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

$$|q_i^T A A^T q_i - \sigma_i(A)^2| \leq \varepsilon \sigma_i(A)^2$$

Algorithm from [Musco & Musco ’15]:

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

1. Sample $B \in \mathbb{R}^{n \times b}$ with i.i.d. $\mathcal{N}(0, 1)$ entries
2. $K = [B, (AA^T)B, \ldots, (AA^T)^t B]$
3. Compute an orthonormal basis $Z$ for $K$
4. Compute $U_k$, the $k$ top left singular vectors of $Z^T A$
5. return $Q = ZU_k$
How should we set the block size $b$ and number of iterations $t$?

In Theory,

- Block size $b = k$ has sublinear convergence for all $A$
- 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 $A$
- 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 $B = [g A^2 g A^4 g]$, then:

$$K = \begin{bmatrix} g & A^2 g & A^4 g & A^6 g & \ldots & A^{2t} g \end{bmatrix}$$

$$\text{span} = \begin{bmatrix} [g A^2 g A^4 g] & [A^2 g A^4 g A^6 g] & \ldots & [A^{2(t-2)} g A^{2(t-1)} g A^{2t}] \end{bmatrix}$$

$$= \begin{bmatrix} B & A^2 B & \ldots & A^{2(t-2)} B \end{bmatrix}$$

Single Vector Krylov for $t$ iterations $\iff$ Block Size 3 Krylov for $t - 2$ iterations starting from $B$
Big Idea: Simulated Block Krylov

Single Vector Krylov simulates all larger-block Krylov

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

$$K = \begin{bmatrix} g & A^2g & A^4g & A^6g & \ldots & A^{2t}g \end{bmatrix}$$

$$\text{span} = \begin{bmatrix} g \ A^2g \ A^4g \ A^6g \ \ldots \ A^{2t}g \end{bmatrix}$$

$$= \begin{bmatrix} B & A^2B & \ldots & A^{2(t-2)}B \end{bmatrix}$$

Single Vector Krylov for $t$ iterations $\iff$ Block Size $b$ Krylov for $t - b + 1$ iterations starting from $B$
Theorem: Initial 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 $Z$ span the columns of $AA^TB$. With probability $1 - \delta$,

$$\|A - ZZ^TA\|_F \leq O\left(\frac{d^2}{\delta g_{min}^b}\right) \|A - A_b\|_F$$

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 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 $\log\left(\frac{1}{g_{\text{min}}}\right)$

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

Is equivalent to

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

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

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

Impact of Krylov Block Size

Linear Plot

Log/Linear Plot

Number of Matrix-Vector Products

Relative Error ($\varepsilon_{\text{empirical}}$)
New topic in Random Matrix Theory: Tiny Gaussian Perturbations shatter eigenvalue gaps [Nguyen et al. ’17]

\[ A + \Delta G \text{ has } g_{min} \geq C_0 \left( \frac{\Delta}{d \| A \|_2} \right)^{17} \]

We can tradeoff convergence and accuracy with \( \Delta \)

Impact of Random Noise on Single Vec Krylov

- Linear Plot
  - Relative Error (\( \epsilon_{\text{empirical}} \)) vs. Number of Matrix-Vector Products

- Log/Linear Plot
  - Relative Error vs. Number of Matrix-Vector Products
  - Lines for different noise levels (SVK, \( 10^{-14} \) Noise, \( 10^{-10} \) Noise, \( 10^{-6} \) Noise, Block Size 2)
Summary / Conclusion

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

\[ g \quad (AA^T)g \quad (AA^T)^2g \quad (AA^T)^{t-b}B \]
THANK YOU