Matrix Product Sketching via Coordinated Sampling

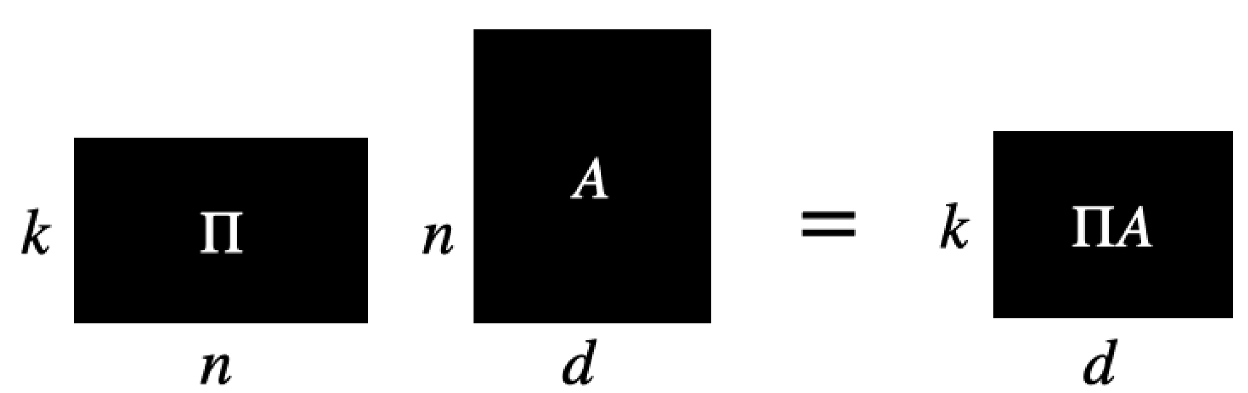
Majid Daliri [1] Juliana Freire [1] Danrong Li [2] Christopher Musco [1]
[1]New York University [2]Pennsylvania State University

Matrix Product Approximation

Prior Work - Sketch Setting

Fact.[1,2,3] Let $\Pi \in^{k\times n}$ be a scaled random Gaussian matrix, random sign matrix, CountSketch matrix [4] or any of a variety of other randomized linear embeddings. If $k = O\left(\frac{1}{\epsilon^2\delta}\right)$, then with probability at least $1-\delta$,

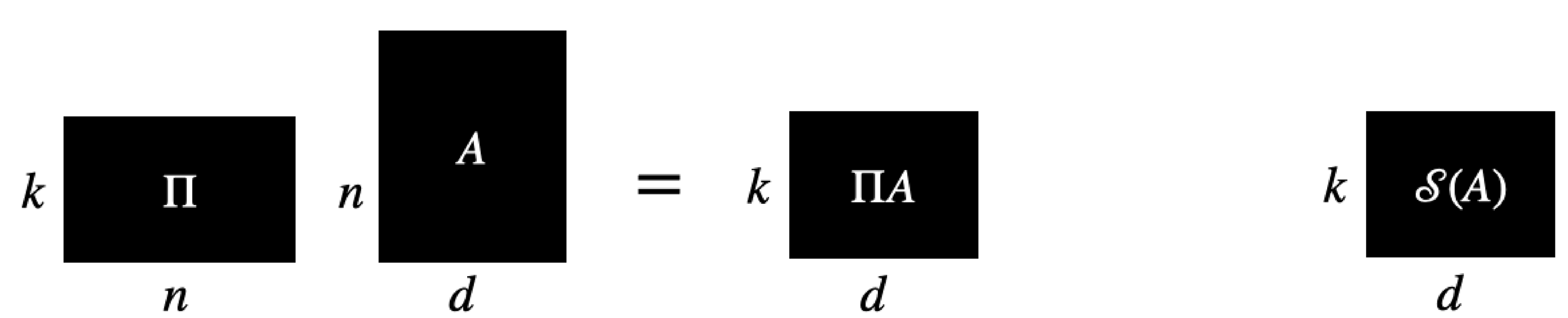
$$\|(\Pi A)^T(\Pi B) - A^T B\|_F \le \epsilon \|A\|_F \|B\|_F.$$



