arXiv:2301.05811v2 [cs.DB] 5 May 2023

加权最小值哈希在内积估计中胜过线性草图技术

AlineBessa[1*] Majid Daliri[1] Juliana Freire[1]ali
ne.bessa@nyu.edu daliri.majid@nyu.edu juliana.freire@nyu.edu
Cameron Musco[2] Christopher Musco[1] Aécio Santos[1]
cmusco@cs.umass.edu cmusco@nyu.edu aecio.santos@nyu.edu
Haoxiang Zhang[1]haoxiang.zhang@nyu.edu

[1]纽约大学[2]马萨诸塞大学阿默斯
特分校

2023年5月8日

摘要

我们提出了一种新方法，可独立计算紧凑草图，用于近似估计高维向量对之间的内积。我们的方法基于加权最小哈希算法，具有很强的准确性保证，在内积估计方面优于流行的线性草图技术 方法，例如 CountSketch 和 Johnson-Lindenstrauss 投影。具体而言，虽然我们的方法在稠密向量上与线性草图技术的表现完全相同，但对于非零条目重叠有限的稀疏向量，其误差更低 。这类向量常见于许多涉及稀疏数据的应用，也常见于日益流行的数据集搜索应用；在这些应用中，内积用于估计数据协方差、条件均值，以及涉及未连接表中列的其他量。我们通过实验证明，我们的方法在稀疏向量上的表现优于现有的线性草图和无权哈希草图，从而补充了理论结果。

1 引言

两个向量 a 和 b 的内积，$\langle a, b\rangle = \sum_{k=1}^{n} a[k]b[k]$，是一种应用广泛的运算。除许多其他应用外，内积还可用于计算文档相似度 [Salton 等人， 1975], 评估学习得到的分类模型，以及估计连接规模 [Alon 等，1999b, Achlioptas，2003，Rusu 和 Dobra， 2008]。然而，在涉及超高维向量的现代应用中，精确计算内积可能难以实现。其计算成本为 $O(n)$，计算 $\langle a, b\rangle$ 需要从内存中读取 $O(n)$ 个数；如果 a 和 b 存储在不同的机器上，则需要通信$O(n)$ 个数。

解决这一问题的常见方法是预先为每个向量计算一个小空间压缩表示（即草图），分别记作 S(a) 和 S(b)。随后使用估计函数 F，通过 F (S(a)，S(b)) ≈ ⟨a，b⟩ 近似内积。草图技术的优势在于，它能同时降低存储、通信和运行时间复杂度。此外，草图一旦计算完成，便可反复用于估计与其他向量的内积。例如，给定另一个向量 c，我们可以通过 F (S(a)，S(c)) 估计 ⟨a，c⟩ ≈ 。

用于近似内积的草图方法早已在计算机科学的各个领域得到广泛应用。在机器学习中，这些方法可用于加速支持向量机或逻辑回归等大规模线性模型的训练 [Arriaga 和 Vempala，2006年，Li等人， 2011]。在关系数据库中，内积草图被用于查询优化器，以便在无需执行涉及大型连接的高成本查询的情况下选择最优查询计划 [Cormode 等人， 2011]。近来，内积草图也已应用于数据集搜索

*作者姓名按字母顺序排列。

方法	大小为 $O(1/\epsilon^2)$ 的草图的误差	假设
Johnson-Lindenstrauss 投影 [Arriaga 和 Vempala， 2006], AMS [Alon 等人, 1999b], CountSketch [Charikar 等人， 2002]	$\epsilon \cdot \|\mathbf{a}\|\|\mathbf{b}\|$	None
MinHash（MH）采样 [Beyer 等人， 2007]	$\epsilon \cdot \max(\|\mathbf{a}_I\|\|\mathbf{b}\|, \|\mathbf{a}\|\|\mathbf{b}_I\|)$	a,b 为二进制，即包含 {0,1} 个条目。
Our 方法：加权 MinHash（WMH）采样	$\epsilon \cdot \max(\|\mathbf{a}_I\|\|\mathbf{b}\|, \|\mathbf{a}\|\|\mathbf{b}_I\|)$	None

表 1：使用各种草图方法估计 $\langle \mathbf{a}, \mathbf{b} \rangle$ 时高概率加性误差保证的比较。令 $I = \{i : \mathbf{a}[i] \neq 0 \text{ 和 } \mathbf{b}[i] \neq 0\}$ 表示a和b的支撑集的交集。$\mathbf{a}_I$ 和 $\mathbf{b}_I$ 表示a 和b 在 I 中索引对应的部分。由于 $\max(\|\mathbf{a}_I\|\|\mathbf{b}\|, \|\mathbf{a}\|\|\mathbf{b}_I\|) \leq \|\mathbf{a}\|\|\mathbf{b}\|$，我们提出的加权 MinHash（WMH）方法的界总是优于线性草图方法。我们的界与无权重 MinHash 的界相匹配，但不需要其严格的限制性假设——即a 和b 必须为二进制；该界适用于所有向量。

以及发现，在这些应用中，它们被用于发现可连接表 [Fernandez 等人， 2019]并估算其他列统计量，例如相关性 [Santos 等人， 2021], 而无需显式执行两个表之间的连接操作。我们将在第 1.2 节中讨论这些应用及其他应用。

此前已知的情况是什么？ 在上述所有应用中，一个主要关注点是优化草图大小（它决定存储、通信和运行效率）与 $\mathcal{F}(\mathcal{S}(\mathbf{a}),\mathcal{S}(\mathbf{b}))$对 $\langle \mathbf{a},\mathbf{b} \rangle$的近似精度之间的权衡。通常，草图越大，近似效果越好，但关键问题在于具体能好多少。目前，对于一般向量（即不假设其值分布或幅值的向量），只有基于线性草图算法的方法能够对这种权衡提供强有力的理论保证。这类算法包括著名的“拔河”草图，又称 AMS 草图 [Alon 等，1999b、Alon 等，1999a], CountSketch 算法 [Charikar 等， 2002],以及基于 Johnson-Lindenstrauss（JL）随机投影的方法 [Achlioptas，2003，Dasgupta 和 Gupta， 2003]。

所有这些方法的形式都类似。我们选择一个随机矩阵 $\Pi \in \mathbb{R}^{m \times n}$（$\Pi$ 可能包含独立同分布的随机元素，也可能具有更复杂的结构），以及集合 $S(\mathbf{a}) = \Pi\mathbf{a}$ 和 $S(\mathbf{b}) = \Pi\mathbf{b}$。每个草图都是一个长度为 m 的向量，由于 S 是一个线性函数，因此称其为线性草图。要估计内积，通常的做法是直接返回草图的内积 $\langle S(\mathbf{a}), S(\mathbf{b}) \rangle$.[1]

文本 内积估计的线性草图技术理论准确性保证 is:

事实 *1*（内积的线性草图 [*Arriaga* 和 *Vempala*， 2006]）。设 $\epsilon, \delta \in$ 分别为 *(0, 1)* 范围内的准确度和失败概率参数，并设 $m = O(\log(1/\delta)/\epsilon^2)$。令 $\Pi \in \mathbb{R}^{m \times n}$ 为一个随机矩阵，其中每个元素都独立地以相同概率取值为 $+\sqrt{1/m}$ 或 $-\sqrt{1/m}$。对于长度为 n 的向量 a、$b \in \mathbb{R}^n$，令 $S(a) = \Pi a$，$S(b) = \Pi b$。其概率至少为 $1 - \delta$，

$$|\langle \mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b}) \rangle - \langle \mathbf{a}, \mathbf{b} \rangle| \leq \epsilon \|\mathbf{a}\|\|\mathbf{b}\|$$

其中，$\|\mathbf{x}\|$ 表示标准欧几里得范数。

除了稠密随机矩阵外，事实 1 的类似结果也可针对稀疏JL矩阵、CountSketch矩阵及其他线性草图 [Cormode 等人， 2011]证明。该事实提供了强有力的准确性保证，且保证会随着草图大小的增加而改善 m，并自然地取决于a 和b的范数。据我们所知，此前只有线性草图方法能够提供如此强有力的理论保证。

1.1 我们的贡献

本文介绍了一种基于加权 MinHash 草图的内积草图新方法[Gollapudi 和 Panigrahy，2006年，Manasse 等人，2010年，Ioe， 2010], 它是经典MinHash方法的一种变体[Broder，1997 年，Broder 等人， 1998]。我们证明，我们的方法比事实 1得到了更精细的保证。特别是，它

[1]其他估计量也有应用，例如对多个近似内积取中位数 [Larsen 等人， 2021]。不过，其理论保证相似，通常仅在对失败概率的依赖方面有所不同 δ

在最坏情况下，当$\mathbf{a}$和$\mathbf{b}$为稠密向量时，其结果与线性草图相当[2]，；但当$\mathbf{a}$和$\mathbf{b}$是非零条目之间重叠有限的稀疏向量时，该方法始终能得到**更好的界**。正如第 1.2 节进一步讨论的那样，在内积草图技术应用于数据库问题和现代数据集搜索应用的许多场景中，这类向量对十分常见。

定理 2（主要结果）。设 $\epsilon, \delta \in (0, 1)$ 分别为准确度和失败概率参数，并且 $m = O(\log(1/\delta)/\epsilon^2)$。存在一个算法 S，可生成大小为 m 的草图（算法 3），以及一个估计过程 F（算法 5），使得对于任意 $a, b \in \mathbb{R}^n$，以至少 $1 - \delta$ 的概率，

$$|\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b})) - \langle \mathbf{a}, \mathbf{b} \rangle| \leq \epsilon \max(\|\mathbf{a}_\mathcal{I}\|\|\mathbf{b}\|, \|\mathbf{a}\|\|\mathbf{b}_\mathcal{I}\|)$$

上述，$\mathcal{I} = \{i : \mathbf{a}[i] \neq 0 \text{ 和 } \mathbf{b}[i] \neq 0\}$ 是$\mathbf{a}$的支撑集与$\mathbf{b}$的支撑集的交集。$\mathbf{a}_\mathcal{I}$ 和 $\mathbf{b}_\mathcal{I}$ 表示将$\mathbf{a}$ 和$\mathbf{b}$ 限制在$\mathcal{I}$中的索引上所得的向量。

我们总有 $\|\mathbf{a}_\mathcal{I}\| \leq \|\mathbf{a}\|$ 和 $\|\mathbf{b}_\mathcal{I}\| \leq \|\mathbf{b}\|$，因此可以对 $\max(\|\mathbf{a}_\mathcal{I}\|\|\mathbf{b}\|, \|\mathbf{a}\|\|\mathbf{b}_\mathcal{I}\|) \leq \|\mathbf{a}\|\|\mathbf{b}\|$ 作出界定。也就是说，定理 2 的保证在最坏情况下与事实 1 相同，但也可能显著更好。例如，考虑 a 和 b 的非零条目数量大致相同，但这些条目中只有 a $\gamma < 1$ 的比例在 a 和 b 中均非零。在这种情况下，合理的预期是 $\|\mathbf{a}_\mathcal{I}\|^2 \approx \gamma\|\mathbf{a}\|^2$ 和 $\|\mathbf{b}_\mathcal{I}\|^2 \approx \gamma\|\mathbf{b}\|^2$，因为 $\mathbf{a}_\mathcal{I}$ 和 $\mathbf{b}_\mathcal{I}$ 仅包含原始向量中 a γ 比例的条目。当然，实际的改进取决于数据；例如，我们可能有 $\|\mathbf{a}_\mathcal{I}\|^2$ 显著小于 $\gamma\|\mathbf{a}\|^2$，也可能并不比 $\|\mathbf{a}\|^2$ 小多少。

然而，考虑“典型情况”，即 γ 一定比例的非零项发生重叠时，我们可以预期，定理 2给出的界会比事实 1 好 $\sqrt{\gamma}$倍。因此，为了获得与线性草图相同的误差，我们的方法可以将 m 设得小 γ倍。在许多应用中，γ 非常小。例如，在第 5 节中，我们考虑一个文档相似度问题，其中 $\gamma \leq .05$ 对于95% 被草图化的向量对都是如此。这意味着，为达到指定的误差水平，所需的草图大小大约可以 $20x$ 降低。

凭借强有力的理论保证，线性草图算法已成为通用内积估计的首选方法 [Cormode 等人， 2011]。我们的结果首次表明，另一种方法也能提供更强的界。我们希望本文能成为进一步研究用于内积草图的基于哈希算法的起点。

1.2 动机性应用：数据集搜索

在介绍研究结果的技术细节并讨论相关工作之前，我们先详细介绍一个可能受益于所提出草图的应用，并借此说明：获得对内积估计的界限时，若能根据 a 和 b 中非零条目的重叠数量进行调整，将有多么重要。具体而言，我们考虑 数据集搜索 问题；近年来，这一问题日益受到关注 [Lehmberget al., 2015、Zhu 等， 2019、Yang 等， 2019、Zhu 等， 2016、Fernandez 等， 2019、Santos 等， 2021、Santos 等， 2022]。

假设一位数据科学家想了解 2022 年纽约市出租车乘客量波动的原因。分析人员手头只有一张包含两列的表：一列是 日期 列，另一列是 出租车乘坐次数，记录当天的乘车次数。为了开展分析，她需要在所在组织的数据湖或纽约市开放数据等公共数据仓库中寻找其他表（其中包含数千个数据集 [纽约市， 2022]），并将这些表与原表连接，以引入其他相关变量。例如，分析人员可能希望找到天气数据，因为天气会影响出租车乘客量。此外，她还希望以 自动化的方式找到一些 自己可能想不到的相关因素。

为了解决这个问题，我们希望有一些方法能够自动发现同时满足以下条件的表：1）能够与目标表连接（即也包含日期为 2022 年的列）；2）与分析人员的数据有实际意义上的关联。例如，如果高降水量日的出租车乘客量显著高于或低于其他日子，就应返回包含降水数据的表。要找到这样的表，暴力搜索并非不可行——我们通常无法承担将分析人员的表与搜索集中的所有表连接起来以寻找合适候选项的代价。因此，我们需要高效估计不同表之间的统计量，而无需实际构造它们的连接 [Santos 等人， 2021]。

草图技术已成为在未连接的数据表之间进行此类估算最受欢迎的方法 [Zhu 等， 2016，Fernandez 等， 2019，Yang 等， 2019，Santos 等， 2021，Santos 等， 2022]。具体而言，

[2]对于稠密向量，事实 1 实际上在常数因子范围内是最优的：近期研究表明，大小为 $m = o(\log(1/\delta)/\epsilon^2)$ 的草图无法在所有输入上以概率 $\epsilon\|\mathbf{a}\|\|\mathbf{b}\|$ 实现误差 $1 - \delta$， [Larsen 和 Nelson， 2017、Alon 和 Klartag， 2017]。我们的结果也达到了这一下界。

$\mathcal{T}_A$

K_A	V_A
1	6.0
3	2.0
4	6.0
5	1.0
6	4.0
7	2.0
8	2.0
9	8.0
11	3.0

$\mathcal{T}_B$

K_B	V_B
2	1.0
4	5.0
5	1.0
8	2.0
10	4.0
11	2.5
12	6.0
15	6.0
16	3.7

表$_{A\bowtie B}$

$K_{A\bowtie B}$	$V_{A\bowtie}$	$V_{B\bowtie}$
4	6.0	5.0
5	1.0	1.0
8	2.0	2.0
11	3.0	2.5

大小$(V_{A\bowtie})$ = 4
求和$(V_{A\bowtie})$ = 12.0
求和$(V_{B\bowtie})$ = 10.5
均值$(V_{A\bowtie})$ = 12.0/4 = 3.0

图2：表 $\mathrm{T}_{A\bowtie B}$ 是表 T_A 与表 T_B 一对一连接的结果。我们希望仅使用内积来近似表 $\mathrm{T}_{A\bowtie B}$ 的连接后统计量（例如连接大小、求和、均值和协方差）。

小空间草图已为搜索集中的所有数据表预先计算完成。分析人员发出查询以查找相关数据时，会将其表的草图与这些预先生成的草图进行比较；与显式物化表连接相比，这只需消耗一小部分计算资源 [Santos 等人， 2021]。

用于数据集搜索的内积草图。 有趣的是，在上述框架中，许多值得关注的问题都可以精确地表述为内积草图问题。为了解其中的原因，请看图 $\mathcal{T}_A$ 和 $\mathcal{T}_B$ 这两个示例表，如图 2所示：每个表都包含一列键， K_A 和 K_B，以及一列值， V_A和 V_B。对这些表按键执行连接 操作后，会生成输出表 $\mathcal{T}_{A\bowtie B}$。 [3]

我们在图2 中列出了一些在搜索相关数据集时可能希望从$\mathcal{T}_{A\bowtie B}$ 中估计的统计量。我们认为，这些统计量都可以通过表的向量表示之间的内积来估计，分别记作$\mathbf{x}^{\mathbb{1}[K_A]}$、$\mathbf{x}^{K_A}$和$\mathbf{x}^{\mathbb{1}[K_B]}$、$\mathbf{x}^{K_B}$，并在图 3中展示。

