arXiv:2611.33139v1 [cs.DS] 24 Nov 2026

动态流中的欧几里得直径

JeffGiliberti MohammadTaghi Hajiaghayi马
里兰大学帕克分校{jeffgili,hajiagha}@umd.edu

摘要

我们给出了一种流式算法，用于在任意插入和删除操作下维护高维欧几里得点集的直径；该算法的更新时间关于点数呈多对数级，并关于维数呈多项式级。对于每个固定的 $c > \sqrt{3}$，存在一个$\epsilon_c > 0$，使得我们的单遍算法使用 $\tilde{O}(N^{1-\epsilon_c}\,\mathrm{poly}(d))$ 空间，具有 $\mathrm{poly}(d,\ \log\ N)$ 的最坏情况更新时间，并维护一个因子为-c的成对见证。

在困难性方面，我们证明，即使对于较短的动态流，任何常数因子近似也需要多项式内存：因子为-c的近似在至多包含 $2N-1$ 次更新的数据流上需要 $N^{\Omega(1/(c^2\log(c+1)))}$ 空间。

我们的上界结合了几何凸包、通过逐渐缩小的球冠进行的高斯哈希，以及局部和全局重建计划。每次更新只会改变哈希层级中的一个叶节点，而答案则通过小型方差证书来维护。下界基于一项通信复杂度论证。这些结果共同揭示了多项式内存机制与处理一次更新所需时间之间的差异。

目录

1 引言 2 2 技术概述 3 3 模型与预备知识 7 4 维护直径点对 9 5 样本及其最远距离凸包 14 6 针对外部点的固定过滤器 17 7 从稀疏叶节点到当前候选集 23 8 显式点的直径证书 29 9 冻结样本中的最远点查找 34 10A 线性长度集合流的下界 37

A 附录 **45**

1 引言

点集的直径由两个点决定，但即使只是维护一个粗略估计，也可能需要掌握许多远离当前极值的点的信息。一次删除可能移除一个见证，而一系列删除可能使某个早已被算法丢弃的点重新显现。在本文中，我们研究全动态流模型下的欧几里得直径，其中一次更新会插入或删除一个点。我们的目标是在尽量减少算法必须保留的信息以及处理每次更新所需时间的同时，维护一个估计值和一个实际的见证点对，并重点争取实现关于维数为多项式、关于点数为多对数级的更新时间。

对于仅含插入操作的欧几里得数据流，Agarwal 和 Sharathkumar [AS15] 获得了 $(\sqrt{2}+\varepsilon)$-近似，并指出在高维情况下，若近似因子低于 $\sqrt{2}$，则可能需要线性空间。Halldórsson、Matsakis 和 Veselý [HMV25]达到了相同的上界，同时仅存储$O(\varepsilon^{-2}\log(1/\varepsilon))$ 个点。为处理删除操作，Padaki、Singal 和 Waingarten [PSW26], 在动态数据流框架中采用 Indyk 的非对称嵌入 [Ind03], , 将空间复杂度降至 $\widetilde{O}(n^{2/(c^2-1)})$，近似因子为 c。其直接实现需要更新每个投影坐标。当近似因子为 2 时，内存和更新时间都呈 $n^{2/3}$ 依赖关系（忽略维数和对数参数上的多项式因子）。鉴于 Khanna、Padaki、Singal 和 Waingarten [KPSW25] 针对一般度量，以及 Padaki、Singal 和 Waingarten [PSW26] 针对~欧几里得空间所证明的下界，对于长度和频率不受限制的数据流，任何正确的算法都必须使用 $n^{\Omega(1/c^2)}$ 空间（否则，固定时间跨度会限制算法必须处理的删除历史）。Padaki 等人明确提出：在多项式时间跨度下，是否仍然需要多项式空间 [PSW26, Section 1.2]。

全动态数据结构为处理删除提供了另一种方式。Chan 和 Pathak [CP14]在高维空间中使用包围球，而 Banihashem 等人 [BGG+25] 针对自适应更新，以多项式最坏情况更新时间维护二倍近似的直径。这些全动态数据结构保留输入。这留下了一个悬而未决的问题：流式算法能否使用次线性空间，以及关于维数呈多项式、关于数据流参数呈多对数级的更新时间，维护一个近似的最远点对。

1.1 我们的贡献

设 P_t d 表示当前支撑集，并令 D_t =diam(P_t)，其中 $D_t = 0$ 当 $|P_t| \le 1$ 时。Wr ite

$$\mathcal{L} = \big(d, \log(N+2), \log(\Delta+2), \log((T+2)/\delta)\big).$$

这里，$N \ge 2$ 是同时处于活动状态的不同点的最大数量的给定上界，而 T 是给定的更新范围。近似约定为 $D_t/c \le L_t \le D_t$。空间以普通机器字为单位；模型和算术运算的处理方式将在下文说明。

定理 1.1 (多项式更新时间)。 对于每个固定的 $c > \sqrt{3}$，存在一个常数 $\epsilon_c > 0$ 以及一个随机化单遍算法，其

$$\text{space } N^{1-\epsilon_c}\,\mathrm{poly}_c(\mathcal{L}), \qquad \text{worst-case update time } \mathrm{poly}_c(\mathcal{L}).$$

对于最多包含 T 次更新的每个无自适应严格转向流，以至少 $1-\delta$ 的概率，该算法始终同时维护 $D_t/c \le L_t \le D_t$ 。答案会被缓存。在每个随机结果下，每个非零答案都是两个当前原始点之间的距离。资源界在每个随机结果下同样成立。

二倍保证	内存	更新时间	输出
高斯嵌入基线	$N^{2/3}$ 多项式($\mathcal{L}$)	$N^{2/3}$ poly($\mathcal{L}$)	缓存标量
定理 1.1	$N^{1-\epsilon_2}$ poly($\mathcal{L}$)	poly($\mathcal{L}$)	当前见证对

表 1：两个上界针对不同的时间—空间权衡。在此次比较中，二者都使用峰值界N 以及无知有限时域。附录 A.3 给出了已确立嵌入基线的实现。定理1.2 排除了多对数内存，但无法区分这些更新时间。

该定理包含因子 2。第 A.1 节给出了一个完整的数值选择，其中 $\epsilon_2 = 10^{-10}$。这个较小的指数是保守的存在性界；我们的贡献在于更新时间中不含 N 的正幂。我们不优化该指数，也不声称由此得到的算法具有实用性。查询可在常数时间内返回指向已存储得分和见证记录的指针。写出精确得分、点键或坐标则需要相应的多项式输出时间。

定理 1.2（短流空间下界）。 对于任意固定的 $c \geq 1$，任何以误差至多为 $1/10$ 返回因子为c 的估计值的随机化流式算法，都需要

$$N^{\Omega(1/(c^2 \log(c+1)))}$$

个内存字（最坏情况下）。该下界适用于峰值基数至多为 N、更新至多为 $2N-1$ 的数据流，其中所有插入都先于所有删除，并且每个前缀上的频率均属于 $\{0, 1\}$。这些点的维数为 $O_c(\log N)$，且整数坐标的绝对值为 $O_c(\sqrt{\log N})$。只要求进行一次最终查询，而且算法可以是非线性的、依赖历史的，且计算能力不受限制。

第二个定理中的字长相对于 N 是多对数级的，这与这些实例在标准 RAM 模型中的情况相同。证明首先对状态中的信息量给出界；除以字长后，所述指数仍然成立。特别是，对于每个固定的近似因子，该定理排除了使用 poly(d, $\log N$, $\log \Delta$, $\log T$) 内存的可能性。这是一个空间下界，并未给出任何正幂级的更新时间下界。

相对于空间下界的新意。下界贡献在于所采用的模型和证明，而非对 c 方面的依赖关系作出改进（相较于 [PSW26]）。我们使用单个插入前缀和单个删除后缀，不涉及重数或漫长的历史记录。任意流式状态都可以转化为有限通信问题中的一条消息。其条件熵直接给出所需的次模函数，因此无需归约到线性草图。上界贡献则与之互补：常数因子近似可以使用多项式内存，而无需在 N 每次更新时花费多项式时间。表1 中的因子二空间指数与下界并不匹配；最优的因子二时空权衡仍有待解决。

2 技术概述

小样本可能会漏掉最远点对的一个端点，但仍能告诉我们应该去哪里寻找这样的点。我们利用样本界定空间区域内的距离，恢复该区域外的点，并维护由样本点和恢复点组成的当前点对。下界

利用一个紧密聚集的点簇和一个远处的点，说明算法必须记住多少关于先前插入的信息。

2.1 样本及其凸包

从当前不同点中随机抽取一个小样本 S 。遗漏许多点的球不太可能包含整个样本。关于欧几里得球的抽样界使这一结论同时适用于所有中心和半径，因此即使球是在看到 S 后才选定的，这一结论仍然适用。

固定一个膨胀因子 $a > 1$。将每个包含 S 的球以其自身中心为基准膨胀 a 倍，并取这些膨胀球的交集。将该交集称为$\mathcal{H}_a(S)$。如果某点位于这个凸包之外，就存在一个球包含 S ，而该点到球心的距离大于球半径的 a 倍。抽样界说明这个球包含了大部分输入。另一个有用性质是引理 5.1给出的直径界。

$$\operatorname{diam}(\mathcal{H}_a(S)) \le \sqrt{2a^2 - 1}\,\operatorname{diam}(S).$$

凸包中的一个点与某个定义在 S 上、方差为 V 的分布的均值之间的距离至多为 $\sqrt{(a^2-1)V}$。对于两个这样的分布，它们均值之间的平方距离加上它们的方差，等于分别从这两个分布中独立抽样所得样本之间的期望平方距离，后者至多为 $\operatorname{diam}(S)^2$。于是，由柯西–施瓦茨不等式可得凸包的直径界。

样本直径涵盖凸包内部的点对。我们将每个外部点恢复到候选集中，从而涵盖凸包外部的点对。对于一对内外混合的点，以外部端点为视点观察凸包，可利用 S 中的最远点界定到内部端点的距离。为找到外部点，我们使用独立于样本抽取的固定过滤器，因此可以针对看到 S 后选定的见证球进行分析。

2.2 恢复凸包外部的点

考虑一个见证球和一个位于球外的目标点。过滤器应将该目标点分配到一个桶中，且该桶内的其他点较少。我们使用多个相互独立的高斯投影。每个投影都配有一个随机平移网格，并通过哈希将网格单元映射到桶。桶标识符组成的元组确定一个带有稀疏恢复草图的叶节点。该分配仅取决于点和预先选定的映射，因此每次插入或删除都会在每个尺度和重复中改变一个叶节点。

分析研究见证球中哪些点可能与目标点落入同一个投影单元。将球归一化为单位半径，并考虑一个与球心距离至少为 $\lambda > \sqrt{2}$ 的目标点。如果球内某点沿目标轴的坐标为 t ，则其垂直分量的平方至多为 $1 - t^2$。在给定目标轴投影的条件下，该垂直分量提供独立的高斯噪声。轴向分离距离的平方超过其方差尺度，且具有正的余量，因为 $(\lambda - t)^2 - (1 - t^2) \ge \lambda^2/2 - 1 > 0$。碰撞分析利用这一余量，将同单元界较弱的点限制在一个朝向目标点的球冠内。该球冠可容纳在一个较小的球中（图 1）。平方半径的上界在每个阶段都会减少一个固定量，因此经过常数个阶段后，难处理区域便为空。分析在每个阶段都选择目标轴上的一个覆盖中心，而算法对所有目标点使用相同的映射。

为了界定目标叶节点上的负载，按照主体点首次离开嵌套困难区域的阶段进行划分。在该阶段，点位于覆盖球内，与目标落入同一网格单元的概率很小。以常数概率，哈希会将目标所在的桶与投影核心中的其他单元隔离开来。两者结合后，会产生一个桶计数因子的倒数，以及一个随投影维度呈指数增长的额外节省。在后续阶段，同单元概率仍然很小，

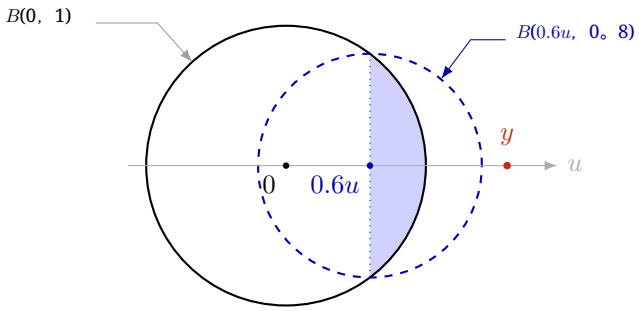


图1：两阶段构造的横截面。第一网格中的困难点位于单位球的阴影帽 $t \geq 0.6$ 中。虚线球包含整个球冠，并确定第二网格的较小半径。其中心和目标方向 u 仅用于分析。

而不同单元只有通过哈希冲突才能共享一个桶。因此，从退出阶段开始产生的桶计数因子的倒数，会抵消我们对所有叶节点的容量求和时相应的因子。我们按几何级数递增投影维度，使退出阶段的额外节省足以抵偿早期阶段的桶计数，并且还多出一个 N 的幂次因子。

叶节点容量还必须使原始球外的每个点都能够发生碰撞。当采样概率为 N^{s-1}（其中 $s > 0$为一个小常数）时，异常点数量至多为 N^{1-s}乘以维数和对数参数的多项式因子。使用 $O(N^{s/2})$ 个叶节点也能使这些异常点所占的空间保持次线性。第6 节给出了负载和空间计算。

我们的恢复论证使用以 $\sqrt{2}$ 见证半径为单位的分离度。取 a 略高于 $\sqrt{2}$，即可得到略高于 $\sqrt{3}$的凸包直径因子 $\sqrt{2a^2 - 1}$ 。

2.3 维护当前见证

一个稀疏叶节点可能包含过多的点，无法在一次更新中完成解码。我们以对其自身叶节点的更新次数来衡量快照的年龄。容量为 K 的叶节点每经过 K 次本地更新就获取一次快照，并在接下来的块中解码和重放变更。其服务快照至多有 $2K$ 次本地更新的延迟。如果该叶节点当前至多有 K 个点，则该快照中至多有 $3K$ 个点，处于解码器的处理预算之内。没有收到更新的叶节点无需进一步工作即可保持其服务字典为最新。

解码器根据精确矩和总质量，验证其快照的标识及完整性。有符号变更日志会将每个恢复的重数转换为其当前值。当一个点出现在多个叶节点中时，其条目共享一个当前计数。原始删除会在后台工作运行前更新该计数，因此，即使解码尝试失败，也只会影响哪些当前点可作为候选项。第7节介绍重构及其调度安排。

记 C 为最终得到的当前候选项集合。每次更新后都扫描 C 会导致超出时间界限。对 C 建立动态层次结构，可维护一个近似于 $\mathrm{diam}(C)$ 的点对，并提供一小组覆盖球。每个球由少数当前点及其凸权重描述，这些点是球的锚点。球心 q 是这些点的加权均值，其半径接近其方差 V的平方根。对于被此球覆盖的候选项 x 和凸包上的点 y，三角不等式可将 $\|x - y\|$ 界定为该球的半径加上 a 乘以距 q最远样本的距离。

我们可以通过查询样本中的一个点 z来近似得到最远距离。要将返回结果转化为一对输入点，请取距离 z 最远的锚点 p 。加权平均

锚点到 z 的距离平方得到

$$\|p - z\|^2 \geq V + \|q - z\|^2.$$

因此，同一对点既决定了半径，也决定了从中心到样本的距离。根据柯西–施瓦茨不等式，随着球和查找的精度提高，混合对因子可以任意接近 $\sqrt{1 + a^2}$ 。对于 C 中的点对，层次结构的因子可以任意接近 $\sqrt{2}$。当 $a > \sqrt{2}$ 时，这两者都不超过凸包因子。

快速最远点查找采用第二种调度方式，以流更新次数为衡量单位。在一个块开始时，我们冻结活动样本。我们计算其直径，并在该块内基于投影网格构建最远点查找表，然后在下一个块中使用它们。样本大小约为 N^s ，相差多项式因子。取 s 较小值，就可以将其多项式预处理分摊到长度约为 N^{1-s} 的一个块中。删除被采样标识会立即使相应的冻结代失效。对于无自适应数据流，在这两个块中被删除的标识是预先固定的。任何标识被采样的概率至多为其数量乘以采样率。将块长度中的常数取小后，样本以常数概率得以保留。独立地重复整个构造，就能在整个数据流中维持近似。算法会缓存样本直径点对、候选层次结构的见证以及通过查找找到的混合点对中最长者（第 4 节）。

2.4 短集合流的直接下界

删除后缀可以使最终的直径查询揭示插入前缀中的一比特。Alice 为单位球面上精细网中的每个点各持有一个独立比特，并插入比特为 1 的点。她将内存状态发送给 Bob。Bob 得到一个小球冠及其对跖目标点。他知道球冠外哪些点已被插入，也知道目标点。他先在球冠中心附近放置一个新锚点，然后删除他知道已被插入的点。如果目标点不存在，剩余集合的直径较小；如果目标点存在，剩余集合的直径较大。半径与$1/c$ 成比例的球冠能以超过 c倍的差距区分这两种情况。

Bob 对其他比特的了解很重要，因为一条消息可以将多个比特编码在一起。固定 Alice 的随机性后，令 M 为消息，并令 $f(A)$表示 Bob 知道 A以外的比特时， M 的条件熵。这个集合函数是单调且次模的，总值至多为消息长度 s。对于球冠 B和目标点 z，增益 $f(B \cup \{z\}) - f(B)$衡量 Bob 获知球冠外的比特和目标点后，消息提供了多少关于目标比特的信息。如果 z 已在 B中，这一增益为零。Bob 能以有界误差解码对跖目标点，因此这些查询的平均信息增益为常数。

我们将这些信息增益与弦长联系起来。给定若干球冠–目标点对，构造一个表，其中球冠 i 与目标点 j 对应的表项就是它们的信息增益。一个次模不等式将对角线之和的平方限制为整个表之和的 $2s$ 倍。其证明沿着目标点的随机链对角线信息增益进行计费，其中交叉项反映了在更早的目标点已被加入后信息增益的减少。

为了应用这个不等式，将成对帽中心和目标点分别放在一个随机定向的仿射子球面直径的两端 t。这样，每条成对弦的长度都是 t，而交叉弦的长度通常接近 $t/\sqrt{2}$ （图3）。如果 a（t）表示弦长为 t时的平均信息增益，那么通过旋转平均可得 a（t）$^2 \leq 2s\,\mathbb{E}[a$（$t\sqrt{Z_0}$）$]$, 其中 Z_0 是随机交叉弦与 t 的平方比值，并且集中在 $1/2$附近。证明允许表中出现重复目标，包括舍入到有限网时产生的重复目标。

从对跖点处的正信息增益出发，只要 s 足够小，递推关系就会迫使较短弦上也有正信息增益。每一步都会将信息增益平方，因此随着信息增益变小，证明会相应调整 Z_0 的尾部截断。这样最终会到达目标点已位于球冠内、信息增益为零的范围，除非消息包含$\exp(\Omega(d/c^2))$ 比特。有限球面网包含$\exp(O(d\log(c+1)))$个点。每个位置最多插入一次、删除一次，因此将比特界转换为空间界只需使用线性长度的集合流（第10节）。

2.5 更多相关工作

几何草图。 Jiang、Krauthgamer 和 Sapir [JKS24] 在动态数据流中采用中等维数缩减来处理 k 中心问题。在多项式长度数据流上，他们获得 $O(\alpha)$ 近似，使用 $n^{1/\alpha^2}\,\mathrm{poly}(kd\log\, n)$ 比特。单中心半径与直径相差至多两倍，因此他们的结果也给出了 $O(\alpha)$ 直径估计。他们的两级采样器还可以返回原始输入标识。我们的过滤器会筛选出彼此间距达到指定阈值的外部点，并且每个层次结构需要进行一次叶节点更新。随后，凸包给出一个从上方任意接近 $\sqrt{3}$ 的显式直径因子。几何哈希也已用于动态数据流设施选址 [CFJ+23]，而流式核则处理 k带离群点的中心问题 [dBBM23]。

删除与代表元。 Banihashem 等人 [BGG+25] 为完全存储的动态集合维护一个鲁棒代表元。在这里，小样本的代表元可能因一次删除而消失。我们会使其所属代失效，并在无自适应数据流上界定这一事件的概率。Blank 等人 [BCH+26] 在过期模型下获得了欧几里得直径的 $(1+\sqrt{3}+\varepsilon)$ 近似；在该模型中，项目的删除时间在其到达时给定。在我们的下界中，选择哪些项目过期会编码解码器的问题。

从流式状态到消息。Khanna 等人的一般度量空间和欧几里得下界 [KPSW25] 以及 Padaki 等人 [PSW26] 将流式计算与线性草图和索引编码联系起来。Li、Nguyen 和 Woodruff [LNW14]，Kallaugher 和 Price [KP20]，以及 Jiang、Liu 和 Yu [JLY26] 研究流式算法何时可以通过线性草图模拟实现。后一项工作涵盖多项式长度的严格转门数据流，并允许中间重数超过 1。我们将状态本身用作一条始终保持二进制的数据流上的消息。条件熵是非线性索引编码中常用的工具 [BKL13]。这里的几何步骤将两个弦长处的边缘分布联系起来，并且在舍入到有限输入全集后仍然成立。

3 模型与预备知识

一次更新为（x，σ），其中 $x\in[-\Delta,\Delta]^d\cap\mathbb{Z}^d$ 且 $\sigma\in\{-1,\,1\}$。数据流从空状态开始，所有频率始终非负。支撑集 P 由频率为正的位置组成。每次更新都会给出其对应点的坐标。算法没有支撑集的外部副本。无自适应数据流独立于所有随机选择和回答预先固定。我们假设 $N\geq 2$、T，Δ，$d\geq 1$和 $0<\delta<1/10$。对于仅取决于固定近似参数的 N 有界范围，可以存储该集合及其计数，并直接计算其直径。

算术运算与调度。在技术章节中，我们将参数简写为

$$\mathcal{L}=(d,\log(N+2),\log(\Delta+2),\log((T+2)/\delta)).$$

我们使用字长为 $w_{\mathrm{word}} = \Theta(\log(d{+}2){+}\log(N{+}2){+}\log(\Delta{+}2){+}\log((T{+}2)/\delta))$ 的字 RAM。我们用$\ln$ 表示自然对数，$\log_2$ 表示二进制对数，用$\log$ 表示渐近界中一个固定底数的对数。我们将多字算术运算、随机比特、平衡字典操作和内存回收计入开销。得分以精确的平方距离和两个点键的形式存储，默认值为零。重复试验使用独立种子。每个有预算限制的后台作业会在更新之间保存其执行位置。对一次更新的所有处理都会在其答案可用之前完成。渐近界指定的预算和样本大小均使用足够大的固定常数，并向上取整为整数，除非另有说明。附录 A.2 给出了随机几何测试所需的精度。

上界构造也支持规范有理坐标，其中分子和正分母至多为 b 比特。此时，不同点之间的距离至少为 2^{-2b} ，且都位于$[-2^b, 2^b]^d$中。第一个半径尺度为 2^{-2b-1} ，尺度数量为 $O_c(b + \log(d + 2))$。键长度、精确有理数算术运算和网格精度均使用 b ，而非整数坐标的比特上界。为便于比较，我们将结果表述为整数网格上的情形。

证明中使用的工具。以下形式给出了我们所用的抽样和极小极大结果。其他几何和恢复例程将在下文构造。

引理 3.1 (球的半网，Haussler 和 Welzl [HW87])。设 A 是 $\mathbb{R}^d$ 中的一个有限集，且$0 < \xi < 1$。无放回均匀抽取大小为 $\min\{|A|, \lceil C(d + \log(1/\xi))\rceil\}$ 的样本，以至少 $1-\xi$的概率具有以下性质：每个包含该样本的欧几里得球都包含 A中至少一半的点。

事实上，球的补集的 VC 维数为 $v = d + 1$。对于 m 次独立抽样，Haussler 和 Welzl 的双重抽样界 [HW87，定理 3.7] 给出了漏掉相对大小超过 $1/2$ 的补集的概率上界 $2(\sum_{j=0}^{v} \binom{2m}{j})2^{-m/4}$ ，其中 $m \geq 16$。对于 $m \geq Cv$ ，总和至多为 $(2em/v)^v \leq 2^{m/8}$，其中 C 是一个绝对常数。取 $m \geq C(v + \log(1/\xi))$ 可使失败概率至多为 ξ。若 $m \geq |A|$，我们就取全部点。否则，可以保留不同的独立抽样结果，并以均匀方式补足至 m 个不同点。根据对称性，这构成无放回均匀样本，而补足过程会保留每一次命中。

