A simple randomized algorithm for approximating the spectral norm of streaming data

Spencer Shortt

[1]Advisor: Stephen Becker
University of Colorado Boulder

[2]Committee Member: Kyle Luh
University of Colorado Boulder

[3]Committee Member: Sean O'Rourke
University of Colorado Boulder

April 2023

Motivation: Why approximate the spectral norm?

Spectral norms can be used as an error estimator when trying to approximate matrices.

Remark 2.1 (Martinson et. al., 2020)

For example, let us consider a variant of the spiked covariance model that is common in statistics applications. Suppose we need to approximate a rank-one matrix contaminated with additive noise: $A = \vec{u}\vec{u}^* + \in \mathbb{R}^{n\times n}$, where $\|\vec{u}\| = 1$ and $G \in \mathbb{R}^{n\times n}$ has independent entries from $\mathcal{N}(0, n^{-1})$ entries. With respect to the Frobenius norm, the zero matrix is almost as good an approximation of A as the rank-one matrix uu^*:

$$\mathbb{E}[\|A - \vec{u}\vec{u}^*\|_F^2] = \varepsilon^2 n \text{ and } \mathbb{E}[\|A - 0\|_F^2] = \varepsilon^2 n$$

An Existing Approach: Power Method

Liberty et. al., (2007)

Suppose A is an $m \times n$ complex-valued matrix and $\vec{\omega}$ is a $n \times 1$ column vector with i.i.d. entries from a complex gaussian distribution. With $\vec{\nu} = \frac{\vec{\omega}}{\|\vec{\omega}\|_2}$, we define

$$p_j(A) = \sqrt{\frac{\|(A^* A)^j \vec{\nu}\|_2}{\|(A^* A)^{j-1} \vec{\nu}\|_2}}.$$

Then $p_j(A) \geq \|A\|/10$ with probability greater than $1 - 4\sqrt{n/(j-1)}100^{-j}$, and $p_j(A) \leq \|A\|$ for all j.

Streaming Data

Let A be an $m \times n$ matrix of data, and suppose we went to append it with an $m \times k$ dataset B.

$$C = \left[A \mid B \right] = \left[A \mid 0_{m \times k} \right] + \left[0_{m \times n} \mid B \right] = A' + B'$$

By writing $C = A' + B'$ as above, we run into a potential storage issue. We must store both the old data A and new data B in order to calculate the spectral norm approximation:

$$p_j(C) = \sqrt{\frac{\left\| \left((A' + B')^* (A' + B') \right)^j \vec{\nu} \right\|_2}{\left\| \left((A' + B')^* (A' + B') \right)^{j-1} \vec{\nu} \right\|_2}}$$

A Different Approach:

Lemma 4.1 (Halko et. al., 2011)

Let A be a real $m \times n$ matrix. Fix a positive integer r and a real number $\alpha > 1$. Draw an independent family $\{\vec{\omega}_i : i = 1, 2, \ldots, N\}$ of standard Gaussian vectors. Then

$$\|A\| \leq \alpha \max_{i=1,2,\ldots,N} \|A\vec{\omega}_i\|$$

except with probability α^{-N}

Efficient Storage for Streaming Data

Let $\Omega_A = \begin{bmatrix} \vec{\omega}_1 & \vec{\omega}_2 & \dots & \vec{\omega}_N \end{bmatrix}$ be an $n \times N$ matrix whose columns are independent standard Gaussian vectors, and define

$$Y_A = A\Omega_A = \begin{bmatrix} A\vec{\omega}_1 & A\vec{\omega}_2 & \dots & A\vec{\omega}_N \end{bmatrix}.$$

To achieve the bound on the previous slide, calculate $\max_{i=1,2,\dots,N} \|A\vec{\omega}_i\|$

Suppose now that we append the $m \times k$ matrix B to A to get $C = \left[A \mid B \right]$. We let Ω_B be a $k \times N$ matrix whose columns are independent standard Gaussian vectors, and define $\Omega_C = \left[\frac{\Omega_A}{\Omega_B} \right]$. Then

$$Y_C = C\Omega_C = \left[A \mid B \right] \left[\frac{\Omega_A}{\Omega_B} \right] = A\Omega_A + B\Omega_B = Y_A + B\Omega_B,$$

implying that we only need to calculate $B\Omega_B$ after storing the $m \times N$ matrix Y_A.