Our Result

Main Theorem. Consider $A \in^{n\times d}$, $B \in^{n\times m}$, and any $\epsilon, \delta \in (0,1)$. There is a sketching procedure (Algorithm 1) that constructs sketches $\mathcal{S}(A)$ and $\mathcal{S}(B)$ consisting of at most $k = \frac{2/\delta}{\epsilon^2} + 1$ rows from A and B, and there is a corresponding estimation procedure (Algorithm 2) that, using the information in these sketches, returns an estimate W such that, with probability $1-\delta$,

$$\|W - A^T B\|_F \le \epsilon \|A\|_F \|B\|_F.$$



$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & 1 \\ -1 & 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 3 \\ 0 & 7 & 0 \\ 9 & 0 & 5 \\ 11 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 21 & 7 & 8 \\ 1 & -7 & -8 \\ 1 & 7 & -8 \end{bmatrix} \qquad \begin{bmatrix} 0 & 7 & 0 \\ 9 & 0 & 5 \\ 11 & 0 & 0 \end{bmatrix}$$

Dense ΠA Compact $\mathcal{S}(A)$

Method	Error Bound	Can Construct Sketch Independently	Saves Storage Space When Matrix Is Sparse
Prior Work - Non Sketch Setting [5]	$\epsilon\|\|A\|\|_F\|\|B\|\|_F$	✗	✓
Prior Work - Sketch Setting	$\epsilon\|\|A\|\|_F\|\|B\|\|_F$	✓	✗
Our Result	$\epsilon\|\|A\|\|_F\|\|B\|\|_F$	✓	✓

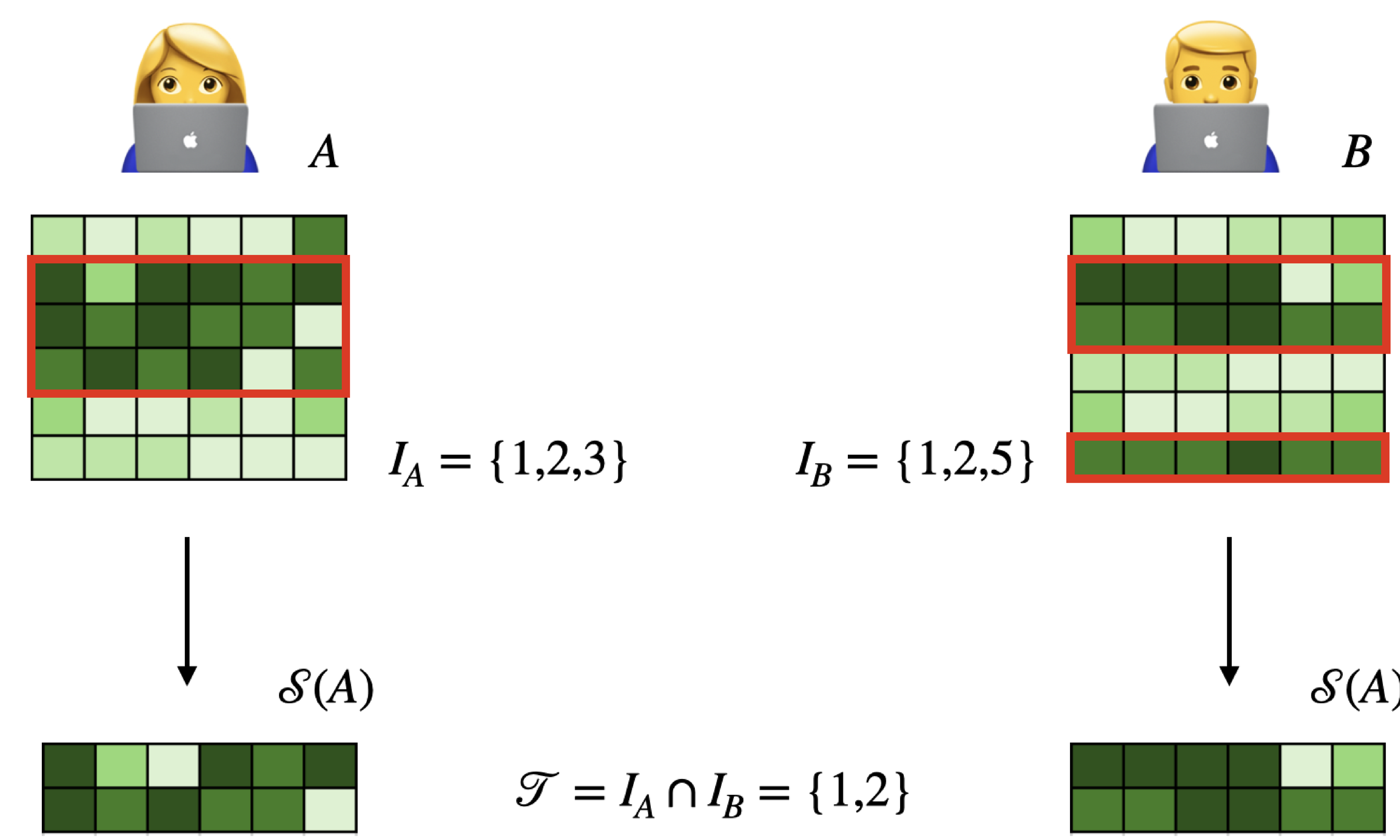
Threshold Sampling - Sketch Size Only Bounded in Expectation

