Sampling Methods for Inner Product Sketching

Christopher Musco, New York University

COLLABORATORS

Collaborators at NYU:

Majid Daliri

Prof. Juliana Freire

Aécio Santos

Haoxiang Zhang

INNER PRODUCT

The inner product between two vectors $\mathbf{a} = [a_1, \ldots, a_d]$ and $\mathbf{b} = [b_1, \ldots, b_d]$ is:

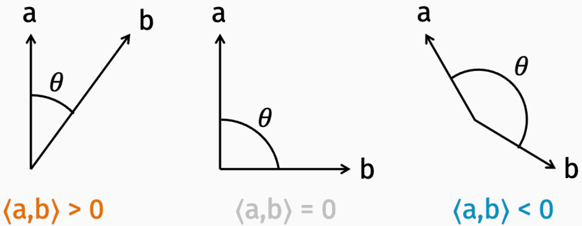
$$\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{i=1}^{d} a_i b_i \qquad .$$

INNER PRODUCT

The inner product between two vectors $\mathbf{a} = [a_1, \ldots, a_d]$ and $\mathbf{b} = [b_1, \ldots, b_d]$ is:

$$\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{i=1}^{d} a_i b_i = \frac{\cos(\theta)}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2}.$$

Natural measure of similarity between vectors:

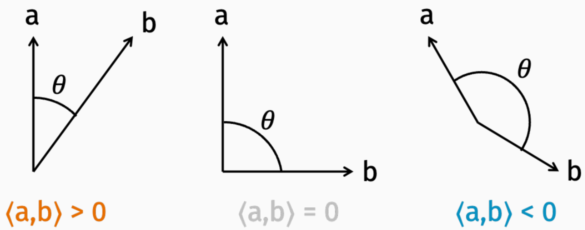


INNER PRODUCT

The inner product between two vectors $\mathbf{a} = [a_1, \ldots, a_d]$ and $\mathbf{b} = [b_1, \ldots, b_d]$ is:

$$\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{i=1}^{d} a_i b_i = \frac{\cos(\theta)}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2}.$$

Natural measure of similarity between vectors:



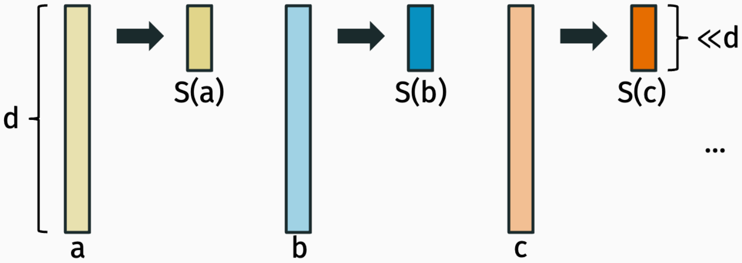
Complexity: can be computed in $O(d)$ time.

PROBLEM SETUP

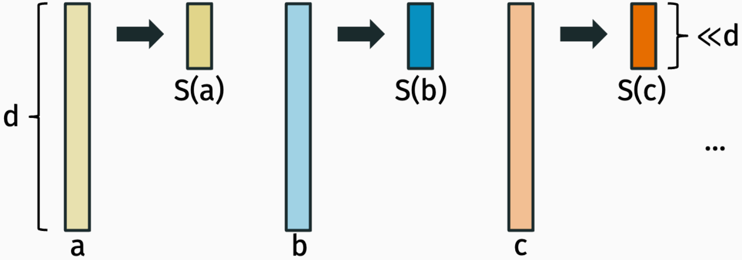
Question: Can we compute inner products faster if pre-processing is allowed?

Question: Can we compute inner products faster if pre-processing is allowed?

Concretely, hope to compute a compression ("sketch") of a vector that contains enough information to estimate that vectors inner product with any other vector.



INNER PRODUCT SKETCHING



We want that, for some estimation procedure F,

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

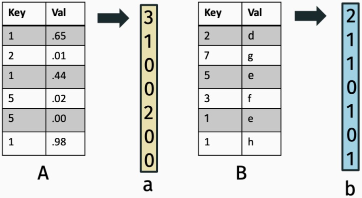
$$F(S(\mathbf{a}), S(\mathbf{c})) \approx \langle \mathbf{a}, \mathbf{c} \rangle$$

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

F should be efficient, running in time that is linear time in size of sketch, i.e. much fast then $O(d)$ time.

