机器学习中若干几何优化（含离群点）问题的次线性时间算法*

Hu Ding

中国科学技术大学计算机科学与工程学院，中国合肥 huding@ustc.edu.cn

arXiv:2301.02870v1 [cs.DS] 7 Jan 2023

摘要。 本文研究机器学习中出现的若干重要几何优化问题。首先，我们重新考察欧几里得空间中的最小包围球（MEB）问题 $\mathbb{R}^d$。尽管该问题此前已得到广泛研究，但现实中的机器学习任务往往需要处理大规模数据集，因此我们甚至无法承担线性时间算法的开销。受近期关于 *beyond worst-case analysis* 的研究启发，我们为最小包围球引入稳定性的概念，该概念直观且易于理解。粗略地说，如果移除少量输入点无法显著减小所得球的半径，那么该最小包围球实例就是稳定的。在稳定性假设下，我们提出两种采样算法，用于计算半径近似最小包围球，其样本复杂度与输入点的数量无关 n。特别是，第二种算法的样本复杂度甚至与维数无关 d。我们还考虑不作稳定性假设的一般情况，并提出一种混合算法，可输出半径近似最小包围球或覆盖近似最小包围球。与以往的次线性最小包围球算法相比，我们的算法降低了运行时间和遍数。我们的方法依赖于两种新颖技术：均匀自适应采样方法和夹逼引理。此外，我们发现，这两种技术还可推广用于设计高维场景下更广泛的含离群点几何优化问题的次线性时间算法，包括带离群点的最小包围球、含离群点的单类和双类线性支持向量机、 k-中心含离群点聚类，以及含离群点的平坦拟合。我们提出的算法也适用于核。

1 引言

许多现实世界中的机器学习任务都可以表述为欧几里得空间中的几何优化问题。我们从一个基础的几何优化问题——最小包围球（*MEB*）开始；近年来，这一问题受到了广泛关注。给定欧几里得空间 $\mathbb{R}^d$ 中由 n 个点组成的集合 P，其中 d 可能相当高，MEB 问题旨在找到一个半径最小、能够覆盖 P [20, 45, 69] 中所有点的球。MEB 在机器学习中有多项重要应用 [76]。例如，若使用核方法将所有映射后的点变换为具有相同范数，则流行的分类模型支持向量机（*SVM*）可以表述为高维空间中的 MEB 问题，例如，使用流行的径向基函数核；这种 SVM 称为“核心向量机（CVM）”，自 2005 [90] 提出以来，它一直是面向大规模数据集最重要的支持向量机训练方法之一。因此，快速的 MEB 算法可用于加速其训练过程 [29, 30, 90]。近年来，MEB 也被用于隐私保护 [44, 77] 和量子密码学 [53]。

通常，我们考虑最小包围球的近似解。如果一个球覆盖了所有 n 个点，但半径大于最优半径，我们称之为**“半径近似解”**；如果一个球的半径不大于最优半径，但覆盖的点少于 n 个，我们则称之为**“覆盖近似解”**（正式定义见第2节）。在大数据时代，数据集可能庞大到令我们甚至无法承担线性时间算法的开销。这促使我们提出以下问题：

* 本研究的部分内容已发表于 [37, 38]。

能否为最小包围球开发出运行时间相对于输入规模为次线性的近似算法？机器学习中出现的其他高维几何优化问题又该如何处理？

通常假设输入数据以 $n \times d$ 矩阵的形式表示，而任何复杂度为 $o(nd)$ 的算法都可视为次线性时间算法。在实践中，数据项通常表示为 $\mathbb{R}^d$ 中的稀疏向量；因此，即使维数 d 很高，执行距离计算等操作也可能很快（例如，如果每个向量有 $s \ll d$ 个非零项，计算距离所需的时间为 $O(s)$，而非$O(d)$；参见 [30] 的结尾注记）。此外，在许多现实场景中，输入点的数量 n 往往远大于维数 d。**因此，我们有兴趣设计复杂度关于 n 呈次线性（或关于 n 呈线性但系数较小）的算法**。

1.1 我们的主要思路和结果

我们设计次线性时间最小包围球算法的思路，受到近期关于稳定实例优化问题的研究启发，这类研究属于超越最坏情况分析 [82]的范畴。例如，近期有多项工作为聚类和最大割等问题引入了稳定性的概念 [8,13,18]。本文提出了最小包围球的**"稳定性"**概念。粗略地说，如果移除少量输入点无法显著减小所得球的半径，那么该最小包围球实例就是稳定的（例如，仅移除 1% 的点时，半径无法减小 10%）。这一概念背后的依据十分自然：如果给定实例不稳定，那么导致半径显著减小的少量点应被视为离群点（或者，我们可能需要多个球来覆盖输入点，就像 k-中心聚类问题 $[51, 61]$ 一样）。据我们所知，这是首次从稳定性的角度研究最小包围球问题。

我们证明了稳定性假设的一个重要推论：粗略地说，如果一个最小包围球实例是稳定的，那么其中心应当体现出空间中某种程度的鲁棒性（第 3 节）。利用这一推论，我们提出了两种采样算法，用次线性时间复杂度计算 $(1+\epsilon)$-半径近似最小包围球（第 4 节）；具体而言，我们的第二种算法的样本量（即 采样点的数量）与输入点的数量 n 和维数 d 无关（据我们所知，这是首个以这种次线性复杂度实现 $(1+\epsilon)$-半径近似的算法）。

此外，我们还有一个有趣的观察：在稳定性假设下提出的思路，即使在放宽要求的情况下，也能帮助我们解决不满足稳定性假设的一般实例。我们提出一**种混合方法**，根据输入实例是否足够稳定，输出半径近似最小包围球或覆盖近似最小包围球[1] （第 5 节）。此外，一个附带结果是，我们可以根据输出推断给定实例的稳定度。值得注意的是，基于 VC 维数 $[58, 92]$ 的简单均匀采样思路只能实现"双准则"近似，即半径和覆盖点数两方面都存在误差（参见第 4.1 节中对我们第一个采样算法的讨论）。与 Clarkson 等人提出的 [30], 次线性时间最小包围球算法相比，我们将总运行时间从 $\tilde{O}(\epsilon^{-2}n + \epsilon^{-1}d + M)$ 降至 $O(n + h(\epsilon,\ \delta) \cdot d + M)$，其中 M 是输入 $n \times d$ 矩阵中的非零项数，而 $h(\epsilon,\ \delta)$是一个取决于预先指定的半径误差界 ϵ 和覆盖误差界 δ的因子。因此，当 $n \gg d$时，我们的改进十分显著。唯一的权衡是，对于不稳定实例，我们允许采用覆盖近似（鉴于 [30], 证明的下界，如果仍将输出限制为 $(1+\epsilon)$-半径近似，就不太可能降低项 $\epsilon^{-2}n$）。此外，我们的算法只需**均匀采样和单遍**

[1] 我们不需要明确地 知道该实例是否稳定， 在运行我们的算法时 .

处理数据；另一方面， [30] 的算法需要 $\tilde{O}(\epsilon^{-1})$ 遍（详情见表 1）。除了稳定性思路外，我们的方法还依赖两项关键技术：新颖的**“均匀自适应采样”**方法和**“夹逼引理”**。粗略地说，均匀自适应采样方法有助于界定每一步“随机贪心选择”所引入的误差；夹逼引理则使我们能够估计每个候选方案的目标值，并以次线性时间选出最优方案。

结果		质量	Time	遍数 可扩展性，用于	带 离群点
Clarkson *et al.* [30]		$(1+\epsilon)$-近似比	$\tilde{O}(\epsilon^{-2}n+\epsilon^{-1}d+M)$	$\tilde{O}(\epsilon^{-1})$	N/A
核集方法 $[20,29,69,79]$		$(1+\epsilon)$-近似比	大约为 $O(\epsilon^{-1}nd)$ 或 $O(\epsilon^{-1}(n+d+M))$ 如果 $M=o(nd)$	$O(\epsilon^{-1})$	双准则 约 [22]
数值方法 [84]		$(1+\epsilon)$ -近似比	$\tilde{O}(\epsilon^{-1/2}nd)$ 或 $\tilde{O}(\epsilon^{-1/2}(n+d+M))$ 如果 $M=o(nd)$	$O(\epsilon^{-1/2})$	N/A
数值方法 [6]		$(1+\epsilon)$-近似比	$\tilde{O}(nd+n\sqrt{d}/\sqrt{\epsilon})$	$\tilde{O}(d+\sqrt{d/\epsilon})$	N/A
流式算法 [4, 25]		1.22-近似比	$O(nd/\epsilon^5)$	单遍	N/A
This 论文	稳定 实例	$(1+\epsilon)$-近似比	$O(C_1\cdot d)$（第 4.2 节）	均匀采样	N/A
	一般 实例	$(1+\epsilon)$-近似比 或 $(1-\delta)$-覆盖。	~~$O((n+C_2)d)$~~ 或 $O(n+C_2\cdot d+M)$ 如果 $M=o(nd)$（第 5.3 节）	均匀采样 再加上一遍	$(1+\epsilon)$ -近似比 或 $(1-\delta)$ -覆盖 （第 6 节）

表 1：高维情形下计算最小包围球的现有结果及我们的结果。表中，“rad.” 和“cov.”分别表示“半径近似”和“覆盖近似”。“M”是输入 $n\times d$ 矩阵中非零项的数量。因子 C_1 取决于 ϵ 和给定实例的稳定度；因子 C_2 取决于 ϵ 和 δ。

最后，我们介绍结果的若干扩展。在实践中，我们可以假设给定的数据集中存在离群点。具体而言，随着机器学习的快速发展，对抗机器学习领域吸引了大量关注 [17,52]。对抗攻击者可能会添加一小组离群点，使模型严重偏离并产生意外错误（关于针对支持向量机的投毒攻击的开创性论文 [16] 刚刚获得了*ICML 2022* 时间检验奖）。为了防御此类投毒攻击，我们通常会设计能够抵御离群点的鲁棒算法 [65]。然而，在高维情形下，离群点的存在不仅使问题变为非凸问题，还使其具有高度组合性；例如，如果输入数据项中有 m 个是离群点（$m<n$），那么在优化目标函数时，我们必须考虑数量呈指数级增长的 $\binom{n}{m}$ 种不同可能情形。因此，我们考虑为以下问题设计次线性时间算法。

带离群点的最小包围球。带离群点的最小包围球是最小包围球问题的一种自然推广，其目标是找到一个最小球，使其覆盖至少一定比例的输入点。我们可以将带离群点的最小包围球应用于解决数据挖掘和数据分析中的许多实际问题（例如，离群点识别） [89]。我们定义了带离群点的最小包围球的稳定性，并提出了次线性时间近似算法。据我们所知，我们的算法是首个用于带离群点的最小包围球问题的次线性时间算法（此前的算法，如 [22]，需要线性时间）。

其他带离群点的包围问题。除了带离群点的最小包围球，我们还发现，所提出的技术可用于解决范围更广的无离群点包围问题。我们定义了一种称为**带离群点的最小包围"x"（MEX）**的一般优化问题，其中"x"可以是指定类型的形状（例如，带离群点的最小包围球中的形状就是球）。我们证明，只要形状"x"满足若干性质，就可以推广均匀自适应采样方法和夹逼引理，使之适用于该形状。我们特别关注带离群点的 MEX 问题，包括平坦拟合、k-中心聚类以及含离群点的支持向量机；这些问题的共同特点是，它们各自的基础版本（不含离群点）都有一种基于贪心选择的迭代算法，与 [20] 的最小包围球算法类似。尽管这些问题此前已得到广泛研究，但关于其次线性时间算法的研究仍相当有限。

注记 1. 由于本文研究的几何优化问题源于机器学习应用，我们也考虑了**核** [85]。我们提出的算法只需对数据项执行基本运算，例如计算距离和内积。因此，这些算法也适用于核。

本文其余部分安排如下。第 1.2 节总结与本文相关的已有结果。第2节给出重要定义，并简要介绍 [20] 中针对最小包围球的核心集构造方法（该方法将在后续算法和分析中使用）。第 3 节证明最小包围球稳定性的推论。随后，在第 4 节中，我们针对稳定实例提出两种次线性时间最小包围球算法。在第 5 节中，我们提出两项关键技术——均匀自适应采样和夹逼引理，并据此给出一种不要求稳定性假设、适用于一般最小包围球问题的次线性时间算法。在第 6 节中，我们将第 5 节的思路扩展到带离群点的最小包围球问题。最后，在第7节中，我们介绍广义均匀自适应采样和夹逼引理，并将其应用于若干带离群点的包围问题（包括平坦拟合、k-中心聚类和含离群点的支持向量机）。

1.2 先前工作

与我们工作最相关的是 [7,30]。Clarkson 等人 [30] 提出了一种优雅的感知机框架，用于求解机器学习中出现的若干优化问题，例如最小包围球。给定由 $\mathbb{R}^d$ 中的 n 个点构成、表示为含有 M 个非零项的 $n \times d$ 矩阵的集合，该框架可在 $\tilde{O}\left(\frac{n}{\epsilon^2}+\frac{d}{\epsilon}\right)$ 时间内计算最小包围球 [2]。请注意，参数"ϵ"是加性误差（即若 r 为最优最小包围球的半径，则所得半径为 $r+\epsilon$），可通过 $O(M)$ 的预处理时间转换为相对误差（即 $(1+\epsilon)r$）。因此，如果 $M = o(nd)$，运行时间仍关于输入规模 nd 呈次线性（见表 1）。 [30] 的框架也启发了用于训练支持向量机 [60] 和近似求解半定规划 [47]的次线性时间算法。Hayashi 和 Yoshida [59] 提出了一种基于采样的方法，用于最小化二次函数；最小包围球的目标函数是其一个特例，但该方法会产生较大的加性误差 $O(\epsilon n^2)$。

Alon等人 [7] 研究了以下性质测试问题：给定某个度量空间中的一个包含 n 个点的集合，判断该实例是否可 (k,b) 聚类，其中，如果一个实例能被半径（或直径）为 $b>0$的 k 个球覆盖，则称其为 (k,b) 可聚类。他们提出了几种采样算法来"近似地"回答这个问题。具体而言，他们区分了实例 (k,b) 可聚类的情况，以及实例距离 (k,b') 可聚类还差ϵ的情况，其中 $\epsilon \in (0,1)$，且 $b' \geq b$。"距离……还差ϵ"是指必须移除超过 ϵn个点，实例才能变为 (k,b') 可聚类。请注意，由于放宽了"距离……还差"这一条件，他们的方法会在半径和覆盖点数上引入不可避免的误差，因此无法为最小包围球问题提供单一准则的半径近似算法或覆盖近似算法

[2] 渐近记号 $\tilde{O}(f) = O\left(f \cdot \text{多对数函数}\left(\frac{nd}{\epsilon}\right)\right)$。

问题，因为放宽“距离.....还差ϵ”会在半径和覆盖点数上引入不可避免的误差。不过，可以将其转化为“双准则”近似，即允许对半径和未覆盖离群点的数量进行近似（例如，丢弃的离群点超过预先指定的数量）。

最小包围球与核心集。核心集 是一个较小的点集，可用于近似表示大得多的点集 [1, 43, 80] 的结构/形状。核心集思想也被用于计算高维空间中的近似最小包围球 [22, 67, 69, 79]。B˘adoiu 和 Clarkson [20]证明，可以找到大小为 $\lceil 2/\epsilon \rceil$ 的核心集，从而得到一个 $(1+\epsilon)$-半径近似最小包围球。其他几种方法可以得到更小的核心集大小，例如 [21, 67]。事实上，计算最小包围球核心集的算法是一种 $Frank\text{-}Wolfe$ 算法 [46]，Clarkson [29] 对其进行了系统研究。其他不依赖核心集的最小包围球算法包括 [6, 45, 84]。Agarwal 和 Sharathkumar [4] 提出了一种用于计算最小包围球的流式处理 $(\frac{1+\sqrt{3}}{2}+\epsilon)$-半径近似算法；随后，Chan 和 Pathak [25] 证明，同一算法实际上可得到小于 1.22 的近似比。最近，Cohen-Addad等人 [31] 提出了一种利用核心集计算高维幂均值（例如，几何中位数和均值点）的次线性时间算法。

带离群点的最小包围球和 k带离群点的 -中心聚类。带离群点的最小包围球问题可视为带离群点的 k-中心聚类问题 [27]的 $k = 1$ 情形。B˘adoiu 等人 [22] 将其核心集思想扩展到最小包围球和带离群点的 k-中心聚类问题，并提出了线性时间双准则近似算法（假设 k 为常数）。Huang 等人 [62]和 Ding 等人 [41]分别证明，简单的均匀采样方法可以得到带离群点的 k-中心聚类的双准则近似。针对低维带离群点的最小包围球，也已提出了若干算法 [5, 42, 54, 71]。此外，还有不少关于流式处理最小包围球和带离群点的 k-中心聚类的研究 [24, 28, 72, 94]。其他相关主题包括鲁棒优化 [14], 、鲁棒拟合 [3, 57], 和不确定性优化 [23]。

含离群点的支持向量机。给定 $\mathbb{R}^d$中的两个点集 P_1 和 P_2 ，支持向量机（*SVM*）问题是寻找能够分隔 P_1 和 P_2 （若二者可分）的最大间隔 [26]。SVM可以表述为一个二次规划问题，过去已有许多高效技术被提出，例如软间隔支持向量机 [32, 81],ν-SVM [33, 86], 和 Core-SVM [91]。此外，也有不少研究致力于为含离群点的支持向量机设计鲁棒算法 [40, 88, 93]。

含离群点的平坦拟合。给定一个整数 $j \geq 0$ 和 $\mathbb{R}^d$中的一个点集，平坦拟合问题旨在找到一个 j维平坦，使其到输入点 [55]的最大距离最小；显然，最小包围球问题是其中的一个特例，此时 $j = 0$。在高维情形下，Har-Peled和Varadarajan [56] 在假设 j 固定的情况下给出了一种线性时间算法；随后，Panigrahy [79] 基于核心集方法进一步降低了其运行时间。也有若干方法研究含离群点的平坦拟合，但仅适用于低维情形 [3, 57]。

稳定性下的优化。 Bilu 和 Linial [18]指出，如果给定实例在边权重扰动下是稳定的，那么最大割问题就会变得更容易。Ostrovsky 等人 [78] 为 k-均值聚类提出了一种分离条件，即对于给定实例， k-均值的聚类代价显著低于 $(k-1)$-均值的情形；他们还证明了 Lloyd启发式算法 [70] 在该分离条件下的有效性。Balcan 等人 [13] 提出了近似稳定性的概念，用于寻找 k-中值和 k-均值聚类的真实聚类。Awasthi等人 [8] 提出了另一种聚类稳定性概念，并为 k-中值和 k-均值聚类给出了一个多项式时间近似方案。 [9–12, 68]研究了更多稳定性假设下的聚类算法。

次线性时间算法。除了前述的次线性最小包围球算法 [30],之外，人们还针对聚类 [35,63, 64,73,74] 和性质测试 [15,50]等问题研究了许多次线性时间算法。关于次线性时间算法的更详细讨论，可参见综述论文 [34,83]。

2 定义与预备知识

我们在单位成本 RAM 模型 [75]中描述并分析我们的算法。假设输入由一个 $n \times d$ 矩阵表示（即，n 个点位于 $\mathbb{R}^d$中）。如 [30], 所述，通常假设矩阵的每个条目都可以在常数时间内恢复。

我们令 $|A|$ 表示给定点集 A 中 $\mathbb{R}^d$的点数，并令 $||x-y||$ 表示 x 和 y 中两点之间的欧几里得距离。我们使用 $\mathbb{B}$（c，r）表示以点 c 为球心、半径为 $r > 0$的球。下面给出最小包围球和稳定性的定义。为使本文结构更加紧凑，我们分别将扩展到第5节、第6节和第7节所需的其他定义放在相应章节中。

定义 1（最小包围球（MEB））。给定 $\mathbb{R}^d$中由 P 个 n 组成的集合，最小包围球问题是寻找一个半径最小的球，以覆盖 P中的所有点。所得的球及其半径分别记为 **最小包围球**（P）和 **半径**（P）。

定义 *2*（半径近似与覆盖近似）。设 $0 < \epsilon, \delta < 1$。若球 $\mathbb{B}(c, r)$ 覆盖 P 中的所有点，且半径为 $r \leq (1+\epsilon)Rad(P)$，则称其为 *MEB(P)* 的（$1+\epsilon$）-半径近似。另一方面，若该球覆盖 P 中至少（$1-\delta$）n 个点，且半径为$r \leq Rad(P)$，则称其为 *MEB(P)* 的（$1-\delta$）-覆盖近似。

半径近似和覆盖近似都是单准则近似。当 ϵ（分别， δ）趋近于 0 时，（$1+\epsilon$）-半径近似（分别，（$1-\delta$）-覆盖近似）将趋近于 **最小包围球**(P)。“覆盖近似”似乎与“带离群点的最小包围球”相似，但实际上两者有很大区别（见第 5 节中的定义 4）。

定义 *3*（(α, β)-稳定）。给定一个由 n 个点组成的集合 P，其中有两个参数 α 和 β，且它们属于 *(0, 1)*，若满足以下条件，则 P 是一个 (α, β)-稳定实例：*(1)* 对任意满足 $|Q| < \beta n$ 的 $Q \subset P$，*Rad(*$P \setminus Q$*)* $>$ *(*$1-\alpha$*)Rad(P)*；*(2)* 存在一个满足 $|Q'| = \lceil \beta n \rceil$ 的 $Q' \subset P$，使得 *Rad(*$P \setminus Q'$*)* $\leq$*(*$1-\alpha$*) Rad(P)*。

定义 3 的直观解释。实际上， β 可以看作 α 的函数，反之亦然。例如，给定一个 $\alpha > 0$, 总是存在一个 $\beta \geq \frac{1}{n}$，使得 P 是一个 (α, β)-稳定实例（$\beta \geq \frac{1}{n}$ 因为我们必须至少移除一个点）。稳定性这一性质表明，除非从 P 中移除足够大比例的点，否则 **半径**(P) 不会显著减小。对于固定的 α， β 越大， P 就应当越稳定。同样，对于固定的 β， α 越小， P 就越稳定。

实际上，我们的稳定性假设在实践中相当合理。例如，如果移除极少一部分点（比如 $\beta = 1\%$）后，半径可以显著缩小（比如缩小 $\alpha = 10\%$），那么将这部分点视为离群点是很自然的。为了更好地理解高维中的稳定性概念，我们考虑以下两个例子。

例（i）。假设 P 在 MEB(P) 内均匀且稠密分布。令 $\alpha \in$ 为 (0, 1) 中的一个固定数值，并研究 P 相应的 β。如果我们希望剩余的 $(1-\beta)n$ 个点的半径尽可能小，直观来说，应该移除最外层的 βn 个点（因为 P 均匀且稠密）。令 Q' 表示最外层的 βn 个点的集合，其 Rad($P \setminus Q'$) 为 $\leq (1-\alpha)$Rad(P)。于是有 $\frac{|P \setminus Q'|}{|P|} \approx Vol\Big($MEB($P \setminus Q'$)$\Big)$ $=$ $Vol\Big($MEB(P)$\Big)$

$\frac{(\text{半径}(P\setminus Q'))^d}{(\text{半径}(P))_d} \le (1-\alpha)^d$，其中 $Vol(\cdot)$ 是体积函数。也就是说，$1-\beta \le (1-\alpha)^d$，这意味着当 α 固定时，$\lim_{d\to\infty}\beta = 1$；也就是说，随着 d 增大，P 趋于稳定。

例（ii）。考虑一个正则的 d 维单纯形 P，其中包含 $d+1$ 个点，且任意两点之间的距离均为 1。不难得到**半径**$(P) = \sqrt{\frac{d}{2(1+d)}}$，记为 r_d。如果从 P 中移除 $\beta(d+1)$ 个点，也就是说，它变成一个具有 $d' = (1-\beta)(d+1) - 1$ 的正则 d' 维单纯形，则新的半径为 $r_{d'} = \sqrt{\frac{d'}{2(1+d')}}$。要在固定 α 的情况下实现 $\frac{r_{d'}}{r_d} \le 1-\alpha$，显然 $1-\beta$ 不应大于 $1+\frac{1}{(2\alpha-\alpha^2)d}$；这意味着 $\lim_{d\to\infty}\beta = 1$。与例（i）类似，随着 d 增大，实例 P 趋于非常稳定。

