,\mathcal{A})+\mathrm{cost}(P,\mathrm{OPT}'))$$

根据马尔可夫不等式，以概率 $4/5$ 有：

$$\sum_{p\text{ b.c.}}\mathrm{dist}(p,\widetilde{p})+\mathrm{dist}(\widetilde{p},\mathrm{OPT}')\le 3\cdot 5\cdot\varepsilon(\mathrm{cost}(P,\mathcal{A})+\mathrm{cost}(P,\mathrm{OPT}')).$$

将该界代入式（2），并利用引理3.1 来界定最优解的代价′，可得：

$$\begin{aligned}\sum_{p\in P}\mathrm{dist}(p,\widetilde{p})+\mathrm{dist}(\widetilde{p},S^*)&\le(1+\varepsilon)\mathrm{cost}(P,\mathrm{OPT})+\varepsilon\mathrm{cost}(P,\mathcal{A})+3\cdot 5\cdot\varepsilon(\mathrm{cost}(P,\mathcal{A})+\mathrm{cost}(P,\mathrm{OPT}'))\\&=(1+O(\varepsilon))\mathrm{cost}(P,\mathrm{OPT})+O(\varepsilon)\mathrm{cost}(P,\mathcal{A}).\end{aligned}$$

□

界定绕行。 对于 k-均值，我们现在得出结构定理，证明通过分解~的门户连接 P 与 S^* 所产生的绕行 $\mathcal{D}$ 不超过定义 A.1中规定的预算。为保证完备性，我们再次给出证明：分类讨论略为简单，因为其中不涉及正方形。

引理 A.5。 设 $P\subset\mathbb{R}^d$，$\mathcal{A}$ 是 *k-median* 在 P 上的一个解，且 $\mathcal{D}$ 是一个随机分解。设 S^*是引理3.7 给出的解。

那么，对于任意 $p\in P$，存在一个在层级 i从 $\widetilde{p}$切出的中心 $s\in S^*$，使得 $\mathrm{dist}(p,\widetilde{p})+\mathrm{dist}(\widetilde{p},s)+\varepsilon 2^i\le\mathrm{dist}(p,\widetilde{p})+\mathrm{dist}(\widetilde{p},S^*)+b(p,\varepsilon)$。

证明。回顾一下，当球 $B(x,3\,\mathrm{dist}(x,S))$ 在高于 $\log(9\,\mathrm{dist}(x,S))+\tau(\varepsilon,d)$ 的层级被切割时，点 $x\in\mathbb{R}^d$ 相对于 $\mathcal{D}$ 和集合 S 会被严重切割。回顾一下 P 和 S^* 的构造：当点 p 相对于 $\mathcal{D}$ 和 $\mathcal{A}$ 被严重切割时，$\widetilde{p}=\mathcal{A}(p)$；当中心 $\ell\in\mathcal{A}$ 相对于 $\mathcal{D}$ 和最优解 OPT' 被严重切割时，$\ell\in S^*$。

为了证明该引理，我们先根据 p 是否相对于 $\mathcal{D}$ 和 $\mathcal{A}$ 被严重切割进行分类讨论。在被严重切割的情况下，有 $\widetilde{p}=\mathcal{A}(p)$。

1. 如果 $\mathcal{A}(p)$ 相对于 $\mathcal{D}$ 和最优解 OPT' 被严重切割，那么 $\mathcal{A}(p)\in S^*$：因此，$\widetilde{p}=\mathcal{A}(p)=S^*(p)$，并且 $\widetilde{p}$ 和 $S^*(p)$ 不会被切割，例如 $i=-\infty$。因此，$s=\mathcal{A}(p)$ 满足该引理，其中 $\mathrm{dist}(p,\widetilde{p})+\mathrm{dist}(\widetilde{p},s)+\varepsilon 2^i=\mathrm{dist}(p,\widetilde{p})$。

2. $\mathcal{A}(p)$ 相对于 $\mathcal{D}$和最优解′未被严重切割。令 $s=S^*(\mathcal{A}(p))$，并令 i 为 $\mathcal{A}(p)$和 s被切割时的层级：记作 $\mathrm{dist}(p,\widetilde{p})+\mathrm{dist}(\widetilde{p},s)+\varepsilon 2^i=\mathcal{A}_p+\mathrm{dist}(\widetilde{p},S^*)+\varepsilon 2^i$。

