arXiv:2601.00272v1 [cs.DS] 1 Jan 2026

对抗鲁棒近似最近邻搜索的高效算法

亚历山大尔·Andoni* Themistoklis Haris† Esty Kelman‡ 克日什托夫·奥纳 k§

2026年1月5日

摘要

我们研究强大的自适应对手控制数据集和一系列 Q 查询时的近似最近邻（ANN）问题。

主要针对 $d = \omega(\sqrt{Q})$ 的高维情形，我们提出了一系列保证逐步增强的算法。首先，我们建立了自适应安全性与公平性之间的新颖联系，利用公平近似最近邻搜索（Aumüller 等人，2022）在具备信息论保证的情况下，向对手隐藏内在随机性。为实现与数据无关的性能，随后我们将搜索问题归约为一个鲁棒决策原语，并在局部敏感哈希（LSH）数据结构上使用差分隐私机制（Hassidim 等人，2022）加以解决。然而，这种方法面临固有的 $\sqrt{n}$ 查询时间瓶颈。为突破这一瓶颈，我们提出了一种新颖的同心环带 LSH 构造，综合运用了公平性与差分隐私技术。分析中引入了一种从底层算法实例中鲁棒地发布时序信息的新方法；作为推论，该方法还改进了公平近似最近邻方面的现有结果。

此外，对于低维情形 $d = O(\sqrt{Q})$，我们提出了专门的算法，可提供强有力的“对所有情况”保证：以高概率确保对每一个可能的查询都正确。我们引入了新颖的度量覆盖构造，简化并改进了以往针对汉明空间和 ℓ_p 空间中近似最近邻问题的方法。

*哥伦比亚大学计算机科学系。电子邮箱：andoni@cs.columbia.edu

†波士顿大学计算机科学系。电子邮箱：tharis@bu.edu

‡波士顿大学计算机科学系和计算与数据科学系，以及麻省理工学院 CSAIL。部分经费由美国国家科学基金会项目（编号 2022446）资助，部分由 NSF TRIPODS 计划（项目编号 DMS-2022448）资助。电子邮箱：ekelman@mit.edu。§波士顿大学计算与数据科学学院。电子邮箱：konak@bu.edu

目录

1 引言 31.1 路线图与讨论.41.2 我们的结果与技术 .4

2 相关及同期工作 72.1 与 Feng 等人(2025) .7

3 预备知识 8**4 公平性蕴含鲁棒性 9****5 假设**

通过分桶实现自由鲁棒搜索 11

5.1 弱判定近似最近邻 .115.2 基于分桶的搜索 .12

6 通过同心 LSH 环带实现宽松公平 ANN 13**7 鲁棒近似最近邻改进 14**7.1 分析 .15**8 结论 18****A 符号表 23**

B 差分隐私背景 23

B.1 差分隐私的定义 .24B.2 拉普拉斯机制及其性质 .24B.3 差分隐私的性质.24B.4 通过保护内部随机性实现鲁棒化 .25

C 定理 C.1 的证明 **25**

D 改进的鲁棒近似最近邻搜索算法 $\forall$ 保证 28

D.1 汉明立方体中的全体保证.29D.1.1 通过采样改进查询运行时间 .31D.1.2 利用最优 LSH 算法 .32D.2 通过度量覆盖对连续空间进行离散化 .32D.2.1 连续空间中的度量覆盖.32D.2.2 鲁棒近似最近邻算法 .34D.2.3 消除对尺度的依赖 .35

1 引言

随机性是算法设计中的重要工具，它能够避开困扰确定性方法的最坏情况，从而以较少资源实现解决方案（Motwani 和 Raghavan，1996）。对此类算法的经典分析假设处于**非自适应**环境中，即数据更新和查询都预先固定。然而，面对**自适应对手**时，这一假设便不再成立：对手可以根据算法先前的输出发起查询。这些输出可能泄露算法内部随机性的信息，使对手得以构造查询序列，恶意破坏算法的性能保证（Hardt 和 Woodruff，2013；Gribelyuk 等人，2024）。

在设计用于估计问题的对抗鲁棒算法方面，已经取得了显著进展；这类问题的输出是一个单一数值（Lai 和 Bayraktar，2020；Hassidim 等人，2022；Chakrabarti 等人，2021；Attias 等人，2024；Ben-Eliezer 等人，2022a；Woodruff 和 Zhou，2022；Cherapanamjeri 等人，2023）。一种常见的防御方法是净化输出，例如通过舍入或添加噪声；这类方法通常借鉴差分隐私技术，以确保输出不会泄露太多有关算法内部状态的信息（Hassidim 等人，2022；Attias 等人，2024；Beimel 等人，2022）。然而，这些技术并不能直接应用于搜索问题。在搜索问题中，算法必须从给定数据集中返回一个特定元素。直接输出原始数据点可能会泄露大量信息，而且显然无法在不违反问题核心约束——即返回一个有效的数据集元素——的情况下添加噪声或以其他方式模糊输出。

或许最基本的搜索问题是近似最近邻（*ANN*）搜索，其应用十分广泛，涵盖从数据压缩、机器人学到 DNA测序、异常检测和检索增强生成（RAG）等领域（SantaLucia 等人，1996；Kalantidis 和 Avrithis，2014；Ichnowski 和 Alterovitz，2015；Verstrepen 和 Goethals，2014；Tagami，2017；Bergman 等人，2020；Han 等人，2024；Kitaev 等人，2020）。形式化地，我们将问题定义如下：

定义 *1.1*（$(c,\ r)$-近似最近邻问题）。给定度量空间 $(\mathcal{M},\ ||\cdot||)$ 中的一个包含 n 个点的数据集 S 以及半径 $r>0$，令 $B_S(q,\ r) := \{p \in S:\ ||p-q|| \le r\}$。给定查询点 $q \in \mathcal{M}$ 和近似参数 $c \ge 1$，目标是构建一个数据结构，当 $B_S(q,\ r) \neq \emptyset$ 时，能够在 $B_S(q,\ cr)$ 中找到一个点。如果 $B_S(q,\ cr) = \emptyset$，则算法需要返回 $\perp$。除查询之外，数据集 S 本身还可能通过添加或删除点进行无感知更新。

用于解决此问题的数据结构，其空间复杂度、查询运行时间和更新运行时间均表示为数据集大小 n、参数 c 以及可能的查询总数的函数。

要实现次线性查询时间和近线性空间这一理想权衡，迄今为止很大程度上只能依靠随机化。事实上，近似最近邻最突出的一类算法基于局部敏感哈希（LSH），该方法在非自适应设置下经历了长期且富有成果的研究（Gionis 等，1999 年；Jafari 等人，2021；Andoni，2009；Andoni 等人，2018、2017b、2016；Andoni 和 Indyk，2017；Andoni 等人，2017a；Indyk 和 Motwani，1998 年；Broder 等人，1998；Andoni 和 Beaglehole，2022）。依赖 LSH 的近似最近邻算法的查询时间复杂度为 $\widetilde{O}(dn^{\rho})$*，空间复杂度为 $\widetilde{O}(n^{1+\rho})$，其中 d 是 $\mathcal{M}$ 的维度，而 $\rho = \rho(c) \in (0,1)$ 是取决于 c 和 LSH 构造† 的固定常数。

*我们使用 $\widetilde{O}$ 记号来隐藏多对数因子。

†例如，当 $\mathcal{M} = \{0,1\}^k$ 采用汉明距离且 $c \ge 1$ 是近似参数时，Andoni 和 Razenshteyn (2015) 给出的最先进构造可达到 $\rho = \frac{1}{2c-1}$。我们将 ρ 和 $\rho(c)$ 互换使用。

这些经典随机化结构的脆弱性最近由 Kapralov 等人 (2024) 揭示，他们对标准 LSH 数据结构实施了一种攻击。他们证明，自适应对手可以利用多对数数量的查询，了解足够多的内部哈希函数信息，从而迫使算法失效。受其工作的启发——他们的工作依赖于数据集的某些结构属性（例如，一个“孤立”的点）——我们考虑一种强大的对抗模型，其中对手同时选择数据集和查询序列。我们研究以下问题：

Can 像近似最近邻这样的问题能否在面对对抗性查询时高效求解 $?$

我们针对（c，r）-近似最近邻问题的对抗模型赋予对手不受限制的计算资源、完全指定数据集 S 的能力、无视算法内部状态而选择该数据集更新序列的能力，以及或许最重要的能力：根据其与算法交互的完整历史来选择每次查询。尽管还可以考虑其他对抗模型，但我们所选择的框架最大限度地提高了普适性，旨在证明即使在广泛的约束条件下，高效的数据结构仍然可以实现。

1.1 路线图与讨论

我们首先将我们的结果置于对抗鲁棒性的更广泛背景下；该领域主要遵循两种范式。第一种是基于隐私的鲁棒化（Hassidim 等人，2022），即使用差分隐私机制聚合算法的多个独立副本。隐私机制会混淆内部随机性，从而阻止自适应对手预测未来行为。第二种是对所有输入均成立的算法（Gilbert 等人，2007），这类算法对每个输入都能成功，但通常会因离散化和并集界而产生较高的运行时间或空间开销，因此具有问题特定性。

近似最近邻问题还带来了另一项挑战：它是搜索问题，而非估计问题，因此向输出添加噪声会破坏正确性。解决这一问题的自然方法，是将搜索问题归约为一系列判定问题，并对这些问题进行基于隐私的鲁棒化。在近似最近邻问题中，对数据集进行划分，并在每个部分中求解弱判定问题，就能在保持次线性查询时间和空间的同时，鲁棒地找回一个有效的近邻。我们认为，将搜索问题归约为鲁棒判定问题这一思路，或许也适用于其他搜索问题。

我们的主要概念性贡献是指出了实现鲁棒性的第三条路径：公平性。简而言之，公平算法不会偏向特定输出，并能在重复查询之间保持统计独立性。这类算法本质上具有鲁棒性，因为它们泄露的有关内部随机性的信息很少。我们将这一直觉形式化，表明公平性可以成为实现鲁棒性的有力工具。从这一视角看，公平性和差分隐私可被视为同一基本原则的两种体现：

鲁棒性源于相对于内部随机性的稳定性。

最后，这些方法可以结合起来。通过将基于公平性的鲁棒性与基于隐私的鲁棒化相结合，并利用近似最近邻的几何结构构造同心环带，我们在鲁棒最近邻搜索中实现了更优的权衡。

1.2 我们的结果与技术

本节将详细介绍我们的技术见解

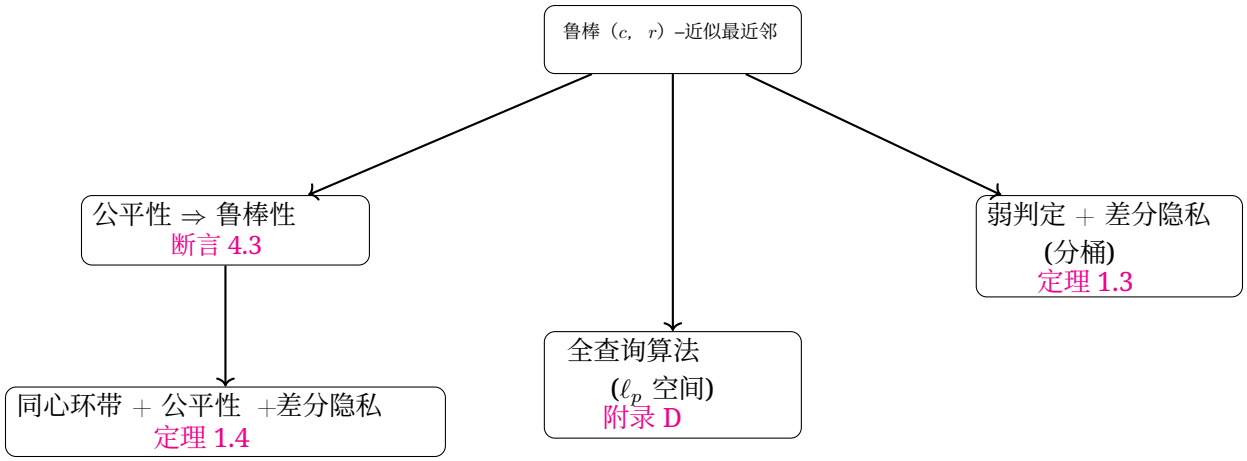

图 1：结果路线图。

鲁棒性与公平性我们首先发现了鲁棒性与公平性之间的联系。公平近似最近邻算法会从有效近邻候选点集合中均匀随机地输出一个点。Aumüller 等人Aumüller 等人(2022）在局部敏感哈希（LSH）的背景下对这类算法进行了严格研究，也研究了近似公平性的概念。我们证明，只需调用一个用于精确公平近似最近邻问题的算法，即可解决鲁棒近似最近邻问题。

定理 *1.2.* 设 $n(q,r) := |B_S(q,r)|$ 为以 $q \in \mathcal{M}$ 为中心的 r-球的 S-密度。存在一种对抗鲁棒的 (c,r)-近似最近邻算法，其使用 $O(n^{1+\rho(c)}\log(nQ))$ 位空间，且每次查询耗时为 $O(d\cdot(n^{\rho(c)}+n(q,cr)\,n(q,r)+1)\log(nQ))$。

尽管这种算法的空间效率很高，但其查询复杂度不幸取决于 D 位于 cr-球与 r-球之间的点的密度比，该球由一次查询 q确定。对手可以构造一个数据集，使这一比值很大，从而严重降低性能。Feng 等人 提出的算法也有这一缺点（(2025），不过它依赖于密度$s := n(q, cr)$，该值严格大于 D （见 表 1 ）。

备注。 公平性与鲁棒性之间的联系并非仅限于近似最近邻。从这一视角来看，公平性并非只是“锦上添花”的属性，而是与安全性密不可分。

通过分桶进行无假设搜索 为减轻数据依赖性，我们提出了一种元算法，将搜索问题归约为弱判定问题。在这个问题中，正实例对应于存在 r与查询距离不超过 q的邻居，而负实例则表明不存在距离不超过 cr的邻居。只需使用经典 LSH 数据结构 $\mathcal{D}$，即可无偏地解决这种弱判定问题。与搜索问题不同，通过应用 (Hassidim 等人，2022) 提出的知名差分隐私（DP）混淆技术，可以使无偏判定器具备鲁棒性：我们维护 $\sqrt{Q}$ 个副本 $\mathcal{D}$ ，并以一种保护每个副本随机比特隐私的方式合并它们的响应。

为了执行搜索，我们接着将 $S \in \mathcal{M}^n$ 划分为大小约为 $\sqrt{n}$的桶，并在每个桶中实例化一个鲁棒弱判定器的副本。我们可以利用这些副本找出一个包含合适输出点的桶，然后在该桶中进行穷举搜索，得到最终答案：

定理 1.3。 存在一种对抗鲁棒算法，可成功回答至多 Q 个 (c,r)-近似最近邻问题查询，成功概率至少为 $1-\Theta(\delta)$。该算法使用 $\widetilde{O}(n^{1+\rho/(2-\rho)}\sqrt{Q})$ 空间，每次查询的时间为 $\widetilde{O}(dn^{1/(2-\rho)})$，其中 $\rho=\rho(c) \in (0,1)$。