注记 2. 在实践中，对于固定的 α，很难知道 β 的精确值。然而，β 的取值只会影响我们在第 4 节所提算法中的样本量，因此只假设一个合理的下界 $\beta_0 < \beta$ 就足够了。在第 5 节中，我们还考虑不作稳定性假设的一般情形，此时所提算法甚至不需要将 β_0 作为输入。

2.1 对 [20] 中核心集构造的更细致分析

我们首先简要介绍最小包围球的核心集构造，因为我们所提算法会用到它。设 $0 < \epsilon < 1$。[20] 中的算法可得到一个大小为 $2/\epsilon$ 的最小包围球核心集（为方便起见，我们始终假设 $2/\epsilon$ 是整数）。不过，他们的论文中有一个小问题。其分析假设每次迭代都计算核心集的精确最小包围球，但实际上，人们可能只计算近似最小包围球。因此，一个直接的问题是，作此改变后，质量是否仍有保证。Kumar 等人 [69]解决了这一问题，并证明每次迭代为核心集计算一个 $(1+O(\epsilon^2))$ 近似最小包围球，仍能保证得到大小为 $O(1/\epsilon)$ 的核心集，其中隐藏常数大于 80。Clarkson [29] 证明，最小包围球的贪心核心集构造算法作为 Frank-Wolfe 算法的一个特例，可得到大小略大于 $4/\epsilon$ 的核心集。注意，还有其他几种方法可以得到更小的核心集大小 [21, 67], ，但其构造算法更复杂，因此不适用于我们的问题。下面我们证明，即使每次迭代只计算近似最小包围球，也能保证得到大小可任意接近 $2/\epsilon$ 的 [20] 核心集。这改进了 [29, 69], 的核心集大小，而新的分析本身也很有意义。

为完整起见，我们先简要介绍 [20] 中核心集构造算法的思路。给定点集 $P \subset \mathbb{R}^d$，该算法是一个简单的迭代过程。首先，它从 P 中任意选取一个点，并将其放入初始为空的集合 T 中。在接下来的每次 $2/\epsilon$ 迭代中，算法都会更新 **最小包围球**(T) 的中心，并将距离当前 **最小包围球**(T) 中心最远的点加入 T。最后，**最小包围球**(T) 的中心可为 $MEB(P)$ 导出一个 $(1+\epsilon)$-近似。所选的 $2/\epsilon$ 个点组成的集合（即，T）称为最小包围球的核心集。为确保每次迭代都能取得预期的改进，他们 [20] 证明了：如果算法始终选取距离当前 **最小包围球**(T) 中心最远的点，则以下两个不等式成立：

$$r_{i+1} \ge (1+\epsilon)\mathbf{Rad}(P) - L_i; \quad r_{i+1} \ge \sqrt{r_i^2 + L_i^2}, \tag{1}$$

其中，r_i 和 r_{i+1} 分别是第 i 次和第 $(i+1)$ 次迭代中 **最小包围球**(T) 的半径，而 L_i 是 **最小包围球**(T) 的中心从第 i 次迭代到第 $(i+1)$ 次迭代的移动距离。

如前所述，我们通常在每次迭代中只计算一个近似的**最小包围球**(T)。在第 i次迭代中，我们分别用 c_i 和 o_i 表示精确**最小包围球**(T)和近似最小包围球的中心。假设 $||c_i - o_i|| \le \xi r_i$，其中 $\xi \in (0, \frac{\epsilon}{1+\epsilon})$（稍后我们会说明为何需要这个

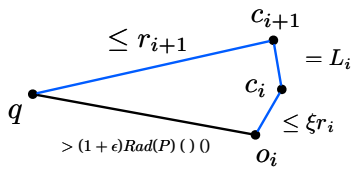

图1：式（2）的示意图。

界）。使用 [20], 中提出的另一种算法，可以在 $O(\frac{1}{\xi^2}|T|d)$时间内得到点 o_i。注意，我们每次迭代只计算 o_i，而非 c_i。因此，我们只能选择距离 o_i最远的点（记作 q）。如果 $||q-o_i|| \leq (1+\epsilon)$**半径**$(P)$，则算法结束，且已得到一个$(1+\epsilon)$-近似最小包围球。否则，我们有

$$(1+\epsilon)\mathbf{Rad}(P) < ||q-o_i|| \leq ||q-c_{i+1}|| + ||c_{i+1}-c_i|| + ||c_i-o_i|| \leq r_{i+1} + L_i + \xi r_i \quad (2)$$

根据三角不等式（见图1）。换言之，我们应将式（1）的第一个不等式替换为"$r_{i+1} > (1+\epsilon)$**半径**(P) $- L_i - \xi r_i$"。此外，式（1）的第二个不等式仍然成立，因为它仅取决于精确最小包围球的性质（见 [20, 引理 2。1]）。因此，我们有

$$r_{i+1} \geq \max\left\{\sqrt{r_i^2 + L_i^2}, (1+\epsilon)\mathbf{Rad}(P) - L_i - \xi r_i\right\}. \quad (3)$$

由此得到以下定理，其证明见附录 A。

定理1。 在 *[20],* 的核心集构造算法中，如果每次迭代都为 T 计算一个近似最小包围球，且所得中心 o_i 到 c_i 的距离小于$\xi r_i = s\frac{\epsilon}{1+\epsilon}r_i$，其中 $s \in (0,\ 1)$，则最终核心集大小有上界 $z = \frac{2}{(1-s)\epsilon}$。此外，当 s 足够小时，该上界可以任意接近 $2/\epsilon$ 。

我们只需将 s 设为 (0，1) 中的任意常数；例如，若 $s = 1/3$，核心集大小将不超过 $z = 3/\epsilon$。由于每次迭代中 $|T| \leq z$，总运行时间为 $O\Big(z(|P|d + \frac{1}{\xi^2}zd)\Big) = O\Big(\frac{1}{\epsilon}(|P| + \frac{1}{\epsilon^3})d\Big)$。

备注 3. 我们还想强调上述核心集构造过程中的一个简单观察，这将在后续的算法和分析中用到。该算法在每次迭代中总是选择距离 o_i 最远的点。然而，这实际上并非必要。只要所选点的距离至少为 $(1+\epsilon)$**半径**(P)，定理1中的结论仍然成立。如果不存在这样的点（即 $P \setminus \mathbb{B}\big(o_i,\ (1+\epsilon)\text{**半径**}(P)\big) = \emptyset$），那么一个 $(1+\epsilon)$-半径近似最小包围球（即 球 $\mathbb{B}\big(o_i,\ (1+\epsilon)$**半径**$(P)\big)$）已经得到。

注记 4（核）。如果每个点 $p \in P$ 通过某个核函数（例如，如 CVM [90]）映射到 $\mathbb{R}^D$ 中的 $\psi(p)$，其中 D 可以是 $+\infty$，我们仍然可以运行 [20],的核心集算法，因为该算法只需要计算距离，并且中心 o_i 在每次迭代中始终是 T 的凸组合；算法不会返回显式的中心，而是输出该中心的凸组合系数。类似地，我们在第 4.2 节中提出的算法 2 对核也同样适用。

3 稳定性性质的含义

在此 本节中，我们展示定义 3 中稳定性的一项重要推论 .

定理 2. 假设 $\epsilon, \epsilon', \beta_0 \in (0,1)$。设 P是一个满足 $\beta > \beta_0$ 的最小包围球问题(ϵ^2, β)-稳定实例，o 是其最小包围球的中心。设 $\tilde{o}$ 是 $\mathbb{R}^d$ 中给定的一个点。假设数量$r \leq (1+\epsilon'^2)$**半径**(P)。如果球 $\mathbb{B}(\tilde{o}, r)$ 覆盖了 P 中至少$(1-\beta_0)n$ 个点，则有以下结论

$$||\tilde{o} - o|| < (2\sqrt{2}\epsilon + \sqrt{3}\epsilon')\mathbf{Rad}(P). \tag{4}$$

定理 2 表明，如果一个球覆盖了 P 中足够大的子集，且其半径有界，那么它的中心应当接近 **最小包围球**(P) 的中心。令 $P' = \mathbb{B}(\tilde{o}, r) \cap P$，并假设 o' 是 **最小包围球**(P') 的中心。为了界定 $\tilde{o}$ 与 o 之间的距离，我们借助点 o' 将两者联系起来（因为 $||\tilde{o} - o|| \leq ||\tilde{o} - o'|| + ||o' - o||$）。以下两个关键引理用于证明定理 2。

引理 1。 距离 $||o' - o|| \leq \sqrt{2}\epsilon$**半径**（$P$）。

证明。 我们考虑两种情况：**最小包围球**(P') 完全被 **最小包围球**(P) 覆盖，以及不完全被覆盖。对于第一种情况（见图 2(a)），不难看出

$$||o' - o|| \leq \mathbf{Rad}(P) - (1-\epsilon^2)\mathbf{Rad}(P) = \epsilon^2\mathbf{Rad}(P) < \sqrt{2}\epsilon\mathbf{Rad}(P), \tag{5}$$

其中，第一个不等式是因为**最小包围球**(P')的半径至少为$(1-\epsilon^2)$**半径**(P)（定义 3）。因此，下面我们可以专注于第二种情况。

设 a 是位于**最小包围球**(P')与**最小包围球**(P)这两个球面交集上的任意一点。于是有以下命题。

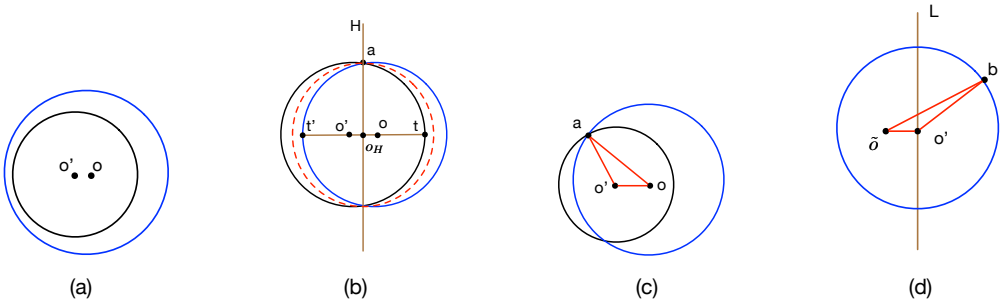

图 2：(a) 情况 $MEB(P') \subset MEB(P)$；(b) 证明命题 1 时在假设 $\angle ao'o < \pi/2$ 下的示意图；(c) 角 $\angle ao'o \geq \pi/2$；(d) 引理 2 的示意图。

命题 1. 角 $\angle ao'o \geq \pi/2$。

证明。 假设 $\angle ao'o < \pi/2$。注意到，由于 $||o - a||$ =**半径**(P) $\geq$**半径**(P') $= ||o' - a||$，$\angle aoo'$ 始终小于 $\pi/2$。因此，o和 o' 被超平面 H 分隔；该超平面垂直于线段 $o'o$，并经过点 a。见图 2(b)。现在我们证明，P' 可以被一个比**最小包围球**(P')更小的球覆盖。设 o_H 为点 $H \cap o'o$，而t（分别为 t'）为与 o 和 o' 共线、位于**最小包围球**(P')球面右侧的点（分别为 位于**最小包围球**(P)球面左侧的点；见图 2(b)）。那么，我们有

$$\begin{aligned} &||t - o_H|| + ||o_H - o'|| = ||t - o'|| = ||a - o'|| < ||o' - o_H|| + ||o_H - a|| \\ &\qquad \Longrightarrow ||t - o_H|| < ||o_H - a||. \end{aligned} \tag{6}$$

类似地，我们有 $||t' - o_H|| < ||o_H - a||$。因此，**最小包围球**$(P) \cap$**最小包围球**$(P')$ 被球 $\mathbb{B}(o_H, ||o_H - a||)$ 覆盖（即图2(b)中的“红色虚线”球）。此外，因为 P'被 **最小包围球**$(P) \cap$ **最小包围球**(P') 和 $||o_H - a|| < ||o' - a||$ = **半径**(P') 覆盖，所以 P' 被球$\mathbb{B}(o_H, ||o_H - a||)$ 覆盖，而该球比 **最小包围球**(P') 更小。这与 **最小包围球**(P') 是 P' 的最小包围球这一事实矛盾。因此，命题 $\angle ao'o \geq \pi/2$ 成立。□

根据命题 1，我们知道 $||o'-o|| \le \sqrt{(\textbf{半径}(P))^2 - (\textbf{半径}(P'))^2}$。见图 2(c)。此外，定义 3 意味着 $\textbf{半径}\,(P') \ge (1-\epsilon^2)\,\textbf{半径}\,(P)$。因此，我们有

$$||o'-o|| \le \sqrt{\big(\mathbf{Rad}(P)\big)^2 - \big((1-\epsilon^2)\mathbf{Rad}(P)\big)^2} \le \sqrt{2\epsilon}\mathbf{Rad}(P). \tag{7}$$

□

引理 2。 距离 $||\tilde{o}-o'|| < (\sqrt{2\epsilon}+\sqrt{3\epsilon'})\textbf{半径}(P)$。

证明。令 L为垂直于线段 $\tilde{o}o'$ 且经过中心o'的超平面。假设 $\tilde{o}$ 位于 L的左侧。那么，总存在一个点 $b \in P'$ 位于由 L 划分的 **最小包围球**(P') 的右侧闭半球上（该结论来自 [22, 引理 2.2]；为完整起见，我们在 B 节陈述该引理）。见图 2(d)。也就是说，角度$\angle bo'\tilde{o} \ge \pi/2$。因此，我们有

$$||\tilde{o}-o'|| \le \sqrt{||\tilde{o}-b||^2 - ||b-o'||^2}. \tag{8}$$

此外，由于 $||\tilde{o}-b|| \le r \le (1+\epsilon'^2)\textbf{半径}(P)$ 且 $||b-o'|| = \textbf{半径}(P') \ge (1-\epsilon^2)\textbf{半径}(P)$，由 (8) 可得 $||\tilde{o}-o'|| \le \sqrt{(1+\epsilon'^2)^2-(1-\epsilon^2)^2}\textbf{半径}(P)$，其中该上界等于

$$\sqrt{2\epsilon'^2+\epsilon'^4+2\epsilon^2-\epsilon^4}\mathbf{Rad}(P) < \sqrt{3\epsilon'^2+2\epsilon^2}\mathbf{Rad}(P) < (\sqrt{2}\epsilon+\sqrt{3}\epsilon')\mathbf{Rad}(P). \tag{9}$$

□

根据三角不等式和引理 1、2，我们立即得到

$$||\tilde{o}-o|| \le ||\tilde{o}-o'|| + ||o'-o|| < (2\sqrt{2}\epsilon+\sqrt{3}\epsilon')\mathbf{Rad}(P). \tag{10}$$

这就完成了定理 2 的证明。

4 稳定性假设下的最小包围球次线性时间算法

假设 $\epsilon \in (0,\ 1)$。我们假设给定实例 P是一个 $(\epsilon^2,\ \beta)$-稳定实例，其中 β大于给定的下界 β_0 (即，$\beta > \beta_0$)。利用定理 2，我们提出两种不同的次线性时间采样算法来计算最小包围球。与大多数关于次线性时间算法的文章（例如，[35, 73, 74]）一样，在我们的算法的每个采样步骤中，我们总是**独立且均匀随机地**抽取样本。

4.1 第一个算法

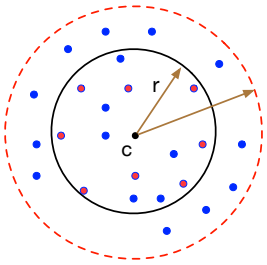

图 3：第一个采样算法的示意图。红色点是样本；我们将 $\mathbb{B}(c,\ r)$ 稍微扩大，较大的球是整个输入点集的半径近似最小包围球。

第一种算法基于 VC 维数和 ϵ-网理论 [58, 92]。粗略地说，我们计算一个小型随机样本（比如 $\mathbb{B}(c,\ r)$）的近似最小包围球，并将球稍微扩大；然后证明这个扩大的球是整个数据集的近似最小包围球（见图 3）。我们的关键思路是证明 $\mathbb{B}(c,\ r)$ 至少覆盖 $(1-\beta_0)n$ 个点，因此根据定理 2，c 接近最优中心。正如第 1.1 节所强调的，我们的结果是单准则近似。如果不采用稳定性假设而直接应用均匀样本的思路（如 [41, 62] 中的思路），则会得到双准则近似；为了实现所需的有界半径，球必须覆盖少于 n 个点。

算法 1 最小包围球算法 I

输入： 两个参数 $0 < \epsilon, \eta < 1$；一个关于最小包围球问题的 (ϵ^2, β)-稳定实例 P，位于 $\mathbb{R}^d$ 中，其中 β 大于给定的下界 $\beta_0 > 0$。1：从 P 中均匀随机地抽取一个集合 S，包含 $\Theta(\frac{1}{\beta_0} \cdot \max\{\log \frac{1}{\eta}, d \log \frac{d}{\beta_0}\})$ 个点。2：应用任意近似最小包围球算法（例如基于核心集的算法 [20]），计算 S 的 $(1+\epsilon^2)$-半径近似最小包围球，并将得到的球记为 $\mathbb{B}(c, r)$。3：输出球 $\mathbb{B}\big(c, \frac{1+(2\sqrt{2}+\sqrt{3})\epsilon}{1-\epsilon^2} r\big)$。

定理 3. 以概率 $1-\eta$，算法 *1* 返回 P 的 λ-半径近似最小包围球，其中

$$\lambda = \frac{\big(1+(2\sqrt{2}+\sqrt{3})\epsilon\big)(1+\epsilon^2)}{1-\epsilon^2} = 1 + O(\epsilon). \tag{11}$$

在证明定理 3 之前，我们先证明以下引理。

引理 3. 设 S 是从给定点集 $P \subset \mathbb{R}^d$ 中随机且独立采样的 $\Theta(\frac{1}{\beta_0} \cdot \max\{\log \frac{1}{\eta}, d \log \frac{d}{\beta_0}\})$ 个点，且 B 是覆盖 S 的任意一个球。那么，以概率 $1-\eta$，$|B \cap P| \geq (1-\beta_0)|P|$。

证明. 考虑范围空间 $\Sigma = (P, \Phi)$，其中每个范围 $\phi \in \Phi$ 都是空间中一个球的补集。在范围空间中，子集 $Y \subset P$ 是一个 β_0-网，当且仅当

$$\text{for any } \phi \in \Phi,\ \frac{|P \cap \phi|}{|P|} \geq \beta_0 \Longrightarrow Y \cap \phi \neq \emptyset. \tag{12}$$

大小为 $|S| = \Theta(\frac{1}{\beta_0} \cdot \max\{\log \frac{1}{\eta}, d \log \frac{d}{\beta_0}\})$，并且由 [58,92] 可知，$S$ 以概率 $1-\eta$ 是 P 的一个 β_0-网。因此，若 $|B \cap P| < (1-\beta_0)|P|$，即 $|P \setminus B| > \beta_0 |P|$，则有 $S \cap \big(P \setminus B\big) \neq \emptyset$。这与 S 被 B 覆盖这一事实矛盾。因此，$|B \cap P| \geq (1-\beta_0)|P|$。□

证明。**(定理 3 的)** 记 o 为 P 的**最小包围球**中心。由于 $S \subset P$ 和 $\mathbb{B}(c,\ r)$ 是 S 的 $(1+\epsilon^2)$-半径近似最小包围球，我们知道 $r \leq (1+\epsilon^2)$**半径**(P)。此外，引理 3 意味着 $|\mathbb{B}(c,\ r) \cap P| \geq (1-\beta_0)|P|$ 的概率为 $1-\eta$。假设这成立，并令 $P' = \mathbb{B}(c,\ r) \cap P$。那么，我们得到距离

$$||c - o|| \leq (2\sqrt{2}+\sqrt{3})\epsilon \mathbf{Rad}(P) \tag{13}$$

根据定理 2（我们设定 $\epsilon' = \epsilon$）。为简单起见，我们用 x 表示 $(2\sqrt{2}+\sqrt{3})\epsilon$。不等式 (13) 表明，点集 P被球 $\mathbb{B}(c,\ (1+x)$**半径**$(P))$ 覆盖。请注意，我们无法直接将 $\mathbb{B}(c,\ (1+x)$**半径**$(P))$ 作为最终结果返回，因为我们不知道 **半径**(P) 的值。因此，我们必须估计半径 $(1+x)$**半径**(P)。

由于 P' 被 $\mathbb{B}(c,\ r)$ 和 $|P'| \geq (1-\beta_0)|P|$ 覆盖，根据定义 3，r 应至少为 $(1-\epsilon^2)$**半径**(P)。因此，我们有

$$\frac{1+x}{1-\epsilon^2} r \geq (1+x)\mathbf{Rad}(P). \tag{14}$$

也就是说，P 被球 $\mathbb{B}(c, \frac{1+x}{1-\epsilon^2} r)$ 覆盖。此外，半径

$$\frac{1+x}{1-\epsilon^2} r \leq \frac{1+x}{1-\epsilon^2}(1+\epsilon^2)\mathbf{Rad}(P). \tag{15}$$

这意味着球 $\mathbb{B}(c, \frac{1+x}{1-\epsilon^2} r)$ 是 P 的 λ-半径近似最小包围球，其中

$$\lambda = (1+\epsilon^2)\frac{1+x}{1-\epsilon^2} = \frac{\big(1+(2\sqrt{2}+\sqrt{3})\epsilon\big)(1+\epsilon^2)}{1-\epsilon^2} \tag{16}$$

且当 ϵ 是 (0,1) 中一个固定的小数时，$\lambda = 1 + O(\epsilon)$。 □

算法 1 的运行时间。为简单起见，我们假设 $\log \frac{1}{\eta} < d \log \frac{d}{\beta_0}$。如果我们使用基于核心集的算法 [20] 来计算 $\mathbb{B}(c, r)$（见备注 3），则算法 1 的运行时间为 $O\big(\frac{1}{\epsilon^2}(|S|d + \frac{1}{\epsilon^6} d)\big) = O\big(\frac{d^2}{\epsilon^2 \beta_0} \log \frac{d}{\beta_0} \mp \frac{d}{\epsilon^8}\big) = \tilde{O}(d^2)$，其中隐藏因子取决于 ϵ 和 β_0。

注记 5. 如果维数 d 过高，可以使用随机投影技术*Johnson-Lindenstrauss (JL)* 变换 [36] 来近似保持包围球的半径 [2, 66, 87]。然而，这对于降低算法 1 的时间复杂度并无帮助。如果我们在步骤 1 中对采样的 $\Theta(\frac{d}{\beta_0} \log \frac{d}{\beta_0})$个点应用 JL 变换，那么 JL 变换这一步本身就需要 $\Omega(\frac{d^2}{\beta_0} \log \frac{d}{\beta_0})$时间。

4.2 第二种算法

我们在第 4.1 节中的第一个算法很简单，但其样本量（即 采样点的数量）取决于维数 d，而**第二种算法的样本量则与和**d **均无关**（当应用核函数时，这一点尤其重要，因为新的维数可能非常高，甚至 $+\infty$）。我们先简要概述一下我们的思路。

