

降维与近邻搜索

用随机表示保留几何与相似性

保持距离、区分邻近、加速搜索

随机投影压缩高维几何；对半径敏感的签名区分近点与远点；局部敏感哈希 则把碰撞概率的间隙变为搜索索引。本讲围绕两个问题展开：随机表示究竟 保留了什么？怎样把这种保证转化为具有明确空间、时间与错误概率的算法？

高级算法课程教学团队

南京大学

第五讲 · 随机算法与随机化技术

学习目标与阅读建议

本讲利用集中性为高维点构造短表示，并在大数据集中进行搜索。读完本讲，应当能够：

1. 从单个向量的尾界与并合界出发，证明 Johnson–Lindenstrauss 定理；
2. 解释高斯、随机子空间与随机符号三种构造，以及各自的归一化因子；
3. 区分近似最近邻与固定半径的近似近邻问题；
4. 分析 Hamming 签名查表结构与局部敏感哈希索引；
5. 组合不同半径的结构，并核算近似因子、失败概率和实际查询成本。

第一次阅读可先学习第 2 节的高斯 JL 证明，再通过第 4 节理解它与搜索的联系。第 5–7 节依次发展 Hamming 签名查表、局部敏感哈希，以及从近邻到最近邻的转化。每个构造都先说明目的，再展开证明和成本分析。

第 3 节的另外两种 JL 证明可以紧接高斯证明阅读，也可以留作第二遍学习。第 8 节比较几种方法并介绍研究联系。附录整理高斯分布、几何与概率工具。主要算法及其实现保证均保留在正文中。

全文中， $[n] = \{1, \dots, n\}$ ， $\ln$ 为自然对数。除非另有说明，输入均独立于算法的随机选择而预先固定，概率针对算法的随机性。需要时分别核算算术、比特与机器字运算。返回点以标识符表示。 $O_c(\cdot)$ 等记号中的常数可以依赖下标参数。

目录

学习目标与阅读建议	i
1 几何、表示与搜索	1
2 Johnson–Lindenstrauss：欧氏降维	2
3 其他欧氏变换：正交性与随机符号	4
4 从距离保持到近似搜索	6
5 Hamming 摘要与多项式空间查找表	8
6 局部敏感哈希：只搜索少数几个桶	10
7 从近邻到最近邻	14
8 把几种方法放在一起看	16
9 习题	17
A 高斯分布与欧氏几何工具	18
B 碰撞与间隙论证中的概率工具	20
文献说明	21

1 几何、表示与搜索

高维向量可以用许多数值特征描述一个对象。随之而来的是两种不同的计算代价：计算两个对象之间的距离，以及在大量候选对象中找到一个近邻。减少坐标数主要降低第一种代价；搜索索引还必须处理第二种代价。本讲分别建立这两类方法，并说明它们的联系与区别。

度量空间由集合 X 和函数 $d_X : X \times X \rightarrow \mathbb{R}_{\geq 0}$ 组成。对任意 $x, y, z \in X$ ，它满足

$$d_X(x, y) = 0 \iff x = y, \quad d_X(x, y) = d_X(y, x), \quad d_X(x, z) \leq d_X(x, y) + d_X(y, z).$$

$\mathbb{R}^d$ 上的欧氏距离是 $\|x - y\|_2$ ，其中 $\|x\|_2 = (\sum_{j=1}^d x_j^2)^{1/2}$ 。在 $\{0, 1\}^d$ 上，Hamming 距离为

$$\text{Ham}(x, y) = \sum_{j=1}^d \mathbf{1}\{x_j \neq y_j\} = \|x - y\|_1 = \|x - y\|_2^2. \quad (1)$$

最后一个等式依赖于向量为 Boolean 向量。对一般实向量，欧氏距离的平方不是度量：直线上的 0, 1, 2 满足 $4 > 1 + 1$ ，违反三角不等式。

嵌入 $\phi : X \rightarrow Y$ 改变点的表示方式。我们用明确的不等式衡量其质量。一种对称约定是

$$\alpha^{-1} d_X(x, y) \leq d_Y(\phi(x), \phi(y)) \leq \alpha d_X(x, y), \quad \alpha \geq 1.$$

另一种约定允许选择缩放因子 $a > 0$ ，写成 $ad_X \leq d_Y \leq Dad_X$ ，并称 D 为失真。前一种对称不等式保证的是 $D \leq \alpha^2$ ，未必是 $D \leq \alpha$ 。直接写出不等式，可以避免这些约定造成的歧义。

欧氏空间降维

给定固定点集 $S = \{x_1, \dots, x_n\} \subseteq \mathbb{R}^d$ 和 $0 < \varepsilon < 1$ ，构造 $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^k$ ，使任意 $x, y \in S$ 满足

$$(1 - \varepsilon) \|x - y\|_2^2 \leq \|\phi(x) - \phi(y)\|_2^2 \leq (1 + \varepsilon) \|x - y\|_2^2. \quad (2)$$

目标是使 $k \ll d$ ，并能高效计算映射。

开平方后，距离的保持因子为 $\sqrt{1 - \varepsilon}$ 与 $\sqrt{1 + \varepsilon}$ ，特别地，也满足较弱的 $1 - \varepsilon$ 与 $1 + \varepsilon$ 因子。反之，距离的 $(1 \pm \eta)$ 保证意味着平方距离的因子为 $(1 \pm \eta)^2$ 。本讲主要分析平方范数，因为它是求和形式；将映射用于搜索时，再明确换算近似比。

Johnson-Lindenstrauss 定理保持的是固定有限点集中的全部两两欧氏距离。Hamming 空间中的搜索摘要则只针对预先选定的半径 r ，区分 $\text{Ham}(x, y) \leq r$ 与 $\text{Ham}(x, y) > cr$ 。它不必保持其他尺度上的距离。最后，局部敏感哈希让近点对比远点对更容易碰撞，从而控制一次查询需要检查的候选点数。

三个问题，三种保证

距离在新表示中有多准确？摘要能否可靠地区分近点对与远点对？检索并验证一个候选点需要多少工作？三者密切相关，但任何一个问题的答案都不会自动解决另外两个。

2 Johnson-Lindenstrauss: 欧氏降维

2.1 定理与算法

定理 2.1 (Johnson-Lindenstrauss)。 设 $S \subseteq \mathbb{R}^d$ 含有 $n \geq 2$ 个点, 且 $0 < \varepsilon, \delta < 1$ 。令

$$k_0 = \left\lceil \frac{8}{\varepsilon^2} \ln \frac{n(n-1)}{\delta} \right\rceil. \quad (3)$$

若 $A \in \mathbb{R}^{k_0 \times d}$ 的各元素独立服从 $\mathcal{N}(0, 1/k_0)$, 则以至少 $1 - \delta$ 的概率, 映射 $x \mapsto Ax$ 对 S 中所有点对同时满足式 (2)。因此, 存在目标维数至多为 $\min\{d, k_0\}$ 的这类线性嵌入。

最后一个结论中, 当 $d \leq k_0$ 时使用恒等映射。若失败概率为常数, 甚至为 $1/n$, 通常将定理简写为 $k = O(\varepsilon^{-2} \log n)$ 。输出坐标数的这个上界不依赖于原始维数 [1, 4]。

高斯 Johnson-Lindenstrauss 变换

按照式 (3) 选择 k 。独立抽取 $G_{ij} \sim \mathcal{N}(0, 1)$, 令 $A = G/\sqrt{k}$ 。保存 A , 将每个输入点 x_i 替换为 Ax_i 。若 $k \geq d$, 也可直接采用恒等映射, 既无需失真, 也不增加维数。

每个新坐标都是全部原始坐标的随机线性组合。这与抽取少数原始坐标不同: 对 $S = \{0, e_1\}$, 均匀选取 k 个坐标会以 $1 - k/d$ 的概率漏掉唯一的非零坐标。在该事件上, 任何缩放都无法挽救产生的零向量。

2.2 从一个向量到全部点对

先考虑一个固定非零向量 v 。由线性性, $Ax - Ay = A(x - y)$ 。对每个不同点对, 定义

$$u_{xy} = \frac{x - y}{\|x - y\|_2}, \quad \|u_{xy}\|_2 = 1.$$

因此, 只需控制固定单位向量 u 的 $\|Au\|_2^2$ 。目标是证明

$$\mathbb{P}(|\|Au\|_2^2 - 1| > \varepsilon) \leq 2e^{-k\varepsilon^2/8}. \quad (4)$$

建立这个界后, 再对 $\binom{n}{2}$ 个向量 u_{xy} 使用并合界。不同点对的坏事件不必相互独立。重复点也不构成问题: 它们之间的零差向量被精确保持。

量词的顺序

保证是: 对每个固定的 u , 新抽取的随机 A 以高概率成功。它不是: 以高概率, A 对所有 $u \in \mathbb{R}^d$ 同时成功。若 $k < d$, 每个线性映射 A 的核中都有非零向量, 所以不可能保持整个空间中所有向量的范数。

2.3 高斯投影产生卡方变量

标准高斯变量 $Z \sim \mathcal{N}(0, 1)$ 的密度为 $(2\pi)^{-1/2}e^{-z^2/2}$, 均值为 0, 方差为 1。更一般地, $\mathcal{N}(\mu, \sigma^2)$ 是 $\mu + \sigma Z$ 的分布, 其中 $\sigma > 0$ 。第二个参数表示方差, 而非标准差。