度量指标	查询时间	空间	更新时间
定理 1.2 (公平性)	$\widetilde{O}(d \cdot (D + n^{\rho}))$	$\widetilde{O}(n^{1+\rho})$	$\widetilde{O}(d \cdot (D + n^{\rho}))$
定理 1.3 (分桶)	$\widetilde{O}(dn^{\frac{1}{2-\rho}})$	$\widetilde{O}(\sqrt{Q} \cdot n^{\frac{2}{2-\rho}})$	$\widetilde{O}(dn^{\frac{\rho}{2-\rho}}\sqrt{Q})$
定理 1.4 (同心环带) $\beta = \Theta(\frac{\log \text{对数 } c}{\text{对数 } c})$	$\widetilde{O}(dn^{\beta})$	$O(\sqrt{Q} \cdot n^{1+\beta})$	$\widetilde{O}(dn^{\beta}\sqrt{Q})$
(Feng 等人，2025)[‡]	$O\left(d \cdot s \cdot n^{\rho}\right)$	$O(\sqrt{Q} \cdot s \cdot n^{1+\rho})$	$\widetilde{O}(dn^{\rho} \cdot s \cdot \sqrt{Q})$

表 1：配备局部敏感哈希的度量空间中 $(c,\ r)$-近似最近邻问题的算法。

借助同心 LSH 环带打破 $\sqrt{n}$ 屏障 分桶方法得到的查询时间复杂度始终至少为 $O(\sqrt{n})$。考虑到通过增大近似参数，局部敏感哈希方法可以使 n 的指数任意接近 0，这并不理想。为突破这一障碍，我们提出一种同心环带构造。

我们将 $(r,\ cr)$-环带划分为若干更小的同心子环带，并在每个子环带中应用公平近似最近邻算法 $\mathcal{A}_{\text{fair}}$。简单的计数论证保证，这些子环带中至少有一个具有较低的点密度比，这意味着 $\mathcal{A}_{\text{fair}}$ 在该环带中的运行时间以至少 0.9 的概率不超过可接受的运行时间阈值。对于每个未超过此阈值的环带，我们估计相应$\mathcal{A}_{\text{fair}}$ 副本快速终止的概率。随后，我们选取一个有利的环带，运行一个留出测试副本 $\mathcal{A}_{\text{fair}}$ 并输出一个点。然而，为了保持鲁棒性，我们必须谨慎选择环带，确保自适应对手无法针对我们的内部随机性进行过拟合。为此，我们将差分隐私鲁棒化框架应用于环带选择过程，估计每个 $\mathcal{A}_{\text{fair}}$ 副本在预设的运行时间阈值内终止的概率。我们的算法无假设，且其运行时间优于 $O(\sqrt{n})$。

定理 1.4。存在一种鲁棒算法，可解决(c, r)-近似最近邻问题，其空间复杂度为 $\widetilde{O}(\sqrt{Q} \cdot n^{1+\beta})$，其中 $\beta = \min_{k \in \mathbb{Z}_{\geq 1}} \max\{\rho(c^{1/k})\ 1/k\}$。查询运行时间为 $\widetilde{O}(dn^{\beta})$ 时间，成功概率至少为 $1 - \delta$。

对于许多度量空间，β 的值可以得到简洁的表达。对于汉明距离下的超立方体 $\{0,1\}^d$，有 $\rho(c) = \frac{1}{2c-1}$，由此得到 $\beta = \Theta(\frac{\log\log c}{\log c}) \to 0$，当 $c \to \infty$时成立；而指数为 $2-\rho(c)$ 时则不成立，这与定理 1.3的情况不同。作为推论，这项技术还使我们能够针对一类“宽松”的公平近似最近邻问题，实现纯次线性时间。

全查询算法 对于低维度量空间，我们开发了用于近似最近邻的算法，并提供强大的全查询保证：以高概率，数据结构能够正确回答每一个可能的查询 $q \in \mathcal{M}$。我们的方法建立在应用于局部敏感哈希数据结构的离散化技术之上，这一范式已在先前工作中得到探索（Cherapanamjeri 和 Nelson，2020，2024；Bateni 等人，2024）。我们通过引入一种新颖且更简单的度量覆盖构造来改进这一研究方向，将空间复杂度降低一个对数因子，并通过采样将时间复杂度提高 d倍。下面我们给出汉明超立方体上的结果，其中还包括 ℓ_p 空间上的结果，详见附录 D。

[‡]关于Feng 等人(2025)同期研究了鲁棒近似最近邻问题。我们在第 2 节中将我们的结果与他们的算法进行了比较，并对相关工作进行了更深入的讨论。

定理 1.5。 对于 d 维汉明超立方体 $\{0,1\}^d$ 中的 (c,r)-近似最近邻问题，存在一种算法，能够以至少 0。99 的概率正确回答所有可能的查询。空间复杂度为 $\widetilde{O}\left(d \cdot n^{1+\rho+o(1)}\right)$，查询时间为 $\widetilde{O}\left(d \cdot n^{\rho}\right)$，其中 $\rho = \frac{1}{2c-1}$。

备注（全查询算法的代价）。尽管全查询算法具有显著的保证，但也存在明显的缺点。它们的空间复杂度会扩大 d倍，因此对于高维度量空间而言难以处理。这是为了确保任意查询发生错误的概率极低而需要大量哈希函数的直接结果。此外，这些算法不具备其自适应对应算法的通用性；它们依赖于度量空间，必须针对所使用的特定度量空间进行定制。

2 相关工作与同期工作

设计能够抵御对抗性查询的算法一直是一个受到广泛研究的挑战，尤其是在隐私和统计学领域（Bassily 等人，2015；Smith，2017；Bassily 等人，2016）。在这些领域中，差分隐私是确保鲁棒性的核心工具（Dwork 等人，2015a；Dinur 等人，2023 年）。Ben-Eliezer 等人（Ben-Eliezer 等人，2022b）正式将对抗鲁棒性问题引入流算法，其动机源于针对线性草图的攻击（Hardt 和 Woodruff，2013），此后又推动了大量关于增强各种流算法鲁棒性的研究（Hassidim 等人，2022；Chakrabarti 等人，2021；Lai 和 Bayraktar，2020；Chakrabarti 和 Stoeckl，2024；Stoeckl，2023；Woodruff 和 Zhou，2022；Ben-Eliezer 等人，2022a）。

我们的工作最直接地受到 Hassidim 等人（Hassidim 等人，2022）提出的框架启发。他们利用差分隐私鲁棒地解决估计问题；此外，Cherapanamjeri 等人（Cherapanamjeri 等人，2023 年）将这一框架应用于查询时间开销较低的情形。尽管我们采用了类似的方法，他们的方法从根本上仅适用于估计问题，无法扩展到最近邻搜索等搜索问题，因为这类问题的输出必须是某个特定的数据集元素。Beimel 等人（Beimel 等人，2022）确立的下界进一步凸显了鲁棒搜索的难度：这些下界表明，某些搜索问题的鲁棒算法本质上比其非自适应对应算法更慢，这也促使我们开展本研究。

其他研究进一步凸显了鲁棒搜索面临的独特挑战。例如，鲁棒图着色也需要采用超越简单噪声添加的技术，因为其输出空间是离散的（Chakrabarti 等人，2021；Behnezhad 等人，2025）。我们的方法也不同于拉斯维加斯局部敏感哈希构造（Pham 和 Pagh，2016 年；Wei，2022）。这些方法虽然保证不存在假阴性，但仍容易受到对手影响，因为对手可以抬高其期望运行时间（Kapralov 等人，2024）。相比之下，我们着重于使传统蒙特卡洛算法具备鲁棒性。

最后，我们的方法基于离散化和基于网的论证，为近似最近邻提供“对所有”保证。这项技术此前曾用于鲁棒距离估计（Cherapanamjeri 和 Nelson，2020）、结合划分树的近似最近邻（Cherapanamjeri 和 Nelson，2024）以及高效质心链接聚类（Bateni 等人，2024）。我们提出了一种更简单、更精简的构造方法，相较于以往工作，性能略有提升。

2.1 与Feng 等人 (2025)的比较

我们的工作与Feng 等人（2025）同期独立开展。我们的方法、假设和性能保证存在显著差异。

方法论 Feng 等人（2025）通过归约到私有选择问题，提出了一种与差分隐私噪声结构紧密耦合的方法。相比之下，我们的“搜索到决策”和公平性框架更具通用性，将差分隐私组件视为黑盒原语。

假设 他们的算法复杂度取决于近邻密度界 s，其中$|B_S(q, cr)| \leq s$。我们提出了首批查询运行时间独立于输入数据集的算法，使其能够应对最坏情况数据分布。

性能 他们的查询时间随环带中点的数量$|B_S(q, cr)|$呈乘法 级增长，而我们的算法要么完全是次线性的，要么其查询时间仅加法地 取决于密度比 $D = \frac{|B_S(q\ cr|,)(,)}{|B_S\ q\ r|}$。关键在于，这种对 D 的依赖不会影响我们的空间复杂度，其增长仍仅额外增加一个 $\sqrt{Q}$ 因子。

3 预备知识

本节介绍我们所研究问题的必要定义，以及所考虑的计算模型。如需查阅完整的符号表，请参见表 2。有关差分隐私定理和概念的背景知识，请参见 附录 B。

对抗鲁棒性模型 如果一个算法能够以高概率正确回答一系列自适应选择的查询，则称该算法具有对抗鲁棒性。此概念通过以下交互式博弈形式化（Ben-Eliezer 等人，2022b）：

定义 3.1. 考虑以下由 算法 $(\mathcal{A})$ 与 对手 $(\mathcal{B})$ 进行的博弈 $\mathcal{G}$：

1. 设置阶段： 对手选择一个数据集 S。算法 $\mathcal{A}$ 随后使用其私有内部随机性 $R_{setup} \in \{0,1\}^*$ 对 S 进行预处理并构建一个数据结构 D。对手可以知道 $\mathcal{A}$ 的代码，但不知道 R_{setup}的具体实例。

2. 查询阶段： 博弈进行 Q 轮。在每一轮 $i \in [Q]$中：
   - 对手自适应地选择一个查询 q_i。这一选择可以取决于数据集 S 以及此前所有查询及其对应答案的历史记录，$(q_1, a_1) \ldots, (q_{i-1}, a_{i-1})$。
   - 算法 $\mathcal{A}$ 使用其数据结构 D，并可能使用新的私有随机性 $R_i \in \{0,1\}^*$，来计算并返回答案 a_i。

3. 获胜条件： 如果存在至少一轮 $i \in [Q]$，使得答案 a_i 不是对查询 q_i的正确回应，则算法失败。

如果对于任意数据集和对手可能采用的任意策略，算法失败的概率至多为 δ，我们就称算法 $\mathcal{A}$ 是 δ-对抗鲁棒的。该概率是针对算法的全部内部随机性（R_{setup}、R_1、……、R_Q）而言的。

备注（无感更新）。我们的框架也支持对数据集进行无感更新。对手会预先选择一系列更新时间步，以添加或删除点的形式进行。这些更新会在博弈过程中与查询阶段交替进行。我们简要讨论如何在算法中高效实现更新；由于我们的算法适用于任意选取的数据集，因此我们不考虑鲁棒性 S。

近似最近邻搜索与局部敏感哈希 在最近邻 问题中，我们的目标是在输入数据集中找到一个点，使其与某个查询点之间的距离最小。

定义 *3.2*（近似最近邻）。设 $c > 1$ 和 $r > 0$ 为正常数。在 (c, r)–近似最近邻问题（近似最近邻）中，输入为一个集合 $S \subset \mathcal{M}$，满足 $|S| = n$，以及一系列查询 $\{q_i\}_{i=1}^{Q}$，满足 $q_i \in \mathcal{M}$。对于每个查询 q_i，若存在 $p \in B_S(q_i, r)$，则要求输出某个点 $p' \in B_S(q_i, cr)$。若 $B_S(q_i, cr) = \emptyset$，则要求输出 $\perp$。若 $B_S(q_i, r) = \emptyset \neq B_S(q_i, cr)$，则可以输出一个来自 $B_S(q_i, cr)$ 的点，或输出 $\perp$。我们的算法应以至少 $2/3$ 的概率成功满足这些要求。

解决近似最近邻问题的一种常用方法是局部敏感哈希（局部敏感哈希）：

定义 3.3（局部敏感哈希，Har-Peled 等人，2012）。函数哈希族 $\mathcal{H}$，其将 $\mathcal{M}$ 映射到一组桶，称为（c，r，p_1，p_2）–局部敏感哈希族（***LSH***），当且仅当满足以下两个条件：

- 如果 $x, y \in \mathcal{M}$ 具有 $||x - y|| \leq r$，那么 概率$_{h\in\mathcal{H}}[h(x) = h(y)] \geq p_1$。
- 如果 $x, y \in \mathcal{M}$ 具有 $||x - y|| \geq cr$，那么 概率$_{h\in\mathcal{H}}[h(x) = h(y)] \leq p_2$。

其中 $p_1 \gg p_2$ 是 $(0, 1)$ 内的参数。我们通常假设，在 d 维度的度量空间中计算 h 需要 $O(d)$ 时间。我们假设所考虑的局部敏感哈希构造是单调的，这意味着 概率$[h(x) = h(y)]$ 会随着 $||x - y||$ 增大而单调递减。

给定度量空间上的 (c, r, p_1, p_2)–局部敏感哈希，我们可以通过放大局部敏感哈希的保证来解决 (c, r)–近似最近邻问题。具体而言，我们采用“与的或”（OR of ANDs）构造：采样 $L := p_1^{-1} = n^{\rho}$ 个哈希函数 $h_1, ..., h_L$，通过将 $\mathcal{H}$ 中 $\rho(c) \in (0, 1)$ 个“原型”局部敏感哈希函数的输出进行拼接来构造它们，如（Indyk 和 Motwani，1998 年）所示。

定理 3.4。如果一个 d–维度度量空间允许存在一个 (c, r, p_1, p_2)–局部敏感哈希族，那么我们可以在该空间上使用 $O(n^{1+\rho})$ 空间和每次查询 $O(dn^{\rho})$ 时间来解决 (c, r)–近似最近邻问题，其中 $\rho = \log(1/p_1)\log(1/p_2)$。

4 公平性蕴含鲁棒性

我们首先建立鲁棒性与公平性之间的联系。我们对 Aumüller 等人 (2022) 给出的近似最近邻问题中的公平性定义进行完善，以便证明后续结论。

定义 4.1（精确公平 (c, r)-近似最近邻）。若在截至目前返回的所有答案均正确的条件下，对每一轮 i 和每个交互记录 T_{i-1}，以下条件均成立，则称一个数据结构解决了精确公平 (c, r) 近似最近邻问题：

1. 若 $B_S(q_i, r) \neq \emptyset$，则条件分布 $\mathcal{L}$ 为均匀分布：

$$\mathcal{L}(a_i \mid T_{i-1}, q_i) = \mathrm{Unif}(B_S(q_i, r)).$$

否则，算法输出 $\perp$。

2. 条件分布 $\mathcal{L}(a_i \mid T_{i-1}, q_i)$ 不依赖于预处理随机性 R_{setup}。

我们首先证明，Aumüller 等人（2022）的精确公平近似最近邻算法符合我们的定义。他们使用标准的 (r, cr, p_1, p_2)-局部敏感哈希族，其中 $\rho = \log(1/p_1)/\log(1/p_2)$。该数据结构由 $L = \Theta(n^{\rho}\log(nQ))$ 个独立哈希表组成，空间开销为 $\widetilde{O}(n^{1+\rho})$。给定查询 q 后，算法从所有哈希表中收集候选集 $C(q)$，并应用 Aumüller 等人（2022 年，第 3.5 节）提出的精确邻域采样过程。该过程使用拒绝采样和全新的逐查询随机性。