第二种算法的总体思路： 回顾备注 3（ii）。如果我们知道 $(1+\epsilon)$**半径**(P) 的值，就可以执行与定理 1 中所述几乎相同的核心集构造过程，从而获得 **最小包围球**(P) 的近似中心。唯一的区别在于，每次迭代中，我们向 o_i 添加一个距离至少为 $(1+\epsilon)$**半径**(P) 的点。这样，我们就无需选择距离 o_i 最远的点，因为这一操作不可避免地具有线性时间复杂度。为了以次线性时间实现这一策略，我们需要先确定 $(1+\epsilon)$**半径**(P) 的值。我们在下文给出引理 4，以估计 **半径** (P) 的范围，然后在该范围内进行二分搜索，近似确定 $(1+\epsilon)$ **半径** (P) 的值。根据稳定性，我们注意到，核心集构造过程可以充当“预言机”，帮助我们猜测 $(1+\epsilon)$**半径**(P) 的值（见算法 3）。设 $h > 0$ 是一个候选值。每次迭代中，我们向 o_i 添加一个距离至少为 h 的点。我们证明，如果 $h \geq (1+\epsilon)$**半径**(P)，该过程不会持续超过 z 次迭代；而如果 $h < (1-\epsilon)$**半径**(P)，则该过程以常数概率持续超过 z 次迭代，其中 z 是定理 1 中所述核心集的大小。此外，在核心集构造过程中，我们通过随机采样而非确定性方式将点加入核心集。这里有一个小问题：我们需要在定理 1 中用 ϵ^2 替换 ϵ，以便在后续分析中实现总体 $(1+O(\epsilon))$ 半径近似。

引理 4。给定参数 $\eta \in (0,\ 1)$，任取一点 $p_1 \in P$，并均匀随机地抽取一个样本 $Q \subset P$，其中 $|Q| = \frac{1}{\beta_0} \log \frac{1}{\eta}$。令 $p_2 = \arg\max_{p \in Q} ||p - p_1||$。则以概率 $1 - \eta$，

$$\mathbf{Rad}(P) \in [\frac{1}{2}||p_1 - p_2||, \frac{1}{1-\epsilon^2}||p_1 - p_2||]. \tag{17}$$

证明。 首先，**半径**(P) 的下界显然，因为 $||p_1 - p_2||$ 始终不大于 2**半径**(P)。接下来，我们考虑上界。令 $\mathbb{B}(p_1, l)$ 为恰好覆盖 P 中 $(1-\beta_0)n$ 个点的球，因此根据定义 3， $l \geq (1-\epsilon^2)$**半径**(P)。为了完成证明，我们还需要以下在 [39] 中给出的经典引理。

引理 5。 *[39]*设 N 是一个元素集合， N' 是 N 的一个子集，其大小为 $|N'| = \tau\,|N|$，其中 $\tau \in (0, 1)$。给定 $\eta \in (0, 1)$，若从 N 中随机抽取 $\ln 1/\eta \ln 1/(1-\tau) \leq \frac{1}{\tau} \ln \frac{1}{\eta}$ 个元素，则样本以至少 $1 - \eta$ 的概率包含 N' 中的至少一个元素。

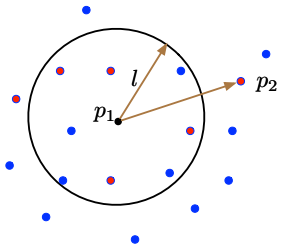

图 4：引理 4 的示意图；红色点是采样集合 Q。

在引理 5 中，令 N 和 N' 分别为点集 P 和子集 $P \setminus \mathbb{B}(p_1,\ l)$。根据引理 5（令 $\tau = \beta_0$），我们知道 Q 至少包含一个来自 N' 的点。也就是说， Q 至少包含一个位于 $\mathbb{B}(p_1,\ l)$ 之外的点。此外，由于 $p_2 = \arg\max_{p \in Q} ||p - p_1||$，我们有 $||p_1 - p_2|| \geq l \geq (1-\epsilon^2)$**半径**$(P)$，也就是说，**半径**$(P) \leq \frac{1}{1-\epsilon^2}||p_1 - p_2||$（示意图见图 4）。□

算法 3 作为算法 2 的子程序。在算法 3 中，我们只需按照定理1中的描述设置 $z = \frac{3}{\epsilon^2}$，其中 $s = 1/3$ （如前所述，我们将 ϵ 替换为 ϵ^2）；在步骤 2(1) 中，我们计算与**最小包围球**(T) 的中心距离小于 $s\frac{\epsilon^2}{1+\epsilon^2}$**半径** (T) 的 o_i。

算法 2 最小包围球算法 II

输入：两个参数 $0 < \epsilon, \eta_0 < 1$；MEB 问题在 $\mathbb{R}^d$中的一个 (ϵ^2, β)-稳定实例 P，其中 β大于给定下界 $\beta_0 > 0$。设定引理 4 得到的 Rad(P) 区间 $[a, b]$。1：在集合 $O = \{1-\epsilon^2 a,\ 1+\epsilon^2 h,\ 1-\epsilon^2 a,\ \cdots,\ 1+\epsilon^2 w,\ 1-\epsilon^2 a = 1+\epsilon^2 b\}$，$w = \lceil \frac{2}{1+\epsilon^2}\frac{2}{1-\epsilon^2}\rceil + 1 = ()$，$()()$，，$()()()$ 中，其中 $\log()(\frac{1}{\epsilon^2})$，使用算法 及 $z = \frac{3}{\epsilon^2}$和 $\eta = \frac{3}{2}\eta_0 \log w$进行二分搜索以求得该值。2：假设当 $h = (1+\epsilon^2)^{i_0}(1-\epsilon^2)a$ 时，算法 3 返回“否”；当 $h = (1+\epsilon^2)^{i_0+1}(1-\epsilon^2)a$时，返回“是”。3：使用 $h = (1+\epsilon^2)^{i_0+2}a$，$z = \frac{3}{\epsilon^2}$和 $\eta = \eta_0/2$再次运行算法 3；循环停止时，令 $\tilde{o}$ 为得到的 T的球心。4：返回球 $\mathbb{B}(\tilde{o}, r)$，其中 $r = (2\sqrt{2} + \frac{2\sqrt{6}}{\sqrt{1-\epsilon^2}})\frac{1}{1+\epsilon^2}h$。

用于测试 h的算法 3 预言机

输入： 一个实例 P、参数 $\eta \in (0, 1)$、 $h > 0$以及一个正整数 z。1：首先，任意选择一个点 $p \in P$，并令 $T = \{p\}$。2： $i = 1$；重复执行以下步骤：

(1) 按照定理1所述，计算 T的一个近似最小包围球，并将球心设为 o_i（将 ϵ 替换为 ϵ^2，并设置 $s = 1/3$）。(2) 均匀随机地抽取一个包含$|Q| = \frac{1}{\beta_0} \log \frac{z}{\eta}$的集合 $Q \subset P$。(3) 选择距离 o_i最远的点 $q \in Q$，并将其加入 T。(4) 如果 $||q - o_i|| < h$，则停止循环并输出“是”。(5) $i = i + 1$；如果 $i > z$，则停止循环并输出“否”。

定理 4。 以 $1 - \eta_0$的概率，算法 *2* 返回 P的一个 λ-半径近似最小包围球，其中

$$\lambda = \frac{(1 + x_1)(1 + x_2)}{1 + \epsilon^2} = 1 + O(\epsilon) \quad with \quad x_1 = O\big(\frac{\epsilon^2}{1 - \epsilon^2}\big), x_2 = O\big(\frac{\epsilon}{\sqrt{1 - \epsilon^2}}\big). \tag{18}$$

运行时间为 $\tilde{O}\big((\frac{1}{\epsilon^2\beta_0} + \frac{1}{\epsilon^8})d\big)$，其中 $\tilde{O}(f) = O(f \cdot$多对数函数$(\frac{1}{\epsilon}, \frac{1}{\eta_0}))$。—

在证明定理 4 之前，我们先给出引理 6。

引理 6。 如果 $h \geq (1 + \epsilon^2)$ **半径**(P)，算法 *3* 返回“是”；否则，如果 $h < (1 - \epsilon^2)$ **半径**(P)，则算法 *3* 以至少 $1 - \eta$ 的概率返回“否”。

证明。 首先，假设 $h \geq (1 + \epsilon^2)$**半径**(P)。回顾定理 1 后面的注记。如果我们始终向 o_i 添加一个与 $h \geq (1 + \epsilon^2)$**半径**(P) 的距离至少为 q 的点，则循环 2(1)-(5) 不会继续超过 z 次迭代，即，算法 3 将返回“是”。

现在，我们考虑 $h < (1 - \epsilon^2)$**半径**(P) 的情形。类似于引理 4 的证明，我们考虑覆盖 P 中恰好 $(1 - \beta_0)n$ 个点的球 $\mathbb{B}(o_i, l)$。根据定义 3，我们知道 $l \geq (1 - \epsilon^2)$**半径**$(P) > h$。此外，根据引理 5，以 $1 - \eta/z$ 的概率，样本 Q 至少包含一个位于 $\mathbb{B}(o_i, l)$ 外部的点。通过取并集界，以概率 $(1 - \eta/z)^z \geq 1 - \eta$， $||q - o_i||$ 始终大于 h，最终算法 3 将返回“否”。□

证明。 **（定理 4 的证明）** 由于算法 3 在 $h = (1 + \epsilon^2)^{i_0}(1 - \epsilon^2)a$ 时返回“否”，并在 $h = (1 + \epsilon^2)^{i_0+1}(1 - \epsilon^2)a$ 时返回“是”，由引理 6 可知

$$(1 + \epsilon^2)^{i_0}(1 - \epsilon^2)a < (1 + \epsilon^2)\mathbf{Rad}(P); \tag{19}$$

$$(1 + \epsilon^2)^{i_0+1}(1 - \epsilon^2)a \geq (1 - \epsilon^2)\mathbf{Rad}(P). \tag{20}$$

上述不等式共同蕴含

$$\frac{(1 + \epsilon^2)^3}{1 - \epsilon^2}\mathbf{Rad}(P) > (1 + \epsilon^2)^{i_0+2}a \geq (1 + \epsilon^2)\mathbf{Rad}(P). \tag{21}$$

因此，在步骤 3 中以 $h = (1 + \epsilon^2)^{i_0+2}a$ 运行算法 3 时，算法返回“是”（根据 (21) 的右端）。接着，考虑球 $\mathbb{B}(\tilde{o}, h)$。我们断言 $|P \backslash \mathbb{B}(\tilde{o}, h)| < \beta_0 n$。否则，在算法 3 的步骤 2(2) 中，样本 Q 以 $1 - \eta/z$ 的概率至少包含一个位于 $\mathbb{B}(\tilde{o}, h)$ 外部的点，即，循环将继续。因此，这与算法返回“是”这一事实矛盾。令 $P' = P \cap \mathbb{B}(\tilde{o}, h)$，则 $|P'| \geq (1 - \beta_0)n$。此外，(21) 的左端表明

$$h = (1 + \epsilon^2)^{i_0+2}a < (1 + \frac{8\epsilon^2}{1 - \epsilon^2})\mathbf{Rad}(P). \tag{22}$$

现在，我们可以应用定理 2，并在该定理中将“ϵ”设为“$\sqrt{\frac{8\epsilon^2}{1-\epsilon^2}}$”。令 o 为**最小包围球**(P) 的中心。因此，我们有

$$||\tilde{o} - o|| < (2\sqrt{2} + 2\sqrt{6}/\sqrt{1-\epsilon^2})\epsilon \cdot \mathbf{Rad}(P). \tag{23}$$

为简便起见，我们令 $x_1 = \frac{8\epsilon^2}{1-\epsilon^2}$ 和 $x_2 = (2\sqrt{2}+2\sqrt{6}/\sqrt{1-\epsilon^2})\epsilon$。因此，(22) 和 (23) 中有 $h \le (1+x_1)$**半径**(P) 和 $||\tilde{o}-o|| \le x_2$**半径**(P)。由 (23) 可知，$P \subset \mathbb{B}(\tilde{o}, (1+x_2)$**半径**$(P))$。由 (21) 右侧可知，$(1+x_2)$**半径**$(P) \le \frac{1+x_2}{1+\epsilon^2} h$。因此，我们有$P \subset \mathbb{B}\left(\tilde{o}, \frac{1+x_2}{1+\epsilon^2} h\right)$ 其中 $\frac{1+x_2}{1+\epsilon^2} h = \frac{1+(2\sqrt{2}+\frac{2\sqrt{6}}{\sqrt{1-\epsilon^2}})\epsilon}{1+\epsilon^2} h$。此外，半径

$$\frac{1+x_2}{1+\epsilon^2} h \underbrace{\le}_{\text{by (22)}} \frac{(1+x_2)(1+x_1)}{1+\epsilon^2} \mathbf{Rad}(P) = \lambda \cdot \mathbf{Rad}(P). \tag{24}$$

因此，$\mathbb{B}\left(\tilde{o}, \frac{1+x_2}{1+\epsilon^2} h\right)$ 是 P的一个 λ半径近似最小包围球，并且 $\lambda = 1 + O(\epsilon)$，其中 ϵ 是(0,1) 中的一个固定小数。

成功概率。 算法 3 的成功概率为 $1-\eta$。在算法 2 中，我们分别在步骤 1 中设定$\eta = \frac{\eta_0}{\log w}$，并在步骤 3 中设定 $\eta = \eta_0/2$。利用并集界，算法 2 的成功概率为$(1 - \frac{\eta_0}{\log w})^{\log w}(1 - \eta_0/2) > 1 - \eta_0$。

运行时间。 作为子程序，算法 3 的运行时间为 $O(z(\frac{1}{\beta_0}(\log \frac{z}{\eta})d + \frac{1}{\epsilon^6}d))$；算法 2 调用该子程序 $O\big(\log(\frac{1}{\epsilon^2})\big)$ 次。注意，$z = O(\frac{1}{\epsilon^2})$。因此，总运行时间为 $\tilde{O}\big((\frac{1}{\epsilon^2\beta_0} + \frac{1}{\epsilon^8})d\big)$。□

5 一般最小包围球的次线性算法

在第 4 节中，我们在稳定性假设下提出次线性时间算法。具体而言，我们假设给定实例是 (ϵ^2, β)-稳定的，并且 β 大于一个合理的已知下界 β_0。然而，当 β_0 的值未知时，我们无法确定该算法的样本量；或者我们可能只知道一个平凡下界，例如，例如 $\frac{1}{n}$，此时样本量可能会过大。因此，本节考虑不作稳定性假设的一般情形。

高层次思想。有趣的是，为稳定实例发展的思想甚至可以帮助我们在稳定性假设不成立时，为最小包围球设计一种混合方法。首先，我们“假设”输入实例是 (α, β)-稳定的，其中“α”和“β”根据预先指定的半径误差界 ϵ 和覆盖误差界 δ 进行设计，并计算一个“潜在的”$(1+\epsilon)$-半径近似（例如一个球 B_1）；然后，我们计算一个“潜在的”$(1-\delta)$-覆盖近似（例如一个球 B_2），其中“覆盖近似”的定义见定义 2；最后，根据它们的半径比确定最终输出。具体而言，我们设置一个由给定半径误差界 ϵ 确定的阈值 τ。如果该比值不大于 τ，我们可以推断 B_1 是一个 “ 真正的 ”$(1+\epsilon)$- 半径近似，并返回它；否则，我们返回 B_2，它是一个 “ 真正的 ”$(1-\delta)$-覆盖近似。此外，对于后一种情况（即返回一个 $(1-\delta)$-覆盖近似），我们将证明，所提算法得到的半径不仅严格小于 $\mathbf{Rad}(P)$，而且与 $\mathbf{Rad}(P)$ 相比还具有 $\Theta(\epsilon^2) \cdot \mathbf{Rad}(P)$ 的间隔（即，返回半径至多为 $\big(1-\Theta(\epsilon^2)\big) \cdot \mathbf{Rad}(P)$）。我们的算法只需要均匀采样，并对输入数据进行单遍处理，其内存空间复杂度为 $O(d)$（隐藏因子取决于 ϵ 和 δ）；如果输入数据矩阵是稀疏的（即 $M = o(nd)$），则时间复杂度为次线性。

在介绍我们的算法之前，我们需要先给出带离群点的最小包围球问题的正式定义，因为它将用于计算（$1-\delta$）-覆盖近似。

定义 4（带离群点的最小包围球）。给定 $\mathbb{R}^d$ 中由 n 个点组成的集合 P，以及一个小参数 $\gamma \in [0,1)$，带离群点的最小包围球问题旨在寻找能够覆盖$(1-\gamma)n$个点的最小球。也就是说，任务是寻找 P 的一个子集，其大小为$(1-\gamma)n$，使得由此得到的最小包围球在所有可能的子集选择中最小。所得的球记为**最小包围球**(P,γ)。

为方便起见，我们用 P_{opt} 表示相对于 **最小包围球**$(P,\ \gamma)$ 的 P 最优子集。也就是说，$P_{\text{opt}} = \arg\min_{Q} \left\{ \text{半径}(Q) \mid Q \subset P,\ |Q| = (1-\gamma)n \right\}$。由定义 4 可知，主要挑战在于确定 P 的子集。与定义 2 类似，我们也为带离群点的最小包围球定义半径近似和覆盖近似。

定义 *5*（半径近似和覆盖近似）。设 $0 < \epsilon, \delta < 1$。若球 $\mathbb{B}(c,r)$ 覆盖 P 中的 $(1-\gamma)n$ 个点，且半径为 $r \le (1+\epsilon)$半径(P_{opt})，则称其为 $MEB(P,\gamma)$ 的 $(1+\epsilon)$-半径近似。另一方面，若该球覆盖 P 中至少 $(1-\delta-\gamma)n$ 个点，且半径为 $r \le$半径(P_{opt})，则称其为 $MEB(P,\gamma)$ 的 $(1-\delta)$-覆盖近似。

双准则$(1+\epsilon, 1-\delta)$-近似是一个至少覆盖 $(1-\delta-\gamma)n$ 个点、且半径至多为$(1+\epsilon)$**半径**(P_{opt})的球。

内容提要。首先，我们在第 5.1 节介绍两种随机采样技术，它们是设计第 5.2 节中用于带离群点的最小包围球的次线性双准则近似算法的关键。基于第 5.2 节中的双准则近似，我们可以在第 5.3 节求解一般最小包围球问题。

5.1 处理离群点的两个关键引理

为了说明我们的思路，考虑使用第 2.1 节中的核心集构造方法，为带离群点的最小包围球的一个实例 (P,γ) 计算双准则 $(1+\epsilon, 1-\delta)$-近似。令 o_i 为当前迭代中得到的球心，Q 为从 P 到 o_i 的最远 $(\delta+\gamma)n$ 个点构成的集合。更新 o_i 的一个关键步骤是找到集合 $P_{\text{opt}} \cap Q$ 中的一个点（形式化分析见第 5.2 节）。实际上，可以通过从 Q 中进行随机采样来实现。不过，这要求预先计算集合 Q，其时间复杂度为 $\Omega(nd)$。为了使运行时间保持次线性，我们需要用更巧妙的方法从 $P_{\text{opt}} \cap Q$ 中找到一个点。由于 P_{opt} 与集合 Q 中的离群点混杂在一起，简单的均匀采样无法实现我们的目标。为了解决这一问题，我们提出了一种称为“均匀自适应采样”的“两层”采样过程。粗略地说，我们首先取一个大小为 n' 的随机样本 A（即均匀采样步骤），然后从 Q' 中随机选择一个点；Q' 是从 o_i 到 A 的最远 $3\,2\,(\delta+\gamma)n'$ 个点构成的集合（即自适应采样步骤）。根据引理 7，以至少 $(1-\eta_1)\,3^{\delta}\,(\delta+\gamma)$ 的概率，选出的点属于 $P_{\text{opt}} \cap Q$；更重要的是，样本量 n' 与 n 和 d 无关。证明引理 7 的关键是说明交集 $Q' \cap \left(P_{\text{opt}} \cap Q\right)$ 的大小足够大。通过为 n' 设定适当的值，我们可以证明 $|Q' \cap \left(P_{\text{opt}} \cap Q\right)|$ 的一个下界。

引理 7（均匀自适应采样）。令 $\eta_1 \in (0,\ 1)$。如果我们从 P 中独立且均匀随机地采样 $n' = O(\frac{1}{\delta}\log\frac{1}{\eta_1})$ 个点，并令 Q' 为样本中距离 o_i 最远的 $3\,2(\delta+\gamma)n'$ 个点的集合，那么，以至少 $1-\eta_1$ 的概率，以下结论成立

$$\frac{\left|Q' \cap \left(P_{opt} \cap Q\right)\right|}{|Q'|} \ge \frac{\delta}{3(\delta+\gamma)}. \tag{25}$$

证明。令 A 表示从 P中采样得到的点集合。首先，我们知道 $|Q| = (\delta+\gamma)n$且 $|P_{\text{opt}} \cap Q| \ge \delta n$（因为 Q中至多有 γn 个离群点）。为便于说明，令

$\lambda = \frac{|P_{\text{opt}} \cap Q|}{n} \geq \delta$。令 $\{x_i \mid 1 \leq i \leq n'\}$ 为 n' 个独立随机变量：若 A 的第 i个采样点属于 $P_{\text{opt}} \cap Q$，则 $x_i = 1$；否则 $x_i = 0$。因此，对每个 i，都有 $E[x_i] = \lambda$。令 σ 为 (0，1) 中的一个小参数。利用 Chernoff界，我们有**概率**$\Big(\sum_{i=1}^{n'} x_i \notin (1 \pm \sigma)\lambda n' \Big) \leq e^{-O(\sigma^2 \lambda n')}$。也就是说，

$$\mathbf{Pr}\Big(|A \cap \big(P_{\text{opt}} \cap Q\big)| \in (1 \pm \sigma)\lambda n' \Big) \geq 1 - e^{-O(\sigma^2 \lambda n')}. \tag{26}$$

同样地，我们有

$$\mathbf{Pr}\Big(|A \cap Q| \in (1 \pm \sigma)(\delta + \gamma)n' \Big) \geq 1 - e^{-O(\sigma^2 (\delta+\gamma) n')}. \tag{27}$$

注意，$n' = O\left(\frac{1}{\delta} \log \frac{1}{\eta_1}\right)$。通过为 (26) 和 (27) 设置 $\sigma < 1/2$，我们有

$$\Big| A \cap \big(P_{\text{opt}} \cap Q\big) \Big| > \frac{1}{2}\delta n' \quad \text{and} \quad \Big| A \cap Q \Big| < \frac{3}{2}(\delta + \gamma)n' \tag{28}$$

以概率 $1 - \eta_1$。注意，Q 包含到 o_i 的所有最远（$\delta + \gamma$）n 个点。记 l_i 为从 P 到 o_i 的（（$\delta + \gamma$）$n + 1$）大距离。则我们有

$$A \cap Q = \{p \in A \mid ||p - o_i|| > l_i\}. \tag{29}$$

此外，由于 Q' 是从 A 到 o_i 的最远 3-2（$\delta + \gamma$）n' 个点的集合，因此存在某个 $l_i' > 0$ 使得

$$Q' = \{p \in A \mid ||p - o_i|| > l_i'\}. \tag{30}$$

(29) 和 (30) 一起意味着，要么（$A \cap Q$）$\subseteq Q'$，要么 $Q' \subseteq$（$A \cap Q$）。由于 $\big|A \cap Q\big| < \frac{3}{2}$（$\delta + \gamma$）$n'$ 且 $|Q'| = \frac{3}{2}$（$\delta + \gamma$）n'，我们知道 $\Big(A \cap Q\Big) \subseteq Q'$。因此，

$$\Big(A \cap \big(P_{\text{opt}} \cap Q\big) \Big) = \Big(P_{\text{opt}} \cap \big(A \cap Q\big) \Big) \subseteq Q'. \tag{31}$$

此外，显然

$$\Big(A \cap \big(P_{\text{opt}} \cap Q\big) \Big) \subseteq \big(P_{\text{opt}} \cap Q\big). \tag{32}$$

由上述（31）和（32）可得

$$\Big(A \cap \big(P_{\text{opt}} \cap Q\big) \Big) \subseteq \Big(Q' \cap \big(P_{\text{opt}} \cap Q\big) \Big). \tag{33}$$

此外，由于 $Q' \subseteq A$，我们有

$$\Big(Q' \cap \big(P_{\text{opt}} \cap Q\big) \Big) \subseteq \Big(A \cap \big(P_{\text{opt}} \cap Q\big) \Big). \tag{34}$$