独立高斯变量的线性组合仍为高斯变量：若 $G_1, \dots, G_d$ 为独立标准高斯变量，则

$$\sum_{j=1}^d v_j G_j \stackrel{d}{=} \|v\|_2 Z. \quad (5)$$

附录 A 用旋转不变性证明这一点。固定单位向量 u 后， G 的不同行使用独立随机性，因此

$$(Au)_i \stackrel{d}{=} \frac{Z_i}{\sqrt{k}}, \quad \|Au\|_2^2 \stackrel{d}{=} \frac{1}{k} \sum_{i=1}^k Z_i^2, \quad (6)$$

其中 Z_i 为独立标准高斯变量。它们的平方和服从自由度为 k 的卡方分布，记为 χ_k^2 。特别地，

$$\mathbb{E}\|Au\|_2^2 = 1, \quad \text{Var}(\|Au\|_2^2) = 2/k.$$

方差公式使用了 $\mathbb{E}Z^4 = 3$ 。仅靠 Chebyshev 不等式只能得到 $2/(k\varepsilon^2)$ ；对全部点对并合后，这不足以支持 $k = O(\log n)$ 。我们需要指数尾界。由于 Z_i^2 无界，不能直接套用有界变量的 Hoeffding 不等式。

2.4 卡方分布的指数尾界

引理 2.2 (高斯变量平方的指数矩)。对 $Z \sim \mathcal{N}(0, 1)$,

$$\mathbb{E}e^{\lambda Z^2} = (1 - 2\lambda)^{-1/2} \quad \text{当 } \lambda < 1/2. \quad (7)$$

若 $\lambda \geq 1/2$ ，该期望为无穷大。

证明。 当 $\lambda < 1/2$ 时，在积分

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-(1-2\lambda)z^2/2} dz$$

中代换 $y = \sqrt{1-2\lambda}z$ 。结果为 $(1-2\lambda)^{-1/2}$ 乘以标准高斯密度的积分。若 $1-2\lambda \leq 0$ ，被积函数不衰减，积分发散。□

引理 2.3 (卡方集中)。设 $Q = \sum_{i=1}^k Z_i^2 \sim \chi_k^2$ 。对 $0 < \varepsilon < 1$,

$$\mathbb{P}(Q \geq (1+\varepsilon)k) \leq e^{-\frac{k}{2}(\varepsilon - \ln(1+\varepsilon))} \leq e^{-k\varepsilon^2/8}, \quad (8)$$

$$\mathbb{P}(Q \leq (1-\varepsilon)k) \leq e^{-\frac{k}{2}(-\varepsilon - \ln(1-\varepsilon))} \leq e^{-k\varepsilon^2/4}. \quad (9)$$

证明。 由独立性及引理 2.2，当 $\lambda < 1/2$ 时， $\mathbb{E}e^{\lambda Q} = (1-2\lambda)^{-k/2}$ 。对 $0 < \lambda < 1/2$ ，指数 Markov 不等式给出

$$\mathbb{P}(Q \geq (1+\varepsilon)k) \leq \exp\left(-\lambda(1+\varepsilon)k - \frac{k}{2}\ln(1-2\lambda)\right).$$

最优参数为 $\lambda = \varepsilon/[2(1+\varepsilon)]$ ，位于允许的区间内。代入得到第一个上尾表达式。又因为

$$\varepsilon - \ln(1+\varepsilon) = \int_0^\varepsilon \frac{t}{1+t} dt \geq \int_0^\varepsilon \frac{t}{2} dt = \varepsilon^2/4,$$

上尾的简化界成立。对下尾取 $\lambda < 0$ 。当 $Q \leq (1-\varepsilon)k$ 时， $e^{\lambda Q} \geq e^{\lambda(1-\varepsilon)k}$ 。相同计算在 $\lambda = -\varepsilon/[2(1-\varepsilon)]$ 处最小，得到式 (9) 中的第一个表达式。最后，

$$-\varepsilon - \ln(1-\varepsilon) = \int_0^\varepsilon \frac{t}{1-t} dt \geq \varepsilon^2/2.$$

至此两侧尾界均已证明；下尾论证并不依赖 Q 的分布对称。□

定理 2.1 的证明. 式 (6) 与引理 2.3 推出式 (4)。对无序不同点对并合, 总失败概率至多为

$$\binom{n}{2} 2e^{-k\epsilon^2/8} = n(n-1)e^{-k\epsilon^2/8} \leq \delta,$$

其中 k 按式 (3) 取整。利用齐次性, 把单位向量的估计换回原始距离。在追求较小维数的版本中, $d \leq k_0$ 时使用恒等映射。□

2.5 计算代价与定理的保证范围

在把高斯抽样作为基本操作的实数算术模型中, 构造稠密矩阵需要 $O(kd)$ 次操作, 变换整个数据集需要 $O(nkd)$ 次操作, 变换一个新向量需要 $O(kd)$ 次操作。保存矩阵与变换后的点需要 $O(kd + nk)$ 个实数。若输入向量只有 s 个非零坐标, 累加对应矩阵列即可将变换代价降到 $O(ks)$ 次算术操作。

这些是算术操作数, 而非有限精度下的位复杂度。精确高斯样本不是有限比特串; 舍入分析需要指定输入精度模型并分配误差预算。第 3.2 节的符号构造避免了高斯抽样, 但输入数值的算术运算仍有代价。

查询点变换后, 扫描全部压缩点需要 $O(nk)$ 时间。这可以改善对 d 的依赖, 但候选点数仍为 n 。降维定理本身并不是次线性查询时间的搜索索引。第 4 节将准确说明它与近似搜索之间的联系。

3 其他欧氏变换: 正交性与随机符号

高斯证明包含两个环节: 一个固定向量的指数尾界, 以及对全部点对的并合界。下面保留第二个环节, 改变随机映射的构造。这样可以看清, 高斯构造中的哪些性质不可缺少, 哪些只是让分析更方便。

3.1 投影到均匀随机子空间

设 $1 \leq k < d$ 。在 $\mathbb{R}^d$ 中选择具有旋转不变分布的标准正交 k 元组 $q_1, \dots, q_k$, 令它们构成矩阵 Q 的各行。于是 $QQ^T = I_k$, Qx 给出 x 在这些向量张成的子空间上的正交投影坐标。定义

$$Tx = \sqrt{d/k} Qx. \quad (10)$$

投影平均保留平方长度的 k/d , 因此需要因子 $\sqrt{d/k}$ 补偿。 Q 的各行正交, 但不独立。一种构造方法是抽取 k 个独立高斯向量, 再进行 Gram-Schmidt 正交化; 它们以概率一线性无关。旋转原始高斯向量会相应旋转正交化结果, 因此产生所需的不变分布。

固定单位向量 u 。由旋转不变性, 将固定向量投影到随机子空间, 与将球面上的均匀随机点投影到前 k 个坐标方向, 具有相同的平方长度分布。附录 A 解释这一对称性。若 $G_1, \dots, G_d$ 为独立标准高斯变量, 则 $G/\|G\|_2$ 在单位球面上均匀分布。因此

$$\|Qu\|_2^2 \stackrel{d}{=} \frac{U}{U+V}, \quad U = \sum_{i=1}^k G_i^2, \quad V = \sum_{i=k+1}^d G_i^2. \quad (11)$$

这里 U 与 V 独立, 但 U 与 $U+V$ 不独立。该比值的均值为 k/d : 均匀单位向量的 d 个坐标平方可交换, 且总和恒为一。

引理 3.1 (随机子空间的尾界)。 设 $0 < \beta < d/k$ 且 $\beta \neq 1$ 。当 $\beta > 1$ 时考虑事件 $U/(U+V) \geq \beta k/d$ ；当 $\beta < 1$ 时考虑事件 $U/(U+V) \leq \beta k/d$ 。相应事件的概率至多为

$$\beta^{k/2} \left(1 + \frac{(1-\beta)k}{d-k}\right)^{(d-k)/2} \leq \exp\left(\frac{k}{2}(\ln \beta + 1 - \beta)\right). \quad (12)$$

证明。 令 $H = (d - \beta k)U - \beta kV$ 。上尾事件为 $H \geq 0$ ，下尾事件为 $H \leq 0$ 。选择

$$t = \frac{\beta - 1}{2\beta(d - \beta k)}.$$

t 与 $\beta - 1$ 同号，因此两种情况下相应坏事件都蕴含 $e^{tH} \geq 1$ 。由卡方矩母函数及独立性，

$$\mathbb{P}(\text{坏事件}) \leq \mathbb{E}e^{tH} = (1 - 2t(d - \beta k))^{-k/2} (1 + 2t\beta k)^{-(d-k)/2}.$$

这两个因子均有限：代入所选 t 后，其底数分别为 $1/\beta > 0$ 与 $(d - k)/(d - \beta k) > 0$ 。代入便得到式 (12) 中的第一个表达式。再使用对 $z > -1$ 成立的 $\ln(1+z) \leq z$ ，即得第二个表达式。□

