The Effectiveness of Single Vector Krylov Methods for Low-Rank Approximation

NYU Tandon School of Engineering, Christopher Musco
COLLABORATORS

Raphael Meyer  
(NYU)

Cameron Musco  
(UMass. Amherst)


Raphael is graduating with his Ph.D. in CS in **Spring 2024** and will be looking for postdocs!
Have a matrix $A \in \mathbb{R}^{d \times n}$ and want to quickly compute an optimal rank-$k$ approximation.

Need to find the top $k$ singular vectors of $A$. 
PROBLEM WE ARE STUDYING

Have a matrix $A \in \mathbb{R}^{d \times n}$ and want to quickly compute an optimal rank-$k$ approximation.

Need to find the top $k$ singular vectors of $A$. Call them $U_k \in d \times k$. Then we can form the approximation:

$$A \approx U_k U_k^T A \overset{\text{def}}{=} A_k$$

$A_k$ is the optimal $k$-rank approximation to $A$. 

\[ \begin{array}{c}
\text{d} & \begin{array}{c}
\text{n} \\
A
\end{array} \\
\end{array} \begin{array}{c}
= \\
\Sigma
\end{array} \begin{array}{c}
\text{U}_k \\
\text{V}_k^T
\end{array} \]
PROBLEM WE ARE STUDYING

Have a matrix $A \in \mathbb{R}^{d \times n}$ and want to quickly compute an optimal rank-$k$ approximation.

Need to find the top $k$ singular vectors of $A$. Call them $U_k \in d \times k$. Then we can form the approximation:

$$A \approx U_k U_k^T A \overset{\text{def}}{=} A_k$$

$A_k$ is the optimal $k$-rank approximation to $A$.

How do we find it quickly without computing a full SVD?
**TYPICAL KRYLOV METHOD**

**Most common approach:** Run a Krylov subspace method.

1. Choose random starting block \( B \in \mathbb{R}^{d \times b} (b = 1 \text{ or larger}) \).
2. Compute orthonormal basis \( Z \) for the Krylov subspace:
   \[
   \mathcal{K} = \begin{bmatrix}
   B & AA^T B & (AA^T)^2 B & \cdots & (AA^T)^t B
   \end{bmatrix}
   \]
3. Return \( Q = Z\tilde{U}_k \) where \( \tilde{U}_k \) contains the top \( k \) eigenvectors of \( Z^T AA^T Z \). **This choice of \( Q \) minimizes** \( \|A - QQ^T A\|_F \) among all rank \( k \) matrices in the span of \( \mathcal{K} \).

**Runtime of this method?** Often dominated by the cost of multiplying vectors by \( A \), so \( O(T_{MV}(A) \cdot b \cdot t) \).
Two parameters to choose\(^1\):

- Number of iterations \(t\).
- Size of starting block \(b\).

**Main question:** For a choice of block-size \(b\), how many matrix vector multiplications \(b \cdot t\) do we need to ensure that:

\[
\|A - QQ^TA\| \leq (1 + \epsilon)\|A - A_k\|
\]

\(^1\)Lots of other choices related to how to orthogonalize and post-process the Krylov subspace. These are not the focus of my talk, but very important!
Most prior theoretical work on $k$-rank approximation focuses on “large block” methods where $b \geq k$. 
LARGE BLOCK METHODS

Most prior theoretical work on $k$-rank approximation focuses on “large block” methods where $b \geq k$. Two main regimes:

**Sketching. Run for single iteration.**

$\begin{pmatrix} A \cdot A^T & B \end{pmatrix} \begin{pmatrix} \pm 1 & \pm 1 & \pm 1 & \pm 1 \\ \pm 1 & \pm 1 & \pm 1 & \pm 1 \\ \pm 1 & \pm 1 & \pm 1 & \pm 1 \\ \pm 1 & \pm 1 & \pm 1 & \pm 1 \end{pmatrix}$

