Randomized algorithms for the approximation of matrices

Luis Rademacher
The Ohio State University
Computer Science and Engineering
(joint work with Amit Deshpande, Santosh Vempala, Grant Wang)

Two topics

- Low-rank matrix approximation (PCA).
- Subset selection:
  Approximate a matrix using another matrix whose columns lie in the span of a few columns of the original matrix.

Motivating example: DNA microarray

- [Drineas, Mahoney] *Unsupervised* feature selection for classification
  - Data: table of gene expressions (features) v/s patients
  - Categories: cancer types
  - Feature selection criterion: Leverage scores (importance of a given feature in determining top principal components)
- Empirically: Leverage scores are correlated with "information gain", a supervised measure of influence. Somewhat unexpected.
- Leads to clear separation (clusters) from selected features.

In matrix form:

- A is $m \times n$ matrix, m patients, n genes (features), find
$$A \approx CX,$$
where the columns of C are a few columns of A (so $X = C^{+}A$).
- They prove error bounds when columns of C are selected at random according to leverage scores (importance sampling).

Question

- *Supervised* feature selection…
  - … for classification and regression with theoretical guarantees, without statistical assumptions?

(P1) Matrix approximation

- Given m-by-n matrix, find low rank approximation …
- … for some norm:
  - $\|A\|_F^2 = \sum_{ij} A_{ij}^2$ (Frobenius)
  - $\|A\|_2 = \sigma_{\max}(A) = \max_x \|Ax\|/\|x\|$ (spectral)

Geometric view

- Given points in R^n, find subspace close to them.
- Error: Frobenius norm corresponds to sum of squared distances.

Classical solution

- Best rank-k approximation A_k in $\|\cdot\|_F^2$ and $\|\cdot\|_2$:
  - Top k terms of singular value decomposition (SVD): if $A = \sum_i \sigma_i u_i v_i^T$ then $A_k = \sum_{i=1}^k \sigma_i u_i v_i^T$
- Best k-dim. subspace: $\mathrm{rowspan}(A_k)$, i.e.
  - Span of top k eigenvectors of $A^T A$.
- Leads to iterative algorithm
  - Convergence: is it a polynomial time algorithm? Dependence on eigenvalue gap.

Want algorithm

- With better error/time guarantees.
- Efficient for very large data:
  - Nearly linear time
  - Pass efficient: if data does not fit in main memory, algorithm should not need random access, but only a few sequential passes.
- Subspace equal to or contained in the span of a few rows (actual rows are more informative than arbitrary linear combinations).

Idea [Frieze Kannan Vempala]

- Sampling rows.
  Uniform does not work (e.g. a single non-zero entry)
- By “importance”: sample s rows, each independently with probability proportional to squared length.

[FKV]

Theorem 1. *Let S be a sample of k/ϵ rows where*

$$\mathbb{P}(\textit{row } i \textit{ is picked}) \propto \|A_i\|^2.$$

Then the span of S contains the rows of a matrix $\tilde{A}$ of rank k for which

$$\mathsf{E}(\|A - \tilde{A}\|_F^2) \leq \|A - A_k\|_F^2 + \epsilon \|A\|_F^2.$$

This can be turned into an efficient algorithm: 2 passes, complexity $O(kmn/\epsilon)$.

(to compute $\tilde{A}$, SVD in span of S, which is fast because n becomes k/ϵ).

Drawback of [FKV]

- Additive error can be large (say if matrix is nearly low rank).
  Prefer relative error, something like

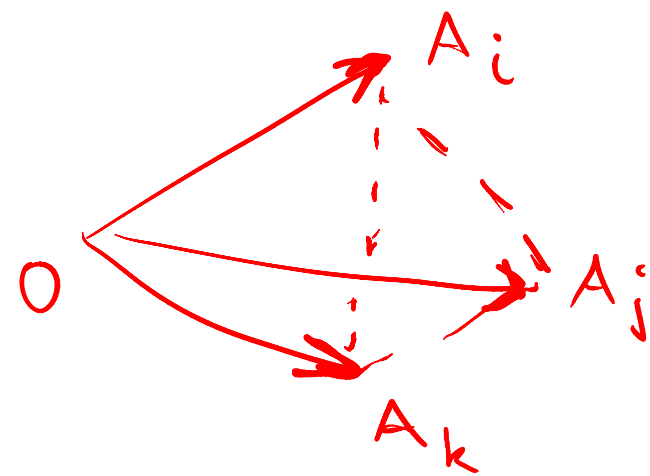
$$\|A - \tilde{A}\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2.$$