分别取 $\beta = 1 - \varepsilon$ 与 $\beta = 1 + \varepsilon$ 。若 $(1 + \varepsilon)k \geq d$ ，上尾事件的概率为零，因为该比值几乎必然严格小于一；否则可使用引理。再利用引理 2.3 中的对数估计，得到

$$\mathbb{P}(|\|Tu\|_2^2 - 1| > \varepsilon) \leq e^{-k\varepsilon^2/4} + e^{-k\varepsilon^2/8} \leq 2e^{-k\varepsilon^2/8}. \quad (13)$$

因此，相同的维数选择与点对并合界证明了映射 (10) 的 JL 保证。若 $k = d$ ，正交矩阵本身就是等距映射，缩放因子等于一。这个基于旋转的证明对应于原始 JL 构造及 Dasgupta 与 Gupta 的初等分析 [1, 2]。

通过 Gram-Schmidt 生成标准正交组需要 $O(dk^2)$ 次算术操作，使用已存储的矩阵变换一个向量需要 $O(dk)$ 次操作。高斯独立行构造省去了正交化步骤。两种构造都不要求输入点集具有随机方向。

3.2 随机符号同样足够

球面对称性并非 JL 所必需。令 $R \in \{-1, 1\}^{k \times d}$ 的各元素为独立公平符号，置 $A = R/\sqrt{k}$ 。构造恰好使用 kd 个随机比特；矩阵向量乘法只需加减法，最后统一缩放。这是二元随机币构造的一种基本形式 [3]。

定理 3.2 (独立符号的 JL 变换)。 对上述符号矩阵、任意固定单位向量 $u \in \mathbb{R}^d$ 及 $0 < \varepsilon < 1$ ，

$$\mathbb{P}(|\|Au\|_2^2 - 1| > \varepsilon) \leq 2e^{-k\varepsilon^2/8}. \quad (14)$$

因此，定理 2.1 及其中相同的充分维数 k_0 也适用于此构造。

证明。 对一行未缩放的结果，记 $Y = \sum_j u_j \varepsilon_j$ 。利用独立性、均值为零，以及四次方展开，得到

$$\mathbb{E}Y^2 = 1, \quad \mathbb{E}Y^4 = \sum_j u_j^4 + 6 \sum_{j < \ell} u_j^2 u_\ell^2 = 3 - 2 \sum_j u_j^4 \leq 3. \quad (15)$$

此外，由 $\cosh s \leq e^{s^2/2}$ ，

$$\mathbb{E}e^{aY} = \prod_j \cosh(au_j) \leq e^{a^2/2}.$$

为控制平方的正指数矩，引入独立标准高斯变量 Z 。利用 $\mathbb{E}_Z e^{bZ} = e^{b^2/2}$ ，对 $0 \leq \lambda < 1/2$ 有

$$\mathbb{E}_Y e^{\lambda Y^2} = \mathbb{E}_Z \mathbb{E}_Y e^{\sqrt{2\lambda}ZY} \leq \mathbb{E}_Z e^{\lambda Z^2} = (1 - 2\lambda)^{-1/2}. \quad (16)$$

被积函数非负，所以可以交换取期望的顺序。各行独立，故得到与式 (8) 相同的上尾界。

对下尾，使用 $z \geq 0$ 时的 $e^{-z} \leq 1 - z + z^2/2$ 。式 (15) 给出

$$\mathbb{E} e^{-\lambda Y^2} \leq 1 - \lambda + \frac{3}{2}\lambda^2 \leq e^{-\lambda + 3\lambda^2/2} \quad (\lambda \geq 0).$$

对于各行的独立结果 $Y_1, \dots, Y_k$ ，指数 Markov 不等式于是给出

$$\mathbb{P} \left(\sum_i Y_i^2 \leq (1 - \varepsilon)k \right) \leq e^{k(-\lambda\varepsilon + 3\lambda^2/2)} \leq e^{-k\varepsilon^2/6},$$

其中取 $\lambda = \varepsilon/3$ 。与上尾界结合，再使用先前的点对并合界即可。一般而言， $\sum_i Y_i^2$ 不服从卡方分布；我们比较的是指数矩，而非断言分布相同。□

3.3 三种构造的共同点

映射	单个向量的分析	随机性与计算
$G/\sqrt{k}$	归一化的卡方和。	独立高斯元素；稠密矩阵乘法。
$\sqrt{d/k}Q$	旋转对称性与两个独立卡方和的比值。	标准正交组；生成时需要正交化。
$R/\sqrt{k}$	对符号和平方的指数矩进行比较。	独立随机比特；加法与减法。

三种构造都给出独立于数据集的随机线性映射，并使用相同的有限点集并合界。目标维数小，并不意味着矩阵稀疏或能快速作用于向量。稀疏与结构化变换研究的是后一个实现问题；第 8 节将简要介绍这些方向。

4 从距离保持到近似搜索

4.1 最近邻与固定半径问题

设 (X, dist) 为度量空间， $S = \{y_1, \dots, y_n\} \subseteq X$ 为非空数据集。对查询点 $x \in X$ ，记

$$d_*(x) = \min_{y \in S} \text{dist}(x, y).$$

精确最近邻搜索返回一个达到最小值的点。对 $c > 1$ ， c -近似最近邻是任意满足 $\text{dist}(x, y) \leq cd_*(x)$ 的 $y \in S$ 。特别地，若 $d_*(x) = 0$ ，乘性保证就要求返回精确匹配。

输出以点的标识符表示；若还需复制完整向量，应另加相应的输出长度成本。

直接扫描计算 n 个距离，代价为 $O(nT_{\text{dist}})$ ，其中 T_{dist} 表示一次距离计算的代价。对稠密欧氏向量或未打包的 Hamming 向量， $T_{\text{dist}} = O(d)$ 。在低维情形，可以用 k -d 树（递归的坐标划分）或 Voronoi 图（最近数据点相同的区域）组织搜索。高维搜索促使我们考虑另一种权衡：不再坚持精确答案，而允许近似与随机化。相应的复杂度权衡必须说明计算模型，以及 n 和 d 各自的作用。

定义 4.1 (近似近邻)。 在预处理之前固定 $c > 1$ 与 $r > 0$ 。 (c, r) -近似近邻查询的要求为

$$\begin{aligned} d_*(x) \leq r & : \text{返回某个 } y \in S, \text{ 满足 } \text{dist}(x, y) \leq cr; \\ d_*(x) > cr & : \text{返回 } \perp; \\ r < d_*(x) \leq cr & : \text{两种结果均允许。} \end{aligned}$$

$\perp$ 表示没有报告数据点。我们的实现会验证返回点，因此即使在两个半径之间的间隙区，任何非 $\perp$ 输出的距离也都不超过 cr 。

近邻与最近邻的区别很重要：近邻结构针对一个预先指定的半径建立，而最近距离本身未知。第 7 节将用明确的搜索不变量，把多个固定半径结构组合起来。

对随机数据结构，成功概率取自预处理及查询中使用的随机性，并要求对每个独立于这些随机选择的固定数据集和固定查询成立。下面的算法具有单侧错误：验证杜绝了返回不合格点的可能，但存在真正 r -近邻时仍可能错误返回 $\perp$ 。间隙区内除验证要求外，不附加其他约束。

4.2 JL 对固定欧氏查询的保证

假设线性映射 A 将固定查询点 x 到全部数据点的平方距离保持在 $1 \pm \varepsilon$ 因子内。对 n 个向量 $x - y_i$ 使用单向量尾界，可知以下维数足够：

$$k \geq \frac{8}{\varepsilon^2} \ln \frac{2n}{\delta}. \quad (17)$$

查询点不必在预处理时给出，但这一概率断言要求它独立于 A 而固定。

命题 4.2 (近似最近邻的转移)。 若 $y \in S$ 在投影后的欧氏空间中，是 Ax 相对于点集 $\{Ay_i\}$ 的 b -近似最近邻，则在距离保持事件上，

$$\|x - y\|_2 \leq b \sqrt{\frac{1+\varepsilon}{1-\varepsilon}} d_*(x). \quad (18)$$

证明。 令 y_* 为原空间中的真正最近邻。依次使用对 y 的失真下界、投影空间的近似保证，以及对 y_* 的失真上界，得到

$$\sqrt{1-\varepsilon} \|x - y\|_2 \leq \|Ax - y\|_2 \leq b \|Ax - Ay_*\|_2 \leq b \sqrt{1+\varepsilon} d_*(x).$$

除以正数 $\sqrt{1-\varepsilon}$ 即可。 □

若投影后进行精确搜索，则 $b = 1$ 。要得到目标近似比 $c > 1$ ，取 $\varepsilon \leq (c^2 - 1)/(c^2 + 1)$ 即可。只保持数据集内部的距离是不够的：查询点到数据点的差向量也必须满足估计。对任意预先固定的 Q 个查询，对 nQ 个差向量并合即可得到同时保证。该论证不适用于观察映射后任意自适应选择的查询。

4.3 超越投影后扫描的两条路线

即使有命题 4.2，投影空间中的直接扫描仍需比较 n 个候选点。下面在 Hamming 空间中建立两类搜索结构。第一类压缩到 $k = O(\log n)$ 比特，并为每个可能的压缩查询预计算答案，支付 2^k 级别的空间。第二类只存储数据实际到达的桶，再查询多个随机表；它支付的是重复哈希与候选验证的代价，而不枚举整个压缩空间。

这两类结构有意放松了几何目标，不要求保持每个距离，只需区分距离不超过 r 与距离大于 cr 的两种情形。两个阈值之间的间隙，为集中性或碰撞概率放大提供了发挥作用的余地。