因此，（33）和（34）共同推出 $Q' \cap \big(P_{\text{opt}} \cap Q\big) = A \cap \big(P_{\text{opt}} \cap Q\big)$，并且 hen Con ce

$$\frac{\Big| Q' \cap \big(P_{\text{opt}} \cap Q\big) \Big|}{|Q'|} = \frac{\Big| A \cap \big(P_{\text{opt}} \cap Q\big) \Big|}{|Q'|} \geq \frac{\delta}{3(\delta + \gamma)}, \tag{35}$$

其中 th 该不等式由 (28) 的第一个不等式以及事实 $|Q'| = \frac{3}{2}(\delta + \gamma)n'$ 得出。 - □

随机采样方法并不总能保证成功。为了提高整体成功概率，我们必须多次重复运行算法，每次运行都会生成一个候选解（即球心）。因此，我们必须从中选出最佳解作为最终解。稍微滥用一下记号，我们仍用 o_i 表示候选球心；由于我们的目标是达到 $(1+\epsilon, 1-\delta)$-近似，因此需要计算从 P 到 o_i 的第 $\big((\delta+\gamma)n+1\big)$ 大距离，记为 l_i。一种直接的方法是对每个候选解在线性时间内计算“l_i”的值，并返回 l_i 最小的那个。本节提出“夹逼引理”，以便在次线性时间内估计 l_i。设 B 为引理 8 中从 P 采样得到的 n'' 的集合，l_i 为从 B 到 o_i 的第 $\big((1+\delta/\gamma)^2\gamma n''+1\big)$ 大距离。若能证明引理 8 的不等式 (37)˜ 和 (38)，则可推出 l_i 是 l_i 的合格估计：如果 $\mathbb{B}(o_i, l_i)$ 是一个 ˜$(1+\epsilon, 1-\delta)$-近似，那么球 $\mathbb{B}(o_i, l_i)$ 应当是一个 $(1+\epsilon, 1-O(\delta))$-近似。其˜ ˜核心思想是证明球 $\mathbb{B}(o_i, l_i)$ 被两个球 $\mathbb{B}(o_i, l_i')$ 和 $\mathbb{B}(o_i, l_i)$“夹住”，˜其中 l_i' 是一个经过精心设计、满足以下条件的值

$$\text{(i) } \tilde{l}_i' \le \tilde{l}_i \le l_i \text{ and (ii) } \Big|P \setminus \mathbb{B}(o_i, \tilde{l}_i')\Big| \le (\gamma + O(\delta))n. \tag{36}$$

˜参见图5。l_i' 的这两个条件可以推出引理 8 中的不等式 (37) 和 (38)。与引理 7 类似，样本量 n'' 也独立于 n 和 d。

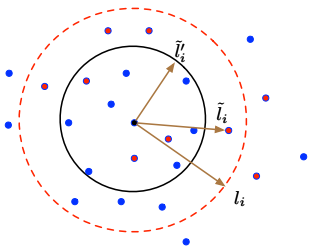

图5：红色点是引理 8 中采样得到的 n'' 点，而 $\big((1+\delta/\gamma)^2\gamma n''+1\big)$第˜远的点位于由球面 $\mathbb{B}(o_i, l_i')$ 和 $\mathbb{B}(o_i, l_i)$ 所界定的环带内。

引理 8（夹逼引理）。令 $\eta_2 \in (0,1)$ 并假设 $\delta < \gamma/3$。如果我们从 P 中独立且均匀随机地抽取 $n'' = O\big(\frac{\gamma}{\delta^2}\log\frac{1}{\eta_2}\big)$ 个点，并令 $\tilde{l}_i$ 为样本到 o_i的 $\big((1+\delta/\gamma)^2\gamma n''+1\big)$大距离，则以概率 $1-\eta_2$，以下结论成立

$$\tilde{l}_i \le l_i; \tag{37}$$

$$\Big|P \setminus \mathbb{B}(o_i, \tilde{l}_i)\Big| \le (\gamma + 5\delta)n. \tag{38}$$

证明。令 B 表示从 P中抽取的 n'' 点集合。为简单起见，令 $t = (\delta+\gamma)n$。假设 $\tilde{l}_i' > 0$ 是满足 $\Big|P \setminus \mathbb{B}(o_i, \tilde{l}_i')\Big| = (\gamma+\delta^2 n)$ 的值。回顾一下，l_i 是从 P到 o_i的第 $\big(t+1\big)$大距离。由于 $(\delta+\gamma)n < (\gamma+\delta^2 n)$，容易得知 $\tilde{l}_i' \le l_i$。下面，我们旨在证明，从 B 出发的第 $\big((1+\delta/\gamma)^2\gamma n''+1\big)$远的点位于由球面 $\mathbb{B}(o_i, l_i')$ 和 $\mathbb{B}(o_i, l_i)$界定的环带内（见图5）。

注意大小 $|B| = n'' = O\big(\frac{\gamma}{\delta^2}\log\frac{1}{\eta_2}\big)$。再次使用 Chernoff界（令 $\sigma = \delta/2$），并采用证明式（28）时的相同思路，可得

$$\Big|B \setminus \mathbb{B}(o_i, \tilde{l}_i')\Big| \geq (1 - \frac{\delta}{2\gamma})\frac{(\gamma+\delta)^2}{\gamma-\delta}n'' > (1 - \frac{\delta}{\gamma})\frac{(\gamma+\delta)^2}{\gamma-\delta}n'' = (1+\delta/\gamma)^2\gamma n''; \tag{39}$$

$$\Big|B \cap Q\Big| \leq (1 + \frac{\delta}{2\gamma})\frac{t}{n}n'' < (1+\delta/\gamma)\frac{t}{n}n'' = (1+\delta/\gamma)^2\gamma n'', \tag{40}$$

以概率 $1 - \eta_2$。假设 (39) 和 (40) 均成立。回顾一下，$\tilde{l}_i$ 是从采样点 B 到 o_i的距离中第 $\big((1+\delta/\gamma)^2\gamma n'' + 1\big)$ 大的距离，因此 $\Big|B \setminus \mathbb{B}(o_i, \tilde{l}_i)\Big| = (1+\delta/\gamma)^2\gamma n''$。结合 (39)，我们有 $\Big|B \setminus \mathbb{B}(o_i, \tilde{l}_i)\Big| \leq \Big|B \setminus \mathbb{B}(o_i, \tilde{l}_i')\Big|$，即

$$\tilde{l}_i \geq \tilde{l}_i'. \tag{41}$$

不等式 (40) 表明，从 B˜到 o_i 的第 $\big((1+\delta/\gamma)^2\gamma n'' + 1\big)$ 个最远点（设为 q_x）不在 Q 中。接着，我们断言 $\mathbb{B}(o_i, l_i) \cap Q = \emptyset$。否则，设 $q_y \in \mathbb{B}(o_i, l_i) \cap Q$。于是我们有

$$||q_y - o_i|| \leq \tilde{l}_i = ||q_x - o_i||. \tag{42}$$

注意，Q 是 P 中到 o_i 的最远 t 点的集合。因此，$q_x \notin Q$ 意味着

$$||q_x - o_i|| < \min_{q \in Q}||q - o_i|| \leq ||q_y - o_i|| \tag{43}$$

˜，这与 (42) 矛盾。因此，$\mathbb{B}(o_i, l_i) \cap Q = \emptyset$。此外，由于 $\mathbb{B}(o_i, l_i)$ 恰好排除了最远的 t 个点（即，Q），「$\mathbb{B}(o_i, l_i) \cap Q = \emptyset$」意味着

$$\tilde{l}_i \leq l_i. \tag{44}$$

综上，由 (41) 和 (44) 可得 $\tilde{l}_i \in [\tilde{l}_i', l_i]$，即，从 B 出发的第 $\big((1+\delta/\gamma)^2\gamma n'' + 1\big)$ 个最远点˜位于球面 $\mathbb{B}(o_i, l_i')$ 和 $\mathbb{B}(o_i, l_i)$ 所围成的环带内，如˜ ˜图5 所示。此外，$l_i \geq l_i'$ 意味着

$$\Big|P \setminus \mathbb{B}(o_i, \tilde{l}_i)\Big| \leq \Big|P \setminus \mathbb{B}(o_i, \tilde{l}_i')\Big| = \frac{(\gamma+\delta)^2}{\gamma-\delta}n < (\gamma + 5\delta)n, \tag{45}$$

其中，最后一个等式来自假设 $\delta < \gamma/3$。因此，(37) 和 (38) 在引理 8 中成立。□

注记 6. 实际上，我们提出的均匀自适应采样方法和夹逼引理相当通用，我们将在第 7 节说明，如何将它们推广到更广泛的带离群点包围问题。

5.2 双准则近似的次线性时间算法

在本节中，我们提出一种次线性时间算法，用于计算输入实例 (P, γ) 的双准则 $(1+\epsilon, 1-\delta)$-近似；也就是说，返回的球至少覆盖 $\big(1-\delta-\gamma\big)n$ 个点，且半径至多为 $(1+\epsilon)$**半径**(P_{opt})。

回顾定理1之后的注记。只要所选点到**最小包围球**(T)中心的距离大于最优半径的$(1+\epsilon)$倍，就始终能保证获得期望的改进。基于这一观察，我们研究以下方法。假设我们运行定理1中描述的核心集构造过程（我们应该

在下面的分析中将 P 替换为 P_{opt} ）。在第 i步，我们从$P_{\text{opt}} \setminus \mathbb{B}(o_i, (1+\epsilon)$**半径**$(P_{\text{opt}}))$中任意选取一个点，并将其加入 T。我们知道，至多经过$^{2}{}_{(1-s)\epsilon}$步后即可获得$(1+\epsilon)$-近似，即对于某个 $i \leq {}^{2}{}_{(1-s)\epsilon}$，有 $P_{\text{opt}} \subset \mathbb{B}\big(o_i, (1+\epsilon)$**半径**$(P_{\text{opt}})\big)$。

然而，要实现上述方法，我们需要解决两个关键问题：（**i**）如何确定**半径**(P_{opt})的值，以及（**ii**）如何从 $P_{\text{opt}} \setminus \mathbb{B}(o_i, (1+\epsilon)$**半径**$(P_{\text{opt}}))$ 中正确选取一个点。实际上，我们可以将 $(1+\epsilon)$**半径**(P_{opt}) 替换为从 P 中的点到 o_i 的距离中第 t 大的距离，从而隐式地避开第一个问题；其中，我们设定 $t = (\delta+\gamma)n$，以保证得到 $(1+\epsilon,\ 1-\delta)$-近似。对于第二个问题，我们从 P 中距离 o_i 最远的 t 个点中随机选取一个点，并证明该点以一定概率属于 $P_{\text{opt}} \setminus \mathbb{B}(o_i, (1+\epsilon)$**半径**$(P_{\text{opt}}))$。

基于上述思路，本节给出一个次线性时间的 $(1+\epsilon,\ 1-\delta)$-近似算法。为更好地理解该算法，我们先介绍一个线性时间算法（第 5.2.1 节中的算法 4）。请注意，B˘adoiu 等人 [22] 也提出了一个 $(1+\epsilon,\ 1-\delta)$-近似算法，但其复杂度更高；关于运行时间的详细分析，请参见第 5.2.1 节末尾。更重要的是，我们可以改进算法 4 的运行时间，使其达到次线性。为此，我们需要避免计算距离 o_i 最远的 t 个点，因为这一操作需要线性时间。此外，算法 4 会生成一个候选解集合，我们需要从中选出最佳解。这一过程同样需要线性时间。利用第 5.1 节提出的技术，我们可以解决这些问题，并在第 5.2.2 节中设计出一种次线性时间算法，其样本复杂度与 n 和 d 无关。

5.2.1 线性时间算法 本节将介绍一种用于带离群点的最小包围球的线性时间 $(1+\epsilon, 1-\delta)$-近似算法。

算法 4 $(1+\epsilon, 1-\delta)$-带离群点的最小包围球近似算法

输入：一个包含 n 个点的点集 P，其中 $\mathbb{R}^d$ 中离群点的比例为 $\gamma \in (0, 1)$，参数为 $0 < \epsilon, \delta < 1$，$z \in \mathbb{Z}^+$。

1：令 $t = (\delta+\gamma)n$。2：初始时，随机选择一个点 $p \in P$，并令 $T = \{p\}$。3：$i = 1$；重复以下步骤，直到 $i > z$：(1) 将 MEB(T) 的精确中心记为 c_i。按照定理1中的描述，计算近似中心 o_i，使其与 c_i 的距离小于 ξ Rad(T) $= s\,{}_{1+\epsilon}^{\epsilon}$Rad($T$)，其中 s 设为 ${}_{2+\epsilon}^{\epsilon}$。(2) 令 Q 为从 P 到 o_i 距离最远的 t 个点组成的集合；将从 P 到 o_i 的第 $t+1$ 大距离记为 l_i。(3) 随机选择一个点 $q \in Q$，并将其加入 T。(4) $i = i+1$。

4：输出球 $\mathbb{B}(o_{\hat{i}},\ l_{\hat{i}})$，其中 $\hat{i} = \arg_i \min\{l_i \mid 1 \leq i \leq z\}$。

定理 5。 若输入参数为 $z = {}^{2}{}_{(1-s)\epsilon}$ （为方便起见，我们假设它是一个整数），则算法 *4* 以概率 $(1-\gamma)({}_{\gamma+\delta}^{\delta})^z$ 输出带离群点的最小包围球问题的$(1+\epsilon, 1-\delta)$近似解。

在证明定理 5 之前，我们先给出以下两个引理。

引理 9。 以概率 $(1-\gamma)({}_{\gamma+\delta}^{\delta})^z$，在算法 *4* 的步骤 *3* 中运行 z 轮次后，得到的集合为 $T \subset P_{opt}$。

证明。最初，由于 $|P_{\text{opt}}|/|P| = 1-\gamma$，步骤 2 中第一个选中的点以 $1-\gamma$ 的概率属于 P_{opt}。在步骤 3 的 z 轮次中，每一轮选中的点属于 P_{opt}

以 $\frac{\delta}{\gamma+\delta}$ 的概率，因为

$$\frac{|P_{\text{opt}} \cap Q|}{|Q|} = 1 - \frac{|Q \setminus P_{\text{opt}}|}{|Q|} \geq 1 - \frac{|P \setminus P_{\text{opt}}|}{|Q|} = 1 - \frac{\gamma n}{(\delta+\gamma)n} = \frac{\delta}{\delta+\gamma}. \tag{46}$$

因此，以概率 $(1-\gamma)(\frac{\delta}{\gamma+\delta})^z$ 整个集合 $T \subset P_{\text{opt}}$。 □

引理10。 在针对 $1 \leq i \leq z$ 的步骤 *3* 的第 i 轮中，以下两个事件至少有一个发生：*(1)* o_i 是一个 $(1+\epsilon, 1-\delta)$-近似的球心；*(2)* $r_{i+1} > (1+\epsilon)$半径$(P_{opt}) - ||c_i - c_{i+1}|| - \xi r_i$，其中 r_i 是最小包围球(T) 的精确半径，是第 i轮。

证明。 如果 $l_i \leq (1+\epsilon)$**半径**(P_{opt})，则结论成立。也就是说，球 $\mathbb{B}(o_i, l_i)$以半径 $l_i \leq (1+\epsilon)$**半径**(P_{opt})覆盖了 $(1-\delta-\gamma)n$个点（第一个事件发生）。否则， $l_i > (1+\epsilon)$**半径**(P_{opt})，我们考虑第二个事件。设 q 是在第 i轮中添加到 T 的点。利用三角不等式，我们有

$$||o_i - q|| \leq ||o_i - c_i|| + ||c_i - c_{i+1}|| + |c_{i+1} - q|| \leq \xi r_i + ||c_i - c_{i+1}|| + r_{i+1}. \tag{47}$$

由于 $l_i > (1+\epsilon)$**半径**(P_{opt})，且 q 位于 $\mathbb{B}(o_i, l_i)$ 的外部，即， $||o_i - q|| \geq l_i > (1+\epsilon)$**半径**$(P_{\text{opt}})$，由 (47) 可知第二个事件发生，证明完成。□

证明。 **(定理 5 的证明)** 假设引理10中的第一个事件始终没有发生。因此，对于每一对半径 r_{i+1} 和 r_i，我们得到一系列不等式，即， $r_{i+1} > (1+\epsilon)$**半径**$(P_{\text{opt}}) - ||c_i - c_{i+1}|| - \xi r_i$。假设引理 9 中的 $T \subset P_{\text{opt}}$，即，算法每次都能从 P_{opt} 中正确添加一个点到 T。利用与第2.1节中证明定理1几乎完全相同的思路，我们知道，经过至多 z 轮次后，就能得到 P_{opt} 的一个 $(1+\epsilon)$-近似最小包围球。成功概率直接由引理 9 得出。综上，我们得到定理 5。□

定理 5 直接推出以下推论。

推论1. 如果反复运行算法 $4\ O(\frac{1}{1-\gamma}(1+\frac{\gamma}{\delta})^z)$ 次，则该算法以常数概率输出带离群点的最小包围球问题的$(1+\epsilon, 1-\delta)$近似解。

运行时间。 在定理 5 中，我们设定 $z = \frac{2}{(1-s)\epsilon}$ 和 $s \in (0,1)$。为了使 z 较小，根据定理1，我们设定 $s = \frac{\epsilon}{2+\epsilon}$，使得 $z = \frac{2}{\epsilon} + 1$（仅比下界 $\frac{2}{\epsilon}$ 大 1）。对于步骤 3 的每一轮，我们需要计算一个近似中心 o_i，其与精确中心的距离小于 $\xi r_i = s\frac{\epsilon}{1+\epsilon} r_i = O(\epsilon^2) r_i$。使用 [20], 中提出的算法，可在 $O(\frac{1}{\xi^2}|T|d)$ $= O(\frac{1}{\epsilon^5} d)$ 时间内完成。此外，利用 [19] 中的算法，可以在线性时间内得到集合 Q。总之，在推论1中获得 $(1+\epsilon, 1-\delta)$ 近似的时间复杂度为

$$O\big(\frac{C}{\epsilon}(n + \frac{1}{\epsilon^5})d\big), \tag{48}$$

其中 $C = O\left(\frac{1}{1-\gamma}\left(1+\frac{\gamma}{\delta}\right)^{\frac{2}{\epsilon}+1}\right)$。如前所述，Badoiu 等人 [22] 也提出了一种线性时间双准则近似算法。然而，他们算法的运行时间中隐藏的常数关于 $\Theta(\frac{1}{\epsilon\delta})$ 呈指数级增长，远大于 $\frac{2}{\epsilon} + 1$。

5.2.2 运行时间改进 本节中，我们将证明，算法 4 的运行时间还可以进一步改进，使其与点的数量 n 无关。首先，我们注意到，在算法的步骤 3(2) 中，无需计算最远的 t 个点构成的集合 Q。实际上，只要所选点 q 属于步骤 3(3) 中的 $P_{\text{opt}} \cap Q$，就仍能保证得到一个（$1+\epsilon,\ 1-\delta$）近似。第 5.1 节提出的均匀自适应采样过程可以帮助我们在不计算集合 Q 的情况下获得一个点 $q \in P_{\text{opt}} \cap Q$。此外，在引理 8 中，我们证明可以通过随机采样来估计每个候选解的半径。总体而言，我们得到了一种次线性时间算法（算法 5）。按照第 5.2.1 节中的分析，我们设置 $s = \frac{\epsilon}{2+\epsilon}$，使得 $z = \frac{2}{(1-s)\epsilon} = \frac{2}{\epsilon} + 1$。我们在定理 6 和推论 2 中给出结果。与定理 5 相比，定理 6 的成功概率中多了一个 $(1-\eta_1)$ 和 $(1-\eta_2)$ 因子，这是由引理7和引理8中的概率导致的。另一个小问题是，应用引理 8 时，覆盖近似误差从 δ 增大到了 5δ。实际上，只需在参数 n'、t'、n'' 和 t'' 中用 $\delta/5$ 替换 δ，即可轻松解决这一问题，而且渐近复杂度不会改变。

算法 5 带离群点的最小包围球的次线性时间 $(1+\epsilon, 1-\delta)$-近似算法 **输入：** 点集 P，其中包含 n 个位于 $\mathbb{R}^d$ 中的点；离群点比例 $\gamma \in (0,1)$；以及参数 $\epsilon, \eta_1, \eta_2 \in (0,1), \delta \in (0, 1/3\gamma)$ 和 $z \in \mathbb{Z}^+$。**1:** 令 $n' = O(\frac{1}{\delta}\log\frac{1}{\eta_1})$，$n'' = O(\frac{\gamma}{\delta^2}\log\frac{1}{\eta_2})$，$t' = \frac{3}{2}(\delta/5+\gamma)n'$ 和 $t'' = (1+\frac{\delta}{5\gamma})^2\gamma n''$。**2:** 初始时，随机选择一个点 $p \in P$，并令 $T = \{p\}$。**3:** $i = 1$；重复以下步骤，直至 $j = z$：**(1)** 计算 T 的近似最小包围球中心 o_i。**(2)** 从 P 中均匀随机地采样 n' 个点，并令 Q' 为相对于 o_i 距离最远的 t' 个点的集合。**(3)** 随机选择一个点 $q \in Q'$，并将其加入 T。**(4)** 从 P 中均匀随机地采样 n'' 个点，并令 $\tilde{l}_i$ 为从采样点到 o_i 的距离中第 $(t''+1)$ 大的距离。**(5)** $i = i+1$。

4: 输出球 $\mathbb{B}(o_{\hat{i}},\ \tilde{l}_{\hat{i}})$，其中 $\hat{i} = \arg_i \min\{\tilde{l}_i \mid 1 \le i \le z\}$。

定理 6。 如果输入参数为 $z = \frac{2}{\epsilon} + 1$，那么以概率 $(1-\gamma)\big((1-\eta_1)(1-\eta_2)^{\delta/5(\gamma+\delta/5)}\big)^z$，算法 5 将为带离群点的最小包围球问题输出一个$(1+\epsilon, 1-\delta)$-近似解。

为了提高定理 6 中的成功概率，我们需要重复运行算法 5，并输出最佳候选解。然而，设置参数时需要谨慎。定理 6 中的成功概率由两部分组成：$\mathcal{P}_1 = (1-\gamma)\big((1-\eta_1)^{\delta/5(\gamma+\delta/5)}\big)^z$ 和 $\mathcal{P}_2 = (1-\eta_2)^z$，其中 $\mathcal{P}_1$ 表示 $\{o_1, \cdots, o_z\}$ 包含一个合格候选解的概率，而 $\mathcal{P}_2$ 表示引理 8 在所有 z 轮次中成功的概率。因此，如果我们运行算法 5 $N = O(\frac{1}{\mathcal{P}_1})$ 次，那么以常数概率（通过取并集界），生成的所有候选解中至少有一个能给出 $(1+\epsilon,\ 1-\delta)$-近似解；此外，为了保证我们能以常数概率通过夹逼引理正确估计所有候选解所对应的半径，我们需要设置 $\eta_2 = O(\frac{1}{zN})$（因为有 $O(zN)$ 个候选解）。

推论 2. 如果将算法 5 $N = O\Big(\frac{1}{1-\gamma}\big(\frac{1}{1-\eta_1}(3+\frac{3\gamma}{\delta/5})\big)^z\Big)$ 运行多次，并将设置设为 $\eta_2 = O(\frac{1}{zN})$，那么以常数概率，该算法会为带离群点的最小包围球问题输出一个$(1+\epsilon, 1-\delta)$-近似。

运行时间的计算与第 5.2.1 节中的 (48) 类似。我们只需将 n 替换为 $\max\{n', n''\} = O\big(\frac{\gamma}{\delta^2}\log\frac{1}{\eta_2}\big) \equiv O\big(\frac{\gamma}{\delta^2}\log(zN)\big) = \tilde{O}\big(\frac{\gamma}{\delta^2\epsilon}\big)$[3]，并将 C 的值改为 $O\Big(\frac{1}{1-\gamma}\big(\frac{1}{1-\eta_1}(3+\frac{3\gamma}{\delta/5})\big)^{\frac{2}{\epsilon}+1}\Big)$。因此，总运行时间与 n 无关。

