

Workshop on Modern Massive Data Matrices,  
Palo Alto 2006.

**General Thoughts** : Bad historical accident  
that Numerical and Scientific Computation and  
Algorithms and Complexity do not have more  
to do with each other.

NA has a couple of centuries work to offer  
Alg.'s.

Alg.'s (once you get beyond our seeming ob-  
session with poly time) has a lot to offer. Ran-  
domization is certainly one of them....



Here simpler notion of **good** :

$$A^T A \approx R^T R.$$

**Notation**  $A$  is  $m \times n$ .

Number of rows in sample =  $s$ . ( $s \ll m, n$ .)

Quickly : Could mean polynomial time.

Here : Massive Matrices. Perhaps cannot be stored in full in RAM. [More generally, models of computation for handling massive data - eg. the [streaming](#) model.....

Quickly : In one or two passes thro'  $A$ .

Randomization will help.

Uniform random sample won't do : All but one row zero !!

Sample with probabilities depending on size of entries in row.

The Length-squared distribution : Pick rows with probabilities proportional to their squared lengths : Make  $s$  i.i.d. trials. In each trial, pick a row  $A_{(i)}$  (the  $i$  th row of  $A$ ) with

$$\text{Probability of picking row } i = \frac{|A_{(i)}|^2}{\|A\|_F^2}.$$

If  $A_{(i)}$  is picked, include a scaled version :  $A_{(i)}/\sqrt{sP_i}$  as the next row of  $R$ .

If all row lengths are equal, uniform sampling will do and no scaling is necessary.

[In fact, same if all row lengths are within  $O(1)$  of each other.]

## Two Properties of the sampling

$$\text{Unbiased } E(R^T R) = A^T A.$$

This distribution **minimizes** the **total variance**

$$E\|A^T A - R^T R\|_F^2.$$

[Measuring  $E\|A^T A - R^T R\|_F^2$  greatly simplifies the expression.]

For most results, **approx length-squared** distribution, where probability of picking row  $i$  is at least  $\frac{c|A_{(i)}|^2}{\|A\|_F^2}$  suffices.

— Frieze, K., Vempala (1998)

Many other properties of the distribution - fast SVD.....

How **good** is this sample ?

$s$  is the number of sampled rows.

**Lemma** For every matrix  $A$ ,

$$E\|R^T R - A^T A\|_F^2 \leq \frac{1}{s}\|A\|_F^4.$$

Only interesting if

$$\frac{\|A\|_F^4}{\|A^T A\|_F^2}$$

is small.

Condition equivalent to

The top  $O(1)$  singular values form  $\Omega(1)$  part of the “spectrum” of  $A$ .

Above for  $A^T A$  can be generalized to multiplying any two matrices. – Drineas, K.

$R^T R \approx A^T A$  implies the singular values of  $R \approx$  the singular values of  $A$ . Can be quantified by Hoffman-Wielandt inequality.

**More difficult questions** Can one also say the same about the singular vectors of  $A, R$  ? Is there a sense in which

$$R \approx A?$$

## No free lunch

We cannot hope to pick from any general  $m \times n$  matrix, a set of  $s \ll m, n$  rows to form an  $R$  with  $R^T R$  close to  $A^T A$ . Call a matrix  $A$  a PCA matrix if for  $k \in O(1)$  :

$$\lambda_1(A^T A) + \lambda_2(A^T A) + \dots + \lambda_k(A^T A) \geq c \|A\|_F^2.$$

Then, above says :  $E \|R^T R - A^T A\|_F^2 \leq \epsilon \|A^T A\|_F^2$  for  $s \in O(1)$ . Myriad applications of Principal Component Analysis (assume matrix is a PCA matrix or more strongly that they are numerically low-rank) include :

Consumer-Product matrices

Document-term matrices

Test scores- Students matrices....

TCS contribution : Low-rank approximations to matrices and their extensions to tensors can also help solve combinatorial optimization problems.