首先，很容易看出，$\mathrm{T}_{A\bowtie B}$ 的大小等于 K_A 和 K_B中的键的交集大小，即$|K_A \cap K_B| = 4$。这又等于内积$\mathbf{x}^{\mathbb{1}[K_A]}$与$\mathbf{x}^{\mathbb{1}[K_B]}$。类似地，连接后 V_A 中的值的 SUM 聚合（即 SUM$(V_{A\bowtie})$）等于内积 SUM$(V_{A\bowtie})$ = $\langle \mathbf{x}^{V_A}$、$\mathbf{x}^{\mathbb{1}[K_B]}\rangle$。若要估计连接后均值（即 MEAN$(V_{A\bowtie})$），我们可以结合连接大小估计值与 SUM 估计值：

$$\mathsf{MEAN}(V_{A\bowtie}) = \frac{\langle \mathbf{x}^{V_A}, \mathbf{x}^{\mathbb{1}[K_B]}\rangle}{\langle \mathbf{x}^{\mathbb{1}[K_A]}, \mathbf{x}^{\mathbb{1}[K_B]}\rangle}.$$

最后，计算连接后的内积， $\langle \mathbf{x}^{V_A}$， $\mathbf{x}^{V_B}\rangle$ 可能会很有用。在上述应用中，对于包含降水数据和出租车乘客量的表，较高的内积可能意味着高降水日与高乘客量日相对应。

不同方法的比较。基于上述归约，JL 投影和 CountSketch 等线性草图方法，以及我们的加权 MinHash 方法，都可以直接应用于数据集搜索问题。我们只需为搜索集中的所有表 T_B预先计算 S$(\mathbf{x}^{\mathbb{1}[K_B]})$ 和 S$(\mathbf{x}^{V_B})$。对其他向量变换（例如 S$((\mathbf{x}^{V_B})^2)$）进行草图处理，还可以估计连接后方差等其他量。

在搜索应用中，我们注意到向量长度 n 可能非常大。不过，计算草图并不需要将向量 $\mathbf{x}^{\mathbb{1}[K_A]}$和 $\mathbf{x}^{\mathbb{1}[K_B]}$完全物化：本文讨论的所有草图方法都只需处理向量的非零条目。此外，无需事先知道 n： 我们只需将 n设得足够大，以覆盖被草图化的键的整个取值范围（例如， $n = 2^{32}$ 或 $n = 2^{64}$）。

为了比较各种方法， 事实 1 和定理 2 表明，我们的 WMH 方法与线性草图技术在性能上的任何渐近差异，都取决于待处理向量非零条目之间的重叠程度。在数据集搜索中，这恰好对应于键集的 *Jaccard* 相似度 K_A 和 K_B。Jaccard 相似度较小时，我们的方法表现更好。例如，在图2中，14 个唯一键中只有4 个同时出现在两张表中，因此相似度为 $\approx .29$。在上述情形中，我们可以设想这个比例要小得多：例如，我们的数据分析师可能只有一张包含 2022 年出租车数据的表，却要将其与一张日期范围从 1960 年至今的天气数据表进行比较。Jaccard 相似度将为$1/63 \approx .016$。在第5 节中，我们将探讨一个涉及世界银行数据的数据集搜索用例 [世界银行， 2022] 其中42%的表对的 Jaccard 相似度为 $< .1$，而35%的表对的 Jaccard 相似度为 $< .05$。

[3]请注意，在图2所述示例中，我们假设是一对一连接。数据集搜索问题也可能涉及多对多连接，不过典型做法是使用数据聚合函数，将问题简化为一对一的情形 [Santos 等， 2021、Santos 等， 2022、Kanter 和 Veeramachaneni， 2015]。

索引	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$\mathbf{x}^{V_A}$	6.0	0	2.0	**6.0**	**1.0**	4.0	2.0	**2.0**	8.0	0	**3.0**	0	0	0	0	0
$\mathbf{x}^{\mathbb{1}[K_A]}$	1	0	1	**1**	**1**	1	1	**1**	1	0	**1**	0	0	0	0	0
$\mathbf{x}^{V_B}$	0	1.0	0	**5.0**	**1.0**	0	0	**2.0**	0	4.0	2.5	6.0	0	0	6.0	3.7
$\mathbf{x}^{\mathbb{1}[K_B]}$	0	1	0	**1**	**1**	0	0	**1**	0	1	**1**	1	0	0	1	1

图 3：表的向量表示 $\mathcal{T}_A$ 与 $\mathcal{T}_B$ 来自图 2。向量 $\mathbf{x}^{\mathbb{1}[K_A]}$（分别为 $\mathbf{x}^{\mathbb{1}[K_B]}$）是连接键 K_A（分别为 K_B）的向量表示，而 $\mathbf{x}^{V_A}$（分别为 $\mathbf{x}^{V_B}$）是列 V_A（分别为V_B）的向量表示。粗体数字表示来自 $\mathcal{T}_{A\bowtie B}$. 的连接结果中的条目。

1.3 论文结构

在第 2 节中，我们回顾相关的先前工作。在第 3 节中，我们概述对标准非加权MinHash 方法进行内积估计的分析。这项分析为我们的主要结果（定理 2）提供了技术上的预热，该结果关于加权 MinHash，并在第 4 节中给出。最后，在第 5 节中，我们通过对方法进行详细的定理 2 实证评估来支持。

2 相关工作

二进制向量内积估计。 除了用于估计一般实值向量 $\mathbf{a}$ 和 $\mathbf{b}$ 之间内积的线性草图方法外，针对包含 {0,1} 条目的二进制 向量这一特殊情形，先前也有大量研究。对于此类向量，近似内积相当于近似两个集合交集的大小。具体而言，任何 $\mathbf{a},\mathbf{b} \in \{0,1\}^n$ 都可以对应于集合 A 和 B，它们包含来自 $\{1,\dots,n\}$ 的整数。我们定义 A 包含所有满足 $\mathbf{a}[i] = 1$ 的 i，类似地，B 包含所有满足 $\mathbf{b}[i] = 1$ 的 i。注意，$\langle \mathbf{a},\mathbf{b}\rangle = |A \cap B|$。

根据事实 1，我们知道，大小为 $m = O(1/\epsilon^2)$ 的线性草图可以将 $\langle \mathbf{a}, \mathbf{b}\rangle$ 的估计误差控制在 $\epsilon\|\mathbf{a}\|\|\mathbf{b}\| = \epsilon\sqrt{|A||B|}$ 的加性误差范围内。然而，使用基于经典 MinHash 草图 [Broder，1997，Broder 等，1998，Manber，1994，Heintze，1996], 的 k-最小值（KMV）草图 [Beyer 等，2007], 或相关技术 [Li 和 König，2010，Li 等人，2012] 的非线性草图方法，可以得到更好的界。使用 $m = O(1/\epsilon^2)$ 的空间，这些方法可达到误差 $\epsilon\sqrt{\max(|A|,|B|)\cdot|A \cap B|}$，该误差始终小于 $\epsilon\sqrt{|A||B|}$ [Beyer 等，2007，Pagh 等，2014]。对于二值向量，这一界已在 [Pagh 等，2014] 中被证明是最优的。

我们的工作受到此前针对二值向量的一项结果启发。事实上，我们的定理 2 将该界严格推广到了所有实值向量。当 $\mathbf{a}$ 和 $\mathbf{b}$ 为二值向量时，有 $\|\mathbf{a}_I\|^2 = \|\mathbf{b}_I\|^2 = |A \cap B|$。因此不难看出，$\epsilon\sqrt{\max(|A|,|B|)\cdot|A \cap B|} = \epsilon\cdot\max(\|\mathbf{a}_I\|\|\mathbf{b}\|,\|\mathbf{a}\|\|\mathbf{b}_I\|)$，这恰好就是我们定理 2 中的界。我们在表 1 中总结了所有先前的内积草图方法与我们结果的比较。

超越二值向量。对于如何更准确地估计含有非二进制条目的向量的内积，相关研究较少。近期一篇论文 [Larsen 等人，2021] 证明了 CountSketch 方法的改进界，该界取决于 ℓ_1 $\mathbf{a}$ 的范数 和$\mathbf{b}$ （而非欧几里得范数）。对于某些向量，尤其是草图大小 m 较大时，这些界可能比事实 1 更紧。然而，这些结果无法与我们的结果直接比较。

我们采用了一种不同的方法，完全跳出了线性草图技术的范畴。我们的主要结果基于一类草图，我们统称为“加权 MinHash”方法 [Chi and Zhu，2017，Shrivastava，2016]。这些方法包括协调随机采样的加权版本 [Cohen and Kaplan，2007，Cohen and Kaplan，2013], 以及“一致加权采样”算法 [Manasse 等人，2010，Gollapudi 和 Panigrahy，2006]及其后续算法；这些算法在本质上等价，但计算成本更低 [Ioe，2010，Wu 等人，2019，Haeupler 等人，2014]。如第 4 节所示，加权 MinHash 草图能够处理条目幅值差异很大的向量（与二值向量不同，二值向量的所有非零条目幅值均为1）。

加权 MinHash 草图已用于许多应用，包括近似加权 Jaccard 相似度 [Wu 等人，2019], 用于检测具有加权特征的近重复项 [Manasse 等人，2010], 用于近似两个向量之间的距离 [Ioe，2010], 以及对图像直方图进行草图处理 [Shrivastava，2016]。在许多此类应用中，加权草图在经验上优于非加权草图。加权 MinHash 草图也用于计算一般的“SUM 聚合”查询，其中内积是一个

算法1 无权重 MinHash 草图

输入： 长度为 n 的向量 a，样本数 m，随机种子 s。

输出： 草图 $H_{\mathrm{a}} = \{H_{\mathrm{a}}^{hash}, H_{\mathrm{a}}^{val}\}$，其中 H_{a}^{hash} 和 $H^{val}{}_{\mathrm{a}}$的长度为 m，且分别包含 $[0,1]$ 中的值和来自 a 的值，分别

使用种子 s初始化随机数生成器。
对于 $\mathrm{i} = 1, \cdots, \mathrm{m}$ **do**
    均匀随机选择哈希函数 $h^i : \{1, ..., n\} \to [0,1]$。
    计算 $j^* = \arg\min_{j \in \{1,...,n\}, \mathrm{a}[j] \neq 0} h^i(j)$。
    设置 $H_{\mathrm{a}}^{hash}[i] = h^i(j^*)$并 $H^{val}\mathbf{a}\ [i] = \mathbf{a}[j^*]$
循环结束
返回 $\{H_{\mathrm{a}}^{hash}, H_{\mathrm{a}}^{val}\}$

特例 [Cohen and Kaplan, 2013]。不过，我们并未发现上述应用具有有力的最坏情况误差保证，更不用说一般内积估计问题了。一致加权采样也曾用于近似内积，见 [Li, 2017], 尽管其使用的估计量与我们的工作不同。然而，该方法并未提供非渐近的最坏情况保证。

局部敏感哈希。最后，我们从草图中估计内积的问题，与余弦相似度和最大内积搜索（MIPS）密切相关。后者的目标是从数据库中检索与给定查询向量具有最高余弦相似度（或内积）的向量。解决这些问题的一种方法是局部敏感哈希 [Gionis 等人, 1999], 此外，还有基于 MinHash 和随机投影的方法，例如 SimHash [Charikar, 2002]。研究发现，对于二进制数据，MinHash 的表现通常优于 SimHash；这与此前关于二进制内积估计的研究结果一致 [Shrivastava 和 Li, 2014]。

3 预热：无权 MinHash

记号。我们用粗体字母表示向量。对于向量 a，$\mathrm{a}[k]$ 表示其 k^{th} 个分量（索引从 1 开始）。对于两个长度为 n的向量 a、b，$\langle \mathrm{a}, \mathrm{b} \rangle = \sum_{k=1}^{n} \mathrm{a}[k]\mathrm{b}[k]$ 表示它们的内积。$\|\mathrm{a}\| = \sqrt{\langle \mathrm{a}, \mathrm{a} \rangle}$ 表示欧几里得范数，$\|\mathrm{a}\|_\infty = \max_{k \in \{1,...,n\}} |\mathrm{a}[k]|$ 表示无穷范数。$\|\mathrm{a}\|_1 = \sum_{k=1}^{n} |\mathrm{a}[k]|$ 表示 ℓ_1 范数。按照文献中的惯例 [Beyer 等人, 2007], 我们假设可以使用映射到实数轴上的均匀随机哈希函数。也就是说，假设我们可以构造一个随机函数 h，使得对于任意输入 $j \in \{1, \ldots, n\}$，$h(j)$在区间 $[0,1]$上独立且均匀分布。实际上，可以用低随机性函数 h 替代，该函数将输入映射到足够大的离散集合 $\{1/U, 2/U \ldots, 1\}$。通常，U取为 n^c，其中常数 c（例如 $c = 3$）[Cormode 等人, 2011]。我们用 $\Pr[E]$ 表示随机事件 E 发生的概率；$\mathbb{1}[E]$ 是指示随机变量，当 E 发生时取值为 1，否则取值为 0。$\mathbb{E}[X]$ 和 $\mathrm{Var}[X]$ 分别表示随机变量 X的期望和方差。

一种无权重方法。在介绍我们的加权 MinHash 草图方法之前，我们先回顾无权 MinHash 算法，并证明可通过该方法获得的内积估计界。该界与针对二值向量的先前工作密切相关 [Beyer 等, 2007 年, Pagh 等人, 2014] 且只有在对草图向量**a** 和**b** 作出严格假设时才成立——具体而言，它们的条目幅值必须有统一上界。尽管如此，这仍可作为我们主要结果的预热；该结果采用类似的策略证明，但通过加权采样消除了这一假设。

给定向量 a，我们通过将 a 中每个非零条目的索引哈希到区间 $[0,1]$，得到标准 MinHash 草图中的一个条目（例如参见 [Broder, 1997]）。然后存储最小的哈希值。使用独立选取的随机哈希函数重复这一过程m 次。对于非零索引集分别为 $\mathrm{A} = \{k : \mathrm{a}[k] \neq 0\}$ 和 $\mathrm{B} = \{k : \mathrm{b}[k] \neq 0\}$ 的二值向量 a 和 b，仅凭最小哈希值就可以估计 Jaccard 相似度 $|\mathrm{A} \cap \mathrm{B}|/|\mathrm{A} \cup \mathrm{B}|$，或并集大小 $|\mathrm{A} \cup \mathrm{B}|$ [（Flajolet 和 Martin，1985 年；Beyer 等，2007 年；Kane 等人, 2010]）。

对于非二值向量，是否通常会扩充标准 MinHash 草图，同时存储值——即哈希值最小的索引对应的值? 这一思路用于“协调采样”或“条件随机采样”草图中 [Cohen 和 Kaplan, 2013 年、Li 等, 2006 年、Cohen, 2016], 并且最近被用于扩展 MinHash 及其密切相关的 k最小值（KMV）草图，以估计向量相关性 [Santos 等人, 2021]。基本的增强型 MinHash 草图方法如算法1所示，该算法返回 H_{a}^{hash}和 H_{a}^{val}，分别作为最小哈希值及其对应向量值组成的向量。

算法 2 无权 MinHash 估计

输入： 草图 $H_{\mathbf{a}} = \{H_{\mathbf{a}}^{hash}, H^{val}{}_{\mathbf{a}}\}$ $H_{\mathbf{b}} = \{H^{hash}{}_{\mathbf{b}}\ H^{val}{}_{\mathbf{b}}\}$，使用算法1并以相同输入 m、s构造。**输出：** 对 $\langle \mathbf{a}$、$\mathbf{b}\rangle$的估计值。

1: Set $\tilde{U} = \frac{m}{\sum_{i=1}^{m} \min(H_{\mathbf{a}}^{hash}[i], H_{\mathbf{b}}^{hash}[i])} - 1$

2: **return** $\frac{\tilde{U}}{m} \sum_{i=1}^{m} \mathbb{1}\left[H_{\mathbf{a}}^{hash}[i] = H_{\mathbf{b}}^{hash}[i]\right] \cdot H_{\mathbf{a}}^{val}[i] \cdot H_{\mathbf{b}}^{val}[i]$

对于任意单个向量 $\mathbf{a}$，增强型 MinHash 草图 $H_{\mathbf{a}}$ 包含对$\mathbf{a}$中非零值进行的均匀子样本（有放回抽取）。这是因为对于所有 $i \in \{1, \ldots, m\}$，哈希的最小值都等概率地来自任意一个非零值对应的索引。更重要的是，草图可用于从交集 $\mathbf{a}$ 和$\mathbf{b}$中获取均匀子样本，也就是说，从两个向量均非零的条目中获取子样本。进而可利用该子样本估计总和 $\langle \mathbf{a},\mathbf{b}\rangle = \sum_{k=1}^{n} \mathbf{a}[k]\mathbf{b}[k]$，因为只有$\mathbf{a}[k]\mathbf{b}[k]$ 在$\mathbf{a}[k]$和$\mathbf{b}[k]$ 均非零时才会对总和有贡献。具体而言，有以下众所周知的事实：