5 Hamming 摘要与多项式空间查找表

固定非空数据集 $S \subseteq \{0,1\}^d$ 、半径 $r \geq 1$ 和近似比 $c > 1$ 。若 $cr \geq d$ ，任意数据点都是合格答案。下面集中讨论非平凡情形 $cr < d$ 。对于 $0 \leq r < 1$ ，真正的 r -近邻必须是精确匹配，可另行查询。

目标是把查询压缩成较短的 Boolean 签名，用来区分近数据点与远数据点。压缩后的 Hamming 距离不是原始距离的等比例缩放。每个输出比特提供一次独立测试，而测试的期望在距离从 r 增长到 cr 时发生明确的变化。

5.1 随机子集的奇偶性

选择 $k \times d$ 矩阵 A ，其元素独立服从 $\text{Bern}(p)$ ，其中 $0 < p < 1/2$ 。本节所有矩阵向量乘法都在 $\mathbb{F}_2$ 上进行：

$$(Ax)_i = \left(\sum_{j=1}^d A_{ij}x_j \right) \bmod 2. \quad (19)$$

对固定的 x, y ，令 $t = \text{Ham}(x, y)$ 。在 $\mathbb{F}_2$ 中计算 $A(x - y)$ 时，只有二者不同的坐标有贡献。

引理 5.1 (一行的不同概率)。 对每一行 i ,

$$\mathbb{P}((Ax)_i \neq (Ay)_i) = q_p(t) := \frac{1 - (1 - 2p)^t}{2}. \quad (20)$$

不同行对应的不同指示变量相互独立。

证明。 令 $B_1, \dots, B_t$ 为不同坐标处相互独立的矩阵元素。两个输出比特不同时且仅当它们的和为奇数。若该概率为 q ，则

$$1 - 2q = \mathbb{E}(-1)^{B_1 + \dots + B_t} = \prod_{j=1}^t \mathbb{E}(-1)^{B_j} = (1 - 2p)^t.$$

解出 q 即可。不同行使用互不相交的独立矩阵元素。 $\square$

还有一种直观的抽样解释。以概率 $2p$ 独立地将每个坐标纳入集合 D ；在 D 上选择独立公平比特作为该行元素，其余位置设为零。这恰好产生独立 $\text{Bern}(p)$ 元素。若 D 避开所有不同坐标，输出相同；否则输出的奇偶差是公平比特。完全避开的概率为 $(1 - 2p)^t$ ，再次得到式 (20)。

5.2 针对一个半径调节摘要

选择

$$p = \frac{1 - 2^{-1/r}}{2}, \quad \gamma_c = \frac{1}{8} - 2^{-(c+2)}, \quad \tau_c = \frac{3}{8} - 2^{-(c+2)} = \frac{1}{4} + \gamma_c. \quad (21)$$

由 $c > 1$ 可知 $\gamma_c > 0$ 。式 (20) 变为关于 t 单调递增的函数 $q_p(t) = (1 - 2^{-t/r})/2$ 。因此

$$t \leq r \implies q_p(t) \leq 1/4, \quad (22)$$

$$t > cr \implies q_p(t) > 1/2 - 2^{-(c+1)} = 1/4 + 2\gamma_c. \quad (23)$$

阈值 τ_c 位于这两个期望界的中点。例如, $c = 2$ 时, $\gamma_c = 1/16$ 、 $\tau_c = 5/16$; 近点对的均值至多为 $1/4$, 远点对的均值大于 $3/8$ 。

令 $Z = \text{Ham}(Ax, Ay)$, 它是 k 个独立 Bernoulli 变量之和。附录 B 回顾并证明的 Hoeffding 不等式给出

$$\text{Ham}(x, y) \leq r \implies \mathbb{P}(Z > \lfloor k\tau_c \rfloor) \leq e^{-2\gamma_c^2 k}, \quad (24)$$

$$\text{Ham}(x, y) > cr \implies \mathbb{P}(Z \leq \lfloor k\tau_c \rfloor) \leq e^{-2\gamma_c^2 k}. \quad (25)$$

确实, 第一个事件中的整数 Z 大于 $k\tau_c \geq \mathbb{E}Z + \gamma_c k$; 第二种情形中, $Z \leq k\tau_c < \mathbb{E}Z - \gamma_c k$ 。两种偏差都至少为 $\gamma_c k$ 。

5.3 为每个可能的压缩查询预计算答案

令 $s = \lfloor k\tau_c \rfloor$ 、 $z_i = Ay_i$ 。对每个 $u \in \{0, 1\}^k$, 定义列表

$$B_s(u) = \{i \in [n] : \text{Ham}(u, z_i) \leq s\}. \quad (26)$$

无论原始维数多大, 可能的签名都只有 2^k 个。列表存储的是原始点的编号, 而非其 d 比特向量的副本。

用完整签名表求 Hamming 近邻

预处理。 选择 A 并计算全部 $z_i = Ay_i$ 。枚举 $u \in \{0, 1\}^k$, 建立 $B_s(u)$ 。保存数据集与查找表。

查询 x 。计算 $u = Ax$ 。若 $B_s(u)$ 为空, 返回 $\perp$ 。否则任取 $i \in B_s(u)$, 计算 $\text{Ham}(x, y_i)$; 若距离不超过 cr , 返回 y_i , 否则返回 $\perp$ 。

定理 5.2 (完整查表搜索)。 对每个大小为 $n \geq 1$ 的固定数据集 S 、每个固定查询 x 以及 $0 < \delta < 1$, 选择

$$k = \left\lceil \frac{\ln(2n/\delta)}{2\gamma_c^2} \right\rceil. \quad (27)$$

算法以至少 $1 - \delta$ 的概率解决 (c, r) -近邻查询。即使成功事件没有发生, 算法报告的点也一定合格。

证明。 对距离不超过 r 的每个数据点, 将其未进入被查询列表视为坏事件; 对距离大于 cr 的每个数据点, 将其进入列表视为坏事件。每个点至多对应一个这类事件。由式 (24)–(25) 及并合界, 任一坏事件发生的概率至多为 $ne^{-2\gamma_c^2 k} \leq \delta/2$ 。

在补事件上, 每个真正的 r -近邻都位于 $B_s(Ax)$ 中, 而列表中每个编号对应的原始距离均不超过 cr 。所以, 只要真正近邻存在, 列表就非空, 任取编号都会通过验证。若全部点都比 cr 更远, 则列表为空。对 A 的任意抽样结果, 验证步骤都杜绝了不合格输出。□

并合界涉及的是一个固定查询到 n 个点的距离, 而不是全部 2^d 个查询。查找表为每个压缩查询设置了地址, 但这本身并没有加强概率保证的量词。

5.4 空间、预处理与更小的表

计算全部签名需要 $O(nkd)$ 次位操作；直接枚举全部列表还需 $O(nk2^k)$ 次比特比较。所有列表中的编号总数至多为 $n2^k$ 。除编号和指针空间外，还应保留 nd 个输入比特与 kd 个矩阵比特。查询需要 $O(kd + d)$ 次位操作，以及常数查表访问；这里采用字长足以寻址 2^k 个表项及数据集的 RAM 模型。计算 k 比特地址的工作已包含在变换代价中。

事实上，每个非空 $B_s(u)$ 只需保存一个代表点编号，空列表保存 $\perp$ 。定理 5.2 的证明不需要其他成员：在成功事件上，被查询单元中的任意代表都合格。因此，表可以缩小为 2^k 个表项。除寻址约定外，其比特空间为

$$O(nd + kd + 2^k \lceil \log_2(n+1) \rceil). \quad (28)$$

在较小的表中，验证尤其有用：不幸选到过远的代表只能导致假阴性，不会产生假阳性。

对固定 $c > 1$ 和 $\delta = n^{-O(1)}$ ，有 $k = O_c(\log n)$ 、 $2^k = n^{O_c(1)}$ 。因此，多项式空间换来了 $O_c(d \log n)$ 查询时间。但空间指数可能很大。当 $c \downarrow 1$ 时， $\gamma_c = \Theta(c-1)$ ，所以 $k = O((c-1)^{-2} \log(n/\delta))$ 。这首先是一个理论上的数据结构，并不自动意味着实用的小索引。这类随机 Hamming 搜索与 Kushilevitz、Ostrovsky 和 Rabani 的工作密切相关 [6]。

抽样概率与有限描述。 精确表述允许按指定参数独立抽取 $\text{Bern}(p)$ 。有限比特版本可把 p 舍入为二进有理数 p' ，满足 $|p' - p| \leq \gamma_c/(4d)$ ，并把 τ_c 舍入到误差 $\gamma_c/4$ 以内。耦合一行的 d 个元素可知，任意一行的不同概率至多改变 $d|p' - p| \leq \gamma_c/4$ 。到阈值的间隔仍至少为 $\gamma_c/2$ ，故将 k 增大四倍即可保留原来的失败概率界。对固定可计算的 c 与整数半径， $O(\log(d/\gamma_c))$ 比特的偏置精度足够；用独立公平比特块就能实现舍入后的 Bernoulli 试验。此构造不需要连续随机矩阵。