$O(k/\varepsilon)$

[ Sarlós, 2006 ] [ Martinsson, Rokhlin, Tygert, 2006 ] [ Halko, Martinsson, Tropp 2011 ] [ Clarkson, Woodruff 2009, 2013 ]
Most prior theoretical work on $k$-rank approximation focuses on “large block” methods where $b \geq k$. Two main regimes:

**Block size $k$. Run for multiple iterations**

[Rokhlin, Szlam, and Tygert, 2009] [Halko, Martinsson, Tropp 2011] [Gu, 2015] [Musco, Musco 2015] [Drineas, Ipsen, Kontopoulous, Magdon-Ismail 2017]
Surely the case of $b = 1$ has also been studied? This is textbook single vector Krylov iteration.

**Yes and no.** Single vector Krylov methods have been studied extensively for the problem of eigenvector/eigenspace approximation. But this problem differs in subtle but important ways from the low-rank approximation problem.
1. **Effectively take advantage of parallelism.** Multiplying $AA^T$ by $k$ vectors all at once might not be that much more expensive than multiplying by 1 vector.

2. **Do not require iteration for very large $b$.** Useful in streaming and distributed computing environments, for constrained low-rank approximation, and more.

3. **Enjoy “gap-independent” theoretical guarantees.**
1. **Effectively take advantage of parallelism.** Multiplying $AA^T$ by $k$ vectors all at once might not be that much more expensive than multiplying by 1 vector.

2. **Do not require iteration for very large** $b$. Useful in streaming and distributed computing environments, for constrained low-rank approximation, and more.

3. **Enjoy “gap-independent” theoretical guarantees.**
**Theorem (Musco, Musco 2015)**

When run for $O \left( \frac{\log d}{\sqrt{\epsilon}} \right)$ iterations with a random starting block with $b = k$ columns, the Krylov method returns a rank $k$ matrix $Q$ satisfying:

$$\|A - QQ^T A\|_{2,F} \leq (1 + \epsilon)\|A - A_k\|_{2,F}$$

$O \left( \frac{k \log d}{\sqrt{\epsilon}} \right)$ matrix-vector multiplications total.
Theorem (Musco, Musco 2015)

When run for \( O \left( \frac{\log d}{\sqrt{\epsilon}} \right) \) iterations with a random starting block with \( b = k \) columns, the Krylov method returns a rank \( k \) matrix \( Q \) satisfying:

\[
\| A - QQ^T A \|_{2,F} \leq (1 + \epsilon) \| A - A_k \|_{2,F}
\]

\( O \left( \frac{k \log d}{\sqrt{\epsilon}} \right) \) matrix-vector multiplications total.

In contrast, guarantees for eigenvector approximation always depend on matrix dependent quantities like \( \frac{\sigma_{i+1}}{\sigma_i - \sigma_{i+1}} \).
The following matrix has top singular vectors $e_1$ and $e_2$:

\[
\begin{pmatrix}
1 & 1 \\
1 & .999 \\
.5 & .4
\end{pmatrix}
\]

Convergence to the top subspace, $\text{span}(e_1, e_2)$, inherently depends polynomially on the inverse gap $\frac{\sigma_3}{\sigma_2 - \sigma_3}$.

But for accurate rank 2 approximation, we can converge to any two vectors in the span of $e_1, e_2, e_3$. No gap dependence necessary.
The following matrix has top singular vectors $e_1$ and $e_2$:

\[
\begin{pmatrix}
1 & 1 \\
.999 & .5 \\
.4 & \\
\end{pmatrix}
\]

Convergence to the top subspace, $\text{span}(e_1, e_2)$, inherently depends polynomially on the inverse gap $\frac{\sigma_3}{\sigma_2 - \sigma_3}$.