事实 3. 考虑使用算法*1* **对向量** $\mathbf{a}$ 和 $\mathbf{b}$ 进行草图化，得到草图 $H_{\mathbf{a}}$ 和 $H_{\mathbf{b}}$。定义集合 $\mathcal{A} = \{i : \mathbf{a}[i] \neq 0\}$ 和 $\mathcal{B} = \{i : \mathbf{b}[i] \neq 0\}$。那么对于所有 $i \in \{1, \ldots, m\}$，有：

1. $H_{\mathbf{a}}^{hash}[i] = H^{hash}{}_{\mathbf{b}}[i]$ 以概率 $\frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|}$。
2. 如果 $H_{\mathbf{a}}^{hash}[i] = H_{\mathbf{b}}^{hash}[i]$，那么 $H^{val}{}_{\mathbf{a}}[i] = \mathbf{a}[j]$ 和 $H^{val}{}_{\mathbf{b}}[i] = \mathbf{b}[j]$ 对于从 $\mathcal{A} \cap \mathcal{B}$ 中均匀随机选取的 j。

事实 3 表明，要从$\mathbf{a}$ 和$\mathbf{b}$的交集中获得一个均匀子样本，我们只需取 $H_{\mathbf{a}}^{val}$ 和 $H^{val}{}_{\mathbf{b}}$中的所有条目，其中 $H_{\mathbf{a}}^{hash}$ 和 $H^{hash}{}_{\mathbf{b}}$中的对应条目相等——并且根据 (1)，它们以较高概率将会相等。

有了 事实 3，我们描述一种基于 MinHash 的内积估计量（算法 2）。该估计量将作为下一节中加权 MinHash 估计器的模板。

考虑 算法 2第 2 行中的求和。利用期望的线性性和 事实 3，我们可以计算其期望：

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{m} \mathbb{1}\left[H_{\mathbf{a}}^{hash}[i] = H_{\mathbf{b}}^{hash}[i]\right] \cdot H_{\mathbf{a}}^{val}[i] \cdot H_{\mathbf{b}}^{val}[i]\right] \\
&= m \cdot \mathbb{E}\left[\mathbb{1}\left[H_{\mathbf{a}}^{hash}[1] = H_{\mathbf{b}}^{hash}[1]\right] \cdot H_{\mathbf{a}}^{val}[1] \cdot H_{\mathbf{b}}^{val}[1]\right] \\
&= m \cdot \sum_{j \in \mathcal{A} \cap \mathcal{B}} \frac{1}{|\mathcal{A} \cup \mathcal{B}|} \mathbf{a}[j]\mathbf{b}[j] = \frac{m}{|\mathcal{A} \cup \mathcal{B}|} \cdot \langle \mathbf{a}, \mathbf{b}\rangle.
\end{aligned}$$

由上述内容可知，如果我们将求和式 $\sum_{i=1}^{m} \mathbb{1}\left[H_{\mathbf{a}}^{hash}[i] = H^{hash}{}_{\mathbf{b}}[i]\right] \cdot H^{val}{}_{\mathbf{a}}[i] \cdot H^{val}{}_{\mathbf{b}}[i]$乘以 $\frac{|\mathrm{A} \cup \mathrm{B}|}{m}$，就能得到 $\langle \mathrm{a}, \mathrm{b}\rangle$ 的无偏估计，正如所需。唯一的问题是，我们不知道$|\mathrm{A} \cup \mathrm{B}|$。无法根据我们的草图 $H_{\mathbf{a}}$ 和 $H_{\mathbf{b}}$精确计算这个并集大小。不过，可以利用 MinHash 草图中包含的相同信息对其进行估计。具体来说，由于 h^i 均匀地哈希到 $[0, 1]$，$\frac{m}{\sum_{i=1}^{m} \min(H_{\mathbf{a}}^{hash}[i], H^{hash}{}_{\mathbf{b}}[i])} - 1$ 可以很好地估计 $|\mathrm{A} \cup \mathrm{B}|$。这实际上是著名的~~Flajolet-Martin 不同~~元素估计器的一种标准变体（ [Flajolet 和 Martin，1985 年，Beyer 等人， 2007]）。在算法 2 的第 1 行，我们将 $\tilde{U}$ 设为该估计量，并在第 2 行乘以 $\frac{\tilde{U}}{m}$，以此作为 $|\mathrm{A} \cup \mathrm{B}|$ 的替代值。由此得到对 $\langle \mathrm{a}, \mathrm{b}\rangle$ 的最终估计量。

总体而言，我们能够证明以下集中界，适用于计算任意一对有界向量内积的估计量。对于二值向量，常数 c 在下文中等于1，并且我们精确地恢复了先前工作 [Pagh 等人， 2014]。

定理 *4*（中间结果：使用无权重 *MinHash* 进行内积草图）。设 $\epsilon, \delta \in$ 是准确度和失败概率参数，并且 $m = O(\log(1/\delta)/\epsilon^2)$。存在一个算法 S，可生成大小为 m 的草图（算法*1*），并配合估计过程 F，使得对于任意 a、$b \in \mathbb{R}^n$（其条目均有界于 $[-c, c]$），以至少 $1 - \delta$ 的概率，

$$|\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b})) - \langle \mathbf{a}, \mathbf{b}\rangle| \leq \epsilon \cdot c^2 \cdot \sqrt{\max(|\mathcal{A}|, |\mathcal{B}|) \cdot |\mathcal{A} \cap \mathcal{B}|}$$

对于$\mathrm{A} = \{i : \mathbf{a}[i] \neq 0\}$ 和 $\mathrm{B} = \{i : \mathbf{b}[i] \neq 0\}$。

算法 3 加权 MinHash 草图

输入： 长度为 n 的向量 $\mathbf{a}$，样本数 m，随机种子 s，整数离散化参数 L。
输出： 草图 $W_{\mathbf{a}} = \{W_{\mathbf{a}}^{hash}, W_{\mathbf{a}}^{val}, \|\mathbf{a}\|\}$，其中 $W_{\mathbf{a}}^{hash}$ 是一个长度为 m、取值属于 $[0,1]$ 的向量，$W_{\mathbf{a}}^{val}$ 是一个长度为 m 包含 $\mathbf{a}$ 的部分条目的向量，而 $\|\mathbf{a}\|$ 是标量，$\mathbf{a}$ 的欧几里得范数。

1: 使用种子 s 初始化随机数生成器。
2: 设置 $\tilde{\mathbf{a}} =$ 舍入$(\mathbf{a}/\|\mathbf{a}\|, L)$，使用算法 4。
3: 对于每个 $i \in \{1,\dots,n\}$，令 $\bar{\mathbf{a}}^{(i)}$ 为一个长度为 L 的向量，其前 $\tilde{\mathbf{a}}[i]^2 \cdot L$ 个条目设为 $\tilde{\mathbf{a}}[i]$。将剩余的条目设为 0。
4: 令 $\bar{\mathbf{a}} = [\bar{\mathbf{a}}^{(1)},\dots,\bar{\mathbf{a}}^{(n)}]$ 为长度为 $n \cdot L$ 的向量，由上述定义的向量拼接而成。
5: **对于** $i = 1,\cdots,m$ **do**
6: 选择均匀随机哈希函数 $h^i: \{1,\dots,nL\} \to [0,1]$。
7: 计算 $j^* = \arg\min_{j\in\{1,\dots,n\cdot L\}, \bar{\mathbf{a}}[j]\neq 0} h^i(j)$。
8: 设定 $W_{\mathbf{a}}^{hash}[i] = h^i(j^*)$ 和 $W_{\mathbf{a}}^{val}[i] = \bar{\mathbf{a}}[j^*]$。
9: **结束循环**
10: **返回** $\{W_{\mathbf{a}}^{hash}, W_{\mathbf{a}}^{val}, \|\mathbf{a}\|\}$.

定理 4 的完整证明见附录 A.1。证明需要两个技术要点。首先，我们必须界定一个“理想”估计量的方差，该估计量使用 $|A \cup B|$ 的精确值。由于 $\mathbf{a}$ 和 $\mathbf{b}$ 的条目均有 $[-c,c]$ 界，因此可以做到这一点。其次，我们可以界定用并集的估计值替代 $|A \cup B|$ 所引入的误差，如上文所述。为此，我们依赖以下标准结果，该结果表明，可以使用 $\mathbf{a}$ 和 $\mathbf{b}$ 的 MinHash 草图计算真实并集 $|A \cup B|$ 的 $(1\pm\epsilon)$ 相对误差近似，条件是 $m = O(1/\epsilon^2)$：

引理 1 (并集大小估计器 [Blum 等, 2020])。设 $\mathcal{A}$ 和 $\mathcal{B}$ 是 $\{1,\dots,n\}$ 的非空子集，且 $h^1,\dots,h^m: \{1,\dots,n\} \to [0,1]$ 是相互独立的均匀随机哈希函数。对于任意 $\epsilon,\delta \in (0,1)$，如果 $m = O\left(\frac{1}{\delta\epsilon^2}\right)$，那么估计量 $U = \frac{m}{\sum_{i=1}^m \text{Min}_{j\in\mathcal{A}\cup\mathcal{B}} h^i(j)} - 1$ 以至少 $1-\delta$ 的概率满足：

$$(1-\epsilon)|\mathcal{A}\cup\mathcal{B}| \le \tilde{U} \le (1+\epsilon)|\mathcal{A}\cup\mathcal{B}|.$$

请注意，尽管写法略有不同，U 引理 1 中的 引理 1 与算法 2 中的 U 完全等价（其中 A 和 B 分别包含 $\mathbf{a}$ 和$\mathbf{b}$的非零索引）。要理解原因，请注意$H_{\mathbf{a}}^{hash}[i] = \min_{j\in\mathcal{A}} h^i(j)$ 且 $H^{hash}{}_{\mathbf{b}}[i] = \min_{j\in\mathcal{B}} h^i(j)$。因此 $\min\left(H_{\mathbf{a}}^{hash}[i], H^{hash}{}_{\mathbf{b}}[i]\right) = \min_{j\in\mathcal{A}\cup\mathcal{B}} h^i(j)$。

4 主要结果：加权 MinHash

我们工作的主要技术挑战，是将上一节（定理 4）的结果推广到条目幅值差异很大的向量。显然，简单的 MinHash 方法不适用于这类向量。例如，考虑一个极端情况：$\mathbf{a}$ 和 $\mathbf{b}$ 在某个索引 i处都包含一个非常大的值，大到使项 $\mathbf{a}[i]\mathbf{b}[i]$ 在内积 $\langle\mathbf{a},\mathbf{b}\rangle = \sum_{k=1}^n \mathbf{a}[k]\mathbf{b}[k]$中占据主导地位。为了正确近似内积，我们需要分别将 $\mathbf{a}[i]$ 和 $\mathbf{b}[i]$ 纳入 $\mathbf{a}$ 和 $\mathbf{b}$ 的草图中。由于 MinHash 草图会从向量的交集中均匀采样条目，因此以低概率才会采到这些值。所以，它对 $\langle\mathbf{a},\mathbf{b}\rangle$的估计会很差。

为了解决大幅值条目带来的问题，我们修改了第 3 节中的方法，采用非均匀采样权重，并使用加权 MinHash 草图 [Manasse 等人, 2010]。这使我们能够以更高的概率采样向量中幅值较大的条目。具体而言，我们的目标是以与平方幅值 i^{th} 成正比的概率采样 $\mathbf{a}$ 中的**条目**$[i]^2$。加权 MinHash 草图以一种简单的方式实现非均匀采样：我们构造一个扩展的向量 $\bar{\mathbf{a}}$，其中包含与$\mathbf{a}$相同的条目，但会将条目重复多次，重复次数与其幅值成正比。然后，我们对 $\bar{\mathbf{a}}$应用标准 MinHash 草图。该方法详见算法 3。

舍入与归一化。加权 MinHash 允许我们以非均匀概率对条目进行采样，但这也带来了另一个挑战：由于 $\mathbf{a}$ 和$\mathbf{b}$ 的草图是独立计算的，我们不再从两个向量中以相同概率进行采样。对于$\mathbf{b}$，加权 MinHash 会以与$\mathbf{b}[i]^2$ 成比例的概率对索引进行哈希采样，而不是$\mathbf{a}[i]^2$。这种不匹配实际上可能降低我们从$\mathbf{a}$ 和$\mathbf{b}$ 中选出索引相同的条目的概率。

算法 4 加权 MinHash 的向量舍入

输入： 长度为 n 的单位向量 $\mathbf{z}$，整数离散化参数 L。

输出： 长度为 n 的单位向量 $\tilde{\mathbf{z}}$，其中 $\tilde{\mathbf{z}}[i]^2$ 对所有 i 均为 $1/L$ 的整数倍。

1：对于所有 $i \in \{1,\ldots,n\}$，$\tilde{\mathbf{z}}[i] = \text{sign}(\mathbf{z}[i]) \cdot \sqrt{\lfloor L\,\mathbf{z}[i]^2 \rfloor / L}$。2：令 $i^* = \arg\max_{i\in 1,\ldots,n} |\mathbf{z}[i]|$。3：固定 $\delta = 1 - \|\tilde{\mathbf{z}}\|^2$，然后令 $\tilde{\mathbf{z}}[i^*] = \text{sign}(\mathbf{z}[i^*]) \cdot \sqrt{\tilde{\mathbf{z}}[i^*]^2 + \delta}$。

4：**返回** $\tilde{\mathbf{z}}$。

算法 5 加权 MinHash 估计

输入： 草图 $W_{\mathbf{a}} = \{W^{hash}{}_{\mathbf{a}}\ W^{val}{}_{\mathbf{a}}\ \|\mathbf{a}\|\}$ 和 $W_{\mathbf{b}} = \{W^{hash}{}_{\mathbf{b}}\ W^{val}{}_{\mathbf{b}}\ \|\mathbf{b}\|\}$ 使用 算法 3 在相同输入下构造 m、s和 L。

输出： $\langle \mathbf{a}、\mathbf{b}\rangle$ 的估计值。

1：对于 $i \in \{1、\ldots\ldots、m\}$，设 $q_i = \min\left(W^{val}{}_{\mathbf{a}}[i]^2, W^{val}{}_{\mathbf{b}}[i]^2\right)$。2：设 $\tilde{M} = \frac{1}{L}\cdot\left(\sum_{i=1}^{m} W^{hash}{}_{\mathbf{a}}[i] = W^{hash}{}_{\mathbf{b}}[i]\right) \cdot \frac{\min(W^{val}{}_{\mathbf{a}}[i],\ W^{val}{}_{\mathbf{b}}[i])}{q_i} - 1\right)$。3：设 $I = \frac{\tilde{M}}{m}\sum_{i=1}^{m} \mathbb{1}\left[W^{hash}{}_{\mathbf{a}}[i] = W^{hash}{}_{\mathbf{b}}[i]\right] \cdot W^{val}{}_{\mathbf{a}}[i] \cdot W^{val}{}_{\mathbf{b}}[i]$。4：**返回** $\|\mathbf{a}\|\|\mathbf{b}\| \cdot I$

我们可以通过一种归一化策略来平衡这个问题。具体而言，算法 3第 2 行执行了一个简单但重要的预处理步骤，将 和 舍入 $\mathbf{a}$ 为单位向量 $\tilde{\mathbf{a}}$ 其平方项均为 $1/L$ 的整数倍，其中 L 是一个较大的整数（稍后确定）。舍入处理了一个小问题：由于我们通过重复$\mathbf{a}[i]$ 来控制每个条目被采样的频率，为了按照与$\mathbf{a}[i]^2$成比例的概率精确采样，所有条目的平方值都必须是同一个固定常数的整数倍。正如我们将证明的那样，L 可以取适当的值，使离散化对最终内积估计的准确度影响很小；而且该参数也不会影响算法1返回的草图大小。[4]

缩放处理了上文讨论的更大问题，即$\mathbf{a}$ 和$\mathbf{b}$之间采样概率的不匹配。令人惊讶的是，我们可以证明，当 $\|\mathbf{a}\| = \|\mathbf{b}\|$时，这种不匹配的影响是可控的。因此，尽管如果直接对$\mathbf{a}$ 和$\mathbf{b}$进行草图处理，确实可能构造出使算法失败的示例，但我们可以通过对$\mathbf{a}/\|\mathbf{a}\|$ 和$\mathbf{b}/\|\mathbf{b}\|$进行草图处理，近似 $\langle \mathbf{a}/\|\mathbf{a}\|, \mathbf{b}/\|\mathbf{b}\|\rangle$，然后将结果右乘 $\|\mathbf{a}\|\|\mathbf{b}\|$，从而得到我们的最终估计量，并获得最坏情况界。

推导内积估计器。 接下来，我们说明算法 2的动机，该算法用于根据我们的草图估计$\langle \mathbf{a}, \mathbf{b}\rangle$。请注意，加权 MinHash 草图（算法 3）实际上会针对扩展向量返回无权重 MinHash˜草图（算法1）。因此，我们可以对扩展向量 ˜$\mathbf{a}$，$\mathbf{b}$应用事实 3 以得到如下结果：