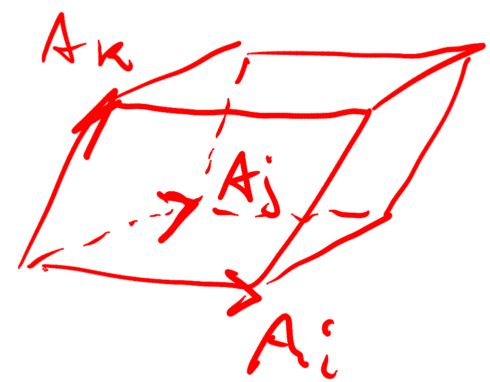
3 ways

- [Har-Peled ‘06] (first linear time relative approximation)
- [Sarlos ‘06]: Random projection of rows onto a $O(k/\epsilon)$-dim. subspace. Then SVD.
- [Deshpande R Vempala Wang] [Deshpande Vempala ‘06] Volume sampling (rough approximation) + adaptive sampling.

(P2) Algorithmic Problems: Volume sampling and subset selection

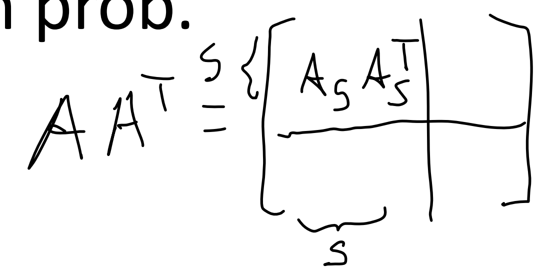
- Given m-by-n matrix, pick set of k rows at random with probability proportional to squared volume of k-simplex spanned by them and origin. [DRVW] (equivalently, squared volume of parallelepiped determined by them)





Volume sampling

- Let S be k-subset of rows of A
  - $[k!\,\mathrm{vol}(\mathrm{conv}(0, A_S))]^2 = \mathrm{vol}(\square(A_S))^2 = \det(A_S A_S^T)$ (*)
  - volume sampling for A is equivalent to: pick k by k principal minor "S × S" of A A^T with prob. proportional to $\det(A_S A_S^T)$



  - For(*): complete A_S to a square matrix B by adding orthonormal rows, orthogonal to span(A_S).

$$\mathrm{vol}\left(\square(A_S)\right)^2 = (\det B)^2 = \det(BB^T) = \det\begin{pmatrix} A_S A_S^T & 0 \\ 0 & I \end{pmatrix} = \det(A_S A_S^T)$$

Original motivation:

- **Relative error** low rank matrix approximation [DRVW]:
  - S: k-subset of rows according to volume sampling
  - A_k: best rank-k approximation, given by principal components (SVD)
  - π_S: projection of rows onto rowspan(A_S)

$$\implies \mathbb{E}_S(\|A - \pi_S(A)\|_F^2) \leq (k+1)\|A - A_k\|_F^2$$

- Factor "k+1" is best possible [DRVW]
- Interesting existential result (there exist k rows…). Alg.?
- Lead to linear time, pass-efficient algorithm for relative approximation of A_k [DV]. $(1+\varepsilon)$ in span of $O^*(k/\varepsilon)$ rows

Where does volume sampling come from?

- **No self-respecting architect leaves the scaffolding in place after completing the building.**

  Gauss?

Where does volume sampling come from?

- Idea:
  - For picking k out of $k+1$ points, k with *maximum* volume is optimal.
  - For picking 1 out of m, random according to squared length is better than max. length.
  - For k out of m, this suggest volume sampling.
- Why does the algebra work? Idea:
  - When picking 1 out of m random according to squared length, expected error is sum squares of areas of triangles. This sum corresponds to certain coefficient of the characteristic polynomial of AA^T

Later motivation [BDM,…]

- (row/column) Subset selection.
  A refinement of principal component analysis:
  Given a matrix A,
  - PCA: find k-dim subspace V that minimizes

$$\|A - \pi_V(A)\|_F^2$$

  - Subset selection: find V *spanned by k rows of A*.
    - Seemingly harder, combinatorial flavor.

(π projects rows)

Why subset selection?