## Approximating $A$ itself

Suppose  $C$  is a random subset of  $s$  columns of  $A$  picked according to the length-squared distribution (and scaled as above) and  $R$  is a subset of  $s$  rows of  $A$  " " ". From just  $C, R$ , we can find an  $s \times s$  matrix  $U$  such that

$$E\|A - CUR\|_F \leq \|A - A_{s^{1/5}}\|_F + \frac{4}{s^{1/5}}\|A\|_F,$$

$$E\|A - CUR\|_2 \leq \|A - A_{s^{1/5}}\|_2 + \frac{4}{s^{1/5}}\|A\|_F,$$

[where,  $A_k$  is the best rank  $k$  approximation to  $A$  and  $\|A\|_2$  denotes the spectral norm.] – Drineas, K., also Drineas, K., Mahoney.

$$\begin{pmatrix} A \end{pmatrix} \approx \quad (2)$$

$$\begin{pmatrix} C \end{pmatrix} \cdot \begin{pmatrix} U \end{pmatrix} \cdot \begin{pmatrix} R \end{pmatrix} \quad (3)$$

Sparsity preserved

Further **matrix-vector products**  $Ax$  can be approximated by  $C(U(Rx))$ .

## Matrix Reconstruction

$m$  users and  $n$  products.  $A_{ij}$  measures the preference of user  $i$  for product  $j$ .

Suppose we have observed some entries of the matrix. Can we infer the other entries? [So, having observed some market behaviour, we want to recommend to users what they would like.]

[Recommendations Systems / Collaborative filtering]

Azar, Fiat, Karlin, McSherry and Saia

Achlioptas and McSherry

Drineas, Kerenidis and Raghavan

Achlioptas and McSherry's algorithm :

$p$  probability. Independently for each entry  $A_{ij}$  of matrix, replace it with  $A_{ij}/p$  with probability (w.p)  $p$  and 0 with probability  $1 - p$ . So, number of non-zero entries reduced by a factor of  $p$ .

$$\hat{A}_{ij} = \begin{cases} 0 & \text{w.p. } 1 - p \\ A_{ij}/p & \text{w.p. } p. \end{cases}$$

$$\begin{pmatrix} 5 & 3 & 3 & -2 & -7 & 8 & 9 \\ 1 & 2 & 2 & -17 & 1 & -8 & 9 \\ 21 & 41 & 22 & -2 & 0 & 0 & 0 \end{pmatrix} \rightarrow \quad (4)$$

$$\begin{pmatrix} 10 & 6 & 0 & 0 & -14 & 16 & 0 \\ 2 & 4 & 0 & 0 & 0 & -16 & 18 \\ 0 & 0 & 44 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (5)$$

If  $|A_{ij}| \leq 1$ , **WHP**,  $\|A - \hat{A}\|_2$  is small.

\*\*\* [Coming Attractions](#) See Achlioptas's talk.

\*\*\*\*

“Exponential convergence” of Kaczmarz equation solver :

$$Ax = b$$

At iteration  $k$  : have  $x_k$ . Get  $x_{k+1}$  by adding to  $x_k$  the l.h.s. of any violated equation suitably scaled :

$$x_{k+1} = x_k + \frac{b_i - (A_{(i)} \cdot x_k)}{|A_{(i)}|} A_{(i)}.$$

Strohmer and Vershynin (2006):  $A$  is  $m \times n$  with rank  $n$ . If at each step,  $i$  is chosen according to the length-squared distribution, then for  $x^*$  with  $Ax^* = b$ ,

$$E|x_k - x^*|^2 \leq \left(1 - \frac{1}{R^2}\right)^k |x^* - x_0|^2,$$

where  $R = \|A\|_F / \sigma_{\min}(A)$ .

What is wrong with the length-squared distribution ?

An Example :  $A$  has the first  $m - 1$  rows all equal and the last row orthogonal to them; all rows are of length 1. [Drineas, Vempala]

Best rank 2 approximation :  $A$  itself. **Error=0.**

**Repeated sampling only yields the first vector.**  
**Error  $\not\sim O(\text{best error})$  !!**

Two Issues : **Relative Error** Get a low-rank approximation  $\hat{A}$  to  $A$  so that

$$\|A - \hat{A}\|_F \leq (1 + \epsilon)\|A - A_k\|_F.$$

(Recall :  $A_k$  best rank  $k$  approx to  $A$ .)