事实 5。 考虑向量 $\mathbf{a}$ 和 $\mathbf{b}$ 使用 算法 *3* 进行草图处理，得到 $W_{\mathbf{a}}$ 和 $W_{\mathbf{b}}$。定义 $\mathcal{A}$ 和 $\mathcal{B}$ 如 事实 *3*所述。对于所有$i \in \{1,\ldots,m\}$，我们有：

1. $W^{hash}{}_{\mathbf{a}}[i] = W^{hash}{}_{\mathbf{b}}[i]$ 的概率等于加权 *Jaccard* 相似度，即 $\bar{J} = \frac{\sum_{j=1}^{n} \min(\tilde{\mathbf{a}}[j]^2, \tilde{\mathbf{b}}[J]^2)}{\sum_{j=1}^{n} \max(\tilde{\mathbf{a}}[j]^2, \tilde{\mathbf{b}}[j]^2)}$。

2. 若 $W^{hash}{}_{\mathbf{a}}[i] = W^{hash}{}_{\mathbf{b}}[i]$，则对于从 $\mathcal{A} \cap \mathcal{B}$ 中选取的 j，我们有 $W^{val}{}_{\mathbf{a}} = \tilde{\mathbf{a}}[j]$ 和 $W^{val}{}_{\mathbf{b}} = \tilde{\mathbf{b}}[j]$ 的概率等于 $\min(\tilde{\mathbf{a}}[j]^2, \tilde{\mathbf{b}}[j]^2)/\sum_{i=1}^{n}\max(\tilde{\mathbf{a}}[j]^2, \tilde{\mathbf{b}}[j]^2)$。

事实 5 的证明见附录 A.2。陈述完毕后，我们给出一种基于$\langle \mathbf{a}、\mathbf{b}\rangle$ 的估计方法，这些草图由算法 3计算得出。该方法见算法 5，与上一节中用于非加权草图的估计量类似。唯一的区别是，由于我们采用非均匀概率进行采样，因此需要在求和时对样本进行反向加权，以确保期望结果正确。具体而言，考虑算法第 3 行中的求和。根据事实 5 和期望的线性性，我们得到：

[4]请注意，我们的舍入方法（算法 4）并不常规：它会将输入向量中的所有条目舍入为幅值更小的值，但向量中幅值最大的条目除外，该条目会被向上舍入。这种方案使我们在舍入时能够实现较小的相对误差，并避免取决于 $1/L$ 的加性误差。

$$\mathbb{E}\left[\sum_{i=1}^{m} \mathbb{1}\left[W_{\mathrm{a}}^{hash}[i] = W_{\mathrm{b}}^{hash}[i]\right] \cdot \frac{W_{\mathrm{a}}^{val}[i] \cdot W_{\mathrm{b}}^{val}[i]}{q_i}\right]$$

$$m \cdot \mathbb{E}\left[\mathbb{1}\left[W_{\mathrm{a}}^{hash}[i] = W_{\mathrm{b}}^{hash}[i]\right]\right] \cdot \frac{W_{\mathrm{a}}^{val}[i] \cdot W_{\mathrm{b}}^{val}[i]}{q_i}$$

$$= m \cdot \sum_{j \in \mathcal{A} \cap \mathcal{B}} \frac{q_j}{\sum_{i=1}^{n} \max(\tilde{\mathbf{a}}[i]^2, \tilde{\mathbf{b}}[i]^2)} \frac{\tilde{\mathbf{a}}[j]\tilde{\mathbf{b}}[j]}{q_j}$$

$$= \frac{m}{\sum_{i=1}^{n} \max(\tilde{\mathbf{a}}[i]^2, \tilde{\mathbf{b}}[i]^2)} \cdot \langle \tilde{\mathbf{a}}, \tilde{\mathbf{b}} \rangle.$$

˜因此，我们得到了一个估计量，其期望值等于 $\langle \tilde{}, \rangle$ m $M = \mathrm{a}\,\mathrm{b}$，再除以 $\sum_{j=1}^{n} \tilde{}[j]^2\ \tilde{}[j]^2$ M $\max(\mathrm{a}, \mathrm{b})$。这一项称为向量之间的加权并集大小。我们可以乘以 $\frac{M}{m}$，从而得到 $\langle$˜a ˜b$\rangle$ 的无偏估计量。由于 ˜a 和 ˜b 是分别通过除以 a 和 b 的欧几里得范数进行缩放得到的（暂且忽略舍入的影响），因此算法 5 第 4 行中的最终估计量会乘以‖a‖‖b‖。 ‖a‖ 和 ‖b‖ 的值分别显式存储在 a 和 b 的草图中（每个草图只需额外存储一个数）。

算法 5 的形式化分析可得出定理 2，相关内容见附录 A.2。该分析包含三个部分。首先，在分析无权估计量时，我们并不知道 M 的确切值，因此必须对其进行估计。我们可以˜ ˜利用这样一个事实： M 恰好等于扩展向量 ˜a 和 b 的非零˜ ˜ ˜索引集 A 和 B 之间的无权并集大小 |A ∪B|。这些扩展向量是在算法 3 中构造的。我们可以直接应用引理 1˜来得到一个估计量，在算法 5 中将其记为 M。其次，我们需要分析和 $\sum_{i=1}^{m} \mathbb{1}\left[W^{hash}\mathrm{a}[i] = W^{hash}\mathrm{b}[i]\right] \cdot \frac{W_{\mathrm{a}}^{val}[i]}{q_i} \cdot W^{val}[i]\mathrm{b}$ 的方差。这项分析利用了 ˜a 和 ˜b 是单位向量这一事实。第三，我们需要严格分析算法 3 第 2 行所执行的舍入过程的影响，以证明对 $\langle$˜a ˜b$\rangle$ 的良好估计确实能得到对 $\langle$ a/‖a‖、b/‖b‖$\rangle$ = ‖ ‖^{1}‖a b‖$\langle$a、b$\rangle$的良好估计。

最后，我们指出，对算法 5 的最终分析要求将 L 设定为 n^6/ϵ^2 的量级（使用算法 3进行草图处理时）。这个数值听起来可能很大，但请注意，该参数不会影响算法 3返回的草图大小，也不会影响我们的估计过程算法 5的运行时间。 L 确实会影响算法 3的运行时间，但如第 5 节所述，先前工作可以用于实现加权 MinHash 草图方法，使其对呈对数级依赖 L ——也就是说，对 $O(\log(n/\epsilon))$。

5 实验

为支持第 4 节中呈现的结果，我们使用合成数据和真实世界数据集进行了实验评估。

基线方法。我们将加权 MinHash 方法与 4 种基线方法进行比较，其中 2 种为线性方法，2 种为基于采样的方法，旨在评估估计内积时草图大小与准确度之间的权衡。这些方法如下：

Johnson-Lindenstrauss 投影（***JL***）：等价于 AMS 草图 [Alon 等，1999b，Achlioptas，2003]。使用一个随机矩阵 Π，其条目经过缩放 ±1（事实 1）。

CountSketch（***CS***）：由 [Charikar 等人 2002], 提出的经典线性草图方法，对应于与一个具有稀疏随机元素的 Π 相乘。我们遵循 [Larsen 等人 2021], 的实现，使用5 次草图重复，并取中位数以提升性能。

MinHash 采样（***MH***）：在算法1中描述的方法；我们使用单个草图，不进行中位数估计。

k-最小值采样（***KMV***）：一种与 MinHash 密切相关的基于采样的草图，但它从待绘制草图的向量中不放回地抽取样本。它也可用于估计并集大小。我们采用 [Beyer 等人 2007] 和 [Santos 等人 2021]的实现。

加权 ***MinHash*** 采样（***WMH***）：我们在算法 3中介绍的方法；我们使用单个草图，不进行任何中位数估计。

存储大小。对于线性草图，我们将矩阵乘法 $\mathbf{\Pi a}$ 的输出存储为 64 位双精度浮点数。我们还将 $W^{val}{}_{\mathbf{a}}$ 和 $H^{val}_{\mathbf{a}}$ 存储为 64 位双精度浮点数。由于基于采样的草图需要存储哈希值（在我们的情况下为 32 位整数），因此含有 m 个样本的基于采样的草图所占空间，与含有 $1.5x$ 行的 JL 草图一样大。在实验中，我们绘制存储大小，表示草图中的总位数除以 64，即草图中使用的 64 位双精度浮点数（或等效数据）的总数。标准的量化技巧可能可以用于缩小所有草图（线性草图和基于采样的草图）中数值的大小，但我们将这类方法的开发留待未来工作。作为起点，我们指出，关于量化 JL 投影已有一些有趣的研究[Jacques, 2015, Li 等人, 2016]，用于估计余弦相似度的 SimHash 方法，可以视为对 JL 草图进行的"1 位"量化 [Charikar, 2002]。

估计误差。对于所有图表，我们报告 $\langle \mathbf{a},\mathbf{b} \rangle$ 与估计值之间的绝对差，并将其除以 $\|\mathbf{a}\|\|\mathbf{b}\|$。这是线性草图准确性保证右侧出现的项 事实 1，因此这种缩放大致能确保误差介于 0 和 1之间，从而更容易比较不同数据集。我们始终报告 10 次独立试验的平均误差。

对 L的选择。注意，算法 3 中 L 的选择不会影响最终草图的大小，因此通常应尽可能取大。引理3给出的界表明 L 应设为 $\geq n^6$，但这些界可能较松（我们并未尝试优化多项式因子）；不过我们发现，至少必须确保 $L > n$。理想情况下，它应再乘以 100 或 1000。原因是，如果 a 是稠密向量，并像算法 1 中那样归一化为单位范数，那么其中大多数条目的平方值可能是 $< 1/n$（因为单位范数向量中，a 的平方条目的平均值始终为 $1/n$）。如果我们设定 $L < 1/n$，那么值为 $< 1/n$ 的任何条目都会被舍入为 0，这可能会降低内积估计的准确度。

高效加权哈希。当 L 很大时，直接实现算法 3 会慢得难以接受。"扩展"向量 $\bar{\mathbf{a}}$ 的长度为 $n \cdot L$，我们必须对该向量中的每个非零条目应用哈希函数。与之前一样，令 $\mathrm{A} = \{i : \mathrm{a}[i] \neq 0\}$，则 $|\mathrm{A}|$ 等于 a 中非零值的数量。若将每次哈希计算视为单位成本，则运行时间为 $O(|\mathrm{A}|m \cdot L)$，这过于庞大，因为 L 的取值大于 n。

幸运的是，我们可以利用加速加权 MinHash 草图的技术，将这一开销改进至 $O(|\mathrm{A}|m \cdot \log L) = O(|\mathrm{A}|m \cdot \log n)$。近年来，这类技术得到了广泛研究 [Ioe, 2010, Wu 等人, 2019, Haeupler 等人, 2014, Shrivastava, 2016]。这种方法节省的开销相当可观，使得对每个 m 样本进行草图处理的计算成本几乎与输入规模呈线性关系。在各种更快的方法中，我们具体采用了简单的"活跃索引"技术，该技术最早由 [Gollapudi 和 Panigrahy, 2006]提出。大致思路是：对 $\bar{\mathbf{a}}$ 中某个长度为 L 的块里的非零条目进行哈希时，无需对该块中的所有非零索引都进行哈希。我们可以跳过大段索引，因为如果 z 是目前生成的最小哈希值，那么下一个产生更小哈希值的索引服从参数为 z 的几何随机变量。我们可以高效地从几何分布中抽样（例如，使用 Python 内置例程），并直接跳到对应索引。可以证明，这种方法对每个块的期望开销仅为 $O(\log L)$。更多细节请参阅 [Manasse 等人, 2010]中的介绍。

自本文最初发布以来，我们了解到加权 MinHash 有更快的实现，可将运行时间降至 $O(|\mathcal{A}| + m \log m)$，这对于待生成草图的向量而言，几乎与其非零元素的数量呈线性关系[Ertl, 2018, Christiani, 2020]。这类方法应该可以改造后用于我们的内积草图应用，不过我们将进一步探索留待未来工作。

哈希函数的选择。在实践中，我们无法从 $\{1, \ldots, n\}$ 到实数范围获得真正均匀随机的哈希函数，因此必须使用近似方法。在实验中，我们采用一种标准的两两独立哈希函数（系数随机的线性函数），将 $\{1, \ldots, n\}$ 映射到 $\{1, \ldots, p\}$，其中 p [是一个 31 位素数 1979]Carter 和 Wegman, [5]。h 然后，我们将 $i(p)/p$ 用作哈希值，它是一个介于 0 和 1 之间的数。由于 h 选取为 31 位，我们可以使用标准的 32 位整数在草图中存储$i()$的值。

5.1 合成数据

我们首先使用合成数据评估我们的方法。我们生成长度为10000 的向量**a** 和**b**，每个向量都有2000 个非零条目。为了模拟表之间具有不同可连接性水平的各种实际场景（见重叠），我们调整**a** 和**b**中非零条目的重叠比例，即同时在两者中非零的条目比例。向量中的非零条目服从取值介于 -1 和1之间的正态随机变量，但随机选取10% 的条目作为离群值，并将其设为介于20 和30之间的随机值。

[5]我们选择使用两两独立哈希函数，是基于加权 MinHash 方法的先前实现 [Wu 等人, 2020] 这些实现也是如此。

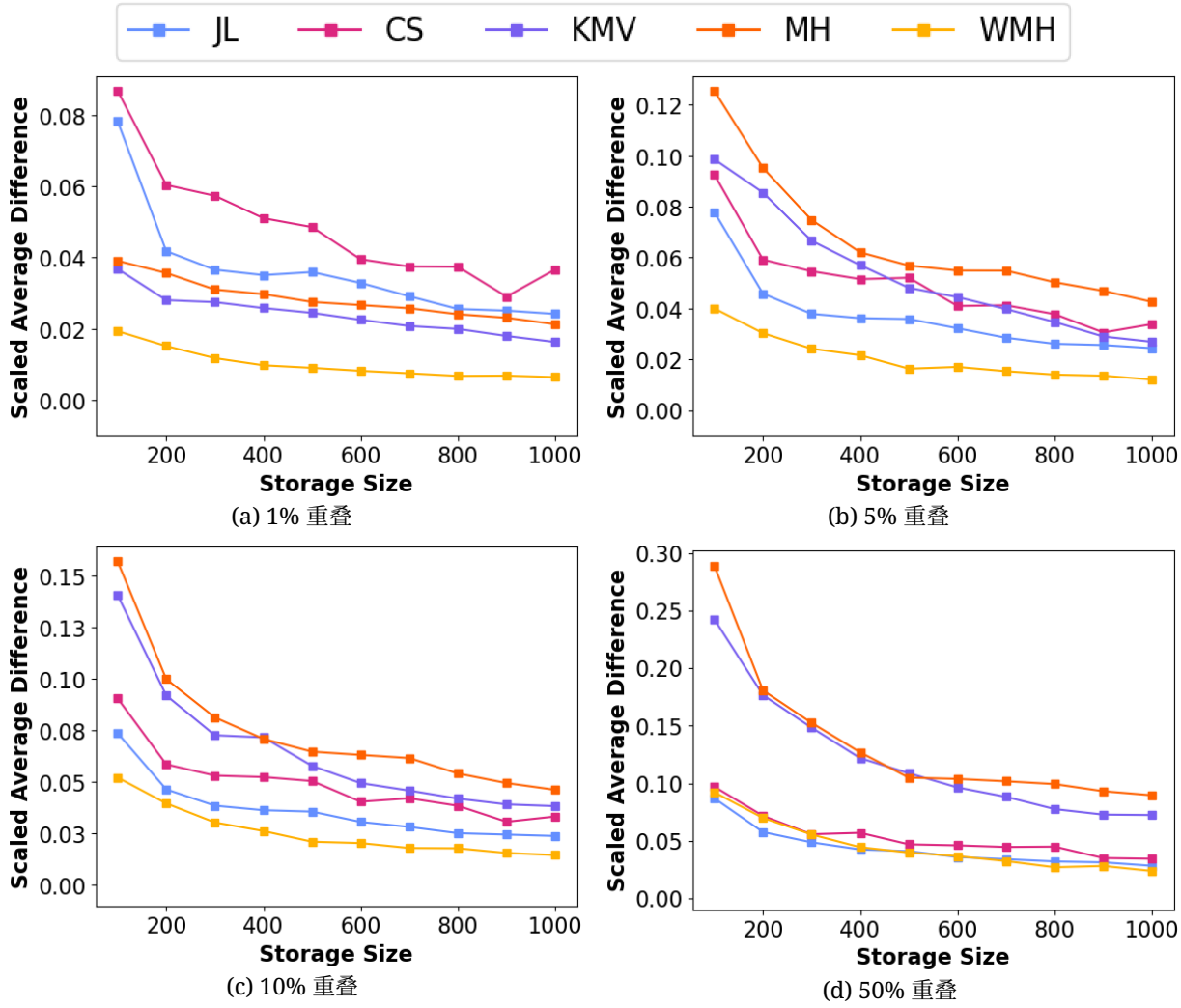


图 4：内积估计（合成数据）。

不同重叠程度下的结果如图 4所示。结果与我们的理论发现高度一致：当重叠较少时，加权 MinHash 的界显著优于线性草图方法的界。因此，当重叠比例为 $\leq 10\%$时，WMH 的表现优于所有其他方法。值得注意的是，在重叠极低（1%）时，基于非加权采样的草图也优于线性草图。但随着重叠增加，定理 2 相较于事实 1 所带来的优势逐渐减小。从图 4(d) 中可以看出这一点：当重叠率为 50% 时，线性草图技术的表现与加权 MinHash 相当。