定理 4.2。 存在一种随机化的 (c, r)-近似最近邻数据结构，其空间开销为 $O(n^{1+\rho}\log(nQ))$，并具有以下性质。对于任意固定且独立于 R_{setup} 的查询 q，以至少 $1-\frac{1}{nQ}$ 的概率（关于 R_{setup}），候选集 $C(q)$ 包含 $B_S(q,r)$ 中的所有点。在给定此事件的条件下，算法返回 $B_S(q,r)$ 中的一个元素，该元素服从精确均匀分布，且其分布独立于 R_{setup}。此外，q 的期望查询时间有如下界：

$$O\left(d\cdot\left(n^{\rho}+\frac{n(q,cr)}{n(q,r)+1}\right)\cdot\log(nQ)\right).$$

证明。 固定任意一个与 R_{setup} 独立的查询 q。对于任意 $p\in B_S(q,\ r)$，p 在所有 L 张表中均未能与 q 发生碰撞的概率至多为 $(1-p_1)^L$。对至多 n 个数据点应用并集界，由我们对 L 的选择可得

$$\Pr_{R_{\text{setup}}}[\exists p\in B_S(q,r)\notin C(q)]\le n(1-p_1)^L\le\frac{1}{nQ}.$$

令 $\mathcal{E}_q$ 表示互补事件，即 $B_S(q,\ r)\subseteq C(q)$，并以 $\mathcal{E}_q$ 为条件。

以 $\mathcal{E}_q$ 为条件时，候选集 $C(q)$ 包含 q 的整个 r-邻域。随后，查询算法对候选集 $C(q)$应用 Aumüller 等人 (2022，第 3.5 节)提出的精确邻域采样过程。根据该过程的正确性，返回的点在 $B_S(q,\ r)$上严格服从均匀分布。

此外，决定输出的随机选择（拒绝采样所用的随机数和随机交换）均在查询时重新生成，且独立于预处理随机性 R_{setup} 和此前的任何交互记录。由于候选集 $C(q)$ 本身仅取决于 q 和 R_{setup}，因此，以 $\mathcal{E}_q$ 为条件，对于任意交互记录 T，条件分布 $\mathcal{L}(a\mid T,\ q)$ 在 $B_S(q,\ r)$ 上严格服从均匀分布，且不依赖于 R_{setup}。 □

令 $\mathcal{A}_{\text{fair}}$ 为定理 4.2给出的精确公平近似最近邻算法。现在我们证明公平性蕴含对抗鲁棒性。直观来说，在给定截至第 $i-1$ 轮均成功的条件下，对手的下一次查询与预处理随机性之间仍保持独立，因此我们可以在每一轮应用逐查询的局部敏感哈希保证。

断言 4.3（公平性蕴含鲁棒性）。算法 $\mathcal{A}_{fair}$ 对于 Q次自适应查询具有$\frac{1}{n}$ 对抗鲁棒性。

证明。不失一般性，我们可以假设对手是确定性的。令 F_i 表示算法在第i轮失败的事件（即候选集未包含完整的 r球），令 $S_i=\bigcap_{j=1}^{i}\neg F_j$ 表示截至第 i轮均成功的事件。

我们首先对 i使用归纳法证明：在给定 S_{i-1}的条件下，交互记录 T_{i-1} =(q_1，a_1、.....、q_{i-1}，a_{i-1})与预处理随机性 R_{setup}相互独立。当 $i=1$时，交互记录为空，结论显然成立。假设结论对 $i-1$ 成立。在给定 S_{i-1}的条件下，精确公平性意味着，给定 (T_{i-2}，q_{i-1})时，a_{i-1} 的分布不依赖于 R_{setup}。因此，将交互记录从 T_{i-2} 扩展到 T_{i-1} 后，它仍与 R_{setup}相互独立。由于对手是确定性的，下一次查询 $q_i=\mathcal{B}(T_{i-1})$是 T_{i-1} 的函数，因此在给定 S_{i-1}的条件下，也与 R_{setup} 相互独立。

现在，我们可以将定理 4.2 的逐查询保证应用于 q_i，得到

$$\Pr[F_i\mid S_{i-1}]\le\frac{1}{nQ}.$$

最后，根据序贯事件的并集界，

$$\Pr\left[\bigcup_{i=1}^{Q} F_i\right] \leq \sum_{i=1}^{Q} \Pr[F_i \mid S_{i-1}] \leq \frac{1}{n}.$$

□

对数据集的更新 S 公平近似最近邻数据结构也支持无感动态更新。插入操作通过使用固定的预处理哈希函数，将新点哈希到每个 L 局部敏感哈希表中来实现。删除操作则通过将该点从相应的桶中移除来实现；这可以通过维护桶指针，或用该点本身查询数据结构并将其从所得候选列表中删除来完成。每次更新都会访问相同的一组哈希表，其渐近开销与一次查询相同。因此，期望更新时间与期望查询时间仅相差常数因子。

这两种操作仅依赖预处理随机性，不使用依赖于交互记录的随机性。此外，精确邻域采样过程保持不变，仍会为每次查询使用新的逐查询随机性。因此，在任意一系列无感更新下，精确公平性和鲁棒性分析依然成立。

5 通过分桶实现无假设鲁棒搜索

公平算法的一个主要局限是，只有当数据集不会导致过高的密度比时，它才能高效运行；如果 S 由对手选取，这一点就无法保证。理想情况下，我们希望获得无需对输入数据集作任何假设即可运行的次线性算法。为此，我们引入一个搜索到判定框架：

5.1 弱判定近似最近邻

定义 5.1（弱判定近似最近邻）。考虑度量空间 $\mathcal{M}$ ，设 $S \subseteq U$ 及$|S| = n$ 为输入点数据集。设 $r > 0$、 $c > 1$ 为两个参数， $q \in \mathcal{M}$ 为自适应选择的查询。如果 $B_S(q, r) \neq \emptyset$，则应回答***1***。如果 $B_S(q, cr) = \emptyset$，则必须回答***0***。其他情况下，任何回答都可以接受。令 $D(S, q, c, r) \subseteq \{0, 1\}$ 表示针对数据集 S、参数 c、 r 和查询 q，此弱判定问题的正确答案集合。

设 $\mathcal{A}$ 是一个用于解决弱判定近似最近邻问题的算法，但不一定具有鲁棒性。我们可以利用 $\mathcal{A}$ 设计一个对抗稳健判定算法_ $\mathcal{A}_{\text{dec}}$ ，而空间仅增加 $\sqrt{Q}$倍。遵循 Hassidim 等人（2022）的框架，我们维护由 $\mathcal{A}$ 使用 L 个独立随机串生成的 $\sim L = \Theta(\sqrt{Q})$ 个数据结构 $\mathcal{D}_1$、.....、 $\mathcal{D}_L$ 的副本；随后，对于每个查询 q ，我们以私有方式合并 $\mathcal{A}$ 的答案。与原始的 Hassidim 等人（2022）框架不同，我们无需使用私有中位数算法，这简化了分析。为了减少查询时间，我们对每次查询使用通过子采样实现的隐私放大，从而只需采样 $k \ll L$ 个副本（定理 B.8）。我们的算法如下所示，见算法 1。

为了分析该算法，我们论证，对于所有 $i \in [Q]$，即使存在对抗性生成的查询， k 个答案中至少有 8 10 是正确的。为此，我们首先需要证明，关于用于指定所选局部敏感哈希函数的输入随机字符串 R ，该算法满足 (ε, δ)-差分隐私，其中 δ 是期望的失败概率， ε 是经过适当选取的常数。因此，如果我们设定 $L = \Theta(\varepsilon^{-1}\log^{1.5}(1/\delta)\cdot\sqrt{2Q})$ 和 $k = \Theta(\log(Q/\delta))$，就能得到一个鲁棒判定器，其成功概率至少为 $1 - \Theta(\delta)$。我们的分析（定理 C.1）

算法 1 鲁棒判定器 $\mathcal{A}_{\text{dec}}$（基于无偏判定器 $\mathcal{A}$）

输入： 随机字符串 $R = \sigma_1 \circ \sigma_2 \circ \cdots \sigma_L$。
参数：查询数量 Q、副本数量 L、采样索引数量 k。
从对手处接收输入数据集 $S \subseteq U$，其中 $n = |S|$。
初始化 $\mathcal{D}_1$、……、$\mathcal{D}_L$，其中 $\mathcal{D}_i \leftarrow \mathcal{A}(S)$，基于随机字符串 σ_i。
对于 $i = 1$ 到 Q，**执行**
    从对手处接收查询 q_i。
    $J_i \leftarrow$ 在 $[L]$ 中有放回地抽取 k 个索引。
    令 $a_{ij} \leftarrow \mathcal{D}_j(q_i) \in \{0,1\}$ 和 $N_i := \frac{1}{k}|\{j \in J_i \mid a_{ij} = 1\}|$。
    Let $\widehat{N}_i = N_i +$第$\left(\frac{1}{k}\right)$层。
    输出 $\mathbb{1}[\widehat{N}_i > \frac{1}{2}]$

全文详见附录 C。该分析运用差分隐私的技术工具，将 Ben-Eliezer 等人 (2022b) 的鲁棒化论证应用于我们的设定。我们与他们分析的主要区别在于，我们可以避免使用黑盒私有中位数算法；拉普拉斯机制足以满足我们的需求。

5.2 基于分桶的搜索

要执行最终搜索，我们将点数据集 S 划分为 $n^{1-\alpha}$ 个分段，其中 $\alpha \in (0, 1)$。然后，在每个分段中各实例化一个 $\mathcal{A}_{\text{dec}}$ 的副本 $\mathcal{A}_i \equiv \mathcal{A}_{\text{dec}}$。当收到查询时，我们将其转发给每个 $\mathcal{A}_i$；如果某个分段回答**1**，则在该分段中执行穷举搜索，以找到一个属于 $B_S(q, cr)$ 的点。

算法 2 通过弱判定和分桶实现鲁棒近似最近邻

参数：错误概率 $\delta > 0$，查询数量 Q
将点集 S 任意划分为 $\kappa = n^{1-\alpha}$ 个分段 L_1，……，L_κ，每个分段大小为 n/κ。
初始化 κ 个 $\mathcal{A}_{\text{dec}}$的独立副本 $\mathcal{A}_1$，……，$\mathcal{A}_\kappa$，每个副本具有 $\delta' = \delta/\kappa$
对 $i = 1$ 到 Q **执行**
    从对手处接收查询 q_i。
    对 $j = 1$ 到 κ，其中 $A_j(q_i) = 1$ **执行**
        对 $p \in L_j$ **执行**
            如果 $p \in B_S(q, cr)$，**则**
                输出 p并继续处理下一个查询。
    输出 $\perp$ 并继续下一个查询。

引理 5.2。 算法 2 是一个 δ-对抗鲁棒算法，用于近似最近邻问题。

证明。 假设对手已经获胜。只有当所有数据结构都回复**0**时，算法才会出错，尽管存在一个点 $p \in B_S(q, r)$。考虑线段 L_i，其中 $p \in L_i$，并单独对其进行分析。由于算法 $\mathcal{A}_{\text{dec}}$ 的所有副本彼此独立地初始化，对手成功诱使 $\mathcal{A}_i$ 出错；根据假设，通过并集界，该事件在所有线段上的概率至多为 $\kappa\delta' = \delta$。 □

现在我们开始证明定理 1.3：

定理 *1.3* 的证明定理 *1.3*。我们创建 $n^{1-\alpha}$ 条线段，每条包含 n^{α} 个点。回顾一下，单个算法 $\mathcal{A}_{\text{dec}}$ 需要 $\widetilde{O}(n^{1+\rho}\sqrt{Q})$ 的预处理时间和空间，每次查询需要 $\widetilde{O}(n^{\rho})$ 的时间和空间。每个副本 $\mathcal{A}_i$ 运行于 n^{α} 个点，因此在预处理阶段，我们的算法使用

$$\widetilde{O}\left(n^{1-\alpha}(n^{\alpha})^{1+\rho}\sqrt{Q}\right)=\widetilde{O}\left(n^{1+\alpha\rho}\sqrt{Q}\right)$$

用于创建 $n^{1-\alpha}$ 个副本 $\mathcal{A}_1$、……、$\mathcal{A}_L$ 所需的空间位数。另一方面，处理单次查询时，算法使用

$$\widetilde{O}\left(n^{1-\alpha}\cdot(n^{\alpha})^{\rho}+n^{\alpha}\right)=\widetilde{O}\left(n^{1-\alpha+\alpha\rho}+n^{\alpha}\right)$$

的时间。为平衡查询复杂度项中的各个加数，我们设

$$n^{1-\alpha+\alpha\rho}=n^{\alpha}\implies\alpha=\frac{1}{2-\rho}$$

这就证明了定理 1.3并完成了对算法2的分析。 □

我们算法最大的优势是不对输入数据集作任何假设。然而，当 $\rho<1$ 时，它可实现次线性查询时间 $n^{\frac{1}{2-\rho}}<1$。此外，当 Q 的取值较小时，我们算法的空间复杂度优于空间复杂度为 $n^{1+\rho}$ 的无视近似最近邻算法。

在我们的框架中，更新也很容易实现：在我们维护的每个 LSH 副本中执行插入和删除操作~~。我们构建 $L=O(\sqrt{Q})$ 个局部敏感哈希数据结构，每个结构都需要更新 $n^{\rho/(2-\rho)}$ 个哈希函数，由此得到最终的更新时间复杂度。我们还需要更新桶，但由于数据集上的划分可以任意进行，因此每次只需更新一个桶。为了提高实际运行效率，我们可以延迟执行更新，只在即将使用数据结构时才对其进行修改。

6 通过 ConcentricLSH 环带实现的放宽公平 ANN

作为预热，我们首先对公平 ANN 的定理 4.2 提出一种算法改进，移除对比值 $n_{(q,\ cr)}/n_{(q,\ r)}$ 的依赖，该比值在查询时间中可能增长到 n。对于宽松公平性保证，我们实现了纯次线性时间：

定义 6.1（ANN 中的放宽公平性）。设 S 为输入数据集，$q\in\mathcal{M}$ 为查询点。若 $B_S(q,r)\neq\emptyset$，则算法旨在从 $B_S(q,r')$中输出一个点，该点独立于先前查询，且均匀随机选取，其中 $r'\in[r,cr]$ 是取决于 q和 S的随机变量。否则，若 $B_S(q,r)=\emptyset$，算法可以回答 $\perp$，或以 $r'\in(r,cr]$的概率从 $B_S(q,r')$ 中输出一个均匀随机点。

考虑 r 与 cr 之间的以下半径序列，各半径交错排列，使得任意两个相邻半径之比保持不变：$r_0=r$，r_1，……r_{k-1}，$r_k=cr$ 定义为 $r_i=c'\cdot r_{i-1}$，其中$i\in\{1,\ldots,k\}$，且 $c'=\sqrt[k]{c}$。我们创建 k 个 $\mathcal{A}_{\text{fair}}$ 实例，其中第 i 个实例使用参数（c'，r_k）进行初始化。我们运行每个实例，以便从 B_S（q，r_i）中均匀地输出一个点。令$\rho(c')$表示取决于新近似参数 c' 的 LSH-ρ 常数；如果观察到某个实例运行时间超过 $100\cdot dn^{\max\{\rho(c'),\ 1/k\}}$ 个时间步，则停止执行并切换到下一个实例。