Frobenius Norm is off by factor of $r^{1/2}$

We can write the Frobenius norm as the ℓ_2-norm of the singular values: $\|A\|_F = \sqrt{\sum_{i=1}^r \sigma_i^2}$. Using this and the fact that the spectral norm of A is the largest singular value of A, we have

$$\|A\| \leq \|A\|_F \leq r^{1/2}\|A\|$$

since

$$\sigma_{\max} \leq (\sum_{j=1}^{r} \sigma_r^2)^{1/2} \leq r^{1/2}\sigma_{\max}.$$

This tells us that the Frobenius norm can be off from the spectral norm by a factor of $r^{1/2}$.

Estimate is greater than Frobenius norm

We show $\mathbb{E}[\|A\vec{\omega}\|^2] = \|A\|_F^2$:

$$\mathbb{E}[\|A\vec{\omega}\|^2] = \mathbb{E}[\vec{\omega}^T A^T A\vec{\omega}] = \mathrm{Tr}(\mathbb{E}[\vec{\omega}^T A^T A\vec{\omega}]) = \mathbb{E}[\mathrm{Tr}(\vec{\omega}^T A^T A\vec{\omega})]$$

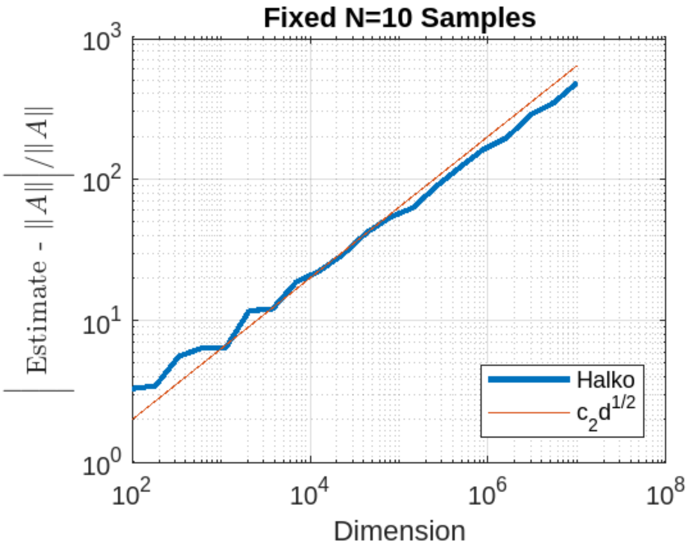
$$= \mathbb{E}[\mathrm{Tr}(A^T A\vec{\omega}\vec{\omega}^T)] = \mathrm{Tr}(\mathbb{E}[A^T A\vec{\omega}\vec{\omega}^T]) = \mathrm{Tr}(A^T A\mathbb{E}[\vec{\omega}\vec{\omega}^T])$$

$$= \mathrm{Tr}(A^T A) = \|A\|_F^2$$

Analyzing the bound given by Halko et. al. (2011), we see

$$\mathbb{E}[\max_{i=1,2,\ldots,N} \|A\vec{\omega}_i\|^2] \geq \mathbb{E}[\frac{1}{N}\sum_{i=1}^{N} \|A\vec{\omega}_i\|^2] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[\|A\vec{\omega}_i\|^2] = \|A\|_F^2$$

Plot of Error



The ℓ_4 norm is better

By the same arguement as before,

$$\|A\| \le \|\vec{\sigma}\|_4 \le r^{1/4}\|A\|.$$

One idea is to approximate the ℓ_4-norm of the singular values since this is a tighter bound.

