Random perturbation of low rank matrices and applications

Van H. Vu

Department of Mathematics
Yale University

Department of Mathematics
NUS

joint works with S. O'rourke (Colorado) and K. Wang (HKUST)

Let A be a large matrix of size $N \times n$; A is *sparse* if the majority of its entries is zero.

Let A be a large matrix of size $N \times n$; A is *sparse* if the majority of its entries is zero.

Given A, one often wants/needs to compute its key spectral parameters: leading singular values, singular vectors etc

It is easier (faster) to compute with sparse matrices.

Matrix sparsification problem. Replace A by a sparse matrix $S(A)$ which approximates the key parameters of A well.

Matrix completion

From a matrix A one can only observe a small (random) subset of the entries. Recover A.

Matrix completion

From a matrix A one can only observe a small (random) subset of the entries. Recover A.

B= observed matrix. B is sparse. Find $R(B)$ that approximate A well.

Matrix completion

From a matrix A one can only observe a small (random) subset of the entries. Recover A.

B= observed matrix. B is sparse. Find $R(B)$ that approximate A well.

Two problems are closely related. We will focus on the Sparsification, but similar results can be obtained for the other problem.

Random sparsification

$A = N \times n$ matrix with real entries a_{ij}.

Random sparsification

$A = N \times n$ matrix with real entries a_{ij}.

Set $p_{ij} := \frac{1}{2}(\frac{a_{ij}^2}{\|A\|_2^2} + \frac{|a_{ij}|}{\|A\|_1})$.

$$\|A\|_2^2 = \sum_{i,j} a_{ij}^2; \|A\|_1 = \sum_{i,j} |a_{ij}|.$$

Random sparsification

$A = N \times n$ matrix with real entries a_{ij}.

Set $p_{ij} := \frac{1}{2}(\frac{a_{ij}^2}{\|A\|_2^2} + \frac{|a_{ij}|}{\|A\|_1})$.

$$\|A\|_2^2 = \sum_{i,j} a_{ij}^2; \|A\|_1 = \sum_{i,j} |a_{ij}|.$$

. The numbers p_{ij} sums up to 1 and form a distribution.

Random sparsification

$A = N \times n$ matrix with real entries a_{ij}.

Set $p_{ij} := \frac{1}{2}(\frac{a_{ij}^2}{\|A\|_2^2} + \frac{|a_{ij}|}{\|A\|_1})$.

$$\|A\|_2^2 = \sum_{i,j} a_{ij}^2; \|A\|_1 = \sum_{i,j} |a_{ij}|.$$

. The numbers p_{ij} sums up to 1 and form a distribution.
Choose one entry ij with respect to this distribution and define by B the matrix with exactly one nonzero entries $b_{ij} := a_{ij} p_{ij}^{-1}$.
Clearly B is very sparse (having only one huge non-zero entry).
Furthermore,

$$\mathbf{E}B = A.$$

Of course B cannot be a good approximation of A.

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \quad vs \quad \begin{bmatrix} N & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

Resampling trick

Repeat the experiment m times: iid matrices $B_1, \ldots, B_m$.

Resampling trick

Repeat the experiment m times: iid matrices $B_1, \ldots, B_m$.

Set $S(A) := \frac{1}{m} \sum_{i=1}^{m} B_i$.

Resampling trick

Repeat the experiment m times: iid matrices $B_1, \ldots, B_m$.

Set $S(A) := \frac{1}{m} \sum_{i=1}^{m} B_i$.

We still have $\mathbf{E}S(A) = A$. But now m non-zeroes entries of $S(A)$ spread out.

Revisiting the identity matrix

Take $m = Nk$. Each diagonal is expected to appear k times. Set $k = \log^2 N$, then with probability $1 - o(1)$ each diagonal entries appears $(1 + o(1))k$ times.

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \quad vs \quad \begin{bmatrix} \frac{(1+o(1))kN}{m} & 0 & 0 & \dots & 0 \\ 0 & \frac{(1+o(1))kN}{m} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & \frac{(1+o(1))kN}{m} \end{bmatrix}$$

An approximation theorem

Assume $N \geq n$.

Theorem (Kundu-Drineas 2014)

For every $\epsilon > 0$, there exists $C > 0$ such that if A is an $N \times n$ matrix and $m \geq N \log N$, then

$$\|A - S(A)\| \leq C\|A\|_2 \sqrt{\frac{N \log N}{m}}$$

with probability at least $1 - \epsilon$.