Useful in any setting where we need to compute many inner products with a vector. Tons of applications in databases.

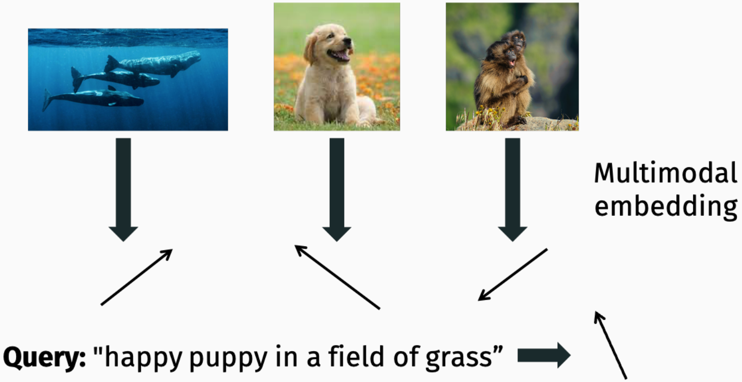
- **Join-size estimation.** Each vector contains key counts for a given table. Inner product equals size of join between two tables.



- **Estimating "post-join statistics."** For example, estimate the correlation between columns in two tables without explicitly joining the tables. Useful in dataset search.
- **Faster search in vector databases.**

MAXIMUM INNER PRODUCT SEARCH (MIPS)

Goal: Find vector representation of database item closest to vector representation of query.



- Pre-process all vectors in the database by sketching.
- Sketch query vector at query time.

GENERAL PROBLEM

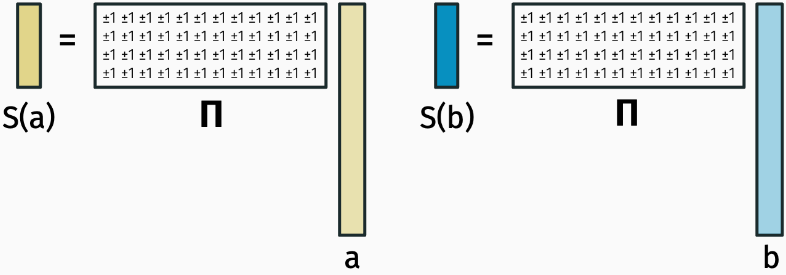
Want a sketching procedure S and estimation procedure, F, such that, for any $\mathbf{a}, \mathbf{b}$,

1. $S(\mathbf{a})$ is much smaller than $\mathbf{a}$, $S(\mathbf{b})$ is much smaller than $\mathbf{b}$.
2. $F(S(\mathbf{a}), S(\mathbf{b})) \approx \langle \mathbf{a}, \mathbf{b} \rangle$.
3. S and F should be efficient to apply, i.e., run in linear time.

Large **Existing Approach: Linear Sketching**

Includes Johnson-Lindenstrauss random projection, AMS sketch, CountSketch, Fast-AGMS sketch, etc.

Main idea: Compress **a** and **b** by multiplying by a random matrix, **Π**. E.g., random ± 1 or Gaussian entries.



Then we simply estimate $\langle \mathbf{a}, \mathbf{b} \rangle$ as:

$$\langle \mathbf{a}, \mathbf{b} \rangle = \langle S(\mathbf{a}), S(\mathbf{b}) \rangle = \langle \mathbf{\Pi a}, \mathbf{\Pi b} \rangle.$$

[Alon, Gibbons, Matias, Szegedy, 1999], [Achlioptas, 2003], [Dasgupta, Gupta, 2003]

Beautifully simple approach with strong guarantee.

