Does machine learning need the power of iterative methods for the SVD?

Andreas Stathopoulos
with Z. Liu, E. Myers, E. Romero, L. Wu
The problems

Given $A$ large, $n \times d$ matrix with $d \leq n$, find $i = 1, \ldots, k$ singular triplets

$$Av_i = \sigma_i u_i, \quad \sigma_1 \geq \ldots \geq \sigma_k$$

1. $k = 1$, largest triplet but fast
2. $k$ is small but high accuracy needed
3. a low rank approximation of $A$ (various measures)
The problems

Given $A$ large, $n \times d$ matrix with $d \leq n$, find $i = 1, \ldots, k$ singular triplets

$$Av_i = \sigma_i u_i, \quad \sigma_1 \geq \ldots \geq \sigma_k$$

1. $k = 1$, largest triplet but fast
2. $k$ is small but high accuracy needed
3. a low rank approximation of $A$ (various measures)

Applications:

- applied math (norm computation, reduction of dimensionality or variance)
- combinatorial scientific computing, social network analysis
- computational statistics and Machine Learning (PCA, SVM, LSI)
- low dimensionality embeddings (ISOMAP, MDS)
- SVD updating of streaming data
Methods for high accuracy

Golub-Kahan Lanczos based:

- unrestarted with full or partial reorthogonalization
- implicitly restarted (IRBLA, PROPACK)
- thick and/or locally optimally restarted and with preconditioning (GKD, JDSVD, PLMR, PRIMME, SVDIFP)

For small $k$, GKD near-optimal and robust

For large $k$, Lanczos based methods may be preferable

Are these methods competitive when low accuracy is needed?
Randomized SVD

For low accuracy or low rank approximations, a random block projection with 1-2 subspace iterations is often sufficient [Rokhlin et al.’09, Halko et al.’11]

Randomized SVD (RSVD)
In: $A$, desired rank $k$, oversampling parameter $p$
Select a random $\tilde{V} \in \mathbb{R}^{d \times (k+p)}$
$[\tilde{U}, R] = \text{qr}(A\tilde{V}, 0)$
for $i = 1 \ldots q$
$[\tilde{V}, R_v] = \text{qr}(A^T\tilde{U}, 0)$
$[\tilde{U}, R] = \text{qr}(A\tilde{V}, 0)$
Out: $\tilde{A} = \tilde{U} \tilde{U}^T A \approx A$

What makes RSVD attractive for ML problems?
Randomized SVD

For low accuracy or low rank approximations, a random block projection with 1-2 subspace iterations is often sufficient [Rokhlin et al.’09, Halko et al.’11]

Randomized SVD (RSVD)
In: $A$, desired rank $k$, oversampling parameter $p$
   Select a random $\tilde{V} \in \mathbb{R}^{d \times (k+p)}$
   $[\tilde{U}, R] = \text{qr}(A\tilde{V}, 0)$
   for $i = 1 \ldots q$
   $[\tilde{V}_i, R_{v_i}] = \text{qr}(A^T\tilde{U}, 0)$
   $[\tilde{U}, R] = \text{qr}(A\tilde{V}, 0)$
Out: $\tilde{A} = \tilde{U}\tilde{U}^T A \approx A$

What makes RSVD attractive for ML problems?

The low rank criterion: $\|A - \tilde{U}\tilde{U}^T A\|_{2/F} = \|A - UU^T A\|_{2/F}(1 + \varepsilon)$
RSVD converges independently of gaps!

Goal: satisfy $\|A - \tilde{U}\tilde{U}^T A\|_{2/F} = \|A - UU^T A\|_{2/F}(1 + \varepsilon)$

Classical NLA:

Subspace Iteration converges in $O \left( \frac{\log(d)}{\sqrt{\frac{\sigma_k}{\sigma_{k+1}} - 1}} \right)$ iterations [Saad]

Randomized NLA: with high probability

RSVD converges in $O\left(\frac{1}{\varepsilon}\right)$ iterations [Rokhlin et al.’09, Halko et al.’11]

Randomized block Lanczos in $O\left(\frac{1}{\sqrt{\varepsilon}}\right)$ iterations [Musco & Musco,’15]

Practically useful when $\varepsilon$ larger than gap (e.g., in ML)

Although surprising, the reason is the weaker condition on $\|A - \tilde{U}\tilde{U}^T A\|_{2/F}$
Gaps not needed for low rank approximation

Example: \( A = \begin{pmatrix} 1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \), the best rank-1: \( \|A - A_1\|_F = \sigma_2 \)

and a perturbed eigenvector: \( \tilde{u} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} + 10^{-3} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \), \( \tilde{u} = \tilde{u}/\|\tilde{u}\| \)