Let $\vec{\omega}, \vec{\nu} \in \mathcal{N}(0, I_r)$ be independent gaussian random vectors. Define the random variable $X = (A\vec{\omega})^T A\vec{\nu}$. We will show $\mathbb{E}[X^2] = \|\vec{\sigma}\|_4^4$

WLOG, use diagonal matrices

Let $A = U\Sigma V^T$ be the singular value composition of our $m \times n$ matrix A. By orthogonality,

$$\mathbb{E}[X^2] = \mathbb{E}[((A\vec{\omega})^T A\vec{\nu})^2] = \mathbb{E}[(\vec{\omega}^T A^T A\vec{\nu})^2] = \mathbb{E}[(\vec{\omega}^T V\Sigma U^T U\Sigma V^T \vec{\nu})^2]$$

$$= \mathbb{E}[(\vec{\omega}^T V\Sigma^2 V^T \vec{\nu})^2] = \mathbb{E}[((V^T\vec{\omega})^T \Sigma^2 V^T \vec{\nu})^2]$$

Since $\vec{\omega}, \vec{\nu} \in \mathcal{N}(0, I_n)$, we have that $V^T\vec{\omega}, V^T\vec{\nu} \in \mathcal{N}(0, V^T V) = \mathcal{N}(0, I_n)$. Thus,

$$\mathbb{E}[((A\vec{\omega})^T A\vec{\nu})^2] = \mathbb{E}[((\Sigma\omega)^T \Sigma\nu)^2]$$

Furthermore, since Σ only has r non-zero values along it's diagonal, without loss of generality, we can let Σ be an $r \times r$ diagonal matrix from here on and have $\vec{\omega}, \vec{\nu} \in \mathcal{N}(0, I_r)$, and later on we will asume A to be the same.

Calculating the ℓ_4 norm:

$$\mathbb{E}[X^2] = \mathbb{E}[((\Sigma\vec{\omega})^T\Sigma\vec{\nu})^2] = \mathbb{E}[(\sum_{j=1}^{r}\sigma_j^2\omega_j\nu_j)^2]$$

$$= \mathbb{E}[\sum_{j=1}^{r}\sum_{k=1}^{r}\sigma_j^2\sigma_k^2\omega_j\omega_k\nu_j\nu_k] = \sum_{j=1}^{r}\sum_{k=1}^{r}\sigma_j^2\sigma_k^2\mathbb{E}[\omega_j\omega_k\nu_j\nu_k]$$

$$= \sum_{j=1}^{r}\sum_{k=1}^{r}\sigma_j^2\sigma_k^2\mathbb{E}[\omega_j\omega_k]\mathbb{E}[\nu_j\nu_k] = \sum_{j=1}^{r}\sigma_j^4 = \|\vec{\sigma}\|_4^4.$$

Practically speaking, we draw random vectors from a Gaussian distribution to create a sample mean to approximate $\mathbb{E}[X^2]$. Thus, we would like to show that the difference $\left|\frac{1}{N}\sum_{j=1}^{N}X_j^2 - \mathbb{E}[X^2]\right|$ is small with high probability.

Sub-Weibull Random Variables

We define X to be sub-Weibull random variable with tail parameter θ if

$$\mathbb{P}(|X| \geq x) \leq a \exp(-bx^{1/\theta}) \text{ for all } x > 0, \text{ for some } \theta, a, b > 0$$

Equivalently, a random variable is a sub-Weibull with tail parameter θ if there exists some constant $K_2 > 0$ such that

$$\|X\|_p := (\mathbb{E}[|X|^p])^{1/p} \leq K_2 p^\theta$$

for all $p \geq 1$.