- PCA unsatisfactory:
  - top components are linear combinations of rows (all rows, generically). Many applications prefer *individual*, most relevant rows, e.g.:
    - feature selection in machine learning
    - linear regression using only most relevant independent variables
    - out of thousands of genes, find a few that explain a disease

Known results

- [Deshpande-Vempala] Polytime k!-approximation to volume sampling, by adaptive sampling:
  - pick a row with probability proportional to squared length
  - project all rows orthogonal to it
  - repeat
- Implies for random k-subset S with that distribution:

$$\mathbb{E}_S(\|A - \pi_S(A)\|_F^2) \leq (k+1)!\|A - A_k\|_F^2$$

Known results

- [Boutsidis, Drineas, Mahoney] Polytime randomized algorithm to find k-subset S:

$$\|A - \pi_S(A)\|_F^2 \leq O(k^2 \log k)\|A - A_k\|_F^2$$

- [Gu-Eisenstat] Deterministic algorithm,

$$\|A - \pi_S(A)\|_2^2 \leq (1 + f^2 k(n-k))\|A - A_k\|_2^2$$

in time $O((m + n \log_f n)n^2)$

Spectral norm:

$$\|A\|_2 = \sup_{x \in R^n} \|Ax\|/\|x\|$$

Known results

- Remember, volume sampling equivalent to sampling k by k minor "S $\times$ S" of AA^T with probability proportional to

  (*) $$\det(A_S A_S^T)$$

- [Goreinov, Tyrtishnikov, Zamarashkin] Maximizing (*) over S is good for subset selection.
- [Çivril , Magdon-Ismail] But maximizing is NP-hard, even approximately to within an exponential factor.

Results

- *Volume sampling:* First polytime exact alg. $O(mn^{\omega} \log n)$ (arithmetic ops.)
- Implies alg. with optimal approximation for *subset selection* under Frobenius norm. Can be *derandomized* by conditional expectation. $O(mn^{\omega} \log n)$
- $1+\varepsilon$ approximations to the previous 2 algorithms in nearly linear time, using volume-preserving random projection [M Z].

Results

- Observation: Bound in Frobenius norm easily implies bound in spectral norm:

$$\begin{aligned}\|A - \pi_S(A)\|_2^2 &\le \|A - \pi_S(A)\|_F^2 \\ &\le (k+1)\|A - A_k\|_F^2 \\ &\le (k+1)(n-k)\|A - A_k\|_2^2\end{aligned}$$

using

$$\|A\|_F^2 = \sum_i \sigma_i^2 \qquad \|A\|_2^2 = \sigma_{\max}^2$$

$\sigma_{\max} = \sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n \ge 0$ are the singular values of A

Comparison for subset selection

$$\text{Find } S \text{ s.t. } \quad \|A - \pi_S(A)\|_?^2 \leq ?\|A - A_k\|_?^2$$

	Frobenius norm sq	Spectral norm sq	Time (assuming m>n) ω: exponent of matrix mult.	
[D R V W]	k+1		Existential	
[Despande Vempala]	(k+1)!		kmn	R
[Gu Eisenstat]		1+k(n-k)	Existential	
[Gu Eisenstat]		$1+f^2k(n-k)$	$((m + n \log_f n)n^2$	D
[Boutsidis Drineas Mahoney]	$k^2 \log k$	$k^2 (n-k) \log k$ (F implies spectral)	mn^2	R
[Desphande R]	k+1 (optimal)	(k+1)(n-k)	$kmn^\omega \log n$	D
[Desphande R]	(1+ε)(k+1)	(1+ε) (k+1)(n-k)	$O^*(mnk^2/\varepsilon^2 + m\, k^{2\omega+1}/\varepsilon^{2\omega})$	R

Proofs: volume sampling

- Want (w.l.o.g.) **k-tuple** S of rows of m by n matrix A with probability

$$\frac{\det(A_S A_S^T)}{\sum_{S' \in [m]^k} \det(A_{S'} A_{S'}^T)}$$

Proofs: volume sampling

- Idea: pick S=(S_1, S_2, …, S_k) in sequence. Need marginal distribution of S_1 to begin:

$$\mathbb{P}(S_1 = i) = \frac{\text{tuples with } S_1 = i}{\text{all tuples}} = \frac{\sum_{S' \in [m]^k, S'_1 = i} \det(A_{S'} A_{S'}^T)}{\sum_{S' \in [m]^k} \det(A_{S'} A_{S'}^T)}$$