引理 3.2 (有限独立矩界 [SSS95])。 若 Y 是 $2k$个均值为 μ的两两独立伯努利变量之和，则存在绝对常数 C 和整数 $k \geq 1$，使得 $\mathbb{E}(Y - \mu)^{2k} \leq (Ck\mu + Ck^2)^k$。

中心矩是次数至多为 $2k$ 的指示变量多项式，因此与完全独立条件下的取值一致。对于独立变量，通常的双侧伯恩斯坦界至多为 $2\exp(-t^2/(4\mu)) + 2\exp(-3t/4)$，在偏差 $t > 0$且 $\mu > 0$时成立。对$2kt^{2k-1}$ 进行积分，分别得到至多为 $(Ck\mu)^k$ 和 $(Ck^2)^k$阶的项。 $\mu = 0$ 的情形是直接的。第 5.3 节给出了有限比特采样族。

引理 3.3 (极小极大，Sion [Sio58])。 对于有限维实向量空间中的紧凸集 U， V 以及一个在 g 上为凹函数、在 $u \in U$ 上为凸函数的连续函数 $v \in V$，有$\max_{u\in U} \min_{v\in V} g(u, v) = \min_{v\in V} \max_{u\in U} g(u, v)$。

我们将这一恒等式应用于样本上的分布单纯形以及可能中心的有界球。它仅用于分析。

4 维护直径对

该算法出于不同目的维护两个集合。冻结样本 S 为其扩展凸包中的点对提供见证。动态集合 C 包含外部点，也可能包含其他当前点。我们为 C 维护得分，并将其球心与 S进行比较。下面的构造证明了定理 1.1。确切的近似因子和空间指数见引理 4.1。

4.1 算法

对于所需的因子 $c > \sqrt{3}$，我们为凸包选择一个有理数 $\sqrt{2} < a < \sqrt{(c^2+1)/2}$，为候选球选择一个有理精度 $0 < \eta \le 1/4$，并为最远点查找选择一个有理数 $\kappa > 1$。我们选择 η 足够小，并使 κ 足够接近于一，具体理由见定理 1.1的证明。如果$B_\kappa \ge 2$ 是引理 9.1的预处理指数，我们固定有理指数

$$0 < s < 1/(B_\kappa + 1), \qquad 0 < \rho \le s/2. \tag{1}$$

样本哈希以固定的二进概率 $p \in [N^{s-1}/4, N^{s-1}]$选择每个位置。输入从一个空的活动样本、一个空的候选层次结构以及零得分开始。在首次更新之前，我们用 $\mathrm{poly}(\mathcal{L})$时间初始化恢复映射、哈希和时钟。矩表和查询条目按需分配。我们使用足够大的常数乘以 $\lceil \log((T+2)/\delta) \rceil$个完整副本。精确的二倍参数见附录A.1。

算法 1 使用以下接口。

- 抽样和最远点查找 （引理 5.2 和 9.1）。我们存储正采样位置及其计数。在每个时期边界 τ（包括时刻零），我们冻结一个持久根 S_τ。在接下来的 $H_{\text{时期}}$ 更新期间，我们计算其精确直径、一个最大化对及其查找表。空样本不需要预言机。生成的代次从$\tau + H_{\text{时期}}$ 开始提供服务，直到下一次安装；此时

$$H_{\text{epoch}} = \max\{1, \lfloor N^{1-s}/100 \rfloor\}.$$

  删除任何被冻结的样本标识都会永久使该代次失效，即使该位置仍保留另一次出现也不例外。我们会同时检查构建中的代次和提供服务的代次。对有效表的查询会返回一个样本标识。它不会更改样本，也不会影响任何后续更新。

- 固定过滤器和局部重构（定理 6.1 和引理 7.2）。我们选择 $\sqrt{2} < \lambda < \min\{a, \sqrt{3}\}$，并维护比率为$1 < q_{\text{scale}} \le a/\lambda$、从 $1/2$ 到不低于 $2\Delta\sqrt{d}/(a-1)$ 的第一个尺度的几何尺度。每次更新都会触及每个尺度和重复中的一个叶节点。每个叶节点都根据自己的更新时钟重构其快照。引理 5.2 的边界异常预算为 $b_0 = CN^{1-s}((d+2)\log(N+2) + \log((T+2)/\delta) + 1)$。对于所有服务时刻，我们使用

$$b = b_0 + 2H_{\text{epoch}}. \tag{2}$$

  容量为 $\lceil b + 2 \rceil$ 的后备桶会处理每次更新。每个活动采样点也有一个候选参考点。所有参考点变更（包括退役操作）都进入同一个有序队列。转换在被处理前会保留其原始点键。

- 动态球证书 （引理8.1）。我们将队列中的每次候选插入或删除应用于一个单独随机化的层次结构。该结构维护一个见证分数 L_C，并提供多项式数量的证书。每个证书包含少量当前锚点 K、有理凸权重、均值 q和方差 V。在该层次结构的成功事件中，这些球 $B(q, (1+\eta)\sqrt{V})$ 覆盖 $C_{\text{+}}$

算法 1 在单点更新后维护见证

输入 合法更新（p，σ）以及前一次更新后的状态
输出 更新后的状态，以及在得分为正时带有当前端点的缓存分数

对于每个启用的独立副本**执行**
  如果已记录，则调整 p 的全局计数，然后更新其活动样本引用
  若 $\sigma = -1$，则使包含 p 的每个冻结代失效
  如果超出样本上限，则禁用此副本并继续处理下一个
  对每个尺度和恢复重复次数**执行**
    将 p 哈希到其叶节点，并应用算法 2 处理 (p, σ)
  结束循环
  将相同的更新应用于全局回退桶
  推进外部引用回收，然后按顺序处理 C 的层次结构中的所有排队的转换
  按照各自的独立预算推进样本预言机任务和私有回收
  如果一个时期结束，**则**
    安装已完成的代及其有效性标志，并冻结下一个样本
  结束如果
  从见证候选分数 L_C 开始（若无则为零）
  如果有效的非空冻结样本 S 正在使用，**则**
    纳入其预先计算的直径和最大化对
    对于每个暴露锚点集 K（中心为 q），包括单例叶节点，**执行**
      在 q 处查询 S，以获取 z，并纳入 $(x,\ z)$，其中 $x \in K$ 是距离 z 最远的
    结束循环
  结束 if
  缓存最大的纳入平方距离及其端点
结束 for
存储启用的副本中的最佳缓存；如果没有副本具有正分数，则存储零值

包括 $q = x$ 和 $V = 0$ 的单元素证书。队列清空后，我们针对当前服务代次为已暴露证书评分。这也涵盖了同一次原始更新中新暴露的叶节点以及纪元切换。

查询可在常数时间内读取由此得到的最大值。其句柄指向当前缓存。复制两个端点的开销为其输出长度。禁用的内部层次结构既不贡献中心，也不贡献得分。如果完整副本中的所有内部层次结构都不可用，我们取 $L_C = 0$，仍可使用该副本有效的样本直径。内部故障不会影响候选转换的顺序。

4.2 分析

引理 4.1 (直径见证的组合)。 对于固定的有理数 $a > \sqrt{2}$， $0 < \eta \le 1/4$，$\kappa > 1$ 以及满足 (1) 的有理指数，存在一个 $\epsilon_0 > 0$，使得算法 *1* 使用的空间为

$$\left(N^{sB_\kappa} + N^{1-s+\rho} + N^{1-\epsilon_0}\right) \operatorname{poly}(\mathcal{L}) \tag{3}$$

并且在每种结果下，最坏情况更新时间为 $\operatorname{poly}(L)$。以至少 $1-\delta$ 的概率，在整个数据流过程中，其见证答案均满足

$$L \le \operatorname{diam}(P) \le c_* L, \qquad c_* = \max\left\{\sqrt{2a^2-1},\ \sqrt{(1+\eta)^2+1},\ \sqrt{(1+\eta)^2+a^2\kappa^2}\right\}. \tag{4}$$

证明。 我们首先说明证书必须覆盖哪些点，然后说明每个成对见证如何对应三个因子之一。证明的最后一部分检查条件化以及两个重建截止期限。

服务时刻的覆盖。 根据引理 5.2，每个包含 S_τ 的球在其边界上至多排除 b_0 个点。服务结束前至多发生 $2H_{\text{epoch}}$ 次插入，因此即使球是在之后选定的，(2) 也能界定其异常点。假设 $|S_\tau| \ge 2$。对于每个 y 位于其凸包之外的点，半径界 (13) 给出了一个见证，其半径为 r，且处于维护范围内。向上舍入可保留球的大部分部分，并使目标点与其中心的距离超过 λ 个舍入后的半径。在给定原始数据流和样本的条件下，我们可以在抽取网格之前固定此见证。对网格进行独立重复，并对点、代次和时间应用并集界，可确保每个这样的 y 都有一个大小至多为其容量的叶节点。给定网格后，局部快照时间便固定下来，且独立于其私有解码哈希。因此，引理 7.2 给出

$$P \setminus \mathcal{H}_a(S_\tau) \subseteq C \subseteq P. \tag{5}$$

右侧的包含关系在每个结果上都成立。在第一个预言机就绪之前，$|P| \le H_{\text{epoch}}$。对于空边界样本或单元素集合边界样本，网事件给出的边界支撑至多为 $b_0 + 1$，因此当前支持集至多为 $b + 1$。在每种情况下，后备桶都在同一恢复事件上给出 $C = P$。

见证分数。 对于有效的服务样本，设 $s_* \in S_\tau$ 为预言机在证书中心 q处的回复。该证书的贡献为

$$F = \|q - s_*\|, \qquad L_{\text{node}} = \max_{z \in K} \|z - s_*\|. \tag{6}$$

我们保留达到 L_{node} 的那一对。其平方距离不小于锚点上的加权平均，即

$$L_{\text{node}}^2 \ge \sum_{z \in K} \lambda_z \|z - s_*\|^2 = V + F^2. \tag{7}$$

因此，对于保留的样本，缓存答案为

$$L = \max\left\{L_C, \operatorname{diam}(S_\tau), \max_{e \text{ exposed}} L_e\right\}, \tag{8}$$

其中 L_e 是已暴露证书 e的得分。若样本缺失、无效或为空，则仅使用 L_C 。每个可用得分都有当前端点。在显式层级结构成功时，$\operatorname{diam}(C) \le \sqrt{(1+\eta)^2+1}\, L_C$。

固定一个时刻，此时冻结样本仍然有效，且其网、恢复、显式层级结构和预言机事件均成立。若冻结样本为空或为单元素集合，或者处于启动阶段，则全局回退给出$C = P$，而显式点层级结构足以满足要求。否则，考虑一个直径点对 x， y。

若两个点都位于$\mathcal{H}_a(S_\tau)$中，则由凸包引理可得

$$\|x - y\| \le \sqrt{2a^2 - 1}\, \operatorname{diam}(S_\tau) \le \sqrt{2a^2 - 1}\, L.$$

若两个点都在凸包之外，则根据（ C 5），它们都属于，且

$$\|x - y\| \le \sqrt{(1+\eta)^2 + 1}\, L_C.$$

对于剩余情形，假设 $x \in C \setminus \mathcal{H}_a(S_\tau)$ 且 $y \in \mathcal{H}_a(S_\tau)$。有效证书覆盖 x。叶节点情形使用 $q = x$ 和 $V = 0$。在证书中心 q 处评估凸包成员关系，可得$\|q - y\| \le a R_{S_\tau}(q)$。最远点保证给出

$$\begin{aligned}
\|x - y\| &\le (1+\eta)\sqrt{V} + a R_{S_\tau}(q) \\
&\le (1+\eta)\sqrt{V} + a\kappa F \\
&\le \sqrt{(1+\eta)^2 + a^2\kappa^2}\,\sqrt{V + F^2} \\
&\le \sqrt{(1+\eta)^2 + a^2\kappa^2}\, L_{\text{node}}.
\end{aligned}$$

方差项不会遭受预言机损失。取这三种情况中的最大值即可证明（4）。

条件化与存活。 在任意固定的服务时刻，无自适应数据流指定的自边界以来被删除的标识不超过 $2H_{\text{epoch}}$ 个。每个标识被选中的概率至多为 p。在 N 的渐近范围内，

$$\Pr[\text{serving sample invalid}] \le 2H_{\text{epoch}} p \le 1/50. \tag{9}$$

我们将这个无条件界与覆盖概率结合使用。看到见证的对手可能会故意删除一个采样端点，而这正是此论证依赖无自适应性的地方。我们首先固定原始的无自适应数据流以及所有样本、网格和稀疏恢复种子。这些共同确定候选序列及其操作调度。每个候选操作都补齐到显式原语已知的最坏情况运行时间界。外部引用退役和内部几何回收使用彼此独立的队列和固定预算；内部私有分配绝不会改变外部清理顺序。因此，候选序列独立于层级结构的私有随机性。显式点定理适用于这个至多包含 $T\text{poly}(\mathcal{L})$ 个操作的固定序列。

接下来固定层级结构种子。它们决定每个已暴露中心、其有效区间及其查询时间。这些都不依赖于静态预言机答案。冻结预言机构造使用其规定的工作预算和安装截止时间，预言机操作也可以进行填充。因此，

对给定冻结预言机的完整有限查询列表独立于该预言机的种子而固定。对整个预言机进行重复，即可为所有这些查询提供所需的保证；其数量的对数是$\mathcal{L}$的多项式。其有限精度针对有界有理中心域进行选择。

我们选择网、恢复、层次结构和预言机失败预算，使得它们在固定时刻的总和至多为 $1/10$。将它们与（9）结合，并应用并集界，可为一个完整副本提供恒定的成功概率。无需这些事件之间相互独立。我们运行 $O\left(\log\left(\left(T+2\right)/\delta\right)\right)$ 个相互独立的完整副本，并返回可用的当前配对分数中的最大值。对于固定数据流，这些副本相互独立，因此在固定时刻所有副本同时失败的概率会随副本数量增加而呈指数级减小。对 $T+1$ 个输出时刻应用并集界，即可证明所声称的同时保证。取最大值不会产生高估。

最坏情况资源界。 表 2 记录了一个副本中所维护结构的开销。

结构	空间至多为 poly$(\mathcal{L})$	计入一次更新的工作量
活动样本	N^{s}	一次成员测试和一次计数变更
冻结查找	N^{sB_κ}	一个时期所分摊的预处理，加上中心查找
几何叶节点	$N^{1-s+\rho}+N^{1-\epsilon_0}$	每个尺度和每次重复中各触及一个叶节点
后备桶	N^{1-s}	一项局部重构预算
候选层次结构	M	多项式数量的转移和已暴露中心

表 2：一个完整副本中的开销。 M 界定当前候选项的数量，包括中间转移。对于冻结查找，重构的节奏由全局更新控制；对于解码器，则由对每个叶节点的更新控制。

我们对候选流的大小（包括中间转移）取确定性上界 M，并将其长度界定为 $T_C \le (T+1)\,\mathrm{poly(L)}$。定理 6.1 和局部编译器给出 $M = O(N^{1-s+\rho}+N^{1-\epsilon_0})\,\mathrm{poly(L)}$，其中包括大小为 $O(b)$ 的回退项和活动样本。此处，$s < 1/(B_\kappa+1)$ 可确保 $s < 1-s+\rho$。因此，引理 8.1 的空间开销为 $M\mathrm{poly(L)}$，会暴露 $O(R_{\mathrm{hier}}(\log(M+2)+H_{\mathrm{leaf}}))$ 个 $=\mathrm{poly(L)}$ 中心，并以最坏情况多项式时间处理每次转换。这里，R_{hier} 是层级副本的数量，H_{leaf} 将精确叶大小界定在一个常数因子以内。每个中心都是有界有理凸组合，因此其编码符合引理 9.1 的查询契约。样本上限为 $N^{s}\,\mathrm{poly(L)}$。因此，构建静态预言机（包括精确样本直径和重复）需要 $N^{sB_\kappa}\,\mathrm{poly(L)}$ 时间和空间。将其总工作量界定为 $W = N^{sB_\kappa}P(\mathrm{L})$，其中 P 是一个固定多项式，并将其分摊到下一个时期。由于 $H_{\mathrm{epoch}} \ge N^{1-s}/200$，每次更新的预算满足

$$\lceil W/H_{\mathrm{epoch}} \rceil \le 1 + 200N^{s(B_\kappa+1)-1}P(\mathcal{L}) = O(1+P(\mathcal{L})) \tag{10}$$

按（1）计算。该界包含快照遍历、高斯映射生成、查找枚举、样本扫描以及所有放大表。我们保留常数数量的代次，以常数元数据工作量退役旧根，并在单独的固定全局预算下回收其记录。由于每次更新的记录分配量是多项式，突发加速率的退役论证最多会使未退役存储量再增加一个常数倍。

叶构造遵循局部时钟。每个被触及的叶节点都会获得固定的多项式预算，与其容量无关；其 $K\mathrm{poly}(\mathcal{L})$ 解码和实体化工作分摊到对该叶节点的 K 次更新中。每次原始更新会触及多项式数量的叶节点。编译器只生成多项式数量的候选转换，每个转换均由最坏情况多项式显式点

原始操作处理。对已暴露证书进行完整枚举，每个证书执行一次预言机查询和多项式数量的锚点比较，也符合此上界。查找任务使用（3）中的第一项，过滤器和候选层次结构则使用另外两项。对样本数量、链深度和工作分配的限制，会在溢出可能导致存储上界增大之前生效。这就证明了该引理。

□

定理 *1.1* 的证明。 对于 $\sqrt{2} < a \leq \sqrt{(c^2+1)/2}$，凸包因子严格小于 c。当 $\eta = 0$且 $\kappa = 1$时，平方混合因子为 $1 + a^2 < 2a^2 - 1$。由连续性可知，存在固定的有理数 $\eta > 0$ 和$\kappa > 1$，使得另外两个因子都小于凸包因子。取 s，ρ 如式（1）所示，我们可以令

$$\epsilon_c = \min\{1 - sB_\kappa,\ s - \rho,\ \epsilon_0\} > 0.$$

引理 4.1 由此给出所需的空间、更新时间和同时保证。有限位实现见附录A.2。

□

5 样本及其最远距离凸包

算法 1 的证明用到了其冻结样本的两个性质。凸包限制了两个端点均为内部端点的点对长度。抽样则限制了每个能够作为外部端点见证的球所对应的异常点数量。凸包和见证球都是分析对象。

5.1 凸包内的距离

对于非空有限的 $S \subset \mathbb{R}^d$ 和 $a > 1$，记

$$R_S(q) = \max_{s \in S} \|s - q\|, \qquad \mathcal{H}_a(S) = \bigcap_{q \in \mathbb{R}^d} B(q, aR_S(q)).$$

由于 $S \subseteq \mathcal{H}_a(S)$，问题在于这个凸包能延伸到样本之外多远。记 $L_S = \mathrm{diam}(S)$。

对于固定的中心，凸包由距离最远的单个样本点决定。转而考虑分布后，我们可以为每个凸包点选取一个均值和一个方差，然后通过分别从这两个分布中独立抽样来比较两个凸包点。

引理 5.1 (凸包的方差表示)。 对于非空有限集 S 上的分布 λ，设其均值为 μ_λ，方差为 $V_\lambda = \mathbb{E}_\lambda \|s - \mu_\lambda\|^2$。对于每个 $a > 1$,

$$\mathcal{H}_a(S) = \bigcup_{\lambda} B\left(\mu_\lambda, \sqrt{(a^2-1)V_\lambda}\right). \tag{11}$$

特别地,

$$\mathrm{diam}(\mathcal{H}_a(S)) \leq \sqrt{2a^2 - 1}\, L_S. \tag{12}$$

证明。 对于固定的点 y，属于凸包等价于

$$\sup_q \min_\lambda \left(\|y - q\|^2 - a^2 \sum_s \lambda_s \|s - q\|^2 \right) \leq 0.$$

该表达式关于 λ 是仿射的，关于 q是凹的，其二次部分为 $-(a^2-1)\|q\|^2$。由于 λ 属于一个紧单纯形，且 q 中的线性系数均有界，因此所有相关的最大化元都位于

同一个紧球内。引理 3.3 因此交换上确界与最小值。随后通过配方得到该值

$$\min_{\lambda} a^2 \left(\frac{\|y-\mu_\lambda\|^2}{a^2-1} - V_\lambda \right).$$

该最小值可以取到，从而证明了（11）。

考虑由分布 λ，γ表示的两个凸包点，其均值分别为 μ，ξ，方差分别为 V，W。设 $h = \|\mu - \xi\|$。从这两个分布中独立抽样可得

$$h^2 + V + W = \mathbb{E}\|s-t\|^2 \le L_S^2.$$

因此，这两个凸包点之间的距离至多为

$$h + \sqrt{a^2-1}(\sqrt{V}+\sqrt{W}) \le h + \sqrt{2(a^2-1)}\sqrt{L_S^2 - h^2} \le \sqrt{2a^2-1}\, L_S.$$

This 这就证明了该引理，其中包括 S 中任意仿射维数或点重合的情形 . □

凸包因子可以取到。设 $h = (2a^2-1)^{-1/2}$ 和 $V = (1-h^2)/2 = (a^2-1)h^2$。取两个正则单纯形，每个单纯形有 $m \ge \max\{2, h^{-2}\}$ 个顶点，外接半径为 $\sqrt{V}$。将它们的中心放置在一条直线上，两中心距离为 h；该直线垂直于两个单纯形，且两个单纯形的方向空间彼此正交。令 S 由它们的所有顶点组成。任意两个来自不同单纯形的顶点之间的距离平方均为 $h^2 + 2V = 1$。每个单纯形内的边长平方为 $2Vm/(m-1) \le 1$，因此 $L_S = 1$。两个单纯形上的均匀分布的方差为 V。根据（11），凸包在每个中心周围都包含一个半径为 $\sqrt{(a^2-1)V}$ 的球。位于中心连线上的两个最外侧点之间的距离为 $h + 2\sqrt{(a^2-1)V} = 1/h = \sqrt{2a^2-1}$。

有界见证半径。对于输入点 $y \notin \mathcal{H}_a(S)$，存在某个 q 满足

$$\|y-q\| > a R_S(q).$$

设 $D_{\text{univ}} = 2\sqrt{d\bar{\Delta}}$。对于任意 $s \in S$，

$$\|y-q\| \le \|y-s\| + \|s-q\| \le D_{\text{univ}} + R_S(q).$$

因此，每个这样的见证球都具有

$$R_S(q) < D_{\text{univ}}/(a-1). \tag{13}$$

如果 S 包含两个不同的整数点，则还有 $R_S(q) \ge L_S/2 \ge 1/2$。因此，只需使用对数数量的几何半径尺度，尽管凸包定义是对 $\mathbb{R}^d$中的所有中心进行量化。算法无需计算见证中心。对于概率论证，可以将其选为 S 和 y的确定性函数；严格不等式也允许从有理中心的固定枚举中选取一个。

5.2 为什么恢复凸包外部就足够

设 $S \subseteq P$，且集合 $C \subseteq P$ 包含 $P \setminus \mathcal{H}_a(S)$ 中的每个点。定义交叉评分

$$L_{\text{cross}} = \max_{p \in C} R_S(p),$$

如果 C 为空，则其值为零。如果见证下界估计 L_C 满足 $\operatorname{diam}(C) \le f L_C$，则

$$\operatorname{diam}(P) \le \max\{\sqrt{2a^2-1}, f\} \max\{L_S, L_{\mathrm{cross}}, L_C\}. \tag{14}$$

事实上，任取 P中的两个点12。如果两者都属于凸包，则使用（ C；如果两者都在凸包外，则两者都属于 C，并使用关于 $x \in C$ 的估计。如果 y 在凸包外，而 $q = x$ 在凸包内，则凸包定义中的特定中心给出

$$\|x - y\| \le a R_S(x) \le a L_{\mathrm{cross}}.$$

由于 $a > 1$，混合对因子 a 小于 $\sqrt{2a^2-1}$。对于非空样本，显式计算 $\operatorname{diam}(S \cup C)$ 因此得到因子 $\sqrt{2a^2-1}$。算法 1 通过适当的证书和预言机精度达到相同的因子，同时每次更新只检查多项式数量的证书中心。

