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

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

NYU UMass Amherst NYU

Low-Rank Approximation

Given $A \in \mathbb{R}^{n \times n}$ (PSD), 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

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

Usually Gaussian

b = block size

2. Build Krylov subspace

$$Z = \text{orth}(\mathcal{K}) = \text{orth}([B \quad AB \quad \cdots \quad A^t B])$$

3. Return a solution

$$Q = Z^T U_K \quad \text{where} \quad U_K = \text{top } K \text{ eigvecs of } Z^T A A^T Z$$

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 result for L.R.A. specifically

Gap-Independent Convergence [Musco Musco '15]

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

Strong theoretical result for L.R.A. specifically

Gap-Independent Convergence

[Musco Musco '15]

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

Let $g_{K\to b}=\frac{\lambda_K-\lambda_{b+1}}{\lambda_K}$

Spectral Decay Convergence

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

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"

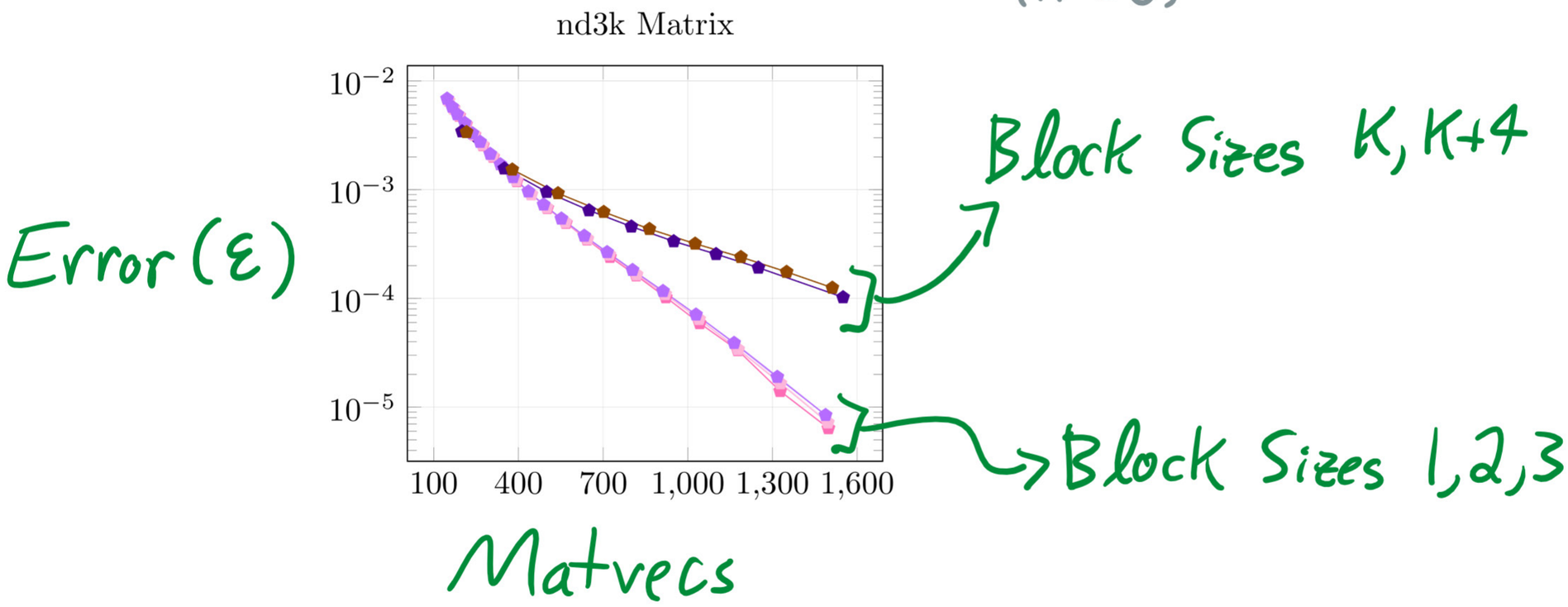
   Classical NLA suggests $b \approx$ size of eigval clusters