Algorithm 4 Threshold Sampling

Input: Matrix $\mathbf{A}$ of size $n \times d$, random seed s, target number of row samples, k.
Output: Sketch $\mathcal{S}(\mathbf{A}) = \{\mathcal{I}_\mathbf{A}, V_\mathbf{A}, \tau_\mathbf{A}\}$, where $\mathcal{I}_\mathbf{A}$ is a subset of row indices from $\{1,\dots,n\}$ and $V_\mathbf{A}$ contains $\mathbf{A}_i$ for all $i \in \mathcal{I}_\mathbf{A}$.

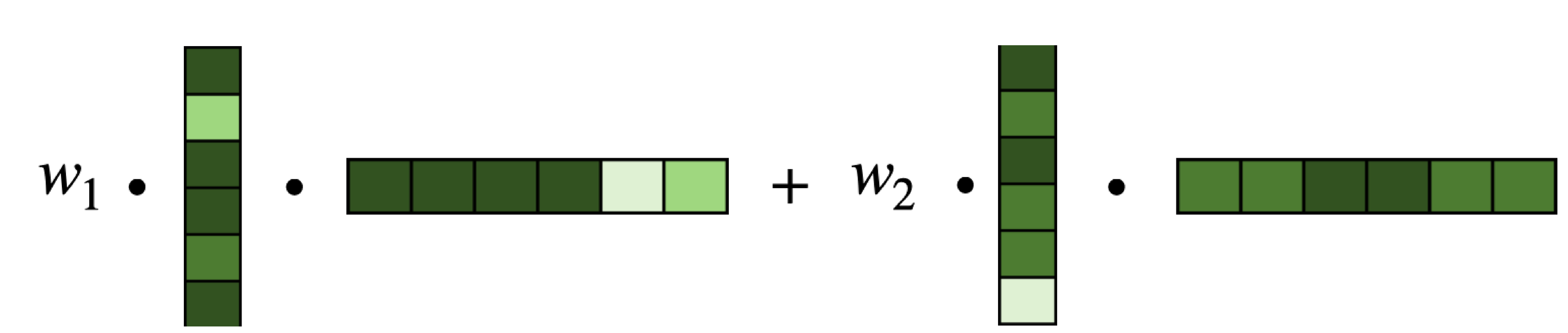
```
1: Use random seed s to select a uniformly random hash function h : {1,...,n} → [0,1].
2: Initialize I_A and V_A to be empty lists.
3: for i ∈ 1,...,n do
4:     Set threshold τ_i = k · ||A_i||_2^2 / ||A||_F^2.
5:     if h(i) ≤ τ_i then
6:         Append i to I_A, append A_i to V_A.
7: return S(A) = {I_A, V_A, τ_A} where τ_A = k/||A||_F^2.
```