5.3 抽样所有包含球

算法。 我们设定 $m = N^s$ 并以公共概率 $p \in [m/(4N),\ m/N]$ 使用固定哈希选择每个点标识。该区间内的二进阈值即可满足要求。我们存储所选位置及其精确频数，并使用确定性大小上限。令

$$Q = (d+2)\log(N+2) + \log((T+2)/\delta) + 1, \qquad b = C N^{1-s} Q.$$

我们使用带有 $k = \lceil C_1 Q \rceil$ 的 $2k$-重独立性。为计算哈希，我们将键嵌入二进制域中，使其具有足够的比特来容纳一个键和二进阈值。对一个次数至多为$2k-1$ 的随机多项式在该键处求值，并将其输出与阈值进行比较。种子由 $2k$ 个域元素组成。通过连续进行二次扩域，可以获得具有足够比特长度的域。[1]

分析。

引理 5.2（同时球网）。 对于任意固定的 $0 < s < 1$，上述维护的样本占用$N^s \operatorname{poly}(L)$ 空间，并具有多项式更新时间。以概率 $1-\delta$，在每个指定的快照时刻，包含该样本的每个球至多排除 $C N^{1-s} Q$ 个当前点。对于任意时期长度$H_{\mathrm{epoch}} \ge 1$，如果冻结快照至多用于后续 $2H_{\mathrm{epoch}}$ 次更新，则相应的异常界为 $C N^{1-s} Q + 2H_{\mathrm{epoch}}$。

证明。 这些提升 $(x,\ \|x\|^2)$ 将球的成员关系表示为 $\mathbb{R}^{d+1}$中的半空间测试。在大小至多为 N 的支撑集上，至多存在 $(N+1)^{d+2}$ 个球的迹。对于任一大小大于 b的迹的补集，其采样计数的均值为 $\mu \ge CQ/4$。引理 3.2 和马尔可夫不等式给出概率$[Y = 0] \le 2^{-k}$ 当 C 相对于 C_1足够大时成立。先取 C_1 足够大，对这些迹以及 $T+1$ 个叶节点应用并集界，失败概率至多为 $\delta/2$。因此，即使是在看到样本后选取的球也满足

$$S \subseteq B(q,r) \quad \Longrightarrow \quad |P \setminus B(q,r)| \le b \qquad \text{for every ball } B(q,r). \tag{15}$$

总样本量的均值至多为 m。在阶数为$2k'$时应用相同的矩界，对于 $k' = \Theta(\log((T+2)/\delta)) \le k$，超出 $m + C_2(m+k')$ 的概率至多为 $2^{-k'}$

[1]从 $\mathbb{F}_2$中的 $a = 1$ 开始，维护一个绝对迹为一的元素。于是 $X^2 + X + a$ 是不可约的，而根 z 具有相对迹一。元素 az 在该扩域中的绝对迹为一。比特次数翻倍，而该塔中的算术运算只需多项式数量级的比特运算。

当 C_2 足够大时。我们禁用一个溢出的副本。这在所有更新过程中至多产生 $\delta/2$ 的开销，同时确保每个结果都满足空间界。对键进行抽样并更新其字典均需多项式时间。

对于冻结样本，任意包含球中的异常点数量，在每次中间插入后最多增加一个。如果样本为空，我们可以使用一个与边界支撑集不相交的球。对于单元素集合，我们可以使用其唯一点处的零半径球。这些情况下，边界支撑集分别至多为 b 和 $b+1$。

□

6 外部点的固定过滤器

更新例程需要过滤器族具备两个性质。它必须能够快速找到包含更新点的过滤器，并且远处的目标必须属于一个足够稀疏、可以进行解码的过滤器。只有第二个性质涉及见证球。对于随机映射的任意选择，第一个性质都成立。

6.1 算法

在一个基准半径 w 处，一个层次结构由 h 个阶段组成。在第 i 个阶段，我们独立选择一个 $k_i \times d$高斯矩阵 A_i、其像空间中边长为 $\delta_g \widetilde{r}_i w$的均匀平移网格，以及一个将单元标识符映射到 B_i 个桶的两两独立哈希。对于点 x，其叶节点是包含 $A_i x$ 的各阶段单元所对应桶标识符组成的元组。一次更新会改变该叶节点处的一个经认证的矩量草图（引理7.1）。这些阶段在该元组的所有前缀之间共享其映射。

以下是一般构造所用的参数。我们假设 $\sqrt{2} < \lambda < \sqrt{3}$，将掩码裕量设为 $0 < \eta_{\mathrm{mask}} \le \min\{m_\lambda/4,\ 1/4\}$，并取

$$m_\lambda = \lambda^2/2 - 1, \qquad r_i^2 = 1 - (i-1)m_\lambda/2.$$

最后阶段 h 是第一个满足 $(1+\eta_{\mathrm{mask}})r_h^2 < m_\lambda$ 的阶段。由此得到 $h = O_\lambda(1)$ 和 $r_h > 0$。每个$\widetilde{r}_i$ 都是 $[r_i, (1+\zeta)r_i]$ 中的一个固定有理数。我们取 ξ，δ_g，$\zeta > 0$ 较小、$\beta > \sqrt{1+\xi}$ 接近于一，并设

$$C_g = \left\lceil \sqrt{2\pi e}\,(\beta + \delta_g)/\delta_g \right\rceil, \qquad A_0 = 100h.$$

我们选择 B_i，使其为不小于 $A_0 C_g^{k_i}$ 的最小二次幂。证明给出了一个常数 $\tau > 0$，它仅取决于这些选择。对于预算指数 $0 < \rho < s$，我们设

$$G > 2 + 4\ln(C_g)/\tau, \qquad W = \sum_{i=1}^{h} G^{i-1}, \qquad 0 < \theta < \frac{\rho}{W \ln C_g}, \qquad k_i = \left\lceil \theta G^{i-1} \ln(N+2) \right\rceil.$$

叶节点容量如下文所示（18）。等价地，实现可以使用其中给出的显式系数。附录 A.1 采用了更精确的两阶段计算，以及显式因子二指数对应的容量（33）。所有选择均在数据流开始前确定。

我们在每个指定基准尺度上维护独立重复。固定过滤器接口会返回被触及的叶节点及其容量。对于独立恢复，解码所有叶节点即可返回其经认证字典的并集。算法 1则将相同的更新输入引理7.2中的本地时钟。

6.2 分析

定理 6.1（球外稀疏恢复）。 固定 $\sqrt{2} < \lambda < \sqrt{3}$、 $0 < \bar{} s < 1$和 $\bar{0} < \rho < s$。存在 $\epsilon_0 > 0$，使得在某个指定的正有理数半径 w 下，固定过滤器的叶节点总容量至多为

$$(N^{1-s+\rho} + N^{1-\epsilon_0})\,\mathrm{poly}_{\lambda,s,\rho}(\mathcal{L}).$$

在尺度编码和 $|$对数 $w|$关于 L 为多项式有界的前提下，它们使用相同形式的空间，并且每次更新需要多项式$\lambda,\ s,\ \rho(L)$ 时间。假设除 $b = O(N^{1-s}$多项式$(L))$ 个当前点外，其余当前点都位于 $B(c, w)$ 中，且当前点 y 满足 $\|y - c\| \geq \lambda w$。对于这个固定的输入、目标点和见证，以常数概率， y 的叶节点支撑集大小不超过其指定容量。随后，其经认证的快照可以以常数成功概率恢复 y 。更新只使用点标识和预先选定的过滤器。对于预先指定的一系列半径，各项开销相加。通过独立重复，可以同时覆盖任意预先指定的有限目标点和时间集合，其开销仅随集合大小及失败概率倒数的对数增长。

证明。 我们固定见证球和目标点，并使其独立于网格。论证分为三个步骤。我们找出球中可能发生过多冲突的部分，界定某个点首次离开该部分时的冲突次数，并对其在各叶节点上的贡献求和。

每个尺度上的球冠

固定一个见证球 $B(c,\ w)$。在几何分析中，对输入点 p使用归一化坐标 $z = (p - c)/w$，并将归一化输入集记为 P° 。目标 y 的归一化坐标为 Ru，其中 u 是一个单位向量。将主体点写作 $z = tu + v$，其中 $v \perp u$、 $q = \|v\|^2$和 $t^2 + q \leq 1$。分离假设给出 $R \geq \lambda$。定义

$$\phi(t, q) = q \exp\left(\frac{(\lambda - t)^2}{q} - 1\right),$$

其在 $q = 0$处的值为 $+\infty$。采用这一约定是因为 $\lambda - t \geq \lambda - 1 > 0$。共线点与目标之间的归一化投影间距为 $(R - t)\|A_i u\|$ ，在下述轴向范数事件上，一旦常数 δ_g 足够小，该间距就超过单元直径 $\delta_g \widetilde{r}_i \sqrt{k_i}$。对于下面的计算，取 $m = m_\lambda$ 和$\eta = \eta_{\mathrm{mask}}$ ，将 R_0 定义为原单位球，并令 $1 \leq i \leq h$,

$$R_i = \{tu + v :\ t^2 + \|v\|^2 \leq 1, \phi(t, \|v\|^2) \leq (1 + \eta) r_i^2\}.$$

这些集合是嵌套的。最后一个集合 R_h 为空。

对于每个 $i < h$，集合 R_i 都包含在一个半径为 r_{i+1} 的球中，该球的中心位于原目标轴上。为证明这一点，首先注意到

$$(\lambda - t)^2 - q \geq \lambda^2 - 2\lambda t + 2t^2 - 1 \geq m.$$

因此 $\phi(t,\ q) \geq q + m$并且，对于固定的 t，函数 $q \mapsto \phi(t,\ q)$ 在 $0 < q \leq 1 - t^2$ 上递减。因此， t在 R_i 中的可能取值满足

$$F(t) := (1 - t^2) \exp\left(\frac{(\lambda - t)^2}{1 - t^2} - 1\right) \leq (1 + \eta) r_i^2.$$

如果该集合非空，令 t_i 为其最小元素。它是正的：对于 $t \le 0$， $F(t) \ge (\lambda - t)^2 > 2$，而右侧至多为 $1+\eta < 2$。连续性，加上端点处的 $F(t) \to \infty$，给出 $F(t_i) = (1+\eta) r_i^2$。 R_i 中的每个点都满足 $t \ge t_i$，因此

$$\|tu + v - t_i u\|^2 \le 1 - 2t_i t + t_i^2 \le 1 - t_i^2.$$

在下边界处，应用 $\phi(t, q) \ge q + m$ 于 $q = 1 - t_i^2$可得 $F(t_i) = (1+\eta) r_i^2 \ge 1 - t_i^2 + m$。因此

$$1 - t_i^2 \le (1+\eta) r_i^2 - m \le r_i^2 - m/2 = r_{i+1}^2,$$

其中 $r_i \le 1$ 和 $\eta \le m/4$。这证明了所声称的覆盖球。若某个掩码已经为空，则其后续所有掩码也都为空，并且它们的覆盖球可以任意选择。最后， $\phi \ge q + m > m$证明了 $R_h = \varnothing$。

对于阶段 i，令 c_i 表示覆盖 R_{i-1}的半径为 r_i 的球的一个仅供分析的中心。因此 $c_1 = 0$，后续的中心可以取相应的 $t_{i-1} u$。所有这些中心都位于同一条轴上；算法既不会计算这些中心，也不会计算目标方向。

沿目标方向的投影

固定一个单位向量 u 和一个高斯矩阵 $A \in \mathbb{R}^{k \times d}$。条件于 Au，向量 $v \perp u$ 具有$Av \sim N(0, \|v\|^2 I_k)$，且独立于 Au。通常的卡方矩生成函数给出了 $\|Au\|^2/k$ 偏离 1 的指数尾部。

共享一个平移网格单元的概率。 在边长为 ℓ 且各坐标平移量相互独立并服从均匀分布的网格中，两个差向量为 $\omega \in \mathbb{R}^k$ 的投影点共享一个单元的概率为

$$\prod_{j=1}^{k} (1 - |\omega_j|/\ell)_+.$$

这个三角核的总积分为 ℓ^k ，支撑集为 $[-\ell, \ell]^k$。若随机差服从均值为 μ 、协方差为 $\sigma^2 I_k$ 的高斯分布，其中 $\sigma > 0$，则对其密度积分可得

$$\Pr(\text{same cell}) \le \left(\frac{\ell}{\sqrt{2\pi}\sigma} \right)^k \exp\left(-\frac{(\|\mu\| - \ell\sqrt{k})_+^2}{2\sigma^2} \right).$$

在层次结构中， $z = tu + v$ ，且目标为 Ru，因此在给定 Au 的条件下，该差的均值为$(t - R)Au$ ，协方差为 qI_k，其中 $q = \|v\|^2$。在 $\|Au\|^2 \ge (1-\xi)k$上，较小的均值间隔保留了<sty le id='11'>带符号的轴向坐标 t。若将其替换为绝对值，就会丢失下文所需的球冠结构。

投影一个包含球。 将包含球平移至原点，并将其半径缩放为1。将其中的点写作 $tu + v$，其中 $v \perp u$ 且 $t^2 + \|v\|^2 \le 1$。固定 $Au = g$ ，使得 $\|g\|^2 \le (1+\xi)k$。给定此向量， $A(tu+v)$服从均值为 tg 、协方差为 $\|v\|^2 I_k$的高斯分布。

我们将每个投影坐标除以 $\sqrt{1+\xi}$。现在，这些坐标具有相同的方差$v_0 = \|v\|^2/(1+\xi)$和均值 $m_j = tg_j/\sqrt{1+\xi}$。它们的平均平方均值满足 $\overline{m^2} = k^{-1}\sum_j m_j^2 \le t^2 \le 1 - v_0$。对于 $0 < \theta < 1/2$，平方范数除以 k 后的条件对数矩生成函数为

$$-\tfrac{1}{2}\ln(1 - 2\theta v_0) + \frac{\theta \overline{m}^2}{1 - 2\theta v_0} \le -\tfrac{1}{2}\ln(1 - 2\theta v_0) + \frac{\theta(1 - v_0)}{1 - 2\theta v_0}.$$

这里，独立性使我们可以将各个坐标对数矩母函数相加，其中每个函数关于 m_j^2 都是仿射的。右侧关于 v_0 的导数为 $2\theta^2(1-v_0)/(1-2\theta v_0)^2 \geq 0$，因此其最大值为 $-\frac{1}{2}\ln(1-2\theta)$，即每个坐标的中心 χ_k^2 界。因此，对于 $\beta > \sqrt{1+\xi}$，

$$\Pr(\|A(tu+v)\| > \beta\sqrt{k} \mid Au) \leq e^{-Jk}, \qquad J = \frac{1}{2}\left(\frac{\beta^2}{1+\xi} - 1 - \ln\frac{\beta^2}{1+\xi}\right) > 0.$$

在原始目标轴上以任意中心进行平移，并按其包围半径重新缩放后，同样的计算仍然适用。

半径为 $\beta r\sqrt{k}$ 的欧几里得球至多与

$$\left(\sqrt{2\pi e}\frac{\beta+\delta_g}{\delta_g}\right)^k$$

边长为 $\delta_g r$ 的单元相交。事实上，这些单元的不相交内部都位于扩大了一个单元直径的球内，而 k 维欧几里得单位球的体积 v_k 满足 $v_k \leq (2\pi e/k)^{k/2}$。在接下来的证明中，这两个界将同单元碰撞与其他投影尾部点区分开来。

隔离投影核心。我们以.....为条件

$$(1-\xi)k_i \leq \|A_i u\|^2 \leq (1+\xi)k_i \quad \text{for every } i.$$

对于所有充分大的 N，该事件的概率至少为 0.99；所有 k_i 都将是 $\log N$ 的正固定倍数。有界的 N 可以通过完整存储来处理，其常数取决于所选的近似。在原始网格中，投影核心是以 $A_i(c+wc_i)$ 为中心、半径为 $\beta\widetilde{r}_i w\sqrt{k_i}$ 的球。根据上述装填界，它至多与 $C_g^{k_i}$ 个单元相交，这些单元称为核心单元。当 z 处于归一化坐标时，用 $\mathrm{cell}_i(z)$ 表示 $A_i(c+wz)$ 所在的单元，并用 $\mathrm{cell}_i(Ru)$ 表示目标单元。在给定投影和平移量的条件下，独立的外层哈希以至少 $1-1/A_0$ 的概率避免将任何其他核心单元发送到目标桶。令 I_i 表示该隔离事件，并令 $I = \bigcap_i I_i$。在给定高斯范数事件的条件下，根据并集界，$\Pr[I] \geq 0.99$。

掩码外的碰撞概率

存在固定的 $\xi, \beta - 1, \delta_g, \zeta > 0$ 以及一个固定的 $\gamma_{\mathrm{col}} > 0$，使得在每个阶段 i，原单位球中满足

$$\phi(t,q) > (1+\eta)r_i^2$$

的点的条件同胞概率至多为

$$\Pr[\mathrm{cell}_i(z) = \mathrm{cell}_i(Ru) \mid A_i u] \leq C_g^{-k_i} e^{-\gamma_{\mathrm{col}} k_i}. \tag{16}$$

该结论并不要求 z 位于阶段-i 覆盖球内；只要属于原单位球且位于掩码外即可。

平移网格估计将碰撞概率界定为.....的 k_i 次方

$$\frac{\delta_g \widetilde{r}_i}{\sqrt{2\pi q}} \exp\left[-\frac{(\sqrt{1-\xi}(\lambda - t) - \delta_g \widetilde{r}_i)^2}{2q}\right].$$

我们选择这些常数，使指数中括号内的表达式保持为正。实际的目标比率 $R \geq \lambda$ 只会使这个上界进一步减小。将底数乘以 C_g，并取极限 $\xi,\ \beta - 1,\ \delta_g,\ \zeta \to 0$，便恰好得到

$$\frac{r_i}{\sqrt{\phi(t,q)}} \leq (1+\eta)^{-1/2} < 1.$$

该界可在 $q = 0$ 处连续延拓为零，因为 $\lambda - t \geq \lambda - 1 > 0$，且这一点均匀成立。原单位球的紧致性以及阶段集合的有限性，因此保证了共同的固定小参数和一个正的 γ_{col}，用于（16）。等价地，可以单独考虑一个较小的 q 区间，在此区间上指数衰减是一致的；再在其紧致补集上使用通常的一致连续性。

对于阶段-i 覆盖球中的归一化点 z，将条件投影界应用于 $(z - c_i)/\widetilde{r}_i$。在归一化坐标下，投影核心尾部为

$$\Pr[\|A_i(z - c_i)\| > \beta \widetilde{r}_i \sqrt{k_i} \mid A_i u] \leq e^{-Jk_i}, \tag{17}$$

其中

$$J = \frac{1}{2}\left[\frac{\beta^2}{1+\xi} - 1 - \ln\left(\frac{\beta^2}{1+\xi}\right)\right] > 0.$$

目标叶节点上的负载

根据大量点首次离开嵌套掩码的位置对其进行划分：

$$G_i = (P^\circ \cap R_{i-1}) \setminus R_i, \qquad 1 \leq i \leq h.$$

这些类别覆盖原见证球中的所有点，因为 $R_h = \varnothing$。它们由固定输入、见证和目标决定，与所有阶段随机性无关。

固定 $z \in G_i$，并以 $A_i u$ 处于其指定范数范围内为条件。令 E_i 表示 z与目标位于同一外部桶的事件，S_i 表示它们位于同一单元的事件，而 T_i 表示式（17）中的投影核心尾部。核心隔离给出如下包含关系

$$E_i \cap I_i \ \subseteq\ S_i \ \cup\ (T_i \cap S_i^c \cap E_i).$$

事实上，由 I_i可知，与投影核心相交的其他单元均被排除。给定投影和平移量后，两个不同的单元键发生外部哈希碰撞的概率为 $1/B_i$。对 $T_i \cap S_i^c$积分，并应用式（16）和式（17），可得

$$\begin{aligned}\Pr(E_i \cap I_i \mid A_i u) &\leq \Pr(S_i \mid A_i u) + B_i^{-1} \Pr(T_i \mid A_i u) \\ &\leq \frac{2A_0 e^{-\gamma_{\mathrm{col}} k_i} + e^{-Jk_i}}{B_i} \leq \frac{(2A_0 + 1)e^{-\tau k_i}}{B_i},\end{aligned}$$

对于 $0 < \tau \leq \min\{\gamma_{\mathrm{col}},\ J\}$，使用 $B_i < 2A_0 C_g^{k_i}$。这给出了联合事件 $E_i \cap I_i$ 的上界。

在每个后续阶段 $j > i$，同一点仍在掩码之外：$\phi(t,\ q) > (1+\eta) r_i^2 \geq (1+\eta) r_j^2$。因此，它与目标位于同一几何单元的概率至多为 $C_g^{-k_j} e^{-\gamma_{\mathrm{col}} k_j}$。如果单元不同，成对哈希以 $1/B_j$ 的概率将其通过。因此，无需在阶段 j 要求发生任何隔离事件，

$$\Pr[z \text{ passes stage } j \mid A_j u] \leq \frac{1 + 2A_0 e^{-\gamma_{\mathrm{col}} k_j}}{B_j} \leq \frac{2}{B_j}$$

对于足够大的 N。

即使在分别对各自的范数事件进行条件化之后，各阶段仍然相互独立。舍弃 i 之前的测试只会增大该界。 z 到达目标叶节点且所有隔离事件均成立的概率至多为

$$C_h \frac{e^{-\tau k_i}}{B_i B_{i+1} \cdots B_h},$$

其中 C_h 仅取决于固定深度。这避免了错误的分析方式：将早期层级的异常负载原封不动地传递到所有后续层级。点一旦落在掩码之外，后续的独立哈希就会以 $O(1/B_j)$的比例将其稀释。

根据期望的线性性，在给定所有范数事件的条件下，

$$\mathbb{E}\left[\mathbf{1}_I \cdot \#\{\text{bulk points in the target leaf}\}\right] \le C_h N \sum_{i=1}^{h} \frac{e^{-\tau k_i}}{B_i \cdots B_h}.$$

马尔可夫不等式表明， I 成立但该负载超过所示期望的一百倍的概率至多为 0.01。我们分别计入范数失效和隔离失效。原有的 b 异常点可能全部落在目标叶节点中。因此，叶节点容量

$$K = \left\lceil C'_h \left[b + 1 + N \sum_{i=1}^{h} \frac{e^{-\tau k_i}}{B_i \cdots B_h} \right] \right\rceil \tag{18}$$

以常数概率恢复实际目标。可以取 $C'_h = 100(2A_0+1)2^{h-1}$：第一阶段的系数至多为 $2A_0+1$，后续阶段的系数总和至多为 2^{h-1}，而因子 100 是马尔可夫不等式的容许倍数。整个过程中，我们都不假设点之间相互独立。

空间、更新时间与重复

算法中维数的增长意味着

$$B_{\text{tot}} := \prod_{i=1}^{h} B_i = O_{\lambda,\rho}(N^{\rho}),$$

并且，对于每个阶段 i,

$$e^{-\tau k_i} \prod_{j<i} B_j \le O_{\lambda,\rho}(N^{-\epsilon_0}), \qquad \epsilon_0 = \tau\theta/2 > 0.$$

后一个不等式由 自然对数 $C_g \sum_{j<i} G^{j-1} \le (\tau/4)\, G^{i-1}$得出；舍入和 A_0 因子只贡献固定常数。将（18）乘以叶节点数可得

$$B_{\text{tot}} K \le O_{\lambda,s,\rho}\left(N^{\rho}(b+1) + N^{1-\epsilon_0}\right).$$

对于 $b = N^{1-s}$ 多项式和 $\rho < s$， N 的两个幂都严格小于一 .