依赖半径的摘要不是 Hamming JL 嵌入

随着 t/r 增大， $q_p(t)$ 趋向 $1/2$ 。因此，差别很大的两个大 Hamming 距离，其压缩期望也可能十分接近。这种饱和不妨碍固定半径近邻查询：我们只需要区分 $t \leq r$ 与 $t > cr$ 。

6 局部敏感哈希：只搜索少数几个桶

完整查找表为空间中的全部 2^k 个压缩签名支付代价。局部敏感哈希则只存储包含数据的桶，并使用多个独立表。分析必须同时兼顾两个目标：真正近邻应当在某处发生碰撞，而远点不能造成过多查询工作。

6.1 碰撞概率的间隔及其放大

定义 6.1 (局部敏感哈希)。 映射 $h: X \rightarrow U$ 上的一个分布称为 (r, cr, p, q) 局部敏感哈希族，若 $c > 1$ 、 $r > 0$ 、 $0 < q < p < 1$ ，并且对每个固定 $x, y \in X$ ，都有

$$\text{dist}(x, y) \leq r \implies \mathbb{P}(h(x) = h(y)) \geq p, \quad (29)$$

$$\text{dist}(x, y) > cr \implies \mathbb{P}(h(x) = h(y)) \leq q. \quad (30)$$

不同于普通哈希，相似对象之间的碰撞在这里是有用的。哈希族是一个分布，概率来自抽取的函数，而不要求输入点本身随机。Indyk 与 Motwani 将 LSH 引入高维近似搜索算法 [4]。

选择 K 个独立哈希函数，将输出拼接为

$$g(x) = (h_1(x), \dots, h_K(x)) \in U^K. \quad (31)$$

对固定点对，完整键相等要求每个分量都相等。由独立性， g 是 (r, cr, p^K, q^K) -敏感的。这种张量化，或 AND 放大，降低了远点碰撞概率，但也同时降低近点碰撞概率。

对 $n \geq 2$ ，取

$$K = \left\lceil \frac{\ln n}{\ln(1/q)} \right\rceil, \quad p_* = p^K, \quad L = \lceil 1/p_* \rceil. \quad (32)$$

于是 $q^K \leq 1/n$ 。单个表被查询的桶中，远点出现次数的期望至多为一，但某个近点进入该桶的概率仅为 p_* 。使用 L 个独立表就提供了 OR 放大：近点只需在其中一个表内发生碰撞。

6.2 数据结构与候选检查预算

在表 j 中，用字典将每个实际出现的键 $g_j(y_i)$ 映射到具有该键的全部点编号列表。相同键必须合并到同一列表中，而不能互相覆盖。原始点只保存一份，在验证时访问。字典进行精确查询；第 6.4 节给出具体实现，并核算长键的代价。

带检查上限的 LSH 近邻搜索

预处理。独立生成 L 个拼接哈希 $g_1, \dots, g_L$ ，每个使用 K 个独立基础哈希。建立对应的 L 个桶字典。

查询 x 。计算 $g_j(x)$ ，检索各表中对应的桶。按固定顺序遍历这些列表，总共检查至多 $10L$ 次候选出现。每次遇到 y_i ，计算 $\text{dist}(x, y_i)$ ，返回第一个距离不超过 cr 的点。若列表已遍历完毕，或检查预算用尽而仍未找到合格点，则返回 $\perp$ 。

同一点可能位于多个被查询的桶中，并被重复检查。按出现次数计数，可以同时明确工作量与概率分析，不需要额外的去重数据结构。

定理 6.2 (LSH 的基本保证)。对每个含 $n \geq 2$ 个点的固定数据集，以及独立于随机哈希的固定查询，上述结构解决 (c, r) -近邻问题的失败概率至多为

$$e^{-1} + \frac{1}{10} < \frac{1}{2}. \quad (33)$$

算法绝不返回距离大于 cr 的点。候选出现的检查次数确定性地不超过 $10L$ 。

证明。假设存在 $y_* \in S$ 满足 $\text{dist}(x, y_*) \leq r$ ，并独立于哈希固定一个这样的点。它在全部 L 个表中都未发生碰撞的概率至多为

$$(1 - p_*)^L \leq e^{-p_* L} \leq e^{-1}.$$

令 F 统计所有被检索桶中的远点出现，即满足 $\text{dist}(x, y_i) > cr$ 且 $g_j(x) = g_j(y_i)$ 的二元组 (j, i) 数。由期望的线性性，

$$\mathbb{E}F = \sum_{j=1}^L \sum_{i: \text{dist}(x, y_i) > cr} \mathbb{P}(g_j(x) = g_j(y_i)) \leq Lnq^K \leq L. \quad (34)$$

这里不需要不同数据点的碰撞事件独立。Markov 不等式给出 $\mathbb{P}(F \geq 10L) \leq 1/10$ 。

若 y_* 在某表中碰撞，且 $F < 10L$ ，则遍历一定在预算用尽前遇到合格点：第一个合格出现之前的全部出现都来自远点，而整个桶集合中的远点出现总数都小于 $10L$ 。所以失败事件包含

在刚才两个坏事件的并中，二者也不必独立。若所有数据点都比 cr 更远，验证迫使输出为 $\perp$ 。在间隙区内，两种结果都允许。 $\square$

这个证明的一个关键特点是：不需要限制整个桶的大小。桶内可以有許多合格点，因为查询遇到第一个就停止。只有远点引起的工作可能延误成功。

6.3 指数与实际资源界

定义

$$\rho = \frac{\ln(1/p)}{\ln(1/q)} = \frac{\ln p}{\ln q} \in (0, 1). \quad (35)$$

对式 (32) 中的 K 取整后，有精确而有用的界

$$p n^{-\rho} \leq p_* \leq n^{-\rho}, \quad n^{\rho} \leq L \leq p^{-1} n^{\rho} + 1. \quad (36)$$

因此，若 p 有正常数下界，则 $L = O(n^{\rho})$ 。我们得到的权衡是：查询 L 个桶、进行 $O(L)$ 次距离验证，并在全部表中保存 nL 次点的出现。

具体实现时，设每个基础哈希值用 b 比特编码，求值时间为 T_h ，一次距离计算的时间为 T_{dist} 。使用第 6.4 节中的精确压缩字典树，查询时间为

$$O(L(KT_h + Kb + T_{\text{dist}})). \quad (37)$$

字典使用 $O(nL)$ 个机器字记录，此外还需保存原始数据集与 LK 个基础哈希函数的描述。长键不必永久保存：叶子记录一个代表数据点编号，在最终等值检查时重新计算其键。更简单的实现若保留所有完整键，会额外使用 $O(nL Kb)$ 比特。字长须足以容纳点编号、地址以及 Kb 比特键中的位置编号。

例如，一种确定性预处理实现先计算全部键，再对每个表的键排序，将相同键分组并建立压缩字典树。除哈希函数生成外，时间为 $O(nLK(T_h + b \log n))$ ；临时键存储可在不同表之间复用。这里明确区分了候选编号的空间、向量空间和哈希描述空间。

若基础哈希与距离计算均为单位代价，哈希值长度为常数，且 p, q 固定，式 (37) 就化为熟悉的 $O(n^{\rho} \log n)$ 查询时间，索引记录数为 $O(n^{1+\rho})$ 。这些假设对高维数据并非自动成立。下面的 Hamming 例子将保留维数相关的代价。

6.4 长签名的精确桶查询

键 $g_j(y)$ 有 $B = Kb$ 比特。二叉字典树依次按照键的各位分支。存储 m 个不同键的普通字典树至多使用 $1 + mB$ 个节点；沿查询的 B 个比特前进，即可找到相等键的桶，或确定它不存在。第 7 节中 Hamming 向量的精确成员查询也可用这一结构实现。

为减少永久索引空间，压缩所有没有分支的内部路径。每个保留的内部节点记录其下方各键下一次发生差异的比特位置，以及两个孩子指针。叶子存一个代表点标识符及桶链表。每条根到叶的路径上，测试位置严格递增，故至多进行 B 次比特测试。每个内部节点都有两个孩子、共有 m 个叶子的二叉树恰有 $2m - 1$ 个节点。因此，每张表使用 $O(m)$ 个树记录及 n 个链表记录。

跳过公共前缀位置后，到达叶子本身不能确认键相等。对叶子中的代表点 y 重新计算 $g_j(y)$ ，将完整键与 $g_j(x)$ 比较，相等时才返回桶。存在的相等键必然沿自己的路径到达对应叶子；不相等的键会被最后的比较拒绝。因此，在查询键计算完成后，一次查找需要 $O(B + KT_h)$ 时间，而完整的存储键可以丢弃。这是精确字典，其正确性不依赖 LSH 碰撞的概率假设。

若完整键可放入常数个机器字，FKS 完美哈希 [5] 也是一种精确字典选择，支持常数时间访问桶。但哈希求值和距离验证的成本仍需单独计算。上面的压缩字典树实现不假设任意长的键都能在常数时间内读取。

提高成功概率。 独立建立 $a = \lceil \log_2(1/\delta) \rceil$ 份完整结构，依次查询，直到得到经过验证的点，或全部副本失败。对存在真正近邻的固定查询，全部失败概率至多为 $2^{-a} \leq \delta$ 。空间、预处理和最坏情况查询工作量均增大 a 倍。每个输出都经过验证，因此返回第一个合格答案仍具有单侧错误。