Theorem (Folklore / Arriaga, Vempala, 2006)

For random Gaussian entries, ± 1 *entries, etc.,*

$$\mathbb{E}[\langle \mathbf{\Pi} \mathbf{a}, \mathbf{\Pi} \mathbf{b} \rangle] = \langle \mathbf{a}, \mathbf{b} \rangle,$$

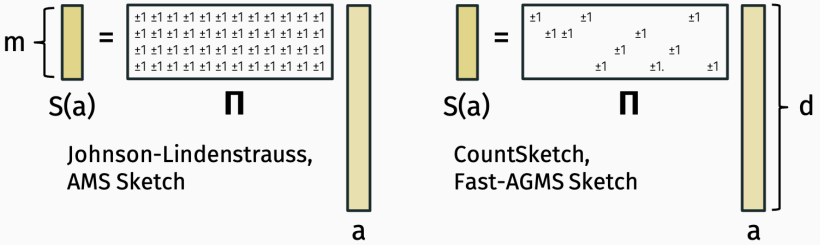
and, if $\mathbf{\Pi}$ *is chosen to have* *m rows, then:*

$$\text{Var}[\langle \mathbf{\Pi} \mathbf{a}, \mathbf{\Pi} \mathbf{b} \rangle] \leq \frac{2}{m} \|\mathbf{a}\|_2 \|\mathbf{b}\|_2.$$

Corollary: If we use sketches of size $m = O(1/\epsilon^2)$ (independent of original dimension d!), then with high probability,

$$|\langle \mathbf{\Pi} \mathbf{a}, \mathbf{\Pi} \mathbf{b} \rangle - \langle \mathbf{a}, \mathbf{b} \rangle| \leq \epsilon \cdot \|\mathbf{a}\|_2 \|\mathbf{b}\|_2.$$

FAST LINEAR SKETCHING



Naive cost of linear sketching is $O(dm)$ time.

This can be accelerated to $O(d)$ (linear) time without sacrificing accuracy by using an ultra-sparse random matrix. [Charikar, Chen, Farach-Colton, 2002], [Cormode, Garofalakis, 2005].

Thanks to strong theoretical guarantees, speed, and simplicity, linear sketching has become ubiquitous for inner product estimation. Is there any hope to do better?

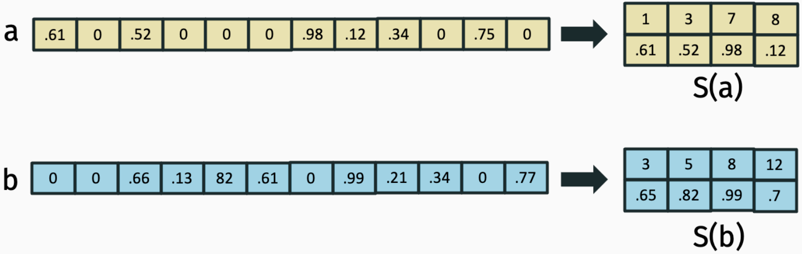
Main result: There is a simple, linear time sketching method that improves on the accuracy of linear sketching both in theory and in practice.

Builds on our work[1] from PODS 2023, which offered improved accuracy in theory.

[1]"Weighted Minwise Hashing Beats Linear Sketching for Inner Product Estimation", Bessa, Daliri, Freire, Musco, Musco, Santos, Zhang, 2013.

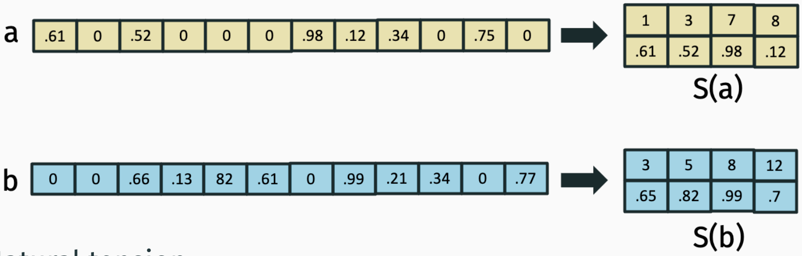
SAMPLING-BASED SKETCH

Sketch consists of subset of index/value pairs from **a** and **b**.



Let $\mathcal{T}$ be the set of indices common to $S(\mathbf{a})$, $S(\mathbf{b})$. Estimate $\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{i=1}^{d} a_i b_i$ based on $\sum_{i \in \mathcal{T}} a_i b_i$, which we can compute from the sketches.