How good is the bound ?

How good is the bound ?

Trivial approximation $S(A) = 0$ gives $\|A - S(A)\| = \|A\|$.

How good is the bound ?

Trivial approximation $S(A) = 0$ gives $\|A - S(A)\| = \|A\|$.

We want

$$R := \|A\|_2 \sqrt{\frac{N \log N}{m}} = o(\|A\|).$$

A has singular values $\|A\| = \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$

$$\|A\|_2^2 = \sigma_1^2 + \cdots + \sigma_n^2.$$

We want $R^2 = o(\|A\|^2)$, which means

We want $R^2 = o(\|A\|^2)$, which means

$$\|A\|_2^2 = o(\frac{m}{N \log N}\|A\|^2).$$

$$\sigma_1^2 + \cdots + \sigma_n^2 = o(\frac{m}{N \log N}\sigma_1^2).$$

If $m = Nn^\delta$, then to guarantee success, we need to assume

$$\sigma_1^2 + \cdots + \sigma_n^2 = o(n^\delta \sigma_1^2).$$

We want $R^2 = o(\|A\|^2)$, which means

$$\|A\|_2^2 = o(\frac{m}{N \log N}\|A\|^2).$$

$$\sigma_1^2 + \cdots + \sigma_n^2 = o(\frac{m}{N \log N}\sigma_1^2).$$

If $m = Nn^\delta$, then to guarantee success, we need to assume

$$\sigma_1^2 + \cdots + \sigma_n^2 = o(n^\delta \sigma_1^2).$$

This is true if A has low rank or is approximately low rank.

How to use the theorem ?

We have

$$\|A - S(A)\| \leq R.$$

How to use the theorem ?

We have

$$\|A - S(A)\| \leq R.$$

Assume that we want to compute $\sigma_1(A) = \|A\|$, we end up computing $\sigma_1(S(A)) = \|S(A)\|$, which is faster.

How to use the theorem ?

We have

$$\|A - S(A)\| \leq R.$$

Assume that we want to compute $\sigma_1(A) = \|A\|$, we end up computing $\sigma_1(S(A)) = \|S(A)\|$, which is faster.

How accurate is the result ?

How to use the theorem ?

We have

$$\|A - S(A)\| \leq R.$$

Assume that we want to compute $\sigma_1(A) = \|A\|$, we end up computing $\sigma_1(S(A)) = \|S(A)\|$, which is faster.

How accurate is the result ?

Weyl's inequality

$$-\|A - B\| \leq \|A\| - \|B\| \leq \|A - B\|.$$

How to use the theorem ?

We have

$$\|A - S(A)\| \leq R.$$

Assume that we want to compute $\sigma_1(A) = \|A\|$, we end up computing $\sigma_1(S(A)) = \|S(A)\|$, which is faster.

How accurate is the result ?

Weyl's inequality

$$-\|A - B\| \leq \|A\| - \|B\| \leq \|A - B\|.$$

So

$$-R \leq \|A\| - \|S(A)\| \leq R.$$

$$\text{(Relative form) } |\frac{\|S(A)\|}{\|A\|} - 1| \leq \frac{R}{\|A\|} := \epsilon.$$

How to use the theorem ?

We have

$$\|A - S(A)\| \leq R.$$

How to use the theorem ?

We have

$$\|A - S(A)\| \leq R.$$

Assume that we want to compute the first singular vector $v_1(A)$, we end up computing $v_1(S(A)) = \|S(A)\|$, which is faster.

How to use the theorem ?

We have

$$\|A - S(A)\| \leq R.$$

Assume that we want to compute the first singular vector $v_1(A)$, we end up computing $v_1(S(A)) = \|S(A)\|$, which is faster.

How accurate is the result ?

How to use the theorem ?

We have

$$\|A - S(A)\| \leq R.$$

Assume that we want to compute the first singular vector $v_1(A)$, we end up computing $v_1(S(A)) = \|S(A)\|$, which is faster.

How accurate is the result ?

Wedin-Davis-Kahan inequality

$$\|v_1(A) - v_1(B)| \leq \frac{\|A - B\|}{\sigma_1(A) - \sigma_2(A)}.$$

How to use the theorem ?

We have

$$\|A - S(A)\| \leq R.$$

Assume that we want to compute the first singular vector $v_1(A)$, we end up computing $v_1(S(A)) = \|S(A)\|$, which is faster.