Examples

Sub-Gaussian random variables have $\theta = 1/2$

Sub-Exponential have $\theta = 1$

$X = (D\vec{\omega})^T D\vec{\nu}$ is sub-exponential

Let A be a diagonal $r \times r$ matrix with positive diagonal entries σ_i, and let $\omega_i, \nu_i \in \mathcal{N}(0,1)$. Since ω_i, ν_i are sub-Gaussian, there exists a constant k such that for all $p \geq 1$,

$$\|\omega_i\|_p \leq kp^{1/2}.$$

Since $\|\cdot\|_p$ is a norm, we can use the triangle inequality on X:

$$\|X\|_p = \|\sum_{i=1}^{r} \sigma_i^2 \omega_i \nu_i\|_p \leq \sum_{i=1}^{r} \sigma_i^2 \|\omega_i \nu_i\|_p = \sum_{i=1}^{r} \sigma_i^2 (\mathbb{E}[|\omega_i|^p |\nu_i|^p])^{1/p}.$$

By independence, the above equals

$$\sum_{i=1}^{r} \sigma_i^2 (\mathbb{E}[|\omega_i|^p])^{1/p} (\mathbb{E}[|\nu_i|^p])^{1/p} \leq \sum_{i=1}^{r} \sigma_i^2 (kp^{1/2})(kp^{1/2}) = k^2 p \|A\|_F^2$$

We care about X^2, but there's a problem

$X^2 = \left((D\vec{\omega})^T D\vec{\nu}\right)^2$ is sub-Weibull with parameter $\theta = 2$:

$$\|X^2\|_p = (\mathbb{E}[|X^2|^p])^{1/p} = \left((\mathbb{E}[|X|^{2p}])^{1/2p}\right)^2 = \left(\|X\|_{2p}\right)^2$$

$$\leq \left(\|A\|_F^2 k^2 (2p)\right)^2 = 4k^4 \|A\|_F^4 p^2.$$

We would like to use concentration properties of sub-Weibull random variables to show the difference $\left|\frac{1}{N}\sum_{j=1}^N X_j^2 - \mathbb{E}[X^2]\right|$ is small with high probability.

Sub-Weibull Theorems

Corollay 3.1 (Vladimirova et. al., 2020)

Let $X_1, \ldots, X_n$ be identically distributed sub-Weibull random variables with tail parameter θ. Then, for all $x \geq NK_\theta$, we have

$$\mathbb{P}(|\sum_{i=1}^{N} X_i| \geq x) \leq \exp(-(\frac{x}{NK_\theta}))$$

for some constant K_θ dependent on θ.

The problem is that for our situation, K_θ is proportional to $1/N$.

Sub-Weibull theorems

Theorem 3.1 (Kuchibhotla et. al., 2022)

If $X_1, \dots, X_n$ are independent mean zero random variables with $\|X_i\|_{\psi_\alpha} < \infty$ for all $1 \leq i \leq n$ and some $\alpha > 0$, then for any vector $(a_1, \dots, a_n) \in \mathbb{R}^n$, then we have

$$\mathbb{P}(|\sum_{i=1}^{n} a_i X_i| \geq 2eC(\alpha)\|b\|_2\sqrt{t} + 2eL_n^*(\alpha)t^{1/\alpha}\|b\|_{\beta(\alpha)}) \leq 2e^{-t}$$

for all $t \geq 0$, where $b = (a_1\|X_1\|_{\psi_\alpha}, \dots, a_n\|X_n\|_{\psi_\alpha}) \in \mathbb{R}^n$.

Another attempt

Theorem:

Let A be an $m \times n$ real-valued matrix with rank $r > 16$. Draw $\vec{\omega}_i$ and $\vec{\nu}_i$ independently from $\mathcal{N}(0, I_n)$ for all $i \in \{1, \ldots, N\}$. If we define $X_i = (A\vec{\omega}_i)^T A\vec{\nu}_i$, then there exists a constant $K > 0$ such that for any $t > 0$,

$$\left| \frac{1}{N} \sum_{i=1}^{N} \left| X_i \right|^{1/2} - \|A\| \right| \leq (r^{1/4} - 1)\|A\| + t,$$

with probability greater than $1 - 2\exp(-\frac{Nt^2}{Kr\|A\|^2})$.