5.2 真实世界数据

评估重叠和离群值的影响。使用大小为 400,[6] 的草图，我们估计了世界银行集团发布的 56 个数据集中 5000 对数值列之间的内积 [世界银行， 2022]。我们将列归一化，使其范数为1，从而使所有内积的幅值都小于1。我们通过图5中的获胜表格展示结果，表格按不同的重叠率（列）和峰度值（衡量离群值的指标，行）对向量对进行筛选。每个单元格显示具有指定重叠率和峰度值的向量对的平均误差差值（WMH 估计误差减去其他方法的误差）。

蓝色单元格（负差异）对应 WMH 表现优于其他方法的组合，而红色单元格（正差异）则表示其他方法胜出。单元格颜色越深，

[6]该大小是通过经验确定的。我们的目标是模拟真实世界的情形：针对给定应用，必须选定一个固定参数。

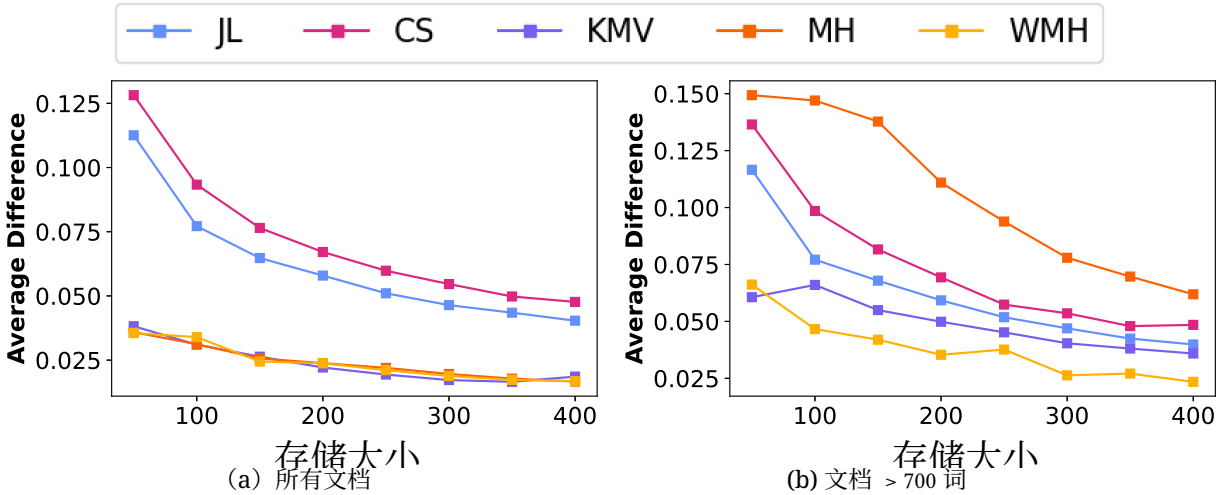
	Kurtosis>0	Kurtosis>10	Kurtosis>50
Overlap<0.05	-0.018	-0.019	-0.018
Overlap<0.1	-0.013	-0.014	-0.014
Overlap<0.25	-0.006	-0.007	-0.010
Overlap<0.5	0.000	-0.001	-0.006
Overlap<0.75	0.005	0.003	-0.003
Overlap<1.0	0.006	0.004	-0.003

（a）WMH 估计误差减去 JL 估计误差。

	Kurtosis>0	Kurtosis>10	Kurtosis>50
Overlap<0.05	-0.012	-0.015	-0.016
Overlap<0.1	-0.009	-0.012	-0.013
Overlap<0.25	-0.004	-0.007	-0.008
Overlap<0.5	-0.008	-0.013	-0.016
Overlap<0.75	-0.014	-0.020	-0.020
Overlap<1.0	-0.020	-0.029	-0.031

（b）WMH 估计误差减去 MH 估计误差。

图5：内积估计（世界银行数据）。不同深浅的蓝色突出显示了 WMH 优于其他方法的组合。



（a）所有文档　　(b) 文档 > 700 词

图6：文本相似度估计（20新组数据集）。请注意，在左侧图中，MH、WMH 和 KMV 的曲线几乎完全重合。

差异越大。峰度较高通常表明存在离群值。根据我们的理论结果，与 JL 或 WMH 方法相比，这会使 MH 等无权重采样方法面临困难。实验结果支持这一点：峰度较高时，WMH 相较于 MH 有显著提升（误差低至 -0.031，而峰度较低时误差至多为 -0.020）。正如 定理 2 所预测的，并且我们的合成实验也表明，对于较低的重叠值，WMH 也明显优于 JL。对于较大的重叠值（大于 0.75），JL 的误差略低（低 0.003 至 0.006）。

这表明，*WMH* 在数据分布未知的应用场景中提供了良好的折中：在许多情况下，它能给出好得多的估计；而在效果不佳时，其估计结果也能与现有草图方法中的最佳结果相媲美。

文档相似度估计。我们还使用以下数据集评估了 WMH 草图在文本相似度估计中的表现：*20*个新闻组数据集 [Mitchell, 1997]。我们将每篇文档表示为一个向量，其中每个元素代表一个词项或两个词项的组合（二元词组），并使用 TF-IDF 权重为其赋值，以表示词项/二元词组的重要性 [Salton等人, 1975]。众所周知，这种设置会产生维度极高的稀疏向量。我们使用余弦相似度作为相似度度量；当向量经过归一化时，它等于内积。我们抽取了 700 篇文档，并估计了超过 200,000 对文档的余弦相似度。 图 6中的结果表明，与先前实验类似，在最坏情况下，*WMH* 的准确度与其他方法相当，但有时会大幅胜出。在这种情况下，对于包含 700 个以上单词的文档，它的表现更好。值得注意的是，即使文档很小，线性投影草图在草图较小时表现也很差；而我们的基于采样的方法则能够显著

在相同的存储预算下获得更高的准确度。最后还要注意，无权重 MinHash (MH) 在长文档上的表现不佳，而加权版本仍能保持良好表现。

致谢。 本研究由 DARPA D3M 计划以及 NSF 的 ISS-2106888 和 CCF-2046235 项目资助。Aline Bessa 获得了 2021 年 CRA/CCC CIFellows 奖资助。Cameron Musco 还获得了 Google Research Scholar Award 的资助。本文所表达的任何观点、发现、结论或建议均属于作者，不一定反映 NSF、DARPA 或其他资助机构的观点。

参考文献

[Achlioptas，2003] Achlioptas，D.（2003）。适用于数据库的随机投影：使用二进制硬币实现Johnson–Lindenstrauss引理。J. Comput. Syst. Sci.，66(4)：671–687。[Alon 等，1999a] Alon，N.、Gibbons，P. B.、Matias，Y. 和 Szegedy，M.（1999a）。在有限存储空间中跟踪连接规模和自连接规模。载于第18届数据库系统原理研讨会（PODS）论文集。[Alon 和 Klartag， 2017] Alon，N. 和 Klartag，B.（2017）。近似内积的最优压缩与降维。载于第58届IEEE计算机科学基础年度研讨会（FOCS）论文集，第639–650页。[Alon 等，1999b] Alon，N.、Matias，Y. 和 Szegedy，M.（1999b）。近似频率矩的空间复杂度。计算机与系统科学杂志，58(1)。[Arriaga 和 Vempala， 2006] Arriaga，R. I. 和 Vempala，S.（2006）。学习的算法理论：鲁棒概念与随机投影。机器学习，63(2)：161–182。[Beyer 等人， 2007] Beyer，K.、Haas，P. J.、Reinwald，B.、Sismanis，Y. 和 Gemulla，R.（2007）。多重集操作下不同值估计的摘要。载于2007年ACM SIGMOD数据管理国际会议论文集，第199–210页。[Blum 等人， 2020] Blum，A.、Hopcroft，J. 和 Kannan，R.（2020）。数据科学基础。剑桥大学出版社。[Broder， 1997] Broder，A.（1997）。文档的相似性与包含率。载于《SEQUENCES 1997压缩与复杂性》会议论文集，第21–29页。[Broder 等人， 1998] Broder，A. Z.、Charikar，M.、Frieze，A. M. 和 Mitzenmacher，M.（1998）。最小值独立置换（扩展摘要）。载于第30届ACM理论计算机年度研讨会（STOC）论文集，第327–336页。[Carter 和 Wegman， 1979] Carter，J. L. 和 Wegman，M. N.（1979）。哈希函数的通用类。计算机与系统科学杂志，18(2)：143–154。[Charikar， 2002] Charikar，M.（2002）。源自舍入算法的相似度估计技术。载于第34届ACM理论计算机年度研讨会（STOC）论文集，第380–388页。[Charikar 等人， 2002] Charikar，M.、Chen，K. 和 Farach-Colton，M.（2002）。在数据流中查找高频项目。载于第29届国际自动机、语言与编程学术研讨会（ICALP）论文集，第693–703页。[Chi and Zhu， 2017] Chi，L. 和 Zhu，X.（2017）。哈希技术：综述与分类。ACM Comput. Surv.，50(1)。[Christiani， 2020] Christiani，T.（2020）。Dartminhash：加权集合的快速草图技术。arXiv：2005.11547。[纽约市， 2022] 纽约市（2022）。纽约市开放数据。https://opendata.cityofnewyork.us/。[Cohen， 2016] Cohen，E.（2016）。Min-Hash草图，第1282–1287页。Springer New York，New York，NY。[Cohen 和 Kaplan，2007] Cohen，E. 和 Kaplan，H.（2007）。使用bottom-k草图汇总数据。载于2007年ACM分布式计算原理研讨会（PODC）论文集，第225–234页。

[Cohen 和 Kaplan， 2013] Cohen，E. 和 Kaplan，H.（2013）。利用协调样本可以完成哪些工作。载于第16届组合优化问题近似算法国际研讨会（APPROX）论文集，第452–467页。[Cormode 等人， 2011] Cormode，G.、Garofalakis，M.、Haas，P. 和 Jermaine，C.（2011）。海量数据摘要：样本、直方图、小波、草图。数据库基础与趋势。NOW出版社。[Dasgupta 和 Gupta， 2003] Dasgupta，S. 和 Gupta，A.（2003）。Johnson–Lindenstrauss定理的一个初等证明。随机结构与算法，22(1)：60–65。[Ertl， 2018] Ertl，O.（2018）。Bagminhash——加权集合的最小值哈希算法。载于第24届ACM SIGKDD知识发现与数据挖掘国际会议（KDD）论文集，第1368–1377页。[Fernandez 等人， 2019] Fernandez，R. C.、Min，J.、Nava，D. 和 Madden，S.（2019）。Lazo：联合估计Jaccard相似度和包含率的基数估计方法。载于第35届IEEE国际数据工程会议（ICDE）论文集，第1190–1201页。[Flajolet 和 Martin， 1985] Flajolet，P. 和 Martin，G. N.（1985）。用于数据库应用的概率计数算法。J. Comput. Syst. Sci.，31(2)：182–209。[Gionis 等人， 1999] Gionis，A.、Indyk，P. 和 Motwani，R.（1999）。通过哈希在高维空间中进行相似性搜索。载于第25届超大型数据库国际会议论文集，第518–529页。[Gollapudi 和 Panigrahy， 2006] Gollapudi，S. 和 Panigrahy，R.（2006）。利用层次主题提取中的非对称性。载于第15届ACM信息与知识管理国际会议（CIKM）论文集，第475–482页。[Haeupler 等人， 2014] Haeupler，B.、Manasse，M. 和 Talwar，K.（2014）。让一致加权采样变得快速、小巧且简单。arXiv：1410.4266。[Heintze， 1996] Heintze，N.（1996）。可扩展的文档指纹识别。载于USENIX电子商务研讨会。[Ioe， 2010] Ioe，S.（2010）。改进的一致采样、加权 MinHash 与L1草图。载于2010年IEEE国际数据挖掘会议（ICDM）论文集，第246–255页。[Jacques， 2015] Jacques，L.（2015）。量化Johnson–Lindenstrauss引理：布丰投针问题的发现。IEEE Trans. Inf.，61(9)：5012–5027。[Kane 等人， 2010] Kane，D. M.、Nelson，J. 和 Woodru，D. P.（2010）。不同元素问题的最优算法。载于第29届数据库系统原理研讨会（PODS）论文集，第41–52页。[Kanter 和 Veeramachaneni， 2015] Kanter，J. M. 和 Veeramachaneni，K.（2015）。深度特征合成：迈向数据科学工作的自动化。载于2015年IEEE数据科学与高级分析国际会议（DSAA）论文集，第1–10页。IEEE。[Larsen 和 Nelson， 2017] Larsen，K. G. 和 Nelson，J.（2017）。Johnson–Lindenstrauss引理的最优性。载于第58届IEEE计算机科学基础年度研讨会（FOCS）论文集，第633–638页。[Larsen 等人， 2021] Larsen，K. G.、Pagh，R. 和 Tětek，J.（2021）。CountSketch、特征哈希与三者中位数。载于第38届国际机器学习会议（ICML）论文集，第6011–6020页。[Lehmberg 等人， 2015] Lehmberg，O.、Ritze，D.、Ristoski，P.、Meusel，R.、Paulheim，H. 和 Bizer，C.（2015）。曼海姆搜索连接引擎。网络语义学杂志，35：159–166。[Li， 2017] Li，P.（2017）。线性化GMM核与归一化随机傅里叶特征。载于第23届ACM SIGKDD知识发现与数据挖掘国际会议（KDD）论文集，第315–324页。[Li等人， 2006] Li，P.、Church，K. 和 Hastie，T.（2006）。稀疏数据的条件随机采样：一种基于草图的采样技术。载于神经信息处理系统进展19（NeurIPS），第19卷。

[Li 和 König， 2010] Li, P. 和 König, A. C.（2010）。b 位最小值哈希。载于第19届国际万维网大会（WWW）论文集。[Li 等人， 2016] Li, P.、Mitzenmacher, M. 和 Slawski, M.（2016）。量化随机投影与余弦相似度的非线性估计。载于《神经信息处理系统进展》第29卷（NeurIPS），第29卷。[Li 等人，2012] Li, P.、Owen, A. 和 Zhang, C.-h.（2012）。单排列哈希。载于《神经信息处理系统进展》第25卷（NeurIPS）。[Li 等人，2011] Li, P.、Shrivastava, A.、Moore, J. 和 König, A.（2011）。面向大规模学习的哈希算法。《神经信息处理系统进展》第24卷（NeurIPS），24。[Manasse 等人， 2010] Manasse, M.、McSherry, F. 和 Talwar, K.（2010）。一致加权采样。技术报告 MSR-TR-2010-73。[Manber， 1994] Manber, U.（1994）。在大型文件系统中查找相似文件。载于 USENIX 1994年冬季技术会议。[Mitchell， 1997] Mitchell, T.（1997）。20新组数据集。https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_20newsgroups.html。[Pagh 等人，2014] Pagh, R.、Stöckel, M. 和 Woodru, D. P.（2014）。最小值哈希对于概括集合交集是否最优？载于第33届数据库系统原理研讨会（PODS）论文集，第109–120页。[Rusu 和 Dobra， 2008] Rusu, F. 和 Dobra, A.（2008）。用于连接估计大小的草图。ACM数据库系统汇刊（TODS），33(3)：1–46。[Salton 等人， 1975] Salton, G.、Wong, A. 和 Yang, C.-S.（1975）。用于自动索引的向量空间模型。ACM通讯，18(11)：613–620。[Santos 等人， 2021] Santos, A.、Bessa, A.、Chirigati, F.、Musco, C. 和 Freire, J.（2021）。近似连接-相关性查询的相关性草图。载于2021年ACM SIGMOD国际数据管理大会论文集，第199–210页。[Santos 等人， 2022] Santos, A.、Bessa, A.、Musco, C. 和 Freire, J.（2022）。用于相关数据集搜索的基于草图的索引。载于第38届IEEE国际数据工程会议（ICDE）论文集，第2928–2941页。IEEE。[Shrivastava， 2016] Shrivastava, A.（2016）。简单高效的加权最小值哈希。载于《神经信息处理系统进展》第29卷（NeurIPS），第1506–1514页。[Shrivastava 和 Li， 2014] Shrivastava, A. 和 Li, P.（2014）。为 Minhash辩护，反对Simhash。载于第17届国际人工智能与统计会议（AISTATS）论文集。[世界银行， 2022] 世界银行（2022）。世界银行集团财务。[Wu 等人， 2020] Wu, W.、Li, B.、Chen, L.、Gao, J. 和 Zhang, C.（2020）。加权Minhash算法综述。IEEE知识与数据工程汇刊，第1–1页。[Wu 等人， 2019] Wu, W.、Li, B.、Chen, L.、Zhang, C. 和 Yu, P. S.（2019）。重新审视改进的一致加权采样。IEEE知识与数据工程汇刊，31(12)：2332–2345。[Yang 等人， 2019] Yang, Y.、Zhang, Y.、Zhang, W. 和 Huang, Z.（2019）。Gb-kmv：用于近似包含相似性搜索的增强型kmv草图。载于第35届IEEE国际数据工程会议（ICDE）论文集，第458–469页。IEEE。[Zhu 等人，2019] Zhu, E.、Deng, D.、Nargesian, F. 和 Miller, R. J.（2019）。Josie：用于在数据湖中查找可连接表的重叠集合相似性搜索。载于2019年ACM SIGMOD国际数据管理大会论文集，第847–864页。[Zhu 等人， 2016] Zhu, E.、Nargesian, F.、Pu, K. Q. 和 Miller, R. J.（2016）。Lsh ensemble：互联网规模的领域搜索。《VLDB基金会会刊》，9(12)。