每次原始更新都会计算 h 高斯投影和网格/哈希标识符，并在相应的叶节点处更改一个稀疏恢复草图。由于 h 是固定的，且每个 $k_i = O_\lambda(\log N)$，因此该过程的时间复杂度关于 d 和对数参数为多项式，且与叶节点数量和叶节点恢复容量无关。只需猜测原始见证半径。在每个$O(\log(d\Delta+2))$个原始几何尺度 w上，所有阶段的网格宽度都是固定常数 $\delta_g \tilde{r}_i w$。不存在依赖数据的重新分类，也不存在尺度选择的乘积。在每个基准尺度上，我们使用 $O(\log(NT/\delta))$个独立层次结构。对所有叶节点进行解码的查询会取其经认证字典的并集

。对目标点和时间取并集界，即可得到同时恢复保证。

为获得有限描述，我们选择满足所显示严格不等式的有理数 λ、η、C_g、G、θ。如果指定的分离度 λ 是无理数，则使用介于 $\sqrt{2}$ 和λ之间的任意有理数分离度。我们选择一个严格大于 $\bar{e}^{-\gamma_{\mathrm{col}}}$ 和 e^{-J}两者的有理数 $q_0 \in (0、1)$，并设 $\tau = -\ln q_0$；这样容量项 $e^{-\tau k_i} = q_0^{k_i}$均为有理数，可以精确地向上舍入。我们取下界之上的有理数 G，并取具有严格总桶指数 $\theta W \ln C_g < \rho$的有理数 θ。从上方以固定加性常数近似 $\ln(N+2)$，只会使每个 k_i 改变一个固定常数。桶范围是 2 的幂，网格半径乘数 $\tilde{r}_i$ 是正 r_i的有理上近似。这些选择保留了正的指数节省。

附录 A.2 证明了成员关系耦合。使用有理尺度乘数以及足够精细的高斯舍入和平移舍入，每个数据流中的点除分配的耦合失败概率之外，都会保持在其理想网格单元中。所需的参数仅需使用多项式数量级的比特。

□

6.3 凸包的应用

我们维护一个活动伯努利样本，其期望大小至多为 N^s，且独立于层次结构。它同时满足球网性质，给出 $b = N^{1-s}\mathrm{poly}(d, \log N, \log(T/\delta))$ 个异常点，对每个包含该样本的球均成立。对于空样本和单元素样本，使用算法 1 中的全局回退桶。

固定 $c > \bar{\sqrt{3}}$ 并选择

$$\sqrt{2} < \lambda < \min\{a, \sqrt{3}\}, \qquad \sqrt{2} < a < \sqrt{(c^2+1)/2}, \qquad 1 < q_{\mathrm{scale}} \le a/\lambda.$$

对于见证 $B(c_0, r)$，其中 $\|y - c_0\| > ar$，将 r 向上取整到第一个受维护的尺度 $w \ge r$。由于$w \le q_{\mathrm{scale}} r$，较大的球仍包含该样本以及 $\|y - c_0\|/w > a/q_{\mathrm{scale}} \ge \lambda$。于是，每个位于扩张后的最远距离凸包 $\mathcal{H}_a(S)$ 之外的点，都可由某个层次结构恢复。若 Z 是恢复点的并集，则精确查询 $L = \mathrm{diam}(S \cup Z)$ 满足

$$L \le \mathrm{diam}(P) \le \sqrt{2a^2 - 1}\, L \le cL.$$

因此，对任何严格大于 $\sqrt{3}$ 的固定因子，尤其是因子二，都存在一种固定幂次、次线性空间的构造，其更新时间为多项式级，但查询代价较高。算法 1 仅查询已暴露中心来维护该近似。计入活动样本后，可以取

$$\alpha = \max\{s, 1 - s + \rho, 1 - \epsilon_0\} < 1$$

对于这一慢查询后果以及所显示的多项式因子。对于整数网格，(13) 为算法中的尺度端点提供了依据1。

7 从稀疏叶节点到当前候选项集

现在可能较稀疏的筛选器，距离上次收到更新或许早在当前全局时期之前。为了及时恢复其中的所有点并为其评分，我们用该筛选器收到的更新次数来衡量其快照的存续时间。由此得到的过程会将原始更新转换为显式候选点的插入和删除有序序列。另一个关键环节是在使用任何存储条目作为见证之前，对每个快照进行验证。

7.1 算法

每个桶 i 都是点标识的固定筛选器，容量为 $K_i \geq 1$。其中每个选中的点都具有其原始的非负频率。每次原始更新最多会触及 $r = \text{poly}(\mathcal{L})$ 个桶。在给定固定数据流和这些筛选器的条件下，每个桶中的序列都独立于其私有恢复哈希。

7.1.1 调度并维护引用计数

图 2 展示了连续两代的构建和服务区间。桶 i 仅对自身的更新进行计数。其块长度为 K_i，以本地时钟计。在每个本地块边界，桶都会捕获其持久化矩草图的根及其总质量。它在下一个块期间解码该快照，在随后一个边界安装所得字典，并在再下一个块后将其退役。如果解码失败，或经认证的支撑集超过 $3K_i$，则该代不可用。时序不依赖于恢复成功与否，也不依赖于恢复点的标识。

最初，空字典可以精确地为该桶提供服务。它可以在前两个本地块内继续提供服务，最多记录 $2K_i$ 个不同键。在时间 K_i 开始执行针对该快照的任务，并在 $2K_i$安装其输出。后续任务遵循相同的调度安排。较小的常数 K_i 可以立即完成所有有界工作，而不会改变渐近界。失败的或

算法2 容量为 K的叶节点的一次更新

输入 单点更新（p，σ）、本地时间 u以及当前代次
输出 已更新的草图和代次，并将候选转换添加到队列中

更新持久化矩草图及其精确总数
递增 u，如果有正在构建的任务，则为其记录（p，σ）
如果服务代次可用，则维护其完整字典；如果有新插入的键，需要
使用 p 的共享全局当前计数刷新物化引用
如果存在正在构建的任务，则按其固定工作预算推进
    解码并验证快照，然后扫描快照和变更映射键
    对于每个扫描到的键 x，使用 $F_\tau(x) + \Delta_\tau(x)$ 物化其引用
如果 u 是 K 的倍数**则**
    如果某个任务到期，则安装其结果，并将前一个根节点加入退役队列
    捕获当前草图的快照，并启动下一个任务
结束 if
按顺序将产生的成员变更追加到副本的候选队列

超大恢复会产生一个不可用代次，调度将继续保持不变。最多只有一个服务代次和一个构建代次需要更新。退役代次不进行本地日志维护：其引用读取全局重数，并由退休队列移除。在插入之前不存在于完整服务字典中的条目，其先前重数为零。因此，可以准确地初始化其新引用。现有条目共享全局当前计数；它们不会再次增加该计数。

我们为每个构建任务存储一份带符号变化映射。它的用途不同于点的共享重数：当任意字典首次发现某个点时，该映射可帮助我们正确初始化该点。具体而言，它可以处理发现过程中被删除后又重新插入的点。对于一个任务

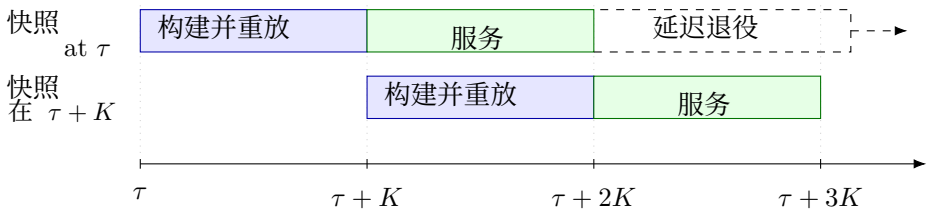


图 2：重构依照叶节点的本地更新计数进行。在一代的服务区间内，其快照至多落后 $2K$ 次本地更新。因此，当前稀疏的叶节点必定有可恢复的服务快照。退役由单独的全局工作队列处理；虚线箭头不设本地截止期限。因此，空闲叶节点不会累积无界的未回收记录。

其快照位于本地时间 τ，保留一份带符号变更字典

$$\Delta_\tau(p) = \sum_{\tau < v \le u} \text{update to point } p \text{ at local time } v.$$

在任务的构建区间内，最多有 K_i 次更新，因此此变更映射及其时间顺序日志占用 $O(K_i\text{poly})$ 空间。一旦经认证解码提供了完整快照字典F_τ，任何相关键的当前重数恰好为

$$f_u(p) = F_\tau(p) + \Delta_\tau(p),$$

其中，完整快照字典中缺失的键的值为零。每次本地更新都分配足够大的固定多项式工作预算。块的前一部分用于完成解码。剩余工作会扫描 F_τ 和变更映射中的键，并使用所示恒等式计算它们当前的重数。扫描期间到达的新变更映射键会加入队列。一个块期间总共只有 $O(K_i)$ 个键需要处理；以快于新键到达的速度处理它们，就能在固定截止期限前完成扫描。扫描追上进度后，由于完整快照字典已可用，新引入的键可以在多项式时间内立即计算出来。每个已计算的键之后都会照常接收原始频率更新。

我们维护一个以原始点标识为键的全局字典。每条记录保存其当前原始重数的精确值、一个引用计数，以及显式点几何结构中的一个标识。引用来自活跃、待处理以及尚未清理的已退役桶代中的物化条目。只有当该点当前的重数和引用计数都为正时，它才会包含在几何候选集中。因此，即使在许多过滤器中反复发现同一个位置，也不会为其生成多个几何标识。首次物化一个点时，我们根据经认证的快照加变更公式初始化其全局重数。如果全局记录已存在，其当前重数已知，并且与该公式一致。此后，每次针对该标识的原始更新，都会在对此次更新进行任何后台物化之前，恰好更新一次全局重数。从正值到零的转换会将该点从几何结构中移除一次，不受字典引用数量影响。从零到正值的转换则会在仍有引用时将其添加一次。原始严格性保证这些计数始终非负。

只有当添加的引用是指向当前计数为正的点的第一个引用时，才会创建几何候选项。删除最后一个引用会从候选结构中移除一个正计数点，并允许释放其全局记录。如果之后再次发现该点，新的经认证的快照和变更映射会根据其当时的重数重新初始化。对于没有已存储引用的原始点，无需维护其计数。新的活跃样本标识可以从其精确的样本频率初始化

同一条记录。每次成员变更都会向公共队列追加一个转换。外部引用回收队列拥有固定预算，该预算独立于候选层次结构执行的任何私有工作。

7.1.2 快照解码器

解码器接收一个只读快照。其成功条件需要检查键的身份，并单独检查是否找到了每个键。冲突可能导致无法发现键，因此两项检查都必须通过。我们将每个原始点编码为一个不同的非负整数键，长度为 u $b_{\mathrm{key}} = O(d \log(\Delta + 2))$ 比特。相同的构造适用于任何多项式位长的单射键编码。我们使用 h_{rec} 个两两独立哈希，将键映射到 B 个桶中，其中 B 是 $[24K, 48K)$ 中的 2 的幂。对定长键字符串使用随机二进制矩阵和独立偏移量，即可生成每个哈希。在每个桶中，我们存储其频率的三个精确整数矩，

$$C_0 = \sum_u f_u, \qquad C_1 = \sum_u u f_u, \qquad C_2 = \sum_u u^2 f_u.$$

符号为 σ 的更新会使 F 增加 σ 一次，并更改每个表中的一个矩三元组。快照会在同一时刻捕获 F 以及所有表。对于经过筛选的数据流， F 精确计数该筛选器中的总量。在单个表中，它也等于 C_0 在各个桶上的总和。我们为预先指定的一组$h_{\mathrm{rec}} = \lceil C(\log(K+2) + \log(M_{\mathrm{snap}}/\delta + 2)) \rceil$ 快照取 M_{snap} 个表。在同一个固定过滤器层次结构中，每个叶节点都可以使用相同的恢复哈希。快照由数据流和外层过滤器决定，论证对这些快照使用并集界。相互独立的完整副本使用独立种子。算法3 会增量扫描这些副本。它唯一可能的输出是完整字典或失败。

算法 3 非负桶快照的认证恢复**输入** 来自同一过滤器快照的矩表和精确总量 F，容量 K
输出 包含精确重数的完整快照字典，或失败1: 初始化一个空字典 $D$2: **对于**每个哈希表中的每个已存储计数器三元组**执行**3: **如果** $C_0 > 0$ 且 $C_0 C_2 = C_1^2$ **则**4: $u \leftarrow C_1 / C_0$5: 存储条目 $D[u] \leftarrow C_0$ 一次，并按键去重6: **结束如果**7: **结束循环**8: **如果** $|D| \le 3K$ 且 $\sum_{u \in D} D[u] = F$ **则**9: **返回** $D$10: **否则**11: **返回**不可用12: **结束如果**

7.2 分析

引理 7.1 （认证稀疏恢复）。 对于峰值支撑集大小至多为N、总质量至多为 T的非负频率向量，容量为 K 的草图占用 K 多项式(L) 空间，更新时间为 多项式(L)，并可在 K多项式(L) 时间内解码。每个被接受的输出都是重数精确的完整快照字典

。草图还会维护 $F=\sum_u f_u$，即其自身频率向量的总质量。对于一个与私有哈希无关的预设快照族，其中每个支撑集大小至多为 $3K$ 的快照都能同时以至少 $1-\delta$的概率被接受。上述资源界假设对数 $(M_{\mathrm{snap}}+2)$ + 对数 $(K+2)$ = 多项式 (L)，与叶节点桶和回退桶中的假设相同。

证明。 单元素集合的检验依据恒等式

$$C_0C_2-C_1^2=\sum_{u<v} f_u f_v (u-v)^2. \tag{19}$$

右侧每一项均非负。因此， $C_0>0$ 和 $C_0C_2=C_1^2$ 当且仅当某个桶中恰有一个占用键时成立。其键为 C_1/C_0 ，快照重数为 C_0。解码器会为纯桶中恢复出的每个键保留一条记录。当字典至多包含 $3K$条记录且这些记录的重数之和为 F时，解码器接受。每个恢复出的计数都是精确的，因此，任何被遗漏的正频率都会使总和严格小于 F。当$F=0$时，这也能确认空字典。对于支撑集大小至多为 $3K$的情况，固定键在一个哈希表中未能被隔离的概率小于 $1/8$。因此，完整恢复失败的概率至多为 $3K8^{-h_{\mathrm{rec}}}$。取 h_{rec} 为$\log(K+2)$的一个充分大的倍数 $+\log(M_{\mathrm{snap}}/\delta+2)$，即可使所有所需的稀疏恢复同时成功。对数据流和外层过滤器进行条件化后，快照族便可固定。并集界只依赖于独立的私有恢复哈希。

由于总质量至多为 T，即使 C_2 也至多包含 $2b_{\mathrm{key}}+\lceil \log_2(T+1)\rceil+1$ 比特。因此，每个计数器占用 $O(1+(b_{\mathrm{key}}+\log(T+2))/w_{\mathrm{word}})$ 个字。对这些多字键和计数器执行奇偶运算、乘法以及精确除法，需要 $\mathrm{poly}(\mathcal{L})$ 次字操作。惰性初始化避免在一次更新中清空大型表。固定的扫描预算，再加上确定性排序或平衡树去重，即可得到所述的解码界。该草图关于频率向量是线性的，并且在存在删除操作时仍然有效，因为每个桶始终构成一个非负频率数据流。 □

被拒绝的快照不会创建新的引用。为了说明即使后续仍有更新，接受一个快照也已足够，我们利用固定的本地截止期限，以及所有指向同一点的引用所共享的精确频率。

引理 7.2 (局部时钟编译器)。 存在一个持续维护的候选集 C ，具有以下性质：

1. 在任何随机结果下，每个候选项都是原始数据流中的一个当前点。*2.* 在恢复成功事件中， C 包含每个桶内当前支撑集大小不超过 K_i的所有点。*3.* 存储的记录至多有 $O(\sum_i K_i)$ 个多项式数量，包括待处理记录和待退役记录。*4.* 在最坏情况下，每次原始更新在 C中仅导致多项式数量的插入或删除，且只需进行多项式数量的其他操作。

证明。 在之后的任何时刻，活跃代都基于至多早于 $2K_i$ 次本地更新的快照。因此，如果当前支持集至多为 K_i，则该快照至多包含 $3K_i$ 个键，并且在成功事件发生时会被接受。一个已接受的代在其整个服务区间内至多包含 $5K_i$ 个当前条目或零频墓碑条目：它开始时至多包含 $3K_i$ 个键，并且之后至多 $2K_i$ 次本地更新可以引入新身份。因此，无论该桶的本地时钟在全局时间中推进得多慢或多快，本地就绪不变量和空间界都成立。

安装时的当前身份。 解码键至多有 $3K_i$ 个，且在构建期间发生变化的不同键至多有 K_i 个。我们为解码预留块的固定首段，然后以超过到达速率的服务速率扫描这些键。新遇到的键会以 $F_\tau(p) + \Delta_\tau(p)$ 进行初始化。扫描赶上后，新插入的键会立即得到处理。提高每次更新常数，可以为截止时间前处理所有 $4K_i$ 个键预留足够的工作量，其中包括字典操作。因此，一个已安装字典已经为每个具有正的当前频率的键包含一个引用。交换服务根节点只需恒定的元数据工作。代际退役采用惰性方式。其条目在后台服务器递减它们之前仍保留为引用，并且可能将仍然存在的点保留为额外候选项。这些额外候选项只会有助于直径下界估计。如果某个原始点在其退役引用得到处理之前被删除，则其全局重数变为零，并立即从几何结构中移除；旧引用无法产生过时距离。重新插入同样安全，因为全局频率记录会继续接收原始更新。

零计数墓碑和未处理的已退役条目均计入存储上界。每个局部字典条目只贡献一个引用。没有引用的点记录可以通过常数次记录操作释放；必要时，使用标准的代际/版本检查。

工作量、内存与独立性。 每次原始更新至多推进 r 个本地时钟。在这些桶中的每个桶内，它都会更新稀疏草图、执行有界的局部记账，以及固定多项式量的解码/物化工作。因此，每次原始更新产生的新引用数量和全局点创建数量，都受允许参数的某个多项式 A_0 所界定。原始频率更新会针对其自身标识至多再执行一次几何插入/删除。所有这些操作都按固定顺序执行。

设 $S_0 = O(\sum_i K_i)$ 为未退役桶、快照和全局引用存储的多项式上界。退役整代时，只需常数工作量即可将其根加入队列。每次原始更新至少处理 $2A_0$ 次常数级的记录清理操作，并适当增大该常数，以涵盖引用递减和几何更新。在任意长度为 u 的更新区间内，最多有 $S_0 + A_0 u$ 条记录进入退役状态：包括区间开始时存活的记录以及新分配的记录。因此，退役积压至多为S_0。特别是，许多桶在退役后可能立即变为空闲状态，而不会因此积累无界的垃圾；清理时钟是全局的，而非局部的。

共享持久化快照采用引用计数，并延迟递减边。每条已分配的指针边只会被处理常数次，因此这些操作遵循相同的摊销收费。几何结构的内存占用相对于候选记录近似线性，其中也包含其自身已有界的退役开销。总内存仍为 $O(\sum_i K_i)$ 多项式，并且对于每个随机结果，总更新工作量均为最坏情况多项式。

为保留显式点原语的无知输入前提，应将其每次更新调用填充至已知的最坏情况上界，并将其计作一次已调度的候选操作。在给定外层投影、哈希和稀疏恢复所用随机性的条件下，候选序列及其插入/删除发生的时刻便只取决于原始固定数据流，而不取决于几何原语的私有随机选择。其失败预算可根据此类内部更新次数的多项式上界来设定。该上界的对数对于允许的对数参数而言仍为多项式。

外部引用回收队列与几何原语的私有回收队列彼此分离。每个队列都有各自的固定工作预算。插入或移除候选项的外部清理操作会执行一次填充的原语调用；该调用内部的私有分配不会消耗外部清理预算，也不会改变其顺序。这样可以防止私有几何随机性改变外部引用丢弃的时机，并使前述独立性声明成为一个明确的

实现不变量。 □

8 显式点的直径证书

此例程的输入是候选转换的有序序列。它的两个输出承担不同的作用。见证分数用于处理候选项对。一个由当前锚点支持的短球列表，使算法 1能够将候选项与冻结样本进行比较。

8.1 算法

8.1.1 球链与精确叶节点

一个前景副本是一条单链。一个节点有一个显式输入集 P_i、一个固定球 $B_i = B(c_i, (1+\eta)\sqrt{V_i})$，以及一个子节点，其输入恰好为 $P_i \setminus B_i$。小型叶节点会被直接存储和处理。中心和方差有一个凸证书

$$c_i = \sum_{a \in K_i} \lambda_a a, \qquad V_i = \sum_{a \in K_i} \lambda_a \|a\|^2 - \|c_i\|^2, \quad \lambda_a \ge 0, \quad \sum_a \lambda_a = 1,$$

其中 K_i 由均匀样本 S_i 中的 $O(\eta^{-2})$ 个点组成，而 B_i 包含整个样本。下一小节将给出一种显式的有理数构造。令 $R_{\mathrm{hier}} = \lceil C_R \ln((T_C+2)/\delta) \rceil$ 为用于放大的独立副本数量。在每个非终端节点处取

$$s_{\mathrm{node}} = \left\lceil C_s \left(d + \ln((T_C+2) R_{\mathrm{hier}} \log(M+2)/\delta) \right) \right\rceil$$

样本点，或者在样本量更小时使用整个输入。在每次抽取时，以确定其输入快照的历史为条件，Lemma3.1 的半网保证适用。每个 T_C 候选转换最多访问 $O(\log(M+2))$ 个作业，每个副本中每项作业最多包含 $O(\log(M+2))$ 个采样层级。因此，即使作业被取消， $O((T_C+1) R_{\mathrm{hier}} \log^2(M+2))$ 也能界定所有副本中的样本数量。取足够大的 C_s 后，并集界可使理想均匀秩实验中任意半网测试失败的概率至多为 $\delta/8$ 。我们使用生成样本球时相同的输入快照来检查该样本球。若半覆盖失败、发生链深度溢出，或某项作业在固定截止期限到达时仍未完成，则受影响的层次结构副本将不可用。深度上限取足够大的常数倍 $\log(M+2)$。这些检查可保证任意结果下的资源使用量有界。每个副本都从一个空的精确叶节点开始。对于每个锚点 $a \in K_i$，存储一个距离堆，其中包含到P_i每个当前点的距离。节点分数为

$$H_i = \max_{a \in K_i,\, p \in P_i} \|a - p\|.$$

如果抽样后删除了整个样本 S_i 中的任何点，则永久标记该节点的得分在该代中不可用。这一做法有意采取了比必要程度更严格的措施；这样做足以保留简单的均匀采样分析。有效节点满足 $K_i \subseteq P_i$，因此 H_i 是当前点对距离。即使球的得分不可用，路由所用的球仍保持不变。所有堆都会持续跟踪当前节点输入。副本报告可用节点得分的最大值，包括其精确叶直径。每个这样的节点样本从抽取时起就受到监控。构建期间发生的删除会在路由或重放时记录下来，因此被失效的节点会将该标记带入服务阶段。已禁用副本不贡献得分或证书。

8.1.2 构造有理数证书

给定一个均匀样本，我们既需要一个包含该样本的球，也需要少量能够见证其方差的原始点。算法 4 从一对点开始；只要球仍未覆盖某个样本点，就将少量权重转移给一个最远点。

算法 4 有限样本的简洁有理方差证书

输入：非空有限样本 S **和有理数** $0 < \eta \le 1/4$。**输出：锚点集** K**、凸权重、均值** c **和方差** V**，满足** $S \subseteq B(c, (1+\eta)\sqrt{V})$。**1：选择** $a_0 \in S$ **和** $b \in \arg\max_{x \in S} \|x - a_0\|$。**2：如果** $b = a_0$**，则 3：返回** $(K = \{a_0\}\ \lambda_{a_0} = 1, c = a_0, V = 0)$。**4：结束 if**。**5：** $\lambda \leftarrow (e_{a_0} + e_b)/2$**；** $\theta \leftarrow (1+\eta)^2 - 1$**；** $\gamma \leftarrow \theta/32$。**6：重复**。**7：** $c \leftarrow \sum_x \lambda_x x$**；** $V \leftarrow \sum_x \lambda_x \|x\|^2 - \|c\|^2$。**8：** $a \leftarrow \arg\max_{x \in S} \|x - c\|$**；** $R^2 \leftarrow \|a - c\|^2$。**9：如果** $R^2 \le (1+\eta)^2 V$**，则 10：返回** $(K = \mathrm{supp}(\lambda)\ \lambda, c, V)$。**11：结束 if**。**12：** $\lambda \leftarrow (1-\gamma)\lambda + \gamma e_a$。**13：结束重复**。

8.1.3 在固定截止时间重建后缀

球一旦安装，就会在路由中保持固定。每个前台节点至多拥有一个任务，该任务从上到下重建其后缀。其截止时间必须在抽取新样本之前确定。这样，重建决策就不会取决于这些样本是否保留下来。我们使用整数调度参数

$$B_{\mathrm{sched}} = \lceil C_B s_{\mathrm{node}} \log(M+2) \rceil, \qquad H_{\mathrm{leaf}} = 128 B_{\mathrm{sched}},$$

