Why Single Vector Krylov is so Effective at Low-Rank Approximation

Raphael A. Meyer (New York University)

With Christopher Musco (New York University) and Cameron Musco (University of Massachusetts Amherst)

Approximate SVD / Low Rank Approximation

- Given $\boldsymbol{A} \in \mathbb{R}^{n \times d}$, target rank k, error tolerance $\varepsilon > 0$
- Return orthonormal matrix $\boldsymbol{Q} \in \mathbb{R}^{n \times k}$ such that

$$|\mathbf{q}_i^\intercal \boldsymbol{A}\boldsymbol{A}^\intercal \mathbf{q}_i - \sigma_i(\boldsymbol{A})^2| \leq \varepsilon \sigma_i(\boldsymbol{A})^2$$

Algorithm from [Musco & Musco '15]:

input: Block size b. Number of iterations t.
output: Orthonormal Matrix $\mathbf{Q} \in \mathbb{R}^{n \times k}$.

1: Sample $\mathbf{B} \in \mathbb{R}^{n \times b}$ with i.i.d. $\mathcal{N}(0,1)$ entries
2: $\mathbf{K} = [\mathbf{B},\ (\mathbf{A}\mathbf{A}^\intercal)\mathbf{B},\ \ldots,\ (\mathbf{A}\mathbf{A}^\intercal)^t\mathbf{B}]$
3: Compute an orthonormal basis $\mathbf{Z}$ for $\mathbf{K}$
4: Compute $\mathbf{U}_k$, the k top left singular vectors of $\mathbf{Z}^\intercal\mathbf{A}$
5: **return** $\mathbf{Q} = \mathbf{Z}\mathbf{U}_k$

Focusing on Block Size

How should we set the block size b and number of iterations t?

In Theory,

- Block size $b = k$ has sublinear convergence for all $\boldsymbol{A}$
- Block size $b = k + 251$ has linear convergence if $\sigma_{k+251} < 0.9\sigma_k$

In Practice,

- Block size $b = 1$ or 2 is good

Why Single Vector Krylov is so Effective at Low-Rank Approximation?

Focusing on Block Size

How should we set the block size b and number of iterations t?

In Theory,

- Block size $b = k$ has sublinear convergence for all $\boldsymbol{A}$
- Block size $b = k + 251$ has linear convergence if $\sigma_{k+251} < 0.9\sigma_k$

In Practice,

- Block size $b = 1$ or 2 is good

Why Single Vector Krylov is so Effective at Low-Rank Approximation?

Big Idea: Simulated Block Krylov

Single Vector Krylov simulates all larger-block Krylovs

To simulate block size $b = 3$, let $\boldsymbol{B} = [\, \mathbf{g}\ \boldsymbol{A}^2\mathbf{g}\ \boldsymbol{A}^4\mathbf{g} \,]$, then:

$$\begin{aligned}
\boldsymbol{K} &= \begin{bmatrix} \mathbf{g} & \boldsymbol{A}^2\mathbf{g} & \boldsymbol{A}^4\mathbf{g} & \boldsymbol{A}^6\mathbf{g} & \dots & \boldsymbol{A}^{2t}\mathbf{g} \end{bmatrix} \\
&\overset{\text{span}}{=} \begin{bmatrix} [\, \mathbf{g}\ \boldsymbol{A}^2\mathbf{g}\ \boldsymbol{A}^4\mathbf{g} \,] & [\, \boldsymbol{A}^2\mathbf{g}\ \boldsymbol{A}^4\mathbf{g}\ \boldsymbol{A}^6\mathbf{g} \,] & \dots & [\, \boldsymbol{A}^{2(t-2)}\mathbf{g}\ \boldsymbol{A}^{2(t-1)}\mathbf{g}\ \boldsymbol{A}^{2t} \,] \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{B} & \boldsymbol{A}^2\boldsymbol{B} & \dots & \boldsymbol{A}^{2(t-2)}\boldsymbol{B} \end{bmatrix}
\end{aligned}$$

