Weighted Minwise Hashing Beats Linear Sketching for Inner Product Estimation

NYU TANDON

Aline Bessa[1] Majid Daliri[1] Juliana Freire[1] Cameron Musco[2] Christopher Musco[1] Aécio Santos[1] Haoxiang Zhang[1]

[1]New York University
[2]University of Massachusetts Amherst

Inner Products and Sketches

- Inner products are used in many applications, e.g., computing document similarity, evaluate classification models, estimate join sizes
- For large vectors, inner product computation can be prohibitively expensive: $O(n)$
- Sketching methods have been proposed to address this challenge:

$$\mathcal{F}(\mathcal{S}(\mathbf{a}), \mathcal{S}(\mathbf{b})) \approx \langle \mathbf{a}, \mathbf{b} \rangle$$

- Sketching simultaneously reduces storage, communication, and runtime complexity

Prior Work

Linear Sketching for Inner Products [Arriaga and Vempala, 2006] Let $\epsilon, \delta \in (0,1)$ be accuracy and failure probability parameters respectively and let $m = O(\log(1/\delta)/\epsilon^2)$. Let $\mathbf{\Pi} \in \mathbb{R}^{m \times n}$ be a random matrix with each entry set independently to $+\sqrt{1/m}$ or $-\sqrt{1/m}$ with equal probability. For length n vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, let $\mathcal{S}(\mathbf{a}) = \mathbf{\Pi a}$ and $\mathcal{S}(\mathbf{b}) = \mathbf{\Pi b}$. With probability at least $1 - \delta$,

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

where $\|\mathbf{x}\|$ denotes the standard Euclidean norm.

Main Result

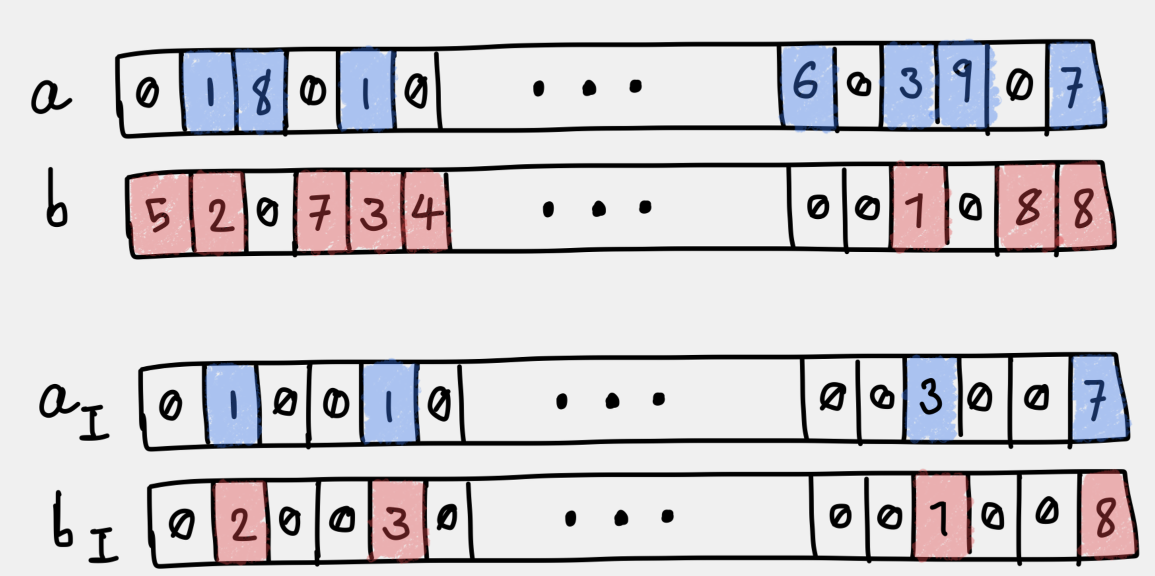
Let $\epsilon, \delta \in (0,1)$ be accuracy and failure probability parameters and let $m = O(\log(1/\delta)/\epsilon^2)$. There is an algorithm $\mathcal{S}$ based on Weighted MinHash sampling that produces size-m sketches, along with an estimation procedure $\mathcal{F}$, such that for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, with probability at least $1 - \delta$,