<table>
<thead>
<tr>
<th>( \sigma_2 )</th>
<th>( \frac{|A - \tilde{u}\tilde{u}^T A|_F - |A - A_1|_F}{|A - A_1|_F} )</th>
<th>( \frac{|A - \tilde{u}\tilde{u}^T A|_F - |A - A_1|_F}{|A|_F} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 10^{-6} )</td>
<td>( 10^3 )</td>
<td>( 10^{-3} )</td>
</tr>
<tr>
<td>( 10^{-5} )</td>
<td>( 10^2 )</td>
<td>( 10^{-3} )</td>
</tr>
<tr>
<td>( 10^{-4} )</td>
<td>9</td>
<td>0.9 ( \times ) 10(^{-3} )</td>
</tr>
<tr>
<td>( 10^{-3} )</td>
<td>0.4</td>
<td>0.4 ( \times ) 10(^{-3} )</td>
</tr>
<tr>
<td>( 10^{-2} )</td>
<td>( 5 \times 10^{-3} )</td>
<td>0.5 ( \times ) 10(^{-4} )</td>
</tr>
<tr>
<td>( 10^{-1} )</td>
<td>( 5 \times 10^{-5} )</td>
<td>0.5 ( \times ) 10(^{-5} )</td>
</tr>
<tr>
<td>0.50</td>
<td>( 2 \times 10^{-6} )</td>
<td>0.7 ( \times ) 10(^{-6} )</td>
</tr>
<tr>
<td>0.99</td>
<td>( 10^{-8} )</td>
<td>0.7 ( \times ) 10(^{-8} )</td>
</tr>
</tbody>
</table>
Angles are not even necessary!

Small angles sufficient (e.g., eigensolvers) but not necessary

Example: \( A = \begin{pmatrix} 1 & 0 \\ 0 & 0.999 \end{pmatrix} \), the best rank-1: \( \| A - A_1 \|_F = 0.999 \)

Using \( \tilde{u} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \), i.e., \( \tilde{u} \perp u_1 \), still gives

\[
\frac{\| A - \tilde{u}\tilde{u}^T A \|_F - \| A - A_1 \|_F }{\| A - A_1 \|_F} = 10^{-3}
\]

\[
\frac{\| A - \tilde{u}\tilde{u}^T A \|_F - \| A - A_1 \|_F }{\| A \|_F} = 7 \cdot 10^{-4}
\]

small gaps may help the projection error!
Is it the choice of norm?

\[ \|A - A_k\|_F^2 = \|A\|_F^2 - \sum_{i=1}^{k} \sigma_i^2 \]

\[ \|A - A_k\|_2^2 = \sigma_{k+1}^2 \] but \( \|A - \tilde{A}\|_2^2 \) expensive to compute

Heavy tail singular value distributions weigh much on Frobenius norm

\[ \Rightarrow \|A - \tilde{A}\|_F \text{ meaningless for approximating singular spaces} \]

Spectral norm often helps distinguish singular values below a noise threshold

However [Ming Gu, ’16]:

\[ \|A - \tilde{A}\|_F^2 < \|A - A_k\|_F^2 + \eta \quad \Rightarrow \quad \|A - \tilde{A}\|_2^2 < \|A - A_k\|_2^2 + \eta \]
Is it the choice of norm?

Recent further analysis for any Schatten-p and unitarily invariant norms

Low rank approximation is well-conditioned [Drineas & Ipsen’18, Saibaba’18]:

\[ \| | A - \tilde{U} \tilde{U}^T A | |_p - \| | A - UU^T A | |_p \leq \| | U - \tilde{U} | |_p \times \| | A | |_p \]

- No gaps in the bound
- Small angles do give good low rank approximations
- But good low rank approximation does not imply small angles

**Conclusion 1:** low rank approximation a weak criterion
The rank of the random block matters

If $k$ wanted rank, $p$ oversampling, $q$ the # of RSVD iterations, then [Ming Gu, ’16]

$$\frac{\|A - \tilde{U}\tilde{U}^TA\|_{2/F}^2 - \|A - UU^TA\|_{2/F}^2}{F^2} \leq k C \left( \frac{\sigma_{k+p+1}}{F} \right)^2 \left( \frac{\sigma_{k+p+1}}{\sigma_k} \right)^{4q}$$

Normalization choices explain the behavior of different norms:

$F = \|A\|_2$ gives $\sigma_{k+p+1} \sigma_1$  \hspace{1cm}  $F = \|A - UU^TA\|_2$ gives $\sigma_{k+p+1}$

$F = \|A\|_F$ gives $\sigma_{k+p+1} \sum_i \sigma_i$  \hspace{1cm}  $F = \|A - UU^TA\|_F$ gives $\sigma_{k+p+1} \sum_{i > k} \sigma_i$
The rank of the random block matters

If $k$ wanted rank, $p$ oversampling, $q$ the # of RSVD iterations, then [Ming Gu, ’16]

$$\frac{\|A - \tilde{U}\tilde{U}^T A\|_2^2/F - \|A - UU^T A\|_2^2/F}{F^2} \leq k C \left( \frac{\sigma_{k+p+1}}{F} \right)^2 \left( \frac{\sigma_{k+p+1}}{\sigma_k} \right)^{4q}$$

Normalization choices explain the behavior of different norms:

- $F = \|A\|_2$ gives $\frac{\sigma_{k+p+1}}{\sigma_1}$
- $F = \|A - UU^T A\|_2$ gives $\frac{\sigma_{k+p+1}}{\sigma_{k+1}}$
- $F = \|A\|_F$ gives $\frac{\sigma_{k+p+1}}{\sum_i \sigma_i}$
- $F = \|A - UU^T A\|_F$ gives $\frac{\sigma_{k+p+1}}{\sum_{i>k} \sigma_i}$

More importantly, the $\frac{\sigma_{k+p+1}}{F}$ factor can be much smaller than $\frac{\sigma_{k+p+1}}{\sigma_k}$

**Conclusion 2:** For low accuracy, sometimes it is faster to just increase $p$!
The rank of the random block matters

In NLA the debate was on asymptotic convergence of Lanczos vs block methods

Many ML problems, however, never enter the asymptotic regime

The RSVD theory adds a non-asymptotic component to the analysis (initial guess)
The streaming problem

Often the rows of $A$ are streamed, and $d$ is too large to form $A^T A$

Several online or incremental SVD (or PCA) algorithms
[Weng et al. ’03, Ross et al.’08, Baker et al.’12, Degas&Cardot’16]
Fast Frequent Directions

A notable online SVD [Gashami-Liberty-Phillips-Woodruff, ’15]

Truncation:

\[
[u,s,v]=\text{svd}(B,'econ');
\]

Truncate \( S(i) = \sqrt{s(i)^2 - s(m)^2} \), \( i=1...m \)

Truncate \( B(1:m, :) = \text{diag}(S)\text{v}(1:m, :)' \);

guarantees:

\[
\|A - \tilde{U}\tilde{U}^T A\|_{2/F} \leq \|A - UU^T A\|_{2/F} (1 + \frac{k}{m-k})
\]

Often quality much better than the bound
Illustrative experiments

A synthetic model

$A(25000 \times 5000)$, dense, constructed as

$$A = SDU + \frac{N(0, 1)}{\zeta}$$

with $S, U$ unitary and $N(0, 1)$ noise.

$$D = \text{diag}(\text{linspace}(1/m, 1)),$$ for rank $m = 200$

When $\zeta = 0.1$ the matrix is dominated by noise

When $\zeta \gg \sqrt{d/m}$ the matrix is basically rank $m$
Methods

PRIMME’s GD+K (soft locking, block size 1)

PRIMME’s LOBPCG

both PRIMME methods stopped when $\|r_i\| < 10^{-2}\|A\|_2$

RSVD (Halko et al.)

performs always 2 iterations

FD (fast frequent directions)

1 pass
\[ A = SDU + N/\zeta, \quad m = 200, \quad \zeta = 0.1 \text{ Noise dominates} \]

Relatively good projection error without singular space, while FD not far behind
\[ A = SDU + N/\zeta, \quad m = 200, \quad \zeta = 3 \text{ Noise less strong} \]
\[ A = SDU + N/\zeta, \ m = 200, \ \zeta = 50 \textbf{Rank-200 dominates} \]

Quality of eigenspaces is unnecessary for low projection error
3D Laplacian 27000, looking for $m = 200$

RSVD faster and competitive in projection error — but not a eigensolver!
Extending iterative methods to various stopping criteria

RSVD is subspace iteration that stops early because high accuracy is not needed

Without intelligent convergence testing, RSVD may not meet user requirements

However, automation is desirable

Goal: extend state-of-the-art iterative methods to implement stopping criteria based on a variety of quality requirements
Various quality requirements