For more discussion, see e.g. Low-Rank Matrix Approximations Do Not Need a Singular Value Gap [Drineas, Ipsen, 2019].
Without additional modifications, the convergence of single vector Krylov iteration must depend inversely on spectral gaps. **Even for low-rank approximation!**

\[
\begin{bmatrix}
1 & 1 \\
1 & 0 \\
0 & 0 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
\cdot.1 \\
\cdot.3 \\
\cdot.4 \\
\cdot.2 \\
\cdot.1
\end{bmatrix}
\begin{bmatrix}
\cdot.1 & \cdot.3 & \cdot.3 & \cdot.3 & \cdot.3 \\
\cdot.3 & \cdot.3 & \cdot.3 & \cdot.3 & \cdot.3 \\
\cdot.4 & 0 & 0 & 0 & 0 \\
\cdot.2 & 0 & 0 & 0 & 0 \\
\cdot.1 & 0 & 0 & 0 & 0
\end{bmatrix}
\]

**Krylov subspace never contains a good low-rank approximation.**

\[\|A - A_2\|_F = 0 \text{ but } \|A - QQ^T A\|_F \text{ is never less than 1, no matter how many iterations we take.}\]
“Despite decades of research on [single vector] Lanczos methods, the theory for [large block] randomized methods is more complete and provides strong guarantees of excellent accuracy, whether or not there exist any gaps between the singular values.” – Slzam, Kluger, Tygert, 2014
Most major numerical linear algebra libraries use a single random starting vector. And they usually work fine!
If you are careful about numerical issues (e.g. build the Krylov subspace with sufficient reorthogonalization) it is hard to find a problem where single vector Krylov methods lose to large block methods (in terms of number of matrix-vector products).
THEORY PRACTICE GAP

Repeated Singular Values

Polynomial Decay ($\beta = 0.5$)

Polynomial Decay ($\beta = 1.5$)

Relative Error ($\varepsilon_{\text{empirical}}$)

Number of Matrix-Vector Products

appu Matrix

human\_gene\_2 Matrix

exdata\_1 Matrix

Block Size 1  Block Size 2  Block Size 3  Block Size 50  Block Size 54
Research question: For low-rank approximation, when and why do single vector Krylov methods require the same or fewer matrix-vector multiplications than large block methods?
Research question: For low-rank approximation, when and why do single vector Krylov methods require the same or fewer matrix-vector multiplications than large block methods?

Answer: For low-rank approximation, single vector methods depend on gaps, but only in a very mild way!

Up to a logarithmic dependence on spectral gaps, single vector methods match (or beat) the performance of large block Krylov iteration run with any block size $b \geq k$. 
**Theorem (Existing Large Block Result)**

When run for $O \left( \frac{\log d}{\sqrt{\epsilon}} \right)$ iterations with block size $b = k$, the Krylov method returns a rank $k$ matrix $Q$ satisfying:

$$
\| A - QQ^T A \|_{2,F} \leq (1 + \epsilon) \| A - A_k \|_{2,F}
$$

Let $g_{\text{min}} = \min_{i=1,\ldots,k-1} \left( \frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+1}} \right)$ be the minimum singular value gap in the top subspace.

**Theorem (Small block result)**

When run for $O \left( \frac{k \log(1/g_{\text{min}}) + \log d}{\sqrt{\epsilon}} \right)$ iterations with block size $b = 1$, the Krylov method returns a rank $k$ matrix $Q$ satisfying:

$$
\| A - QQ^T A \|_{2,F} \leq (1 + \epsilon) \| A - A_k \|_{2,F}
$$
EXAMPLE RESULTS

**Block size** $k$ **method:** $O\left( \frac{k \log d}{\sqrt{\epsilon}} \right)$ matvecs for error $\epsilon$.

**Single vector method:** $O\left( \frac{k \log(1/g_{\text{min}}) + \log d}{\sqrt{\epsilon}} \right)$ matvecs for error $\epsilon$. 

### Theorem (Existing Large Block Result)

When run for $O \left( \frac{\log(d/\epsilon)}{\sqrt{(\sigma_k-\sigma_{k+p})/\sigma_{k+p}}} \right)$ iterations with block size $b = k + p$, Krylov iteration returns a rank $k$ Q satisfying:

$$\|A - QQ^T A\|_{2,F} \leq (1 + \epsilon)\|A - A_k\|_{2,F}$$

Let $g_{\text{min}} = \min_{i=1,...,k+p-1} \left( \frac{\sigma_i-\sigma_{i+1}}{\sigma_{i+1}} \right)$.