Single Vector Krylov for t iterations $\iff$ Block Size 3 Krylov for $t-2$ iterations starting from $\boldsymbol{B}$

Big Idea: Simulated Block Krylov

Single Vector Krylov simulates all larger-block Krylovs

To simulate block size $b = 3$, let $\boldsymbol{B} = [\, \mathbf{g}\ \boldsymbol{A}^2\mathbf{g}\ \boldsymbol{A}^4\mathbf{g} \,]$, then:

$$\begin{aligned}\boldsymbol{K} &= \begin{bmatrix} \mathbf{g} & \boldsymbol{A}^2\mathbf{g} & \boldsymbol{A}^4\mathbf{g} & \boldsymbol{A}^6\mathbf{g} & \dots & \boldsymbol{A}^{2t}\mathbf{g} \end{bmatrix} \\ &\overset{\text{span}}{=} \begin{bmatrix} [\, \mathbf{g}\ \boldsymbol{A}^2\mathbf{g}\ \boldsymbol{A}^4\mathbf{g} \,] & [\, \boldsymbol{A}^2\mathbf{g}\ \boldsymbol{A}^4\mathbf{g}\ \boldsymbol{A}^6\mathbf{g} \,] & \dots & [\, \boldsymbol{A}^{2(t-2)}\mathbf{g}\ \boldsymbol{A}^{2(t-1)}\mathbf{g}\ \boldsymbol{A}^{2t} \,] \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{B} & \boldsymbol{A}^2\boldsymbol{B} & \dots & \boldsymbol{A}^{2(t-2)}\boldsymbol{B} \end{bmatrix}\end{aligned}$$

Single Vector Krylov for t iterations	$\iff$	Block Size b Krylov for $t - b + 1$ iterations starting from $\boldsymbol{B}$

How to Analyzed Simulated Blocks

Theorem: Inital Block isn't that Bad

Let b be the simulated block size. Let $g_{min} := \min_{i \in [b]} \frac{|\sigma_i - \sigma_{i+1}|}{\sigma_{i+1}}$.

Let $\boldsymbol{Z}$ span the columns of $\boldsymbol{A}\boldsymbol{A}^\intercal \boldsymbol{B}$. With probability $1 - \delta$,

$$\|\boldsymbol{A} - \boldsymbol{Z}\boldsymbol{Z}^\intercal \boldsymbol{A}\|_F \leq O\left(\frac{d^2}{\delta g_{min}^b}\right) \|\boldsymbol{A} - \boldsymbol{A}_b\|_F$$

Proof via bounds on Legendre interpolating polynomials [Saad '80]

Via existing iterative analysis, Block Krylov depends on

$$\log\left(\frac{d^2}{\delta g_{min}^b}\right) = b \log\left(\frac{1}{g_{min}}\right) + \log\left(\frac{d}{\delta}\right)$$

Sublinear & Linear Convergence [Musco & Musco '15]

Sublinear Convergence

We simulate block size $b = k$, so

$$t = O\left(\frac{k}{\sqrt{\varepsilon}} \log\left(\frac{1}{g_{min}}\right) + \frac{1}{\sqrt{\varepsilon}} \log\left(\frac{d}{\delta}\right)\right)$$

iterations suffice.

Linear Convergence

We simulate all block sizes $b \in [k+1, t]$, so if $\sigma_b \leq 0.9\sigma_k$,

$$t = O\left(\frac{b}{\sqrt{0.1}} \log\left(\frac{1}{g_{min}}\right) + \frac{1}{\sqrt{0.1}} \log\left(\frac{d}{\delta\varepsilon}\right)\right)$$

iterations suffice.