其中， C_B 足够大，且 $B_{\mathrm{sched}} \ge 128$。因此，即使重新计算其所有两两距离，大小为 $O(H_{\mathrm{leaf}})$的精确叶节点也能符合多项式工作预算。前台节点安装后，立即令 m_{start} 表示其输入的当前大小。其下一次重建从安装时开始，截止时间为

$$\ell = \lfloor m_{\mathrm{start}} / B_{\mathrm{sched}} \rfloor$$

该节点输入的更新。该任务会重建其整个后缀。如果某个祖先节点先替换了此节点，则取消该任务。重建时机从不取决于该节点的样本是否保留、其得分是否可用，或该样本中的标识。精确叶节点在其首次超过 H_{leaf} 时启动后缀任务；如果安装时已经超过该阈值，则立即启动。它会一直保持精确状态，直到该任务提交。其初始大小至多为 $2H_{\mathrm{leaf}}$，并且最多还会接收 $2H_{\mathrm{leaf}}/B_{\mathrm{sched}}$ 次输入更新，因此直接处理仍保持多项式复杂度。每个前台节点可以拥有一个任务。关键是，未安装的待处理链中的节点不会启动重建任务。该规则避免了递归重放旧层次结构和待处理层次结构而产生的递推关系 $U(m) \simeq 2U(m/2)$。

当截止时间 COBII合时，我们首先让更新经过活动前景链及其待处理前缀，并执行它们的预算工作。随后，我们按照祖先优先顺序提交到期任务，取消每个被替换后缀中的任务。新安装的节点已经通过重放包含了当前更新，将自身时钟从零开始计算，并且只接收未来的更新。因此，后缀交换不会重复处理一次输入更新，也不需要将大量数据重放到新启动的任务中。

发起节点的输入更新就是任务的宿主时钟。其固定截止期限为 D。该任务由新节点的已完成前缀和一个精确的当前残余字典 Q 组成；后者是未由该前缀服务的集合。在尚不存在任何前缀之前，它可以与宿主字典建立别名。一次宿主更新至多改变 $|Q|$ 一个单位。该任务更新其已完成前缀中的堆，并在变更位于所有已完成球之外时，将其路由至 Q 。该任务内部不会同时递归运行第二个待处理层次结构。

令 $m = |Q|$ ，并令 r 统计 D 之前剩余的宿主更新次数。对于 $m > H_{\mathrm{leaf}}$ ，任务在 $r > m/B_{\mathrm{sched}}$ 时等待。在首次越过阈值时，它从当前 Q中抽取一个新的均匀样本，计算其锚点，并为该层启动堆和子残余字典。样本计算的工作量关于 s_{node}，d 是多项式的；扫描/复制/堆构建和追赶所需的工作量为 mW ，其中多项式固定于所显示的参数 W 。我们为每次宿主更新分配足够大的 $B_{\mathrm{sched}}W$ 工作量常数倍，从而使一个层能够在 $\lceil m/(16B_{\mathrm{sched}})\rceil$ 次宿主更新内完成。该任务遍历一个持久快照，同时记录新的变更，以快于变更到达的速度按顺序重放这些变更，然后将其已完成前缀切换为直接更新。它在残余集上等待下一次越过阈值。如果 $m \le H_{\mathrm{leaf}}$ 且 $r > 2H_{\mathrm{leaf}}/B_{\mathrm{sched}}$，它只维护残余字典。一旦 $m \le H_{\mathrm{leaf}}$ 和 $r \le 2H_{\mathrm{leaf}}/B_{\mathrm{sched}}$ 均成立，它便以多项式工作量构建一个精确叶节点，并在截止期限前持续使其保持最新。在等待期间，如果残余集增长超过 H_{leaf} ，则恢复大层级规则。

有界工作量的均匀抽取。 若要获得严格的最坏情况运行时间界，可以使用固定数量的随机比特生成均匀秩，而不是使用无界的拒绝循环。对于大小为 m 的当前字典，取一个均匀分布的 b_0 位整数 U，并使用秩 $1 + \lfloor mU/2^{b_0} \rfloor$。它与均匀秩分布之间的全变差距离至多为 $m/2^{b_0}$。我们取 $b_0 = \lceil \log_2(100(M+1)(D+1)/\delta) \rceil$，其中 D 是所有副本中样本抽取次数的上界，并且关于更新时域和工作量参数为多项式。无放回抽样可以通过从剩余秩中抽取，并临时排除顺序统计量来实现。随后进行并合耦合，除分配给此项的单独小失败概率外，所有样本均与理想均匀样本相同。每次抽取的工作量及其随机比特消耗均确定性地为多项式。

8.2 分析

引理 8.1 (显式点证书层级)。 设一个无知序列至多包含 T_C 次插入和删除，并维护 $\mathbb{Q}^d$ 中至多 M 个显式点，其坐标位长至多为 b。对于有理数 $0 < \eta \le 1/4$，一种随机化数据结构维护得分 L ，并在 $L > 0$ 时维护一个当前见证。以至少 $1 - \delta$ 的概率，在每次更新时以下结论同时成立：

$$L \le \operatorname{diam}(P) \le \sqrt{(1+\eta)^2 + 1}\, L.$$

在此事件下，每个时刻至少有一个副本由有效节点球以及其精确叶节点中的单点所覆盖。每个副本仅暴露 $O(\log(M+2))$ 个节点球和多项式数量的精确叶点。一个有效暴露球有 $O(\eta^{-2})$ 个当前锚点，其非负有理权重之和为一。它们的均值 q 和方差 V 给出半径 $(1+\eta)\sqrt{V}$。权重的编码 q，

以及 V 的长度均为更新界参数的多项式。叶节点点以单位权重和零方差暴露自身。更新时间在 d、b、η^{-1}、 η的编码长度、$\log(M+2)$和 $\log((T_C+2)/\delta)$中为最坏情况多项式。空间为 M 乘以这些参数的多项式，答案会被缓存。对 $\eta > 1/4$ 的保证可通过以 $1/4$运行该结构得到。

证明。 我们首先证明有效链的质量。然后说明，其固定截止期限为构建留出了充足时间，并且在任意固定输出时刻，独立采样链以常数概率有效。

方差构造的进展。 如果初始最远点就是起点，则样本是一个单元素集合。否则，令 $D_0 = \|a_0 - b\|$ 表示算法 4中的初始点对。初始方差为 $V_0 = D_0^2/4$。令 $\theta = (1+\eta)^2 - 1$ 且 $\gamma = \theta/32$，若停止测试未通过，则中心从 c 移至 $(1-\gamma)c + \gamma a$，其中 $R = \|a - c\|$ 是当前最远距离。方差变化恰好为

$$V' - V = \gamma(R^2 - V) - \gamma^2 R^2.$$

S 的直径至多为 $2D_0$。由于 $c \in$ 属于 $\mathsf{conv}(S)$，因此 $R^2 \le 4D_0^2 = 16V_0 \le 16V$。当停止测试未通过时， $R^2 - V > \theta V$，所以

$$V' - V \ge \gamma\theta V - 16\gamma^2 V = \theta^2 V/64.$$

另一方面， S 上的每个凸分布都具有

$$V = \frac{1}{2}\sum_{a,b\in S} \lambda_a\lambda_b\|a-b\|^2 \le \frac{1}{2}\operatorname{diam}(S)^2 \le 2D_0^2 = 8V_0.$$

因此，循环在 $O(\theta^{-2}) = O(\eta^{-2})$ 次迭代后停止。每次迭代最多添加一个支撑点，并且只扫描 S，因此运行时间关于 s_{node}， d， η^{-1} 和输入位长是多项式的。由于混合系数 γ 是固定的有理数，所有权重的公分母在每次迭代中只增加一个固定因子。其比特长度以及 c， V 的比特长度因此关于 b， d， η^{-1} 和 η 的编码长度是多项式的。所有比较都使用有理数平方距离。特别地，服务球具有有理中心和有理平方半径。

见证分数与覆盖。 在有效链上，每个点要么位于覆盖它的第一个样本球中，要么保留在精确叶节点中。在覆盖节点处，从其锚点到任意点 y 的加权平方距离平均为 $V_i + \|c_i - y\|^2$。因此，同一组锚点既能证明节点分数，也能证明第 4节中使用的交叉评分。如果每个前台节点的节点分数都有效，则该副本具有所需的近似。对于直径点对 x， y，取第一个节点，使得其中一个端点（设为 x）位于其服务球中。两个端点仍属于其节点输入。如果 $c = c_i$， $r = \sqrt{V_i}$，且 $t = \|c - y\|$，则凸证书给出 $H_i^2 \ge \sum_a \lambda_a\|a-y\|^2 = V_i + t^2$。因此

$$\|x - y\| \le (1+\eta)r + t \le \sqrt{(1+\eta)^2 + 1}\, H_i.$$

精确叶节点处理此前两个端点均未被服务的情况。

完成与几何收缩。 一个大层级作业会扫描并追赶，在 mW 时间内完成固定多项式 W 所需的工作，包括堆和字典操作。按照选定的预算，其主机更新耗时至多为

$$\left\lceil \frac{m}{16B_{\mathrm{sched}}} \right\rceil \le \frac{m}{16B_{\mathrm{sched}}} + 1 \le \frac{m}{8B_{\mathrm{sched}}} \qquad (m \ge 64B_{\mathrm{sched}})$$

主机更新。重放的处理速度快于新变化的到达速度，因此，已完成的堆和残余集都是最新的。每个阶段都有充足的时间。由于 m 每次主机更新最多变化一，因此在首次越过时

$$r \geq m/B_{\mathrm{sched}} - 2.$$

对于 $m \geq 64B_{\mathrm{sched}}$，阶段结束后至少还剩 $0.8m/B_{\mathrm{sched}}$ 次更新。在覆盖事件发生时，新的残余集至多包含 $m/2+m/(8B_{\mathrm{sched}})$ 个 $< 0.51m$ 点（对于我们的较大 B_{sched}）。因此，在该残余集自身的阈值 $r \leq |Q|/B_{\mathrm{sched}}$ 达到之前，还有时间再次等待。如果残余集足够小，可以安装一个精确叶节点，则在截止时间之前其大小至多为 $H_{\mathrm{leaf}} + 2H_{\mathrm{leaf}}/B_{\mathrm{sched}} < 2H_{\mathrm{leaf}}$。如果一个等待中的小残余集增长超过 H_{leaf}，则前一步具有 $r > 2H_{\mathrm{leaf}}/B_{\mathrm{sched}}$。由于 m 最多变化一，越过该阈值时仍有超过 m/B_{sched} 的时间。因此，同样的首次越界估计适用于下一个较大级别。这些情况涵盖了安装前的最后一次更新。因此，从 m 个点中抽样得到的每个新非终端节点，在<用固定>安装截止时间</用固定>之前，最多经历 $m/B_{\mathrm{sched}} + 2$ 次主机更新。该节点的输入在安装前最多变化这么多个点。安装后，其下一项任务会在其自身输入的 $m_{\mathrm{install}} \leq m + m/B_{\mathrm{sched}} + 2$范围内完成提交，即在 $m_{\mathrm{install}}/B_{\mathrm{sched}}$ 次更新之内。因此，从其抽样时刻到其替换时，至多

$$m/B_{\mathrm{sched}} + 2 + m_{\mathrm{install}}/B_{\mathrm{sched}} \leq 3m/B_{\mathrm{sched}}$$

可能会发生相关删除；必要时可增大小集合阈值。

在节点的整个活跃期内，输入规模与子节点规模满足

$$|P_i| \geq m - 3m/B_{\mathrm{sched}}, \qquad |P_{i+1}| \leq m/2 + 3m/B_{\mathrm{sched}}.$$

对于 $B_{\mathrm{sched}} \geq 128$，可得 $|P_{i+1}| < 0.6|P_i|$。因此，所有前台链和待处理前缀的深度均为 $O(\log(M+2))$。待处理前缀的构造规模呈几何递减，只有受控的小幅变化例外，因此其显式存储总量与宿主输入规模呈线性关系，至多相差多项式级别的簿记开销。

固定截止时间下的存活。在预先确定的原始序列中固定一个输出时刻 t。对于在 t时刻处于活跃状态的节点，给定其所有祖先节点的样本以及该节点新抽样之前的调度历史。此时，其输入快照已固定，其样本是在该快照上均匀抽取的。样本时刻和原始任务的固定安装截止时间在此次抽样之前就已确定；之后的重建截止时间取决于其输入规模和时钟，而这些又取决于祖先节点，不取决于该节点自身的样本。子节点构造时间、子节点规模以及该节点自身的样本存活情况都不会改变这些截止时间。祖先节点的替换只会缩短该节点的存活时间。该节点自身的样本会影响其服务球及其后代节点，但不会影响其自身的输入序列。

因此，安装前主机更新窗口与安装后自身输入窗口的并集，能够独立于该样本确定至多 $3m/B_{\mathrm{sched}}$ 个在节点处于活动状态时可能被删除的快照点。均匀分布的 s_{node} 个点命中该集合的概率至多为 $3s_{\mathrm{node}}/B_{\mathrm{sched}}$。反过来以节点存活为条件化依据并无必要，应当避免。固定截止时间至关重要：若每当一条待处理链恰好完成时就提交整条链，可能会选出依赖于样本的生命周期，因此需要采用不同的论证。

为处理依赖于样本的覆盖，即使覆盖测试失败，我们也会将该逻辑时隙计入其固定的安装截止时间。通过覆盖测试且丢失一个样本点的事件，包含于无条件样本命中事件中。覆盖失败单独计入。条件化

祖先样本是独立于该节点新抽样的，因此以祖先样本为条件是合理的。这样就能将全局覆盖失败事件与存活论证分开，同时不会使样本产生偏差。

在固定的输出时刻，从根节点向下展开活动链。在每个深度，祖先历史和输入时钟共同确定正在服务的逻辑时隙。其样本命中界为 $3s_{\mathrm{node}}/B_{\mathrm{sched}}$。对 $O(\log(M+2))$ 个深度求和可知，对于足够大的 C_B，在通过覆盖测试的时隙中命中的概率至多为 $1/8$。因此，我们将副本中覆盖测试失败或有限秩耦合的总概率限制为至多 $1/8$。由此，在该时刻，给定副本拥有完整有效覆盖的概率至少为 $3/4$。这 R_{hier} 个副本在固定输入上使用独立种子。在该时刻所有副本均失败的概率至多为 $4^{-R_{\mathrm{hier}}}$。对 T_C+1 个输出应用并集界，即可证明最大可用得分的同时保证。无论结果如何，每个可用得分都是一个当前点对距离。

最坏情况工作量与存储。 一次前台更新只沿着某个副本的单条前台链进行。在每个访问到的节点，它都会按固定的多项式预算推进该节点唯一的待处理作业，并更新该作业已完成前缀，其深度为 $O(\log\ M)$。因此，每次单点更新只处理 $O(\log^2 M)$ 个前缀/节点关联；待处理作业不会递归调用其他待处理作业。样本选择使用顺序统计字典。快照捕获、堆扫描、重放队列和退役通过持久化或写时复制结合后台工作来实现。替换后缀只需切换指针；其旧存储区可在相同的多项式预算下延迟回收（也可使用代际标记循环利用）。初始和最终的小型集合直接处理。

下面给出明确的全局回收开销。在一次前台更新中，只执行有界的多项式数量 a 个字分配/初始化；大型快照和数组采用增量方式分配，或者使用惰性代际标记。我们将所有已退役记录保存在回收队列中，并在每次更新时至少处理 $2a$ 单位的回收工作，同时调整常数以计入指针/引用清理。如果非退役存储量有界于 $S = M\,\mathrm{poly}(d,\ b,\ s_{\mathrm{node}},\ \log\ M)$，那么在任意长度为 u 次更新的区间内，进入回收队列的体积至多为 $S+au$：它包括区间开始时仍存活的记录，以及区间内分配的记录。因此，速率为 $2a$ 的服务在任何时刻的积压量至多为 S（取以该时刻为终点的所有区间上的上确界即可）。因此，已退役存储只会额外增加 S，而回收开销为最坏情况多项式。共享的不可变快照可以采用引用计数，并将递减操作延迟处理；每条已分配的指针边只会产生常数级的总清理工作。在祖先节点提交时取消大量子作业，不会违反这一突发量加速率的界。

几何递减的前台输入规模之和为 $O(M)$。每个前台节点拥有一条待处理链，其存储量在多项式因子范围内与其输入规模呈线性关系。因此，一个副本使用 $M\,\mathrm{poly}(d,\ b,\ s_{\mathrm{node}},\ \log\ M)$ 空间，其中包括当前和即将退役的内存区。复制仅增加 $O(\log(T_C/\delta))$。所有前台工作都明确受多项式预算限制，因此无论随机选择如何，均可得到最坏情况界。堆中的距离和球测试可以通过有理平方距离进行比较。最终输出可以保留精确的平方见证距离，因此并未将任意实数精度隐藏在一个机器字中。 □

9 冻结样本中的最远点查找

主算法可以承受样本大小的多项式预处理，但必须在关于维数和对数参数的多项式时间内回答每个中心查询。我们为数量有界的每个投影网格点预先计算一个最远样本点。在给定查询处达到精度要求，需要整个样本的上畸变界，以及仅针对一个

最远点。

9.1 算法

对于 $m = 1$，我们返回唯一的样本点。否则，我们通过检查所有样本点对来计算直径 D 及其见证，并确定 $s_0 \in S$。对于 $R = \|q - s_0\| \geq D/(\kappa - 1)$，答案$s_0$ 已经足够，因为每个样本点到 q 的距离至多为 $R + D \leq \kappa R$。其他查询使用下表。

我们选取固定的有理数 $b_0 > 1$、e_0、$\nu > 0$，使得 $\kappa(1 - \nu - 2e_0) > b_0 + \nu$。对于每个固定的 $\kappa > 1$，这样的常数都存在，必要时可在内部使用更小的有理数因子。对于 $I(b) = \frac{1}{2}(b^2 - 1 - \ln b^2)$，我们使用有理数 $A_\kappa > I(b_0)^{-1}$，并取 $k = \lceil A_\kappa \ln(240m) \rceil$，使用误差不超过 1 的对数上近似。我们选取有理数网格步长 $e_0 D/(2\sqrt{k}) \leq h \leq e_0 D/\sqrt{k}$ 和常数 $R_0 > (b_0 + \nu)/(\kappa - 1) + e_0$。一个表使用规范化的高斯投影 A，由有理数矩阵 $\widetilde{A}$实现，其条目误差至多为 $\nu/\sqrt{kd}$，除非发生概率为 $1/480$的情况。附录 A.2 给出了有限比特采样过程。

我们枚举 $[R_0 D, (R_0 + 1)D]$ 中以 As_0为中心、距离不超过有理数半径的格点 $As_0 + hj$。在每个格点处，我们存储投影距离最远的原始样本标识。对于一次查询，我们将 Aq 舍入到最近的格点（采用固定的平局规则），并返回存储的点；如果该格点不存在，则返回 s_0。我们构建 $O(\log((Q+1)/\delta))$ 个相互独立的表，并返回各表答复中按原始距离最远的点。

9.2 分析

引理 9.1 (最远点查找)。 对于任意固定的 $\kappa > 1$，任意非空的有界整数点集合都存在一种随机化数据结构，其预处理时间和空间为 m^{B_κ} 多项式$_\kappa(L\ b_q \log(Q+2))$，其中 $B_\kappa \geq 2$ 为固定值。查询时间为 多项式$_\kappa(L\ b_q \log(m+2), \log(Q+2))$。对于给定的 $[-\Delta, \Delta]^d$ 中任意一列有理数查询，其坐标编码长度至多为 b_q，且该查询列独立于种子；对于每个查询 q，该结构以 $1 - \delta$ 的概率返回一个满足 $\|z - q\| \geq$ 最大值$_{x \in S} \|x - q\|/\kappa$ 的标识 z，其中采用对数级重复因子。无论随机结果如何，每个答复都是原始样本标识。对于 $m = 1$，唯一的样本点能够精确回答每个查询。对于比特长度有界的规范有理数输入坐标，同样的结论也成立，其比特长度计入多项式因子。在直径应用中，b_q 和 $\log(Q+2)$ 均为 L 的多项式。

证明。 我们先界定固定查询的畸变，再计算网格点的数量。

真实最远点的投影。 令 $Y \sim \chi^2_k$ 和 $X = Y - k$，其中 $k \geq 1$为任意整数。其矩为

$$\mathbb{E}X = 0, \qquad \mathbb{E}X^2 = 2k, \qquad \mathbb{E}X^4 = 12k^2 + 48k \leq 60k^2.$$

对第一、第二和第四阶绝对矩进行插值，再应用柯西–施瓦茨不等式，可得

$$\mathbb{E}|X| \geq \frac{(\mathbb{E}X^2)^{3/2}}{(\mathbb{E}X^4)^{1/2}}, \qquad \mathbb{E}X_+ = \tfrac{1}{2}\mathbb{E}|X|, \qquad (\mathbb{E}X_+)^2 \leq \Pr[X \geq 0]\, \mathbb{E}X^2.$$

因此

$$\Pr[Y \geq k] \geq \frac{(\mathbb{E}X^2)^2}{4\mathbb{E}X^4} \geq \frac{1}{60}. \tag{20}$$

对于归一化高斯映射 $A\colon\ \mathbb{R}^d \to \mathbb{R}^k$，对于任意固定的非零向量 v，这表示 $\Pr[\|Av\| \geq \|v\|] \geq 1/60$。

对于上乘数 $b_0 > 1$，通常的精确 Chernoff界为

$$\Pr[\|Av\| > b_0\|v\|] \le e^{-kI(b_0)}, \qquad I(b_0) = \tfrac{1}{2}(b_0^2 - 1 - \ln b_0^2).$$

初等不等式 $x - \ln(1+x) \ge x^2/[2(1+x)]$ 给出了有用的显式下界

$$I(b_0) \ge \frac{(b_0^2-1)^2}{4b_0^2}.$$

固定查询的查找准确率。条目误差给出 $\|A - \widetilde{A}\|_{\mathrm{F}} \le \nu$。它使任意投影差 v 的范数最多改变 $\nu\|v\|$。对于固定查询，令 $p_* \in S$ 为按规范方式选出的真实最远点，并设 $R = \|p_* - q\|$。总共只有 m 个向量 $p - q$，其中包括 $s_0 - q$。我们对 k 的选择将它们合并后的理想上尾失败概率限制为 $1/240$。由（20）和并集界，联合事件

$$\|A(p_* - q)\| \ge R, \qquad \|A(p - q)\| \le b_0\|p - q\| \quad (p \in S),$$

再加上有限比特耦合成功，其概率至少为 $1/60 - 1/240 - 1/480 > 1/200$。这里没有假设上尾事件与下尾事件相互独立。令 u 为舍入查询点。它与 $\widetilde{A}q$ 的距离至多为 $h\sqrt{k}/2 \le e_0 D/2$，因此在良好事件下，根据 $s_0 - q$ 的上估计以及 R_0的定义，u 位于表中。如果存储标识为 x，那么 $\|Ax - u\| \ge \|Ap_* - u\|$。在两端分别应用舍入界可得

$$(b_0 + \nu)\|x - q\| \ge (1 - \nu)R - e_0 D \ge (1 - \nu - 2e_0)R,$$

其中 $D \le 2R$。因此，只需满足

$$\kappa(1 - \nu - 2e_0) \ge b_0 + \nu. \tag{21}$$

由于 $\kappa(1 - \nu - 2e_0) > b_0 + \nu$，这就证明了固定查询的查找保证。对于每个独立的表，该保证以至少 $1/200$ 的概率成立。取最远的实际返回点不会破坏任何成功结果。对 Q 个预先指定的查询应用并集界，即可证明同时保证。无论结果如何，返回的都是一个输入点。

枚举与查询成本。对于固定的 C_κ，表键满足 $\|j\| \le C_\kappa\sqrt{k}$。枚举过程递归地扩展可行整数前缀，并在某个前缀的平方范数超过平方半径时将其丢弃。在深度 i 处，以可行前缀为中心的互不相交单位立方体位于一个半径为 $C_\kappa\sqrt{k} + \sqrt{i}/2$ 的i维球内。由体积界可知，前缀数量至多为 $(C'_\kappa\sqrt{k/i})^i$。由于 $i\ln(k/i) \le k/e$，整个枚举过程需要 $\exp(O_\kappa(k))\,\mathrm{poly}(k) = m^{O_\kappa(1)}$ 次操作。对于每个完整键，算法扫描样本并存储取得最大值的原始身份。计算样本直径、生成映射以及构建查找字典，在增大 B_κ后也具有相同的多项式界。