5.3 一般最小包围球问题

本节中，我们考虑在没有稳定性假设的情况下求解一般最小包围球问题。令 $0 < \epsilon$、$\delta < 1$ 为两个给定参数。首先，我们将输入 P 视为带离群点最小包围球的一个实例，即 $(P,\ \delta/2)$。然后，我们应用第 5.2 节的算法，得到一个双准则 $(1+\epsilon^2/2,\ 1-\delta/2)$-近似解 $\mathbb{B}(c,\ r_c)$（将“ϵ”替换为 $\epsilon^2/2$，并将“δ”替换为 $\delta/2$）。所得球 $\mathbb{B}(c,\ r_c)$ 至少覆盖 P 中 $(1-\delta/2-\delta/2)n = (1-\delta)n$ 个点，且其半径

$$r_c \le (1+\frac{1}{2}\epsilon^2)\cdot r_{-\delta/2}, \tag{49}$$

其中，$r_{-\delta/2}$ 表示能够覆盖 P 中至少 $(1-\delta/2)n$ 个点的最小球的半径。

其次，我们假设输入 P 是一个满足 $\alpha = \epsilon^2$ 和 $\beta = \delta/2$ 的（$\alpha,\ \beta$）稳定实例；然后运行算法 2，得到一个候选球心 $\tilde{o}$（当然，我们也可以使用算法 1，唯一的区别是样本复杂度会更高）。为了计算由 $\tilde{o}$ 得到的实际半径 $r_{\tilde{o}}$（因为 P 可能并非真正的（$\alpha,\ \beta$）稳定实例），我们只需单遍读取整个数据集 P。最后，根据比率 $r_{\tilde{o}}/r_c$ 确定最终输出。

算法 6 最小包围球的混合近似

输入： 最小包围球问题的一个实例 P，其数据位于 $\mathbb{R}^d$ 中；两个参数 $0 < \epsilon,\ \delta < 1$。1：将输入视为带离群点最小包围球的一个（$P,\ \delta/2$）实例；应用推论 2 中的方法，得到一个关于（$P,\ \delta/2$）的双准则（$1+\epsilon^2/2,\ 1-\delta/2$）近似解 $\mathbb{B}(c,\ r_c)$。2：假设输入 P 是一个满足 $\alpha = \epsilon^2$ 和 $\beta = \delta/2$ 的（$\alpha,\ \beta$）稳定实例；然后运行算法 2，得到一个候选球心 $\tilde{o}$。3：单遍读取整个输入数据集 P，并计算半径 $r_{\tilde{o}} = \max_{p\in P}||\tilde{o}-p||$。4：如果比率 $\frac{r_{\tilde{o}}}{r_c} \le \frac{1+\epsilon}{1-\epsilon^2/2}$，则返回球 $\mathbb{B}(\tilde{o},\ r_{\tilde{o}})$，并称其为“$(1+\epsilon)$ 半径近似”。5：否则，返回球 $\mathbb{B}(c,\ r_c)$，并称其为“$(1-\delta)$ 覆盖近似”。

定理 7. 以常数成功概率，算法 6 返回一个 $(1+\epsilon)$-半径近似或一个 $(1-\delta)$-覆盖近似，其运行时间为 $O\Big(\big(n+h(\epsilon,\delta)\big)\cdot d\Big)$，其中 $h(\epsilon,\delta) = O\big(\frac{1}{1-\delta/2}\,exp(O(1/\epsilon^2))\big)$。该算法只需要对输入数据进行均匀采样和单遍扫描，内存空间复杂度为 $O(h(\epsilon,\delta)\cdot d)$。此外，如果输入数据矩阵（表示输入 P 的 $n\times d$ 矩阵）至多包含 $M \ll nd$ 个非零项，总运行时间将为 $O\big(n+h(\epsilon,\delta)\cdot d+M\big)$。

注 7。 在下面的证明中，我们将看到，当算法返回一个（$1-\delta$）覆盖近似时，返回半径不仅是 $\le$ **半径**(P)，而且至多为 $\big(1-\Theta(\epsilon^2)\big)\cdot$ **半径**(P)（见（52）和（54））。

证明。 我们首先研究时间和空间复杂度。推论 2 中的方法只需要进行均匀采样，而算法 6 的步骤 2 只需对输入数据进行单遍扫描。根据

[3] The asymptotic notation $\tilde{O}(f) = O\big(f\cdot \mathtt{polylog}(\frac{\gamma}{\eta_1\delta(1-\gamma)})\big)$.

推论 2，我们知道空间复杂度为 $O(h(\epsilon,\ \delta)\cdot d)$，其中 $h(\epsilon,\ \delta) = O\big(\frac{1}{1-\delta/2}\exp(O(1/\epsilon^2))\big)$。总运行时间为 $O\Big(\big(n + h(\epsilon,\ \delta)\big)\cdot d\Big)$。此外，我们考虑输入矩阵为稀疏矩阵的情况。在步骤 3 中，我们需要计算值 $r_{\tilde{o}} = \max_{p\in P}||\tilde{o} - p||$。对于每个点 $p \in P$，我们知道

$$||\tilde{o} - p||^2 = ||\tilde{o}||^2 + ||p||^2 - 2\langle\tilde{o}, p\rangle, \tag{50}$$

其中 $\langle\tilde{o}, p\rangle$ 表示它们的内积。$||\tilde{o}||^2$ 的值可在 $O(d)$ 时间内计算得到；若输入数据矩阵至多有 $M \ll nd$ 个非零项，则计算 $\{||p||^2 - 2\langle\tilde{o}, p\rangle \mid p \in P\}$ 这些值的复杂度为 $O(n + M)$。总体而言，算法 6 的复杂度为$O\big(n + h(\epsilon,\ \delta)\cdot d + M\big)$。

现在，我们证明解的质量。令 $\alpha = \epsilon^2$ 和 $\beta = \delta/2$，并考虑以下两种情形。

情形 1： 实例 P 是 $(\alpha,\ \beta)$-稳定的。于是我们直接得到

$$r_{\tilde{o}} \le (1+\epsilon)\cdot\mathbf{Rad}(P). \tag{51}$$

若 $\frac{r_{\tilde{o}}}{r_c} > \frac{1+\epsilon}{1-\epsilon^2/2}$，结合 (51)，可得

$$r_c < \Big(1 - \epsilon^2/2\Big)\cdot\mathbf{Rad}(P). \tag{52}$$

那么我们可以返回球 $\mathbb{B}(c,\ r_c)$，并称其为“$(1-\delta)$-覆盖近似”。另一方面，当 $\frac{r_{\tilde{o}}}{r_c} \le \frac{1+\epsilon}{1-\epsilon^2/2}$ 时，根据 (51)，我们可以返回球 $\mathbb{B}(\tilde{o},\ r_{\tilde{o}})$，并称其为“$(1+\epsilon)$-半径近似”。

情况 2： P 不是一个 $(\alpha,\ \beta)$-稳定实例。那么，根据稳定性的定义，我们知道实例 $(P,\ \delta/2)$ 的最优半径不大于

$$(1-\epsilon^2)\cdot\mathbf{Rad}(P). \tag{53}$$

因此我们有

$$r_c < (1+\frac{1}{2}\epsilon^2)(1-\epsilon^2)\cdot\mathbf{Rad}(P) < \Big(1-\epsilon^2/2\Big)\cdot\mathbf{Rad}(P). \tag{54}$$

如果 $\frac{r_{\tilde{o}}}{r_c} \le \frac{1+\epsilon}{1-\epsilon^2/2}$，结合 (54)，则可得

$$r_{\tilde{o}} < (1+\epsilon)\cdot\mathbf{Rad}(P). \tag{55}$$

那么我们可以返回球 $\mathbb{B}(\tilde{o},\ r_{\tilde{o}})$，并称其为“$(1+\epsilon)$-半径近似”。另一方面，当 $\frac{r_{\tilde{o}}}{r_c} > \frac{1+\epsilon}{1-\epsilon^2/2}$ 时，根据 (54)，我们可以返回球 $\mathbb{B}(c,\ r_c)$，并称其为“$(1-\delta)$-覆盖近似”。

由于第 5.2 节的方法的成功概率为常数，算法 7 的整体成功概率也为常数。□

关于算法 6 结果的进一步分析。我们进一步考虑一个“逆向”问题：能否根据算法 6 的输出推断给定实例 P 的稳定度？在步骤 2 中，我们假设 P 是一个$(\epsilon^2, \delta/2)$-稳定实例，但实际情况未必如此。回顾定义 3 中“(α,β)-稳定”的定义。我们知道，始终存在一个值 $\hat{}\alpha \in [0,1)$，使得 P是一个$(\hat{}\alpha, \delta/2)$-稳定实例。对于固定的“$\delta/2$”，我们可以用“$\hat{}\alpha$”表示 P的稳定度。下面的定理表明，我们可以通过算法 6 推断$\hat{}\alpha$ 的值。

定理 8。如果算法 *6* 返回一个$(1+\epsilon)$-半径近似，那么$\hat{\alpha}<\epsilon$；否则，算法返回一个$(1-\delta)$-覆盖近似，这意味着$\hat{\alpha}>\frac{\epsilon^2}{2}$。

换言之，当$\hat{\alpha}\geq\epsilon$时，算法能够区分这种情况（它必须返回$(1-\delta)$-覆盖近似）与$\hat{\alpha}\leq\frac{\epsilon^2}{2}$这种情况（它必须返回$(1+\epsilon)$-半径近似）；但若$\frac{\epsilon^2}{2}<\hat{\alpha}<\epsilon$，算法可以返回$(1-\delta)$-覆盖近似，也可以返回$(1+\epsilon)$-半径近似。

证明。回顾一下，我们在算法 6 中设定了 $\alpha=\epsilon^2$ 和 $\beta=\delta/2$。首先，假设输出是一个$(1+\epsilon)$-半径近似。一种可能情况是，实例 P 是一个真实的(α,β)-稳定实例，此时$\hat{\alpha}=\alpha<\epsilon$。另一种可能情况是，$P$ 不是(α,β)-稳定实例，但比值为$\frac{r_{\tilde{\delta}}}{r_c}\leq\frac{1+\epsilon}{1-\epsilon^2/2}$结合（49），我们得到

$$\frac{\mathbf{Rad}(P)}{r_{-\delta/2}}\leq\frac{r_{\tilde{\delta}}}{\frac{1}{1+\epsilon^2/2}r_c}\leq\frac{(1+\epsilon)(1+\epsilon^2/2)}{1-\epsilon^2/2}. \tag{56}$$

因此$\hat{\alpha}=1-\frac{r_{-\delta/2}}{\text{半径}(P)}\leq 1-\frac{1-\epsilon^2/2}{(1+\epsilon)(1+\epsilon^2/2)}<\epsilon$。总的来说，只要输出是一个 $(1+\epsilon)$-半径近似，$\hat{\alpha}$ 就应小于 ϵ。

接下来假设输出是一个 $(1-\delta)$-覆盖近似。一种可能的情况是，实例 P 不是 (α,β)-稳定的，此时 $\hat{\alpha}>\alpha=\epsilon^2$。另一种可能的情况是，$P$ 是 (α,β)-稳定的，但比率 $\frac{r_{\tilde{\delta}}}{r_c}>\frac{1+\epsilon}{1-\epsilon^2/2}$ 结合 (51)，我们有

$$\frac{\mathbf{Rad}(P)}{r_{-\delta/2}}\geq\frac{\frac{1}{1+\epsilon}r_{\tilde{\delta}}}{r_c}>\frac{1}{1-\epsilon^2/2}. \tag{57}$$

因此$\hat{\alpha}=1-\frac{r_{-\delta/2}}{\text{半径}(P)}>1-(1-\epsilon^2/2)=\epsilon^2/2$。总的来说，只要输出是一个 $(1-\delta)$-覆盖近似，$\hat{\alpha}>\min\{\epsilon^2,\epsilon^2/2\}=\frac{\epsilon^2}{2}$。 □

6 扩展 I：带离群点的最小包围球的混合近似

本节将第 5.3 节的思路扩展，提出一种带离群点的最小包围球问题的混合近似算法（P，γ）。首先，我们将最小包围球的定义 3 扩展到带离群点的最小包围球。

定义 *6*（带离群点的最小包围球的（α,β）-稳定）。令 $0<\alpha,\beta<1$。给定定义 *4* 中带离群点的最小包围球问题的一个实例（P,γ），若（P,γ）满足以下条件，则称其为一个（α,β）-稳定实例：（*1*）对于任意满足 $|Q|<(\gamma+\beta)n$ 的 $Q\subset P$，$\mathit{Rad}(P\setminus Q)>(1-\alpha)\mathit{Rad}(P_{\text{opt}})$；（*2*）存在一个 $Q'\subset P$，满足 $|Q'|=\lceil(\beta+\gamma)n\rceil$，且 $\mathit{Rad}(P\setminus Q')\leq(1-\alpha)\mathit{Rad}(P_{\text{opt}})$。

定义 6 直接推出以下命题。

命题 2. 如果 (P,γ) 是带离群点的最小包围球问题的 (α,β)-稳定实例，那么相应的 P_{opt} 是带 $\tilde{\beta}\geq\frac{\beta}{1-\gamma}$ 的最小包围球的 $(\alpha,\tilde{\beta})$-稳定实例。

注意，定义 6 隐含地要求 $\beta<1-\gamma$。因此，它意味着命题 2 中 $\tilde{\beta}$ 的下界 $\frac{\beta}{1-\gamma}$ 不能大于 1。为证明命题 2 的正确性，我们可以使用矛盾法。假设存在一个子集 $P'\subset P_{\text{opt}}$，使得 $|P'|>(1-\frac{\beta}{1-\gamma})|P_{\text{opt}}|=(1-\gamma-\beta)n$，且 $\text{半径}(P')\leq(1-\alpha)\text{半径}(P_{\text{opt}})$。那么，这就与 (P,γ) 是带离群点的最小包围球问题的 (α,β)- 稳定实例这一事实相矛盾。

要应用第 5.3 节的思路，一个重大挑战在于集合 P_{opt} 与离群点混杂在一起，因此我们无法像算法 6 那样轻易获得 $(1+\epsilon)$-半径近似。我们的起点仍然是此前提出的次线性时间双准则近似算法，见

第 5.2 节。具体而言，给定任意两个小参数 $0 < \epsilon$、 $\delta < 1$，该算法通过均匀自适应采样过程返回一组候选球心。我们用 Ξ 表示该集合。以常数概率， Ξ 中至少有一个候选项（记为 s）满足以下不等式：

$$\left|\mathbb{B}\big(s, (1+\epsilon)\cdot \mathbf{Rad}(P_{\mathrm{opt}})\big) \cap P\right| \geq \big(1-\delta-\gamma\big)n. \tag{58}$$

也就是说，它是一个“$(1+\epsilon,\ 1-\delta)$-近似解”。为了选出这样的合格候选解，可以使用基于均匀采样的技术“夹逼引理”来估计 $\mathbb{B}\big(s,\ (1+\epsilon)\cdot$ **半径**$(P_{\mathrm{opt}})\big) \cap P$ 的大小，而无需读取整个数据集 P。值得注意的是，第 5.2 节定理 5 中包含一个隐含事实。实际上，证明表明，在候选集 Ξ 中，存在一个解 s，使得球 $\mathbb{B}\big(s,\ (1+\epsilon)\cdot$ **半径**$(P_{\mathrm{opt}})\big)$ 至少覆盖来自 P_{opt} 的 $\big(1-\delta-\gamma\big)n$ 个点（因为集合 $T \subset P_{\mathrm{opt}}$ 和解 s 是由 T 生成的（见引理 9））。因此，解 s 应满足

$$\left|\mathbb{B}\big(s, (1+\epsilon)\cdot \mathbf{Rad}(P_{\mathrm{opt}})\big) \cap P_{\mathrm{opt}}\right| \geq \big(1-\delta-\gamma\big)n, \tag{59}$$

这比 (58) 更强。但夹逼引理可能会忽略这样一个更强的解，因为只需选取满足 (58) 的解，就足以保证一个 $(1+\epsilon,\ 1-\delta)$-近似。基于这一观察，我们为带离群点的最小包围球提出以下新算法。

混合近似算法。设 ϵ 和 δ 是给定的两个参数。首先，我们应用第 5.2 节的方法。但我们不会将参数对 $(\epsilon,\ \delta)$ 直接输入双准则近似算法；相反，我们使用 $\big(\frac{1}{2(2\sqrt{2}+\sqrt{3})_2}\epsilon^2,\ \delta\big)$（我们将在分析中解释为何会出现系数“$\frac{1}{2(2\sqrt{2}+\sqrt{3})^2}$”）。也就是说，我们通过第 5.2 节的均匀自适应采样，计算出候选球心集合 Ξ，其中至少有一个中心能为实例 $(P,\ \gamma)$ 产生 $\big(1+\frac{1}{(2\sqrt{2}+\sqrt{3})_2}\epsilon^2,\ 1-\delta\big)$-近似。然后，对每个候选项 $q \in \Xi$，我们定义两个值：

$$r_q = \min\Big\{r > 0 \mid \big|\mathbb{B}(q,r) \cap P\big| \geq (1-\gamma)n\Big\}; \tag{60}$$

$$r'_q = \min\Big\{r > 0 \mid \big|\mathbb{B}(q,r) \cap P\big| \geq \big(1-\delta-\gamma\big)n\Big\}. \tag{61}$$

我们可以通过单遍扫描输入 P，为 Ξ 的所有候选项计算这两个值（而不使用夹逼引理）。我们选取两个点 $s_1 = \arg\min_{q\in\Xi} \frac{r_q}{1+\epsilon}$ 和 $s_2 = \arg\min_{q\in\Xi} r'_q$（它们可能是同一个点，也可能不是）。如果比值 $\frac{r_{s_1}}{r'_{s_2}} \leq \frac{1-\epsilon^2}{\big(2 - \frac{1}{(2\sqrt{2}+\sqrt{3})^2}\big)}$，则返回球 $\mathbb{B}(s_1,\ r_{s_1})$，并称其为“$(1+\epsilon)$-半径近似”；否则，返回球 $\mathbb{B}(s_2,\ r'_{s_2})$，并称其为“$(1-\delta)$-覆盖近似”。

算法7：带离群点的最小包围球混合近似

输入： 一个实例 $(P,\ \gamma)$，属于 $\mathbb{R}^d$ 中的带离群点的最小包围球问题；两个参数 $0 < \epsilon, \delta < 1$。

1：应用第 5.2 节的均匀自适应采样方法，得到候选球心集合 Ξ，其中至少有一个中心能为实例 $(P,\ \gamma)$ 提供 $\big(1+\frac{1}{(2\sqrt{2}+\sqrt{3})_2}\epsilon^2,\ 1-\delta\big)$-近似。2：单遍读取整个输入数据集 P，并针对每个 $q \in \Xi$，按照公式 (60) 和 (61) 计算数值 r_q 和 r'_q。3：令 $s_1 =$ 为 $\arg\min_{q\in\Xi} r_q$， $s_2 =$ 为 $\arg\min_{q\in\Xi} r'_q$。4：若比值 $\frac{r_{s_1}}{r'_{s_2}} \leq \frac{1-\epsilon^2}{\big(2 - \frac{1}{(2\sqrt{2}+\sqrt{3})^2}\big)}$，则返回球 $\mathbb{B}(s_1,\ r_{s_1})$，并称其为「$(1+\epsilon)$-半径近似」。5：否则，返回球 $\mathbb{B}(s_2,\ r'_{s_2})$，并称其为「$(1-\delta)$-覆盖近似」。

定理 9. 以常数成功概率，算法 7 返回半径近似 $(1+\epsilon)$ 或覆盖近似 $(1-\delta)$，运行时间为 $O(g(\epsilon,\delta,\gamma)\cdot nd)$，其中 $g(\epsilon,\delta,\gamma)=O(\frac{1}{1-\gamma}(\frac{\gamma+\delta}{\delta})^{O(1/\epsilon^2)})$。该算法只需进行均匀采样，并对输入数据进行单遍处理；内存中的空间复杂度为 $O(g(\epsilon,\delta,\gamma)\cdot d)$。此外，如果输入数据矩阵（表示输入 P 的 $n\times d$ 矩阵）至多包含 $M\ll nd$ 个非零项，则总运行时间为 $O\big(g(\epsilon,\delta,\gamma)\cdot(n+d+M)\big)$。

注 8. 与定理 7 类似，我们将看到，当算法返回一个 $(1-\delta)$-覆盖近似时，返回半径至多为 $\big(1-\Theta(\epsilon^2)\big)\cdot$**半径**$(P_{\text{opt}})$（见 (63) 和 (64)）。

证明。（定理 9 的证明）我们首先研究时间和空间复杂度。推论 2 中的方法只需要均匀采样，而算法 7 的步骤 2 只需对输入数据进行单遍扫描。根据推论 2，Ξ 的大小为 $g(\epsilon,\delta,\gamma)=O(\frac{1}{1-\gamma}(\frac{\gamma+\delta}{\delta})^{O(1/\epsilon^2)})$。总体而言，空间复杂度为 $O(g(\epsilon,\delta,\gamma)\cdot d)$。生成 Ξ 的复杂度为 $O\big(|\Xi|\cdot \text{poly}(\frac{1}{\epsilon},\frac{1}{\delta})d\big)$，相对于输入规模 nd 而言是次线性的。显然，步骤 2 的复杂度主导了整体复杂度。因此，总运行时间为 $O(g(\epsilon,\delta,\gamma)\cdot nd)$。此外，我们考虑输入矩阵为稀疏矩阵的情况。与定理 7 的证明类似，我们知道，如果输入数据矩阵的非零项至多为 $M\ll nd$，则算法 7 的复杂度为 $O\big(g(\epsilon,\delta,\gamma)\cdot(n+d+M)\big)$。

现在，我们证明解的质量。令 $\alpha=\frac{1}{(2\sqrt{2}+\sqrt{3})^2}\epsilon^2$ 和 $\beta=(1-\gamma)\delta$，并考虑以下两种情形。

情形 1: 实例 (P,γ) 是 (α,β)-稳定的（即 P_{opt} 是一个 MEB 的 (α,β)-稳定实例，且满足 $\beta\geq\delta$；根据命题 2）。记 o 为 $\mathbf{MEB}(P_{\text{opt}})$ 的最优中心。我们假设 Ξ 的一个候选球心 q_0 满足公式 (59)。因此，根据定理 2，我们知道 $||q_0-o||\leq(2\sqrt{2}+\sqrt{3})\sqrt{\alpha}\cdot$**半径**$(P_{\text{opt}})=\epsilon\cdot$**半径**$(P_{\text{opt}})$。也就是说，

$$r_{s_1}\leq r_{q_0}\leq(1+\epsilon)\cdot\mathbf{Rad}(P_{\text{opt}}). \tag{62}$$

若 $\frac{r_{s_1}}{r'_{s_2}}>\frac{1+\epsilon}{1-\epsilon^2/\big(2(2\sqrt{2}+\sqrt{3})^2\big)}$，结合 (62)，我们得到

$$r'_{s_2}<\Big(1-\epsilon^2/\big(2(2\sqrt{2}+\sqrt{3})^2\big)\Big)\cdot\mathbf{Rad}(P_{\text{opt}}). \tag{63}$$

然后我们可以返回球 $\mathbb{B}(s_2,\ r'_{s_2})$，并称“它是一个 $(1-\delta)$-覆盖近似”。另一方面，当 $\frac{r_{s_1}}{r'_{s_2}}\leq\frac{1+\epsilon}{1-\epsilon^2/\big(2(2\sqrt{2}+\sqrt{3})^2\big)}$ 时，根据 (62)，我们可以返回球 $\mathbb{B}(s_1,\ r_{s_1})$，并称“它是一个 $(1+\epsilon)$-半径近似”。

