funNyström: Randomized low-rank approximation of monotone matrix functions

David Persson, Daniel Kressner

10 March 2023

EPFL

Problem statement

Given

1. **LARGE** $\boldsymbol{A} \succeq \boldsymbol{0}$ (numerically low rank);
2. *Operator monotone* f s.t. $f(0) = 0$ $(\boldsymbol{A} \succeq \boldsymbol{B} \Rightarrow f(\boldsymbol{A}) \succeq f(\boldsymbol{B}))$.

Find **LOW RANK** $\widehat{\boldsymbol{B}}$ such that

$$\widehat{\boldsymbol{B}} \approx f(\boldsymbol{A})$$

Operator monotone functions:

$$x, \quad \sqrt{x}, \quad x^r \text{ for } r \in [0,1], \quad \log(1+x), \quad \frac{x}{x+\mu} \text{ for } \mu > 0, \ldots$$

sums, compositions, positive scalings, ...

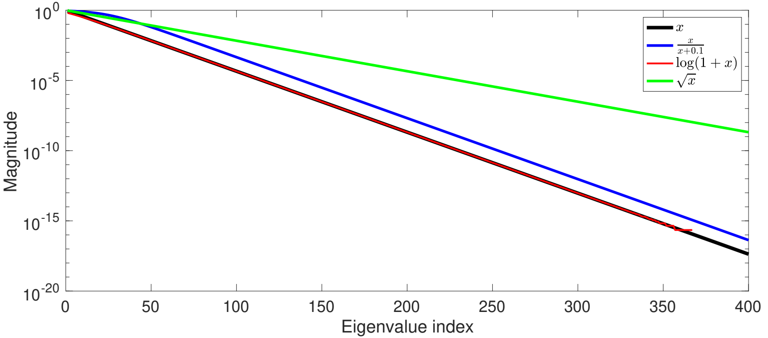
Problem statement

Why monotone and $f(0) = 0$? (Comment on *operator* monotonicity later)

f is continuous (implied by operator monotonicity)

$\Rightarrow f(\boldsymbol{A})$ is (numerically) low rank if $\boldsymbol{A}$ is

$\Rightarrow$Low rank approximation makes sense



Applications

1. Trace estimation:

$$\mathrm{tr}(\widehat{\boldsymbol{B}}) \approx \mathrm{tr}(f(\boldsymbol{A}))$$

- Nuclear norm estimation: $\sqrt{x}$;
- Statistical learning: $\log(1+x)$;
- Inverse problems: $\log(1+x)$, $\frac{x}{x+\mu}$.

2. Fast matvecs with $f(\boldsymbol{A})$:

$$\widehat{\boldsymbol{B}}\boldsymbol{x} \approx f(\boldsymbol{A})\boldsymbol{x}$$

- Sampling from elliptical distributions: $\sqrt{x}$.

3. Diagonal estimation:

$$\mathsf{diag}(\widehat{\boldsymbol{B}}) \approx \mathsf{diag}(f(\boldsymbol{A}))$$

- Ridge leverage scores: $\frac{x}{x+\mu}$.

Low rank approximation of matrix functions - First ideas

Method 1: Optimal approach via `eig`/`svd` $O(n^3)$
Method 2: Construct low rank approximation via matvecs.

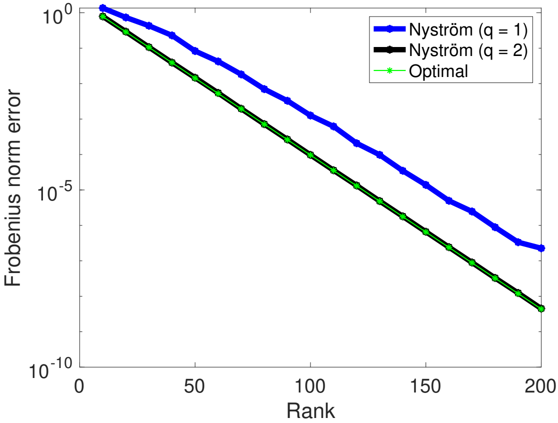
- Randomized SVD [Halko/Martinsson/Tropp'11]
- Nyström approximation $(f(\boldsymbol{A}) \succeq \mathbf{0})$
  [Gittens/Mahoney'13, Tropp/Yurtsever/Udell/Cevher'17]

