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).
(P1) Matrix approximation

• Given $m$-by-$n$ matrix, find low rank approximation ...

• ... for some norm:
  
  \[- \|A\|_F^2 = \sum_{ij} A_{ij}^2 \quad \text{(Frobenius)}\]

  \[- \|A\|_2 = \sigma_{\text{max}}(A) = \max_x \|Ax\|/\|x\| \quad \text{(spectral)} \]
Geometric view

- Given points in $\mathbb{R}^n$, find subspace close to them.
- Error: Frobenius norm corresponds to sum of squared distances.
Classical solution

- Best rank-k approximation $A_k$ in $\|\cdot\|^2_F$ 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: rowspan($A_k$), i.e.
  - Span of top $k$ eigenvectors of $A^T A$.
- Leads to iterative algorithm. Essentially, in time $mn^2$. 
Want algorithm

• With better error/time trade-off.
• 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 $k/\epsilon$ rows where

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

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

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

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 $k/\epsilon$).
One drawback of [FKV]

• Additive error can be large (say if matrix is nearly low rank).

Prefer relative error, something like

$$\|A - \tilde{A}\|_F^2 \cdot (1 + \epsilon) \|A - A_k\|_F^2.$$
Several 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 ‘06] [Deshpande Vempala ‘06] Volume sampling (rough approximation) + adaptive sampling.
Some more relevant work

- [Papadimitriou Raghavan Tamaki Vempala ‘98]: Introduced random projection for matrix approximation.
- [Achlioptas McSherry ‘01][Clarkson Woodruff ’09] One-pass algorithm.
- [Woolfe Liberty Rokhlin Tygert ’08] [Rokhlin Szlam Tygert ‘09] Random projection + power iteration to get very fast practical algorithms. Read survey [Halko Martinsson Tropp ‘09].
- D’Aspremont, Drineas, Ipsen , Mahoney, Muthukrishnan, ...
(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 $A$
  - $[k! \ vol(\text{conv}(0, A_S))]^2 = \ vol(\square(A_S))^2 = \ det(A_S A_S^T)$ (*)
  - volume sampling for $A$ is equivalent to: pick $k$ by $k$ principal minor “$S \times S$” of $A A^\top$ 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($A_S$).

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

\[
\implies E_S \left( \left\| A - \pi_S(A) \right\|_F^2 \right) \cdot (k + 1) \left\| A - A_k \right\|_F^2
\]

• 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+ε) in span of O^*(k/ε) 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.
Where does volume sampling come from?

- Why does the algebra work? Idea:
  - When picking 1 out of $m$ random according to squared length, expected error is sum of squares of areas of triangles:
    \[
    \text{E(error)} = \sum_s \frac{\|A_s\|^2}{\sum_t \|A_t\|^2} \sum_i d(A_i, \text{span}(A_s))^2
    \]
  - 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 $V$ \textit{spanned by $k$ rows of $A$}.

  \begin{itemize}
  \item Seemingly harder, combinatorial flavor.
  \end{itemize}

($\pi$ 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:

$$E_S(\|A - \pi_S(A)\|_F^2) \cdot (k + 1)!\|A - A_k\|_F^2$$
Known results

- [Boutsidis, Drineas, Mahoney] Polytime randomized algorithm to find $k$-subset $S$:
  \[ \| A - \pi_S(A) \|_F^2 \cdot O(k^2 \log k) \| A - A_k \|_F^2 \]

- [Gu-Eisenstat] Deterministic algorithm,
  \[ \| A - \pi_S(A) \|_2^2 \cdot (1 + f^2 k(n - k)) \| A - A_k \|_2^2 \]
  in time $O((m + n \log f n)n^2)$

Spectral norm:
\[ \| A \|_2 = \sup_{x \in \mathbb{R}^n} \| Ax \| / \| x \| \]
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] [see also Koutis ‘06] But maximizing is NP-hard, even approximately to within an exponential factor.
Results

• *Volume sampling:* Polytime exact alg. $O(mn^{o} \log n)$ (arithmetic ops.) (some ideas earlier in [Houges Krishnapur Peres Virág])

• Implies alg. with optimal approximation for *subset selection* under Frobenius norm. Can be *derandomized* by conditional expectation. $O(mn^{o} \log n)$

• $1+\varepsilon$ 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:

\[
\|A - \pi_S(A)\|_2^2 \cdot \|A - \pi_S(A)\|_F^2 \\
\cdot (k + 1)\|A - A_k\|_F^2 \\
\cdot (k + 1)(n - k)\|A - A_k\|_2^2
\]

using

\[
\|A\|_F^2 = \sum_i \sigma_i^2, \quad \|A\|_2^2 = \sigma_{\text{max}}^2
\]

\(\sigma_{\text{max}} = \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0\) are the singular values of A
# Comparison for subset selection

Find $S$ s.t. $\|A - \pi_S(A)\|^2 \cdot \|A - A_k\|^2$

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

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

- Remember characteristic polynomial:

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

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

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

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

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

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

intuition for numerator:

$$|\Box(A_1, A_2, A_3)| = \|A_1\| |\Box(\pi_{A_1}(A_2, A_3))|$$
Proofs: volume sampling

• So:

\[ P(S_1 = i) = \frac{\|A_i\|^2 |c_{m-k+1}(C_i C_i^T)|}{k |c_{m-k}(A A^T)|} \]

\[ C_i = A - \pi_{A_i}(A) \]

• Can be computed in polytime.

• After \( S_1 \), project rows orthogonal to picked row, repeat marginal computation for \( S_2, \ldots \) (use intuition for numerator)

• “flops”:\nk * m * (m^2n + m^ωlog m)
Proofs: volume sampling

• Faster: (we assume \( m > n \))
  
  – use \( p(AA^T) = x^{m-n} p(A^T A) \)
  
  as \( A^T A \) 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 $A \in \mathbb{R}^{m \times n}$, $1 \leq k \leq n$, $\varepsilon < 1/2$ there is a $d = O(k^2 \varepsilon^{-2} \log m)$ s.t. for all $S$, $k$-subset of $[m]$:
  
  $\det A_S A_S^T \cdot \det \tilde{A}_S \tilde{A}_S^T \cdot (1 + \varepsilon) \det A_S A_S^T,$

  $\tilde{A} = AG, \quad G \in R^{n \times d}$ 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 (FOCS ‘11)]:
  improved subset selection using [BSS].
- [Guruswami, Sinop (SODA ‘12)]:
  - 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:

\[
E_S(\|A - \pi_S(A)\|_F^2) \cdot \frac{r + 1}{r + 1 - k} \|A - A_k\|_F^2
\]
Open question

• For a subset of rows $S$ according to volume sampling, lower bound (in expectation?):
  
  $$\sigma_{\text{min}}(A_S)$$

  – Want $A_S$ to be well conditioned.