情况 2： $(P,\ \gamma)$ 不是一个 $(\alpha,\ \beta)$-稳定实例。因此可得

$$\begin{aligned}r'_{s_2}&<(1+\frac{1}{2(2\sqrt{2}+\sqrt{3})^2}\epsilon^2)(1-\frac{1}{(2\sqrt{2}+\sqrt{3})^2}\epsilon^2)\cdot\mathbf{Rad}(P_{\text{opt}})\\&<\Big(1-\epsilon^2/\big(2(2\sqrt{2}+\sqrt{3})^2\big)\Big)\cdot\mathbf{Rad}(P_{\text{opt}}).\end{aligned} \tag{64}$$

如果 $\frac{r_{s_1}}{r'_{s_2}}\leq\frac{1+\epsilon}{1-\epsilon^2/\big(2(2\sqrt{2}+\sqrt{3})^2\big)}$，结合 (64)，则可得

$$r_{s_1}<(1+\epsilon)\cdot\mathbf{Rad}(P_{\text{opt}}). \tag{65}$$

然后我们可以返回球 $\mathbb{B}(s_1,\ r_{s_1})$，并称“它是一个 $(1+\epsilon)$-半径近似”。另一方面，当 $\frac{r_{s_1}}{r'_{s_2}}>\frac{1+\epsilon}{1-\epsilon^2/\big(2(2\sqrt{2}+\sqrt{3})^2\big)}$ 时，根据 (64)，我们可以返回球 $\mathbb{B}(s_2,\ r'_{s_2})$，并称“它是一个 $(1-\delta)$-覆盖近似”。

由于第 5.2 节方法的成功概率是常数，因此算法 7 的总体成功概率也是常数。□

我们还得到了以下关于推断实例（P，γ）稳定性的定理，其证明与定理 8 的证明几乎相同。

定理 10. 假设(P,γ)是一个$(\hat{}\alpha, (1-\gamma)\delta)$-稳定实例。如果算法 7 返回一个$(1+\epsilon)$-半径近似，则$\hat{}\alpha < \epsilon$；否则，该算法返回一个$(1-\delta)$-覆盖近似，并由此推出$\hat{}\alpha > {}_{2}{}^{\epsilon^2}{}_{(2\sqrt{2}+\sqrt{3})^2}$。

7 扩展 II：带离群点的 MEX 双准则近似

在本节中，我们将定义 4 扩展到带离群点的最小包围球，并定义一个更一般的问题，称为**带离群点的最小包围"x"（MEX）**。接着我们将展示，只要形状"x"满足若干性质，引理 7 和 8 的思想就可以推广到带离群点的 MEX 问题。**为了描述形状"x"，我们需要明确三个基本概念：中心、大小和距离函数**。

令 $\mathcal{X}$ 为 $\mathbb{R}^d$ 中指定形状的集合。我们要求每个形状 $x \in \mathcal{X}$ 都由以下两个组成部分唯一确定："$c(x)$"，即 x 的中心；以及"$s(x) \geq 0$"，即 x 的大小。对于任意两个形状 x_1、$x_2 \in \mathcal{X}$，当且仅当 $c(x_1) = c(x_2)$ 且 $s(x_1) = s(x_2)$ 时，$x_1 = x_2$。此外，给定一个中心 o_0 和一个值 $l_0 \geq 0$，我们用 $x(o_0, l_0)$ 表示满足 $c(x)$ 为 $= o_0$ 且 $s(x)$ 为 $= l_0$ 的形状 x。对于不同的形状，中心和大小的定义也各不相同。例如，若 x 是一个球，则 $c(x)$ 和 $s(x)$ 分别应为球心和半径；给定 $o_0 \in \mathbb{R}^d$ 和 $l_0 \geq 0$，$x(o_0, l_0)$ 应为球 $\mathbb{B}(o_0, l_0)$。再举一个更复杂的例子，考虑带离群点的 k-中心聚类问题：该问题要求找到 k 个球，在排除一定数量离群点的情况下覆盖输入点集，并使最大半径最小（不失一般性，我们可以假设这 k 个球的半径相同）。对于该问题，形状"x"是 $\mathbb{R}^d$ 中 k 个球的并集；中心 $c(x)$ 是这 k 个球心的集合，大小 $s(x)$ 则是半径。

对于任意点 $p \in \mathbb{R}^d$ 和任意形状 $x \in \mathcal{X}$，我们还需要定义中心 $c(x)$ 与 p之间的**距离函数** $f(c(x), p)$。例如，如果 x是一个球，$f(c(x), p)$ 就等于 $||p - c(x)||$；如果 x是以中心 $c(x) = \{c_1, c_2, \cdots, c_k\}$为中心的 k个球的并集，距离应为 $\min_{1 \leq j \leq k} ||p - c_j||$。请注意，距离函数仅用于对点进行排序，以便$c(x)$，并不一定非负（例如，在第7.3节中，我们为支持向量机定义了距离函数$f(c(x), p) \leq 0$）。利用这个距离函数，我们可以在推广下文的引理 7 和 8 时定义集合"Q"和值"l_i"。为保证它们的正确性，我们还要求 $\mathcal{X}$ 满足以下三个性质。

性质1。 对于任意两个形状 $x_1 \neq x_2 \in \mathcal{X}$，如果 $c(x_1) = c(x_2)$，那么

$$s(x_1) \leq s(x_2) \iff x_1 \text{ is covered by } x_2, \tag{66}$$

其中，"x_1 被 x_2覆盖"是指"对于任意点 $p \in \mathbb{R}^d$，$p \in x_1 \Rightarrow p \in x_2$"。

性质2。 给定任意形状 $x \in \mathcal{X}$和任意点 $p_0 \in x$，集合

$$\{p \mid p \in \mathbb{R}^d \text{ and } f(c(x), p) \leq f(c(x), p_0)\} \subseteq x. \tag{67}$$

性质3。 给定任意形状中心 o_0 和任意点 $p_0 \in \mathbb{R}^d$，令 $r_0 = \min\{r \mid r \geq 0, p_0 \in x(o_0, r)\}$。那么，对于任意 $r < r_0$，$p_0 \in x(o_0, r_0)$ 以及 $p_0 \notin x(o_0, r)$。（注意：通常，值 r_0 就是从 p_0 到形状中心 o_0的距离；但在某些情况下，例如第7.3节中的支持向量机问题，形状大小和距离函数具有不同的含义）。

直观而言，性质1表明 $s(x)$ 为共享同一中心 $c(x)$ 的形状定义了一种顺序。性质2表明，距离函数 f 为点定义了一种顺序，使得

给定形状中心 $c(x)$。性质3表明，一个中心 o_0 和一个点 p_0 可以定义一个恰好与 p_0“相切”的形状。我们可以以 $\mathcal{X}=\{$所有 d维球$\}$ 为例。对于任意两个同心球，较小的球总是被较大的球覆盖（性质1）；如果一个点p_0 位于球 x内，那么任何距离 $||p-c(x)|| \leq ||p_0-c(x)||$ 的点 p 也应位于 x 内（性质2）；此外，给定一个球心 o_0和一个点 p_0，对于任意 $r<||p_0-o_0||$， $p_0 \in \mathbb{B}(o_0,||p_0-o_0||)$ 和$p_0 \notin \mathbb{B}(o_0,r)$（性质3）。

现在，我们正式定义带离群点的MEX问题及其双准则近似。

定义 7（带离群点的 MEX）。假设形状集 $\mathcal{X}$ 满足性质*1*、*2*和*3*。给定由 P 个点组成的集合 n，这些点位于 $\mathbb{R}^d$ 中，且给定一个小参数 $\gamma \in (0,1)$，带离群点的*MEX*问题是寻找能够覆盖$(1-\gamma)n$ 个点的最小形状 $x \in \mathcal{X}$。也就是说，任务是寻找 P 的一个子集，其大小为$(1-\gamma)n$，使得该子集的最小包围形状 $\mathcal{X}$ 在所有可能的子集选择中最小。所得解记为 $\mathbf{MEX}(P,\gamma)$。

定义8（双准则近似）。给定带离群点的*MEX*问题的一个实例(P,γ) 以及两个小参数 $0<\epsilon$, $\delta<1$，$(1+\epsilon,1-\delta)$-近似是(P,γ) 的一个解$x \in \mathcal{X}$，该解至少覆盖 $\big(1-\delta-\gamma\big)n$ 个点，且其大小至多为$(1+\epsilon)s(x_{opt})$，其中x_{opt} 为最优解。

不难看出，带离群点的最小包围球的定义 4 实际上是定义 7 的特例。与带离群点的最小包围球类似，我们仍使用 P_{opt}，其中 $P_{\text{opt}} \subset P$和 $|P_{\text{opt}}|=(1-\gamma)n$，表示带离群点的 MEX 最优解所覆盖的子集。

现在，我们给出引理 7 和引理 8 的推广版本。与第 2.1 节中的核心集构造方法类似，我们假设存在一个迭代算法 Γ，用于计算 MEX（不含离群点）；实际上，这是在我们的框架下设计次线性时间算法的重要前提（在接下来的小节中，我们将讨论用于带离群点的 MEX 问题的迭代算法，包括平坦拟合、 k-中心聚类和支持向量机）。在 Γ 的第 i次迭代中，它维护一个形状中心 o_i。此外，令 Q 为从 P 到 o_i、关于距离函数 f 的 $(\delta+\gamma)n$ 个最远点的集合。首先，我们需要在下面的命题中将值“l_i”定义为 Q 。

断言 3. 存在一个值 $l_i \geq 0$，满足 $P \setminus x(o_i,\ l_i) = Q$。

证明。可以根据 P 中各点到 o_i 的距离对这些点进行排序。不失一般性，设 $P=\{p_1,p_2,\cdots,p_n\}$，且 $f(o_i,p_1)>f(o_i,p_2)>\cdots>f(o_i,p_n)$（为方便起见，我们假设任意两个距离都不相等；如果出现平局，可以任意确定它们的顺序，使其满足 o_i）。于是集合 $Q=\{p_j \mid 1 \leq j \leq (\delta+\gamma)n\}$。此外，根据性质3，我们知道每个点 $p_j \in P$ 都对应一个值 r_j，满足 $p_j \in x(o_i,r_j)$ 且对于任意 $r<r_j$，有 $p_j \notin x(o_i,r)$。用 x_j 表示形状 $x(o_i,r_j)$。我们选择点 p_{j_0}，使得 $j_0=(\delta+\gamma)n+1$。根据性质2，我们知道对于任意 $j \geq j_0$，都有 $p_j \in x_{j_0}$，即 (a) $P \setminus Q \subseteq x_{j_0}$。我们还需证明，对于任意 $j<j_0$，都有 $p_j \notin x_{j_0}$。假设存在某个 $p_{j_1} \in x_{j_0}$，使得 $j_1<j_0$。于是有 $r_{j_1}<r_{j_0}$，因此 $p_{j_0} \notin x_{j_1}$（根据性质3）。根据性质2， $p_{j_0} \notin x_{j_1}$ 意味着 $f(o_i,p_{j_0})>f(o_i,p_{j_1})$，这与事实 $f(o_i,p_{j_0})<f(o_i,p_{j_1})$ 矛盾。因此有 (b) $Q \cap x_{j_0}=\emptyset$。

上述（**a**）和（**b**）意味着 $\{P \cap x_{j_0}, Q\}$ 是 P的一个划分，即，$(P \cap x_{j_0}) \cup Q = P$且 $(P \cap x_{j_0}) \cap Q = \emptyset$。因此我们知道 $P \setminus x_{j_0}=Q$。所以，我们可以将值设为 $l_i=r_{j_0}$，然后 $P \setminus x(o_i,\ l_i)=Q$。□

引理 11（广义均匀自适应采样）。设 $\eta_1 \in (0,1)$。若从 P 中独立且均匀随机地采样 $n'=O(\frac{1}{\delta}\log\frac{1}{\eta_1})$ 个点，并令 Q' 为相对于样本距离 o_i 最远的 $3\,2(\delta+\gamma)n'$ 个点构成的集合，则以至少 $1-\eta_1^-$ 的概率，

以下结论成立

$$\frac{\left|Q' \cap \left(P_{opt} \cap Q\right)\right|}{|Q'|} \geq \frac{\delta}{3(\gamma+\delta)}. \tag{68}$$

证明。令 A 表示从 P 中抽样得到的 n' 个点的集合。与 (28) 类似，我们有 ve

$$\left|A \cap \left(P_{\text{opt}} \cap Q\right)\right| > \frac{1}{2}\delta n' \quad \text{and} \quad \left|A \cap Q\right| < \frac{3}{2}(\delta+\gamma)n' \tag{69}$$

以概率 $1-\eta_1$。与 (29) 类似，我们有

$$A \cap Q = \{p \in A \mid f(o_i, p) > f(o_i, p_{j_0})\}, \tag{70}$$

其中，p_{j_0} 是断言 3 的证明中选取的点。按照断言 3 中相同的方法，我们也可以选取一个满足下式的点 $p_{j_0'} \in A$

$$Q' = \{p \in A \mid f(o_i, p) > f(o_i, p_{j_0'})\}. \tag{71}$$

接下来，我们可以证明

$$\Big(A \cap \left(P_{\text{opt}} \cap Q\right)\Big) = \Big(Q' \cap \left(P_{\text{opt}} \cap Q\right)\Big). \tag{72}$$

方法与 (33) 相同。因此，

$$\frac{\left|Q' \cap \left(P_{\text{opt}} \cap Q\right)\right|}{|Q'|} = \frac{\left|A \cap \left(P_{\text{opt}} \cap Q\right)\right|}{|Q'|} \geq \frac{\delta}{3(\gamma+\delta)}, \tag{73}$$

最后一个不等式由 (69) 中的第一个不等式以及事实 $|Q'| = \frac{3}{2}(\delta+\gamma)n'$ 得出。 –

□

引理 12（广义夹逼引理）。令 $\eta_2 \in (0,\ 1)$，并假设 $\delta < \gamma/3$。l_i 是来自断言 *3* 的值。我们从 P 中独立且均匀随机地采样 $n'' = O\big(\frac{\gamma}{\delta^2}\log\frac{1}{\eta_2}\big)$ 个点，并令 q 为从采样点到 o_i 中距离最远的第 $\big((1+\delta/\gamma)^2\gamma n''+1\big)$ 个点。如果 $l_i = \min\{r \mid r \geq 0, q \in x(o_i,\ r)\}$（类似于性质 *3*中定义“$r_0$”的方式），那么，以概率 $1-\eta_2$，以下结论成立

$$\tilde{l}_i \leq l_i; \tag{74}$$

$$\left|P \setminus x(o_i, \tilde{l}_i)\right| \leq (\gamma+5\delta)n. \tag{75}$$

证明。令 B 表示来自 P 的采样 n'' 点集合。通过采用与断言 3 相同的方法，我们知道存在一个值 $\tilde{l}_i' > 0$，满足 $\left|P \setminus x(o_i,\ \tilde{l}_i')\right| = \frac{(\gamma+\delta)^2}{\gamma-\delta}\gamma n$。类似于 引理 8 的证明，我们可以证明 $l_i \in [l_i', l_i]$。根据性质 1，我们知道 $x(o_i,\ l_i)$ 被两个形状 $x(o_i,\ l_i')$ 和 $x(o_i,\ l_i)$“夹住”。此外，由于 $x(o_i,\ l_i')$ 被 $x(o_i,\ l_i)$ 覆盖，因此有

$$\left|P \setminus x(o_i, \tilde{l}_i)\right| \leq \left|P \setminus x(o_i, \tilde{l}_i')\right| = \frac{(\gamma+\delta)^2}{\gamma-\delta}\gamma n = (\gamma+5\delta)n, \tag{76}$$

其中，最后一个等式由假设 $\delta < \gamma/3$ 得出。因此，(74) 和 (75) 成立。 □

利用引理 11 和引理 12，我们在以下小节中研究若干应用。

7.1 k-中心离群点聚类

设 $\gamma \in (0, 1)$。给定 $\mathbb{R}^d$ 中由 n 个点构成的集合 P，带离群点的 k-**中心聚类** 问题是寻找 k 个球来覆盖 $(1-\gamma)n$ 个点，并使这些球的最大半径最小化（不失一般性，我们可以假设 k 个球的半径相同）。给定一个实例 (P, γ)，令 $\{C_1, \cdots, C_k\}$ 为构成 P_{opt} 的 k 个簇（即产生最优解的 P 子集），并令 r_{opt} 为最优半径；也就是说，每个 C_j 都由一个半径为 r_{opt} 的独立球覆盖。与第 5.2 节类似，我们先介绍一种线性时间算法，然后说明如何利用引理 11 和引理 12 将其修改为次线性时间算法。

线性时间算法。若假设 k 为常数，则第 5.2.1 节中的算法可推广为一种线性时间双标准算法，用于解决带离群点的 k-中心聚类问题。我们的思路如下。在算法 4 中，我们维护集合 T 作为 P_{opt}的核心集；这里，我们改为维护 k 个集合 T_1、T_2、$\cdots$、T_k，分别作为 C_1、C_2、$\cdots$、C_k的核心集。因此，对于 $1 \le j \le k$ 的每个 T_j，在步骤 3 的第i轮中都有一个近似最小包围球中心 o_i^j，并令 $O_i = \{o_i^1, \cdots, o_i^k\}$。初始时，$O_0$ 以及 $1 \le j \le k$ 对应的 T_j 均为空；我们随机选取一个点 $p \in P$，并以概率 $1-\gamma$令 $p \in P_{\text{opt}}$（不失一般性，假设 $p \in C_1$ 并将其加入 T_1；因此，此步骤之后 $O_1 = \{p\}$）。令 Q 为到 O_i距离最远的$t =$（$\delta + \gamma$）n 个点，并令 l_i 为从 P到 O_i的距离中第（$t+1$）大的距离（点 $p \in P$到 O_i的距离为 $\min_{1 \le j \le k} ||p - o_i^j||$）。然后，我们随机选取一个点 $q \in Q$，并以概率 $\frac{\delta}{\gamma+\delta}$令 $q \in P_{\text{opt}}$（如引理 9 中的式（46））。为便于说明，我们假设 $q \in P_{\text{opt}}$ 发生，并且有一个“预言机”可以猜出 q 属于哪个最优簇，例如 $q \in C_{j_q}$；接着，我们将 q加入 T_{j_q}，并更新 T_{j_q}的近似最小包围球中心。由于对于 $1 \le j \le k$，每个最优簇 C_j都有一个大小为 $\frac{2}{\epsilon} + 1$的核心集（在定理1中令 $s = \frac{\epsilon}{2+\epsilon}$），因此，至多添加 k（$\frac{2}{\epsilon} + 1$）个点后，距离 l_i将小于（$1+\epsilon$）r_{opt}。因此，当 $i \ge k$（$\frac{2}{\epsilon} + 1$）时，即可得到一个（$1+\epsilon$，$1-\delta$）-近似解。注意，基于上述随机采样的方法可能会遗漏一些“小”簇，因此 $|O_i|$ 可能小于 k；但始终可以保证遗漏的内点总数至多为 δn，即始终可以保证（$1+\epsilon$，$1-\delta$）-近似（否则，比率 $\frac{|P_{\text{opt}} \cap Q|}{|Q|} > \frac{\delta}{\gamma+\delta}$，我们可以继续从$P_{\text{opt}}$ 中采样一个点，然后更新 O_i）。

为了去除用于猜测包含 q 的簇的预言机，我们可以枚举所有可能的 k 情况；由于我们向 $T_1, T_2, \cdots, T_k$ 添加了 $k(\frac{2}{\epsilon}+1)$ 个点，因此总共会生成 $k^{k(\frac{2}{\epsilon}+1)} = 2^{k \log k(\frac{2}{\epsilon}+1)}$ 个解，并且至少有一个以概率 $(1-\gamma)(\frac{\delta}{\gamma+\delta})^{k(\frac{2}{\epsilon}+1)}$得到 $(1+\epsilon, 1-\delta)$-近似（证明方式与定理 5 相同）。

定理 11。设(P, γ) 是一个带有离群点的 k-中心聚类实例。给定两个参数 $\epsilon, \delta \in (0, 1)$，存在一种算法，以$(1-\gamma)(\frac{\delta}{\gamma+\delta})^{k(\frac{2}{\epsilon}+1)}$的概率输出一个$(1+\epsilon, 1-\delta)$-近似解。运行时间为 $O(2^{k \log k(\frac{2}{\epsilon}+1)}(n + \frac{1}{\epsilon^5})d)$。

如果以常数概率重复运行算法 $O(\frac{1}{1-\gamma}(\frac{\gamma+\delta}{\delta})^{k(\frac{2}{\epsilon}+1)})$ 次，该算法会输出一个$(1+\epsilon, 1-\delta)$-近似解。

与我们在第 5.2.1 节对带离群点的最小包围球运行时间的讨论类似，B˘adoiu等人 [22] 也针对带离群点的 k-中心聚类问题实现了线性时间双准则近似（见他们论文中的第 4 节）。然而，他们的运行时间中隐藏的常数关于 $(\frac{k}{\epsilon\delta})^{O(1)}$ 呈指数级增长，远大于定理 11 中的“$k \log k(\frac{2}{\epsilon}+1)$”。

次线性时间算法。线性时间算法还可以进一步改进为次线性时间算法；其思路与第 5.2.2 节中设计带离群点的最小包围球次线性时间算法的思路类似。首先，我们遵循定义 7，定义形状集 $\mathcal{X}$，其中每个$x \in \mathcal{X}$都是空间中 k 个球的并集；中心 $c(x)$应为其 k 个球心组成的集合，记为$c(x) = \{o_x^1, o_x^2, \cdots, o_x^k\}$，而大小 $s(x)$为半径，即，$x = \cup_{j=1}^k \mathbb{B}(o_x^j, s(x))$。显然，如果

x 是实例$(P,\ \gamma)$的一个可行解，则大小 $\left|P \cap (\cup_{j=1}^{k} \mathbb{B}(o_x^j,\ s(x)))\right|$ 至少应为$(1-\gamma)n$。此外，定义距离函数 $f(c(x)\ p) = \min_{1\le j\le k} ||p - o_x^j||$。容易验证，形状集 $\mathcal{X}$ 满足性质1、2和3。根据引理 11，我们知道可以以概率$(1-\eta_1)\,3^{\delta}_{(\gamma+\delta)}$在 $P_{\text{opt}} \cap Q$ 中获得一个点。此外，根据引理 12，我们可以估算值 l_i 并选出最佳候选解。综上，我们得到以下定理。

定理 *12.* 设 (P,γ) 是一个带离群点的 k-中心聚类实例。给定参数 $\epsilon,\delta,\eta_1,\eta_2 \in (0,1)$，存在一种算法，以概率 $(1-\gamma)\big((1-\eta_1)(1-\eta_2)\,3^{\delta}_{(\gamma+\delta)}\big)^{k(\frac{2}{\epsilon}+1)}$ 输出一个 $(1+\epsilon, 1-\delta)$-近似解。运行时间为 $\tilde{O}(2^k\, log\ k(\frac{2}{\epsilon}+1)(\frac{\gamma}{\delta^2}+\frac{1}{\epsilon^5})d)$。

如果重复运行算法 $N = O\Big(\frac{1}{1-\gamma}\big(\frac{1}{1-\eta_1}\big(^{3}_{(\gamma+\delta)\,\delta}\big)\Big)^{k(\frac{2}{\epsilon}+1)}\Big)$ 次，并设置为 $\eta_2 = O(_{2^k}{}^{1}_{k}{}^{1}_{\log(\frac{2}{\epsilon}+1)N})$，则该算法以常数概率输出一个$(1+\epsilon, 1-\delta)$-近似解。

7.2 含离群点的平坦拟合

令 j 为介于 0 和 d 之间的固定整数。给定一个 j 维平坦 $\mathcal{F}$ 和一个点 $p \in \mathbb{R}^d$，我们将它们之间的距离 $dist(\mathcal{F},\ p)$ 定义为 p 到其在 $\mathcal{F}$ 上的投影之间的欧几里得距离。设 P 是 $\mathbb{R}^d$ 中由 n 个点构成的集合。**平坦拟合**问题是寻找使 $\max_{p\in P} dist(\mathcal{F},\ p)$ 最小的 j 维平坦 $\mathcal{F}$。不难看出，最小包围球问题是平坦拟合问题在 $j=0$ 情况下的特例。此外，给定参数 $\gamma \in (0,\ 1)$，**含离群点的平坦拟合**问题是寻找一个大小为 $(1-\gamma)n$ 的子集 $P' \subset P$，使得 $\max_{p\in P'} dist(\mathcal{F},\ p)$ 最小化。与带离群点的最小包围球问题类似，我们也用 P_{opt} 表示最优子集。在介绍含离群点的平坦拟合算法之前，我们先介绍 Har-Peled 和 Varadarajan [56] 提出的适用于普通版本（不含离群点）的线性时间算法。