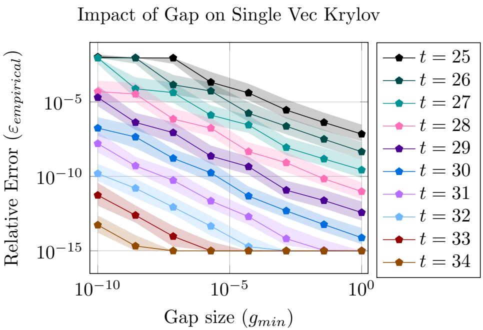
Verifying $\log(\frac{1}{g_{min}})$

$$C_0\, t = \frac{b}{\sqrt{0.1}} \log\left(\frac{1}{g_{min}}\right) + \frac{1}{\sqrt{0.1}} \log\left(\frac{1}{\varepsilon}\right)$$

Is equivalent to

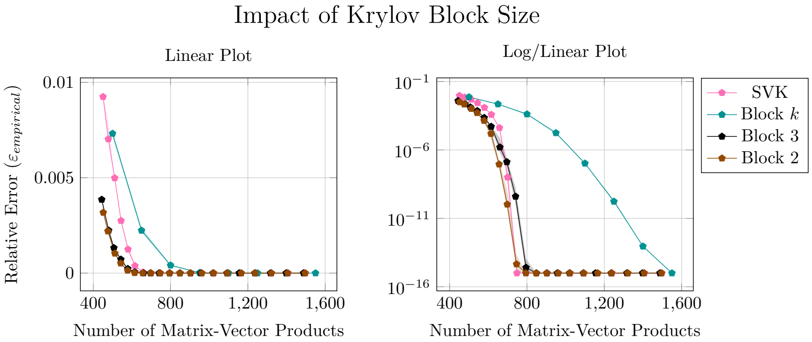
$$\log(\varepsilon) = -C_1 t - b \log(g_{min})$$

So we should see a line on a $\log(\varepsilon)/\log(g_{min})$ plot:



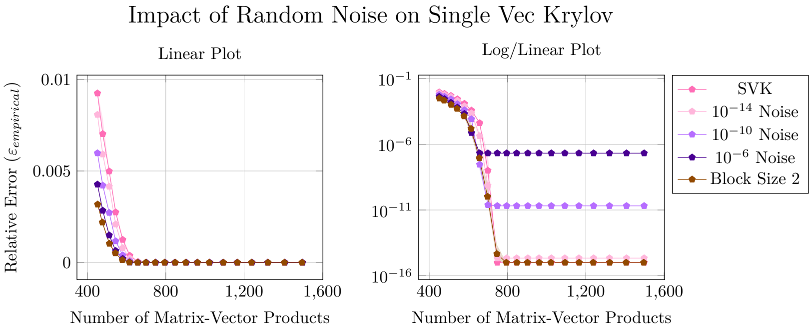
Small Block Sizes

- Actually using block size 2 simulates all even block sizes
- Robust to pairs of overlapping singular values



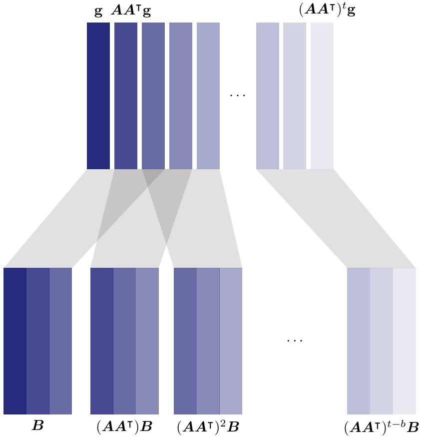
Eigenvalue Repulsion

- New topic in Random Matrix Theory: Tiny Gaussian Perturbations shatter eigenvalue gaps [Nguyen et al. '17]
- $\boldsymbol{A} + \Delta \boldsymbol{G}$ has $g_{min} \geq C_0 (\frac{\Delta}{d \|\boldsymbol{A}\|_2})^{17}$
- We can tradeoff convergence and accuracy with Δ



Summary / Conclusion

1. Single Vector Krylov simulates all larger block sizes
2. Explains slow-then-fast convergence
3. Extensions to larger blocks, random perturbations



THANK
YOU