This theorem is far from ideal.
If $\|A\| \leq \frac{1}{N}\sum_{i=1}^{N} \left|X_i\right|^{1/2}$, we have that $\frac{1}{N}\sum_{i=1}^{N} \left|X_i\right|^{1/2} \leq r^{1/4}\|A\| + t$ and is actually a slightly better approximation than our estimator $\frac{1}{N}\sum_{i=1}^{N} X_i^2$.
However, it is not guaranteed that $\|A\| \leq \frac{1}{N}\sum_{i=1}^{N} \left|X_i\right|^{1/2}$.

(Proof) Concave Jensen

We use the concave version of Jensen's inequality:

$$\mathbb{E}[|X|^{1/2}] = \mathbb{E}[|X|^{2/4}] \leq \left(\mathbb{E}[X^2]\right)^{1/4} = \|\vec{\sigma}\|_4$$

If $\|A\| \leq \mathbb{E}[|X|^{1/2}]$,

$$\mathbb{E}[|X|^{1/2}] - \|A\| \leq r^{1/4}\|A\| - \|A\| = (r^{1/4} - 1)\|A\|,$$

and if $\|A\| \geq \mathbb{E}[|X|^{1/2}]$,

$$\|A\| - \mathbb{E}[|X|^{1/2}] \leq \|A\| \leq (r^{1/4} - 1)\|A\|$$

Thus we have a bound on the absolute value of the error.

(Proof) $X^{1/2}$ is sub-Gaussian

The advantage of using $|X|^{1/2}$ is that it is sub-Gaussian with constant proportional to $\|A\|_F$. Using Jensen's inequality again, we see

$$\||X|^{1/2}\|_p = (\mathbb{E}[|X|^{p/2}])^{1/p} \leq \left((\mathbb{E}[|X|^p])^{1/p}\right)^{1/2} = \left(\|X\|_p\right)^{1/2} \leq k\|A\|_F p^{1/2}$$

Thus, we will apply general Hoeffding's inequality to show $\mathbb{E}[|X|^{1/2}]$ can be closely approximated by $\frac{1}{N}\sum_{j=1}^{N} \left|X_j\right|^{1/2}$ with high probability.

(Proof) General Hoeffding's Inequality

Given a random variable X, we define the sub-Gaussian norm of X to be

$$\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}[\exp(X^2/t^2) \leq 2]$$

General Hoeffding's Inequality (Vershynin, 2018)

Let $X_1, X_2, ..., X_N$ be independent, mean zero, sub-gaussian random variables, and $a = (a_1, a_2, ..., a_N) \in \mathbb{R}^N$. Then for every $t \geq 0$

$$\mathbb{P}\Big(\Big|\sum_{j=1}^{N} a_j X_j\Big| \geq t\Big) \leq 2\exp\left(-\frac{ct^2}{K^2\|a\|_2^2}\right)$$

where $K = \max_j \|X_j\|_{\psi_2}$

(Proof) Applying Hoeffding

Using the triangle inequality,

$$\||X|^{1/2} - \mathbb{E}[|X|^{1/2}]\|_p \le \||X|^{1/2}\|_p + \|\mathbb{E}[|X|^{1/2}]\|_p \le k\|A\|_F p^{1/2} + \mathbb{E}[|X|^{1/2}]$$

$$\le k\|A\|_F p^{1/2} + r^{1/4}\|A\| p^{1/2} \le r^{1/2}(k+1)\|A\| p^{1/2}.$$

We can assert that $\||X|^{1/2} - \mathbb{E}[|X|^{1/2}]\|_{\psi_2} = Cr^{1/2}(k+1)\|A\|$ for some constant $C > 0$.