命题 6.2. 考虑一个查询 q，并假设 $B_S(q,r)\neq\emptyset$。存在 $i\in\{0,\ldots,k-1\}$，使得：

$$\frac{n(q,r_{i+1})}{n(q,r_i)}\leq n^{\frac{1}{k}} \tag{1}$$

证明。由于 $n(q,r) \geq 1$，因此对于所有 $i \in \{0,\ldots,k\}$，也有 $n(q,r_i) \geq 1$。假设对于所有 $i \in \{0,\ldots,k-1\}$，都有 $n_{(q,r_{i+1})}/n(q,r_i) > n^{\frac{1}{k}}$。那么，通过裂项乘积，我们得出矛盾：

$$\frac{n(q,cr)}{n(q,r)} = \frac{n(q,r_1)}{n(q,r_0)} \cdot \frac{n(q,r_2)}{n(q,r_1)} \cdots \frac{n(q,r_{k-1})}{n(q,r_{k-2})} \cdot \frac{n(q,cr)}{n(q,r_{k-1})} > \left(n^{\frac{1}{k-1}}\right)^{k-1} = n \qquad \square$$

命题 6.2 表明，如果 $B_S(q,\ r) \neq \emptyset$，我们会从某个 $B_S(q,\ r_i)$ 中输出一个均匀采样点，其中 r_i 是一个随机变量，R 取决于 $S,\ q$ 以及算法随机性。另一方面，如果 $B_S(q,\ r) = \emptyset$，我们要么在所有副本 $\mathcal{D}_i$ 超时的情况下输出 $\perp$，要么从某个球面 $B_S(q,\ r_i)$ 中输出一个均匀采样点。无论哪种情况，我们都满足定义 6.1中的宽松公平性保证。至于运行时间，我们的算法在预处理时使用空间 $O(kn^{1+\rho(c')})$，每次查询的耗时为 $\widetilde{O}(dk \cdot \max\{n^{\rho(c')}, n^{1/k}\})$。选择使该指数最小的 k，即可得到以下定理：

定理 6.3. 存在一种算法，可以求解宽松公平的(c,r)-近似最近邻问题；该算法每次查询耗时为 $\widetilde{O}(dn^{\beta})$，预处理所需空间为 $\widetilde{O}(n^{1+\rho(c^{1/\bar{k}})})$，其中 $\bar{k} = \arg\min_{k\in\mathbb{Z}} \max\{\rho(c^{1/k}),\ 1/k\}$，且 $\beta = \max\{\rho(c^{1/\bar{k}}),\ 1/\bar{k}\}$。

求解 β 取决于度量空间。对于超立方体，我们可以使用 $\rho(c) = \frac{1}{2c-1}$，通过粗略估算可得 $k = \Theta(\log c \log\log c)$。为了精确确定常数，我们选取：

$$\beta = \min\left\{\max\left\{\frac{1}{\lfloor k^* \rfloor}, \frac{1}{2c^{1/\lfloor k^* \rfloor} - 1}\right\}, \max\left\{\frac{1}{\lceil k^* \rceil}, \frac{1}{2c^{1/\lceil k^* \rceil} - 1}\right\}\right\}$$

我们的算法运行时间为 $\widetilde{O}(dn^{\beta})$，空间复杂度为 $\widetilde{O}(\sqrt{Q} \cdot n^{1+\beta})$。例如，如果$c = 4$，则有 $k^* = 2.48$，因此 $\beta = 1/3$；而对于 $c = 10$，则有 $k^* \approx 3.15$，因此 $\beta = 1/3$。我们绘出了 β 在 $c \in [2, 100]$ 中的解，如图 2所示，分别对应汉明距离和 ℓ_2 度量指标。注意，$\beta \to 0$ 随着 $c \to \infty$。

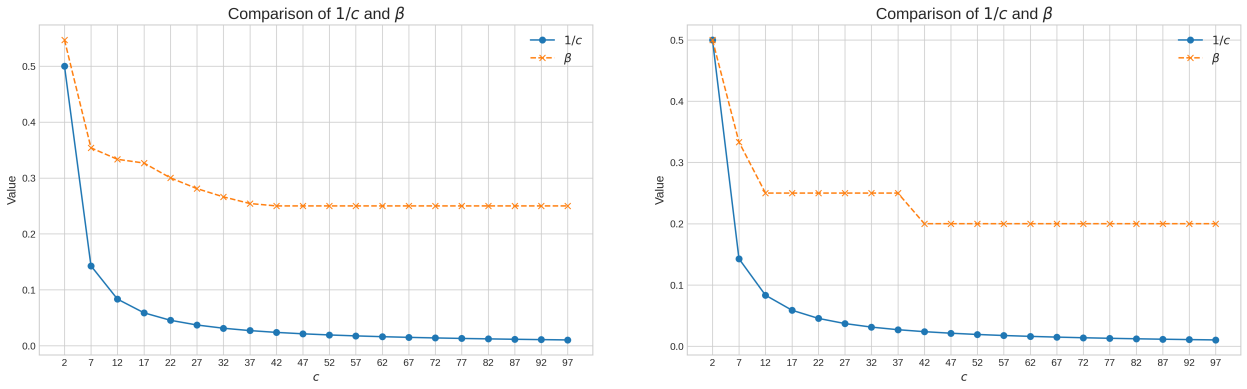

图 2：超立方体（左）和 ℓ_2（右）域中，不同 c 取值对应的 β 解。

7 鲁棒近似最近邻改进

现在，我们将同心环带技术与公平 ANN 相结合，以开发出更高效、更鲁棒的算法。核心思想是使用一组并行实例，动态识别出一个

可能为给定查询带来快速解决方案的环带，然后谨慎地公开这一信息，以维护自适应安全性。

我们再次将空间划分为 $k \geq 1$ 同心环带（r_{i-1}， r_i]，其中 $r_0 = r$ 且 $r_i = c' \cdot r_{i-1}$，对于 $c' = \sqrt[k]{c}$。对于每个环带 i，我们实例化基础算法的两个独立副本：一个测试实例 $\mathcal{A}_i \leftarrow \mathcal{A}_{\text{fair}}$（$c'$，$r_{i-1}$），以及一个留出的执行实例 $\mathcal{A}^{(i)} \xleftarrow{\text{公平}} \mathcal{A}_{\text{fair}}$（$c'$， r_{i-1}）。我们的目标是找到一个环带，使其算法以较大概率在预先设定的运行时间阈值内终止。我们将这一概念形式化如下。

定义 7.1（良好环带）。令 T_i 为测试实例 $\mathcal{A}_i$的随机运行时间。第i个环带是一个良好环带，当且仅当其快速终止的概率 p_i 足够高：

$$p_i := \Pr[T_i \leq 4d(n^{\rho(c')} + n^{1/k}) \log(nQ)] \geq 0.999$$

收到查询 q 后，我们通过观察 $\mathcal{A}_i$ 的 $\Theta(\eta^{-2}\log(kQ/\delta))$ 个独立子试验中在时间限制内终止的比例，以不超过 η 的加性误差估计每个概率 p_i。令 $\hat{}p_i$ 为针对 p_i的这一经验估计。我们使用指示向量 $\hat{}\alpha \in \{0,1\}^k$标记候选环带，其中：$\hat{}\alpha_i = \mathbb{1}[\hat{}p_i \geq 0.998]$。

以高概率，$\hat{}\alpha_i = 1$ 意味着环带 i 是良好的。类似于命题 6.2的论证保证至少存在一个良好的环带。因此，我们找到第一个满足 $\hat{}\alpha_{i^*} = 1$ 的索引 i^*，并将相应的执行实例 $\mathcal{A}^{(i^*)}$ fair 运行至完成。该方法以至少 0.998 的概率，在 $\widetilde{O}(d(n^{\rho(c')} + n^{1/k}))$ 时间内得到一个解

为确保鲁棒性，发布向量$\hat{}\alpha$ （从而选择 i^*）不得泄露有关我们算法实例内部随机性的信息。因此，我们采用Hassidim 等人（2022）提出的基于差分隐私的鲁棒化框架，私密地发布$\hat{}\alpha$。虽然这会使空间复杂度增加$\sqrt{Q}$倍，但使我们能够在无需任何假设的情况下，实现显著更优且纯次线性的查询时间。算法 3详细介绍了我们的方法。

为处理更新，我们再次只需计算所维护的局部敏感哈希数据结构的数量，并按哈希函数的数量进行缩放（见 表 1）

7.1 分析

首先，我们证明向量 $\hat{}a(q)$ 是以鲁棒方式生成的。

引理 7.2。 以至少 $1 - \Theta(\delta)$ 的概率，对于所有查询和环带，向量 $\hat{}p$ 由 算法 3计算得到，并满足：

$$||\hat{p} - p||_\infty \leq \eta$$

尽管对手采取行动试图证明相反结论，上述结论仍然成立。

证明。 我们的论证仿照 定理 C.1 的证明，因为它采用了Hassidim 等人 (2022)的鲁棒化框架。

首先，生成 $\hat{}p$ 经过 Q 时间步后，对于算法使用的随机字符串而言，它满足 (ε, δ)-差分隐私。这是由应用敏感度为 $\frac{1}{s}$ (定理 B.2) 的拉普拉斯机制以及通过子采样实现的隐私放大（定理 B.8）得出的。此外，如果每个 $\hat{}p_i$ 相对于输入随机性都是以隐私方式生成的，我们可以应用差分隐私的并行组合定理（定理 B.5），证明整体发布 $\hat{}p$ 满足隐私性，且无需额外消耗隐私参数。我们知道，相对于输入随机性满足隐私性意味着鲁棒性，因此现在需要计算近似算法中的隐私成本。

改进的鲁棒近似最近邻搜索算法 3

Input: Query $q \in \mathcal{M}$, parameters $c, r, k \geq 1$ and $\delta \in (0, 0.0025)$
procedure INITIALIZE
  Let $c' \leftarrow \sqrt[k]{c}$ and $r_0 \leftarrow r$.
  Let $\eta = 0.001$, $m = \eta^{-2}\log(Qk/\delta)$ and $L = 2400\log^{1.5}(1/\delta)\sqrt{2Q}$.
  for $i = 1, \ldots, k$ **do**
    Let $N = m \times L$ and $r_i \leftarrow c' \cdot r_{i-1}$. ▷ Testing Instance Grid
    Instantiate N copies $\mathcal{A}_{i,j_m,j_L} \leftarrow \mathcal{A}_{\text{fair}}(c', r_{i-1})$ for $(j_m, j_L) \in [m] \times [L]$.
    Instantiate $\mathcal{A}^{(i)}_{\text{fair}} \leftarrow \mathcal{A}_{\text{fair}}(c', r_{i-1})$ ▷ Execution Instances
procedure QUERY(q)
  for $i \in \{1, \ldots, k\}$ **do**
    Let $S_{\text{trunc}} \leftarrow 4d(n^{1/k} + n^{\rho(c')})\log(nQ)$ ▷ Let $p_j \leftarrow \Pr[T_j < S_{\text{trunc}}]$.
    $J_i \leftarrow$ Sample $s = \frac{2}{\eta}\log(Qk/\delta)$ indices $(j_m, j_L) \in [m] \times [L]$ with replacement.
    for $j := (j_m, j_L) \in J_i$ **do**
      ▷ T_{i,j_m,j_L} is the execution time of testing instance A_{i,j_m,j_L}.
      Let $\widetilde{p}_{ij} \leftarrow \mathbb{1}[T_{i,j_m,j_L} < S_{\text{trunc}}]$.
    Let $\hat{p}_i \leftarrow \frac{1}{s}\sum_{j \in J_i} \widetilde{p}_{ij} + \text{Lap}(\frac{1}{s})$.
    Set $\hat{a}(q)_i \leftarrow \mathbb{1}[\hat{p}_i \geq 0.998]$.
  if $\hat{a}(q) = \vec{0}$ **then**
    return $\perp$
  else
    $i^* \leftarrow \min\{i \in \{1, \ldots, k\} \mid a(q)_i = 1\}$ ▷ Find most significant bit index
    return $\mathcal{A}^{(i^*)}_{\text{fair}}(q)$

只需证明对于固定的 $i \in [k]$，有 $|\hat{p}_i - p_i| \leq \eta$。固定一个查询和 $i \in [k]$，根据三角不等式以及标准的切尔诺夫界（因为 $m = \Omega(\eta^{-2}\log(Qk/\delta))$），我们知道：

$$|p_i - \hat{p}_i| \leq \left| p_i - \frac{1}{s}\sum_{j \in J_i} \widetilde{p}_i \right| + \left| \frac{1}{s}\sum_{j \in J_i} \widetilde{p}_{ij} - \widehat{p}_i \right| \leq \frac{\eta}{2} + \left| \frac{1}{s}\sum_{j \in J_i} \widetilde{p}_{ij} - \widehat{p}_i \right|$$

上述求和式的后一项是隐私化过程所引入的误差。我们可以利用已知的拉普拉斯噪声幅度上界来对其进行界定（引理 B.3）：

$$\left| \frac{1}{s}\sum_{j \in J_i} \widetilde{p}_{ij} - \widehat{p}_i \right| = \left| \frac{1}{s}\sum_{j \in J_i} \widetilde{p}_{ij} - \frac{1}{s}\sum_{j \in J_i} \widetilde{p}_{ij} + \text{Lap}\left(\frac{1}{s}\right) \right| \leq \frac{1}{s} \cdot \frac{\eta}{2} \cdot s = \frac{\eta}{2}$$

此情况发生的概率至少为 $1 - e^{-\eta s} \geq 1 -$ [1] 多项式(kQ/δ)。对 k 个环带和 Q 个查询应用并集界，即可证明该引理。 □

推论 7.3。由于 算法 3 通过后处理生成向量 $\vec{a} \in \{0,1\}^k$，这也意味着 a是以鲁棒方式生成的。

接下来，我们说明算法的输出点是正确生成的，并且满足所声称的运行时间和空间复杂度。

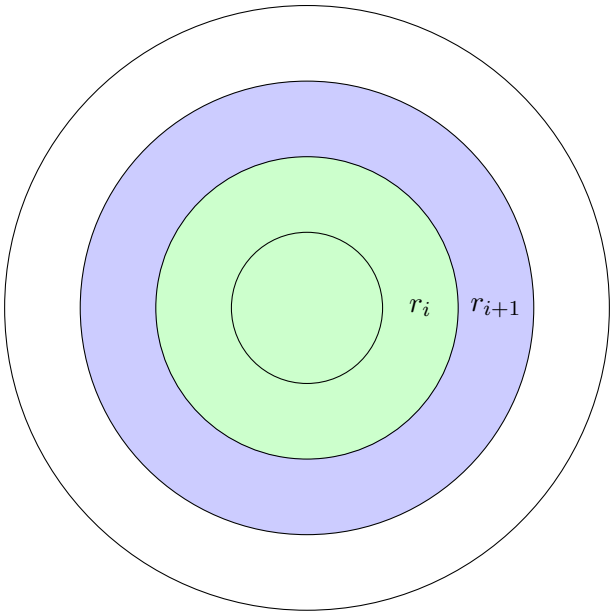

图 3：我们的同心局部敏感哈希构造。绿色区域是集合 $B_S(q,\ r_i)$，蓝色区域表示延伸至 $B_S(q,\ r_{i+1})$ 的环带

定理 7.4。 算法 3 是一个$(\delta + \ ^{1\text{多项式}(n)}$-鲁棒$(c, r)$–近似最近邻算法，使用的预处理空间为 $\widetilde{O}(\sqrt{Q} \cdot n^{1+\rho(c')})$，其中 $c' = c^{1/\bar{k}}$ 且 $\bar{k} = \arg\min_{k \in \mathbb{Z}_{\geq 1}} \left(\beta :\ = \max\{\rho(c^{1/k})\ 1/k\}\right)$。每次查询耗时 $\widetilde{O}(dn^{\beta})$，其概率至少为0.998。

