Numerical Linear Algebra in the Sliding Window Model

Vladimir Braverman[1], Petros Drineas[2], Jalaj Upadhyay[1], David P. Woodruff[4], Samson Zhou[3]

samsonzhou@gmail.com

[1]Johns Hopkins University, [2]Purdue University, [3]Carnegie Mellon University, [4]Indiana University

PRELIMINARIES

- Input: Elements of an underlying data set S, which arrives sequentially
- Sliding Window: "Only the W most recent updates form the underlying data set S"
- Output: Evaluation (or approximation) of a given function
- Goal: Use space *sublinear* in the size of the input S

1 0 1 1 1 0 0 1 1 0 1

Question:
Are there space efficient algorithms for numerical linear algebra in the sliding window model?

1 3 5 -2
0 0 -1 3
2 5 6 1
8 7 2 1
-5 3 -4 -1
7 1 3 2

- Rows arrive one-by-one in the data stream
- Matrix $A \in R^{W\times n}$, $W \gg n$
- Recent interactions, time sensitive

RESULTS

Problem	Space
Deterministic ℓ_2 Spectral $(1+\varepsilon)$ Approximation (Sliding Window)	$\tilde{O}\left(\frac{n^3}{\epsilon}\right)$
ℓ_2 Spectral $(1+\varepsilon)$ Approximation (Sliding Window)	$\tilde{O}\left(\frac{n^2}{\epsilon^2}\right)$
$(1+\varepsilon)$ Rank k Approximation (Sliding Window)	$\tilde{O}\left(\frac{nk}{\epsilon^2}\right)$
$(1+\varepsilon)$ Rank k Approximation (Online)	$\tilde{O}\left(\frac{nk}{\epsilon^2}\right)$
Covariance Matrix Approximation (Sliding Window, Frobenius Norm Error)	$\tilde{O}\left(\frac{n}{\epsilon^2}\right)$

- Also results for ℓ_1 Spectral $(1+\varepsilon)$ Approximation when entries of A and x are bounded integers

- ℓ_2 Spectral $(1+\epsilon)$ Approximation: Given $\epsilon > 0$ and $A \in R^{W\times n}$, find matrix $M \in R^{m\times n}$ with $m \ll W$, such that for *every* $x \in R^n$,
$$(1-\epsilon)\|Ax\|_2 \le \|Mx\|_2 \le (1+\epsilon)\|Ax\|_2$$
- $(1+\epsilon)$ Rank k Approximation: Given $\epsilon > 0$ and $A \in R^{W\times n}$, find matrix $M \in R^{m\times n}$ with $m \ll W$ such that
$$(1-\epsilon)\|A-A_k\|_F \le \|M-M_k\|_F \le (1+\epsilon)\|A-A_k\|_F$$
- Covariance Matrix Approximation: Given $\epsilon > 0$ and $A \in R^{W\times n}, W \gg n$, find $B \in R^{d\times n}$ such that
$$\|A^\top A - B^\top B\|_F \le \epsilon\|A^\top A\|_F$$

APPROXIMATE MATRIX MULTIPLICATION

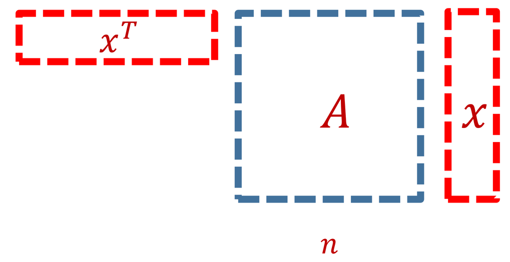
- Intuition: Large entries in $A^\top A$ come from large entries in A and suppose we know $\|A\|_F$
- Importance sampling: B = Sample row a_k of A with probability $p_k \propto \frac{\|a_k\|_2^2}{\|A\|_F^2}$ and rescale by $\frac{1}{\sqrt{p_k}}$.
- Analyze $\mathrm{E}[\|A^\top A - B^\top B\|_F^2]$ [DK01]

- Step 1: Show that $B^\top B$ is an unbiased estimator:
$$\mathrm{E}[B^\top B] = \sum p_k \left(\frac{1}{\sqrt{p_k}} a_k^\top \frac{1}{\sqrt{p_k}} a_k\right) = A^\top A$$
- Step 2: Bound the variance of $(B^\top B)_{i,j}$:
$$\mathrm{Var}[(B^\top B)_{i,j}] \le \sum \frac{1}{p_k}\left(a_k^\top a_k\right)_{i,j}^2$$
- Bound the expected error
$$\mathrm{E}[\|A^\top Awef - B^\top B\|_F^2] \le \sum_{i,j,k} \frac{1}{p_k}\left(a_k^\top a_k\right)_{i,j}^2 = \sum_k \frac{1}{p_k}\|a_k\|_2^4$$
- For $p_k = \frac{c\|a_k\|_2^2}{\|A\|_F^2}$, $\mathrm{E}[\|A^\top A - B^\top B\|_F^2] \le \frac{1}{c}\|A\|_F^4$.
- $\sum p_k = c := \frac{1}{\epsilon^2}$, so total number of sampled rows is $O\left(\frac{1}{\epsilon^2}\log n\right)$ w.h.p.