**“Interpolative” approximation** Get an  $\hat{A}$  which is in the span of at most  $s$  ( $s$  small) rows of  $A$ .

Deshpande, Rademacher, Vempala, Wong; Drineas, Mahoney, Muthukrishnan; Sarlos; Har-Peled; Martinson, Rokhlin and Tygert all address these questions.

\*\*\*COMING ATTRACTIONS \*\*\*\*\*

Sarlos : Take  $s$  random (i.i.d.) linear combinations of the rows of  $A$ . Find best approximation  $\hat{A}$  to  $A$  in their span. Then with high probability :

$$\|A - \hat{A}\|_F \leq (1 + \epsilon) \|A - A_k\|_F,$$

provided  $s \geq ck^2 \log m/\epsilon$ .

One intuition : If one performs a random rotation of  $A$  on the left, one gets vectors all of roughly the same length. So

length-squared distribution  $\approx$  picking first  $s$ ....

More direct proof using classic and recent results on random projections in Sarlos. One issue : Random vectors are dense. But sparse random vectors with same properties recently developed....

Also tackles  $l_2$  linear regression.

Martinson, Rokhlin, Tygert : Independent development on similar lines. But  $s \approx k + 20$ . (No oversampling !). But weaker error bounds of the form (for  $m = n$ ) :

Error in spectral norm at most  $O(kn)\sigma_{k+1}(A)$ .  
Much better Empirical results.

## Tensors

**Max-3-SAT** : Given a Boolean CNF formula with 3 literals per clause, find an assignment to the variables satisfying as many clauses as possible.  $x_1, x_2, \dots, x_n$  0-1 variables. Let  $S = \{(x_1, x_2, \dots, x_n, 1 - x_1, 1 - x_2, \dots, 1 - x_n) : x_i \in \{0, 1\}\}$ . Max-3-SAT can be formulated as :

$$\text{Max}_{y \in S} : \sum_{i,j,k} A_{i,j,k} y_i y_j y_k.$$

Rank 1 3-tensor : Outer product of 3 vectors  
:  $u \otimes v \otimes w = (u_i v_j w_k)$ .

**Low Rank Approximation (LRA) of tensors** : Approximate by a sum of a small number of rank 1 tensors. If we can find a LRA  $B$ , replace  $A$  by  $B$ ; solve exploiting the low rank of  $B$ .

**Existence, Computation ???** — Golub and Lim; SATURDAY

**Existence Lemma** For any  $r$ -tensor  $A$ ,  $\epsilon > 0$ , there exist  $k \leq 1/\epsilon^2$  rank-1 tensors,  $B_1, B_2, \dots, B_k$  such that

$$\|A - (B_1 + B_2 + \dots + B_k)\|_2 \leq \epsilon \|A\|_F.$$

**Computation Theorem** For any  $r$ -tensor ( $r$  fixed)  $A$ ,  $\epsilon > 0$ , we can find  $k$  rank 1 tensors  $B_1, B_2, \dots, B_k$ , where  $k \leq 100/\epsilon^2$ , in time  $(n/\epsilon)^{O(1/\epsilon^4)}$  such that with high probability we have

$$\|A - (B_1 + B_2 + \dots + B_k)\|_2 \leq \epsilon \|A\|_F.$$

— de la Vega, K., Karpinski and Vempala

Notation :  $\|A\|_2$  is the spectral norm =

$\text{Max}_{u,v,w} A(u, v, w) = \sum_{ijk} A_{ijk} u_i v_j w_k$  over all unit length vectors.

Finding LRA for 3-tensors Enough to find  $u, v, w$  to maximize

$$A(u, v, w) = \sum_{ijk} A_{ijk} u_i v_j w_k.$$

Point 1 If  $u, v$  are known,

$$w = A(u, v, \cdot) = \sum_{ij} A_{ij \cdot} u_i v_j \quad (6)$$

suffices.

Point 2 We can **estimate** the sum in the r.h.s. of (6) if we have just  $O(1)$  terms picked according to the **length-squared distribution**. For this, need only  $O(1)$   $u_i, v_j$  !!

Point 3 We can enumerate all possible values of these  $O(1)$   $u_i, v_j$  and find all candidate  $w$ .

Point 4 We can check which candidate  $w$  is best by finding **maximum eigenvalue** of each **matrix**  $A(\cdot, \cdot, w)$  !!