Coordinated Sampling Method - Algorithm 1



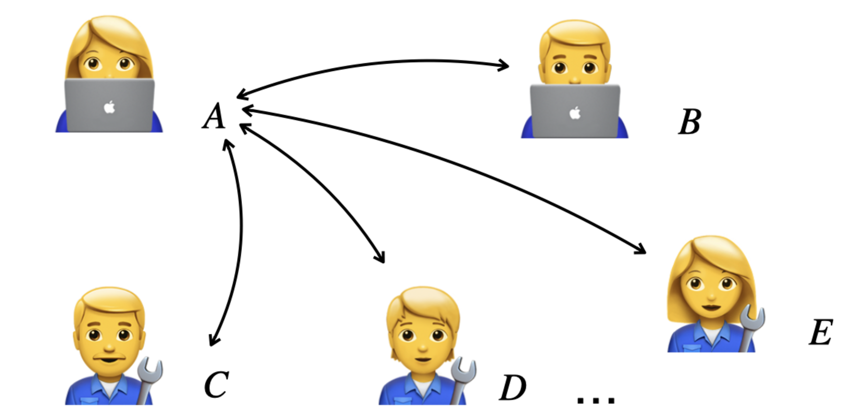
Like threshold sampling, coordinated sampling selects rows with probability proportional to row norms. Intuition: Rows with higher norms tend to contribute more to the product. However, coordinated sampling method dynamically set the threshold to collect exactly k samples.

Routine - Algorithm 2



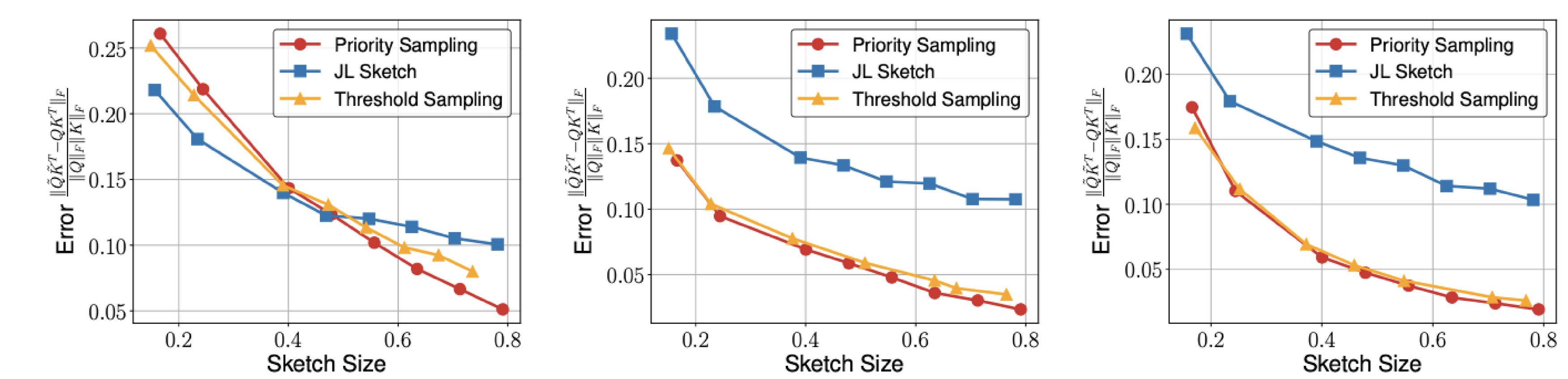
Intuition: $A^T B = \Sigma_{i\in[n]} A_i B_i^T \approx \Sigma_{i\in[T]} w_i A_i B_i^T = W$