一次查询会投影其中心、对网格键进行取整，并在平衡字典中查找存储标识。随后，它会计算原始距离。所有这些操作的耗时均为关于 $\mathcal{L}$ 和查询编码上界的多项式时间。独立地重复该表，只会使预处理和查询时间增加一个对数因子。明确选择 $B_{1.09} = 2000$ 的验证见附录 A.1。 □

10 线性长度集合流的下界

一次最终直径查询就能检测长插入前缀中的一位。我们首先说明定理 1.2所需的位数界。然后构造有限实例，并推导出强制满足该界的信息不等式。

令 S 表示插入前缀之后状态编码的比特长度，并将较短的描述填充至相同长度。因此， W 个机器字给出 $S = O(W w_{\mathrm{word}})$。对于固定的 $c \geq 1$ ，设$g = \lceil c \rceil + 1$、 $r = 1/g$、 $\delta_0 = 1/(100g)$、 $r_0 = r - 2\delta_0$ 和 $\eta = 1 - h_2(1/10)$，其中 h_2 为二元熵。所有熵的单位均为比特。

定理 10.1 (插入前缀中的信息)。 对于每个整数 $d \geq 4$ ，存在一个大小为 U 的有限有理点集 $P \subset \mathbb{R}^d$ ，满足

$$(25g)^{d-1} \leq U \leq (802g)^d, \tag{22}$$

使得以下结论成立。假设某个随机化单遍算法在该族中的每个数据流上都以至多 $1/10$ 的错误率返回 $D/c \leq L \leq D$ 。那么，它在插入前缀之后的状态满足

$$S + \lceil \log_2(U+1) \rceil > \frac{\eta}{4} \exp\left(\frac{(d-1) r_0^2}{1000} \right). \tag{23}$$

这些数据流的维数为 $d+1$，最大支撑集大小为 $U+1$ 或更小，长度至多为 $2U+1$。每个点最多插入一次、删除一次，且所有插入均先于删除。将所有坐标乘以相同的因子，可得到绝对值为 $O(g^2\sqrt{d})$ 的整数。 (23)中的对数项允许使用免费的更新时钟；若将时钟计入状态，则可以省去该项。

证明。 第10.1.1 节构造该网，第10.1.2节将流式状态转换为长度为 $s = S + \lceil \log_2(U+1) \rceil$ 比特的消息。解码器在弦长为 η 处强制边缘量至少为 $[0, r_0]$，并在整个 $[0, r_0]$处为零。第10.2 节中的对角不等式和旋转平均通过（25）约束这些端点之间的取值。第10.3 节表明，除非 $s > (\eta/4)\exp((d-1) r_0^2/1000) r_0^2/1000$，否则该递推关系与两个端点条件不一致。 □

10.1 有限输入与解码器

10.1.1 无理数对跖网

我们取单位球面的一个实数 $\delta_0/2$-网$\mathcal{V}$ $\mathbb{S}^{d-1}$ ，其中

$$|\mathcal{V}| \leq (1 + 4/\delta_0)^d.$$

例如，极大的 $\delta_0/2$-分离集具有这一大小界：以其各点为中心、半径为-$\delta_0/4$ 的球彼此不相交，并且都容纳在半径为-$(1 + \delta_0/4)$的球内。极大性给出覆盖性质。我们加入所有选定点的负点。对于每一对对跖点，我们将其中一个代表舍入到最近的网格点 p，并使用 $-p$作为其对跖点。所有位置都使用相同的网格，其分母为

$$J = \left\lceil \frac{2\sqrt{d}}{\delta_0} \right\rceil.$$

舍入误差至多为 $\sqrt{d}/(2J) \leq \delta_0/4$。 去除重复项后，对称有理点集 $\mathcal{P}$ 的基数至多为 $2(1 + 4/\delta_0)^d \leq (802g)^d$，且每个单位向量与 $\mathcal{P}$中的某个点的距离都不超过 δ_0 。$\mathcal{P}$ 中的每个点的范数介于 $1 - \delta_0$ 和 $1 + \delta_0$之间。

我们固定一个可测的最近网映射

$$Q : \mathbb{S}^{d-1} \longrightarrow \mathcal{P}$$

其中 $\|Q(x) - x\| \le \delta_0$ 和 $Q(-x) = -Q(x)$。这样的奇性平局规则确实存在：先按固定顺序选取最先出现的最近反极点对，再选择距离较近的一侧；若两侧距离相同，则采用任意固定的奇性半球约定。下面使用的旋转分布下，平局无关紧要，但这样可以使映射精确。

为完整起见，球面的任意 δ_0-网的规模也必须关于 d 指数增长。以各网点为球心、半径为 $2\delta_0$ 的球覆盖了半径介于 $[1, 1+\delta_0]$ 的球壳。比较体积可得

$$U(2\delta_0)^d \ge (1+\delta_0)^d - 1 \ge \delta_0.$$

对于 $d \ge 2$，这意味着 $U \ge (1/(4\delta_0))^{d-1} = (25g)^{d-1}$，从而证明了（22）。

$\mathcal{P}$ 的所有坐标的分母均为 J。下文使用的额外锚点高度为 $r = 1/g$。将所有坐标乘以 Jg，即可使每个点的坐标都成为整数，且坐标绝对值上界为 $O(gJ) = O(g^2\sqrt{d})$。这种缩放会保留所有直径比。只对选定的网点进行舍入，可使基数上界不受环境有理网格大小的影响。

10.1.2 一个球冠、一个目标和一个新锚点

Alice 拥有独立公平比特 X_p，每个 $p \in \mathcal{P}$ 对应一个。她在 $X_p = 1$ 时按规范顺序插入 $(p, 0)$，并将流式状态发送给 Bob；如需指定自由时钟，则同时发送她的插入计数。独立地在 $\mathbb{S}^{d-1}$ 上均匀分布地选择 x。Bob 的目标是

$$y = Q(-x) = -Q(x).$$

他知道外部的比特

$$\mathcal{B}_x \cup \{y\}, \qquad \mathcal{B}_x = \mathcal{P} \cap B(Q(x), r).$$

目标位于 $\mathcal{B}_x$ 外部，因为 $2\|Q(x)\| \ge 2(1-\delta_0) > r$。Bob 首先插入新锚点（$Q(x)$ r），其最后一个坐标将其与每个 Alice 点区分开来。然后，他删除该已知补集中的每个现有点。因此，完整的数据流中所有插入都发生在所有删除之前，并且没有任何点被重新插入。

最多有 $2U + 1$ 次更新，每个前缀都是一个集合，且最多有 $U + 1$ 个点处于活动状态。一旦选定 Alice 的比特和 Bob 的查询，整个更新序列就独立于算法的随机硬币而固定；因此，这个下界对非自适应数据流同样成立。如果 $X_y = 0$，则只有锚点和插入的球冠点会保留下来。若没有球冠点，锚点就是一个单元素集合，直径为零。否则，任意两个球冠点之间的距离都不超过 $2r$，且任意球冠点到锚点的距离都不超过 $\sqrt{2}\, r$。因此 $D \le 2r$。如果 $X_y = 1$，则 $(y, 0) = (-Q(x), 0)$ 和锚点都会保留下来，因此

$$D \ge \sqrt{4\|Q(x)\|^2 + r^2} \ge 2\|Q(x)\| \ge 2(1 - \delta_0).$$

第二种情形中的下界与第一种情形中的上界之比为

$$\frac{2(1-\delta_0)}{2r} = \frac{1-\delta_0}{r} = g - 1/100 > c.$$

在近似事件 $D/c \le L \le D$ 上，Bob 因此恰好在 $L > 2r$ 时解码出 $X_y = 1$（在未缩放的坐标中）。尽管使用 x 来描述查询分布，Bob 只需要有限数据 Q（x）和 y；解码器不会获得实值侧信道。

通过取平均，固定公共随机性和 Alice 的私有随机硬币，使平均错误率至多为 $1/10$。Bob 可以使用新的独立随机性，也可以用针对该分布的最优确定性解码器取而代之。由此得到的确定性消息 $M = M(X)$ 至多为

$$s = S + \lceil \log_2(U+1) \rceil$$

比特。定义

$$f(A) = H(M \mid X_{\mathcal{P}\setminus A}), \qquad A \subseteq \mathcal{P}.$$

函数 f 是规范化、单调且次模的，其中 $f(\mathcal{P}) \le s$，并且单点边际属于 $[0,1]$。归一化是因为 M 是比特的确定性函数；单调性则可通过移除条件化得出。对于次模性，写作

$$f(A) = H(M, X_{\mathcal{P}\setminus A}) - |\mathcal{P} \setminus A|.$$

熵具有次模性，补集运算会交换并集与交集，而被减去的基数是模的。单点边际是与一个独立公平比特的条件互信息，因此介于零和一之间。

为明确说明随机性记账，可以将 Alice 的私有硬币表示为一个独立的随机转移硬币表，并以更新时间、当前状态和更新符号为索引。沿任意一个前缀，每个时间索引只会使用一次，因此这与使用新鲜私有硬币的分布完全相同。固定公共种子和该前缀表后，M 便是仍然相互独立的输入比特的确定性函数。Bob 在后续过程中使用新鲜转移硬币，并根据传输的前缀长度确定一个自由时钟（如果存在）。对前缀表和公共种子取平均，可保证存在一个固定选择，使其达到规定的平均错误率。无需私有随机带头或其他未计入的历史信息。

对于连续球面点 x、z，令

$$a(x,z) = f(\mathcal{B}_x \cup \{Q(z)\}) - f(\mathcal{B}_x).$$

若 $\|x - z\| \le r_0 = r - 2\delta_0$，则 $\|Q(x) - Q(z)\| \le r$，因此 $a(x,z) = 0$。令 $a(t)$ 表示球面上弦长为 $t \in [0,2]$ 的旋转不变有序点对上的 $a(x,z)$ 平均值。因此，对 $t \le r_0$，$a(t) \in [0,1]$且 $a(t) = 0$。对于 $z = -x$，在固定 x 时，边缘量等于 $I(M; X_y \mid X_{\mathcal{P}\setminus(\mathcal{B}_x \cup \{y\})})$。在看到 M 之前，即使给定这些旁信息，目标比特仍是均匀随机的。看到 M 后，Bob 对其解码的平均错误率至多为 $1/10$。对于固定的 x，令 e_x 表示确定性解码器的条件错误率。给定消息和旁信息后，目标比特由其错误指示变量确定，因此其剩余熵至多为 $h_2(e_x)$。对其取平均并利用 $\mathbb{E}e_x \le 1/10$ 以及 h_2 的凹性，可得二元 Fano 界 [CT06]

$$a(2) \ge \eta = 1 - h_2(1/10). \tag{24}$$

10.2 从对跖点到较短的弦

10.2.1 次模边缘分布的对角不等式

对角线上的一项表示解码器需要了解的关于其自身目标的信息。非对角线项衡量的是：如果较早的目标已经被揭示，该信息增益可能减少多少。选取一个随机子集后，只有在两个指标都被选中时，我们才需计入后者造成的损失。

引理 10.2（加权对角线不等式）。 对于任意总值不超过 s 的规范化单调次模 f、任意集合 B_1、……、B_m，以及任意元素 z_1、……、z_m（允许重复），令

$$A_{ij} = f(B_i \cup \{z_j\}) - f(B_i), \qquad D = \sum_i A_{ii}, \quad Z = \sum_{i,j} A_{ij}.$$

于是 $D^2 \le 2sZ$。

证明。 选取一组指标并将其排序。令 U_i 包含在指标i之前选取的元素。递减边际收益给出下面的第一个不等式。对于第二个不等式，将 B_i 收缩，写作$g_i(U) = f(B_i \cup U) - f(B_i)$。单调性给出 $g_i(U_i \cup \{z_i\}) \ge g_i(\{z_i\}) = A_{ii}$；而将信息增益沿前面的指标求和并消去中间项，可得 $g_i(U_i) \le \sum_{j \text{ 前面的 } i} g_i(\{z_j\})$。因此

$$\begin{aligned} f(U_i \cup \{z_i\}) - f(U_i) &\ge f(B_i \cup U_i \cup \{z_i\}) - f(B_i \cup U_i) \\ &\ge A_{ii} - \sum_{j \text{ preceding } i} A_{ij}. \end{aligned}$$

如果 z_i 已经出现过，这仍然成立：链增量为零，而前面相同的列会从右侧减去 A_{ii}。我们以概率 q 独立选取指标，并将它们均匀分布地排序。每个对角增益被选取的概率为 q。每一对不同指标按该顺序被选取的概率为 $q^2/2$。链上的求和并消去中间项后至多为 s，从而得到 $s \ge qD - q^2(Z - D)/2 \ge qD - q^2 Z/2$。如果 $D > 0$，则使用 $q = D/Z \le 1$ 得到 $D^2 \le 2sZ$。$D = 0$ 的情形显然。 □

10.2.2 在仿射子球面上取平均

我们利用一个仿射子球面的直径，实现任意指定弦长的对角线。所有数对的中心相同，因此选择两个不同的指标时，它们之间的距离变化仅取决于两个独立方向（图3）。

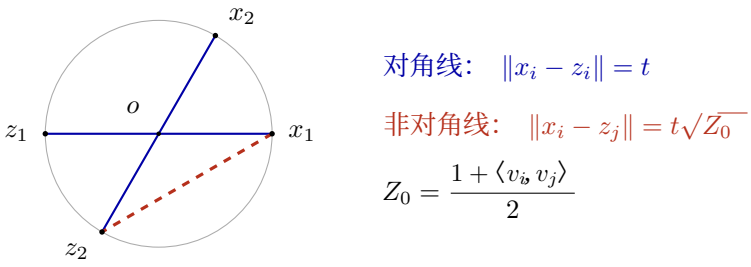


图 3：仿射子球面的二维截面，其中心为 $o = \sqrt{1 - R^2}\, u$，半径为 $R = t/2$。每一对 $(x_i,\ z_i)$ 在此子球面上互为对跖点，因此它们之间的距离恰为 t。对于独立方向，非对角线距离为 $t\sqrt{Z_0}$，其中 Z_0 集中在 $1/2$ 附近。这一步几何推导将熵边缘量从距离 t 传递到距离 $t/\sqrt{2}$ 附近。 –

固定 $t \in [0, 2]$，记 $R = t/2$，并选择一个均匀分布的单位向量 u。在正交超平面$u^{\perp}$中，取其单位球面 $\mathbb{S}^{d-2}$ 上相互独立且均匀分布的向量 v_1、……、v_m。定义

$$x_i = \sqrt{1 - R^2}\, u + R v_i, \qquad z_i = \sqrt{1 - R^2}\, u - R v_i.$$

这些点位于原单位球面上。每一对对角线点 (x_i, z_i) 的弦长为 t，且其分布具有旋转不变性。我们将引理10.2应用于有限表

$$A_{ij} = a(x_i, z_j).$$

其底层元素 $Q(z_j)$可能重合；前述引理明确允许这种情况。

对于 $i \neq j$,

$$\|x_i - z_j\| = R\|v_i + v_j\| = t\sqrt{Z_0}, \qquad Z_0 = \frac{1 + \langle v_i, v_j \rangle}{2}.$$

$\mathbb{S}^{d-2}$ 上的内积密度表明

$$Z_0 \sim \mathrm{Beta}(\alpha, \alpha), \qquad \alpha = (d-2)/2.$$

由旋转不变性可知，在弦长 v 的条件下，期望项为 $a(v)$。对 $v = t\sqrt{Z_0}$ 取平均后，每个非对角线项均为 $\mathbb{E}a(t\sqrt{Z_0})$。若 D_m, Z_m 分别表示对角线和表的总和，则

$$\mathbb{E}D_m = ma(t), \qquad \mathbb{E}Z_m = ma(t) + m(m-1)\mathbb{E}a(t\sqrt{Z_0}).$$

利用 $(\mathbb{E}D_m)^2 \le \mathbb{E}D_m^2 \le 2s\mathbb{E}Z_m$ 并令辅助表大小 m 趋于无穷，得到

$$a(t)^2 \le 2s\,\mathbb{E}a(t\sqrt{Z_0}). \tag{25}$$

该极限仅涉及一个含有重复项的数学表；它不会引入长数据流或扩大的输入空间。

10.3 迭代至零带

10.3.1 随机弦的尾界

该递推关系对下一条弦的长度取平均。我们利用一个球面坐标的偶数阶矩来界定其过长的概率。记为 $n = d - 1$ 和 $W = 2Z_0 - 1$。变量 W 服从球面 $\mathbb{S}^{n-1}$上均匀点的第一坐标的分布。对称性给出所有奇数阶矩均为零。对于每个整数 $j \ge 1$，贝塔积分给出

$$\mathbb{E}W^{2j} = \frac{\Gamma(j + 1/2)\Gamma(n/2)}{\sqrt{\pi}\,\Gamma(j + n/2)} = \frac{(2j-1)!!}{n(n+2)\cdots(n+2j-2)} \le \frac{(2j-1)!!}{n^j}.$$

因此，展开双曲余弦并利用单调收敛可知，对于任意实数 λ,

$$\mathbb{E}e^{\lambda W} = \mathbb{E}\cosh(\lambda W) \le \sum_{j=0}^{\infty} \frac{\lambda^{2j}(2j-1)!!}{(2j)!\,n^j} = \exp\left(\frac{\lambda^2}{2n}\right).$$

此处将 $j = 0$ 项解释为 1。结合 $\lambda = 2n\delta$，切尔诺夫不等式给出有用的依赖维数的上尾界

$$\Pr[Z_0 > 1/2 + \delta] \le \exp(-2n\delta^2), \qquad \delta \ge 0. \tag{26}$$

10.3.2 选择每一步的截断

每次应用递推关系都会使边缘量平方，因此固定尾部截断所带来的误差可能超过它本应界定的量。随着边缘量减小，我们提高截断值，并追踪距离中累积的损失。定义单调包络

$$A(t) = \sup_{0 \le v \le t} a(v), \qquad 0 \le t \le 2.$$

它是单调的，因而是可测的，位于 $[0,1]$，上消失 $[0, r_0]$，并满足 $A\left(2\right) \ge \eta$。对（25）取上确界，并利用 A的单调性，得到

$$A(t)^2 \le 2s\, \mathbb{E} A(t\sqrt{Z_0}). \tag{27}$$

数量 消息比特 s 是一个整数，并且根据（24）必须为正；因此 $s \ge 1$。令

$$L = \ln(4s/\eta), \qquad h = \left\lceil \log_2 \frac{16}{r_0^2} \right\rceil, \qquad \ell_j = 4s \exp(-2^j L), \qquad \delta_j = \sqrt{\frac{2^j L}{n}}.$$

注意， $\ell_0 = \eta$ 和 $\ell_{j+1} = \ell_j^2/\left(4s\right)$。假设存在矛盾，则

$$L \le \frac{n r_0^2}{1000}. \tag{28}$$

由于 $2^h < 32/r_0^2$,

$$\begin{aligned} \sum_{j=0}^{h-1} \delta_j &= \sqrt{L/n}\, \frac{2^{h/2} - 1}{\sqrt{2} - 1} \\ &< (\sqrt{2}+1)\sqrt{2^h L/n} \\ &< (\sqrt{2}+1)\sqrt{32/1000} < \tfrac{1}{2} < \ln 2. \end{aligned}$$

特别地，每个 $\delta_j < 1/2$ 对于 $j < h$。

从 $t_0 = 2$ 开始，定义

$$t_{j+1} = t_j \sqrt{1/2 + \delta_j}, \qquad 0 \le j < h.$$

所有这些值都属于 $[0,2]$。将（27）中的期望按 $Z_0 \le 1/2 + \delta_j$ 拆分，并利用 $A \le 1$，可得

$$A(t_{j+1}) \ge \frac{A(t_j)^2}{2s} - \Pr[Z_0 > 1/2 + \delta_j].$$

由（26），被减去的项至多为

$$\exp(-2^{j+1}L) \le 4s \exp(-2^{j+1}L) = \frac{\ell_j^2}{4s}.$$

若 $A(t_j) \ge \ell_j$，主项至少比尾部大

$$A(t_{j+1}) \ge \frac{\ell_j^2}{2s} - \frac{\ell_j^2}{4s} = \frac{\ell_j^2}{4s} = \ell_{j+1} > 0.$$

由 $A(2) \ge \eta = \ell_0$ 归纳可得 $A(t_h) > 0$。另一方面，$\ln(1+2\delta) \le 2\delta$ 可得

$$t_h = 2^{1-h/2} \prod_{j=0}^{h-1} \sqrt{1+2\delta_j} \le \frac{r_0}{2} \exp\left(\sum_{j=0}^{h-1} \delta_j \right) < r_0.$$

这与 A的零带相矛盾。因此，28）不成立，并且

$$s > \frac{\eta}{4} \exp\left(\frac{(d-1) r_0^2}{1000} \right),$$

这就证明了（23）。

阈值的这种选择保留了对 c的依赖。平方偏差 δ_j^2 的增长速度与 2^j相同，恰好等于 $-$自然对数$(\ell_j/(4s))$的增长速度。与此同时，$\sum_j \delta_j$ 保持有界，这足以在 h 步内达到零带。

10.4 转换为数据流参数

定理*1.2*的证明。对于 $d \ge 4$，有 $d-1 \ge d/2$。网的上界给出 $d \ge$ 自然对数 $U/$自然对数$(802g)$，因此

$$S + \lceil \log_2(U+1) \rceil > \frac{\eta}{4} U^{\beta_g}, \qquad \beta_g = \frac{2401}{5000000 g^2 \ln(802g)}.$$

由于 $g \le c+2 \le 3c$,

$$\beta_g = \Omega\left(\frac{1}{c^2 \log(c+1)} \right).$$

对于固定的 c，对数时钟项与 U^{β_g} 相比可以忽略不计。令 $N = U+1$，即可在最大活跃基数方面得到相同幂次的下界，数据流长度至多为 $2N-1$，维数为 Θ_c（log N）。

对于每个充分大的指定容量 N，不只是该困难族的基数，取 $d = \lfloor \ln(N-1)/\ln(802g) \rfloor$。于是同一构造具有 $U \le N-1$，长度至多为 $2N-1$，并且为 $d = \Theta_c$(log N)。其关于 d 的指数下界因此给出 $N^{\Omega(1/(c^2 \log(c+1)))}$，对每个此类容量均成立。上面显示的数值指数 β_g 是直接进行困难族转换所得；这一填充陈述不必保留该确切的数值常数。

在这些实例中，维数为 $O_c(\log N)$，坐标具有 $O_c(\log\log N)$ 比特，数据流包含 $O(N)$ 次更新。因此，一个标准字具有 $O_c(\log N)$ 比特。将信息下界除以该字长，并将多对数因子吸收到固定的c指数中，即可证明定理 1.2。

□

对 $[D, cD]$ 中的一个估计除以 c 后，可得到此处采用的约定。对于对称约定 $[D/c, cD]$，除以 c 后会产生单侧因子 c^2。在该约定下，每个固定因子都需要多项式空间的结论同样成立。

人工智能使用声明

这项工作源于作者与 OpenAI Astra 之间的广泛交流。作者通过反复讨论提出建议，并对 Astra 的回复进行严格审查。作者还通过提供详细提示，使用 OpenAI Astra 撰写和修改稿件，部分内容由 Gemini Argon 协助完成。整个过程由作者主导，提出的想法和论证也由作者评估。形式化主要定理的 Lean 文件可在 GitHub 上获取。