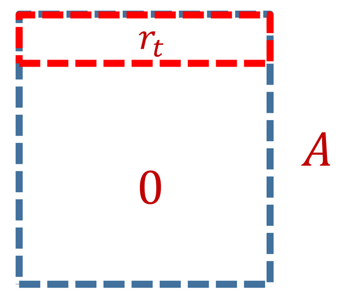
- Note it suffices to have $\widehat{A}$ a 2-approximation of $\|A\|_F^2$
- Why? Sample row a_i of A with probability $p_i \propto \frac{2\|a_i\|_2^2}{\widehat{A}}$
- Frobenius norm is *smooth*, can use smooth histogram to maintain $\widehat{A}$ [BO07]
- Suppose we have sampled row a_i of A with probability $p_i \propto \frac{\|a_i\|_2^2}{\|A\|_F^2}$
- New row arrives a_t: $\|A\|_F^2$ increases by $\|a_t\|_2^2$
- What do we do with a_i?
- Downsample: keep a_i with probability $\frac{\|A\|_F^2}{\|A\|_F^2 + \|a_t\|_2^2}$
- Sampled a_i with probability $p_i \propto \frac{\|a_i\|_2^2}{\|A\|_F^2 + \|a_t\|_2^2}$

- Separate instance of matrix multiplication streaming algorithm for each instance tracking the Frobenius norm
- Total space: $O\left(\frac{1}{\epsilon^2}\log n\right)$ rows $\rightarrow O\left(\frac{n}{\epsilon^2}\log^2 n\right)$ bits of space
- Can decrease to $O\left(\frac{n}{\epsilon^2}\log n\left(\log\log n + \log\frac{1}{\epsilon}\right)\right)$ with bit representation tricks
- Also give $\Omega\left(\frac{n}{\epsilon^2}\log n\right)$ space lower bound

ℓ_2 SPECTRAL APPROXIMATION

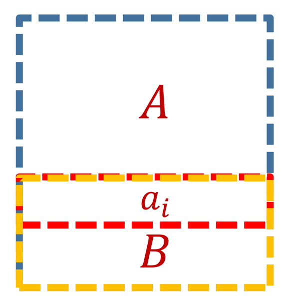


- Find a matrix B so that for all vectors x, $x^T Bx$ is a good approximation for $x^T Ax$
- Approximates *all* cuts of a graph
- Smooth histogram does not work!
- Johnson-Lindenstrauss based compression techniques also do not seem to help
- Intuition: If we tried to build a histogram, a lot of similar structure between instances: most rows are shared!
- Squared row norm sampling does not work!

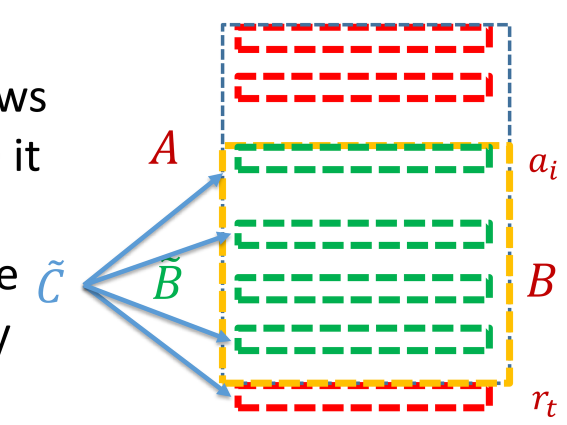


- We should *always* store the most recent row
- Need a new sense of importance for both recency AND uniqueness of a row
- Leverage score sampling does not work!