1. Accurate subspace means small angles with exact space, but we settle for residual norms \( \leq \varepsilon \| A \| \), say \( \varepsilon \leq O(10^{-5}) \)
   assuming the singular value gap from unwanted spectrum is \( \gg \varepsilon \| A \| \)
Various quality requirements

1. **Accurate subspace** means small angles with exact space, but we settle for residual norms \( \leq \varepsilon \| A \| \), say \( \varepsilon \leq O(10^{-5}) \) assuming the singular value gap from unwanted spectrum is \( \gg \varepsilon \| A \| \)

2. Low rank approximation relative to \( \| A \| \)
   \[
   \| A - \tilde{U} \tilde{U}^T A \|_2/F = \| A - UU^T A \|_2/F + \alpha \| A \|_2/F \quad \text{say } \alpha = O(10^{-1})
   \]

3. Low rank approximation relative to optimal low rank
   \[
   \| A - \tilde{U} \tilde{U}^T A \|_2/F = \| A - UU^T A \|_2/F (1 + \varepsilon) \quad \text{say } \varepsilon = O(10^{-3})
   \]
Various quality requirements

4. Low accuracy of singular values relative to $\|A - UU^T A\|_2 = \sigma_{k+1}$

$$\sigma_i - \tilde{\sigma}_i = \varepsilon \sigma_{k+1}, \quad \text{say } \varepsilon \leq 10^{-1}$$
Various quality requirements

4. Low accuracy of singular values relative to \( \|A - \mathbf{U}\mathbf{U}^T A\|_2 = \sigma_{k+1} \)
   \[ \sigma_i - \tilde{\sigma}_i = \varepsilon \sigma_{k+1}, \quad \text{say } \varepsilon \leq 10^{-1} \]

5. Given \( \alpha \), find also \( k \) such that \( \|A - \mathbf{\tilde{U}}\mathbf{\tilde{U}}^T A\|_{2/F} < \alpha \|A\|_{2/F} \)

6. Given a \( \gamma \) find also \( k \) such that \( \sigma_k - \sigma_{k+1} > \gamma \) or \( \sigma_k > \gamma \)

7. None of the above! Decide quality based on downstream application
Prototyping these criteria

Implemented criteria 1–6 in a subspace iteration (SI) in MATLAB

SI provides error estimates for all $k$ singular triplets

Error bounds help dynamically determine blockSize and convergence

Preliminary experiments show that guarantees are met

Adaptivity is more expensive than knowing $k$ and target $\alpha \|A\|_F$

Also a PRIMME implementation of $\|A - \tilde{U} \tilde{U}^T A\|_F < \alpha \|A\|_F$
RSVD always 2 iterations, others iterate to \( \| A - \tilde{U}\tilde{U}^T A \|_F < 0.001 \| A \|_F \)

For slowly-decaying spectra, better converging methods pay off.
Conclusions

Some ML problems do not need quality singular spaces/values

- Low rank projection a weak criterion
- A variety of quality requirements

Classical vs randomized Krylov methods is a little misleading

- Block iterative methods are randomized
- The practical difference is the stopping criteria
  - Important to identify criteria for each problem
  - Many criteria implementable but need to be optimized
  - How to choose block size?

Our goal: Robust unified software for any requirements
Stopping PRIMME by \(|A - \tilde{U} \tilde{U}^T A|_F < \alpha |A|_F\) and varying the block size

**Synthetic model with \(m=200\)**

- Matvecs increase linearly with blockSize \(> 20\)
- Min time: blockSize \(\in [30, 40]\)
- blockSize=1 took 60-200sec
- LOBPCG is blockSize = 200
- For sparse matrices optimal blockSize may differ
Model 1

$A(25000 \times 5000)$, dense, constructed as

$$A = SDU$$

with $S, U$ unitary. Consider two singular value distributions:

$$D = \text{diag}(\text{logspace}(0, t)), \quad t = -0.5, \quad t = -3.5$$

i.e., singular values from $10^0$ to $10^t$

Look for largest sing. space with rank $m = 10$ and $m = 200$
A = SDU, time, projection error, and residual norms, m = 10, t = -0.5

Relatively good projection error without singular space
\( A = SDU \), time, projection error, and residual norms, \( m = 10, t = -3.5 \)

Qualitatively similar
\[ A = S D U, \text{ time, projection error, and residual norms, } m = 200, t = -0.5 \]
\[ A = SDU, \text{ time, projection error, and residual norms, } m = 200, t = -3.5 \]
Model 3: $A(25000 \times 5000)$ Human genotype data, $m = 200$

Eigensolvers get better residuals but worse projection error!