引用

[AS15] Pankaj K. Agarwal 和 R. Sharathkumar。高维范围问题的流式算法。Algorithmica，72(1)：83–98，2015。初步版本发表于 SODA 2010。[BCH+26] Lotte Blank、Sergio Cabello、Mohammad Taghi Hajiaghayi、Robert Krauthgamer、Sepi- deh Mahabadi、André Nusser、Jeff M. Phillips 和 Jonas Sauer。过期流式模型：直径、 k-中心、计数、抽样及相关问题。载于第 53 届自动机、语言与编程国际研讨会（ICALP），莱布尼茨国际信息学会议论文集第 374 卷，第 37:1–37:24 页，2026年。[BGG+25] Kiarash Banihashem、Jeff Giliberti、S amira Goudarzi、MohammadTaghi Hajiaghayi、Peyman Jabbarzade 和 Morteza Monemizadeh。针对自适应对手及更广泛情形的高维动态直径。载于神经信息处理系统进展（NeurIPS），第 38 卷，2025年。完整版本：arXiv:2511.01065。[BKL13] Anna Blasiak、Robert Kleinberg 和 Eyal Lubetzky。带旁信息的广播：广播率的界定与近似。IEEE信息论汇刊，59(9)：5811–5823，2013。[CFJ+23] Artur Czumaj、Arnold Filtser、Shaofeng H.-C. Jiang、Robert Krauthgamer、Pavel Veselý 和 Mingwei Yang。通过几何哈希实现高维流式设施选址，2023年。arXiv:2204.02095v5。初步版本发表于 FOCS 2022。[CP14] Timothy M. Chan 和 Vinayak Pathak。高维最小包围球的流式算法与动态算法。计算几何，47：240–247，2014。初步版本发表于 WADS 2011。[CT06] Thomas M. Cover 和 Joy A. Thomas。信息论基础。John Wiley & Sons，第二版，2006年。[dBBM23] Mark de Berg、Leyla Biabani 和 Morteza Monemizadeh。MPC 与流式模型中的带离群点的 k-中心聚类。载于第 37 届 IEEE国际并行与分布式处理研讨会（IPDPS），第 853–863 页，2023年。[HMV25] Magnús M. Halldórsson、Nicolaos Matsakis 和 Pavel Veselý。高维点的流式直径。载于第 33 届欧洲算法年会（ESA），莱布尼茨国际信息学会议论文集第 351 卷，第 58:1–58:10 页，2025年。[HW87] David Haussler 和 Emo Welzl。 ε-网与单纯形范围查询。《离散与计算几何》，2：127–151，1987。[Ind03] Piotr Indyk。通过非对称嵌入改进高维邻近问题的算法。载于第十四届 ACM-SIAM离散算法研讨会（SODA）论文集，第 539–545 页，2003年。[JKS24] Shaofeng H.-C. Jiang、Robert Krauthgamer 和 Shay Sapir。用于 k-中心聚类的中等程度降维。载于第 40 届计算几何国际研讨会（SoCG），莱布尼茨国际信息学会议论文集第 293 卷，第 64:1–64:16 页，2024年。[JLY26] Cheng Jiang、Yinchen Liu 和 Huacheng Yu。对于多项式长度数据流，转盘流式算法或许（仍然）等价于线性草图，2026年。arXiv:2604.22052v2。

[KP20] John Kallaugher 和 Eric Price。转盘流式处理与线性草图之间的分离与等价性。载于 第*52*届*ACM SIGACT*计算理论研讨会（*STOC*）论文集，第1223–1236页，2020年。[KPSW25] Sanjeev Khanna、Ashwin Padaki、Krish Singal 和 Erik Waingarten。动态流中直径估计的多项式空间下界。载于 第*66*届*IEEE*计算机科学基础年会（*FOCS*），第119–148页，2025年。完整版见 arXiv:2510.04918。[LNW14] Yi Li、Huy L. Nguyen 和 David P. Woodruff。转盘流式算法不妨就是线性草图。载于第*46*届*ACM*计算理论研讨会（*STOC*）论文集，第174–183页，2014年。[PSW26] Ashwin Padaki、Krish Singal 和 Erik Waingarten。动态流中欧几里得直径估计的近最优空间下界，2026年。 arXiv:2609.19089。[Sio58] Maurice Sion。关于一般极小极大定理。《太平洋数学杂志》，8(1):171– 176，1958 年。[SSS95] Jeanette P. Schmidt 、 Alan Siegel 和 Aravind Srinivasan。适用于有限独立性的 Chernoff–Hoeffding界。 《*SIAM*离散数学杂志》，8(2):223–250，1995年。

A 附录

A.1 选择因子二的常数

现在我们指定所有近似参数和指数参数。定理 1.1的证明只需要严格不等式。这里我们验证一种有理数选择。所得资源界为

$$\begin{array}{|rl|}\hline \text{Space:} & N^{1-10^{-10}}\,\mathrm{poly}(\mathcal{L}), \\ \text{Worst-case update:} & \mathrm{poly}(\mathcal{L}). \\ \hline\end{array}$$

该算法维护一个缓存分数 L。每个非零得分都有一对当前输入点作为见证点对；一次不成功的副本可能报告零。以至少 $1-\delta$ 的概率，在每次更新之后同时满足：

$$L \le \mathrm{diam}(P) < 2L$$

只要直径为正；直径为零的情况可以精确处理。资源界和下界不等式在每个随机结果上都成立。下面的所有常数都是固定的，与N、 d、 Δ、 T无关。表 3 列出了决定近似和空间节省的不等式。

A.1.1 几何常数与恢复常数

我们使用以下凸包参数、半径尺度比和目标比率

$$a = \frac{1581}{1000}, \qquad q_{\mathrm{scale}} = \frac{5001}{5000}, \qquad \lambda_0 = \frac{7903}{5000}, \qquad \lambda = \frac{79}{50}.$$

它们满足 $a/q_{\mathrm{scale}} > \lambda_0 > \lambda > \sqrt{2}$ 且

$$2a^2 - 1 = 3.999122 < 4.$$

所需不等式	经验证的取值
凸包因子小于二	$2a^2-1=3.999122<4$
混合对不劣于凸包	$(1+\eta_{\mathrm{in}})^2+a^2\kappa^2=3.9898284241<3.999122$
候选因子不劣于凸包	$(1+\eta_{\mathrm{in}})^2+1=2.0201<3.999122$
预言机准确率	$\kappa(1-\nu-2e_0)\ =1.084441>1.0801$
桶指数	$\theta(1+G)$ 自然对数 $C_g<\rho$
预言机每次更新的工作量	$s(B_\kappa+1)-1=-2/2003<0$
恢复性节省	$\epsilon_0>1/6{,}345{,}504{,}000>10^{-10}$

表3：因子二参数检查。所有参数均独立于 N，d，Δ，T 固定。该指数有意取保守值。

掩蔽松弛量为 $1/20$，半径为 $r_1=1$，$r_2=4/5$，势函数为

$$\phi(t,q)=q\exp\left(\frac{(\lambda-t)^2}{q}-1\right),\qquad t^2+q\le 1.$$

下面的双帽计算表明，第一个掩码 $\phi\le 21/20$ 位于以 $(3/5)$ 乘以目标方向为中心、半径为 $4/5$ 的球内，并且第二个掩码 $\phi\le(21/20)(4/5)^2$ 为空。事实上，$F(3/5)>79/75>21/20$，而全局势函数下界为 $\phi>16133/24000>84/125=(21/20)(4/5)^2$。

对于两个相互独立的高斯网格阶段，分别选择

$$\xi=10^{-6},\qquad \delta_g=10^{-4},\qquad \beta=\frac{1023}{1000},\qquad C_g=42300,\qquad q_0=\frac{1999}{2000},\qquad \tau=-\ln q_0.$$

带符号的目标到主体投影间隙至少为 $\lambda-t$：

$$\sqrt{1-\xi}(\lambda_0-t)-\delta_g r_i\ge(1-10^{-6})(\lambda_0+1)-1-\delta_g-t>\lambda-t.$$

核心单元界为 $C_g^{k_i}$，因为 $4.132<\sqrt{2\pi e}<4.1\overline{33}$，且 $4。133\cdot 10231=42284.723<C_g$。阶段掩蔽集之外的点，其同单元概率至多为 $q_0^{k_i}/C_g^{k_i}$，因为

$$\frac{C_g\delta_g}{\sqrt{2\pi e}\sqrt{21/20}}<\frac{2115/2066}{\sqrt{21/20}}<\frac{10575000}{10584118}<q_0.$$

对于条件投影核心尾部，$x=\beta^2/(1+\xi)\ -1$ 位于 $(0。0465，0。047)$ 内，且其 Chernoff 速率满足

$$J\ge\frac{x^2}{4(1+x)}>\frac{0.0465^2}{4(1.047)}>\frac{1}{1950}>\frac{1}{1999}>\tau.$$

因此，首个安全阶段论证中的两项都以有理数速率 $q_0^{k_i}$ 衰减。

A.1.2 两个掩码的验证

与层级证明中一样，将原单位球中的一个主体点写作 $tu+v$，其中 $q=\|v\|^2$ 和$t^2+q\le 1$，并令

$$\phi(t,q)=q\exp\left(\frac{(\lambda-t)^2}{q}-1\right).$$

令 $\eta_{\mathrm{mask}} = 1/20$。第一个掩码为 $\phi \le 21/20$。其可能的纵向坐标由以下条件确定：

$$F(t) = (1-t^2)\exp\left(\frac{(\lambda - t)^2}{1-t^2} - 1\right).$$

微分得到

$$(\log F)'(t) = \frac{2g(t)}{(1-t^2)^2}, \qquad g(t) = t\lambda^2 - \lambda t^2 + t^3 - \lambda,$$

$$g'(t) = (\lambda - t)^2 + 2t^2 > 0, \qquad g(3/5) = -0.43496 < 0.$$

因此，F 在（-1，$3/5$]内始终递减。在其右端点处，

$$F(3/5) = \frac{16}{25}\exp(0.500625) > \frac{16}{25}\left(1 + \frac{1}{2} + \frac{1}{8} + \frac{1}{48}\right) = \frac{79}{75} > \frac{21}{20}.$$

因此，每个第一个掩码中的点都有 $t > 3/5$。整个球冠 $t \ge 3/5$, $t^2 + q \le 1$ 都位于$B((3/5)u,\ 4/5)$内。这是第二层的仅分析球。

第二个掩码为空。事实上，

$$(\lambda - t)^2 - q \ge m := \lambda^2/2 - 1 = \frac{1241}{5000},$$

因此

$$\phi(t, q) \ge qe^{m/q} \ge em.$$

最后一个不等式是通过对 $qe^{m/q}$ 在 $q > 0$ 上取最小值得出的，其最小值在 $q = m$ 处取得。由于 $e > 1 + 1 + 1/2 + 1/6 + 1/24 = 65/24$,

$$em > \frac{16133}{24000} > \frac{84}{125} = (1 + \eta_{\mathrm{mask}})\, r_2^2.$$

这证明了带有明确松弛量的两阶段终止条件。

A.1.3 将较早的密度与实际较晚的核半径耦合

对于一个归一化原始球点 $z = tu + v$，令 u 为目标方向，$q = \|v\|^2$、$t^2 + q \le 1$，并写作 $\phi = \phi(t, q)$。令

$$D_0 = C_g\delta_g/\sqrt{2\pi e} < 1.024, \qquad \sigma(\phi) = \max\{0, \ln(D_0/\sqrt{\phi})\}.$$

阶段 i 的条件同单元概率至多为 $C_g^{-k_i}(D_0 r_i/\sqrt{\phi})^{k_i}$。全局势函数下界 $\phi > 84/125$，结合 $D_0 < 128/125$ 和 $(128/125)^2 < (5/4)^2(84/125)$，得到$\sigma(\phi) < $ 自然对数$(5/4) < 9/40$。我们令 $A_0 = 200$，并令 B_i 为不小于 $A_0 C_g^{k_i}$的最小 2 的幂。给定 $A_1 u$ 处于其指定范数范围内，第一阶段通过概率至多为

$$\frac{C}{B_1} e^{\sigma(\phi)k_1}.$$

事实上，不同单元以概率 $1/B_1$ 通过，而共享单元的贡献至多为 $C_g^{-k_1} e^{\sigma(\phi)k_1}$。因此，均匀分布地有 $C = 401$。该指数将与点到第二核心中心的实际距离所对应的尾概率配对。

势函数的两个更强下界。对于任意实数 w，不等式 $e^x \ge 1 + x$蕴含

$$qe^{(\lambda-t)^2/q-1} \ge e^w\big((\lambda - t)^2 - wq\big).$$

Ta取 $w = 7/8$，使用 $q \le 1 - t^2$，并将所得的 t 二次式最小化，可得 s

$$\phi \ge e^{7/8}\frac{7}{8}\left(\frac{\lambda^2}{15/8} - 1\right) > \frac{119}{50}\frac{21749}{75000} = \frac{2588131}{3750000} > \frac{69}{100}.$$

为完整起见， $e^{7/8} > 119/50$ 可通过保留其泰勒级数的前五个非负项得到。

此外， $q \mapsto \phi\ (t,\ q)$在 $0 < q \le 1 - t^2$ 上递减，因为 $(\lambda - t)^2 - q \ge \lambda^2/2 - 1 > 0$。恒等式

$$(\lambda - t)^2 - (\lambda^2 - 1)(1 - t^2) = (\lambda t - 1)^2$$

因此可得

$$\phi \ge e^{\lambda^2-2}(1 - t^2) > \frac{41}{25}(1 - t^2). \tag{29}$$

最后一个严格不等式例如可由 $\lambda^2 - 2 = 0.4964 > 0.495$ 和 $e^{0.495}$ 的前五个泰勒项推出，其和大于 $1.6402 > 41/25$。

距第二核心中心的实际距离。考虑第二偏离类中的一个点 G_2。第一掩码几何已经证明了 $t > 0.6$。它距第二核心中心的实际平方距离为

$$v_*^2 = (t - 0.6)^2 + q \le 1.36 - 1.2t < 0.64.$$

Put

$$f_* = (41/25)(16/25) = 1.0496, \qquad u_* = (f_* - \phi)_+.$$

当 $\phi \le f_*$时，（29）和 $t > 0$ 推出

$$v_*^2 \le h(\phi) := 1.36 - 1.2\sqrt{1 - \phi/1.64}.$$

这里 $h(f_*) = 0.64$，并且在整个 $0.69 < \phi \le f_*$ 中，

$$h'(\phi) = \frac{0.6}{1.64\sqrt{1 - \phi/1.64}} > \frac{0.6}{1.64(0.77)} > 0.47.$$

对 $\sqrt{1 - 0.69/1.64 < 0.77}$ 两边平方即可得到该界。对于 $\phi > f_*$，使用独立不等式$v_*^2 < 0.64$。两种情况都给出均匀的逐点估计

$$v_*^2 \le 0.64 - 0.47u_*. \tag{30}$$

由于 $D_0^2 < 1.048576 < f_*$，当 $\phi \ge f_*$时，先验指数消失。否则，自然对数$(1 + x) \le x$且 $\phi > 0.69$ 给出

$$\sigma(\phi) \le \tfrac{1}{2}\ln(f_*/\phi) \le \frac{f_* - \phi}{2\phi} < 0.75u_*. \tag{31}$$

使用实际半径的范数尾部速率。我们以第二阶段目标轴范数事件为条件。记 $g = A_2 u$；则 $\|g\|^2 \le (1+\xi)k_2$ 并且

$$A_2(z - 0.6u) = (t - 0.6)g + \sqrt{q}\, Z, \qquad Z \sim N(0, I_{k_2}).$$

对于 $v_* > 0$，应用第 6.2 节中对 $(z - 0.6u)/v_*$ 的计算。投影范数平方的矩母函数有一个上界，该上界由 $(1+\xi)v_*^2\chi^2_{k_2}$ 的矩母函数给出。因此，对于

$$a_0 = \frac{\beta^2 r_2^2}{1+\xi}, \qquad 0.6697 < a_0 < 0.67,$$

落在投影后的第二核心之外的概率至多为 $\exp(-J_* k_2)$，其中，对于 $v_*^2 > 0$，

$$J_* = \frac{1}{2}\left[\frac{a_0}{v_*^2} - 1 - \ln\frac{a_0}{v_*^2}\right] \ge \frac{(a_0 - v_*^2)^2}{4a_0 v_*^2} \ge \widehat{J}(u_*) := \frac{(0.0297 + 0.47u_*)^2}{1.7152}.$$

当 $v_*^2 = 0$ 时，投影点位于核心中心，且尾概率恰好为零，因此，带有 J 的相同上界对^仍然成立。下文所有舍入估计均使用后一个有限函数；不需要无界速率。

特别地，

$$\widehat{J}(u_*) \ge \frac{0.0297^2}{1.7152} > \frac{1}{1950}.$$

由初等平方不等式可得

$$(0.0297 + 0.47u_*)^2 \ge 4(0.0297)(0.47)u_*.$$

由于

$$0.75(1.7152) = 1.2864 < 24 \cdot 4(0.0297)(0.47) = 1.340064,$$

（31）蕴含关键的耦合界

$$\sigma(\phi) < 24\widehat{J}(u_*). \tag{32}$$

^当 $u_* = 0$ 时也成立，因为此时 $\sigma = 0$ 且 $J > 0$。

结合两个独立阶段。第二阶段同单元项具有单独且更强的一致速率：

$$\frac{D_0 r_2}{\sqrt{\phi}} < \frac{1.024(0.8)}{\sqrt{0.69}} < 0.99.$$

最后一个比较等价于 $0.67108864 < 0.676269$。令 $\nu_{\mathrm{cell}} = -\ln(0.99) > 1/100$。在第二核心隔离成立时，一个 G_2 点只有通过与目标单元共享单元，或作为投影尾部点且其不同单元哈希到目标桶，才能通过第二阶段。因此，在两个隔离事件都成立时，它到达目标叶节点的概率至多为

$$\frac{C}{B_1 B_2}\left[e^{\sigma(\phi)k_1 - \nu_{\mathrm{cell}}k_2} + e^{\sigma(\phi)k_1 - \widehat{J}(u_*)k_2}\right].$$

第一阶段因子给出了通过的概率上界，且不要求第一阶段隔离成立。两个阶段随机种子相互独立，因此只需对各自独立的轴范数事件进行条件化后，便可将概率相乘。此处不对隔离事件进行条件化。

我们取 $G = 32$ 和

$$k_1 = \lceil \theta \ln(N+2) \rceil, \qquad k_2 = \lceil 32\theta \ln(N+2) \rceil.$$

在第二类上均匀分布，

$$32\nu_{\mathrm{cell}}-\sigma>0.32-0.225=0.095>\tau/2,$$

$$32\widehat{J}-\sigma>8\widehat{J}>8/1950>\tau/2.$$

k_1 中的向上取整至多带来一个 $e^{\sigma}<5/4$倍的因子； k_2 中的向上取整只会改善该界。因此，在包括实际核心半径为零的情形在内，两个项分别都是带有统一常数的 $O(N^{-\tau\theta/2})/(B_1B_2)$。第一离开类仍满足其先前的 $O(N^{-\tau\theta})/(B_1B_2)$界。

对范数失效和隔离失效分别计费，采用相同的指标加权马尔可夫论证，因而得到

$$B_1B_2K\le C\left[B_1B_2(b+1)+N^{1-\epsilon_0}\right],\qquad \epsilon_0=\tau\theta/2.$$

对于完全有理的容量参数，取

$$K=\left\lceil C'\left[b+1+\frac{Nq_0^{\lfloor k_1/2\rfloor}}{B_1B_2}\right]\right\rceil \tag{33}$$

其中显式常数为 $C'=10^9$。为验证这一选择，第一阶段通过系数至多为 401，两个第二阶段系数分别至多为 400 和 1。每个指数乘积至多为 $2q_0^{\lfloor k_1/2\rfloor}$ ，其中包括向上取整以及误差不超过 1 的向上对数近似。因此，两个离开类界之和至多为 $3\cdot 401^2<10^6$ 倍 $Nq_0^{\lfloor k_1/2\rfloor}/(B_1B_2)$。采用容量系数 10^9 的马尔可夫不等式可知，失败概率至多为 10^{-3}。事实上， $q_0^{\lfloor k_1/2\rfloor}$ 与 $\exp(-\tau\theta\ln(N+2)/2)$之比至多相差一个固定因子，因此，在同一成功事件上，该容量支配所需负载，并具有所显示的总空间指数。其所有输入均为有限整数或有理常数。更明确地，对于 $t_0=\theta\ln(N+2)$和 $k_1=\lceil t_0\rceil$，差值 $|\lfloor k_1/2\rfloor-t_0/2|$ 至多为 $1/2$，从而所得因子介于 $q_0^{1/2}$ 和 $q_0^{-1/2}$ 之间。有界误差的有理数对数近似只会再引入一个固定因子。这个逐点论证取代了两个互不相关的最坏情况之积；它不会改变任何维护中的草图或其更新。

上述系数中使用的二次幂桶数允许采用附录 A.2 中的仿射二进制哈希。

A.1.4 预处理指数为 2000 的显式形式

Take

$$\kappa=\frac{109}{100},\qquad b_0=\frac{27}{25},\qquad \nu=\frac{1}{10000},\qquad e_0=\frac{1}{400}.$$

Then

$$\kappa(1-\nu-2e_0)=1.084441>1.0801=b_0+\nu,$$

因此，近似条件在固定松弛量下成立。此外，

$$I(b_0)\ge\frac{0.1664^2}{4(1.1664)}>\frac{1}{170},$$

因为 $0.02768896\cdot 170=4.7071232>4.6656$。选择

$$k=120\left\lceil\log_2(240m)\right\rceil.$$

由于 $\ln\ 2 < 0.7$，我们有 $kI(b_0)\ > (120/119)\ln(240m)$；$m$ 个查询向量的同时理想上尾失败至多为 $1/240$。

一次近查询有 $\|q - s_0\| < (100/9)D$。其舍入投影点到

$$1.0801(100/9)D + D/800 < 13D$$

As_0的距离小于 $\tilde{}$。我们使用介于 $13D$ 和 $14D$之间的有理数表半径。网格步长界由此保证每个表键都满足 $\|j\| \le 11200\sqrt{k}$。在枚举深度 i处，围绕可行整数前缀的互不相交单位立方体位于半径为 $11200\sqrt{k} + \sqrt{i}/2$的 i维球内。体积界$v_i \le (2\pi e/i)^{i/2}$ 和 $\sqrt{2\pi e} < 4.14$ 给出至多

$$(50000\sqrt{k/i})^i \le 50000^k e^{k/(2e)}$$

pre个固定项，其中用到了 $i\ln(k/i)\ \le k/e$。由于 $e > 8/3$ 且 $\ln(5/4)\ \ge 1/5 > 3/16$ ，

$$50000e^{1/(2e)} < 62500 < 65536 = 2^{16}.$$

所有前缀　因此可以在 $(k+1)2^{16k}$ 步内完成枚举，忽略多项式因子，　d

$$2^{16k} \le 2^{1920}(240m)^{1920} = Cm^{1920}.$$

每个网格点进行一次样本扫描会增加一个 m因子。精确样本直径计算以及所有坐标和键运算都符合相同的界。因此，特意向上取整的预处理指数

$$B_\kappa = 2000 \tag{34}$$

有效。其余因子关于 d、$\log\ m$、$\log\ \Delta$、b_q 和对数重复参数均为多项式。查询时间关于这些参数仍为多项式，并且每个返回的点都是 S 中的实际成员。

A.1.5 预言机与缓存得分

取候选证书精度 $\eta_{\mathrm{in}} = 1/100$。对于每个有效的活动节点中心 c （其方差为 V），查询冻结样本的静态 1.09 最远点预言机。如果它返回 $s_* \in S$，则为实际点对计分：

$$\max_{z\ \text{an anchor}} \|z - s_*\|.$$

该平方得分至少为 $V + \|c - s_*\|^2$。因此，中心查询方差论证给出的平方混合情形因子为

$$(1+\eta_{\mathrm{in}})^2 + a^2(1.09)^2 = 3.9898284241 < 3.999122.$$

凸包内点对的平方因子为 $2a^2 - 1 = 3.999122$。凸包外点对的平方因子为 $(1+\eta_{\mathrm{in}})^2 + 1 = 2.0201$。因此，凸包因子仍然最大，且严格小于二。因此，式（34）和（21）中证明的选择适用于算法 1中的每一次中心比较。

A.1.6 样本指数、截止期限与最终空间指数

我们选择有理指数和维度比

$$s = \frac{1}{2003}, \qquad \rho = \frac{s}{2}, \qquad G = 32, \qquad \theta = \frac{\rho}{12(G+1)}.$$

我们维护一个速率至多为 N^{s-1} 的活动样本，在全局间隔 $H_{\mathrm{epoch}} = \max\{1,\ \lfloor N^{1-s}/100 \rfloor\}$ 处将其冻结，并在一个区间内构建其预言机，然后在下一个区间为其提供服务。样本上限和球网定理给出 $b = N^{1-s}\,\mathrm{poly}$ 个异常点；此外，至多 $2H_{\mathrm{epoch}}$ 次介于其中的更新具有相同的数量级，并计入 b。当 $N^{1-s} < 100$ 时，适用有界-N 约定。在有限实现中，采样概率可以是介于 $N^{s-1}/4$ 和 N^{s-1} 之间的二进有理数，该概率由 N 的整数二进制长度和有理常数 s 获得。这至多使异常界扩大四倍，同时保持样本量上界和存活概率不变，也不会改变任何指数。由于 s 是有理数，因此可以通过整数幂比较计算时期长度。