6.5 Hamming 空间中的坐标抽样

假设 $1 \leq r < d/c$ 。从 $[d]$ 均匀选择 J ，令 $h(x) = x_J$ 。若 $t = \text{Ham}(x, y)$ ，则

$$\mathbb{P}(h(x) = h(y)) = 1 - t/d. \quad (38)$$

因此，这是参数为 $p = 1 - r/d$ 、 $q = 1 - cr/d$ 的 (r, cr, p, q) 哈希族，其指数满足

$$\rho = \frac{\ln(1 - r/d)}{\ln(1 - cr/d)} \leq \frac{1}{c}. \quad (39)$$

证明时令 $u = r/d \in (0, 1/c)$ 。由

$$-\ln(1 - cu) = c \int_0^u \frac{dt}{1 - ct} \geq c \int_0^u \frac{dt}{1 - t} = -c \ln(1 - u),$$

除以正分母即可得到式 (39)。

每个基础哈希只读一位，所以 $T_h = O(1)$ 、 $b = 1$ 。完整 Hamming 距离验证需要 $O(d)$ 次位操作。必须注意，拼接长度为

$$K = \left\lceil \frac{\ln n}{-\ln(1 - cr/d)} \right\rceil \leq 1 + \frac{d}{cr} \ln n. \quad (40)$$

它不对所有 r/d 都具有统一的 $O(\log n)$ 上界。由于 $p > 1 - 1/c$ ，有 $L = O_c(n^{1/c})$ ，从而查询时间为

$$O_c \left(n^{1/c} \left(d + \frac{d}{r} \ln n + 1 \right) \right). \quad (41)$$

索引含 $O(nL)$ 个记录；哈希描述由 LK 个坐标编号构成；原始数据为 nd 比特。若 $cr \geq d$ ，直接返回任意数据点即可。半径小于一时可用精确成员查询处理。在指定的字模型下，比特打包可改善 T_{dist} ，但不能在不说明模型的情况下直接省略这一代价。

AND 压低不希望的碰撞，OR 恢复希望的碰撞

拼接使 $q^K \leq 1/n$ ，控制每个表中远点工作的期望。多个独立表让固定近邻以常数概率至少出现一次。检查上限将期望界转化为有界工作量的 Monte Carlo 算法，而精确距离验证杜绝假阳性。

7 从近邻到最近邻

固定半径结构回答的是阈值问题。要近似最近距离，可以在几何增长的多个尺度上分别建立结构。但有两点必须处理：半径范围应当覆盖需要考虑的情形，而间隙区内的近似答案未必单调。

7.1 已知正距离范围时的归约

假设查询满足已知范围 $0 < a \leq d_*(x) \leq b$ 。选择近邻因子 $c_0 > 1$ 与尺度比 $\eta > 1$ ，在以下半径建立 (c_0, r_j) 结构：

$$r_j = a\eta^j, \quad 0 \leq j < J, \quad J = 1 + \lceil \log_\eta(b/a) \rceil. \quad (42)$$

YES 表示返回了经过验证、距离不超过 $c_0 r_j$ 的点；NO 表示 $\perp$ 。暂时假设每个固定半径查询都满足其保证。

寻找 YES-NO 边界

先查询 r_0 。若回答 YES，返回相应点；否则查询 r_{J-1} 。由于 $r_{J-1} \geq b \geq d_*(x)$ ，后者一定回答 YES。

维护下标 $\ell < h$ ，使 r_ℓ 已回答 NO， r_h 已回答 YES，并保留后者返回的点。只要 $h - \ell > 1$ ，就查询 $j = \lfloor (\ell + h)/2 \rfloor$ 。若回答 NO，令 $\ell = j$ ；若回答 YES，令 $h = j$ 并保留新返回的点。下标相邻时，返回 h 处保存的点。

当 $J = 1$ 时，第一个半径已经覆盖整个距离范围，只需第一次查询。

命题 7.1 (由区间边界得到近似)。 若所有固定半径查询均正确，上述过程用 $O(1 + \log J)$ 次半径查询返回一个 $c_0\eta$ -近似最近邻。

证明。 若第一次回答 YES，其返回点的距离至多为 $c_0 a \leq c_0 d_*(x)$ 。否则，无论其他半径如何回答，过程始终在 ℓ 处保存一个实际的 NO，在 h 处保存一个实际的 YES。正确的 NO 蕴含 $d_*(x) > r_\ell$ ，因为存在该半径内的近点就必须回答 YES。经过验证的 YES 则给出距离不超过 $c_0 r_h$ 的点。终止时 $r_h = \eta r_\ell$ ，故其距离至多为

$$c_0 r_h = c_0 \eta r_\ell < c_0 \eta d_*(x). \quad (43)$$

每一步至少将中间的下标数减半。 □

较小半径处的 YES 未必意味着所有更大半径都回答 YES：两个查询都可能处于各自允许任意回答的间隙区。证明没有使用这种单调性，而只维护实际收到的答案所提供的两个事实。

对目标近似比 $C > 1$ ，取 $c_0 = \eta = \sqrt{C}$ 。固定半径结构的近似比与半径间隔都对最终因子有贡献；若两者都取 C ，一般只能得到 C^2 ，而不是 C 。

7.2 用端点处理覆盖任意度量查询

数据集内部的两两距离，本身不能界定查询的最近距离。例如， $\mathbb{R}$ 中的数据集 $\{0, 1\}$ 的最小与最大点间距离均为一，但查询既可能任意接近零，也可能离两个点都很远。正确的数据相关归约必须处理这两种极端情况。

设数据集包含 $n \geq 2$ 个不同点, 定义

$$D_{\min} = \min_{i < j} \text{dist}(y_i, y_j) > 0, \quad D_{\max} = \max_{i < j} \text{dist}(y_i, y_j), \quad R = D_{\max}/D_{\min}.$$

对目标 $C > 1$, 令 $c_0 = \eta = \sqrt{C}$, 并选择

$$a = \frac{D_{\min}}{2(c_0 + 1)}, \quad b = \max \left\{ a, \frac{D_{\max}}{C - 1} \right\}. \quad (44)$$

按照式 (42) 建立半径网格。先查询最小半径, 若返回点则直接返回; 若回答 NO, 再查询最大半径。这里最大半径处若回答 NO, 则返回任意数据点; 若回答 YES, 则启动先前的边界搜索。

定理 7.2 (带端点处理的近邻到最近邻归约)。 若固定半径查询均正确, 上述过程对任意位置的查询点都返回 C -近似最近邻, 包括最近距离为零的查询。需要存储的尺度数为

$$J = 1 + \left\lceil \log_{\sqrt{C}} \max \left\{ 1, \frac{2R}{\sqrt{C} - 1} \right\} \right\rceil. \quad (45)$$

对固定 $C > 1$, 该值为 $O_C(1 + \log R)$ 。

证明。 若 $d_*(x) \leq a$, 固定一个最近点 y_* 。任何不同数据点 y 均满足

$$\text{dist}(x, y) \geq \text{dist}(y, y_*) - \text{dist}(x, y_*) \geq D_{\min} - a > c_0 a.$$

因此, 最小半径查询只能返回 y_* 本身。这也处理了 $d_*(x) = 0$ 。若 $d_*(x) > a$, 而最小半径处返回了点, 则其距离至多为 $c_0 a < c_0 d_*(x)$ 。

最小半径处回答 NO 后, 有 $d_*(x) > a$ 。若最大半径也回答 NO, 则 $d_*(x) > r_{J-1} \geq D_{\max}/(C - 1)$ 。对任意数据点 y , 由三角不等式,

$$\text{dist}(x, y) \leq d_*(x) + D_{\max} < C d_*(x).$$

若最大半径回答 YES, 则可使用命题 7.1 中的边界论证; 该论证只需要 NO 与 YES 两个端点, 不依赖预先知道 d_* 的上界。最后, $b/a = \max\{1, 2R/(\sqrt{C} - 1)\}$, 得到式 (45)。□

若数据集只有一个点, 直接返回该点。计算 $D_{\min}$ 前可合并重复点。若这些参数尚未给出, 直接计算数据集的两个距离极值需要 $O(n^2 T_{\text{dist}})$ 预处理时间, 不能漏计。

7.3 失败概率与 Hamming 特例

对一个固定查询, 让每个半径结构的失败概率至多为 δ/J 。对全部已存储半径并合, 就能以至少 $1 - \delta$ 的概率保证自适应边界搜索所需的每个答案都正确。这避免了把单次查询保证直接条件化到一个自适应选择的尺度上, 而且不要求不同尺度之间独立。

若每个固定半径结构在该置信度下使用空间 s 、查询时间 t , 则总空间为 $O(Js)$, 查询时间为 $O(t(1 + \log J))$, 另计共享数据存储与端点参数的建立代价。若用独立 LSH 副本提高置信度, s 与 t 必须包含 $O(\log(J/\delta))$ 因子, 这部分代价并非免费。

在 Hamming 空间中, 有一种无需计算 $D_{\min}$ 与 $D_{\max}$ 的更简单方法。先检查查询是否精确属于 S 。对已存储的 d 比特向量建立二元字典树, 使用 $O(nd)$ 个结点, 可在 $O(d)$ 时间完成精确成员查询; 第 6.4 节解释其基本操作。若不属于 S , 则 $1 \leq d_*(x) \leq d$ 。在式 (42) 中取 $a = 1$ 、 $b = d$ 、 $c_0 = \eta = \sqrt{C}$, 便得到使用 $J = 1 + \lceil \log_{\sqrt{C}} d \rceil$ 个尺度的 C -近似最近邻搜索。零距离与半径端点的情形因而都有了明确处理。