证明。 我们维护 $\Theta(k\log^{2.5}(kQ)\sqrt{Q})$ 个测试实例，以及 k 个执行实例。这意味着我们的总空间复杂度为：

$$\Theta\left(k \log^{2.5}(kQ/\delta)\sqrt{Q} \cdot n^{1+\rho(c')} + kn^{1+\rho(c')}\right) = \widetilde{O}(k\sqrt{Q} \cdot n^{1+\rho(c')}).$$

对于查询运行时间，假设 $B_S(q,\ r) \neq \emptyset$。正如我们在 命题 6.2 中论证的那样，必然存在某个环带 ℓ，其密度比至多为 $n^{1/k}$。对于该环带，定理 4.2 意味着：

$$\Pr\left[T_\ell < 4(n^\rho + n^{1/k}) \log(nQ)\right] \geq 1 - \frac{1}{\text{poly}(n)} \gg 0.999$$

因此，当 $B_S(q,\ r) \neq \emptyset$ 时，始终存在一个优良环带。

根据引理 7.2，一个优良环带以高概率会被算法 3捕获。反之，如果 $\hat{}a(q)_i = 1$，那么 $p_i \geq 0.999 - \eta = 0.998$。因此，如果 i^* 是 a 的最高有效位，则相应的执行实例 $\mathcal{A}^{(i^*)}$ 公平地 以 $O(dn^{\beta})$ 的运行时间运行，且概率至少为 0。998。总的来说，为了处理一次查询，我们运行 $ks = O(\log(Qk))$ 个 $\mathcal{A}_{ij}$ 的截断副本。因此，正如最初所声称的那样，我们的算法每次查询耗时 $O(d\log(nQ) \cdot n^{\beta})$。

最后，为了论证鲁棒性，根据引理 7.2，我们知道以鲁棒的方式发布向量 $\hat{}a$。此外，断言 4.3 告诉我们，留出的执行副本具有鲁棒性，因为最高有效位 i^* 是通过固定函数（后处理）从 a 生成的。总而言之，算法 3的输出以至少 $1 - \delta - \ ^{1\text{多项式}(n)}$的概率具有对抗鲁棒性，其中已计入任一公平近似最近邻算法失效的概率。 □

最后，我们使用中位数放大技巧，以高概率保证运行时间。我们维护上述算法的 $t = \Theta(\log(Q/\delta))$ 个独立副本，并将运行时间定义为

中位运行时间。标准的切尔诺夫界论证表明，任意 Q 个查询中有查询失败的概率至多为 δ/Q；根据并集界，失败概率至多为 δ，这正是我们所需的结果。

8 结论

本研究提出了一系列用于应对自适应对手的近似最近邻算法。我们的方法融合了公平性和隐私原则以及新颖的数据构造，具有高效且与输入无关的特点。我们的工作也引出了若干值得未来研究探讨的问题：我们能否为鲁棒算法确立时间和空间下界？公平性与鲁棒性之间这种强有力的联系，能否扩展到估计问题等其他领域？最后，在实际系统中实现我们的方法，以及研究实现过程中涉及的种种细节，也是值得探讨的重要方向。

致谢

作者谨向 Simons Institute 的 Summer Program on Sublinear Algorithms 组织者致谢，本研究正是在该项目期间启动的。我们感谢 Sofya Raskhodnikova 在早期讨论中帮助我们明确问题，也感谢 Adam Smith 介绍自适应数据分析技术的最新进展。我们还感谢 Christian Sohler 就其论文展开的宝贵讨论，以及 Rathin Desay 提供的有益意见。

参考文献

Andoni, A.（2009）。最近邻搜索：旧方法、新方法与不可能之事。博士论文，麻省理工学院。Andoni, A. 和 Beaglehole, D.（2022）。有保证的鲁棒哈希学习。载于国际机器学习会议，第 599–618 页。PMLR。Andoni, A. 和 Indyk, P.（2017）。高维空间中的近邻。载于离散与计算几何手册，第 1135–1155 页。Chapman and Hall/CRC。Andoni, A.、Indyk, P. 和 Razenshteyn, I.（2018）。高维近似最近邻搜索。载于《国际数学家大会论文集：里约热内卢 *2018*》，第 3287–3318 页。World Scientific。Andoni, A.、Laarhoven, T.、Razenshteyn, I. 和 Waingarten, E.（2016）。近似近邻搜索的时间-空间权衡下界。*arXiv* 预印本 *arXiv:1605.02701*。Andoni, A.、Laarhoven, T.、Razenshteyn, I. 和 Waingarten, E.（2017a）。近似近邻搜索的最优基于哈希的时间-空间权衡。载于第二十八届 *ACM-SIAM* 离散算法年会论文集，第 47–66 页。SIAM。Andoni, A. 和 Razenshteyn, I.（2015）。用于近似近邻搜索的最优数据依赖哈希。载于第四十七届 *ACM* 计算理论研讨会论文集，第 793–801 页。

Andoni, A.、Razenshteyn, I. 和 Nosatzki, N. S.（2017b）。LSH 森林：将实用算法理论化。载于第二十八届 *ACM-SIAM* 离散算法年会论文集，第 67–78 页。SIAM。Attias, I.、Cohen, E.、Shechner, M. 和 Stemmer, U.（2024）。一种通过差分隐私和差值估计器实现对抗式流处理的框架。*Algorithmica*，第 1–56 页。Aumüller, M.、Har-Peled, S.、Mahabadi, S.、Pagh, R. 和 Silvestri, F.（2022）。在高维空间中采样一个近邻——谁最公平？*ACM* 数据库系统汇刊 *(TODS)*，47(1):1–40。Bassily, R.、Nissim, K.、Smith, A.、Steinke, T.、Stemmer, U. 和 Ullman, J.（2016）。自适应数据分析的算法稳定性。载于第四十八届 *ACM* 计算理论研讨会论文集，第 1046–1059 页。Bassily, R.、Smith, A.、Steinke, T. 和 Ullman, J.（2015）。自适应数据分析中，更一般的查询与更低的泛化误差。*arXiv* 预印本 *arXiv:1503.04843*。Bateni, M.、Dhulipala, L.、Fletcher, W.、Gowda, K. N.、Hershkowitz, D. E.、Jayaram, R. 和 Lacki, J.（2024）。高效质心链接聚类。神经信息处理系统进展，37:49649–49683。Behnezhad, S.、Rajaraman, R. 和 Wasim, O.（2025）。针对自适应对手的全动态 $(\Delta + 1)$ 着色。载于 Azar, Y. 和 Panigrahi, D.（编），*2025* 年 *ACM-SIAM* 离散算法年会论文集，*SODA 2025*，美国路易斯安那州新奥尔良，*2025* 年 *1* 月 *12–15* 日，第 4983–5026 页。SIAM。Beimel, A.、Kaplan, H.、Mansour, Y.、Nissim, K.、Saranurak, T. 和 Stemmer, U.（2022）。针对自适应对手的动态算法：通用构造与下界。载于第 *54* 届 *ACM SIGACT* 计算理论年会论文集，第 1671–1684 页。Ben-Eliezer, O.、Eden, T. 和 Onak, K.（2022a）。通过稠密-稀疏权衡实现对抗鲁棒流处理。载于算法简约性研讨会 *(SOSA)*，第 214–227 页。SIAM。Ben-Eliezer, O.、Jayaram, R.、Woodruff, D. P. 和 Yogev, E.（2022b）。对抗鲁棒流算法的框架。*ACM* 通讯 *(JACM)*，69(2):1–33。Bergman, L.、Cohen, N. 和 Hoshen, Y.（2020）。深度近邻异常检测。*arXiv* 预印本 *arXiv:2002.10445*。Broder, A. Z.、Charikar, M.、Frieze, A. M. 和 Mitzenmacher, M.（1998）。最小值独立置换。载于第三十届 *ACM* 计算理论研讨会论文集，第 327–336 页。Bun, M.、Nissim, K.、Stemmer, U. 和 Vadhan, S.（2015）。阈值函数的差分隐私发布与学习。载于 *2015 IEEE* 第 *56* 届计算机科学基础年会，第 634–649 页。IEEE。Chakrabarti, A.、Ghosh, P. 和 Stoeckl, M.（2021）。图流的对抗鲁棒着色。*arXiv* 预印本 *arXiv:2109.11130*。Chakrabarti, A. 和 Stoeckl, M.（2024）。查找缺失项需要强随机性形式。载于第 *39* 届计算复杂性大会 *(CCC 2024)*。Schloss Dagstuhl–Leibniz 信息学中心。

Charikar, M. S. (2002年). 舍入算法中的相似性估计技术. 收录于第三十四届 *ACM* 计算理论研讨会论文集，第 380–388 页。Cherapanamjeri, Y. 和 Nelson, J. (2020). 关于自适应距离估计. 神经信息处理系统进展，33:11178–11190。Cherapanamjeri, Y. 和 Nelson, J. (2024). 次线性时间的终端嵌入. *TheoretiCS*，3。Cherapanamjeri, Y., Silwal, S., Woodruff, D. P., Zhang, F., Zhang, Q., 和 Zhou, S. (2023). 有界对手下针对自适应输入的鲁棒算法. *arXiv* 预印本 *arXiv:2304.07413*。Datar, M., Immorlica, N., Indyk, P., 和 Mirrokni, V. S. (2004). 基于 p-稳定分布的局部敏感哈希方案. 收录于第二十届计算几何研讨会论文集，第 253–262 页。Dinur, I., Stemmer, U., Woodruff, D. P., 和 Zhou, S. (2023). 关于差分隐私以及有界空间下的自适应数据分析. 收录于密码学理论与应用国际年会，第 35–65 页。施普林格。Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., 和 Roth, A. (2015a). 自适应数据分析中的泛化与留出集复用. 神经信息处理系统进展，28。Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., 和 Roth, A. L. (2015b 年). 在自适应数据分析中保持统计有效性. 收录于第四十七届 *ACM* 计算理论研讨会论文集，第 117–126 页。Dwork, C., McSherry, F., Nissim, K., 和 Smith, A. (2006). 隐私数据分析中针对敏感度校准噪声. 收录于密码学理论：第三届密码学理论会议，*TCC 2006*，美国纽约州纽约市，*2006* 年 *3* 月 *4–7* 日。论文集 *3*，第 265–284 页。施普林格。Dwork, C., Rothblum, G. N., 和 Vadhan, S. (2010). 提升与差分隐私. 收录于*2010* 年 *IEEE* 第 *51* 届计算机科学基础年度研讨会，第 51–60 页。IEEE。Feng, S., Feng, Y., Li, G. Z., Song, Z., Woodruff, D., 和 Zhang, L. (2025). 关于通过草图自适应求解搜索问题的差分隐私. 收录于第四十二届国际机器学习会议。Gilbert, A. C., Strauss, M. J., Tropp, J. A., 和 Vershynin, R. (2007). 一份草图适用于所有问题：压缩感知的快速算法. 收录于第三十九届 *ACM* 计算理论研讨会论文集，第 237–246 页。Gionis, A., Indyk, P., 和 Motwani, R. (1999). 通过哈希在高维空间中进行相似性搜索. 收录于第 *25* 届超大型数据库国际会议论文集，VLDB ’99，第 518–529 页，美国加利福尼亚州旧金山。摩根·考夫曼出版社。Gribelyuk, E., Lin, H., Woodruff, D. P., Yu, H., 和 Zhou, S. (2024). 对抗鲁棒 ℓ_0 估计（针对线性草图）的强分离结果. 收录于*2024 IEEE* 第 *65* 届计算机科学基础年度研讨会（*FOCS*），第 2318–2343 页。IEEE。Han, I., Jayaram, R., Karbasi, A., Mirrokni, V., Woodruff, D. P., 和 Zandieh, A. (2024). 超长上下文注意力：近线性时间的长上下文注意力. 收录于国际学习会议

表示（ *ICLR)*. arXiv preprint arXiv : 310 05869 : 2. .

Har-Peled, S., Indyk, P., and Motwani, R. (2012). 近似最近邻：迈向消除维数灾难。计算理论，1(8):321–350。Hardt, M. and Woodruff, D. P. (2013). 线性草图对自适应输入有多鲁棒？载于第四十五届 *ACM* 计算理论研讨会论文集，第 121–130 页。Hassidim, A., Kaplan, H., Mansour, Y., Matias, Y., and Stemmer, U. (2022). 通过差分隐私实现对抗鲁棒流算法。*ACM*期刊，69(6):1–14。Ichnowski, J. and Alterovitz, R. (2015). 基于采样的运动规划中 se (3) 的快速最近邻搜索。载于机器人学算法基础 *XI*：机器人学算法基础第十一届国际研讨会精选论文，第 197–214 页。施普林格。Indyk, P. and Motwani, R. (1998). 近似最近邻：迈向消除维数灾难。载于第三十届 *ACM* 计算理论研讨会论文集，第 604–613 页。Jafari, O., Maurya, P., Nagarkar, P., Islam, K. M., and Crushev, C. (2021). 局部敏感哈希算法及其应用综述。*arXiv* 预印本 *arXiv:2102.08942*。Kalantidis, Y. and Avrithis, Y. (2014). 用于近似最近邻搜索的局部优化乘积量化。载于*IEEE*计算机视觉与模式识别会议论文集，第 2321–2328 页。Kapralov, M., Makarov, M., and Sohler, C. (2024). 汉明空间中局部敏感哈希的对抗鲁棒性。*arXiv* 预印本 *arXiv:2402.09707*。Kitaev, N., Kaiser, Ł., and Levskaya, A. (2020). Reformer：高效的 Transformer。载于国际学习表征会议 *(ICLR)*。arXiv 预印本 arXiv:2001.04451。Lai, L. and Bayraktar, E. (2020). 鲁棒估计量的对抗鲁棒性。*IEEE*信息论汇刊，66(8):5097–5109。Motwani, R. and Raghavan, P. (1996). 随机化算法。*ACM*计算调查（*CSUR*），28(1):33–37。Pham, N. and Pagh, R. (2016). 借助快速 coveringlsh 实现可扩展性与完全召回。载于第 *25* 届 *ACM* 信息与知识管理国际会议论文集，第 1109–1118 页。SantaLucia, J., Allawi, H. T., and Seneviratne, P. A. (1996). 用于预测 DNA 双链稳定性的改进最近邻参数。《生物化学》，35(11):3555–3562。Shalev-Shwartz, S. and Ben-David, S. (2014). 理解机器学习：从理论到算法。剑桥大学出版社。Smith, A. (2017). 自适应数据分析中的信息、隐私与稳定性。*arXiv* 预印本 *arXiv:1706.00820*。Stoeckl, M. (2023). 缺失项查找问题的流算法。载于 *2023* 年 *ACM-SIAM* 离散算法年度研讨会论文集（*SODA*），第 793–818 页。SIAM。

Tagami, Y. (2017). Annexml：用于极端多标签分类的近似最近邻搜索。载于第*23*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第455–464页。Verstrepen, K. 和 Goethals, B. (2014). 统一近邻协同过滤。载于第*8*届*ACM*推荐系统会议论文集，第177–184页。Wei, A. (2022). ℓ_p中的最优拉斯维加斯近似近邻。*ACM*算法汇刊（*TALG*），18(1):1–27。Woodruff, D. P. 和 Zhou, S. (2022). 基于差值估计器的对抗鲁棒数据流与滑动窗口紧致界。载于*2021*年第*62*届*IEEE*计算机科学基础年会（*FOCS*），第1183–1196页。IEEE。