### Theorem (New small block result)

When run for $O \left( \frac{(k+p) \log(1/g_{\text{min}}) + \log(d/\epsilon)}{\sqrt{(\sigma_k-\sigma_{k+p})/\sigma_{k+p}}} \right)$ iterations with block size $b = 1$, Krylov iteration returns a rank $k$ Q satisfying:

$$\|A - QQ^T A\|_{2,F} \leq (1 + \epsilon)\|A - A_k\|_{2,F}$$
EXAMPLE RESULTS

Block size $k + p$: $O \left( \frac{(k+p) \log(d/\epsilon)}{\sqrt{(\sigma_k - \sigma_{k+p})/\sigma_{k+p}}} \right)$ matvecs for error $\epsilon$.

Single vector: $O \left( \frac{(k+p) \log(1/g_{\min}) + \log(d/\epsilon)}{\sqrt{(\sigma_k - \sigma_{k+p})/\sigma_{k+p}}} \right)$ matvecs for error $\epsilon$, where we can minimize over $p$.

Single-vector method obtains the “best of all worlds” without having to do any parameter selection.
[Bakshi, Clarkson, Woodruff, 2022] shows how to compute $Q$ satisfying:

$$\|A - QQ^T A\|_F \leq (1 + \epsilon)\|A - A_k\|_F$$

using $\tilde{O}(k/\epsilon^{1/3})$. They run Krylov iteration with block sizes $k$ and $k/\epsilon^{1/3}$ and take the best result in both subspaces.

**Corollary of our result:** Up to a logarithmic dependence on the minimum gap, we obtain the same bound with a simpler, parameter free method – simply run single vector Krylov.
All of these theorems can be proven using a single “gray box” reduction.
All of these theorems can be proven using a single “gray box” reduction. **Start with a naive observation:**

\[
\text{span} \left[ \begin{bmatrix} b & \mathbf{A} \mathbf{A}^T b & (\mathbf{A} \mathbf{A}^T)^2 b & \cdots & (\mathbf{A} \mathbf{A}^T)^t b \end{bmatrix} \right]
\]

is equal to

\[
\text{span} \left[ \begin{bmatrix} \mathbf{S}_k & \mathbf{A} \mathbf{A}^T \mathbf{S}_k & (\mathbf{A} \mathbf{A}^T)^2 \mathbf{b} & \cdots & (\mathbf{A} \mathbf{A}^T)^{t-k+1} \mathbf{S}_k \end{bmatrix} \right]
\]

where \( \mathbf{S}_k = \begin{bmatrix} b & \mathbf{A} \mathbf{A}^T b & \cdots & (\mathbf{A} \mathbf{A}^T)^{k-1} b \end{bmatrix} \) contains the first \( k \) columns of the Krylov subspace.
Single vector iteration is equivalent to large block iteration with a particular choice of starting block.
**Upside:** Every additional “iteration” of the block method only requires one matrix-vector multiply with $\mathbf{AA}^T$. Typically would need $kt$ matrix-vector multiplies to iterate a block of size $k$ for $t$ iterations. Only need $k + t$ if our starting block is $\mathbf{S}_k$.

**Downside:** The starting block $\mathbf{S}_k$ looks very different from a random matrix. E.g. will typically be numerically low-rank.

$$\mathbf{S}_k = \begin{bmatrix} \mathbf{b} & \mathbf{AA}^T \mathbf{b} & \cdots & \mathbf{(AA}^T)^k \mathbf{b} \end{bmatrix}$$

This is huge upside and a huge downside.
Almost all prior analysis of block Krylov methods depend on showing that the random starting block \( \mathbf{B} \) satisfies a condition similar to the following:

**Definition** \(((k, L))-good starting matrix\):  
A starting matrix \( \mathbf{B} \in \mathbb{R}^{d \times k} \) is a \((k, L)\)-good starting matrix for \( \mathbf{A} \) if,  
\[
\mathbf{Q} = \text{span}(\mathbf{B}) \text{ satisfies } \| (\mathbf{U}_k^T \mathbf{Q})^{-1} \|_2^2 \leq L.
\]

We want \( L \) to be as small as possible. Requires that \( \mathbf{Q} \) has good inner product with any vector in the span of \( \mathbf{U}_k \).