8 把几种方法放在一起看

8.1 短表示的三种不同含义

随机表示需要保留的是查询所需的信息，而未必是输入的全部特征。本讲的三类构造给出了不同的保证。

构造	保留什么	怎样帮助搜索
欧氏 JL 变换	固定有限点集中所有点对的距离近似不变。	降低一次距离计算的代价；避免扫描全部数据仍需要索引。
Hamming 奇偶映射	在一个指定尺度上区分近距离与远距离。	小的布尔空间允许预先计算所有压缩查询的答案。
局部敏感哈希	近点与远点之间的碰撞概率间隙。	通过串接、重复和截断，限制候选点验证的工作量。

奇偶映射不是 Hamming 立方体的等距映射，LSH 碰撞保证也不是 JL 距离保证。不保留全部距离的映射仍可能有利于搜索；反过来，近似保距的映射本身也未必给出次线性查询算法。

8.2 欧氏维度能降到多低？

当失真固定且不允许距离完全塌缩时，最坏情况下对数维度不可避免。下面的初等论证解释了维度为什么必须依赖点数。

命题 8.1 (装球下界)。 设 $0 < \varepsilon < 1$ 。若把 $\mathbb{R}^n$ 中的 n 个标准基向量映到 $\mathbb{R}^k$ ，并满足式 (2)，则

$$k \geq \frac{\ln n}{\ln \left(1 + 2\sqrt{(1+\varepsilon)/(1-\varepsilon)} \right)}. \quad (46)$$

证明。 不同标准基向量的距离为 $\sqrt{2}$ 。像点之间的距离介于 $a = \sqrt{2(1-\varepsilon)}$ 与 $b = \sqrt{2(1+\varepsilon)}$ 之间。以各像点为中心、半径为 $a/2$ 的开球两两不相交，并都包含于以第一个像点为中心、半径为 $b + a/2$ 的球中。比较 k 维体积得到 $n(a/2)^k \leq (b + a/2)^k$ ，再取对数。若 $k = 0$ 且 $n \geq 2$ ，距离下界本身就不可能满足。□

对固定 $\varepsilon < 1$ ，这给出 $\Omega_\varepsilon(\log n)$ ，即使允许非线性嵌入也成立。但这个证明没有给出对 ε 的精确依赖。Larsen 与 Nelson [7] 研究了更强的下界；其参数范围不能忽略，因为保留原来的 d 个坐标始终是可行方案。

稀疏变换与快速变换。 稠密高斯矩阵和符号矩阵对每个向量需要 $O(kd)$ 时间。进一步的目标有两类：减少每列非零元数，或者利用矩阵结构加速乘法。稀疏 JL 变换 [8] 与快速 JL 变换 [9] 分别研究这些问题。后一类方法的一个思想是先施加随机符号，再做归一化 Hadamard 变换，将向量能量分散到各个坐标后再抽样。维度为二的幂时，Hadamard 变换的递归求和与作差结构支持 $O(d \log d)$ 次加减运算。抽样阶段仍需单独的集中性论证；JL 定理的存在性结论并不意味着可以任意删除坐标。这些改进关注求值成本，而不是改变近似保距的含义。

8.3 研究联系：嗅觉回路中的稀疏标签

Dasgupta、Stevens 与 Navlakha [10] 从果蝇嗅觉回路研究相似性搜索。他们给出的计算描述分为三步：归一化输入响应，通过稀疏二元连接扩展表示，再仅保留最强的响应。在所研究的回路中，约 50 个输入投射神经元连接到约 2000 个 Kenyon 细胞，每个后者汇总约六个输入；响应最强的约 5% 细胞构成稀疏标签。

这里与降维形成对照：中间表示先从约 50 维扩展到 2000 维，然后稀疏化。该工作把标签之间的相似性与相似性搜索联系起来，并评估了受此组织方式启发的算法。这种联系本身并不证明定义 6.1 中的特定 (r, cr, p, q) 碰撞不等式，也不意味着该回路实现了欧氏 JL 映射。它展示了更一般的设计思路：中间坐标可以更多，但最后活跃或需要存储的坐标很少。

检查随机搜索保证时应问什么

先明确度量、查询问题、随机性及概率保证覆盖的点对或查询，再核算表示的求值成本。候选点的检索与验证应分开分析。低维度、低碰撞指数和低总查询时间不是同一个结论。

9 习题

以下习题从距离保持逐步过渡到完整搜索结构。陈述概率保证时，请明确哪些输入是在抽取随机映射之前固定的。

欧氏降维

1. **距离平方与近似因子。** 从式 (2) 推导距离本身的上下界。某算法在投影空间中返回 b -近似最近邻，求能够保证原空间近似因子为 $C > b$ 的最大 ϵ 。最近距离为零时，需要注意什么？
2. **数据点与查询点。** 构造一个数据集及线性映射，使全部数据点之间的距离精确保持，却有某个查询到数据点的距离不被保持。再解释为什么从 $\mathbb{R}^d$ 到 $\mathbb{R}^k$ 的线性映射在 $k < d$ 时，无法对整个空间的全部距离保持一个正的乘性下界。
提示：投影可以保持包含数据集的一条直线，同时抹去一个正交方向。
3. **选择失败概率预算。** 对 n 个固定数据点与 Q 个固定查询，给出充分的高斯目标维度，使全部查询到数据点的距离平方在总失败概率至多为 δ 时保持于 $1 \pm \epsilon$ 范围。与保持数据和查询并集内的所有点对距离比较。
4. **随机子空间的计算。** 验证引理 3.1 中最优 t 对应的两个矩母函数因子。分别代入 $\beta = 1 - \epsilon$ 与 $1 + \epsilon$ ，并单独处理 $(1 + \epsilon)k \geq d$ 的情况。为什么不能把式 (11) 的分子与分母当作独立变量？
5. **随机符号不是高斯。** 直接证明式 (15)。当 $u = e_1$ 时，精确计算符号变换下的 $\|Au\|_2^2$ 。说明为什么正参数的矩母函数比较本身不能证明下尾界。

对半径敏感的压缩

6. **奇偶性与具体间隙。** 通过对随机支撑集取条件概率推导式 (20)，而不使用 $(-1)^Z$ 。当 $c = 2$ 时，计算 γ_c 、 τ_c ，以及使失败概率至多为 $1/n$ 的一个充分整数 k 。解释 $s = \lfloor k\tau_c \rfloor$ 中取整的作用。
7. **每个签名只存一个代表。** 证明在定理 5.2 使用的好事件上，每个非空表项仅存一个标识符就够了。说明为什么这并不意味着布尔立方体中的全部查询同时成功。在好事件之外，原空间验证仍能保证什么？

哈希与候选点验证

8. **改变候选预算。** 将 $L = \lceil 1/p_* \rceil$ 改为 $L = \lceil a/p_* \rceil$ ，将截断量 $10L$ 改为 BL ，其中 $a > 0$ 、 $B > 1$ 。必要时使用整数截断量，证明失败概率上界 $e^{-a} + 1/B$ 。区分随机碰撞事件与确定性的距离验证。
9. **出现次数与不同点数。** 一个远点可能出现在很多张表中。解释为什么不同远点数量的上界，本身不是遍历链表工作量的上界。给出按出现次数计数的实现，再给出避免重复计算距离、但仍核算每条遍历记录的实现。
10. **对半径的依赖。** 证明式 (39)，并推导 $r/d \rightarrow 0$ 时的 K 。固定 c ，比较 $r = 1$ 与 $r = \Theta(d)$ 。当一次完整 Hamming 距离计算需要 d 次比特运算时，写出对应的查询成本。

组合不同半径的结构

11. **答案不单调时的搜索。** 构造间隙区域中一组合法但不单调的 YES/NO 答案。证明边界搜索仍然得到 $c_0\eta$ 近似。指出它实际维持的两个不变式。
12. **最小尺度与最大尺度。** 对数据集 $\{0, 1\} \subset \mathbb{R}$ 与目标 $C = 4$ ，计算带端点保护的半径序列。分别检查查询等于数据点、非常接近数据点，以及远离数据集的情形。再说明为什么给每个存储尺度分配 δ/J 失败概率，就足以保护自适应选择的半径查询序列。

A 高斯分布与欧氏几何工具

A.1 归一化、矩与线性组合

高斯积分满足

$$\int_{-\infty}^{\infty} e^{-x^2/2} dx = \sqrt{2\pi}.$$

一种验证方法是将积分平方，再在 $\mathbb{R}^2$ 中使用极坐标，得到 $\int_0^{2\pi} \int_0^{\infty} e^{-r^2/2} r dr d\theta = 2\pi$ 。所有被积函数非负，因此迭代积分合法。对标准正态变量 Z ，由对称性有 $\mathbb{E}Z = 0$ 。密度 $\phi(z) = (2\pi)^{-1/2} e^{-z^2/2}$ 满足 $\phi'(z) = -z\phi(z)$ ，分部积分得到

$$\mathbb{E}Z^{2m} = (2m-1)\mathbb{E}Z^{2m-2} \quad (m \geq 1).$$

边界项为零，因为多项式乘以 $e^{-z^2/2}$ 趋于零。因此 $\mathbb{E}Z^2 = 1$ 、 $\mathbb{E}Z^4 = 3$ ，且 $\text{Var}(Z^2) = 2$ 。

