On the **Unreasonable** Effectiveness Of Single Vector Krylov for Low-Rank Approximation

By: Raphael A. Meyer, Cameron Musco, Christopher Musco

NYU UMass Amherst NYU

NEW *Accepted at SODA 2024!*

Low-Rank Approximation

Given $\overset{\text{SPSD}}{A} \in \mathbb{R}^{n \times n}$, k, $\varepsilon > 0$ find ortho $Q \in \mathbb{R}^{n \times k}$ with

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

Ideally, $Q =$ top k eigenvectors of A, so use Krylov

[Rokhlin et al. '09], [Halko et al. '11], [Drineas Ipsen '19], [Tropp '22], ...

Block Krylov

1. Pick a start block Usually Gaussian

   $B \in \mathbb{R}^{n \times b}$ b = block size

2. Build Krylov subspace

   $$Z = \mathrm{orth}(\mathbb{K}) = \mathrm{orth}([B \;\; AB \;\; \cdots \;\; A^t B])$$

3. Return a solution

   $Q = Z^T U_k$ where U_k = top k eigvecs of $Z^T A A^T Z$

But how should we pick b?

How should we pick b?

1. Large block size $b \geq K$

Rich line of work [Tropp, Halko, Martinson, Gu, Drineas, Ipsen, Woodruff, ...]

Strong theoretical results for L.R.A. specifically

Gap-Independent Convergence [Musco Musco '15]

$b=K,\ [B]_{i,j} \sim N(0,1) \implies t = O\left(\frac{1}{\sqrt{\varepsilon}} \log\left(\frac{n}{\varepsilon}\right)\right)$ suffices

Let $g_{k \to b} = \frac{\lambda_k - \lambda_{b+1}}{\lambda_k}$

Spectral Decay Convergence

[Musco Musco '15]

$b \geq k$, $[B]_{ij} \sim N(0,1)$ $\Longrightarrow$ $t = O\left(\frac{1}{\sqrt{g_{k \to b}}} \log\left(\frac{n}{\varepsilon}\right)\right)$ suffices

Let $b = k+2,\ k+5,\ k+10$

How should we pick b?

2. Small block size $b \ll k$

$b=1$ is called "Single Vector Krylov"

Classical NLA suggests $b \approx$ size of eigval clusters

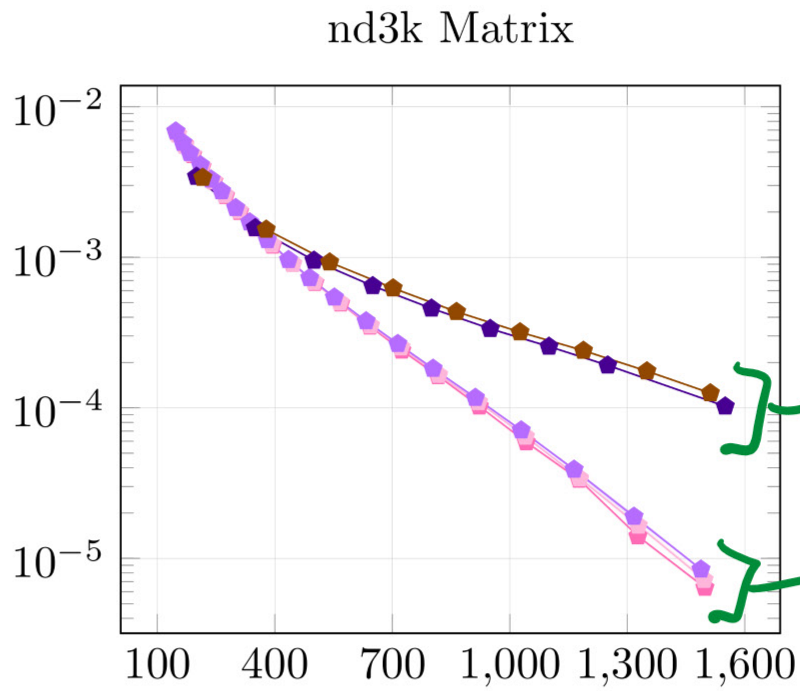
Lack results* for Low-Rank Approximation

Cannot be gap-independent

In practice, $b=1$ often just works well

(K=50)

Error (ε)



Block Sizes K, K+4

Block Sizes 1,2,3

Matvecs

A theory/practice gap!

When and why do small block methods match/outperform large block methods for low-rank approximation?