How accurate is the result ?

Wedin-Davis-Kahan inequality

$$\|v_1(A) - v_1(B)| \leq \frac{\|A - B\|}{\sigma_1(A) - \sigma_2(A)}.$$

So

$$\|v_1(A) - v_1(S(A))\| \leq \frac{R}{\sigma_1(A) - \sigma_2(A)}.$$

New bound for low rank matrices

Recall the relative bound

$$\left|\frac{\|S(A)\|}{\|A\|} - 1\right| \leq \frac{R}{\|A\|} := \epsilon.$$

Assume that A has rank $r = o(\log N)$. Then, with probability $1 - o(1)$

New bound for low rank matrices

Recall the relative bound

$$|\frac{\|S(A)\|}{\|A\|} - 1| \leq \frac{R}{\|A\|} := \epsilon.$$

Assume that A has rank $r = o(\log N)$. Then, with probability $1 - o(1)$

$$|\frac{\|S(A)\|}{\|A\|} - 1| \leq \frac{R}{\|A\|} = O(\epsilon^2).$$

(O'rourke-V.-Wang 2017-2018)

Further new results

- Improvements for other leading singular values, singular vectors etc.
- Larger rank r.

Matrix completion. In the observable matrix B, each entry $b_{ij} = a_{ij}$ with prob. p and 0 otherwise.

Further new results

- Improvements for other leading singular values, singular vectors etc.
- Larger rank r.

Matrix completion. In the observable matrix B, each entry $b_{ij} = a_{ij}$ with prob. p and 0 otherwise.

Thus $b_{ij} = pa_{ij} + e_{ij}$ where e_{ij} is a random variable with mean 0. So

$$B = pA + E.$$

Spectral norm of random matrices

How to prove the Kundu-Drineas bound ?

Recall $\mathbf{E}S(A) = A$. So

$$S(A) = A + E$$

where E is a zero-mean random matrix.

Tools from random matrix theory (matrix Chernoff bound) show that with high probability

$$\|A - S(A)\| = \|E\| \leq R.$$

This bound is sharp.

The key observation

The use of Weyl's inequality is *wasteful*.

The key observation

The use of Weyl's inequality is *wasteful*.

Weyl's inequality

$$\|A + B\| \leq \|A\| + \|B\|.$$

The key observation

The use of Weyl's inequality is *wasteful*.

Weyl's inequality

$$\|A + B\| \leq \|A\| + \|B\|.$$

This is sharp, equality can happen if A and B (symmetric matrices) share the leading eigenvector v. Then

$$\|A + B\| = v^T(A + B)v = v^T Av + v^T Bv = \|A\| + \|B\|.$$

We have

$$S(A) = A + E.$$

The leading eigenvectors of A and E have no reason to point to the same direction, as E is random.

The key observation

The use of Weyl's inequality is *wasteful*.

Weyl's inequality

$$\|A + B\| \leq \|A\| + \|B\|.$$

This is sharp, equality can happen if A and B (symmetric matrices) share the leading eigenvector v. Then

$$\|A + B\| = v^T(A + B)v = v^T Av + v^T Bv = \|A\| + \|B\|.$$

We have

$$S(A) = A + E.$$

The leading eigenvectors of A and E have no reason to point to the same direction, as E is random.

In general the leading eigenvectors of E is far from the space spanned by the eigenvectors $v_1, \ldots, v_r$ of A.

Perturbation theory

Classical. How does a perturbation E changes the key spectral parameters of A: $A + E$ vs A.

Perturbation theory

Classical. How does a perturbation E changes the key spectral parameters of A: $A + E$ vs A.

Random. How does a random perturbation E changes to key spectral parameters of A.

Perturbation theory

Classical. How does a perturbation E changes the key spectral parameters of A: $A + E$ vs A.

Random. How does a random perturbation E changes to key spectral parameters of A.

In the case that A has low rank, one can do much better in the random setting than just applying results from the classical literature.

Perturbation theory

Classical. How does a perturbation E changes the key spectral parameters of A: $A + E$ vs A.

Random. How does a random perturbation E changes to key spectral parameters of A.

In the case that A has low rank, one can do much better in the random setting than just applying results from the classical literature.

In many problems in data science, the solutions (or inputs) have the form $A + E$, with A being deterministic, low rank, and E random.