1. Sample random $n \times (k + p)$ matrix $\boldsymbol{\Omega}$;
2. $\boldsymbol{Q} = \texttt{orth}(f(\boldsymbol{A})^{q-1}\boldsymbol{\Omega})$
3. Return $\widehat{\boldsymbol{B}} = f(\boldsymbol{A})\boldsymbol{Q}(\boldsymbol{Q}^T f(\boldsymbol{A})\boldsymbol{Q})^{\dagger}(f(\boldsymbol{A})\boldsymbol{Q})^T$.

Nyström costs $q(k+p)$ matvecs with $f(\boldsymbol{A})$!

Low rank approximation of matrix functions - First ideas

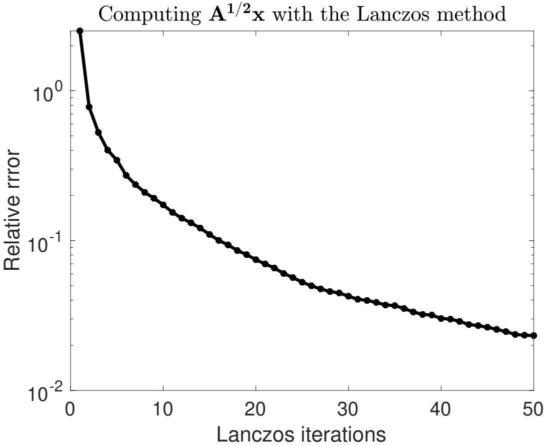
Nyström is very good!



But...

Low rank approximation of matrix functions - First ideas

Computing $f(\boldsymbol{A})\boldsymbol{\Omega}$ is expensive (compared to $\boldsymbol{A\Omega}$)!



(Rational Krylov, and other methods, are also 'expensive'.)

Low rank approximation of matrix functions - Better ideas?

We want: obtain low rank approximation using much fewer matvecs.

Back to the setting...

1. $\boldsymbol{A} \succeq \boldsymbol{0}$;
2. f is (operator) monotone and $f(0) = 0$.

Lemma: Let $\boldsymbol{A}_k$ be rank-k truncated SVD. Then...

$f(\boldsymbol{A}_k)$ is a best rank-k approximation to $f(\boldsymbol{A})$!

Idea: Compute Nyström approximation $\widehat{\boldsymbol{A}}$ of $\boldsymbol{A}$ and use approximation

$$f(\widehat{\boldsymbol{A}}) \approx f(\boldsymbol{A}).$$

Bypasses the need for matrix-vector products with $f(\boldsymbol{A})$!

Similar idea in trace estimation for $f(x) = x, \log(1+x), \frac{x}{x+1}$ [Saibaba/Alexanderian/Ipsen'17, Herman/Alexanderian/Saibaba'20].

funNyström

1. Sample random $n \times (k + p)$ matrix $\mathbf{\Omega}$ and obtain $\boldsymbol{Q} = \texttt{orth}(\boldsymbol{A}^{q-1}\mathbf{\Omega})$.
2. Obtain eigenvalue decomposition of Nyström approximation

$$\widehat{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q})^{\dagger}(\boldsymbol{A}\boldsymbol{Q})^T = \widehat{\boldsymbol{U}}\widehat{\mathbf{\Lambda}}\widehat{\boldsymbol{U}}^T.$$

3. Return low-rank approximation of $f(\boldsymbol{A})$

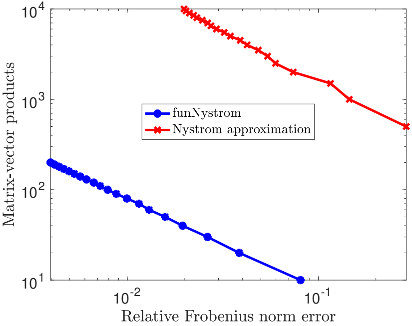
$$f(\widehat{\boldsymbol{A}}) = \widehat{\boldsymbol{U}} f(\widehat{\mathbf{\Lambda}})\widehat{\boldsymbol{U}}^T.$$