Caveats: Infinite Precision, Matvec Complexity

Main Result

For all $b \geq K$

Number of matvecs needed by Single Vector Kyrlov

is less than

Number of matvecs needed by block size b Krylov

If any $b \geq k$ is fast, then single vector is fast

Up to log dependence on eigengaps

Main Result (Rigorous)

Let g_{min} = smallest gap between any of top K eigs

$$= \min_{i=1,\cdots,K-1} \frac{\lambda_i - \lambda_{i+1}}{\lambda_{i+1}}$$

Then,

$b=1$ converges in $t = O\left(\frac{K}{\sqrt{\varepsilon}} \log\left(\frac{1}{g_{min}}\right) + \frac{1}{\sqrt{\varepsilon}} \log\left(\frac{n}{\varepsilon}\right)\right)$

For all $l \geq K$, converges in $t = O\left(\frac{l}{\sqrt{g_{K \to l}}} \log\left(\frac{1}{g_{min}}\right) + \frac{1}{\sqrt{g_{K \to l}}} \log\left(\frac{n}{\varepsilon}\right)\right)$

If some $b \preceq K$ gets linear convergence,

$$O\left(\frac{K}{\sqrt{g_{K\to b}}}\log\left(\frac{n}{\varepsilon}\right)\right) \quad \text{vs} \quad O\left(\frac{K}{\sqrt{g_{K\to b}}}\log\left(\frac{1}{g_{min}}\right)+\frac{1}{\sqrt{g_{K\to b}}}\log\left(\frac{n}{\varepsilon}\right)\right)$$

Upside: separate K from ε

Downside: depends on g_{min}

Key Observation: A Silly Manipulation

Suppose $b=1$, so for $\underline{x} \sim N(0, I)$

$$\mathrm{span}(K) = \mathrm{span}([\underline{x} \ A\underline{x} \ A^2\underline{x} \cdots A^t\underline{x}])$$

Now, repeat some columns

$$= \mathrm{span}([\underbrace{\underline{x} \ A\underline{x} \cdots A^{\ell}\underline{x}}_{S_\ell} \ \underbrace{A\underline{x} \ A^2\underline{x} \cdots A^{\ell+1}\underline{x}}_{AS_\ell} \ \underbrace{A^2\underline{x} \ A^3\underline{x} \cdots A^{\ell+2}\underline{x}}_{A^2S_\ell} \cdots\cdots \underbrace{A^{t-\ell}\underline{x} \cdots A^t\underline{x}}_{A^{t-\ell}S_\ell}])$$

$$= \mathrm{span}([S_\ell \quad AS_\ell \quad A^2S_\ell \quad \cdots \quad A^{t-\ell}S_\ell])$$

Where $S_\ell = [\underline{x} \ A\underline{x} \cdots A^{\ell}\underline{x}]$ is our Simulated Start Block

$b=1$ Krylov Subspace of degree t starting from $\underline{x} \sim N(0, I)$ $=$ $b=\ell$ Krylov Subspace of degree $t-\ell$ starting from S_ℓ

Upside: 1 matvec = 1 iteration of block krylov

Downside: S_ℓ is a bad starting block

$$S_\ell = [\underline{x} \;\; A\underline{x} \;\; \cdots \;\; A^\ell \underline{x}]$$

Let $B \in \mathbb{R}^{n \times b}$ be an L-good Starting Matrix. Then, [Musco Musco '15]

$b = K$ converges in $O\left(\frac{1}{\sqrt{\varepsilon}} \log\left(\frac{nL}{\varepsilon}\right)\right)$ iterations

$b \geq K$ converges in $O\left(\frac{1}{\sqrt{g_{K \to b}}} \log\left(\frac{nL}{\varepsilon}\right)\right)$ iterations

$[B]_{ij} \sim N(0,1)$ has $L = O(nb)$

$b = K \Rightarrow O\left(\frac{1}{\sqrt{\varepsilon}} \log\left(\frac{n}{\varepsilon}\right)\right)$

[New Result]

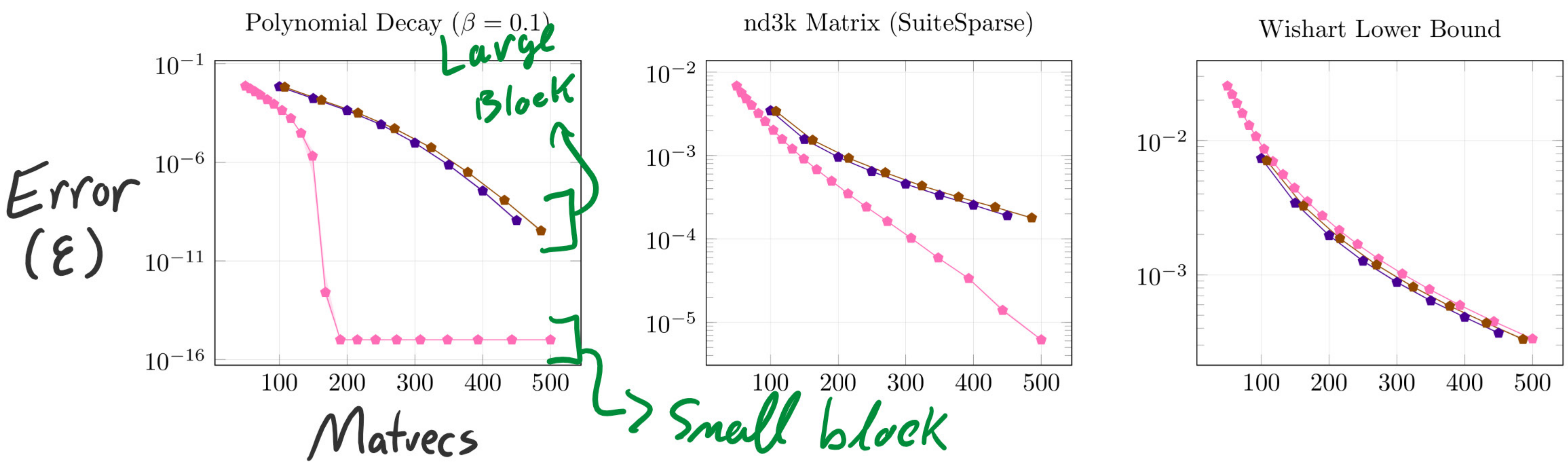
S_ℓ has $L = O\left(\frac{n \ell^3}{g_{min}^{4\ell}}\right)$