A 补充证明

A.1 无权 MinHash 分析

本节将给出定理 4的完整证明。

定理 定理 4的证明。 令 $\overline{\mathcal{F}}(H_\mathrm{a}, H_\mathrm{b})$ 表示算法 2 中的估计量算法 2。最终，我们将在 $\mathcal{F}$ 定理 4 中将其设为 F，但重复 $O(\log(1/\delta))$ 次，以获得失败概率 $1-\delta$。

我们首先着重证明，$\mathrm{F}(H_\mathrm{a}, H_\mathrm{b})$ 以概率 $\geq 2/3$ 达到误差 $\epsilon \cdot c^2 \cdot \sqrt{\max(|\mathrm{A}|, |\mathrm{B}|) \cdot |\mathrm{A} \cap \mathrm{B}|}$。为此，令 $\mathrm{F}^*(H_\mathrm{a}, H_\mathrm{b})$ 表示另一个理想化估计量，其中我们将算法 2 第 1 行的 U 替换为真实的并集大小 $U = |\mathrm{A} \cup \mathrm{B}|$:

$$\mathcal{F}^*(H_\mathrm{a}, H_\mathrm{b}) = \frac{U}{m} \sum_{i=1}^{m} \mathbb{1}\left[H_\mathrm{a}^{hash}[i] = H_\mathrm{b}^{hash}[i]\right] \cdot H_\mathrm{a}^{val}[i] \cdot H_\mathrm{b}^{val}[i].$$

我们首先分析 $\mathcal{F}^*$，然后证明 $\mathcal{F}$ 能得到本质上同样好的估计。正如第 3 节 第 3 节中所确立的，利用 事实 3的性质，我们有

$$\mathbb{E}\left[\mathcal{F}^*(H_\mathrm{a}, H_\mathrm{b})\right] = U \cdot \frac{1}{|\mathcal{A} \cup \mathcal{B}|} \cdot \langle \mathbf{a}, \mathbf{b} \rangle = \langle \mathbf{a}, \mathbf{b} \rangle.$$

因此，我们转而对估计量的方差进行界定。定义随机变量 $Z_i = \mathbb{1}\left[H_\mathrm{a}^{hash}[i] = H^{hash}{}_\mathrm{b}[i]\right] . H^{val}{}_\mathrm{a} \quad [i] . \ H^{val}{}_\mathrm{b}[i]$，并注意到 $\mathcal{F}^*(H_\mathrm{a}, H_\mathrm{b}) = \frac{U}{m} \sum_{i=1}^{m} Z_i$。 根据 事实 3，我们有:

$$Z_i = \begin{cases} 0 & \text{with probability } 1 - \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|} \\ \mathbf{a}[j]\mathbf{b}[j] & \text{with probability } \frac{1}{|\mathcal{A} \cup \mathcal{B}|} \text{ for all } j \in \mathcal{A} \cap \mathcal{B}. \end{cases}$$

由于每个 Z_i 彼此独立，我们可以给出如下界:

$$\mathrm{Var}\left[\mathcal{F}^*(H_\mathrm{a}, H_\mathrm{b})\right] = \frac{U^2}{m^2} \sum_{i=1}^{m} \mathrm{Var}\left[Z_i\right].$$

根据我们的假设，对于所有 k，$\mathbf{a}[k]$，$\mathbf{b}[k] \leq c$，我们有

$$\mathrm{Var}\left[Z_i\right] \leq \mathbb{E}\left[Z_i^2\right] = \sum_{j \in \mathcal{A} \cap \mathcal{B}} \frac{1}{|\mathcal{A} \cup \mathcal{B}|} \cdot \mathbf{a}[j]^2 \mathbf{b}[j]^2 \leq c^4 \cdot \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|},$$

对于所有 Z_i。因此我们得出 $\mathrm{Var}\left[\mathrm{F}^*(H_\mathrm{a}, H_\mathrm{b})\right] \leq \frac{1}{m} \cdot c^4 \cdot |\mathrm{A} \cap \mathrm{B}||\mathrm{A} \cup \mathrm{B}|$.。接着，我们将期望和方差的界代入切比雪夫不等式。若 $m = O(1/\epsilon^2)$，则我们可以得出，以概率 $\geq 5/6$,

$$|\mathcal{F}^*(H_\mathrm{a}, H_\mathrm{b}) - \langle \mathbf{a}, \mathbf{b} \rangle| \leq \epsilon \cdot c^2 \sqrt{|\mathcal{A} \cap \mathcal{B}||\mathcal{A} \cup \mathcal{B}|}. \tag{1}$$

证明几乎完成了；我们只需将此界推广到非理想化估计量 $\mathcal{F} = \frac{U}{\tilde{U}} \cdot \mathcal{F}^*$。我们通过注意到 U 是 $\tilde{U}$ 的良好近似来做到这一点。具体而言，根据 引理 1，取 $\delta = 1/6$，当 $m = O(1/\epsilon^2)$ 时，$(1-\epsilon)U \leq \tilde{U} \leq (1+\epsilon)U$，以概率 $\geq 5/6$。由此可得

$$(1-\epsilon)\mathcal{F}^*(H_\mathrm{a}, H_\mathrm{b}) \leq \overline{\mathcal{F}}(H_\mathrm{a}, H_\mathrm{b}) \leq (1+\epsilon)\mathcal{F}^*(H_\mathrm{a}, H_\mathrm{b}). \tag{2}$$

根据并集界，以至少 2/3 的概率，(1) 和 (2) 同时成立。最后，根据三角不等式以及 $\langle a, b \rangle \leq c^2 |\mathcal{A} \cap \mathcal{B}| \leq c^2 \sqrt{|\mathcal{A} \cap \mathcal{B}||\mathcal{A} \cup \mathcal{B}|}$ 可得:

$$|\overline{\mathcal{F}}(H_\mathrm{a}, H_\mathrm{b}) - \langle \mathbf{a}, \mathbf{b} \rangle| \leq 3\epsilon \cdot c^2 \cdot \sqrt{|\mathcal{A} \cap \mathcal{B}||\mathcal{A} \cup \mathcal{B}|}.$$

注意到 $|\mathrm{A} \cap \mathrm{B}||\mathrm{A} \cup \mathrm{B}| \leq 2\max(|\mathrm{A}|, |\mathrm{B}|) \cdot |\mathrm{A} \cap \mathrm{B}|$，并将 ϵ 乘以一个常数因子进行调整，我们因此得到：当 $m = O(1/\epsilon^2)$ 时，$\mathrm{F}(H_\mathrm{a}, H_\mathrm{b})$ 以至少 2/3 的概率满足定理 4 的保证。若要将成功概率提升至 $1-\delta$，我们可以使用定理 2 证明中完全相同的中位数技巧：不再为输入 a、b 计算单个草图对 $H_\mathrm{a}, H_\mathrm{b}$，而是拼接 $O(\log(1/\delta))$ 个草图，每个草图都使用独立的随机种子构造。如果我们将 F 应用于每一对独立草图，并返回 ⟨a 、 b⟩ 的中位数估计，则以至少 $1-\delta$ 的概率，它将满足我们所需的保证。□

A.2 加权 MinHash 分析

在本节中，我们完成对 算法 5（在第 4节中介绍）的分析，该算法给出了我们的主要结果，即定理 2。我们首先正式证明事实 5，它是事实 3的加权对应形式。

事实 5 的证明。设 $A = \{i : \tilde{a}[i] \neq 0\}$ 和 $B = \{i : b[i] \neq 0\}$。由于 $\tilde{a}$、b 各由 n 块、每块含 L 个元素组成，且 i^{th} 块中前 $\tilde{a}[i]^2 \cdot L$ 个条目和 $\tilde{b}[i]^2 \cdot L$ 个条目为非零，因此有以下等式：

$$|\bar{\mathcal{A}} \cap \bar{\mathcal{B}}| = L \cdot \sum_{j=1}^{n} \min(\tilde{a}[j]^2, \tilde{b}[j]^2) \tag{3}$$

$$|\bar{\mathcal{A}} \cup \bar{\mathcal{B}}| = L \cdot \sum_{j=1}^{n} \max(\tilde{a}[j]^2, \tilde{b}[j]^2). \tag{4}$$

由于 $W_a^{hash}[i]$ 和 $W^{hash}{}_b[i]$ 恰好是由 $\tilde{a}$ $\tilde{b}$ 构造的无权重 MinHash 草图，根据 事实 3 的断言（1），$W_a^{hash}[i] = W^{hash}{}_b[i]$ 的概率为 $|\bar{\mathcal{A}} \cap \bar{\mathcal{B}}| / |\bar{\mathcal{A}} \cup \bar{\mathcal{B}}| = \bar{J}$。由此得到断言（1）。

为证明断言（2），我们注意到，这等价于声称：无论 $W_a^{hash}[i] = W^{hash}b[i]$, $W^{val}a = \tilde{a}[j]$ and $W^{val}b = \tilde{b}[j]$ 是否成立，对于某个共同的 $j \in A \cap B$，以概率 $\min(\tilde{a}_{[j]^2}, \tilde{b}_{[j]^2}) / \sum_{i=1}^{n} \max(\tilde{a}_{[j]^2}, \tilde{b}_{[j]^2})$ 都有。为证明这一结论，我们利用事实 3：对于任意 $\ell \in \bar{A} \cap \bar{B}$，$W_a^{hash}[i] = W^{hash}b[i] = h^i(\ell)$，$W_a^{val}[i] = \tilde{a}[\ell]$，且 $W^{val}b[i] = \tilde{b}[\ell]$ 的概率为 $|\bar{A} \cup \bar{B}|^{-1} = L \sum_{k=1}^{n} \max(a_{[k]^2}, \tilde{b}_{[k]^2})^{-1}$。现在，根据构造（算法 3 的第 3 行），只要 ℓ 位于 $\tilde{a}$ 中长度为 j^{th}length L 的条目块内，就有 $\tilde{a}[\ell] = \tilde{a}[j]$ 和 $\tilde{b}[\ell] = \tilde{b}[j]$。对于给定的 j，使得 $\tilde{a}[\ell] = \tilde{a}[j]$、$\tilde{b}[\ell] = \tilde{b}[j]$ 成立的 ℓ 的值的数量恰好为 $L \cdot \min(\tilde{a}[j]^2, \tilde{b}[j]^2)$。因此，对这些条目求和可得，$W^{hash}a[i] = W^{hash}b[i]$，$W_a^{val}[i] = \tilde{a}[j]$，且 $W^{val}b[i] = \tilde{b}[j]$ 的概率为 $\min(\tilde{a}_{[j]^2}, \tilde{b}_{[j]^2}) / \sum_{k=1}^{n} \max(\tilde{a}_{[k]^2}, \tilde{b}_{[k]^2})$。□

离散向量的分析。 接下来，为了证明 定理 2，我们证明一个适用范围较窄的中间结果，引理 2，它仅适用于这样的向量：将其缩放至单位范数后，其条目已经是$1/L$ 的整数倍，其中离散化参数 L固定。此时，舍入过程在算法 3 中不起作用：它只是原样返回$\mathbf{a}/\|\mathbf{a}\|$。作此假设可以简化分析。之后，我们将引入舍入误差分析，以获得适用于任意向量的结果。

引理 *2*。*考虑任意整数离散化参数 L、精度参数 $\epsilon \in (0, 1)$，以及满足以下条件的 $a, b \in \mathbb{R}^n$：对于所有 i，$a_{[i]^2}$ 和 $b_{[i]^2}$ 都是 $1/L$ 的整数倍。使用样本量 $m = O\left(1/\epsilon^2\right)$ 和离散化参数L运行时，算法 3 返回草图W_a和 W_b。令 F 表示算法 5 的估计过程，则以至少 2/3 的概率，*

$$|\mathcal{F}(W_a, W_b) - \langle \mathbf{a}, \mathbf{b} \rangle| \leq \epsilon \max\left(\|\mathbf{a}_{\mathcal{I}}\|\|\mathbf{b}\|, \|\mathbf{a}\|\|\mathbf{b}_{\mathcal{I}}\|\right).$$

这里，$\mathcal{I} = \{i : \mathbf{a}[i] \neq 0 \text{ 和} \mathbf{b}[i] \neq 0\}$ 是$\mathbf{a}$和$\mathbf{b}$的支撑集的交集，而$\mathbf{a}_{\mathcal{I}}$,$\mathbf{b}_{\mathcal{I}}$ 表示将$\mathbf{a}$ 和$\mathbf{b}$ 限制在$\mathcal{I}$中的索引上所得的向量。

注意，引理 2 也弱于定理 2，因为它只能以常数概率（2/3）$1 - \delta$给定任意 δ时得到准确解。这同样是为了简化分析；稍后我们将展示如何使用标准的“中位数技巧”来提高成功概率至 $1 - \delta$ [Cormode 等, 2011, Larsen 等, 2021]。

证明。根据假设，由于$\mathbf{a}/\|\mathbf{a}\|$ 和$\mathbf{b}/\|\mathbf{b}\|$ 的平方项都是$1/L$ 的整数倍，在算法 3 的第 2 行中，$\text{Round}(\mathbf{a}/\|\mathbf{a}\|, L)$只是将 $\tilde{a} = \mathbf{a}/\|\mathbf{a}\|$。类似地，它将 $b = \mathbf{b}/\|\mathbf{b}\|$。令 $A = \{i : \mathbf{a}[i] \neq 0\}$ 和 $B = \{i : \mathbf{b}[i] \neq 0\}$ 分别表示 a 和 b 的支撑集。我们有 $I = A \cap B$。

归约到单位向量。 我们首先注意到，为了证明该定理，只需考虑单位向量 $\tilde{a}$ 与 $\mathbf{b}$之间的内积 $\tilde{}$。具体而言，我们将证明：

$$\left|\frac{\mathcal{F}(W_a, W_b)}{\|\mathbf{a}\|\|\mathbf{b}\|} - \langle \tilde{\mathbf{a}}, \tilde{\mathbf{b}} \rangle\right| \leq \epsilon \sqrt{\sum_{i \in \mathcal{A} \cap \mathcal{B}} \max(\tilde{\mathbf{a}}[i]^2, \tilde{\mathbf{b}}[i]^2) \sum_{i=1}^{n} \max(\tilde{\mathbf{a}}[i]^2, \tilde{\mathbf{b}}[i]^2)}. \tag{5}$$

利用 $\|\tilde{\mathbf{a}}\|^2 + \|\tilde{\mathbf{b}}\|^2 = 2$，由于 $\tilde{\mathbf{a}}$ $\tilde{\mathbf{b}}$是单位向量，我们有：

$$\begin{aligned}
&\sqrt{\sum_{i\in\mathcal{A}\cap\mathcal{B}} \max(\tilde{\mathbf{a}}[i]^2, \tilde{\mathbf{b}}[i]^2) \sum_{i=1}^{n} \max(\tilde{\mathbf{a}}[i]^2, \tilde{\mathbf{b}}[i]^2)} \\
&\quad \le \sqrt{\left(\|\tilde{\mathbf{a}}_{\mathcal{I}}\|^2 + \|\tilde{\mathbf{b}}_{\mathcal{I}}\|^2\right)\left(\|\tilde{\mathbf{a}}\|^2 + \|\tilde{\mathbf{b}}\|^2\right)} \\
&\quad = \sqrt{2\left(\|\tilde{\mathbf{a}}_{\mathcal{I}}\|^2 + \|\tilde{\mathbf{b}}_{\mathcal{I}}\|^2\right)} = \sqrt{2\left(\frac{\|\mathbf{a}_{\mathcal{I}}\|^2}{\|\mathbf{a}\|^2} + \frac{\|\mathbf{b}_{\mathcal{I}}\|^2}{\|\mathbf{b}\|^2}\right)}.
\end{aligned}$$