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

Above, $\mathcal{I} = \{i : \mathbf{a}[i] \neq 0 \text{ and } \mathbf{b}[i] \neq 0\}$ is the intersection of $\mathbf{a}$'s and $\mathbf{b}$'s supports. $\mathbf{a}_{\mathcal{I}}$ and $\mathbf{b}_{\mathcal{I}}$ denote $\mathbf{a}$ and $\mathbf{b}$ restricted to indices in $\mathcal{I}$.

Low Overlap and Data Sparsity: Why do we care?

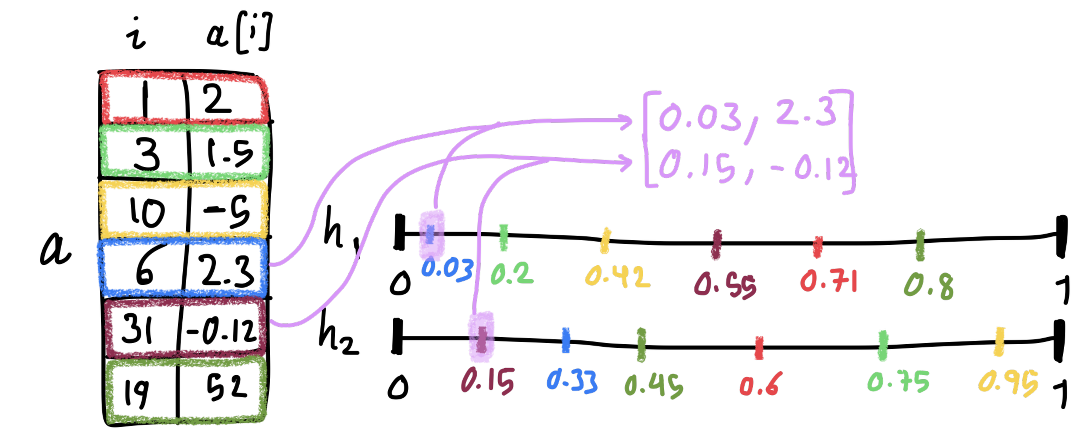
Error-bound improvement hinges on data sparsity, a common trait in practice. The provided image demonstrates how exploiting this sparsity significantly reduces errors.



MinHash Algorithm

- MinHash Sampling:
  1. Map indexes i of non-zero entries to $[0,1]$ with a hash function h;
  2. Select the index with $\min h(i)$;
  3. Repeat m times with different hashing functions h
- Store selected values $\mathbf{a}[i]$ and $\mathbf{b}[i]$ in $H_{\mathbf{a}}, H_{\mathbf{b}}$
- Estimate the inner product as:

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



Weighted Minhash Algorithm

- We propose a Weighted MinHash (WMH) algorithm that samples entries with probability proportional to $\mathbf{a}[i]^2$ and $\mathbf{b}[i]^2$:
  1. Normalize the vectors $\mathbf{a}$ and $\mathbf{b}$:

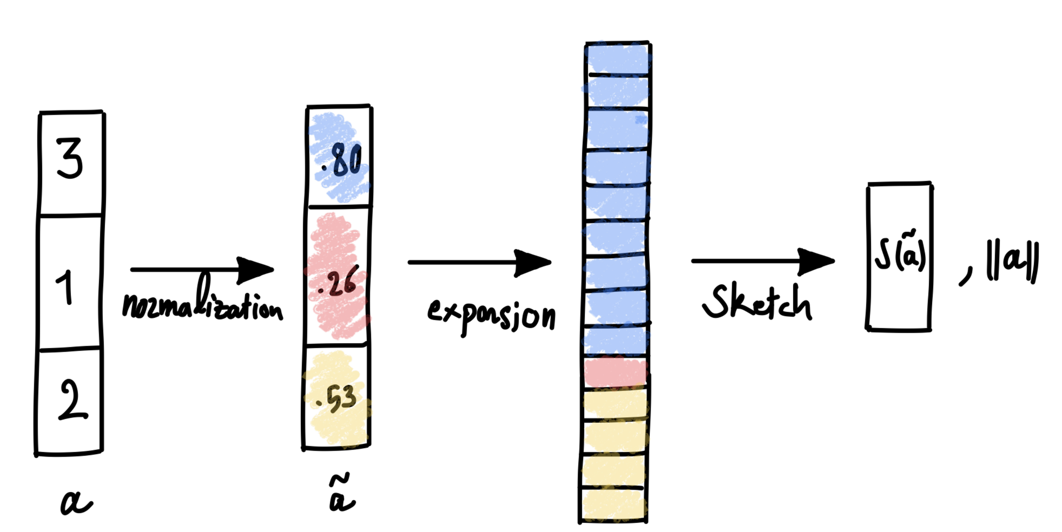
$$\tilde{\mathbf{a}} = \mathbf{a}/\|\mathbf{a}\| \qquad \tilde{\mathbf{b}} = \mathbf{b}/\|\mathbf{b}\|$$

  2. Sample entries with probability:

$$\frac{\min(\tilde{\mathbf{a}}[j]^2, \tilde{\mathbf{b}}[j]^2)}{\sum_{i=1}^{n} \max(\tilde{\mathbf{a}}[j]^2, \tilde{\mathbf{b}}[j]^2)}$$

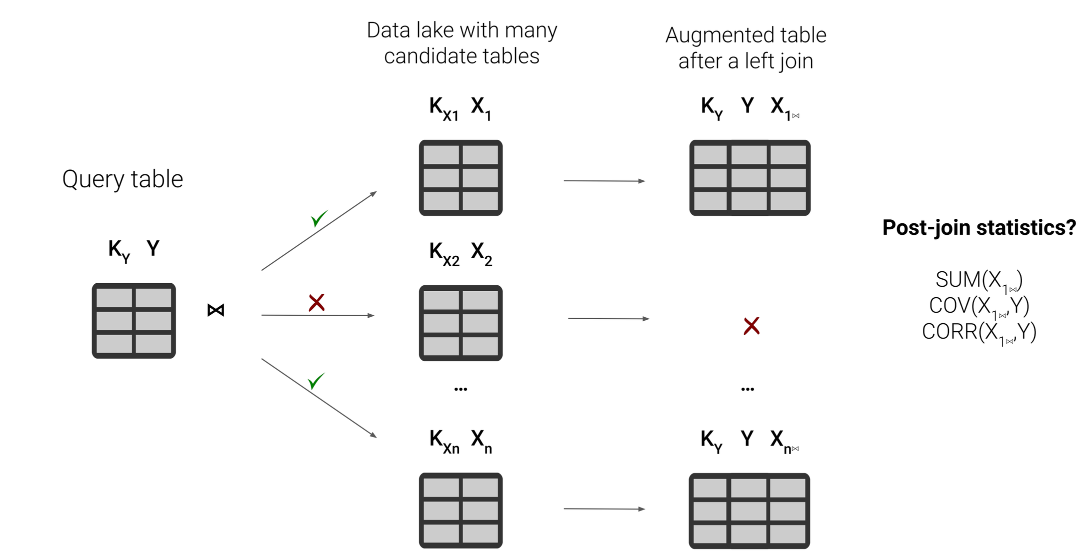
  3. Estimate the inner product by computing:

$$\langle \mathbf{a}, \mathbf{b} \rangle \approx \|\mathbf{a}\| \|\mathbf{b}\| \langle \tilde{\mathbf{a}}, \tilde{\mathbf{b}} \rangle$$