我们从 $j=1$ 的情形开始，也就是说，平坦 $\mathcal{F}$ 是空间中的一条直线。粗略地说，他们的算法是一种迭代过程，逐轮更新解，直到它足够接近最优直线 l_{opt}。该算法分为两部分。（**1**）它选取任意一点 $p_\Delta \in P$，并令 q_Δ 为 P 中距离 p_Δ 最远的点；可以证明，经过 p_Δ 和 q_Δ 的直线（记为 l_0）是一个良好的初始解，其目标函数值可达到 4- 近似。（**2**）在之后的每一轮中，算法都将解从 l_{i-1} 更新为 l_i，其中 $i \ge 1$ 是当前轮次：令 p_i 为 P 中距离 l_{i-1} 最远的点，并令 h_i 表示由 p_i 和 l_{i-1} 张成的二维平坦；随后，算法在 h_i 上计算一个包含 $O(\frac{1}{\epsilon^8}\log^2\frac{1}{\epsilon})$ 条直线的集合，并通过“预言机”从中选取一条作为 l_i。他们证明了从 l_{i-1} 到 l_i 的改进足够显著；因此，经过运行 $\nu = O(\frac{1}{\epsilon^3}\log\frac{1}{\epsilon})$ 轮后，算法能够达到 $(1+\epsilon)$-近似。若要移除“预言机”，算法可以枚举 h_i 上的所有 $O(\frac{1}{\epsilon^8}\log^2\frac{1}{\epsilon})$ 条直线，因此总运行时间为 $O\big(2^{\frac{1}{\epsilon^3}\log^2\frac{1}{\epsilon}} nd\big)$。

线性时间算法。现在我们考虑将上述算法调整为适用于带离群点的情形；事实上，其思路与第 5.2.1 节中针对带离群点的最小包围球所提出的思路类似。为简单起见，我们仍使用与上文相同的记号。先考虑部分 **(1)** 。如果我们从 P 中随机选取一个点 p_Δ，那么它属于 P_{opt} 的概率为 $1-\gamma$；接着，我们从 p_Δ 到 P 的最远点中随机选取一个点，记为 q_Δ，其中这些最远点构成的集合大小为 $(\delta_0+\gamma)n$，而 δ_0 的值将在下文确定。显然，概率为 $\frac{\delta_0}{\gamma+\delta_0}$ 时，$q_\Delta \in P_{\text{opt}}$。记为 $P_0 = \{p \in P_{\text{opt}} \mid ||p-p_\Delta|| \le ||q_\Delta - p_\Delta||\}$。于是我们得到以下引理。

引理 13. 将经过 p_Δ 和 q_Δ 的直线记为 l_0。那么，以概率 $(1-\gamma)(\frac{\delta_0}{\gamma+\delta_0})$，

$$\max_{p\in P_0} dist(l_0,p) \le 4\max_{p\in P_0} dist(l_{opt},p) \le 4\max_{p\in P_{opt}} dist(l_{opt},p). \tag{77}$$

此外， P_0 的大小至少为 $\big(1-(\delta_0+\gamma)\big)n$。

显然可以得到 P_0 的大小。不等式（77）直接由 [56], 的上述结果推出，只要满足 p_Δ 和 $q_\Delta \in P_{\text{opt}}$。因此，我们可以将直线 l_0 作为初始解。接着，我们可以采用相同的随机采样思路，在第 i 轮选取点 p_i。具体而言，我们从 P 中距离 l_i 最远的 $(\delta_0+\gamma)n$ 个点组成的集合中，随机选取一个点作为 p_i。此外，我们需要将集合 P_{i-1} 缩小到 $P_i = \{p \in P_{i-1} \mid dist(l_{i-1}, p) \le dist(l_{i-1}, p_i)\}$。与引理 13 类似，我们可以证明，从 l_{i-1} 到 l_i 的改进以概率 $(1-\gamma)\left(\frac{\delta_0}{\gamma+\delta_0}\right)^{i+1}$ 足够显著，且 P_i 的大小至少为 $\big(1-((i+1)\delta_0+\gamma)\big)n$。经过 ν 轮后，我们得到直线 l_ν，使得 $\max_{p\in P_\nu} dist(l_\nu, p) \le (1+\epsilon)\max_{p\in P_{\text{opt}}} dist(l_{\text{opt}}, p)$，并且 $|P_\nu| \ge \big(1-((\nu+1)\delta_0+\gamma)\big)n$。因此，若我们将 $\delta_0 = \frac{\delta}{\nu+1}$ 设为给定的 $\delta \in (0,1)$，则直线 l_ν 将是实例 (P,γ) 的一个双准则 $(1+\epsilon, 1-\delta)$-近似。利用 [56], 中的思路，我们可以将结果扩展到满足 $\nu = \frac{e^{O(j^2)}}{\epsilon^{2j+1}}\log\frac{1}{\epsilon}$ 的情形 $j>1$。更多细节请参阅 [56]。

定理 *13.* 设 (P,γ) 是一个 j 维含离群点的平坦拟合实例。给定两个参数 $\epsilon, \delta \in (0,1)$，存在一种算法，以概率 $(1-\gamma)\left(\frac{1}{2}\right)^{g(j,\epsilon)}$ 输出一个 $(1+\epsilon, 1-\delta)$-近似，其中 $g(j,\epsilon) = poly(e^{O(j^2)}\frac{1}{\epsilon^j})$。运行时间为 $O(2^{g'(j,\epsilon)}nd)$，其中 $g'(j,\epsilon) = poly(e^{O(j^2)}\frac{1}{\epsilon^j})$。

如果重复运行算法 $\frac{2^{g(j,\epsilon)}}{1-\gamma}$ 次，以常数概率，该算法会输出一个$(1+\epsilon, 1-\delta)$-近似解。

次线性时间算法。我们可以将带离群点的平坦拟合问题视为带离群点的 MEX 问题。设 $r \ge 0$ 和 $\mathcal{F}$ 是一个 j 维平坦。于是，我们可以定义一个 j 维“板层” $SL(\mathcal{F}, r) = \{p \in \mathbb{R}^d \mid dist(\mathcal{F}, p) \le r\}$，其“中心”和“大小”分别为 $\mathcal{F}$ 和 r（例如，球是一个 0 维板层）；距离函数 $f(\mathcal{F}, p) = dist(\mathcal{F}, p)$。容易看出，板层的形状集满足性质1、2和3。此外，求最优平坦等价于求覆盖 P 的 $(1-\gamma)n$ 个点的最小板层。因此，利用引理 11 和 12，我们得到以下定理。

定理*14*。设 (P,γ) 是一个 j 维带离群点的平坦拟合实例。给定参数 ϵ、δ、η_1、$\eta_2 \in (0,1)$，存在一种算法，以概率 $(1-\gamma)\big((1-\eta_1)(1-\eta_2)^3\frac{\delta}{(\gamma+\delta)}\big)^{g(j,\epsilon)}$ 输出一个 $(1+\epsilon, 1-\delta)$-近似，其中 $g(j,\epsilon) = poly(e^{O(j^2)}\frac{1}{\epsilon^j})$。运行时间为 $O(2^{g'(j,\epsilon,\delta,\gamma)}nd)$，其中 $g'(j,\epsilon) = poly(e^{O(j^2)}\frac{1}{\epsilon^j}, \frac{1}{\delta}, \frac{1}{\gamma})$。

如果反复运行该算法 $N = O\Big(\frac{1}{1-\gamma}\big(\frac{1}{1-\eta_1}\big(\frac{3(\gamma+\delta)}{\delta}\big)\big)^{g(j,\epsilon)}\Big)$ 次，并采用设置$\eta_2 = O\big(\frac{1}{2^{g(j,\epsilon)}N}\big)$，则该算法以常数概率输出一个（$1+\epsilon, 1-\delta$）-近似解。

7.3 带离群点的单类支持向量机

实际中，数据集通常包含离群点。离群点可能会显著降低支持向量机的分离间隔。如 [40], 所述， [88, 93] 现有的大多数支持向量机离群点去除技术都是数值方法（例如，在目标函数中添加某种惩罚项），且只能保证得到局部最优解。Ding 和 Xu [40] 将含离群点的支持向量机建模为一个组合优化问题，并提出了一种名为“随机梯度下降树”的算法。我们首先关注带离群点的单类支持向量机，并在第7.4节说明如何扩展到带离群点的二类支持向量机。下面给出 [40]提出的带离群点的单类支持向量机问题的定义。

定义 9（带离群点的单类支持向量机）。给定由 $\mathbb{R}^d$ 中 P 个点组成的集合以及一个小参数 $\gamma \in (0,\ 1)$，带离群点的单类支持向量机问题是寻找一个子集 $P' \subset P$

算法 8 Gilbert 算法 [40, 49]

输入： $\mathbb{R}^d$ 中的一个点集 P，以及 $N \in \mathbb{Z}^+$。**输出：** v_i，作为原点与 P 之间多面体距离的近似解。

1. 将 $i = 1$ 和 v_1 初始化为 P 中距离原点 o 最近的点。2. 迭代执行以下步骤，直到满足 $i = N$。

(a) 找到点 $p_i \in P$，使其在 ov_i 线段的支撑直线上的正交投影到 o 的距离最近（称为 p_i 的投影距离），*i.e.*, $p_i = \arg\min_{p \in P}\{\frac{\langle p, v_i\rangle}{||v_i||}\}$，其中 $\langle p, v_i\rangle$ 是 p 和 v_i 的内积（见图 6）。(b) 令 v_{i+1} 为线段 $v_i p_i$ 上距离原点 o 最近的点；更新 $i = i + 1$。

其大小为 $(1-\gamma)n$，且存在一个超平面 $\mathcal{H}$ 将原点 o 与 P' 分隔开，使得 o 与 $\mathcal{H}$ 之间的距离最大。

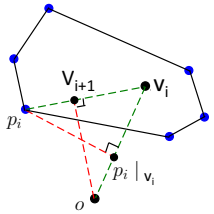

图 6：算法 8 中步骤 2 的示意图；$p_i|_{v_i}$ 是 p_i 在 ov_i 上的投影

线性时间算法。我们简要介绍 [40] 的算法。他们还考虑了具有两个小参数 ϵ，$\delta \in (0, 1)$ 的"双准则近似"：一个超平面 $\mathcal{H}$ 将原点 o 与一个大小为 $(1-\delta-\gamma)n$ 的子集 $P' \subset P$ 分隔开，其中 o 与 $\mathcal{H}$ 之间的距离至少为最优值的 $(1-\epsilon)$。[40] 的思路基于这样一个事实：支持向量机（不含离群点）问题等价于计算几何中的多面体距离问题 [48]。

令 o 为原点，P 为 $\mathbb{R}^d$ 中给定的点集。多面体距离问题是寻找凸包 P 内的一个点 q，使距离 $||q-o||$ 最小。

对于单类支持向量机的一个实例 P，可以证明，若 q_{opt} 是 o 与 P 之间多面体距离的最优解，则向量 $q_{opt} - o$ 是最优超平面的法向量。更多细节请参阅 [40, 48]。多面体距离问题可以通过 *Gilbert* 算法 [46, 49] 高效求解。为完整起见，我们在算法 8 中给出该算法。

与第2.1节中的最小包围球核心集构造方法类似，该算法也会在每次迭代中通过选择某个点 p_i 来贪心地改进当前解。令 ρ 为 o 与 P 之间的多面体距离，$D = \max_{p,q \in P} ||p-q||$，且为 $E = \frac{D^2}{\rho^2}$。已证明，给定 $\epsilon \in (0, 1)$，最多运行算法 8 $2\lceil 2E/\epsilon \rceil$ 步 [29, 48]，即可获得单类支持向量机的 $(1-\epsilon)$-近似解（即分离间隔的宽度至少为最优值的 $(1-\epsilon)$）。为处理离群点，[40] 的算法遵循第 5.2.1 节中的类似思路；它将贪心选择点 p_i 的步骤替换为从集合 Q 中随机采样一个点。该集合包含投影距离最小的 $(\delta+\gamma)n$ 个点（即算法 8 的步骤 2(a) 中函数 $\frac{\langle p, v_i\rangle}{||v_i||}$ 的取值）。为以常数成功概率获得 $(1-\epsilon, 1-\delta)$-近似解，该算法需要 $O\big(\frac{1}{1-\gamma}(1+\frac{\gamma}{\delta})^z \frac{D^2}{\epsilon\rho^2} nd\big)$ 时间，其中 $z = O(\frac{D^2}{\epsilon\rho^2})$。

次线性时间算法。我们将 $\mathcal{X}$ 定义为所有不包含原点 o 的闭半空间在 $\mathbb{R}^d$ 中的集合；对于每个 $x \in \mathcal{X}$，令 $\mathcal{H}_x$ 为包围 x 的超平面，并令 h_x

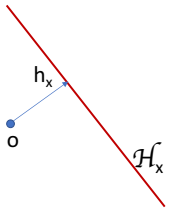

图 7： $\mathcal{H}_x$ 和 h_x的示意图。

为 o 在 $\mathcal{H}_x$ 上的投影（见图 7）。我们假设给定实例 (P, γ) 有可行解。也就是说，至少存在一个半空间 $x \in \mathcal{X}$，其超平面 $\mathcal{H}_x$将原点 o与大小为 $(1-\gamma)n$的子集 P'分隔开。我们定义中心 $c(x) = \frac{h_x}{||h_x||}$；由于定义 7 中的带离群点的 MEX 问题是一个最小化问题，因此我们设计大小函数 $s(x) = \frac{1}{||h_x||}$。显然，含离群点的支持向量机问题的 $(1-\epsilon)$-近似等价于关于大小函数 $s(x)$ 的 $\frac{1}{1-\epsilon}$-近似。我们还定义距离函数 $f(c(x), p) = -\langle p, \frac{h_x}{||h_x||}\rangle$。很容易验证，形状集 $\mathcal{X}$满足性质1、2和3。

回顾算法 8，在每次迭代中选择点 $p_i = \arg\min_{p\in P}\{\frac{\langle p, v_i\rangle}{||v_i||}\}$。实际上，向量 $\frac{v_i}{||v_i||}$ 可以视为形状中心，而 p_i 是基于距离函数 $f(c(x), p)$相对于 $\frac{v_i}{||v_i||}$ 的最远点。此外，前述线性时间算法中提到的集合 Q 实际上是从 P到 $\frac{v_i}{||v_i||}$的最远 $(\delta+\gamma)n$ 个点的集合。因此，我们可以应用引理 11从 $P_{opt} \cap Q$中抽取一个点，并应用引理 12估计每个候选解 $\frac{v_i}{||v_i||}$对应的 l_i 值。总的来说，我们可以将算法的运行时间 [40] 改进为与 n无关。

定理 15。令(P, γ)为含离群点的支持向量机的一个实例。给定参数 ϵ、δ、η_1、$\eta_2 \in (0, 1)$，存在一种算法，以概率$(1-\gamma)$输出一个$(1-\epsilon, 1-\delta)$近似$(1-\eta_1)(1-\eta_2)\frac{\delta}{(\gamma+\delta)}^z$，其中 $z = O(\frac{D^2}{\epsilon\rho^2})$。运行时间为 $\tilde{O}(\frac{D^2\gamma}{\delta^2\epsilon^2\rho^2}d)$。

如果将算法 $N = O\left(\frac{1}{1-\gamma}\left(\frac{1}{1-\eta_1}\left(3 + \frac{3\gamma}{\delta}\right)\right)^z\right)$ 次，并采用设置$\eta_2 = O\left(\frac{1}{zN}\right)$，那么该算法以常数概率输出一个 $(1-\epsilon, 1-\delta)$-近似解。

7.4 带离群点的二类支持向量机

下面给出文献 [40]中提出的带离群点的二类支持向量机问题的定义。

定义 10（带离群点的二类支持向量机）。给定两个点集 P_1 和 P_2，它们位于 $\mathbb{R}^d$中，以及两个小参数 γ_1、$\gamma_2 \in$（0，1），带离群点的二类支持向量机问题旨在找到两个子集 $P_1' \subset P_1$ 和 $P_2' \subset P_2$，满足 $|P_1'| =$（$1-\gamma_1$）$|P_1|$ 和 $|P_2'| =$（$1-\gamma_2$）$|P_2|$，并找到一个将 P_1' 和 P_2'分隔开的间隔，使该间隔的宽度最大化。

我们用 P_1^{opt}和 P_2^{opt}表示由最优间隔分隔的 P_1 和 P_2 的子集，其中 $|P_1^{opt}| = (1-\gamma_1)|P_1|$和 $|P_2^{opt}| = (1-\gamma_2)|P_2|$。普通二类支持向量机（不含离群点）问题等价于计算原点 o 与 $\mathcal{M}(P_1, P_2)$之间的多面体距离，其中 $\mathcal{M}(P_1, P_2)$是 P_1和 P_2 [48]的闵可夫斯基差。注意，无须显式计算集合 $\mathcal{M}(P_1, P_2)$。相反，算法 8 只需在每次迭代中从 $\mathcal{M}(P_1, P_2)$中选取一个点，整体运行时间仍与输入规模呈线性关系。为处理带离群点的二类支持向量机，Ding 和 Xu [40] 对单类情形的算法做了少许修改。每次迭代中，算法会考虑两个子集 $Q_1 \subset P_1$和 $Q_2 \subset P_2$，它们分别由 P_1 中投影距离最小的 $(\delta+\gamma_1)|P_1|$ 个点以及投影距离最大的 $(\delta+\gamma_2)|P_2|$ 个点组成

向量 v_i上的 P_2 中的点；然后，算法随机选取两个点 $p_i^1 \in Q_1$ 和$p_i^2 \in Q_2$，其差向量 $p_i^2 - p_i^1$ 将在算法 8 的步骤 2(a)中用作 p_i，以更新当前解 v_i。这种方法能在线性时间内得到 $(1-\epsilon, 1-\delta)$-近似。

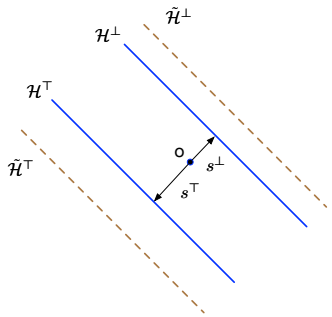

图 8：二分类支持向量机示意图。从 o 到 $\mathcal{H}^{\perp}$ 和 $\mathcal{H}^{\top}$ 的距离分别为 $s^{\perp}$ 和 $s^{\top}$。超平面 $\tilde{\mathcal{H}}^{\perp}$ 和 $\tilde{\mathcal{H}}^{\top}$ 分别是 $\mathcal{H}^{\perp}$ 和 $\mathcal{H}^{\top}$ 的估计值，o 到它们的距离分别为 $\tilde{s}^{\perp}$ 和 $\tilde{s}^{\top}$。

为了将算法改进为次线性时间，我们需要对之前针对单类情况的思路进行一些修改。首先，我们将距离函数改为：

$$f(p,c) = \begin{cases} -\langle p, \frac{h_x}{||h_x||} \rangle \text{ if } p \in P_1; \\ \langle p, \frac{h_x}{||h_x||} \rangle \quad \text{if } p \in P_2. \end{cases}$$

利用这一新的距离函数，我们可以应用引理 11，在次线性时间内分别得到点 $p_i^1 \in Q_1 \cap P_1^{opt}$ 和 $p_i^2 \in Q_2 \cap P_2^{opt}$。给定一个向量（即候选中心） $\frac{v_i}{||v_i||}$，假设 $\mathcal{H}^{\perp}$ 和 $\mathcal{H}^{\top}$ 是与 $\frac{v_i}{||v_i||}$ 正交的平行超平面，由它们形成的间隔将 P_1' 和 P_2' 分隔开，其中 $P_1' \subset P_1$ 和 $P_2' \subset P_2$ 分别具有 $|P_1'| = (1-\gamma_1)|P_1|$ 和 $|P_2'| = (1-\gamma_2)|P_2|$。不失一般性地，假设原点 o 位于间隔内部。设 o 到 $\mathcal{H}^{\perp}$ 和 $\mathcal{H}^{\top}$ 的距离分别为 $s^{\perp}$ 和 $s^{\top}$。于是，我们得到两个形状（闭半空间） $x^{\perp} = (-\frac{v_i}{||v_i||}, \frac{1}{s^{\perp}})$ 和 $x^{\top} = (\frac{v_i}{||v_i||}, \frac{1}{s^{\top}})$，其具有 $P_1' \subset x^{\perp}$ 和 $P_2' \subset x^{\top}$。因此，我们可以两次应用引理 12，得到两个值 $\frac{1}{\tilde{s}^{\perp}} \leq \frac{1}{s^{\perp}}$ 和 $\frac{1}{\tilde{s}^{\top}} \leq \frac{1}{s^{\top}}$，其中 $\left| P_1 \setminus x\,(-\frac{v_i}{||v_i||}, \frac{1}{\tilde{s}^{\perp}}) \right| \leq (O(\delta) + \gamma_1)|P_1|$，且 $\left| P_2 \setminus x\,(\frac{v_i}{||v_i||}, \frac{1}{\tilde{s}^{\top}}) \right| \leq (O(\delta) + \gamma_2)|P_2|$。因此，我们可以使用值 $\tilde{s}^{\perp} + \tilde{s}^{\top}$ 作为 $s^{\perp} + s^{\top}$ 的估计值。示意图见图 8。总体而言，我们可以在次线性时间内实现 $(1-\epsilon, 1-O(\delta))$-近似。

8 未来工作

基于我们的工作，未来还有若干值得研究的有趣问题。例如，与半径近似不同，目前关于最小包围球覆盖近似的研究仍不充分。具体而言，我们能否像 [30] 证明半径近似最小包围球的下界结果那样，给出计算覆盖近似最小包围球的复杂度下界？此外，能否将稳定性概念推广到结构更复杂的其他几何优化问题？在第7节中，我们只为带离群点的 MEX 问题给出了双准则近似。因此，将稳定性概念推广到这些几何优化问题，并据此为其设计混合近似算法，是一个值得探讨的方向。

9 致谢

本研究部分得到中国国家重点研发计划项目（资助编号：2021YFA1000900）和安徽省自然科学基金（资助编号：2208085MF163）的支持。作者还要感谢徐晋辉教授对本稿提出的宝贵意见。