REVERSE ONLINE LEVERAGE SCORES



- Leverage score of row a_i is $\ell_i = a_i(A^\top A)^{-1}a_i^\top$
- Rows before a_i might be deleted so they shouldn't count towards the importance of a_i
- Reverse online leverage score of row a_i is $\tau_i = a_i(B^\top B)^{-1}a_i^\top$ where B are the rows after a_i



- Algorithm: sample (and rescale) a number of rows
- New row arrives – store it
- For each sampled (and rescaled) row a_i, sample the row with probability $\propto \tau_i \leftrightarrow$ downsampling
- How to ensure rows remain after repeated sampling? How to deal with compounding error?
- Correctness: Show an invariant that each row a_i is sampled with probability $\propto$ *final* reverse online leverage score (Suffices by [SS08])
- a_i remains with probability $\propto \tau_{\tilde{C}}\left(\frac{a_i}{\sqrt{p_i}}\right)$
- Reverse online leverage score:
$$\left(\frac{a_i}{\sqrt{p_i}}\right)(\tilde{C}^\top\tilde{C})^{-1}\left(\frac{a_i}{\sqrt{p_i}}\right)^\top = \left(\frac{a_i}{\sqrt{p_i}}\right)(\tilde{B}^\top\tilde{B} + r_t^\top r_t)^{-1}\left(\frac{a_i}{\sqrt{p_i}}\right)^\top$$
- Recall $(1-\epsilon)B^\top B \preccurlyeq \tilde{B}^\top\tilde{B} \preccurlyeq (1+\epsilon)B^\top B$, so $(1-\epsilon)C^\top C \preccurlyeq \tilde{C}^\top\tilde{C} \preccurlyeq (1+\epsilon)C^\top C$
- a_i survives w.p. $c_1\tau_C(a_i) \le p_i \le c_2\tau_C(a_i)$

LOW-RANK APPROXIMATION

- Reverse online leverage score: Sample each row a_i with probability $p_i \propto \tau_i = a_i(B^\top B + \lambda I_n)^{-1}a_i^\top$
- Issues: Compute $\lambda = \frac{\|A-A_k\|_F^2}{k}$, Bound $\sum\tau_i$
- Observation: it suffices to have a constant factor approximation of $\lambda = \frac{\|A-A_k\|_F^2}{k}$
- Use projection-cost preserving sketch [CEMMP15] to reduce the dimension of each row and feed reduced rows into spectral approximation algorithm
- Space used by the algorithm $\rightarrow$ Bounding the sum of the reverse online leverage scores

$$\begin{aligned}\det(A^\top A + \lambda I_n) &= \det(A_{W-1}^\top A_{W-1} + \lambda I_n)\,(1 + a_W(A_{W-1}^\top A_{W-1} + \lambda I_n)^{-1}a_W^\top)\\ &= \det(A_{W-1}^\top A_{W-1} + \lambda I_n)\,(1+\tau_W)\\ &\ge \det(A_{W-1}^\top A_{W-1} + \lambda I_n)(1 + e^{\tau_W/2})\end{aligned}$$
$$\det(A^\top A + \lambda I_n) \ge \lambda^n\, e^{\sum\tau_i/2}$$
$$\det(A^\top A + \lambda I_n) = \prod \sigma_i(A^\top A + \lambda I_n)$$

- Small singular values: $\sigma_{k+1} + \ldots + \sigma_n = \|A-A_k\|_F^2 + \lambda(n-k)$
- By AM-GM, $\prod_{i=k+1}^{i=n}\sigma_i \le \left(\frac{\|A-A_k\|_F^2 + \lambda(n-k)}{n-k}\right)^{n-k}$
- Large singular values: $\sigma_i \le \|A\|_2^2 + \lambda$ for $1 \le i \le k$
$$\log\det(A^\top A + \lambda I_n) = O(k\log n)$$
- $\sum\tau_i = O(k\log n)$
- Also gives a space efficient *online* algorithm for low-rank approximation!

ℓ_1 SPECTRAL APPROXIMATION

- Can show that if $\|Ax\|_1$ increases by $(1+\epsilon)$, $\|Ax\|_2^2$ must increase by $\left(1+\frac{\epsilon}{\mathrm{poly}(n)}\right)$
- Can use deterministic algorithm to find these breakpoints
- Use separate instances of streaming ℓ_1 spectral approximation algorithm starting at each of these breakpoints

REFERENCES

- [BO07] Vladimir Braverman, Rafail Ostrovsky. Smooth histograms for sliding windows. FOCS 2007
- [CEMMP16] Michael B. Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. Dimensionality reduction for k-means clustering and low rank approximation. STOC 2015
- [DK01] Petros Drineas, Ravi Kannan. Fast Monte-Carlo Algorithms for Approximate Matrix Multiplication. FOCS 2001
- [SS08] Daniel A. Spielman, Nikhil Srivastava. Graph sparsification by effective resistances. Funda Ergün, Hossein Jowhari, and Mert Saglam. Periodicity in streams. STOC 2008