Equivalent to requiring that all principal angles between subspaces \( \mathbf{U}_k \) and \( \mathbf{Q} \) have \( \cos(\theta_i) \geq 1/\sqrt{L} \).
Almost all prior analysis of block Krylov methods depend on showing that the random starting block block $B$ satisfies a condition similar to the following:

**Definition ($(k, L)$-good starting matrix)**

A starting matrix $B \in \mathbb{R}^{d \times k}$ is a $(k, L)$-good starting matrix for $A$ if, $Q = \text{span}(B)$ satisfies $\| (U_k^T Q)^{-1} \|_2^2 \leq L$.

We want $L$ to be as small as possible. Requires that $Q$ has good inner product with any vector in the span of $U_k$.

**Now standard result [Rudelson, Vershynin 2010]:** A random Gaussian matrix is $(k, L)$ good with probability $1 - \delta$ for:

$$L = O(kd/\delta).$$
RESULT FOR GENERIC STARTING BLOCKS

Theorem (Musco, Musco 2015)

When run for \( O \left( \frac{\log dL}{\sqrt{\epsilon}} \right) \) iterations with a \((k, L)\)-good starting block, the block Krylov method returns a rank \( k \) matrix \( Q \) satisfying:

\[
\|A - QQ^TA\|_{2,F} \leq (1 + \epsilon)\|A - A_k\|_{2,F}
\]

Plugging in \( L = O(kd/\delta) \) gives the desired bound for a random starting block.
Main question: Is the following matrix \((k, L)\) good?

\[
S_k = \begin{bmatrix}
b & AA^Tb & \cdots & (AA^T)^k b
\end{bmatrix}
\]
Main question: Is the following matrix \((k, L)\) good?

\[
S_k = \begin{bmatrix} b & AA^T b & \cdots & (AA^T)^k b \end{bmatrix}
\]

Claim: With probability \((1 - \delta)\), \(S_k\) is \((k, L)\)-good for:

\[
L = \text{poly}(k, d, 1/\delta) \cdot \frac{1}{g_{\min}^{4k}},
\]

where \(g_{\min} = \min_{i=1,\ldots,k-1} \left( \frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+1}} \right)\).
Main question: Is the following matrix \((k, L)\) good?

\[
S_k = \begin{bmatrix}
  b & AA^T b & \cdots & (AA^T)^k b
\end{bmatrix}
\]

Claim: With probability \((1 - \delta)\), \(S_k\) is \((k, L)\)-good for:

\[
L = \text{poly}(k, d, 1/\delta) \cdot \frac{1}{g_{\text{min}}^{4k}},
\]

where \(g_{\text{min}} = \min_{i=1,\ldots,k-1} \left( \frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+1}} \right)\).

This is exponentially worse than what we know for random matrices.
Which is great!!
### Theorem (Musco, Musco 2015)

When run for \( O\left(\frac{\log d L}{\sqrt{\epsilon}}\right) \) iterations with a \((k, L)\)-good starting block, the block Krylov method returns a rank \(k\) matrix \(Q\) satisfying:

\[
\|A - QQ^T A\|_{2,F} \leq (1 + \epsilon)\|A - A_k\|_{2,F}
\]

Plugging in \(L = \text{poly}(k, d, 1/\delta) \cdot \frac{1}{g_{\min}^{4k}}\) gives the desired bound for the single vector starting block \(S_k\) with:

\[
t = O\left(\frac{k \log(1/g_{\min}) + \log(d/\delta)}{\sqrt{\epsilon}}\right)
\]

iterations.
Theorem (Musco, Musco 2015)