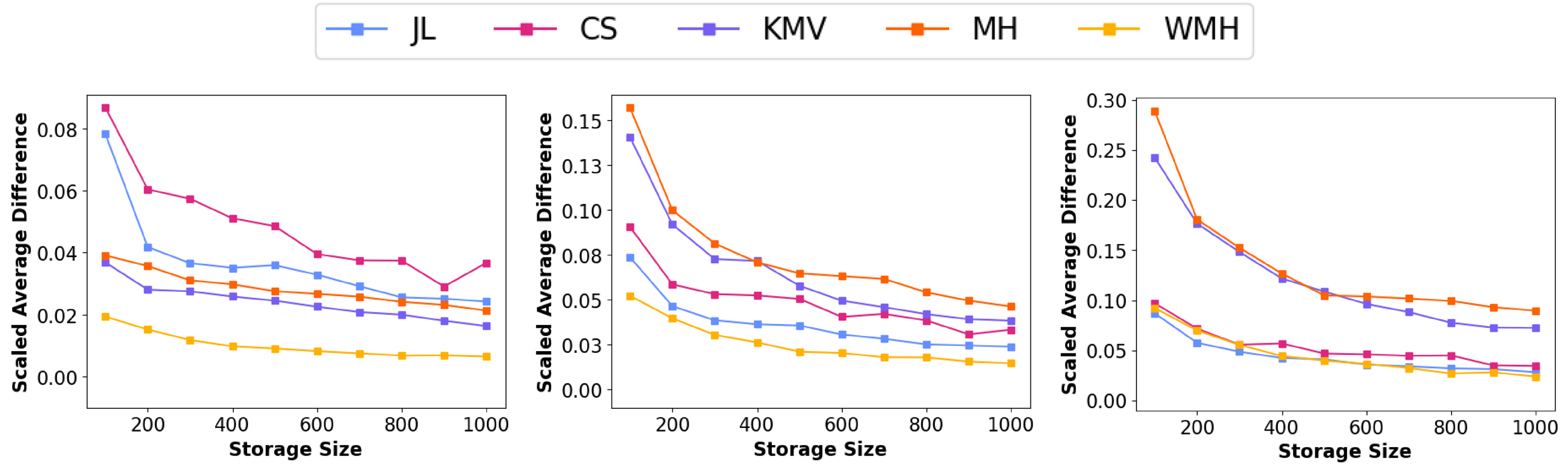


Applications

Weighted MinHash sketches enable efficient estimation of inner products for numerous applications, including (1) estimating the size of join of two tables, (2) estimating the similarity of text documents, and (3) estimating statistics (e.g. sum, covariance and correlation) of columns generated after a join *without* materializing table joins.



Experimental Result



(a) 1% overlap (b) 10% overlap (c) 50% overlap

Figure 2. Across all three plots, the Weighted MinHash method clearly outperforms or matches the accuracy of competing approaches. The anticipated substantial performance gap between JL and Weighted MinHash is particularly evident when the overlap is minimal.

Summary

Table 1. Our bounds generalize existing minwise hashing bounds for binary vectors to general vectors.

Method	Error for sketches of size $O(1/\epsilon^2)$	Assumptions
Linear Sketches	$\epsilon \cdot \|\mathbf{a}\| \|\mathbf{b}\|$	None
MinHash (MH) Sampling	$\epsilon \cdot \max(\|\mathbf{a}_{\mathcal{I}}\| \|\mathbf{b}\|, \|\mathbf{a}\| \|\mathbf{b}_{\mathcal{I}}\|)$	$\mathbf{a}, \mathbf{b} \in \{0,1\}$
MinHash (MH) Sampling	$\epsilon \cdot c^2 \cdot \max(\|\mathbf{a}_{\mathcal{I}}\| \|\mathbf{b}\|, \|\mathbf{a}\| \|\mathbf{b}_{\mathcal{I}}\|)$	$\mathbf{a}, \mathbf{b} \in [-c, c]$
Weighted MH Sampling	$\epsilon \cdot \max(\|\mathbf{a}_{\mathcal{I}}\| \|\mathbf{b}\|, \|\mathbf{a}\| \|\mathbf{b}_{\mathcal{I}}\|)$	None

Acknowledgments

NSF DARPA

Paper URL