BASIC IDEA



Natural tension:

- Larger entries in **a** and **b** contribute more to $\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{i=1}^{d} a_i b_i$. I.e., choice of indices should depend on magnitude of entries in vector being sketched.
- Want $S(\mathbf{a})$ and $S(\mathbf{b})$ to have many of the same indices. I.e., choice of indices should be independent of the vectors.

Have to balance these two goals.

Threshold Sampling (our method):

- Set target sketch size m.
- Draw uniform random numbers $u_1, \ldots, u_d \sim [0, 1]$.
- For $i \in 1, \ldots, d$:
  - Add (i, a_i) to $S(\mathbf{a})$ if $u_i \leq m \cdot \frac{a_i^2}{\|\mathbf{a}\|_2^2}$.
  - Add (i, b_i) to $S(\mathbf{b})$ if $u_i \leq m \cdot \frac{b_i^2}{\|\mathbf{b}\|_2^2}$.

Estimation:

- Let $\mathcal{T}$ be the set of indices common to $S(\mathbf{a}), S(\mathbf{b})$.
- Return $F(S(\mathbf{a}), S(\mathbf{b})) = \sum_{i \in \mathcal{T}} \frac{1}{p_i} a_i b_i$, where $p_i = \min\left(1, m \cdot \frac{a_i^2}{\|\mathbf{a}\|_2^2}, m \cdot \frac{b_i^2}{\|\mathbf{b}\|_2^2}\right)$.

Similar method has been used for join-size estimation under the name “End-Biased Sampling” [Estan, Naughton, 2006].

THEORETICAL GUARANTEE

Theorem

Let $S(\mathbf{a}), S(\mathbf{b})$ be sketches for $\mathbf{a}, \mathbf{b}$ obtained via Threshold Sampling and let $F(S(\mathbf{a}), S(\mathbf{b}))$ be the corresponding estimate for $\langle \mathbf{a}, \mathbf{b} \rangle$ obtained from those sketches.

We have that $\mathbb{E}[|S(\mathbf{a})|] = m$, $\mathbb{E}[|S(\mathbf{b})|] = m$, and:

$$\mathbb{E}[F(S(\mathbf{a}), S(\mathbf{b}))] = \langle \mathbf{a}, \mathbf{b} \rangle$$

$$\mathrm{Var}[F(S(\mathbf{a}), S(\mathbf{b}))] \leq \frac{2}{m} \underline{\max(\|\mathbf{a}_{\mathcal{I}}\|_2^2 \|\mathbf{b}\|_2^2, \|\mathbf{a}\|_2^2 \|\mathbf{b}_{\mathcal{I}}\|_2^2)}$$

Corollary: If $m = O(1/\epsilon^2)$, then with high probability, $|F(S(\mathbf{a}), S(\mathbf{b})) - \langle \mathbf{a}, \mathbf{b} \rangle| \leq \epsilon \cdot \underline{\max(\|\mathbf{a}_{\mathcal{I}}\|_2 \|\mathbf{b}\|_2, \|\mathbf{a}\|_2 \|\mathbf{b}_{\mathcal{I}}\|_2)}$.

MASKED VECTORS

$\mathbf{a}$	.61	0	.52	0	0	0	.98	.12	.34	0	.75	0
$\mathbf{b}$	0	0	.66	.13	82	.61	0	.99	.21	.34	0	.77
$\mathbf{a}_{\mathcal{I}}$	0	0	.52	0	0	0	0	.12	.34	0	0	0
$\mathbf{b}_{\mathcal{I}}$	0	0	.66	0	0	0	0	.99	.21	0	0	0

Linear sketching variance: $\frac{2}{m} \cdot \|\mathbf{a}\|_2^2 \|\mathbf{b}\|_2^2$

Threshold sampling variance: $\frac{2}{m} \cdot \max(\|\mathbf{a}_{\mathcal{I}}\|_2^2 \|\mathbf{b}\|_2^2, \|\mathbf{a}\|_2^2 \|\mathbf{b}_{\mathcal{I}}\|_2^2)$

Can be a significant improvement. E.g., if $\mathbf{a}$ and $\mathbf{b}$ overlap on 5% of entries, we expect that $\|\mathbf{a}_{\mathcal{I}}\|_2^2 \approx .05\|\mathbf{a}\|_2^2$ and $\|\mathbf{b}_{\mathcal{I}}\|_2^2 \approx .05\|\mathbf{b}\|_2^2$.