When run for \( O \left( \frac{\log dL}{\sqrt{\epsilon}} \right) \) iterations with a \((k, L)\)-good starting block, the block Krylov method returns a rank \( k \) matrix \( Q \) satisfying:

\[
\| A - QQ^T A \|_{2,F} \leq (1 + \epsilon) \| A - A_k \|_{2,F}
\]

Plugging in \( L = \text{poly}(k, d, 1/\delta) \cdot \frac{1}{g_{\text{min}}^{4k}} \) gives the desired bound for the single vector starting block \( S_k \) with:

\[
t = O \left( k \log \left( \frac{1}{g_{\text{min}}} \right) + \log \left( \frac{d}{\delta} \right) \right) \sqrt{\frac{\epsilon}{\delta}} \text{ iterations.}
\]

And running this many iterations only requires \( t + k \) matrix-vector multiplications with \( AA^T \).
How to show that $S_k = \begin{bmatrix} b & AA^T b & \cdots & (AA^T)^k b \end{bmatrix}$ is $(k, L)$-good?

Reduce to a problem about polynomials, by taking advantage of the fact that:

$$\| (U_k^T Q)^{-1} \|_2^2 = \| S_k (U_k^T S_k)^{-1} \|_2^2 \max_x \frac{\| S_k x \|_2^2}{\| U_k^T S_k x \|_2^2}.$$
How to show that $S_k = \begin{bmatrix} b & AA^T b & \cdots & (AA^T)^k b \end{bmatrix}$ is $(k, L)$-good?

Reduce to a problem about polynomials, by taking advantage of the fact that:

$$\| (U_k^T Q)^{-1} \|_2^2 = \| S_k (U_k^T S_k)^{-1} \|_2^2 \max_x \frac{\| S_k x \|_2^2}{\| U_k^T S_k x \|_2^2}.$$  

**After applying some Gaussian anti-concentration, need to bound:**

$$\max_{\text{degree } k-1 \text{ poly } p} \frac{\max_{i \in 1, \ldots, d} p(\sigma_i^2)}{\max_{i \in 1, \ldots, k} p(\sigma_i^2)}.$$
Need to bound:

\[ \max_{\text{degree } k-1 \text{ poly } p} \frac{\max_{i \in k+1, \ldots, d} p(\sigma_i^2)}{\max_{i \in 1, \ldots, k} p(\sigma_i^2)} \]
SINGLE VECTOR ANALYSIS

Need to bound:

\[
\max_{\text{degree } k-1 \text{ poly } p} \frac{\max_{i \in k+1, \ldots, d} p(\sigma_i^2)}{\max_{i \in 1, \ldots, k} p(\sigma_i^2)}
\]

Obtain a bound of \(O(1/g_{min}^{4k})\) following an approach from [Saad, 1980].
**Punchline:** For low-rank approximation, single vector Krylov iteration matches large block methods (in terms of matrix-vector products) up to a logarithmic factor.

**Theorem**

When run for \( O \left( \frac{k \log(1/g_{\text{min}}) + \log d}{\sqrt{\epsilon}} \right) \) iterations with block size \( b = 1 \), the Krylov method returns a rank \( k \) matrix \( Q \) satisfying:

\[
\| A - QQ^T A \|_{2,F} \leq (1 + \epsilon) \| A - A_k \|_{2,F}
\]
Other results in the paper:

- Bounds for block size $1 < b < k$. Depend on minimum $b^{th}$-order gap.
- By adding a small random perturbation to $A$, we can replace the $\log(1/g_{min})$ with $\log(d)$ without hurting accuracy.
- More experiments.
**Future work:**

- Better understanding finite precision. Can we obtain similar results with partial orthogonalization?
- Improving bounds for block size $1 < b < k$. For $b = c \cdot k$, we require $O(k^2)$ matrix-vector multiplies. Not sure if tight.
- Are our current low-rank approximation algorithms optimal in terms of matrix-vector products? [Bakshi, Narayanan 2023].
questions?