$\ell = K \Rightarrow O\left(\frac{K}{\sqrt{\varepsilon}} \log\left(\frac{1}{g_{min}}\right) + \frac{1}{\sqrt{\varepsilon}} \log\left(\frac{n}{\varepsilon}\right)\right)$

$\ell \geq K \Rightarrow O\left(\frac{\ell}{\sqrt{g_{K \to \ell}}} \log\left(\frac{1}{g_{min}}\right) + \frac{1}{\sqrt{g_{K \to \ell}}} \log\left(\frac{n}{\varepsilon}\right)\right)$

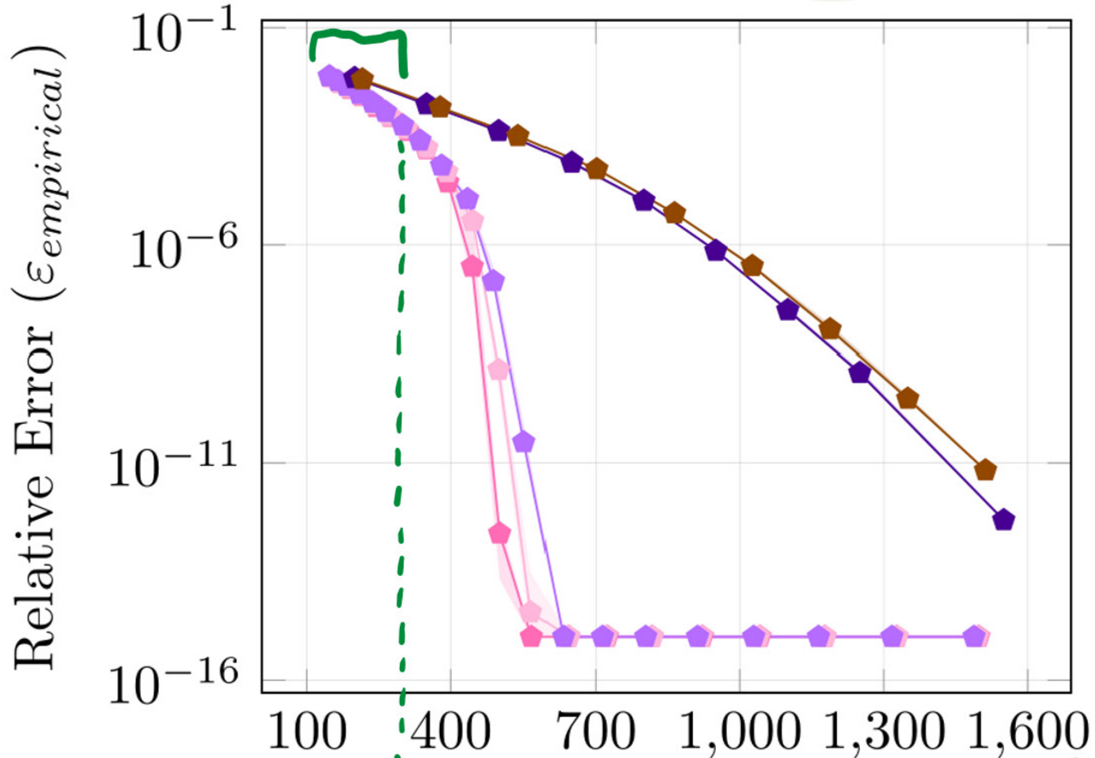
Q: *When are single vector methods faster?*

A: *When some block size achieves linear convergence*



Simulated blocks may explain slow-then-fast convergence

Can only simulate small blocks



Can now simulate large blocks

In the paper: Grab bag of more implications

- Beyond $b=1$
- Smoothed Analysis shatters g_{min}
- Simplify Fast-Frobenius L.R.A. [Bakshi et al. '22]
- Faster-ish Schatten-norm L.R.A
- Single Vector Subspace Iteration
- Experiments

Any questions?