   根据事实 3.6，$s\in B(\mathcal{A}(p),2\text{最优解}'_{\mathcal{A}(p)})$，因此 $i\le i':=\mathrm{lvl}\left(\mathcal{A}(p),3\text{最优解}'_{\mathcal{A}(p)}\right)$。结合 $\mathcal{A}(p)$ 未被严重切割这一事实，可得 $\varepsilon 2^i\le\varepsilon 2^{i'}\le b_3(p)$。

现在转向 p 未被严重切割的情况。在这种情况下，$\widetilde{p}=p$：因此我们只需对 $\mathrm{dist}(p,s)+\varepsilon 2^i$进行界定。此外，球 $B(p,3\mathcal{A}_p)$ 在层级 $\mathrm{lvl}(p,3\mathcal{A}_p)\le\log(3\mathcal{A}_p)+\tau(\varepsilon,d)$处被切割。我们根据球 $B(p,3(\mathcal{A}_p+\text{最优解}'_p))$ 是否被严重切割进行分类讨论。注意，根据事实3.5，$S^*(p)\in B(p,3(\mathcal{A}_p+\text{最优解}'_p))$。

1. 如果 $B(p, 3(\mathcal{A}_p + \text{最优解}'_p))$ 未被严重切割，那么 p 和 $s = S^*(p)$ 在不高于 $i = \text{lvl}(p, 3(\mathcal{A}_p + \text{最优解}'_p))$的层级处被切割，并且 s 满足引理： $\text{dist}(p, s) + \varepsilon 2^i \le \text{dist}(p, S^*) + b_2(p)$。

2. 如果 $B(p, 3(\mathcal{A}_p + \text{OPT}'_p))$ 被严重切割，我们证明 $s = S^*(\mathcal{A}(p))$ 满足该引理。特别地，

(a) 如果 $\mathcal{A}(p)$ 被严重切割，则 $s = \mathcal{A}(p)$；在这种情况下，由于 p 未被严重切割， p 和 s 在不超过 $i = \text{lvl}(p, 3\mathcal{A}_p)$ 的层级被切割，而 $\text{dist}(p, s) + \varepsilon 2^i \le \text{dist}(p, S^*) + b_1(p)$也如此。(b) 如果 $\mathcal{A}(p)$ 未被严重切割，则可以将 p 分配给 $s = S^*(\mathcal{A}(p))$。这里的关键观察是：它们在 $i = \max\left(\text{lvl}(p, \mathcal{A}_p), \text{lvl}(\mathcal{A}(p), 3\text{OPT}'_{\mathcal{A}(p)})\right)$层级被切割：事实上，根据事实 3.6 s 位于球 $B(\mathcal{A}(p), 3\text{OPT}'_{\mathcal{A}(p)})$中，因此 s 和 $\mathcal{A}(p)$ 在层级 i处属于同一个单元；类似地，p 和 $\mathcal{A}(p)$在层级 i处属于同一个单元，因此， p 和 s 在不超过 i的层级被切割。此外，根据事实 3.6， p 和 s 满足 $\text{dist}(p, s) \le 3\mathcal{A}_p + 2\text{OPT}'_p$。因此，

$$\text{dist}(\widetilde{p}, s) + \varepsilon 2^i \le 3\mathcal{A}_p + 2\text{OPT}'_p + \varepsilon 2^i$$

现在，由于 $B(p, 3(\mathcal{A}_p + \text{OPT}'_p))$ 被严重切割，因此 $3\mathcal{A}_p + 2\text{OPT}'_p \le b_2(p)$。此外，根据 i 的上界，我们有 $\varepsilon 2^i \le b_1(p) + b_3(p)$，因为 p 和 $\mathcal{A}(p)$ 均未被严重切割。因此，$\text{dist}(\widetilde{p}, s) + \varepsilon 2^i \le b_1(p) + b_2(p) + b_3(p)$，由此完成该情形的证明，也证明了引理。

□

k 中位数算法的结论。 鉴于 k-中位数问题具有这一结构性结果，与求解 k-均值问题时相同的动态规划方法，可以让我们计算出一个 $(1 + \varepsilon)$-近似解。