   Cannot be gap-independent

   Lack results* for Low-Rank Approximation

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

(K=50)



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 Pitch

Up to log dependence on eigengaps,

Single Vector's convergence rate (in matvecs)

is bounded by

Large block Krylov's convergence rate (in matvecs)
for all block sizes $b \geq K$

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

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

Proof by Black-box-ish Reduction

Let $g_{min} = \min_{i=1,\dots,K-1} \frac{\lambda_i - \lambda_{i+1}}{\lambda_{i+1}}$

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

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

Key Observation: A Silly Manipulation

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

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

Now, repeat some columns

$$= \text{span}([\underline{x} \; A\underline{x} \cdots A^{\ell}\underline{x} \;\; A\underline{x} \; A^2\underline{x} \cdots A^{\ell+1}\underline{x} \;\; A^3\underline{x} \; A^4\underline{x} \cdots A^{\ell+2}\underline{x} \;\; \cdots \;\; A^{t-\ell}\underline{x} \cdots A^t\underline{x}])$$

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

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 $\underline{x} \sim \mathcal{N}(0, I)$ $=$ $b=K$ Krylov Subspace of degree $t-K$ starting from S_K

Upside: 1 matvec = 1 iteration of block krylov

Downside: S_K is a bad starting block

$$S_K = [\underline{x} \; A\underline{x} \; \cdots \; A^K \underline{x}]$$

[Musco Musco '15]

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

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

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

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

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

[New Result]

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

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

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

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

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

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

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

[New Result]

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

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

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

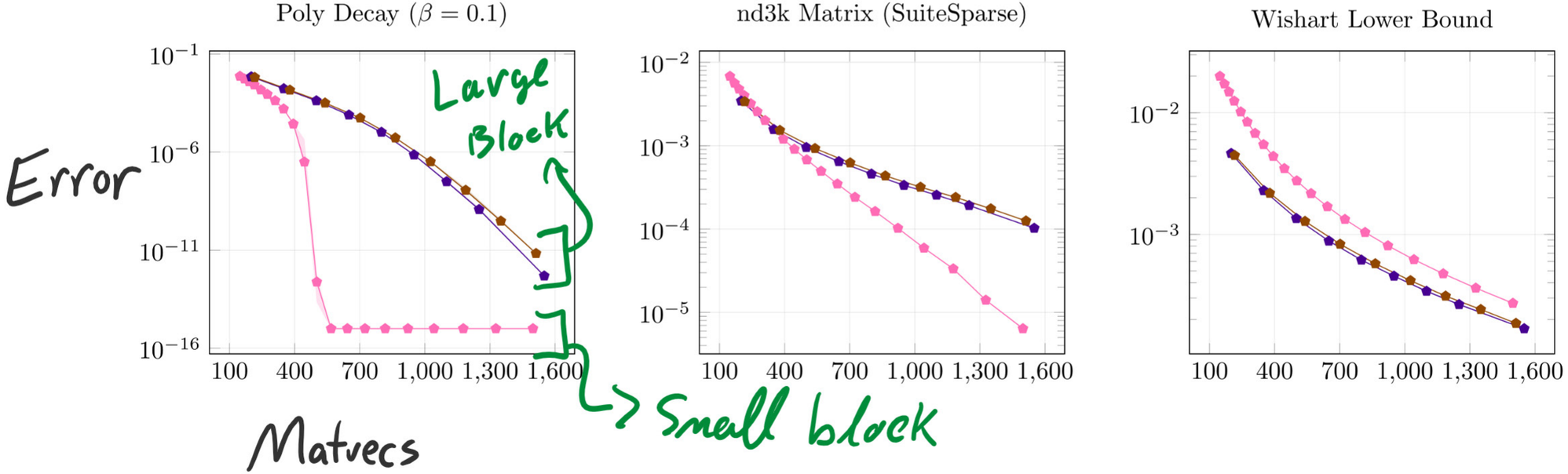
If some $b \asymp K$ gets linear convergence,