符号表

符号	描述
n	数据集中的点数 $P \subseteq \mathbb{R}^d$
d	环境空间的维度
Q	（可能自适应的）查询数量
r	(c, r) 近似最近邻问题的目标半径
$c > 1$	近似因子
q	一个查询点
$B_S(q, r)$	以 q 为中心、半径为 r 的球与数据集 S 的交集
$n_{(q,r)}$	B_S 中的数据集点数（q， r）
$\rho(c)$	近似因子的 LSH 指数 c
A_{dec}	弱（c， r）近似最近邻判定算法
A_{fair}	输出分布独立于先前查询的公平近似最近邻算法的
α	分桶参数（分段大小指数）
S_i	数据集的第 i个桶/分段
k	同心环带的数量（或子采样参数，取决于上下文）
r_i	第 i个环带的半径， $r_i = c^{i/k} r$
c'	每个环带的近似因子， $c' = c^{1/k}$
β	查询时间指数 $\beta = \min_k \max\{\rho(c^{1/k}),\ 1/k\}$
L	独立算法副本的数量（用于鲁棒性）
m	每个副本的运行时间试验次数
$T_{i,j}$	第 j 个副本在环带 i上的运行时间
$S_{\text{截断}}$	运行时间截断阈值
N_i	在环带 i上于时间 S_{trunc} 内停止的副本比例
$\hat{N}_i$	通过拉普拉斯机制发布的 N_i 带噪估计
ε	差分隐私参数
δ	失败概率参数 / 差分隐私参数

表 2：全文使用的符号汇总。

B 差分隐私背景

我们的工作大量依赖差分隐私领域的研究成果，因此在此给出必要的定义和结果。

B.1 差分隐私的定义

定义 B.1（差分隐私）。设 $\mathcal{A}$ 是一个随机算法，它在元素来自某个全集的数据库上运行。对于参数 $\varepsilon > 0$ 和 $\delta \in [0,1]$，若对于任意两个相邻数据库 $S \sim S'$（它们仅有一行不同），算法在 S 和 S' 上运行时的输出分布都非常接近，则算法 $\mathcal{A}$ 是（ε, δ）–差分隐私的（*DP*）。也就是说，对于任意 $S \sim S'$ 以及输出空间 $\mathcal{A}$ 的任意结果子集 T，都有：

$$\Pr[\mathcal{A}(S) \in T] \leq e^{\varepsilon} \cdot \Pr[\mathcal{A}(S') \in T] + \delta$$

B.2 拉普拉斯机制及其性质

定理 B.2（拉普拉斯机制（Dwork 等人，2006））。设 $f: X^* \to \mathbb{R}$是一个函数。定义其敏感度 ℓ 为以下变化幅度的一个上界：f 在相邻数据库上的变化量。

$$\forall S \sim S': \quad |f(S) - f(S')| \leq \ell$$

该算法在输入 $S \in X^*$时返回 $f(S) + Lap\left(\frac{\ell}{\varepsilon}\right)$，并且是（$\varepsilon$, 0）–差分隐私的，其中

$$Lap(\lambda; x) := \frac{1}{2\lambda} \exp\left(-\frac{|x|}{\lambda}\right)$$

是在 $\mathbb{R}$上的拉普拉斯分布。

我们将利用拉普拉斯分布的以下集中性质：

引理 B.3。对于 $m \geq 1$，令 Z_1、$...Z_m \sim Lap(\lambda)$ 为独立同分布的随机变量。我们有：

$$\Pr\left[\max_{i=1}^{m} Z_i > \lambda(\ln(m) + t)\right] \leq e^{-t}$$

B.3 差分隐私的性质

差分隐私具有许多有助于算法设计的性质。以下定理称为“高级自适应组合”，描述了差分隐私算法以自适应方式依次连接的情形。

定理 *B.4*（高级组合，（*Dwork* 等人，*2010*））。假设算法 $\mathcal{A}_1$，...，$\mathcal{A}_k$均满足*(*ε*,* δ*)*–差分隐私。令 $\mathcal{A}'$ 为这些算法的自适应组合：对于输入数据库 x，算法$\mathcal{A}_i$ 会获得 x；对于 $i \geq 2$，则会获得 $\mathcal{A}_{i-1}$的输出 y_{i-1}。那么，对于任意 $\delta' \in$ *(0, 1)*，~算法 $\mathcal{A}$ 满足*(*$\widetilde{\varepsilon}$*,* δ*)*–差分隐私，其中：

$$\widetilde{\varepsilon} = \varepsilon \cdot \sqrt{2k\ln(1/\delta')} + 2k\varepsilon^2 \quad and \quad \widetilde{\delta} = k\delta + \delta'$$

此外，对于数据集被划分的情形，还有一个组合定理：

定理 B.5（并行组合）。设 f_1, ..., f_k 是（ε, 0）-差分隐私机制，X 是一个数据集。假设 X 被划分为 k 个部分 X_1, ..., X_k，并令 $f(X) = (f_1(X_1)$ ..., $f_k(X_k))$。那么 f 是（ε,0）-差分隐私。

下一个定理指出，只要处理过程不使用原始数据库中的信息，对差分隐私算法的输出进行后处理就不会削弱其隐私保证。

定理 B.6（差分隐私在后处理中保持不变）。设 $\mathcal{A}\colon U^n \to Y^m$ 和 $\mathcal{B}\colon Y^m \to Z^r$ 是随机算法，其中 U、Y、Z 是任意集合。若 $\mathcal{A}$ 满足$(\varepsilon,\ \delta)$–差分隐私，则复合算法 $\mathcal{B}(\mathcal{A}(\cdot))$ 也满足该性质。

下面的定理展示了差分隐私算法在学习中的强大能力。

定理 *B.7*（差分隐私与泛化，（*Bassily* 等人，*2016* 年；*Dwork* 等人，*2015b* 年））。令$\varepsilon \in (0, 1/3)$ 和 $\delta \in (0, \varepsilon/4)$。设 $\mathcal{A}$ 是一个（ε, δ）差分隐私算法，它在 X^n 中的数据库上运行，并输出 m 个谓词函数 h_i：$X \to \{0,1\}$，其中 $i \in [m]$。那么，如果 D 是 X上的任意分布，且 S 由从 D中抽取的 $n \geq \frac{1}{\varepsilon^2} \cdot log\left(\frac{2\varepsilon m}{\delta}\right)$ 个独立同分布样本组成，则对所有$i \in [m]$，有：

$$\Pr_{\substack{S\sim D^n \\ h_i \leftarrow \mathcal{A}(S)}}\left[\left|\frac{1}{|S|}\sum_{x\in S} h_i(x) - \mathop{\mathbb{E}}_{x\sim D}[h_i(x)]\right| \geq 10\varepsilon\right] \leq \frac{\delta}{\varepsilon}$$

换句话说，私下生成的谓词是输入数据上任意分布下其期望的良好估计量。我们将使用的隐私的最后一个性质，是通过子采样实现的提升技术：

定理 B.8（通过子采样实现的隐私放大，（Bun 等人，2015；Cherapanam- jeri 等人，2023 年）。设 $\mathcal{A}$ 是一个（ε，δ）-差分隐私算法，作用于大小为 m的数据库。对于 $n \geq 2m$，考虑这样一个算法：输入大小为 n的数据库后，它从数据库中（有放回地）采样 m 行，并在采样结果上运行 $\mathcal{A}$ 。那么，该算法满足（ε'，δ'）-差分隐私，其中

$$\varepsilon' = \frac{6\varepsilon m}{n} \ \ and \ \ \delta' = \exp\left(\frac{6\varepsilon m}{n}\right) \cdot \frac{4m}{n} \cdot \delta$$

B.4 通过内部随机性的隐私实现鲁棒化

本工作反复使用的一种工具是：通过要求算法相对于其内部随机性（而非输入数据集）保持稳定，我们可以获得针对自适应查询的鲁棒性（Hassidim 等人，2022）。具体而言，我们将算法使用的随机硬币视为一个（隐藏的）数据库，并要求可观测到的交互记录相对于随机性的变化满足差分隐私。差分隐私由此保证，任何自适应选择的查询序列都无法显著依赖或过拟合于设置或执行过程中所做的特定随机选择。我们在附录 C的证明中明确运用并完善了这类论证。

C 定理 C.1 的证明

本节对 算法 1 的构造进行形式化分析。我们证明以下定理：

定理 C.1。设 $\mathcal{A}$ 是一个用于近似最近邻的无偏判定器算法，其使用 $s(n)$ 空间，且每次查询耗时 $t(n)$。设 、，并假设我们令 $\log^{1.5}(1/\delta)\cdot\sqrt{2Q}$，且 $k = \log(Q/\delta)$。那么，算法 $\overline{\mathcal{A}_{dec}}$ 是一个对抗稳健判定算法，其成功概率至少为$1 - \Theta(\delta)$，使用 $(n) \cdot \widetilde{O}\left(\sqrt{Q}\right)$ 位空间，且每次查询耗时 $\widetilde{O}(t(n))$。

首先，我们证明该算法关于其输入随机性满足差分隐私。

引理 C.2. 令 $\varepsilon = 0.01$ 和 $\delta < \varepsilon/4$。算法 $\mathcal{A}_{dec}$ 关于随机字符串 R 满足 (ε, δ)–差分隐私。

证明。 我们分析算法 $\mathcal{A}_{\text{dec}}$ 的隐私，该算法见算法 1 ，我们将随机字符串 R视为其输入。假设我们令

$$\varepsilon' = \frac{\varepsilon}{2\sqrt{2Q\ln(1/\delta)}}$$

对于所有 $i \in [Q]$，我们声称，查询 q_i 的响应相对于 R 是 $(\varepsilon', 0)$–差分隐私。这是因为算法 1 第 8 行定义的统计量 N_i 的敏感度为 $1/k$ ，因此根据定理 B.2，在第 9 行应用拉普拉斯机制后，发布 N_i 对字符串 R 是 $(1, 0)$–差分隐私，其中 $\widehat{\ }$ 相对于 $\widehat{\ }$。基于将 N_i 与常数阈值 $1/2$ 进行比较而得出的二进制输出，通过后处理（定理 B.6）仍满足 $(1, 0)$-差分隐私。

由于 $L \geq 2k$，根据子采样放大性质（定理 B.8），我们得到每次迭代都满足 $(\varepsilon', 0)$–差分隐私，因为当 Q 足够大时，我们有：

$$\begin{aligned}
\frac{6k}{L} &= \frac{6 \cdot 0.01(\log Q + \log \frac{1}{\delta})}{24 \log \frac{1}{\delta}\sqrt{2Q \log \frac{1}{\delta}}} \\
&= \frac{6\varepsilon \log \frac{1}{\delta} + 6\varepsilon \log Q}{24 \cdot \log \frac{1}{\delta}\sqrt{2Q \ln\left(\frac{1}{\delta}\right)}} && (\text{Since } \varepsilon = 0.01) \\
&< \frac{2\varepsilon}{4\sqrt{2Q \ln\left(\frac{1}{\delta}\right)}} \\
&= \varepsilon'
\end{aligned}$$

最后，根据自适应组合（定理 B.4），经过 Q 个自适应步骤后，我们得到的算法满足(ε'', δ)-差分隐私，其中：

$$\varepsilon'' = \varepsilon'\sqrt{2Q \ln\left(\frac{1}{\delta}\right)} + Q(\varepsilon')^2 = \frac{\varepsilon}{2} + \frac{\varepsilon^2}{4 \ln\left(\frac{1}{\delta}\right)} \leq \varepsilon$$

对于 $\varepsilon \leq 2\ln \delta^{-1}$，当 $\delta \in (0, 0.0025)$ 时，该条件成立。因此，就其输入——随机字符串 R——而言，算法 $\mathcal{A}_{\text{dec}}$ 是 (ε, δ)–差分隐私的。 □

接下来，我们将证明，即使面对对抗性生成的查询，大多数数据结构 $\mathcal{D}_i$ 也能以很高的概率输出准确的判定结果。

引理 C.3. 以至少 $1 - \delta$ 的概率，对于所有 $i \in [Q]$，至少 $0.8L$ 的回答 a_{ij} 是对查询 q_i 所对应判定问题的准确回答。

证明。 证明的核心思想源自（Hassidim 等人，2022），即设想对手 $\mathcal{B}$ 是一个后处理机制，它试图猜出哪些随机字符串会导致 $\mathcal{A}$ 出错。

设有一个包装元算法 $\mathcal{C}$，如算法 4 所示，它以随机字符串 $R = \sigma_1 \circ r_2 \circ \cdots \circ \sigma_L$ 作为输入；该字符串按照某种未知的任意分布$\mathcal{R}$生成。该算法 $\mathcal{C}$ 模拟 $\mathcal{A}_{\text{dec}}$ 与 $\mathcal{B}$之间的博弈：首先运行 $\mathcal{B}$ ，向 $\mathcal{A}_{\text{dec}}$提供某个输入数据集 $S \subseteq U$ ，该数据集以 R中的随机字符串作为种子。随后， $\mathcal{C}$ 使用 $\mathcal{B}$ 对 $\mathcal{A}_{\text{dec}}$进行自适应查询，查询内容为 $(q_1, ..., q_Q)$。与此同时，它模拟 $\mathcal{A}_{\text{dec}}$，接收答案 $a_1, ..., a_Q$，并将这些答案反馈给 $\mathcal{B}$。根据引理 C.2，无论对手如何进行查询，输出 $(a_1, ..., a_Q)$ 都是相对于 R私密生成的。

在每一步 i， $\mathcal{B}$向 $\mathcal{A}_{\text{dec}}$提供 $\vec{q_i} = (q_1, \ldots\ldots, q_i)$并收到 i 个答案 $(a_1, \ldots\ldots, a_i)$ 后，我们的元算法 $\mathcal{C}$ 对交互记录 $\{(q_j, a_j)\}_{j=1}^{i}$ 进行后处理，以生成一个谓词$h_{\vec{q_i}}$： $\{0,1\}^* \to \{0,1\}$。该谓词指出，在判定问题情形下，对于输入数据集 S，哪些字符串 $\sigma \in \{0,1\}^*$能使算法 $\mathcal{A}$成功回答查询前缀 $\vec{q_i}$。更正式地说：

$$h_{\vec{q_i}}(\sigma) := \bigwedge_{1 \le j \le i} \{\mathcal{A}(\sigma)(S, q_j) \in D(S, q_j, c, r)\} \tag{2}$$

请注意，此定义涵盖了一种中间情形：算法给出的任何答案都被视为正确。

算法 4 元算法 $\mathcal{C}$，运行了 i 步

1: 输入：随机字符串 $R = \sigma_1 \circ \sigma_2 \circ \cdots \sigma_L$、算法 $\mathcal{A}_{\text{dec}}$ 和 $\mathcal{B}$ 的描述。 2: 模拟 B，得到一个数据集 $S \subset U$。 3: 用随机字符串（$\sigma_1, ..., \sigma_L$）和数据集 S 初始化 $\mathcal{A}_{\text{dec}}$。 4: 对 $i \in Q$ 执行 5: 模拟 $\mathcal{B}$，根据此前的查询和回答历史生成一个查询 q_j。 6: 在查询 q_j 上模拟 $\mathcal{A}$，生成一个答案。 7: 根据公式 2，通过对查询/回答历史进行后处理，计算谓词 $h_{\vec{q_i}}(\cdot)$。 8: 输出 $(h_{\vec{q_1}}, ..., h_{\vec{q_Q}})$。