因此，将两边同时乘以（5） $\|\mathbf{a}\|\|\mathbf{b}\|$，得到：

$$\begin{aligned}
|\mathcal{F}(W_{\mathbf{a}}, W_{\mathbf{b}}) - \langle \mathbf{a}, \mathbf{b}\rangle| &\le \epsilon\sqrt{2}\|\mathbf{a}\|\|\mathbf{b}\| \cdot \sqrt{\frac{\|\mathbf{a}_{\mathcal{I}}\|^2}{\|\mathbf{a}\|^2} + \frac{\|\mathbf{b}_{\mathcal{I}}\|^2}{\|\mathbf{b}\|^2}} \\
&= \epsilon\sqrt{2}\sqrt{\|\mathbf{a}_{\mathcal{I}}\|^2\|\mathbf{b}\|^2 + \|\mathbf{b}_{\mathcal{I}}\|^2\|\mathbf{a}\|^2} \\
&\le 2\epsilon \cdot \max\left(\|\mathbf{a}_{\mathcal{I}}\|\|\mathbf{b}\|, \|\mathbf{b}_{\mathcal{I}}\|\|\mathbf{a}\|\right).
\end{aligned}$$

最后一个不等式成立，是因为总和至多为最大值的两倍。将 ϵ 调整一个常数因子，即可得到所需的界 引理 2。因此，我们转而证明（5）。

单位向量的分析。我们先分析算法 5 计算的估计量的一个理想化版本，其中 M 被替换为精确的加权并集大小 $M = \sum_{i=1}^{n}$ 最大值($\tilde{\mathbf{a}}[i]^2$， $\tilde{\mathbf{b}}[i]^2$)。具体而言，定义：

$$\mathcal{F}^* = \frac{M}{m}\sum_{i=1}^{m} \mathbb{1}\left[W_{\mathbf{a}}^{hash}[i] = W_{\mathbf{b}}^{hash}[i]\right] \cdot \frac{W_{\mathbf{a}}^{val}[i] \cdot W_{\mathbf{b}}^{val}[i]}{q_i}, \tag{6}$$

其中 $q_i = \min\left(W^{val}{}_{\mathbf{a}}[i]^2,\ W^{val}{}_{\mathbf{b}}[i]^2\right)$如算法 5 第 1 行所示。

$\tilde{}$我们首先证明 $\mathbb{E}[\mathrm{F}^*] = \langle \tilde{}\mathbf{a}、\mathbf{b}\rangle$，然后界定 F^*的方差。对于每个 $i \in \{1, \ldots\ldots, m\}$，定义随机变量 Z_i 为

$$Z_i = \mathbb{1}\left[W_{\mathbf{a}}^{hash}[i] = W_{\mathbf{b}}^{hash}[i]\right] \cdot \frac{W_{\mathbf{a}}^{val}[i] \cdot W_{\mathbf{b}}^{val}[i]}{q_i}.$$

回想一下， $\bar{J} = \frac{\sum_{j=1}^{n} \min(\tilde{}\mathbf{a}[j]^2\tilde{\mathbf{b}}[j]^2)}{\sum_{j=1}^{n} \max(\tilde{}\mathbf{a}[j]^2, \tilde{}\mathbf{b}[j]^2)}$是 加权 Jaccard 相似度，介于 $\tilde{}\mathbf{a}$和 ，应用 事实 5 我们有 ve:

$$Z_i = \begin{cases} 0 & \text{with probability } 1 - \bar{J} \\ \frac{\tilde{\mathbf{a}}[j]\tilde{\mathbf{b}}[j]}{\min(\tilde{\mathbf{a}}[j]^2, \tilde{\mathbf{b}}[j]^2)} & \text{with probability } \frac{\min(\tilde{\mathbf{a}}[j]^2, \tilde{\mathbf{b}}[j]^2)}{\sum_{k=1}^{n} \max(\tilde{\mathbf{a}}[k]^2, \tilde{\mathbf{b}}[k]^2)} \\ & \text{for all } j \in \mathcal{A}\cap\mathcal{B}. \end{cases}$$

因此， $\mathbb{E}[Z_i] = \frac{\langle \tilde{}\mathbf{a}}{\sum_{k=1}^{n}}$ $\tilde{}$ $\max(\mathbf{a}[k]^2, \tilde{}\mathbf{b}[k]^2) = \frac{\langle \tilde{}\mathbf{a}, \rangle \mathbf{b}}{M}$。 由于 $\mathrm{F}^* = \frac{M}{m}\sum_{i=1}^{m} Z_i$， 根据期望的线性性可得：

$$\mathbb{E}[\mathcal{F}^*] = \frac{M}{m}\sum_{i=1}^{m} \mathbb{E}[Z_i] = \langle \tilde{\mathbf{a}}, \tilde{\mathbf{b}}\rangle. \tag{7}$$

接下来我们界定 F^*的方差。对于每个 Z_i， 有：

$$\begin{aligned}
\mathrm{Var}[Z_i] &\le \sum_{j\in\mathcal{A}\cap\mathcal{B}} \frac{\min(\tilde{\mathbf{a}}[j]^2, \tilde{\mathbf{b}}[j]^2)}{\sum_{k=1}^{n} \max(\tilde{\mathbf{a}}[k]^2, \tilde{\mathbf{b}}[k]^2)} \cdot \frac{\tilde{\mathbf{a}}[j]^2\tilde{\mathbf{b}}[j]^2}{\min(\tilde{\mathbf{a}}[j]^2, \tilde{\mathbf{b}}[j]^2)^2} \\
&= \sum_{j\in\mathcal{A}\cap\mathcal{B}} \frac{\max(\tilde{\mathbf{a}}[j]^2, \tilde{\mathbf{b}}[j]^2)}{\sum_{k=1}^{n} \max(\tilde{\mathbf{a}}[k]^2, \tilde{\mathbf{b}}[k]^2)} \\
&= \frac{\sum_{j\in\mathcal{A}\cap\mathcal{B}} \max(\tilde{\mathbf{a}}[j]^2, \tilde{\mathbf{b}}[j]^2)}{\sum_{k=1}^{n} \max(\tilde{\mathbf{a}}[k]^2, \tilde{\mathbf{b}}[k]^2)} = \frac{\sum_{j\in\mathcal{A}\cap\mathcal{B}} \max(\tilde{\mathbf{a}}[j]^2, \tilde{\mathbf{b}}[j]^2)}{M}
\end{aligned}$$

由于每个 Z_i 彼此独立，因此可得：

$$\begin{aligned}\mathrm{Var}[\mathcal{F}^*] &= \frac{M^2}{m^2}\sum_{i=1}^{m}\mathrm{Var}\left[Z_i\right] \\ &\le \frac{1}{m}\sum_{j\in\mathcal{A}\cap\mathcal{B}}\max(\tilde{\mathbf{a}}[j]^2,\tilde{\mathbf{b}}[j]^2)\cdot\sum_{j=1}^{n}\max(\tilde{\mathbf{a}}[j]^2,\tilde{\mathbf{b}}[j]^2).\end{aligned} \tag{8}$$

结合 (7) 和 (8)，利用切比雪夫不等式可得，当 $m = O(1/\epsilon^2)$ 时，以至少5/6的概率：

$$\left|\mathcal{F}^* - \langle\tilde{\mathbf{a}},\tilde{\mathbf{b}}\rangle\right| \le \epsilon\sqrt{\sum_{j\in\mathcal{A}\cap\mathcal{B}}\max(\tilde{\mathbf{a}}[j]^2,\tilde{\mathbf{b}}[j]^2)\sum_{j=1}^{n}\max(\tilde{\mathbf{a}}[j]^2,\tilde{\mathbf{b}}[j]^2)}. \tag{9}$$

我们希望将这个界从理想化估计量 $\mathcal{F}^*$ 扩展到实际的估计量 $\mathcal{F}$，它等于 $\frac{M}{\tilde{M}}\cdot\mathcal{F}^*$。为此，我们利用 $\tilde{M}$ 是 M 的良好近似这一点。如 第 4，$\tilde{M}$ 是 $\frac{1}{L}$ 节所述，这是因为它恰好等于将不同元素估计器应用于扩展向量 $\mathcal{A}$ 和 $\mathcal{B}$ 的支撑集所得到的结果，其中扩展向量为 $\tilde{\mathbf{a}},\tilde{\mathbf{b}}$。由 (4) 和引理 1可知，对于 $m = O(1/\epsilon^2)$，

$$(1-\epsilon)M \le \tilde{M} \le (1+\epsilon)M,$$

以至少 5/6 的概率。因此：

$$(1-\epsilon)\mathcal{F}^* \le \frac{\mathcal{F}(W_{\mathbf{a}},W_{\mathbf{b}})}{\|\mathbf{a}\|\|\mathbf{b}\|} \le (1+\epsilon)\mathcal{F}^*. \tag{10}$$

由并集界可知，以至少 2/3 的概率，(9) 和 (10) 同时成立。最后，根据柯西-施瓦茨不等式，

$$\langle\tilde{\mathbf{a}},\tilde{\mathbf{b}}\rangle \le \sqrt{\sum_{j\in\mathcal{A}\cap\mathcal{B}}\max(\tilde{\mathbf{a}}[j]^2,\tilde{\mathbf{b}}[j]^2)\sum_{j=1}^{n}\max(\tilde{\mathbf{a}}[j]^2,\tilde{\mathbf{b}}[j]^2)}.$$

结合（9）和（10），根据三角不等式可得

$$\begin{aligned}&\left|\frac{\mathcal{F}(W_{\mathbf{a}},W_{\mathbf{b}})}{\|\mathbf{a}\|\|\mathbf{b}\|} - \langle\tilde{\mathbf{a}},\tilde{\mathbf{b}}\rangle\right| \\ &\le 3\epsilon\sqrt{\sum_{j\in\mathcal{A}\cap\mathcal{B}}\max(\tilde{\mathbf{a}}[j]^2,\tilde{\mathbf{b}}[j]^2)\sum_{j=1}^{n}\max(\tilde{\mathbf{a}}[j]^2,\tilde{\mathbf{b}}[j]^2)}.\end{aligned}$$

将 ϵ 调整1/3 倍即可证明引理 2。 □

连续向量的舍入。有了引理 2 之后，我们通过分析算法 5 中舍入步骤的影响，完成定理 2的证明。在引理 3中，我们证明，若 L 设定为约 n^6/ϵ^2，则可以界定这一步对内积估计准确度的影响。形式化地，我们有：

引理 3（舍入）。任取 a、$b \in \mathbb{R}^n$ 及离散化参数 L。令 $\tilde{a} = Round(a/\|a\|, L)$ 和 $\tilde{b} = Round(b/\|b\|, L)$，如算法 3 第 2 行所示。令 $a' = \|a\|\cdot\tilde{a}$ 和 $b' = \|b\|\cdot b$，并令 B 表示 $B = max\left(\|a_I\|\|b\|, \|a\|\|b_I\|\right)$。

1. $\mathbf{a}'$ 和 $\mathbf{b}'$ 满足引理 2的假设：对于所有 i，$\frac{\mathbf{a}'[i]^2}{\|\mathbf{a}'\|^2}$ 和 $\frac{\mathbf{b}'[i]^2}{\|\mathbf{b}'\|^2}$ 都是$1/L$的整数倍。

2. 对于任意离散化参数 L、草图大小 m和随机种子 s，算法 3 对以下对象生成相同的输出：$\mathbf{a}$、$\mathbf{b}$以及$\mathbf{a}'$、$\mathbf{b}'$。即，$W_{\mathbf{a}} = W_{\mathbf{a}'}$ 以及 $W_{\mathbf{b}} = W_{\mathbf{b}'}$。

3. 对于 $L \ge 9n^6/\epsilon^2$，$|\langle\mathbf{a},\mathbf{b}\rangle - \langle\mathbf{a}',\mathbf{b}'\rangle| \le \epsilon B$.

4. 对于 $L \geq n^3$, 最大值 $\left(\|\mathbf{a}'_I\|\|\mathbf{b}'\|, \|\mathbf{a}'\|\|\mathbf{b}'_I\|\right) \leq 2B$.

证明。 我们依次证明引理的四个结论。对于前两个结论，我们关注$\mathbf{a}$ 和$\mathbf{a}'$。对$\mathbf{b}$ 和$\mathbf{b}'$也有相同的结论。

结论 1：对于所有 i，$\mathbf{a}'_{\|}[i]^2 / \|\mathbf{a}'_{\|}\|^2 \mathbf{a}$ 是 $1/L$ 的整数倍。首先注意到，$\tilde{\mathbf{a}} = \text{Round}(\mathbf{a}/\|\mathbf{a}\|, L)$ 是一个单位向量。这由算法 4 的第 3 行保证。因此，$\|\mathbf{a}'\| = \|\mathbf{a}\| \cdot \|\tilde{\mathbf{a}}\| = \|\mathbf{a}\|$，且 $\mathbf{a}'_{\|}[i]^2 / \|\mathbf{a}'_{\|}\|^2 = \mathbf{a}'_{\|}[i]^2 / \|\mathbf{a}_{\|}\|^2 = \tilde{\mathbf{a}}[i]^2$。所以，要证明该结论，只需说明对于所有 i，$\tilde{\mathbf{a}}[i]^2$ 是 $1/L$ 的整数倍。算法 4 保证了这一点。执行第 1 行后，我们可以看到，对于所有 i，$\tilde{\mathbf{z}}[i]^2$ 是 $1/L$ 的整数倍。由于 L 是整数，1 显然也是 $1/L$ 的整数倍。因此，第 2 行中设定的 $\delta = 1 - \|\tilde{\mathbf{z}}\|^2$ 是 $1/L$ 的整数倍。最后，这保证了第 3 行中设定的 $\tilde{\mathbf{z}}[i^*]^2 = \tilde{\mathbf{z}}[i^*]^2 + \delta$ 是 $1/L$ 的整数倍，从而完成该结论的证明。

断言 2：$W_{\mathbf{a}} = W_{\mathbf{a}'}$。如上所示，$\|\mathbf{a}'\| = \|\mathbf{a}\|$。因此，要证明该断言，只需证明 $\text{Round}\left(\mathbf{a}/\|\mathbf{a}\|, L\right) = \text{Round}\left(\mathbf{a}'_{\|}/\|\mathbf{a}'_{\|}\|, L\right)$。这确保算法 3 对输入 a 和 a′ 的处理完全相同。根据断言（1），$\text{Round}(\mathbf{a}'/\|\mathbf{a}'\|, L) = \mathbf{a}'/\|\mathbf{a}'\| = \mathbf{a}'/\|\mathbf{a}\| = \tilde{\mathbf{a}}$。根据定义，$\tilde{\mathbf{a}} = \text{Round}(\mathbf{a}/\|\mathbf{a}\|, L)$。断言得证。

断言 3： 对于 $L \geq 9n^6/\epsilon^2$，$|\langle \mathbf{a}、\mathbf{b}\rangle - \langle \mathbf{a}'、\mathbf{b}'\rangle| \leq \epsilon B$. 令 $\hat{\mathbf{a}} = \mathbf{a}/\|\mathbf{a}\|$，以及 $\hat{\mathbf{b}} = \mathbf{b}/\|\mathbf{b}\|$。于是 $\tilde{\mathbf{a}} = \text{舍入}(\hat{\mathbf{a}}\ L)$，且 $\tilde{\mathbf{b}} = \hat{}\text{舍入}(\mathbf{b}\ L)$。我们将证明

$$\left|\langle \hat{\mathbf{a}}, \hat{\mathbf{b}}\rangle - \langle \tilde{\mathbf{a}}, \tilde{\mathbf{b}}\rangle\right| \leq \epsilon \cdot \sqrt{\|\hat{\mathbf{a}}_I\|^2 + \|\hat{\mathbf{b}}_I\|^2}. \tag{11}$$

将（11）的两边分别乘以 $\|\mathbf{a}\|\|\mathbf{b}\|$，得到：

$$\begin{aligned}
|\langle \mathbf{a}, \mathbf{b}\rangle - \langle \mathbf{a}', \mathbf{b}'\rangle| &\leq \epsilon \cdot \|\mathbf{a}\|\|\mathbf{b}\| \sqrt{\|\hat{\mathbf{a}}_I\|^2 + \|\hat{\mathbf{b}}_I\|^2} \\
&= \epsilon \cdot \|\mathbf{a}\|\|\mathbf{b}\| \sqrt{\left(\frac{\|\mathbf{a}_I\|^2}{\|\mathbf{a}\|^2} + \frac{\|\mathbf{b}_I\|^2}{\|\mathbf{b}\|^2}\right)} \\
&= \epsilon \sqrt{(\|\mathbf{a}_I\|^2\|\mathbf{b}\|^2 + \|\mathbf{b}_I\|^2\|\mathbf{a}\|^2)} \\
&\leq \sqrt{2}\epsilon \cdot \max\left(\|\mathbf{a}_I\|^2\|\mathbf{b}\|^2, \|\mathbf{b}_I\|^2\|\mathbf{a}\|^2\right),
\end{aligned}$$