Set

$$k_1 = \lceil \theta \ln(N+2) \rceil, \qquad k_2 = \lceil G\theta \ln(N+2) \rceil.$$

在进行这些向上取整之前，可以使用与对数相差不超过一的有理上近似。这只会使维数改变固定常数。桶乘积为 $O(N^{\rho})$，因为 $\theta(G+1)\ln C_g < \rho$，其中使用了 $\ln C_g < 11 < 12$。刚刚证明的逐点耦合带来了指数节省

$$\epsilon_0 = \frac{\tau\theta}{2} > \frac{1}{96000 \cdot 2003 \cdot 33} = \frac{1}{6345504000} > 10^{-10}.$$

其余空间指数为

$$sB_{\kappa} = \frac{2000}{2003}, \qquad 1 - s + \rho = 1 - \frac{1}{4006}.$$

这三个值都严格小于 $1 - 10^{-10}$。预言机的构造可纳入一个全局时期，并且每次原始更新的工作量为多项式，因为

$$s(B_{\kappa} + 1) - 1 = -\frac{2}{2003} < 0.$$

将这些指数代入(3)和每次更新界(10)，即可证明明确的二倍因子结论。所有重复、局部重构和有限精度开销都通过多项式因子体现。

A.2 有限精度实现

高斯论证可以用预先指定的有限个随机比特和最坏情况多项式级的算术运算来实现。对于恢复网格，我们将实现与理想连续构造耦合，使得数据流中的每个点都具有完全相同的单元标识符。相比之下，静态最远点预言机只需要对投影距离的扰动有一个一致的界。它的查找单元无需与理想查找表一致。

在本节中，数据流独立于算法的随机种子而固定。这与上界定理中的无知流假设相同。对所得有理坐标、网格索引和原始平方距离执行的所有算术运算都是精确的。

A.2.1 预算与需要近似的量的数量

回顾一下， $|p_j| \le \Delta$ 对每个输入坐标都成立。现在令 $\ell_{\min} > 0$ 为每个恢复网格边长的一个已知下界。构造中的网格具有有理数边长

$$\ell_i = \delta_g \widetilde{r}_i w,$$

其中， w 属于指定的有限有理数尺度列表。尺度参数的编码以及 $\log(1/\ell_{\min})$ 的大小均由输入参数多项式有界。

设 K 是标量网格行总数的上界；若某行在不同尺度或使用不同平移量时被使用，则重复计数。即使某个标识被删除后重新插入，在包含 T 次更新的数据流中，出现的点标识也至多有 T 个。因此

$$E = \max\{1, TK\}$$

这是需要保留的标量点/网格成员测试数量的一个充分上界。可以对每个输入标识和每个维护的网格进行并集界定，无论该特定组合是否曾被评估。

设 $R \ge 1$ 是整个计算过程中生成的独立标量高斯条目数量的上界，其中包括每次冻结样本生成和每次重复中的静态预言机映射。该上界可以预先确定。恢复矩阵由其所在阶段的所有叶节点共享；不会针对每个叶节点分别生成。恢复矩阵的数量关于对数参数是多项式级的。每次静态样本生成都会使用多项式数量的映射，每个映射含有 $d\,O(\log\ N)$ 个条目；每个副本至多有 $T+1$ 代。枚举其大型查找表不会额外抽取高斯条目。因此，一个足够宽松的上界为

$$R \le (N + T + d + 2)^C \big(1 + \log\Delta + \log(1/\delta)\big)^C$$

对于一个固定常数 C，它仅取决于固定的近似参数。相关事实是，$\log R$ 关于允许的维数和对数参数是多项式的。采用更大的上界只会使精度按对数增长。

高斯映射可以按其副本、尺度、阶段和样本代进行索引，并预留允许的最大行数，忽略未使用的行。这样便提供了一个至多含 R 个条目的固定族以供耦合。结构更新计划从不取决于静态预言机返回的答案。或者，也可以按顺序抽取新条目，并使用相同的确定性数量上界；下文的条件并集界保持不变。

A.2.2 以有界工作量生成舍入高斯变量

固定条目误差 $h = 2^{-b}$、 $b \ge 1$，以及所需的失败预算 $\varepsilon_{\mathrm{cpl}} \in (0,\ 1)$。Pu t

$$L = \lceil \log_2(8R/\varepsilon_{\mathrm{cpl}}) \rceil, \qquad M = \left\lceil \sqrt{2L} \right\rceil.$$

只需进行整数比 较即可得到这些上界。对于一个标准高斯变量 G ，

$$\Pr[|G| > M] \le 2e^{-M^2/2} \le 2\,2^{-L} \le \varepsilon_{\mathrm{cpl}}/(4R).$$

因此，所有 R 理想高斯条目以至多 $\varepsilon_{\mathrm{cpl}}/4$ 的概率落在 $[-M, M]$ 中 .

下面给出一种与一个理想条目耦合的有限构造。令 U 在 $(0,\ 1)$ 上均匀分布，并设$G = \Phi^{-1}($ $U)$，其中 Φ 是标准高斯分布函数。该实现仅读取前

$$r = b + M^2 + 8$$

U 位，得到一个均匀整数 $J \in \{0, \dots, 2^r - 1\}$，并使用 $u_J = (J + 1/2)2^{-r}$。这只是一个耦合描述：实际算法直接抽取 r 独立比特。

截断逆函数

$$g_M(u) = \max\{-M, \min\{M, \Phi^{-1}(u)\}\}$$

是全局利普希茨连续的，常数至多为

$$\sqrt{2\pi} e^{M^2/2} \le 3\, 2^{M^2}.$$

由此可得

$$|g_M(U) - g_M(u_J)| \le 3\, 2^{M^2 - r - 1} < h/8.$$

我们通过下文所述的有界二分法计算 $g_M(u_J)$，使其加性误差至多为 $h/4$，并将~结果舍入为网格间距至多为 $h/4$的二进有理数。所得有理数 G 满足

$$|\widetilde{G} - g_M(U)| < h, \qquad |\widetilde{G}| \le M + h,$$

~对每个结果都成立。特别地，当 $|G| \le M$ 时，$|G - G| < h$。每个条目恰好使用 r 个随机比特，且其算术运算量有一个指定的多项式上界。不存在拒绝循环。

经过认证的 **逆累积分布函数计算**。在 $[-M, M]$ 上，高斯密度至少为 st

$$m_0 = \tfrac{1}{3}\, 2^{-M^2}.$$

在二分中点 x 处，以至多 $hm_0/100$ 的绝对误差计算 $\Phi(x)$。如果所得区间完全位于 u_J 下方或上方，就保留相应的半区间。如果该区间包含 u_J，则根据密度下界，中点距截断逆函数不超过 $h/50$；返回该中点。如果没有提前终止，至多经过 $\lceil \log_2(32(M+1)/h) \rceil$ 次迭代，剩余的夹逼区间长度就会小于 $h/8$。当未截断的逆函数位于 $[-M, M]$ 之外时，此论证同样适用：二分法会收敛至相应的端点。迭代次数有一个与所有随机结果无关的上界。

为完整起见 ，计算 Φ 所需的求值也只需多项式时间。对于有理数 $x \in [-M, M$] use

$$\Phi(x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2}\, dt$$

并对泰勒多项式积分

$$\sum_{j=0}^{J_0} \frac{(-1)^j t^{2j}}{2^j\, j!}$$

逐项计算。逐点绝对余项至多为 $(M^2/2)^{J_0+1}/(J_0+1)!$。取 $J_0 + 1 \ge 3M^2$ 并令 $J_0 + 1 \ge \log_2(8($ $M + 1)/\epsilon)$，即可在使用 $j! \ge (j/e)^j$ 和$e < 3$的情况下，使积分误差低于 $\epsilon/8$。因此，所需项数为 $O(M^2 + \log(1/\epsilon) + \log(M + 1))$。精确有理中间值具有多项式位长，尽管可能发生消去。常数 $1/\sqrt{2\pi}$ 可以在多项式时间内界定到所需的附加精度：例如，在 $\pi = 16 \arctan(1/5) - 4\arctan($ $1/239)$ 中使用几何收敛的反正切级数，然后

使用有理二分法求平方根。这些公式给出了明确的可验证误差界和固定的多项式运算次数。

归一化映射 $A = G/\sqrt{k}$ 的处理方式相同。将每个 G 近似到足够精细的程度，通过有理二分法近似 $1/\sqrt{k}$ ，然后进行乘法，并为每项操作各预留一半的目标条目误差。这只会增加 $O(\log k + \log(M+1))$ 个精度比特。在任一归一化方式下，实现都能在共同的高斯截断事件上保证规定的逐项误差 h 。

A.2.3 保留所有恢复网格标识符

将标量理想单元标识符写为

$$\left\lfloor \frac{a \cdot p}{\ell} + U_0 \right\rfloor,$$

其中 a 是一个标准高斯行， U_0 是 $[0,1$中的独立均匀随机数）。它与边长为 ℓ的通常均匀平移网格等价。将实现的平移量 U_0 与 U_0的前 b_s 个比特关联，并向下取整。于是 $|\widetilde{U}_0 - U_0| \le 2^{-b_s}$ 确定地成立，而实现的实际平移量 ℓU_0 是有理数。

Set

$$\eta = \frac{\varepsilon_{\mathrm{cpl}}}{8E}, \qquad h \le \frac{\eta \ell_{\min}}{2d\Delta}, \qquad 2^{-b_s} \le \eta/2.$$

逐项网格间距可以向下舍入到下一个二进有理数值。在共同的高斯截断事件上，对于每个输入点和每个网格行，

$$\left| \frac{\widetilde{a} \cdot p}{\ell} + \widetilde{U}_0 - \frac{a \cdot p}{\ell} - U_0 \right| \le \frac{d\Delta h}{\ell_{\min}} + 2^{-b_s} \le \eta.$$

对于一个固定的输入点，在给定理想高斯矩阵的条件下， $a \cdot\ p/\ell + U_0$的小数部分服从均匀分布。其与某个整数的距离不超过 η 的概率至多为 2η。对至多 E 个标量测试应用并集界，可得任意相关理想坐标过于接近边界的概率至多为 $2E\eta = \varepsilon_{\mathrm{cpl}}/4$ 。这些边界事件之间无需相互独立。在其补集上，每个向下取整的结果，以及由此得到的每个完整网格键和外层桶标识符，都完全不变。

结合边界界和高斯截断界，以至少 $1 - \varepsilon_{\mathrm{cpl}}$ 的概率，数据流中的所有网格成员关系都与理想的连续构造一致。重复的插入和删除使用相同的存储有限精度矩阵和平移量，因此，无论结果如何（包括耦合失败的结果），一个点的键始终是其身份的固定函数。因此，每个桶内的严格转门合法性都得到保持，且无需附加任何概率条件。

正是由于使用了均匀平移量，因此无需对点的坐标或某个特定的高斯点积作出反集中假设。固定的预先确定的输入序列对于所展示的边界并集界至关重要；这并不意味着对于在检查网格种子后才选择的点，也能得到相同的保证。

A.2.4 静态预言机中的有理中心查询

静态预言机使用其实现的有理数矩阵 A构建并接受查询。其距离证明不要求查询的查找单元与理想矩阵下的查找单元一致。只需将理想范数上下界事件转移到 A，并使用预留的余量 ν。

令 $k_{\max}$ 表示归一化预言机映射中的最大行数。在共同截断事件上，我们使用逐项误差 $h \le \nu/\sqrt{k_{\max} d}$ 。对于一个 $k \times d$ 映射以及任意 $v \in \mathbb{R}^d$，

$$\|(\widetilde{A} - A)v\|_2 \le \|\widetilde{A} - A\|_{\mathrm{F}} \, \|v\|_2 \le \sqrt{kd}\, h \, \|v\|_2 \le \nu \|v\|_2.$$

这对 $v=0$也成立。因此，任何理想上界 $b_0\|v\|$ 都可迁移至 $(b_0+\nu)\|v\|$，理想非收缩性也可迁移至 $(1-\nu)\|v\|$。扰动估计对所有向量都一致成立。对于与理想映射的种子相互独立的查询，理想映射的概率范数界同样适用，如引理 9.1所述。证书中心是 $[-\Delta,\Delta]^d$中的输入点的有理凸组合。其多项式位数界 b_q 控制投影查询、网格舍入和精确距离比较的成本。

表格网格步长及其外接半径可以通过整数平方比较或有理数平方根二分法，在静态预言机引理所规定的区间内有理地选取。这些区间中的严格常数松弛量使其成为一项多项式位数任务。所有表格条目都是原始样本标识，最终分数是精确的原始距离平方，因此投影舍入不会产生伪见证或高估当前直径。

A.2.5 资源与概率核算

对于共同精度，取满足两个条目误差限制的最大二进有理数 $h\le 1/2$ （在因子二以内取更小的二进有理数也足够）。所需的比特数至多为

$$\begin{aligned}O\big(&\log(E+2)+\log(R+2)+\log(1/\varepsilon_{\mathrm{cpl}})\\&+\log(d\Delta+2)+\log(k_{\max}+2)+\log(1+1/\ell_{\min})+\log(1/\nu)\big).\end{aligned}$$

平移精度、高斯截断长度、每个条目所需的随机位数以及算术精度，均关于相同参数呈多项式关系。在直径构造中，它们都关于 d、输入坐标位数界、$\log N$ 和 $\log(T/\delta)$ 呈多项式关系，其中常数仅取决于所选的近似因子。精确的网格索引也具有多项式位长，因为

$$\left\|\left\lfloor \widetilde{a}\cdot p/\ell+\widetilde{U}_0\right\rfloor\right\|\le 2+d\Delta(M+1)/\ell_{\min}.$$

对于外层的两两独立仿射哈希，它们可以无碰撞地编码为有界二进制字符串。

更明确地说，选择一个 Q_0 大于所显示上界的整数，并使用定长有符号整数编码对$[-Q_0,Q_0]^k$中的每个网格键进行编码。从该有限比特域到桶位的仿射二进制线性映射是两两独立的：对于两个不同的键，它们的差非零，因此每个独立采样的矩阵行及其独立的仿射偏移量都会产生两个相互独立且均匀分布的输出位。对于完整格上的理想比较，通过为每个域外单元赋予一个独立且均匀分布的桶标签，将这一随机哈希扩展到完整格。该扩展族是两两独立的，因此理想隔离和碰撞估计适用于每个网格。实际算法仅生成有限仿射映射。在耦合事件上，所有流式单元键都属于实现的域，并且在两个过程中保持一致。在每个结果下，截断实现都具有有界的键。

生成每个有限矩阵时，每个条目只需固定多项式量的工作。初始恢复矩阵在允许的预处理中生成；冻结的静态预言机映射则在已计入预算的后台构造任务中生成。这不会将N 的正幂引入原始更新，也不会改变任何空间指数。这里不假设实数运算具有常数时间，也不假设存在无限的随机带。

例如，将 $\varepsilon_{\mathrm{cpl}}=\delta/10$ 分配给整个有限精度耦合，并在理想概率引理之间分配剩余的误差预算。在理想实验中应用独立副本放大，然后再计入耦合失败的代价。在耦合事件上，所有原始点恢复成员关系都与连续情形证明中分析的一致，并且每个静态

范数事件都会连同其预留的松弛量一起转移。通过并集界，可以对所有 T 更新得到所需的同时保证。无论该耦合是否成功，资源界、固定的桶成员关系、精确的稀疏恢复认证，以及每个报告的非零当前点对距离的有效性，都会在每个随机结果上成立。

A.3 高斯嵌入基线

作为比较，考虑形式为 $N^{\alpha}\,\mathrm{poly}(d,\log N)$的更新界，其中 $\alpha<1$固定。若空间和更新时间都允许达到该指数，则在指数 $2/3$ 下存在一个简单得多的基线。该基线基于 Indyk 的高斯嵌入，并采用 Padaki 、 Singal 和 Waingarten 提出的动态流式表述 [PSW26]。下文的实现使用本文中经认证的稀疏恢复原语，明确给出了更新和报告的开销。

定理 A.1 (三分之二次幂基线**)。** 在相同的有界整数、无知、严格转储模型中，存在一个单遍二倍直径估计器，满足

$$\begin{aligned} \mathit{space} &= N^{2/3}\,\mathrm{poly}(d,\log(N+2),\log(\Delta+2),\log((T+2)/\delta)),\\ \mathit{worst\text{-}case\ update\ time} &= N^{2/3}\,\mathrm{poly}(d,\log(N+2),\log(\Delta+2),\log((T+2)/\delta)). \end{aligned}$$

每次更新后，它都会缓存一个标量答案。以至少 $1-\delta$ 的概率，所有答案都满足$D_t/2\le L_t\le D_t$。这两个近似不等式均以该成功事件为条件。

A.3.1 高斯最大嵌入

Put

$$q=\lceil\log_2(N+2)\rceil,\qquad \gamma=\frac{1}{100q},\qquad a=\frac{2}{1+8\gamma},\qquad k=\left\lceil N^{2/(a^2-1)}q^{16}\right\rceil.$$

展开式 $2/(a^2-1)=2/3+O(1/q)$ 表明 $k=N^{2/3}\,\mathrm{poly}(\log(N+2))$。令 G 具有 $\mathbb{R}^d$中的 k个相互独立的标准高斯行，并令

$$\rho=\frac{1}{a}\sqrt{2\ln(40kN^2)}.$$

对于大小至多为 N 的任意固定点集 P，其概率至少为 0.9，

$$\rho\|x-y\|\le\|Gx-Gy\|_\infty\le a\rho\|x-y\|\qquad(x,y\in P). \tag{35}$$

为完整起见，对于固定的非零差值，规范化的行内积是相互独立的标准高斯变量。利用并集界和 $\Pr(|Z|\ge t)\le 2e^{-t^2/2}$，可得失败概率至多为 $1/(20N^2)$。当 $\rho\ge 2$ 时，高斯尾部下界至少为 $C\rho^{-1}e^{-\rho^2/2}$。因此，超过 ρ 的坐标的期望数量至少为

$$C\rho^{-1}k^{1-1/a^2}N^{-2/a^2}=\Omega(q^{16(1-1/a^2)-1/2}).$$

由于 $a>1.9$，该指数始终大于 1。对于所有充分大的 N，它都超过 $\ln(20N^2)$，因此该差值对应的下侧失败概率至多为 $1/(20N^2)$。对所有点对应用并集界。对于有界的 N，可通过存储活动集来处理，其中固定常数被吸收到界中。

我们使用附录 A.2中的有限比特高斯构造，将 G 替换为一个条目有界的二进矩阵 A，其中

$$|A_{ij}-G_{ij}|\le h,\qquad 2d\Delta h\le\gamma\rho.$$

对于一个完整矩阵，其耦合失败概率可以降至 $1/100$ 以下。任意两个不同的整数点之间的距离至少为一。因此，在嵌入和舍入事件发生时，

$$(1-\gamma)\rho\|x-y\| \le \|Ax-Ay\|_\infty \le (a+\gamma)\rho\|x-y\|. \tag{36}$$

条目绝对精度和截断界在所显示参数下具有多项式位长。不使用针对数据流的并集界来增大 k；放大将置于此嵌入之外。

A.3.2 维护单个投影坐标范围

固定 A 的一行，并写作 $z(x) = \langle A_i, x\rangle$。这些是具有公分母 2^b 的精确二进有理数。对于固定范围内的每个二进长度 ℓ，维护区间标识符的频率向量

$$j(x) = \lfloor z(x)/\ell \rfloor.$$

一次原始更新会改变该行及该尺度的恰好一个区间计数。所有这些计数均为非负数。设 $K = \lceil 4/\gamma + 3 \rceil$，并维护第 7.1.2 节中针对区间标识符的经认证 K-稀疏恢复草图。使用 $O(\log(kK(J+1)))$ 个相互独立的哈希表，其中 $J+1$ 是长度的数量。在固定时刻，对于至多具有 K 个占用区间的每个向量，都能以足够高的概率恢复，使得所有此类行和尺度共同成功的概率至少为 0.99。这是在 A条件下成立的，因为稀疏草图种子与 A相互独立。

对于一个经过认证的，非空区间字典，令 $j_{\min}$ 和 $j_{\max}$ 为其极值标识符，且 pu t

$$U_{i,\ell} = (j_{\max} - j_{\min} + 1)\ell.$$

这始终是实际坐标范围 w_i的上界。如果解码器未能认证完整性，则忽略该尺度。此外，保持 z的精确总质量和前两阶矩；零方差恒等式可认证 $w_i = 0$，其中包括任意多个投影坐标相同的原始点。

令最小长度为不超过 $\gamma 2^{-b}/4$的最大二进有理数，最大长度为不小于投影坐标绝对值上界两倍的最小二进有理数。数量$J+1$ 关于对数和维数参数是多项式的。对于每个 $w_i > 0$，都存在一个长度

$$\gamma w_i/4 \le \ell \le \gamma w_i/2.$$

其占用区间数至多为 $w_i/\ell + 2 \le K$，且恢复出的上界满足 $U_{i,\ell} \le w_i + 2\ell \le (1+\gamma)w_i$。令 $\mathcal{C}_i$ 为针对行 i具有经过认证的区间字典的长度集合。如果 $w_i > 0$ 且$\mathcal{C}_i \neq \varnothing$，则定义

$$U_i := \min_{\ell \in \mathcal{C}_i} U_{i,\ell}.$$

当支撑集为空或精确时刻测试得到 $w_i = 0$ 时，我们使用 $U_i = 0$。若 $w_i > 0$ 且$\mathcal{C}_i = \varnothing$，则使用已知绝对投影坐标界的两倍。在同时发生的恢复事件上，只要 $w_i > 0$，上述尺度就属于 $\mathcal{C}_i$，因此在所有情况下都有 $w_i \le U_i \le (1+\gamma)w_i$。

稀疏解码器的成本为 $K\,\mathrm{poly}$，而这里的 K 本身是 $\log\ N$ 的多项式。因此，我们可以在每次更新后立即解码每一行的每个尺度，并维护 $U = \max_i U_i$。不存在延后的大型解码任务，也无需在查询时进行扫描。

A.3.3 归一化、放大与实际更新界

我们选择一个有理数 $\widetilde{\rho}$，使得 $\rho \le \widetilde{\rho} \le (1+\gamma)\rho$，并缓存

$$L = \frac{U}{(1+\gamma)(a+\gamma)\widetilde{\rho}}.$$

在前述良好事件成立时，如果 $D = \text{直径}(P) > 0$，

$$\frac{1-\gamma}{(1+\gamma)^2(a+\gamma)} D \le L \le D.$$

下方的系数大于 $1/2$。事实上，将所需的不等式乘以 $1+8\gamma$后，可化为证明下式为正：

$$9\gamma - 28\gamma^2 - 17\gamma^3 - 8\gamma^4,$$

这对 $0 < \gamma \le 1/100$ 成立。

在每个固定时刻，单个完整副本成功的概率大于 0.8。我们运行$O(\log((T+2)/\delta))$ 个相互独立的完整副本，并缓存它们的中位数。通过 Chernoff界和对预定时间范围取并集界，可得到所述的同时保证。中位数能够容忍任一方向的标量误差。算法1中取最大值则有不同的依据：每个正分数都有当前见证。

一次原始点更新会计算 k 个点积，对每个点积更新多项式数量的区间草图，只解码多项式大小的草图，并更新缓存的最大值和中位数。因此，最坏情况下，每项操作的开销均为 $k\mathrm{poly}$。存储空间也为 $k\mathrm{poly}$。计数的比特长度为 $O(\log(T+2))$，区间键、高斯近似、矩计数器和归一化值的位长均为多项式级别。整个过程中不会使用过期时间、原始点重放，也不会使用未经计费的输入副本。

所有参数选择均可通过多项式比特开销有效计算。特别地，$2/(a^2-1)$ 是有理数，其分子和分母关于 q 均为多项式；其幂和向上取整可通过整数幂比较求得。对数和平方根的有理数界足以用于 $\widetilde{\rho}$ 以及条目精度选择。

与我们的多项式更新构造相比，该基线具有优得多的内存指数，但在更新时间上付出了相同的 $N^{2/3}$ 次幂代价。这两个保证分别回答了时空权衡的不同版本。