Applying Hoeffding

This lets us apply Hoeffding to the subgaussian random variables $\tilde{X}_j = |X_j|^{1/2} - \mathbb{E}[|X|^{1/2}]$ with $a_j = 1/N$ for all j and $K = C^2(k+1)^2/c$:

$$\mathbb{P}\Big(|\frac{1}{N}\sum_{j=1}^{N}|X_j|^{1/2} - \mathbb{E}[|X|^{1/2}]| \geq t\Big) \leq 2\exp\left(-\frac{Nt^2}{Kr\|A\|^2}\right)$$

(Proof) Conclusion

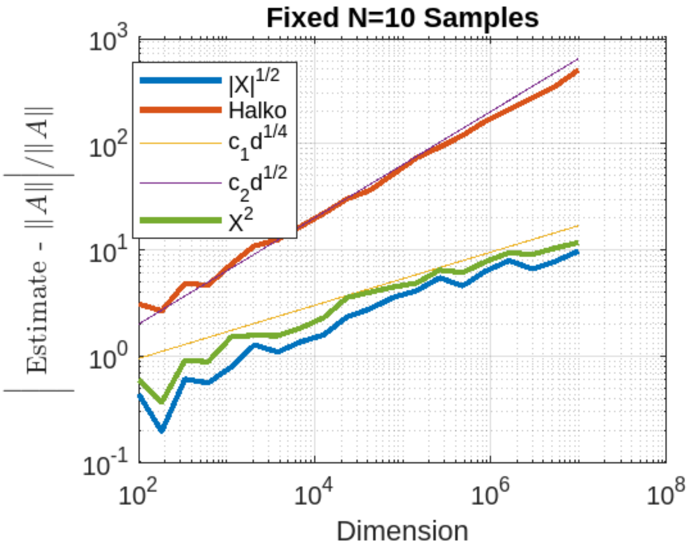
Finally, by the triangle inequality,

$$\left|\frac{1}{N}\sum_{i=1}^{N}\left|X_i\right|^{1/2} - \|A\|\right| \leq \left|\mathbb{E}[|X|^{1/2}] - \|A\|\right| + \left|\frac{1}{N}\sum_{i=1}^{N}\left|X_i\right|^{1/2} - \mathbb{E}[|X|^{1/2}]\right|$$

$$\leq (r^{1/4} - 1)\|A\| + t$$

with probability greater than $1 - 2\exp(-\frac{Nt^2}{Kr\|A\|^2})$.

Conclusion



Fixed N=10 Samples
|X|^1/2
Halko
c_1 d^1/4
c_2 d^1/2
X^2
|Estimate - ||A||| / ||A||
Dimension
10^-1
10^0
10^1
10^2
10^3
10^2
10^4
10^6
10^8

- N. Halko, P. G. Martinsson, J. Tropp,n *Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions* , SIAM Review **53**(2), (2011), 217-288.
- E. Liberty, F. Woolfe, P.G. Martinsson, V. Rokhlin, M. Tygert, *Randomized algorithms for the low-rank approximation of matrices*, Proc. Natl. Acad. Sci. **104**(51) (2007), 20167–20172.
- A. Kuchibhotla, A. Chakrabortty, *Moving beyond sub-Gaussianity in high-dimensional statistics: applications in covariance estimation and linear regression*, Inf. Inference **11**(4) (2022), 1389-1456.

- P. Martinsson, J. Tropp, Randomized numerical linear algebra: Foundations and algorithms. Acta Numerica, **29** (2020), 403–572.
- R. Vershynin, *High dimensional probability. An introduction with applications in Data Science*. Cambridge University Press, 2018.
- M. Vladimirova, S. Girard, H. Nguyen, J. Arbel *Sub-Weibull distributions: Generalizing sub-Gaussian and sub-Exponential properties to heavier tailed distributions.* Stat. **9** (2020)

Highlighting text

In this slide, some important text will be highlighted because it's important. Please, don't abuse it.

Remark

Sample text

Important theorem

Sample text in red box

Examples

Sample text in green box. The title of the block is "Examples".