参考文献

1. P. K. Agarwal、S. Har-Peled 和 K. R. Varadarajan。通过核心集实现几何近似。 *Combinatorial and Computational Geometry* 52 1 30, 2005 , : – 。2. P. K. Agarwal、S. Har-Peled 和 H. Yu。欧几里得空间中曲面、曲线和移动点的嵌入。载于 *Proceedings of the 23rd ACM Symposium on Computational Geometry, Gyeongju, South Korea, June 6-8 2007* 381 389, 2007 , , 第 – 页。3. P. K. Agarwal、S. Har-Peled 和 H. Yu。通过剥离和栅格核集进行鲁棒形状拟合。 *Discrete & Computational Geometry* 39 1 3 38 58, 2008 , (-): – 。4. P. K. Agarwal 和 R. Sharathkumar。高维范围问题的流式算法。*Algorithmica*, 72(1):83–98, 2015年。5. A. Aggarwal、H. Imai、N. Katoh 和 S. Suri。寻找直径最小的 k 个点及相关问题。 *Journal of algorithms*, 12(1):38–56, 1991。6. Z. Allen Zhu、Z. Liao 和 Y. Yuan。加速计算几何的优化算法。载于 *43rdInternational Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15*, 2016, *Rome Italy* 53 1 53 6, 2016 , , 第 : – : 页。7. N. Alon、S. Dar、M. Parnas 和 D. Ron。聚类测试。 *SIAM Journal on Discrete Mathematics*, 16 3 393( ): –417, 2003。8. P. Awasthi、A. Blum 和 O. Sheffet。稳定性为 k-中位数和 k-均值聚类带来多项式时间近似方案。载于 *51th Annual IEEE Symposium on Foundations of Computer Science FOCS 2010 October 23-26 2010 Las* , , , , , , , 第 – 页。9. P. Awasthi、A. Blum 和 O. Sheffet。扰动稳定性下的基于中心的聚类。 *Vegas Nevada USA* 309 318, 2010 *Inf. Process. Lett.* 112 1 2 49 54, 2012 , (-): – 。10. M. Balcan 和 M. Braverman。寻找低误差聚类。载于 *COLT 2009 - The 22nd Conference on Learning Theory Montreal Quebec Canada June 18-21 2009* 2009 , , , , , , 。11. M. Balcan、N. Haghtalab 和 C. White。扰动韧性下的 k-中心聚类。载于 *43rdInternational Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15*, 2016年, *Rome Italy* 68 1 68 14, 2016 , , 第 : – : 页。12. M. Balcan 和 Y. Liang。扰动韧性下的聚类。 *SIAM J. Comput.*, 45(1):102–155, 2016年。13. M.-F. Balcan、A. Blum 和 A. Gupta。近似稳定性下的聚类。*(JACM)* 60 2 8, 2013 *Journal of the ACM* , ( ): 。14. D. Bertsimas 和 M. Sim。鲁棒性的代价。 *Oper. Res.*, 52(1):35–53, 2004 年。15. A. Bhattacharyya 和 Y. Yoshida。 *Property Testing - Problems and Techniques*。施普林格，2022年。16. B. Biggio、B. Nelson 和 P. Laskov。针对支持向量机的投毒攻击。载于 *Proceedings of the 29th International Conference on Machine Learning ICML 2012 Edinburgh Scotland UK June 26 - July 1 2012* , , , , , , , , 。17. B. Biggio 和 F. Roli。异常模式：对抗机器学习兴起十年之后。 *Pattern Recognition* 84 317 331, 2018 , : – 。18. Y. Bilu 和 N. Linial。稳定实例容易求解吗? *Combinatorics Probability & Computing* 21 5 643 660, , , ( ): – 。19. M. Blum、R. W. Floyd、V. Pratt、R. L. Rivest 和 R. E. Tarjan。选择问题的时间界。 *Journal of Computer and System Sciences* 7 4 448 461, 1973 , ( ): – 。20. M. Bˇadoiu 和 K. L. Clarkson。更小的球核心集。载于 , 第 – 页。21. M. Bˇadoiu 和 K. L. Clarkson。球的最优核心集。 *Computational Geometry*, 40(1):14–22, 2008。22. M. Bˇadoiu、S. Har-Peled 和 P. Indyk。通过核心集实现近似聚类。载于 , 第 – 页。23. G. C. Calafiore 和 M. C. Campi。不确定凸规划：随机化解和置信水平。 *Math. Program.*, 102(1):25–46, 2005年。24. M. Ceccarello、A. Pietracaprina 和 G. Pucci。在 MapReduce 和流式处理中求解带离群点的 k-中心聚类，精度几乎与顺序算法相当。 *PVLDB*, 12(7):766–778, 2019年。25. T. M. Chan 和 V. Pathak。高维最小包围球的流式和动态算法。 *Comput. Geom.*, 47(2):240–247, 2014年。26. C.-C. Chang 和 C.-J. Lin。LIBSVM：支持向量机库。 *ACM TIST*, 2(3), 2011 年。 *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)* 801 802, 2003 *Proceedings of the ACM Symposium on Theory of Computing (STOC)* 250 257, 2002

27. M. Charikar、S. Khuller、D. M. Mount 和 G. Narasimhan。带离群点的设施选址问题算法。载于 *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*, 第 642–651 页。工业与应用数学学会，2001年。28. M. Charikar、L. O'Callaghan 和 R. Panigrahy。更优的聚类问题流式算法。载于 *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, 第 30–39 页。ACM，2003年。29. K. L. Clarkson。核心集、稀疏贪心近似与 Frank-Wolfe 算法。 *ACM Transactions on Algorithms* 6 4 63, 2010 , ( ): 。

30. K. L. Clarkson、E. Hazan 和 D. P. Woodruff。用于机器学习的次线性优化。 *J. ACM*, 59 5 23 1 23( ): : – :49, 2012年。31. V. Cohen-Addad、D. Saulpic 和 C. Schwiegelshohn。欧几里得空间中幂均值的改进核心集和次线性算法。 *Advances in Neural Information Processing Systems*, 34:21085–21098, 2021年。32. C. Cortes 和 V. Vapnik。支持向量网络。 *Machine Learning*, 20:273, 1995年。33. D. J. Crisp 和 C. J. C. Burges。v-SVM分类器的几何解释。载于 S. A. Solla、T. K. Leen 和 K.-R. M¨uller 编， *NIPS*, 第 244–250 页。麻省理工学院出版社，1999年。34. A. Czumaj 和 C. Sohler。次线性时间算法。35. A. Czumaj 和 C. Sohler。通过随机采样实现聚类的次线性时间近似。载于 *International Colloquium on Automata Languages and Programming* 396 407 2004 , , , 第 – 页。

施普林格，。36. S. Dasgupta 和 A. Gupta。Johnson–Lindenstrauss 定理的初等证明。 *Random Structures & Algorithms* 22 1 60 65, 2003 , ( ): – 。

37. H. Ding。高维含离群点几何优化的次线性时间框架。载于 F. Grandoni、G. Herman 和 P. Sanders 编， *28th Annual European Symposium on Algorithms ESA* , , , , , , 卷 *2020 September 7-9 2020 Pisa Italy (Virtual Conference)* 173 *LIPIcs* 38 1 38 21 of , 第 : – : 页。Schloss Dagstuhl - Leibniz-Zentrum f¨ur Informatik，2020年。38. H. Ding。稳定性为机器学习中的几何优化带来次线性时间算法。载于 P. Mutzel、R. Pagh 和 G. Herman 编， *29th Annual European Symposium on Algorithms, ESA 2021, September 6-8 2021 Lisbon Portugal (Virtual Conference)* 204 *LIPIcs* 38 1 38 19 , , , , 卷 of , 第 : – : 页。Schloss Dagstuhl - Leibniz-Zentrum f¨ur Informatik，2021年。39. H. Ding 和 J. Xu。最小截尾平方估计量及相关问题的次线性时间混合近似。载于 *Proceedings of the International Symposium on Computational geometry (SoCG)* 110, , 第 页。2014
40. H. Ding 和 J. Xu。随机梯度下降树：一种针对含离群点支持向量机的组合方法。载于 *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 第 2561–2567 页，2015年。41. H. Ding、H. Yu 和 Z. Wang。贪心策略适用于含离群点的 k-中心聚类和核心集构造。载于 *27th Annual European Symposium on Algorithms, ESA 2019, September 9-11*, 2019年, *Munich/Garching* 40 1 40 16, 2019 , *Germany.*, 第 : – : 页。42. A. Efrat、M. Sharir 和 A. Ziv。计算最小 k 包围圆及相关问题。 *Computational Geometry* 4 3 119 136, 1994 , ( ): – 。

43. D. Feldman。核心集：最新综述。 *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 10(1), 2020年。44. D. Feldman、C. Xiang、R. Zhu 和 D. Rus。差分隐私 k 均值聚类的核心集及其在移动传感器网络隐私保护中的应用。载于 *Proceedings of the 16th ACM/IEEE InternationalConference on Information Processing in Sensor Networks, IPSN 2017, Pittsburgh, PA, USA, April 18-21*, 2017 3, 第 –15 页，2017年。45. K. Fischer、B. G¨artner 和 M. Kutz。高维最小包围球的快速计算。载于*Algorithms - ESA 2003, 11th Annual European Symposium, Budapest, Hungary, September 16-19*, 2003年, *Proceedings* 630 641, 2003 , 第 – 页。
46. M. Frank 和 P. Wolfe。二次规划算法。 *Naval Research Logistics Quarterly*, 3 1 2 95 110, 1956年 (-): – 。47. D. Garber 和 E. Hazan。次线性时间近似半定规划。载于 J. Shawe-Taylor、R. S. Zemel、P. L. Bartlett、F. C. N. Pereira 和 K. Q. Weinberger 编， *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011 Granada Spain* 1080 1088, 2011 , , , 第 – 页。

48. B. G ¨ artner 和 M. Jaggi。多面体距离的核心集。载于 *Proceedings of the International Symposium on Computational geometry (SoCG)* 33 42, 2009 , 第 – 页。49. E. G. Gilbert。在凸集上计算二次型最小值的迭代过程。 *SIAM Journal on Control*, 4(1):61–80, 1966年。50. O. Goldreich、S. Goldwasser 和 D. Ron。性质测试及其与学习和近似的联系。 *J. ACM*, 45(4):653–750, 1998年。51. T. F. Gonzalez。使最大簇间距离最小化的聚类。 *Theoretical Computer Science*, 38 293 306, 1985年 : – 。52. I. J. Goodfellow、P. D. McDaniel 和 N. Papernot。增强机器学习对抗对抗输入的鲁棒性。 *Commun. ACM*, 61(7):56–66, 2018年。53. L. Gyongyosi 和 S. Imre。量子密码学中物理允许的量子克隆变换的几何分析。 *Information Sciences*, 285:1–23, 2014年。

54. S. Har-Peled 和 S. Mazumdar。快速计算最小 k 包围圆的算法。 *Algorithmica*, 41 3 147 157, 2005 ( ): – 。55. S. Har-Peled 和 K. R. Varadarajan。通过线性化进行近似形状拟合。载于 *42nd Annual Symposium on Foundations of Computer Science, FOCS 2001, 14-17 October 2001, Las Vegas, Nevada, USA* 66 73, 2001 , 页码 – 。56. S. Har-Peled 和 K. R. Varadarajan。线性时间高维形状拟合。 *Discret. Comput.* , ( ): – 。57. S. Har-Peled 和 Y. Wang。含离群点的形状拟合。 *SIAM Journal on Computing*, 33(2):269–285, 2004。58. D. Haussler 和 E. Welzl。ε-网与单纯形范围查询。 *Discrete & Computational Geometry*, 2 2 127 151, 1987 ( ): – 。59. K. Hayashi 和 Y. Yoshida。常数时间内最小化二次函数。载于 D. D. Lee、M. Sugiyama、U. von Luxburg、I. Guyon 和 R. Garnett 编, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10*, 2016, *Barcelona, Spain* 2217 2225, 2016 , 页码 – 。60. E. Hazan、T. Koren 和 N. Srebro。超越随机梯度下降：次线性时间学习支持向量机。载于 J. Shawe-Taylor、R. S. Zemel、P. L. Bartlett、F. C. N. Pereira 和 K. Q. Weinberger 编, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011 Granada Spain* 1233 1241, 2011 , , , 页码 – 。61. D. S. Hochbaum 和 D. B. Shmoys。k 中心问题的最优启发式算法。 *Mathematics of operations research* 10 2 180 184, 1985 , ( ): – 。62. L. Huang、S. Jiang、J. Li 和 X. Wu。倍增度量空间中用于聚类（含离群点）的 ε-核心集。载于 *59th IEEE Annual Symposium on Foundations of Computer Science FOCS 2018 Paris France October 7-9 2018* 814 825, 2018 , , , , , , 页码 – 。63. P. Indyk。度量空间问题的次线性时间算法。载于 *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing May 1-4 1999 Atlanta Georgia USA* 428 434, 1999 , , , , , , 页码 – 。64. P. Indyk。度量空间聚类的次线性时间近似方案。载于 *40th Annual Symposium on Foundations of Computer Science FOCS '99 17-18 October 1999 New York NY USA* 154 159, , , , , , , , 页码 – 。1999
65. M. Jagielski、A. Oprea、B. Biggio、C. Liu、C. Nita-Rotaru 和 B. Li。操纵机器学习：回归学习中的投毒攻击与对策。载于 *2018 IEEE Symposium on Security and Privacy SP 2018 Proceedings 21-23 May 2018 San Francisco California USA* 19 35, 2018 , , , , , , , 页码 – 。66. M. Kerber 和 S. Raghvendra。通过随机投影实现投影聚类的近似算法与流式算法。载于 *Proceedings of the 27th Canadian Conference on Computational Geometry CCCG 2015 Kingston Ontario Canada August 10-12 2015* 2015 , , , , , , , , 。67. M. Kerber 和 R. Sharathkumar。低维和高维中的近似切赫复形。载于 *Algorithms and Computation - 24th International Symposium, ISAAC 2013, Hong Kong, China, December 16-18*, 2013, *Proceedings* 666 676, 2013 , 页码 – 。68. A. Kumar 和 R. Kannan。利用谱范数进行聚类与 k 均值算法。载于 *2010 IEEE 51st Annual Symposium on Foundations of Computer Science* 299 308 2010 , 页码 – 。IEEE, 。69. P. Kumar、J. S. B. Mitchell 和 E. A. Yildirim。利用核心集在高维中近似求解最小包围球。 *ACM Journal of Experimental Algorithmics*, 8, 2003。70. S. Lloyd。脉冲编码调制中的最小二乘量化。 *IEEE transactions on information theory* 28 2 129 137, , ( ): – 。71. J. Matoušek。用圆包围 k 个点。 *Information Processing Letters*, 53(4):217–221, 1995。72. R. M. McCutchen 和 S. Khuller。含离群点且具有匿名性的 k 中心聚类流式算法。载于 *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, 页码 165–178。施普林格，2008。73. A. Meyerson、L. O'callaghan 和 S. Plotkin。运行时间与数据集大小无关的 k 中位数算法。 *Machine Learning*, 56(1-3):61–87, 2004。74. N. Mishra、D. Oblinger 和 L. Pitt。次线性时间近似聚类。载于 *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms* 439 447 , 页码 – 。工业与应用数学学会，2001。75. R. Motwani 和 P. Raghavan。 *Randomized Algorithms*。剑桥大学出版社，美国，1995。76. F. Nielsen 和 R. Nock。近似求解最小包围球及其在机器学习中的应用。 *Int. J. Comput. Geom. Appl.* 19 5 389 414, 2009 , ( ): – 。77. K. Nissim、U. Stemmer 和 S. P. Vadhan。以隐私方式定位一个小簇。载于 *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2016, San Francisco, CA USA June 26 - July 01 2016* 413 427, 2016 , , , , 页码 – 。78. R. Ostrovsky、Y. Rabani、L. J. Schulman 和 C. Swamy。Lloyd 型方法对于 k 均值问题的有效性。 *Journal of the ACM (JACM)*, 59(6):28, 2012。79. R. Panigrahy。高维最小包围多胞体。 *arXiv preprint cs/0407020*, 2004。80. J. M. Phillips。核心集与草图。 *Computing Research Repository*, 2016。

81. J. Platt。使用序列最小优化快速训练支持向量机。载于 B. Sch¨olkopf、C. J. C. Burges 和 A. J. Smola 编，*Advances in Kernel Methods — Support Vector Learning*, 页码 185–208, 马萨诸塞州剑桥，1999年。麻省理工学院出版社。82. T. Roughgarden。超越最坏情况分析。 *Commun. ACM*, 62(3):88–96, 2019年。83. R. Rubinfeld。次线性时间算法。Citeseer, 2006年。84. A. Saha、S. V. N. Vishwanathan 和 X. Zhang。新的最小包围凸形状近似算法。载于 *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2011 San Francisco California USA January 23-25 2011* 1146 1160, 2011 , ,,,,, 页码 – 。85. B. Sch¨olkopf 和 A. J. Smola。 *Learning with Kernels: support vector machines, regularization, optimization, and beyond* 2002 。自适应计算与机器学习丛书。麻省理工学院出版社，。86. B. Scholkopf、A. J. Smola、K. R. Muller 和 P. L. Bartlett。新型支持向量算法。*Neural Computation* 12 1207 1245, 2000 ,: – 。87. D. R. Sheehy。随机投影下距离函数的持久同调。载于 *30th Annual Symposium on Computational Geometry SOCG'14 Kyoto Japan June 08 - 11 2014* 328, 2014 , ,,,,, 页码。88. S. Suzumura、K. Ogawa、M. Sugiyama 和 I. Takeuchi。离群点路径：用于鲁棒支持向量机的同伦算法。载于 T. Jebara 和 E. P. Xing 编，*Proceedings of the 31st International Conference on Machine Learning (ICML-14)* 1098 1106, 2014 , 页码 – 。

89. P.-N. Tan、M. Steinbach 和 V. Kumar。 *Introduction to Data Mining*。2006年。90. I. W. Tsang、J. T. Kwok 和 P. Cheung。核心向量机：在超大数据集上快速进行支持向量机训练。*Journal of Machine Learning Research*, 6:363–392, 2005年。91. I. W. Tsang、J. T. Kwok 和 P.-M. Cheung。核心向量机：在超大数据集上快速进行支持向量机训练。*Journal of Machine Learning Research*, 6:363–392, 2005年。92. V. N. Vapnik 和 A. Y. Chervonenkis。事件相对频率一致收敛至其概率。载于 *Measures of complexity*, 页码 11–30。施普林格，2015年。93. L. Xu、K. Crammer 和 D. Schuurmans。通过凸离群点消除实现鲁棒支持向量机训练。载于 *AAAI*, 页码 536–542。AAAI出版社，2006年。94. H. Zarrabi-Zadeh 和 A. Mukhopadhyay。在高维中对含离群点的 1 *Proceedings* -中心进行流式处理。载于 *of the Canadian Conference on Computational Geometry (CCCG)* 83 86, 2009 , 页码 – 。

定理1的证明

为确保算法 [20], 在每次迭代中都能取得预期的改进，他们证明了：如果算法始终选择距当前 $\mathbf{MEB}(T)$ 中心最远的点，则以下两个不等式成立：

$$r_{i+1} \geq (1+\epsilon)\mathbf{Rad}(P) - L_i; \quad r_{i+1} \geq \sqrt{r_i^2 + L_i^2}, \tag{78}$$

其中， r_i 和 r_{i+1} 分别是第 i 次和第 $(i+1)$ 次迭代中 $\mathbf{MEB}(T)$ 的半径，而 L_i 是从第 i 次迭代到第 $(i+1)$ 次迭代期间 $\mathbf{MEB}(T)$ 中心的移动距离。

然而，我们通常在每次迭代中只计算近似的 $\mathbf{MEB}(T)$。在第 i 次迭代中，我们分别用 c_i 和 o_i 表示精确和近似的 $\mathbf{MEB}(T)$ 的中心。假设 $||c_i - o_i|| \leq \xi r_i$，其中 $\xi \in (0, \frac{\epsilon}{1+\epsilon})$（稍后我们会说明为何需要这个界）。注意，我们在每次迭代中只计算 o_i，而不是 c_i。因此，我们只能选择距离 o_i 最远的点（记作 q）。如果 $||q - o_i|| \leq (1+\epsilon)\mathbf{Rad}(P)$，则问题已解决，并且已经得到 MEB 的 $(1+\epsilon)$-半径近似。否则，我们有

$$\begin{aligned}(1+\epsilon)\mathbf{Rad}(P) &< ||q - o_i|| \\ &\leq ||q - c_{i+1}|| + ||c_{i+1} - c_i|| + ||c_i - o_i|| \\ &\leq r_{i+1} + L_i + \xi r_i\end{aligned} \tag{79}$$

根据三角不等式。换言之，我们应将（78）的第一个不等式替换为$r_{i+1} > (1+\epsilon)$**半径**$(P) - L_i - \xi r_i$。此外，（78）的第二个不等式仍然成立，因为它只依赖于精确最小包围球的性质（见 [20]中的引理 2.1）。因此，我们有

$$r_{i+1} \geq \max\left\{\sqrt{r_i^2 + L_i^2}, (1+\epsilon)\mathbf{Rad}(P) - L_i - \xi r_i\right\}. \tag{80}$$

与 [20], 中的分析类似，我们令 $\lambda_i = \quad r_i(1+\epsilon)$半径$(P)$。由于 r_i是T的**最小包围球**的半径，且 $T \subset P$，我们知道 $r_i \leq$**半径**(P)，进而有 $\lambda_i \leq 1/(1+\epsilon)$。通过简单计算可知，当 $L_i = \big($$(1+\epsilon \quad P-\xi r_i\big)^2 - r_i^2\big)$半径 $2\big(1+\epsilon \quad P-\xi r_i\big)$半径时，（80）中 r_{i+1} 的下界达到最小值。将 L_i 的这个值代入（80），得到

$$\lambda_{i+1}^2 \geq \lambda_i^2 + \frac{\big((1-\xi\lambda_i)^2 - \lambda_i^2\big)^2}{4(1-\xi\lambda_i)^2}. \tag{81}$$

为简化不等式 (81)，我们考虑函数 $g(x) = (1-x)_{1-x}^{2-\lambda_i^2}$，其中 $0 < x < \xi$。其导数 $g'(x) = -1 - \quad \lambda_i^2(1-x)^2$ 始终为负，因此我们有

$$g(x) \geq g(\xi) = \frac{(1-\xi)^2 - \lambda_i^2}{1-\xi}. \tag{82}$$

由于 $\xi < \frac{\epsilon}{1+\epsilon}$ 和 $\lambda_i \leq 1/(1+\epsilon)$，我们知道 (82) 的右侧始终非负。利用 (82)，不等式 (81) 可简化为

$$\begin{aligned}\lambda_{i+1}^2 &\geq \lambda_i^2 + \frac{1}{4}\big(g(\xi)\big)^2 \\ &= \lambda_i^2 + \frac{\big((1-\xi)^2 - \lambda_i^2\big)^2}{4(1-\xi)^2}.\end{aligned} \tag{83}$$

(83) 可进一步改写为

$$\begin{aligned}&\Big(\frac{\lambda_{i+1}}{1-\xi}\Big)^2 \geq \frac{1}{4}\Big(1 + (\frac{\lambda_i}{1-\xi})^2\Big)^2 \\ \Longrightarrow\ &\frac{\lambda_{i+1}}{1-\xi} \geq \frac{1}{2}\Big(1 + (\frac{\lambda_i}{1-\xi})^2\Big).\end{aligned} \tag{84}$$

现在，我们可以对 λ_i 应用与 [20]中相同的变换。令 $\gamma_i = \frac{1}{1-\frac{\lambda_i}{1-\xi}}$。我们知道 $\gamma_i > 1$（注意 $0 \leq \lambda_i \leq \frac{1}{1+\epsilon}$ 和 $\xi < \frac{\epsilon}{1+\epsilon}$）。于是，由 (84) 可得

$$\begin{aligned}\gamma_{i+1} &\geq \frac{\gamma_i}{1-\frac{1}{2\gamma_i}} \\ &= \gamma_i\big(1 + \frac{1}{2\gamma_i} + (\frac{1}{2\gamma_i})^2 + \cdots\big) \\ &> \gamma_i + \frac{1}{2},\end{aligned} \tag{85}$$

其中，该方程源于以下事实：$\gamma_i > 1$，因此 $\frac{1}{2\gamma_i} \in (0, \frac{1}{2})$。注意，$\lambda_0 = 0$，因此 $\gamma_0 = 1$。由此可得 $\gamma_i > 1 + \frac{i}{2}$。此外，由于 $\lambda_i \leq \frac{1}{1+\epsilon}$，即$(1+\epsilon)(1-\xi)$，我们有

$$i < \frac{2}{\epsilon - \xi - \epsilon\xi} = \frac{2}{(1 - \frac{1+\epsilon}{\epsilon}\xi)\epsilon}. \tag{86}$$

因此，我们得到该定理。

B [22] 中的引理 2.2

引理 14（[22]）。设 $\mathbb{B}(c, r)$ 是点集 $P \subset \mathbb{R}^d$的一个最小包围球，那么任何包含 c的闭半空间都必须包含 P 中的至少一个点，且该点与 c的距离为r 。