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

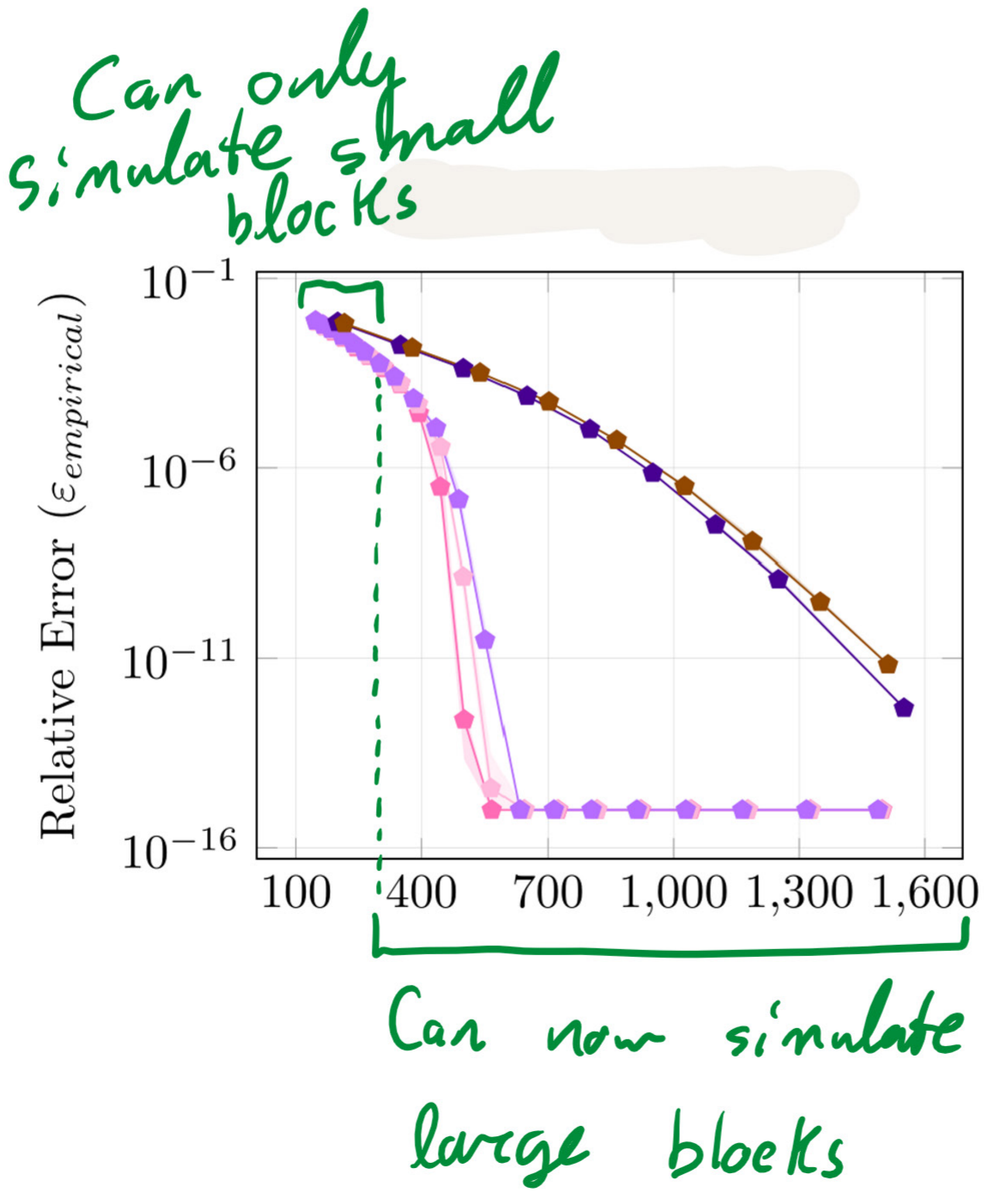
Upside: separate K from ε

Downside: depends on g_{min}

Single Vector Methods outperform large block methods if some block size achieves linear convergence*



Simulated blocks may explain slow-then-fast convergence



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
- Experiments

Any questions?