Why Sketching Is Important



- We want a sketch for A to be interoperable with sketches for B, C, D, E, and any other matrices we might see in the future.
- Multi-vector retrieval application.
- Regression-based dataset search application.

Product Approximation on Attention Matrices



KV Cache

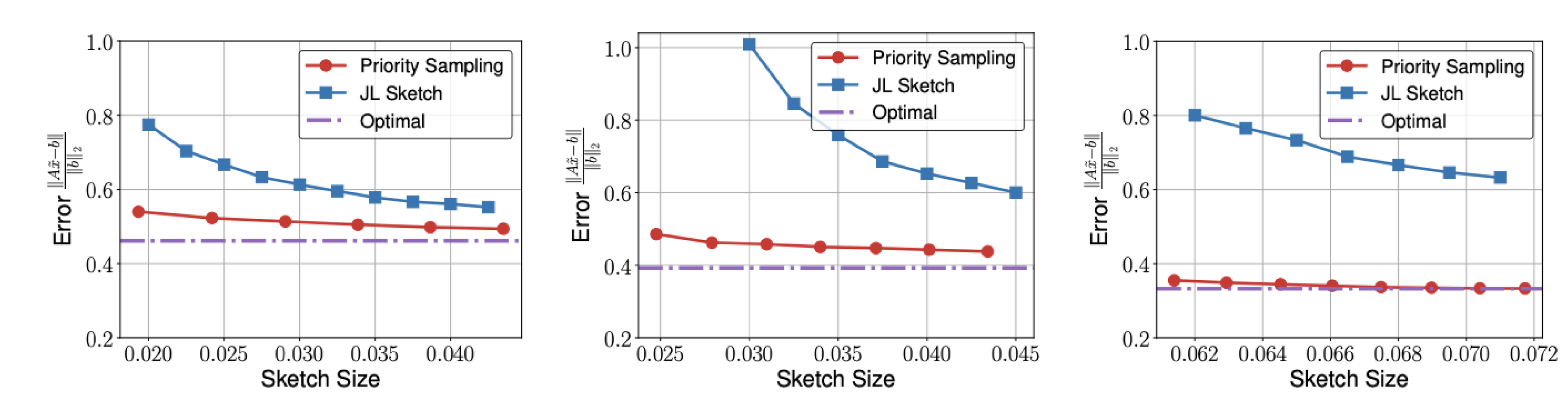
The matrices Q and K are generated from prompt tokens. As matrices sparsity increases from left to right, our method outperforms prior work.

Extension to Sketched Regression

Sketched Regression. There is a procedure that constructs sketches $\mathcal{S}(A)$ and $\mathcal{S}(b)$ consisting of $O(d/\epsilon)$ row samples from $A \in^{n\times d}$ and $b \in^n$ such that, using only the information in those sketches, we can compute $\tilde{x} \in^d$ satisfying, with probability at least 99/100,

$$\|A\tilde{x} - b\|_2^2 \le \|Ax^* - b\|_2^2 + \epsilon\|b\|_2^2.$$

Regression on Real World Dataset



Android Reviews

Matrix A is generated via SPLADE [6] on 10,000 random reviews. Vector b represents the review scores. As the matrices become sparser from left to right, our method improves over the best-known linear sketching methods.

References

[1] Sarlós, 2006 [2] Kane and Nelson, 2014 [3] Cohen et al., 2016 [4] Charikar et al., 2002
[5] Drineas et al., 2006 [6] Formal et al., 2022