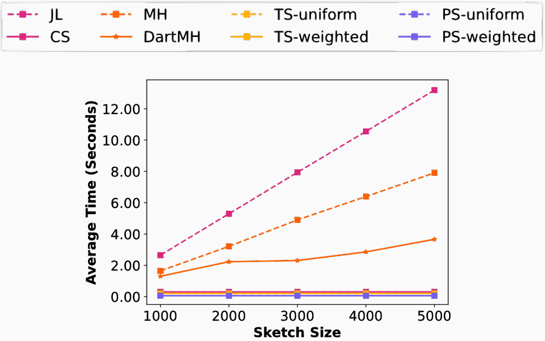
Equates to 20x reduction in variance.

OTHER POINTS

1. Analysis of the method is completely elementary.
2. Sketches can be improved to have size exactly m, instead of just m in expectation.[2] Essentially no loss in accuracy.
3. The method actually works really well in experiments.

[2] Using well known Priority Sampling method of [Duffield, Lund, Thorup, 2004].

EXPERIMENTAL EVALUATION: SKETCHING SPEED



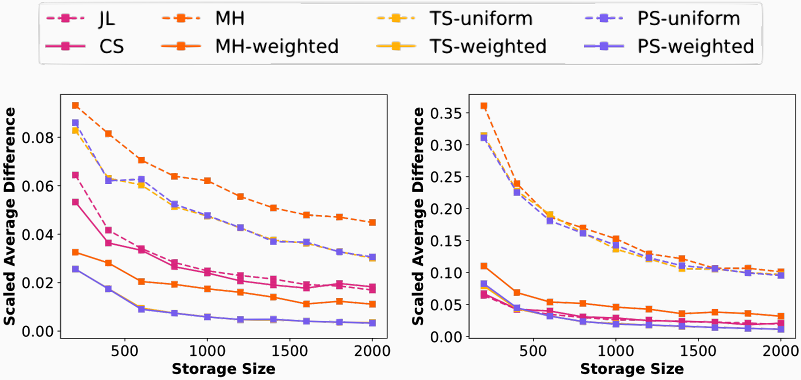
Linear Sketching Methods: JL = Johnson-Lindenstrauss, CS = CountSketch

Our PODS 2023 Methods: MH, DartMH = "MinHash" based sampling.

This work: TS-weighted, PS-weighted = Threshold and Priority Sampling.

Takeaway: Sketching time is $O(d)$, does not scale with size of sketch.

EXPERIMENTAL EVALUATION: ACCURACY



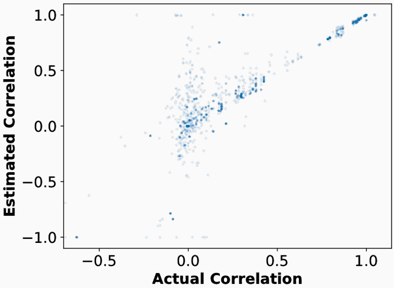
10% non-zero overlap

100% non-zero overlap

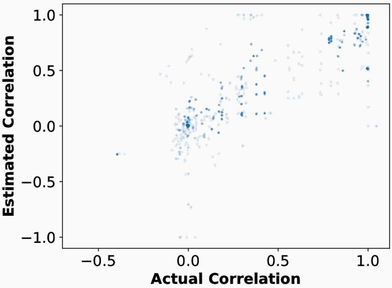
- Bigger accuracy improvement for sparse vectors $\mathbf{a}, \mathbf{b}$ with less overlap between non-zero entries.
- When 100% of non-zeros overlap, performance matches linear sketching methods, as predicted by our theory.

EXPERIMENTAL EVALUATION: ACCURACY

Priority Sampling



CountSketch



- Post-join correlation estimation for World Bank Data. Our best sampling method outperforms the best linear sketching method (CountSketch) with the same size sketch.

See paper for experiments on document similarity, join size estimation, and more!

questions?