B 低维算法的其他扩展 m

本节将介绍针对 [13]，中讨论的两种聚类变体——奖金收集聚类和带离群点的聚类——的改进近似方案。

B.1 奖品收集 k-均值和 k-中位数

在奖品收集 k-中位数和 k-均值问题中，每个客户点 $p \in P$ 都关联一个惩罚费用 $\pi(p)$，可以支付这笔费用来代替将 p 连接到某个中心。该问题的一个解是一组 k 个中心 S以及一组离群点 $O \subseteq P$：该解的代价为 $\sum_{p \in P \setminus O} \text{cost}(p, S) + \sum_{p \in O} \pi(p)$。

[13] 的方法如下：计算一个常数因子近似 $(\mathcal{A}_S, \mathcal{A}_O)$；如果点 $p \in \mathcal{A}_O$ 相对于 $\mathcal{D}$ 和最优解 OPT 被严重切割（即与 $\mathcal{A}_S$中心的定义相同），就称该点被严重切割。离群点被严重切割的概率为 ε。严重切割的中心集合为 $B_D \subseteq \mathcal{A}_S$，严重切割的离群点集合为 $BO_D \subseteq \mathcal{A}_0$。现在，当一个实例具有较小的失真时，

1. 关于 $\mathcal{A}$ 和最优解′ 的总预算有界：$\sum_{p \in P} b(p, \varepsilon) \le \varepsilon(\text{cost}(P, \text{最优解}) + \text{cost}(P, \mathcal{A}))$。

2. 存在一个解 (S^*, O^*)，使得 S^* 包含 B_D， O^* 包含 BO_D，且 $\sum_{p \in P \setminus O^*} \left(\text{dist}(p, \widetilde{p}) + \text{dist}(\widetilde{p}, S^*)\right)^2 \le (1 + \varepsilon)\text{cost}(P, \text{最优解}) + \varepsilon\text{cost}(P, \mathcal{A})$,

3. 对于每个点 $p \in P \setminus O^*$，其到 S^* 的绕行至多为 $b(p, \varepsilon)$，即存在一个在层级 i从 $\widetilde{p}$切割出的中心 $s \in S^*$， 使得 $(\text{dist}(p, \widetilde{p}) + \text{dist}(\widetilde{p}, s) + \varepsilon 2^i)^2 \le \left(\text{dist}(p, \widetilde{p}) + \text{dist}(\widetilde{p}, S^*)\right)^2 + b(p, \varepsilon)$。

[13] 展示了如何修改 S^* 的构造以考虑离群点：本质上，每当最优解′ 的一个中心在构造的步骤 1 中被关闭时，它所服务的、同时也在 $\mathcal{A}_0$中的所有客户点都会被加入 O^*。此外，所有严重切割的离群点都会被加入 O^*。由于这些情况各自发生的概率至多为 ε，以这种方式加入的离群点的代价至多为 $\varepsilon \sum_{p\in\mathcal{A}_O} \pi(p) = O(\varepsilon)\text{cost}(P, \text{最优解})$。

对于其他性质，结构定理定理 3.3 的证明完全采用相同的方式，表明实例以至少 $2/3$的概率具有较小的失真。第 3.4 节的动态规划可以轻松调整，用于求解这些问题的奖品收集版本。

B.2 k-均值和带离群点的 k-中位数

在这种聚类版本中，给定一个整数 z，代价函数忽略代价最高的 z 个点。更正式地说，该问题的解是一个包含 k 个中心 S的集合，以及一个包含 z 个离群点 $O \subseteq P$的集合：该解的代价为 $\sum_{p\in P\setminus O} \text{cost}(p, S)$。

[13] 展示了如何利用奖品收集的思路获得一个双准则解，即一个包含 $(1+\varepsilon)z$ 个离群点且代价为 $(1+\varepsilon)\text{cost}(P, \text{OPT})$ 的解。其思路是，先前针对奖品收集的论证表明，新加入的离群点出现的概率仅为 ε：从一个包含 z 个离群点的集合开始，这样新增的离群点数量仅为 $O(\varepsilon z)$。分析的其余部分——对非离群点的绕行进行界定——与标准的 k-中位数和 k-均值问题相比保持不变；因此，我们改进后的分析可直接推广到带离群点的情形。