- Remember characteristic polynomial:

$$p_{AA^T}(x) = \det(xI - AA^T) = \sum_i c_i(AA^T) x^i$$

$$|c_{m-k}(AA^T)| = \sum_{S \subseteq [m], |S| = k} \det(A_S A_S^T)$$

Proofs: volume sampling

- So, for $C_i = A - \pi_{A_i}(A)$

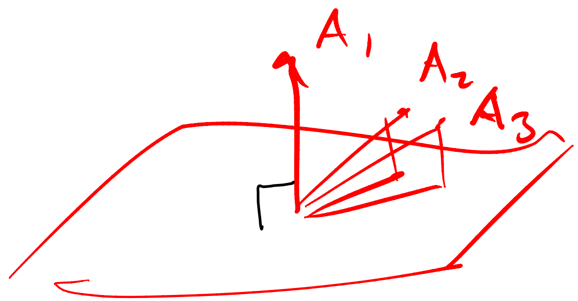
$$\mathbb{P}(S_1 = i) = \frac{\sum_{S' \in [m]^k, S'_1 = i} \det(A_{S'} A_{S'}^T)}{\sum_{S' \in [m]^k} \det(A_{S'} A_{S'}^T)}$$

$$= \frac{(k-1)! \|A_i\|^2 \sum_{S' \subseteq [m], |S'| = k-1} \det((C_i)_{S'} (C_i)_{S'}^T)}{k! \sum_{S' \subseteq [m], |S'| = k} \det(A_{S'} A_{S'}^T)}$$

$$= \frac{\|A_i\|^2 |c_{m-k+1}(C_i C_i^T)|}{k |c_{m-k}(AA^T)|}$$

intuition for numerator:

$$|\square(A_1, A_2, A_3)| = \|A_1\| |\square(\pi_{A_1^\perp}(A_2, A_3))|$$



Proofs: volume sampling

- So: $$\mathbb{P}(S_1 = i) = \frac{\|A_i\|^2 |c_{m-k+1}(C_i C_i^T)|}{k|c_{m-k}(AA^T)|}$$

$$C_i = A - \pi_{A_i}(A)$$

- Can be computed in polytime.
- After S_1, project rows orthogonal to picked row, repeat marginal computation for S_2,… (use intuition for numerator)
- “flops”: k * m * ($m^2 n + m^\omega \log m$)

Proofs: volume sampling

- Faster: (we assume m > n)
  - use $p(AA^T) = x^{m-n} p(A^T A)$

  as A^TA is n by n (smaller than m by m)
  - use rank-1 updates:

$$C_i = A - \frac{1}{\|A_i\|^2} A A_i A_i^T,$$

$$C_i^T C_i = A^T A - \frac{A^T A A_i A_i^T}{\|A_i\|^2} - \frac{A_i A_i^T A^T A}{\|A_i\|^2} + \frac{A_i A_i^T A^T A A_i A_i^T}{\|A_i\|^4}.$$

  - total flops: $mn^2 + km(n^2 + n^\omega \log n) = k\, m\, n^\omega \log n$

Even faster

- Volume sampling only cares about volumes of k-subsets,
  $\Rightarrow$ can get $1+\varepsilon$ approximation using a volume preserving random projection [Magen, Zouzias] (generalizing Johnson Lindenstrauss, not same as Feige).

Even faster

- [Magen Zouzias]:
  For any $A \in R^{m \times n}$, $1 \leq k \leq n$, $\varepsilon < 1/2$ there is a $d = O(k^2 \varepsilon^{-2} \log m)$ s.t. for all S, k-subset of [m]:

$$\det A_S A_S^T \leq \det \tilde{A}_S \tilde{A}_S^T \leq (1+\epsilon) \det A_S A_S^T,$$

$$\tilde{A} = AG, \quad G \in R^{n \times d} \text{ random Gaussian matrix, scaled}$$

- k=1 is JL
- This as preprocessing implies $1+\varepsilon$ volume sampling in time

  $O^*(mnk^2/\varepsilon^2 + m\, k^{2\omega+1}/\varepsilon^{2\omega})$

Recent news

- Boutsidis, Drineas, Magdon-Ismail: near optimal subset selection.
- Guruswami, Sinop:
  - Volume sampling in time $O(kmn^2)$
  - Relative $(1+\epsilon)$ matrix approximation with one round of $r = k - 1 + k/\epsilon$ rows of volume sampling. More precisely, for S a sample of size $r \geq k$ according to volume sampling:

$$\mathsf{E}_S(\|A - \pi_S(A)\|_F^2) \leq \frac{r+1}{r+1-k}\|A - A_k\|_F^2$$

Open question

- For a subset of rows S according to volume sampling, lower bound (in expectation?):

$$\sigma_{\min}(A_S)$$

  - Want A_S to be well conditioned.