之所以能够生成这些谓词，是因为 $h_{\vec{q_i}}$ 仅依赖于 $\vec{q_i}$，而 $\vec{q_i}$ 是输出历史的子串， $\mathcal{C}$ 可以访问该历史。因此， $\mathcal{C}$ 可以通过（例如）穷举计算 R 的每个取值，来生成 $h_{\vec{q_i}}$[§]。由于 $\mathcal{C}$ 只能对查询/回答向量（$q_1, a_1, ..., q_i, a_i$）进行后处理，根据定理 B.6，输出谓词 $h_{\vec{q_i}}$ 相对于 $\sigma_1, ..., \sigma_L$ 也以 (ε, δ)–差分隐私方式生成。

根据这些 Q 私下生成的谓词，并且由于 $L > \frac{1}{\varepsilon^2} \log \frac{2\varepsilon Q}{\delta}$ 对于足够大的~~Q~~，根据差分隐私的泛化性质（定理 B.7），我们有[¶] 其成立的概率至少为 $1 - \frac{\delta}{\varepsilon} = 1 - \Theta(\delta)$ 对于任意分布 $\mathcal{R}$ 以及所有 $i \in [Q]$，都有：

$$\left| \mathop{\mathbb{E}}_{\sigma \sim \mathcal{R}} \left[h_{\vec{q_i}}(\sigma) \right] - \frac{1}{L} \sum_{j=1}^{L} h_{\vec{q_i}}(\sigma_j) \right| \le 10\varepsilon = \frac{1}{10} \tag{3}$$

但是，如果 $\mathcal{R}$ 是均匀分布，那么 $\mathbb{E}_{\sigma \sim \mathcal{R}} \left[h_{\vec{q_i}}(\sigma) \right]$ 就是 $\mathcal{A}$ 在固定的查询序列 $\vec{q_i}$上给出准确答案的概率。由于 $\mathcal{A}$ 是一个无偏判定器，公式 3意味着：

$$\mathop{\mathbb{E}}_{\sigma \sim \mathcal{R}} \left[h_{\vec{q_i}}(\sigma) \right] \ge \frac{9}{10} \tag{4}$$

此外，$\frac{1}{L} \sum_{j=1}^{L} h_{\vec{q_i}}(\sigma_j)$ 是使 $\mathcal{A}_2$ 正确的随机字符串所占的比例。因此，根据公式 4，对于所有 $\left(\frac{9}{10} - \frac{1}{10}\right) L = 0.8L$，该比例至少为 ~~$i$~~ $\in [Q]$。 □

现在我们可以证明本节的主定理了。

[§]我们假设 $\mathcal{C}$ 拥有无限的计算能力。

[¶]假设 $\delta \in (0, \varepsilon/4)$。

定理 定理 C.1 的证明。我们假设事件 引理 C.3 成立，该事件发生的概率至少为 $1-\Theta(\delta)$。那么，对于所有 $i \in [Q]$，查询 q_i 的正确答案比例要么至少为 0.8，当 $B_S(q_j, r) \neq \emptyset$时；否则至多为 $1-0.8=0.2$。现在，我们需要考虑子采样引入的误差，这将通过以下引理来处理：

引理 C.4. 令 $L \in \mathbb{N}$，并假设至少有 α 比例的 L 结构是良好的，其中$\alpha \in \{0.8, 0.2\}$。令 $j_1, \ldots, j_k$从 $[L]$ 中独立且均匀地（有放回地）抽样，并定义

$$\bar{N} \;=\; \frac{1}{k}\sum_{t=1}^{k} X_t, \qquad X_t := \mathbf{1}[\textit{structure } j_t \textit{ is good}].$$

那么对于 $k \geq 40\log(Q/\delta)$，

$$\Pr\left[\begin{array}{ll} \bar{N} \leq 0.6 & \textit{if } \alpha = 0.8, \\ \bar{N} \geq 0.4 & \textit{if } \alpha = 0.2 \end{array}\right] \;\leq\; \delta/Q.$$

证明。 随机变量 $X_1, \ldots, X_k$是独立同分布的伯努利(α)随机变量，因此 $\mathbb{E}[N] = \alpha$。 **情况 1：** $\alpha = 0.8$。 我们应用乘法型切尔诺夫界。对于任意 $0 < \gamma < 1$,

$$\Pr[\bar{N} \leq (1-\gamma)\alpha] \;\leq\; \exp\left(-\frac{\gamma^2 \alpha k}{2}\right).$$

令 $(1-\gamma)\alpha = 0.6$，得到 $\gamma = 1/4$，因此

$$\Pr[\bar{N} \leq 0.6] \;\leq\; \exp\left(-\frac{(1/4)^2 \cdot 0.8}{2}\, k\right) \;=\; \exp(-k/40).$$

情况 2： $\alpha = 0.2$。 由对称性,

$$\Pr[\bar{N} \geq 0.4] \;\leq\; \exp(-k/40).$$

取 $k \geq 40\log(Q/\delta)$，即可得到所声称的界。 □

最后，根据引理 B.3，我们要求最大拉普拉斯噪声以高概率不超过0。1：

$$\Pr\left[|Z_i| > 0.1\right] = \Pr\left[|Z_i| > \frac{1}{k}\left(\ln(1) + 0.1k\right)\right] \leq e^{-0.1k} \tag{5}$$

由于我们的判定阈值为 $N_i := N_i + Z_i \geq 0.5$，可以看出，设置 $k = \Omega(\log(Q/\delta))$会使公式 5 中的概率至多为$\frac{\delta}{Q}$，由并集界可知， $\mathcal{A}_{\text{dec}}$ 在每个时间步都输出正确答案 $i \in [Q]$ 的概率很高。 □

D 改进的鲁棒 ANNS 算法，具有 $\forall$ 保证

在本节中，我们将讨论实现搜索问题对抗鲁棒性的另一种途径——提供对所有情形都适用的保证。本节将重点讨论近似最近邻问题，因为该问题应用广泛且十分重要，也适合采用我们将要讨论的技术。

D.1 汉明超立方体中的全体 保证

我们首先介绍汉明距离近似最近邻情形，因为这是最自然的一种对所有情形都适用的保证。这是因为我们处理的空间是离散的，因此可以轻松地对所有可能的查询进行并合界定，而付出的代价仅为度量空间维度的多项式级别 d 。

定理 D.1. 存在一种对抗鲁棒算法，能够在d维汉明超立方体中解决(c, r)–近似最近邻问题，并且以至少 $1 - 1/n^2$ 的概率正确回答每一个可能的查询。空间需求为 $\widetilde{O}(d \cdot n^{1+\rho+o(1)})$，每次查询所需的时间为 $\widetilde{O}(d^2 \cdot n^{\rho})$，其中 $\rho = 1/c$。

证明。 首先，回顾汉明超立方体中的标准局部敏感哈希：给定一个包含 $|S| = n$的点集 $S \subseteq \{0,1\}^d$。我们接收查询 $q \in \{0,1\}^d$。我们的局部敏感哈希族 $\mathcal{H}$定义如下：选取某个坐标 $i \in [d]$，并根据 $x_i \in \{0,1\}$对 $x \in \{0,1\}^d$进行哈希。该函数 h 根据第 i 个坐标，将超立方体中的点分割成两个相等的部分，起到超平面的作用。从 $\mathcal{H}$ 中均匀随机采样 h ，等价于均匀随机采样 $i \in [d]$ 。我们很容易看出， $\mathcal{H}$ 是一个 (r, cr, p_1, p_2)–局部敏感哈希族，因为：

$$\Pr_{h\sim\mathcal{H}}[h(p) = h(q)] = \frac{d - ||p-q||}{d} = \begin{cases} \geq 1 - \frac{r}{d} := p_1, & \text{when } ||p-q|| \leq r \\ \leq 1 - \frac{cr}{d} := p_2, & \text{when } ||p-q|| \geq cr \end{cases}$$

现在我们来介绍局部敏感哈希族的典型放大过程（Gionis 等人，1999）。我们不再只采样一个坐标，而是采样 k。我们也不再只采样一个哈希函数，而是采样 L 个不同的哈希函数 h_1，.....，$h_L \in \mathcal{H}^k$并要求一个近点至少与 q碰撞一次。采用这一方案，我们知道，如果固定 $q \in \{0,1\}^d$和 $p \in B_S(q,\ r)$，则有：

$$\Pr\left[\exists i \in [L] \,:\, h_i(p) = h_i(q)\right] \geq 1 - (1 - p_1^k)^L$$

此外，如果 $||p-q|| \geq cr$，则必有：

$$\Pr\left[\exists i \in [L] \,:\, h_i(q) = h_i(p)\right] \leq L p_2^k$$

现在，我们希望保证以高概率不存在任何查询 $q \in \{0,1\}^d$，使得对于所有点 $p \in B_S(q,\ r)$，对所有 $i \in [L]$，都有 $h_i(q) \neq h_i(p)$。换言之，我们希望：

$$\Pr\left[\exists q \in \{0,1\}^d \,:\, \forall p \in B_S(q,r)\, \forall i \in [L] : h_i(p) \neq h_i(q)\right] \leq \frac{1}{n}$$

我们可以利用并集界得到：

$$\begin{aligned} &\Pr\left[\exists q \in \{0,1\}^d \,:\, \forall p \in B_S(q,r)\, \forall i \in [L] : h_i(p) \neq h_i(q)\right] \\ &\leq \sum_{q\in\{0,1\}^d} \Pr\left[\forall p \in B_S(q,r)\, \forall i \in [L] : h_i(p) \neq h_i(q)\right] \end{aligned}$$

因此，只需证明对于固定的 $q \in \{0,1\}^d$，有：

$$\Pr\left[\forall p \in B_S(q,r)\, \forall i \in [L] : h_i(p) \neq h_i(q)\right] \leq \frac{1}{n2^d}$$

我们可以放宽这一陈述，并按如下方式使用并集界：

$$\begin{aligned}\Pr\left[\forall p \in B_S(q,r)\, \forall i \in [L] : h_i(p) \neq h_i(q)\right] &\leq \Pr\left[\exists p \in B_S(q,r)\ \nexists i \in [L] : h_i(p) = h_i(q)\right] \\ &\leq \sum_{p \in B_S(q,r)} \Pr\left[\nexists i \in [L] : h_i(p) = h_i(q)\right] \\ &\leq |B_S(q,r)| \cdot (1-p_1^k)^L \\ &\leq n(1-p_1^k)^L\end{aligned}$$

因此，只需满足：

$$(1-p_1^k)^L \leq \frac{1}{n^2 2^d} \tag{6}$$

另一方面，我们将在 $S \setminus B_S(q,\ cr)$ 中看到的、与 q 落入同一桶的点的期望数量为：

$$\mathbb{E}\left[|p \in S \setminus B_S(q,cr) \mid \exists i \in [L]\ :\ h_i(p) = h_i(q)|\right] = \sum_{p \in S \setminus B_S(q,cr)} \Pr\left[\exists i \in [L] \mid h_i(p) = h_i(q)\right] \tag{7}$$

$$\leq nLp_2^k \tag{8}$$

现在，我们可以结合 公式 6 和 公式 8 来求出 k 和 L的取值。首先，我们希望期望 $O(L)$ 时间，因此要求 $p_2^k \leq 1/n$，由此可得：

$$k \geq \log_{1/p_2}(n)$$

现在，令 $p_1 = p_2^\rho$。代入后，可求得 L 的值为：

$$L \geq n^\rho d \log n$$

由此可知，我们的算法以高概率在 $O(L)$ 时间内运行。具体而言，令 X 为 $S \setminus B_S(q,\ cr)$ 中与 q 被哈希到同一个桶的点数。根据简化的切尔诺夫界，我们有：

$$\Pr\left[X \geq 10L\right] \leq 2^{-10L} = \frac{1}{n^{10dn^\rho}} \ll \frac{1}{n^{\Omega(1)}}$$

这意味着我们的算法以高概率使每次查询的运行时间为 $O(L)$。至于常数 ρ 的取值，根据定义有：

$$\rho := \frac{\log p_1}{\log p_2} = \frac{\log\left(1-\frac{r}{d}\right)}{\log\left(1-\frac{cr}{d}\right)} \approx \frac{1}{c}$$

~总的来说，计算我们的哈希函数需要 $O(\log n)$ 时间，而计算点之间的距离需要 $O(d)$ 时间。我们维护 $O(d \cdot n^\rho \log n)$ 个哈希表，这意味着处理单次查询需要 $O(d^2 \cdot n^\rho \log\ n)$ 时间。对于预处理，除了将整个数据集存储在 dn 空间中，我们还需要 $O(d \cdot n^{1+\rho+o(1)})$ 空间来构建数据结构。 □

D.1.1 通过采样改进查询运行时间

我们可以通过采样找到一个合适的桶，从而改善查询运行时间对 d 的依赖。下面的定理概括了这一发现，将运行时间复杂度降低了 d 倍：

定理 D.2。存在一种对抗鲁棒算法，能够在 d 维汉明超立方体中求解 (c, r)–近似最近邻问题，并且能够以至少 $1 - 1/n^2$ 的概率正确回答所有可能的查询。空间需求为 $\widetilde{O}(d \cdot n^{1+\rho+o(1)})$，每次查询所需的时间为 $\sim O(d \cdot n^{\rho})$，其中 $\rho = 1/c$。

证明。根据上述分析，我们知道需要取 $L = n^{\rho} \cdot d\log\ n$ 个不同的哈希函数。考虑某个查询 q。我们分析包含某个点$p \in B_S(q, r)$ 的桶的期望数量。令 X_q 为一个随机变量，表示满足以下条件的桶的数量：$i \in [L]$ 中的某个点位于桶 i，其中 $B_S(q, r)$。定义以下指示随机变量：

$$\mathbb{1}_i = \begin{cases} 1, & \text{if some point } p \in B_S(q,r) \text{ lies in bucket } i \in [L] \\ 0, & \text{otherwise} \end{cases}$$

根据期望的线性性，我们现在可以写出：

$$\begin{aligned} \mathbb{E}[X_q] &= \sum_{i=1}^{L} \Pr[\mathbb{1}_i = 1] \\ &= \sum_{i=1}^{L} \Pr\left[\bigcup_{p \in B_S(q,r)} \{h_i(p) = h_i(q)\}\right] \\ &\geq L \cdot p_1^k \\ &= L \cdot (p_2)^{\rho k} \\ &\geq \frac{L}{n^{\rho}} \\ &= d \log n \end{aligned}$$

通过切尔诺夫 界，我们可以看到，以高概率， X_q 接近其期望值 ：

$$\Pr\left[X_q \leq \frac{1}{2} d \log n\right] \leq e^{-\frac{d \log n}{8}} = \frac{1}{n^{d/8}} \ll \frac{1}{n}$$

那么，我们以 $X_q > \frac{1}{2} d\log\ n$ 为条件。在查询时间，我们只需从 $[L]$ 中均匀随机抽取 $m = \Theta(n^{\rho}\log\ n)$ 个桶。我们知道，单个随机选取的桶以至少 $^{d}{}_{\log 2n^{\rho} d}{}^{n}{}_{\log\ n} = \frac{1}{2n^{\rho}}$ 的概率包含来自 $B_S(q, r)$ 的某个点。因此，所有 m 次抽取都不包含此类点的概率至多为：

$$\left(1 - \frac{1}{n^{\rho}}\right)^{n^{\rho} \log n} \leq e^{-\log n} = \frac{1}{n}$$