调整 ϵ 一个常数后，即完成了该断言的证明。

$\hat{}$ 我们继续证明（11）。注意，对于任意 $i \notin \mathrm{I}$，$\hat{\mathbf{a}}[i]$ 或 $\mathbf{b}[i]$ 中至少有一个等于 0。相应地，$\tilde{\mathbf{a}}[i]$ 或 $\mathbf{b}[i]$ 中也至少有一个为 0，因为在算法 4 的舍入过程中，$\mathbf{z}$ 中任何等于 0 的条目都会在 $\tilde{\mathbf{z}}$ 中被设为 0。$\hat{}\ \hat{}\ \tilde{}$ $\tilde{}$ 因此我们可以得出，$\langle \hat{\mathbf{a}}, \mathbf{b}\rangle = \langle \hat{\mathbf{a}}_I、\mathbf{b}_I\rangle$，同样地，$\langle \tilde{\mathbf{a}}, \mathbf{b}\rangle = \langle \tilde{\mathbf{a}}_I、\mathbf{b}_I\rangle$。由此可得：

$$\left|\langle \hat{\mathbf{a}}, \hat{\mathbf{b}}\rangle - \langle \tilde{\mathbf{a}}, \tilde{\mathbf{b}}\rangle\right| = \left|\langle \hat{\mathbf{a}}_I, \hat{\mathbf{b}}_I\rangle - \langle \tilde{\mathbf{a}}_I, \tilde{\mathbf{b}}_I\rangle\right|.$$

因此，要证明 (11)，只需对上述等式右侧进行界定。我们考虑两种情况：

情况 1：$\max\left(\|\hat{\mathbf{a}}_I\|, \|\hat{\mathbf{b}}_I\|\right) \geq \frac{1}{\sqrt{L}}$. 对于 $i \in \mathrm{I}$，若 $|\hat{\mathbf{a}}[i]| < \frac{1}{\sqrt{L}}$ and $L \geq n$，则 $\lceil \hat{\mathbf{a}}[i] \rceil < \frac{1}{\sqrt{n}}$，因此 $i \neq \arg\max_{i \in 1,\ldots,n} \hat{\mathbf{a}}[i]$，因为 $\hat{\mathbf{a}}$ 是单位向量，故至少有一个分量的幅值不小于 $\geq 1/\sqrt{n}$。因此，$\hat{\mathbf{a}}[i]$ 在算法 4 的第 1 行进行了舍入，而没有在第 3 行进行舍入。我们有 $\lfloor \hat{\mathbf{a}}[i]^2 \cdot L \rfloor = 0$，因此 $|\tilde{\mathbf{a}}[i] - \hat{\mathbf{a}}[i]| = \lceil \hat{\mathbf{a}}[i] \rceil < \frac{1}{\sqrt{L}}$。或者，若 $\lceil \hat{\mathbf{a}}[i] \rceil \geq \frac{1}{\sqrt{L}}$ and $i \neq \arg\max_{i \in 1,\ldots,n} \hat{\mathbf{a}}[i]$（因此 $\hat{\mathbf{a}}[i]$ 在算法 4 的第 1 行进行了舍入，但没有在第 3 行进行舍入），则：

$$\begin{aligned}
|\tilde{\mathbf{a}}[i] - \hat{\mathbf{a}}[i]| &\leq \frac{1}{\sqrt{L}} \cdot \left|\sqrt{\hat{\mathbf{a}}[i]^2 \cdot L} - \sqrt{\hat{\mathbf{a}}[i]^2 \cdot L - 1}\right| \\
&= \frac{1}{\sqrt{L}} \cdot \frac{1}{\sqrt{\hat{\mathbf{a}}[i]^2 \cdot L} + \sqrt{\hat{\mathbf{a}}[i]^2 \cdot L - 1}} \leq \frac{1}{\sqrt{L}}.
\end{aligned}$$

如果 $i = \arg\max_{i \in 1,\ldots,n} \hat{\mathbf{a}}[i]$，那么$\hat{\mathbf{a}}[i]$ 会在第 3 行进行舍入，因此

$$|\tilde{\mathbf{a}}[i] - \hat{\mathbf{a}}[i]| \leq \left|\sqrt{\hat{\mathbf{a}}[i]^2 + \delta} - |\hat{\mathbf{a}}[i]|\right| \leq \frac{\delta}{2|\hat{\mathbf{a}}[i]|}, \tag{12}$$

其中我们利用了 $\sqrt{x}$是凹函数，其导数为 $\frac{1}{2|\hat{\mathbf{a}}[i]|}$处的 $\hat{\mathbf{a}}[i]^2$。在 算法 4 的第 2 行中，我们设定 $\delta = 1 - \|\tilde{\mathbf{a}}\|^2$，其中 $\tilde{\mathbf{a}}$ 由第 1 行中将 $\hat{\mathbf{a}}$的条目向下舍入得到。每个平方后的条目最多向下舍入 $1/L$，因此，回想一下 $\hat{\mathbf{a}}$ 是一个单位向量，$\delta \le n/L$。代入（12），并回想一下我们假设 $\hat{\mathbf{a}}[i] \ge 1/\sqrt{L}$，

$$|\tilde{\mathbf{a}}[i] - \hat{\mathbf{a}}[i]| \le \frac{n/L}{2/\sqrt{L}} \le \frac{n}{\sqrt{L}}. \tag{13}$$

总之，我们可以得出 $\|\tilde{\mathbf{a}}_I - \hat{\mathbf{a}}_I\|_\infty \le \frac{n}{\sqrt{L}}$。类似地，我们有 $\|\mathbf{b}_I - \mathbf{b}_I\|_\infty \le \frac{n}{\sqrt{L}}$。因此，

$$\begin{aligned}\left|\langle\tilde{\mathbf{a}},\tilde{\mathbf{b}}\rangle - \langle\hat{\mathbf{a}},\hat{\mathbf{b}}\rangle\right| &= \left|\langle\tilde{\mathbf{a}}_I,\tilde{\mathbf{b}}_I\rangle - \langle\hat{\mathbf{a}}_I,\hat{\mathbf{b}}_I\rangle\right| \\ &\le \frac{n}{\sqrt{L}}\left(\|\hat{\mathbf{a}}_I\|_1 + \|\hat{\mathbf{b}}_I\|_1\right) + \frac{|I|\cdot n^2}{L}.\end{aligned}$$

根据柯西-施瓦茨不等式，我们有 $\|\hat{\mathbf{a}}_I\|_1 \le \sqrt{|I|}\cdot\|\hat{\mathbf{a}}_I\|$，且 $\|\hat{\mathbf{b}}_I\|_1 \le \sqrt{|I|}\cdot\|\hat{\mathbf{b}}_I\|$。总之，这给出：

$$\begin{aligned}\left|\langle\tilde{\mathbf{a}},\tilde{\mathbf{b}}\rangle - \langle\hat{\mathbf{a}},\hat{\mathbf{b}}\rangle\right| &\le \frac{n\sqrt{|I|}}{\sqrt{L}}\left(\|\hat{\mathbf{a}}_I\| + \|\hat{\mathbf{b}}_I\|\right) + \frac{|I|\cdot n^2}{L} \\ &\le \frac{n^3}{\sqrt{L}}\cdot\left(\|\hat{\mathbf{a}}_I\| + \|\hat{\mathbf{b}}_I\| + \max\left(\|\hat{\mathbf{a}}_I\|,\|\hat{\mathbf{b}}_I\|\right)\right),\end{aligned}$$

其中，在最后一行中，我们利用了 $|I| \le n$，以及情况 1 中 $\max\left(\|\hat{\mathbf{a}}_I\|,\|\hat{\mathbf{b}}_I\|\right) \ge \frac{1}{\sqrt{L}}$ 的假设。令 $L \ge \frac{9n^6}{\epsilon^2}$，可得

$$\left|\langle\tilde{\mathbf{a}},\tilde{\mathbf{b}}\rangle - \langle\hat{\mathbf{a}},\hat{\mathbf{b}}\rangle\right| \le \epsilon\cdot\max\left(\|\hat{\mathbf{a}}_I\|,\|\hat{\mathbf{b}}_I\|\right) \le \epsilon\cdot\sqrt{\|\hat{\mathbf{a}}_I\|^2 + \|\hat{\mathbf{b}}_I\|^2}.$$

这就证明了11）适用于情况 1。

情形2：$\max\left(\|\hat{\mathbf{a}}_I\|,\|\hat{\mathbf{b}}_I\|\right) < \frac{1}{\sqrt{L}}$。在这种情况下，对于所有 $i \in \mathrm{I}$，$|\hat{\mathbf{a}}[i]| < \frac{1}{\sqrt{L}}$且 $|\hat{\mathbf{b}}[i]| < \frac{1}{\sqrt{L}}$。因此，对于 $L > n$，不存在满足 $i = \arg\max_{i\in1,\ldots,n}\hat{\mathbf{a}}[i]$或 $i = \arg\max_{i\in1,\ldots,n}\mathbf{b}[i]$的 $i \in \mathrm{I}$。所以，对于所有 $i \in \mathrm{I}$，$\hat{\mathbf{a}}[i]$和 $\mathbf{b}[i]$在算法 4 的第 1 行中都被舍入为 0。也就是说，$\tilde{\mathbf{a}}_I$ 和 $\mathbf{b}_I$ 都是全零向量。因此，要证明 (11)，我们必须证明 $\left|\langle\hat{\mathbf{a}}_I,\hat{\mathbf{b}}_I\rangle\right| \le \epsilon\cdot\sqrt{\|\hat{\mathbf{a}}_I\|^2 + \|\hat{\mathbf{b}}_I\|^2}$.。这可由柯西-施瓦茨不等式以及我们的假设 $\|\mathbf{a}_I\|,\|\mathbf{b}_I\| < \frac{1}{\sqrt{L}}$得出。

$$\begin{aligned}\left|\langle\hat{\mathbf{a}}_I,\hat{\mathbf{b}}_I\rangle\right| &\le \|\hat{\mathbf{a}}_I\|\|\hat{\mathbf{b}}_I\| \le \frac{1}{\sqrt{L}}\max\left(\|\hat{\mathbf{a}}_I\|,\|\hat{\mathbf{b}}_I\|\right) \\ &\le \frac{1}{\sqrt{L}}\sqrt{\|\hat{\mathbf{a}}_I\|^2 + \|\hat{\mathbf{b}}_I\|^2}.\end{aligned}$$

设置$L \ge \frac{1}{\epsilon^2}$给出（11），从而完成引理的断言（3）。

断言4： 对于 $L \ge n^3$，$\max\left(\|\mathbf{a}'_I\|\|\mathbf{b}'\|,\ \|\mathbf{a}'\|\|\mathbf{b}'_I\|\right) \le 2B$. 回忆一下，根据构造，$\|\mathbf{a}'\| = \|\mathbf{a}\|$且 $\|\mathbf{b}'\| = \|\mathbf{b}\|$。因此，将不等式两边分别除以 $\|\mathbf{a}\|\|\mathbf{b}\|$ 后，只需证明：

$$\max\left(\frac{\|\mathbf{a}'_I\|}{\|\mathbf{a}\|},\frac{\|\mathbf{b}'_I\|}{\|\mathbf{b}\|}\right) \le 2\max\left(\frac{\|\mathbf{a}_I\|}{\|\mathbf{a}\|},\frac{\|\mathbf{b}_I\|}{\|\mathbf{b}\|}\right).$$

也就是说，我们必须证明 $\max(\|\tilde{\mathbf{a}}_I\|,\ \|\mathbf{b}_I\|) \le 2\max(\|\hat{\mathbf{a}}_I\|,\ \|\mathbf{b}_I\|)$。只需证明 $\|\tilde{\mathbf{a}}_I\| \le 2\|\hat{\mathbf{a}}_I\|$，以及$\|\mathbf{b}_I\| \le 2\|\mathbf{b}_I\|$。我们着重证明 $\mathbf{a}$ 的情形。$\mathbf{b}$ 的界限可用相同方法证明。我们分两种情况讨论。令 $i^* = \arg\max_{i\in1,\ldots,n}|\hat{\mathbf{a}}[i]|$。

情况 1： $i^* \notin I$. 在这种情况下，$\hat{\mathbf{a}}_I$ 中的所有条目都只在算法 4的第 1 行进行舍入。因此，它们都向下舍入，所以 $\|\tilde{\mathbf{a}}_I\| \le \|\hat{\mathbf{a}}_I\|$，结论成立。

情形2：$i^* \in \mathrm{I}$. 在这种情况下，由于 $\hat{\mathbf{a}}$ 是一个单位向量，我们有 $\|\hat{\mathbf{a}}_I\| \ge |\hat{\mathbf{a}}[i^*]| \ge 1/\sqrt{n} \ge 1/\sqrt{L}$当 $L > n$ 时。此外，$\hat{\mathbf{a}}_I$中的所有条目都向下舍入，除了 $\hat{\mathbf{a}}[i^*]$。但如式 (13) 所示，$|\tilde{\mathbf{a}}[i^*]| \le |\hat{\mathbf{a}}[i^*]| + \frac{n}{\sqrt{L}}$. 因此，只要 $L \ge n^3$，就有 $\|\tilde{\mathbf{a}}_I\| \le \|\hat{\mathbf{a}}_I\| + \frac{n}{\sqrt{L}} \le 2\|\hat{\mathbf{a}}_I\|$。

至此，命题 (4) 得证，进而引理得证。 □

令 e ver将所有内容结合起来。 最后，我们通过结合引理3 与引理 2，

定理 2 的证明。给定任意 $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$，令 $\mathbf{a}'$ 和 $\mathbf{b}'$ 如引理3 中所定义。考虑应用算法 3 计算大小为 $m = O(1/\epsilon^2)$ 的草图 $W_{\mathbf{a}}, W_{\mathbf{b}}, W_{\mathbf{a}'}, W_{\mathbf{b}'}$，使用离散化参数 $L = O(n^6/\epsilon^2)$。根据引理3 的第一个断言，我们可以对 $\mathbf{a}', \mathbf{b}'$ 应用引理 2，从而证明其成立的概率为 $\geq 2/3$，

$$|\mathcal{F}(W_{\mathbf{a}'}, W_{\mathbf{b}'}) - \langle \mathbf{a}', \mathbf{b}' \rangle| \leq \epsilon \max\left(\|\mathbf{a}'_I\|\|\mathbf{b}'\|, \|\mathbf{a}'\|\|\mathbf{b}'_I\|\right).$$

结合三角不等式与引理3的断言 (2) 和 (4)，我们可以得出其成立的概率为 $\geq 2/3$，

$$\begin{aligned}|\mathcal{F}(W_{\mathbf{a}}, W_{\mathbf{b}}) - \langle \mathbf{a}, \mathbf{b} \rangle| \leq\ & |\langle \mathbf{a}, \mathbf{b} \rangle - \langle \mathbf{a}', \mathbf{b}' \rangle| \\ & + 2\epsilon \max\left(\|\mathbf{a}_I\|\|\mathbf{b}\|, \|\mathbf{a}\|\|\mathbf{b}_I\|\right).\end{aligned}$$

最后，应用引理3的断言 (3) 可得

$$|\mathcal{F}(W_{\mathbf{a}}, W_{\mathbf{b}}) - \langle \mathbf{a}, \mathbf{b} \rangle| \leq 3\epsilon \max\left(\|\mathbf{a}_I\|\|\mathbf{b}\|, \|\mathbf{a}\|\|\mathbf{b}_I\|\right).$$

将 ϵ 调整为原来的 1/3 后，即可得到定理 2 的界。成功概率为 2/3。使用一种标准技巧，我们可以通过使用算法 3 和独立随机种子 [，计算 a、b 的 $t = O(\log(1/\delta))$ 个独立草图，从而提高成功概率（Cormode 等人， 2011]）。将这些草图记为 $W(1)\mathbf{a} \cdots, W(t)\mathbf{a}$ 和 $W(1)\ \mathbf{b} \cdots, W(t)\ \mathbf{b}$。对于任意 i，其成立的概率为 $\geq 2/3$，

$$|\mathcal{F}(W_{\mathbf{a}}^{(i)}, W_{\mathbf{b}}^{(i)}) - \langle \mathbf{a}, \mathbf{b} \rangle| \leq \epsilon \max\left(\|\mathbf{a}_I\|\|\mathbf{b}\|, \|\mathbf{a}\|\|\mathbf{b}_I\|\right).$$

根据标准的切尔诺夫界，以至少 $1 - \delta$ 的概率，该界对 $> t/2$ 个独立草图成立。因此，如果我们取这些草图产生的中位数估计，它以$\geq 1 - \delta$ 的概率满足所需的界。将我们的 t 个独立草图串接成一个草图，可以看出总草图大小为$t \cdot m = O(\log(1/\delta)/\epsilon^2)$，由此得到 定理 2。 □