对任意实数 a , 配方给出

$$\mathbb{E}e^{aZ} = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{az - z^2/2} dz = e^{a^2/2}. \quad (47)$$

若 $G_1, \dots, G_d$ 为独立标准正态变量, 则联合密度为

$$f_G(g) = (2\pi)^{-d/2} e^{-\|g\|_2^2/2}. \quad (48)$$

对任意正交矩阵 O , $\|Og\|_2 = \|g\|_2$ 且 $|\det O| = 1$, 故换元公式说明 $OG \stackrel{d}{=} G$ 。特别地, 对非零 $v \in \mathbb{R}^d$, 选择首行为 $v/\|v\|_2$ 的正交矩阵。观察第一个坐标可得

$$\langle v, G \rangle \stackrel{d}{=} \|v\|_2 G_1.$$

对独立标准正态变量平移与缩放, 即可证明线性组合规则 (5)。高斯投影的不同行使用互不相交的独立矩阵元素, 因此相应输出坐标独立。

A.2 均匀方向与随机正交标架

事件 $G = 0$ 的概率为零。在极坐标下, 密度 (48) 只依赖半径, 因此 $G/\|G\|_2$ 关于单位球面面积均匀分布, 方向与半径也独立。一维情形对应两个等概率方向。

下面直接说明第 3.1 节使用的对称性。设随机正交标架 Q 的分布在 $\mathbb{R}^d$ 的旋转下不变, 则对每个固定单位向量 u , $\|Qu\|_2$ 的分布都相同。因此, 它也等于先取独立均匀单位向量 U , 再计算 $\|QU\|_2$ 所得到的分布。给定任意标架 Q , 旋转坐标可将其张成的子空间变为前 k 个坐标方向, 而 U 仍然均匀。因此该分布就是 $\sqrt{U_1^2 + \dots + U_k^2}$ 的分布。代入 $U = G/\|G\|_2$ 即证式 (11), 不需要声称球面坐标相互独立。

对独立高斯行向量做 Gram-Schmidt 正交化, 就得到这样的标架: 这些向量几乎必然线性无关, 且正交化过程与旋转可交换。失败概率为零, 因为每个新的高斯向量具有密度, 而真线性子空间的 Lebesgue 测度为零。

A.3 随机符号所需的两个指数估计

由 $(2m)! = \prod_{j=1}^m (2j)(2j-1) \geq 2^m m!$, 逐项比较可得符号和所需的界:

$$\cosh s = \sum_{m \geq 0} \frac{s^{2m}}{(2m)!} \leq \sum_{m \geq 0} \frac{(s^2/2)^m}{m!} = e^{s^2/2}.$$

下尾论证使用 $e^{-z} \leq 1 - z + z^2/2$ ($z \geq 0$)。事实上, 差值 $f(z) = 1 - z + z^2/2 - e^{-z}$ 满足 $f(0) = f'(0) = 0$ 及 $f''(z) = 1 - e^{-z} \geq 0$, 积分两次即得 $f(z) \geq 0$ 。

B 碰撞与间隙论证中的概率工具

B.1 一阶矩、并合界与独立性

对非负随机变量 Z 与 $a > 0$, 由 $Z \geq a \mathbf{1}\{Z \geq a\}$ 得到 Markov 界 $\mathbb{P}(Z \geq a) \leq \mathbb{E}Z/a$ 。对事件 $E_1, \dots, E_m$, 逐点不等式 $\mathbf{1}\{\cup_i E_i\} \leq \sum_i \mathbf{1}\{E_i\}$ 给出

$$\mathbb{P}\left(\bigcup_i E_i\right) \leq \sum_i \mathbb{P}(E_i).$$

这两个界都不需要独立性。独立性用在别处: 独立坐标或独立行使指数矩分解, 独立生成的哈希表使固定近点错过所有表的概率相乘。即使不同数据点的碰撞事件相依, 期望的线性性仍然成立。

B.2 Bernoulli 变量的 Hoeffding 界

引理 B.1. 设 $k \geq 1$, $X_1, \dots, X_k \in \{0, 1\}$ 相互独立, $X = \sum_i X_i$ 。对任意 $t > 0$,

$$\mathbb{P}(X - \mathbb{E}X \geq t) \leq e^{-2t^2/k}, \quad \mathbb{P}(X - \mathbb{E}X \leq -t) \leq e^{-2t^2/k}. \quad (49)$$

证明. 对均值为 p 的 Bernoulli 变量 B , 令 $h(s) = \ln(1 - p + pe^s) - ps$ 。若 $p = 0$ 或 1 , 则 $B - p = 0$ 。否则 $h(0) = h'(0) = 0$, 且

$$h''(s) = q_s(1 - q_s) \leq \frac{1}{4}, \quad q_s = \frac{pe^s}{1 - p + pe^s}.$$

带积分余项的 Taylor 公式对正负 s 都给出 $h(s) \leq s^2/8$, 故 $\mathbb{E}e^{s(B-p)} \leq e^{s^2/8}$ 。由独立性, 将各界相乘得到 $\mathbb{E}e^{s(X-\mathbb{E}X)} \leq e^{ks^2/8}$ 。对 $s > 0$, 指数 Markov 不等式将上尾界为 $e^{-st+ks^2/8}$; 取 $s = 4t/k$ 。下尾则取 $s = -4t/k$, 或对 $\sum_i (1 - X_i)$ 应用同一论证。□

因此, 大小为 γk 的均值间隙, 使每个尾事件的概率至多为 $e^{-2\gamma^2 k}$ 。这正是区分 Hamming 近距离与远距离的估计。双侧偏差需要合并两个尾事件, 前面有因子 2; 这里的两个单侧分离事件各自只使用一侧尾界。

B.3 取整、置信度与固定输入

若统计量 Z 取整数值, 则 $Z > \lfloor s \rfloor$ 蕴含 $Z > s$, 而 $Z \leq \lfloor s \rfloor$ 蕴含 $Z \leq s$ 。这些关系保证 Hamming 阈值取整后仍可使用原来的界。向上取整重复次数只会改善递减的失败概率界, 但资源界中的常数可能依赖于基础碰撞概率。

对固定查询, 可对全部 J 个存储尺度的失败事件使用并合界, 使它们同时正确。这里不要求各尺度独立, 也不要求算法实际访问全部尺度。相反, 把重复结构的失败概率相乘需要相互独立的随机性。对于许多预先固定的查询, 应将它们的坏事件都纳入并合界; 依赖存储随机性而选择的新查询, 并不会自动继承单个固定查询的保证。

文献说明

欧氏嵌入定理由 Johnson 与 Lindenstrauss 提出 [1]。高斯、随机子空间与符号矩阵提供了互补的证明视角 [4, 2, 3]。随机表示下的 Hamming 搜索及 LSH 框架见 [6, 4]。其余文献对应第 8 节中介绍的改进方向与研究联系。

参考文献

- [1] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz Mappings into a Hilbert Space. *Contemporary Mathematics*, 26:189–206, 1984.
- [2] Sanjoy Dasgupta and Anupam Gupta. An Elementary Proof of a Theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003. [doi:10.1002/rsa.10073](https://doi.org/10.1002/rsa.10073).
- [3] Dimitris Achlioptas. Database-Friendly Random Projections: Johnson–Lindenstrauss with Binary Coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003. [doi:10.1016/S0022-0000\(03\)00025-4](https://doi.org/10.1016/S0022-0000(03)00025-4).
- [4] Piotr Indyk and Rajeev Motwani. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In *Proceedings of STOC*, pages 604–613, 1998. [Author manuscript](#).
- [5] Michael L. Fredman, János Komlós, and Endre Szemerédi. Storing a Sparse Table with $O(1)$ Worst Case Access Time. *Journal of the ACM*, 31(3):538–544, 1984.
- [6] Eyal Kushilevitz, Rafail Ostrovsky, and Yuval Rabani. Efficient Search for Approximate Nearest Neighbor in High Dimensional Spaces. *SIAM Journal on Computing*, 30(2):457–474, 2000. [doi:10.1137/S0097539798347177](https://doi.org/10.1137/S0097539798347177).
- [7] Kasper Green Larsen and Jelani Nelson. Optimality of the Johnson–Lindenstrauss Lemma. In *Proceedings of FOCS*, 2017. [arXiv:1609.02094](https://arxiv.org/abs/1609.02094).
- [8] Daniel M. Kane and Jelani Nelson. Sparser Johnson–Lindenstrauss Transforms. *Journal of the ACM*, 61(1), Article 4, 2014. [arXiv:1012.1577](https://arxiv.org/abs/1012.1577).
- [9] Nir Ailon and Bernard Chazelle. The Fast Johnson–Lindenstrauss Transform and Approximate Nearest Neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009. [doi:10.1137/060673096](https://doi.org/10.1137/060673096).
- [10] Sanjoy Dasgupta, Charles F. Stevens, and Saket Navlakha. A Neural Algorithm for a Fundamental Computing Problem. *Science*, 358(6364):793–796, 2017. [doi:10.1126/science.aam9868](https://doi.org/10.1126/science.aam9868).