因此，以至少 $1 - \frac{1}{n}$ 的概率，我们能找到一个包含优质点的桶。由于以高概率，任意桶中的 $P \setminus B_S(q, cr)$ 个点的数量为 $O(L)$，因此我们看到，这种采样方法将查询运行时间改进为 $O(n^{\rho} \log n)$。

□

D.1.2 利用最优 LSH 算法

我们之前的阐述使用了针对汉明超立方体的原始局部敏感哈希构造（Indyk 和 Motwani，1998），其达到 $\rho = 1/c$。我们也可以使用（Andoni 和 Razenshteyn，2015）中的最先进的方法，该方法达到 $\rho = \frac{1}{2c-1}$，以替代定理 D.1。这会略微改进 n 上的指数：

定理 D.3. 存在一种对抗鲁棒算法，可以解决 d 维汉明超立方体上的(c, r)–近似最近邻问题，并能以至少0.99的概率正确回答所有可能的查询。其空间复杂度为 $O(d \cdot n^{1+\rho+o(1)})$，每次查询所需时间为 $O(d \cdot n^{\rho})$，其中 $\rho = \frac{1}{2c-1}$。这些运行时间保证以高概率成立。

分析过程完全相同，因此不再赘述：由于该算法以常数概率成功，而我们希望它对所有 2^d 可能的查询都成功，因此将其成功概率提升至 $1 - \frac{1}{100 \cdot 2^d}$。这样，经过并集界后，任意查询成功的概率至少为 0.99。此外，定理 D.2 中用于改善查询运行时间的采样算法分析也保持不变。使用标准汉明范数 LSH 与使用最优 LSH 算法之间唯一的变化是比率 $\rho := \log(1/p_1)\ \log(1/p_2)$。

D.2 通过度量覆盖对连续空间进行离散化

该全称 算法（如定理 D.2 所述）无法应用于离散空间之外的情况，这是因为我们进行分析的关键在于对所有可能查询应用并集界。

为了模拟类似的论证来解决连续 ℓ_p 空间中的近似最近邻问题，我们可以考虑对空间进行离散化的策略。我们放置特殊的“标记”点，并保证在这些点附近可以解决某种形式的近似最近邻问题。然后，当查询到来时，我们找到与其对应的标记点，并针对该点解决近似最近邻问题。我们证明，只要标记点周围的“邻域”足够小，所得答案对原始查询同样有效。类似的策略和覆盖构造曾出现在（Cherapanamjeri 和 Nelson，2024）的工作中，不过他们并未利用可以将任意查询点投影到覆盖集上的这一能力来设计算法。相反，他们的算法认为，只要在整个覆盖集中的每个点上都能成功就足够了。

D.2.1 连续空间中的度量指标覆盖

为了开始我们的探究，我们需要给出度量指标覆盖的定义：

定义 D.4. 考虑度量空间 $\mathcal{M} = (\mathbb{R}^d, ||\cdot||_p)$，其度量指标为 μ。令 $U \subset \mathbb{R}^d$ 为一个有界子集。集合 $\widehat{S} \subseteq \mathbb{R}^d$ 称为 U 的 Δ-覆盖，如果对于所有 $q \in U$，都存在某个 $\widehat{s} \in \widehat{S}$，使得

$$||q - \widehat{s}||_p \leq \Delta$$

假设 U 是 $\mathbb{R}^d$ 的一个有界子集。我们可以构造如下的 Δ-覆盖：令 $C := \sup_{x \in U} ||x||_\infty$，并假设 $\{u_i\}_{i=1}^d$ 是张成 U 的正交归一基。我们知道，对于所有 $x \in U$，$||x||_\infty \leq C$，因此定义：

$$\widehat{S} = \sum_{i=1}^{d} \widehat{\alpha}_i u_i, \quad \text{where}$$
$$\widehat{\alpha}_i \in \{-C, -C + \varepsilon, ..., C - \varepsilon, C\}$$

取某个 ε，具体取值稍后确定。这是针对 ℓ_2 的标准构造，我们现在将其推广到 ℓ_p（Shalev-Shwartz 和 Ben-David，2014）。根据定义，我们有：

$$\left|\widehat{S}\right| = \left(\frac{2C}{\varepsilon}\right)^d$$

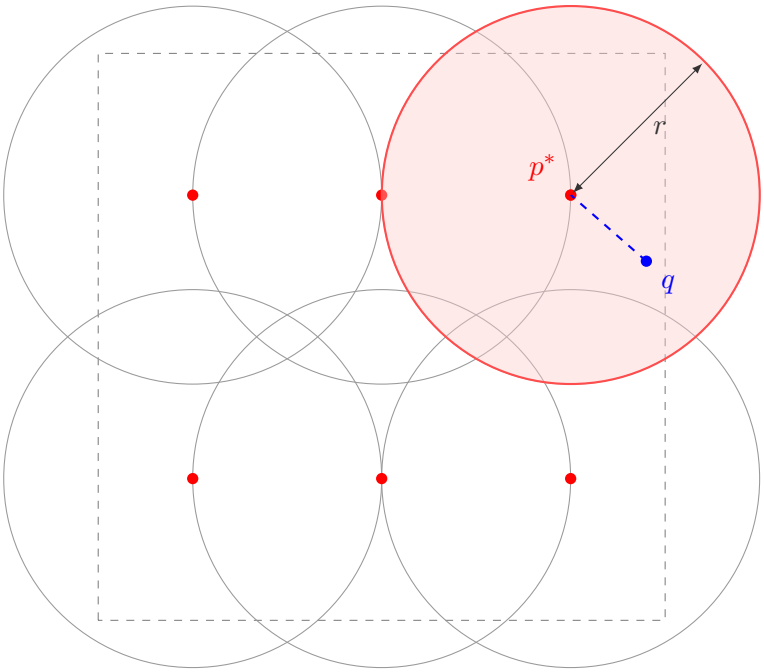

图 4： r-覆盖的示意图。

现在，固定某个 $q \in U$。我们可以写出：

$$q = \sum_{i=1}^{d} \alpha_i u_i$$

对所有 $i \in [d]$，令 $\widehat{\alpha}_i$ 满足 $\alpha_i \in \widehat{\alpha}_i \pm \varepsilon$。令 $\widehat{s} := \sum_{i=1}^{d} \widehat{\alpha}_i u_i$。现在我们有：

$$||q - \widehat{s}||_p^p = \left|\left|\sum_{i=1}^{d} (\alpha_i - \widehat{\alpha}_i) u_i\right|\right|_p^p = \sum_{i=1}^{d} |\alpha_i - \widehat{\alpha}_i|^p \leq d\varepsilon^p$$

现在，我们设：

$$\varepsilon = \frac{\Delta}{d^{1/p}} \implies ||q - \widehat{s}||_p \leq \Delta$$

因此，我们构造的规模为：

$$|\widehat{S}| = \left(\frac{2Cd^{1/p}}{\Delta}\right)^d$$

D.2.2 鲁棒近似最近邻算法

基于这一构造，我们针对鲁棒$(c,\ r)$–近似最近邻的算法，其在 ℓ_p 空间中的实现如下，见算法5。该算法不依赖于具体采用哪种局部敏感哈希数据结构来在 ℓ_p 度量空间中以非自适应方式求解近似最近邻（Charikar，2002年；Datar 等人，2004年），但假设可以通过增加所用哈希函数的数量，提高该数据结构对一组查询的成功概率。汉明范数也满足这一条件。

算法 5 通过离散化实现鲁棒 ℓ_p 近似最近邻

1：参数：最大范数 C、运行时间与准确率权衡 $\Delta > 0$、局部敏感哈希参数 c、$r > 0$。2：从对手处接收带有 $|S| = n$ 的点数据集 $S \subset U$。3：令 $\widehat{S}$ 为按照 D.2.1 节中的构造得到的 Δ覆盖，并令 $c' \leftarrow \frac{cr-\Delta}{r+\Delta}$。4：初始化局部敏感哈希数据结构 $\mathcal{D}$，用于求解 $(c', r+\Delta)$–近似最近邻，并以高概率正确回答 S中的所有查询。5：当对手提供查询时，执行 6：从对手处接收查询 $q \in U$。7：找到满足 $||q - \widehat{s}||_p \leq \Delta$的 $\widehat{s} \in S$。8：在 $\widehat{s}$ 上查询 $\mathcal{D}$，并输出其输出的任何内容。

定理 D.5. 存在一种对抗鲁棒算法，可在(c, r)–近似最近邻问题中，针对$(\mathbb{R}^d, \ell_p)$ 度量空间，回答无限次对抗性查询。假设输入数据集和查询均为 $U = \{x \in \mathbb{R}^d \mid ||x||_p \leq C\}$ 中的元素，其中 $C > 0$，$\widetilde{\ }$，则预处理空间为 $O(nT)$，每次查询的时间为 $O(T)$，其中：

$$T = O\left[d \cdot n^{\rho'} \log\left(\frac{Cd^{1/p}}{cr}\right)\right] \tag{9}$$

其中：

$$\rho' = \frac{(10+c)^2}{161c^2 - 20c - 100}$$

证明。首先，为了论证正确性，令 q 为任意查询。假设存在某个点 $x \in S$，满足 $||x - q||_p \leq r$。那么，根据三角不等式，有：

$$||x - \widehat{s}||_p \leq ||x - q||_p + ||\widehat{s} - q||_p \leq \Delta + r$$

因此，以高概率，$\mathcal{D}$ 会找到某个满足 $||x' - \widehat{s}||_p \leq cr - \Delta$ 的点 $x' \in S$。对于该点，我们有：

$$||x' - q||_p \leq ||x' - \widehat{s}||_p + ||\widehat{s} - q||_p \leq cr - \Delta + \Delta = cr$$

因此，算法 5 将输出正确答案。如果不存在这样的点 x，我们的算法输出 $\perp$ 是有效的，因此证明完成。

对于运行时间，回想一下 $|\widehat{S}| \leq O(2Cd^{1/p}/\Delta)^d$。因此，为了保证 S 中所有 $\widehat{\ }$ 查询都能成功，类似于汉明超立方体的分析表明，$\mathcal{D}$ 占用：

$$O\left[d \cdot n^{1+\frac{1}{2c'^2-1}} \log\left(\frac{2Cd^{1/p}}{\Delta}\right)\right]$$

用于预处理的空间，以及

$$O\left[n^{\frac{1}{2c'^2-1}}\log\left(\frac{2Cd^{1/p}}{\Delta}\right)\right]$$

每次查询处理的时间，其中

$$c' := \frac{cr-\Delta}{r+\Delta}$$

请注意，对于 ℓ_p 空间，我们使用最优 LSH 算法，这保证了 $\rho = \frac{1}{2c^2-1}$。我们的唯一约束是必须满足 $\Delta < cr$。若设置 $\Delta = \frac{c}{10}r$，则每次查询的运行时间为：

$$O\left[n^{1+\frac{1}{2c'^2-1}}\log\left(\frac{20Cd^{1/p}}{cr}\right)\right], \quad \text{where } c' = \frac{9c}{10+c}$$

□

D.2.3 消除对尺度的依赖

定理 D.5 中的算法关键依赖于 $\log C$，其中 C 是查询和输入点空间在 ℓ_p 范数下的包围盒。我们可以通过设计数据依赖的覆盖来消除对 C 的依赖，但需要额外付出一个对数因子。

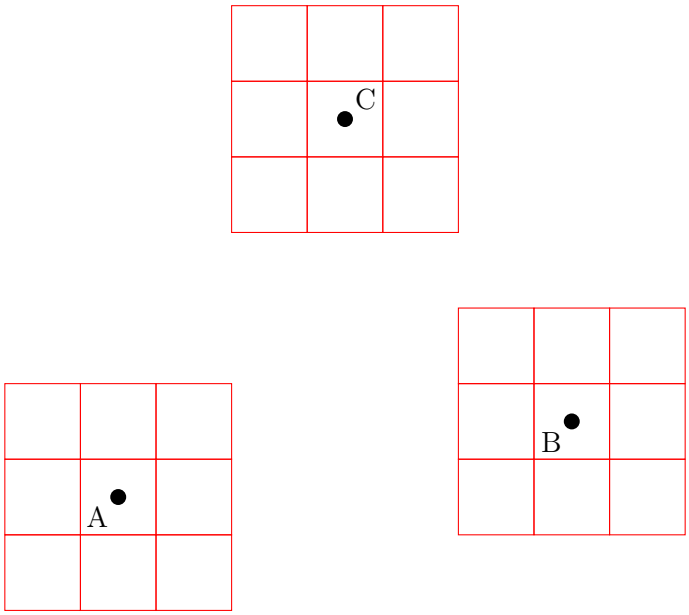

图 5：输入查询空间的数据依赖离散化。

我们的新覆盖 S' 将由 n Δ-覆盖构成，如算法 5所构造，其中每个覆盖都会离散化以点 $p \in S$为中心的 r-球 $p \in S$。这个新覆盖中的点数为：

$$|\widehat{S'}| \le O\left[n \cdot \left(\frac{r \cdot d^{1/p}}{cr}\right)^d\right] = O\left[n \cdot \left(\frac{d^{1/p}}{c}\right)^d\right] \tag{10}$$

请注意，这一覆盖的大小优于 $(nd)^d$ 给出的覆盖大小（见 Chera- panamjeri 和 Nelson, 2024），因此运行时间略有改善。这一新覆盖的显著特点是

它并未覆盖所有可能的查询。不过，它恰好覆盖了我们关心的查询。这一改进后的覆盖为鲁棒近似最近邻提供了如下全称量词保证：

定理 D.6. 存在一种对抗鲁棒算法，可在(c, r)–近似最近邻问题中求解$(\mathbb{R}^d, \ell_p)$ 度量空间，并能回答任意数量的对抗性查询。预处理时间 / 空间为 $O(nT)$，每次查询的时间为 $O(T/d)$，其中：

$$T = O\left[d \cdot n^{\rho'} \left(d \log d + \log n\right)\right] \tag{11}$$

其中：

$$\rho' = \frac{1}{2c'^2 - 1} = \frac{(10 + c)^2}{161c^2 - 20c - 100}$$

证明。 我们区分以下两种情况：

1. 如果查询 q 不属于任何 $B_S(p, r)$（对任意 $p \in S$ 均如此），那么可以放心地将答案设为 $\perp$，因为 $B_S(q, r)$ 必然成立。因此，我们只需运行默认的 LSH 算法，并直接输出其结果。

2. 否则，查询 q 可能属于某个 $B_S(p, r)$（对任意 $p \in S$）（对某个 $s' \in S'$）。那么，假设 $\widehat{s'}$ 是覆盖中的一个点，且满足 $||q - s'||_p \leq \Delta$。则：

$$||p - \widehat{s'}||_p \leq ||p - q||_p + ||\widehat{s'} - q||_p \leq r + \Delta \tag{12}$$

因此，正如我们之前所论证的，高概率下，$\mathcal{D}$ 会找到某个点 $x \in S$，满足 $||x - s'||_p \leq cr - \Delta$，对于该点，我们有：

$$||x - q||_p \leq ||x - \widehat{s'}||_p + ||\widehat{s'} - q||_p \leq cr - \Delta + \Delta = cr \tag{13}$$

这意味着我们的算法将输出正确答案。

与之前一样，我们算法的空间和运行时间保证随 $\log |S'|$ 成比例增长。 □