Potential benefits:

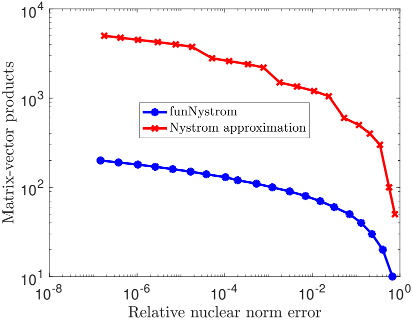
- No approximations of $f(\boldsymbol{A})\boldsymbol{x}$
  $\Rightarrow$ funNyström is much cheaper than Nyström on $f(\boldsymbol{A})$.
- It can even be more accurate!

Numerical results

How many matvecs do we save?



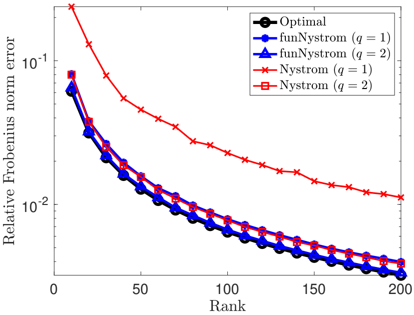
(a) $f(x) = x^{1/2}$ and $\lambda_i = i^{-3}$.



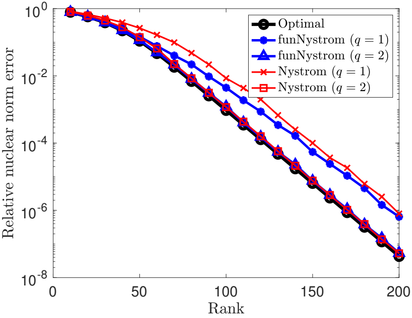
(b) $f(x) = \frac{x}{x+1}$ and $\lambda_i = 10e^{-i/10}$.

Numerical results

What if $f(\boldsymbol{A})\boldsymbol{x}$ is very cheap?



(a) $f(x) = x^{1/2}$ and $\lambda_i = i^{-3}$.



(b) $f(x) = \frac{x}{x+1}$ and $\lambda_i = e^{-i/10}$.

Theoretical results

Let $\gamma = \lambda_{k+1}/\lambda_k$, $q \geq 2$ and $\|f(\mathbf{\Lambda}_2)\|$ = best rank-k approx. error

$$\mathbb{E}\|f(\boldsymbol{A}) - f(\widehat{\boldsymbol{A}})\|_F^2 \leq \left(1 + \gamma^{2(q-3/2)}\frac{5k}{p-1}\right)\|f(\mathbf{\Lambda}_2)\|_F^2$$

Assumption $q \geq 2$ can be removed at the cost of a weaker bound.

Theoretical results

Let $\gamma = \lambda_{k+1}/\lambda_k$, $q \geq 2$ and $\|f(\boldsymbol{\Lambda}_2)\|$ = best rank-k approx. error

$$\mathbb{E}\|f(\boldsymbol{A}) - f(\widehat{\boldsymbol{A}})\|_F^2 \leq \left(1 + \gamma^{2(q-3/2)}\frac{5k}{p-1}\right)\|f(\boldsymbol{\Lambda}_2)\|_F^2$$

Assumption $q \geq 2$ can be removed at the cost of a weaker bound.

If $q \geq 1$

$$\mathbb{E}\|f(\boldsymbol{A}) - f(\widehat{\boldsymbol{A}})\|_* \leq \left(1 + \gamma^{2(q-1)}\frac{k}{p-1}\right)\|f(\boldsymbol{\Lambda}_2)\|_*$$

Theoretical results

Let $\gamma = \lambda_{k+1}/\lambda_k$, $q \geq 2$ and $\|f(\mathbf{\Lambda}_2)\| =$ best rank-k approx. error

$$\mathbb{E}\|f(\boldsymbol{A}) - f(\widehat{\boldsymbol{A}})\|_F^2 \leq \left(1 + \gamma^{2(q-3/2)} \frac{5k}{p-1}\right) \|f(\mathbf{\Lambda}_2)\|_F^2$$

Assumption $q \geq 2$ can be removed at the cost of a weaker bound.

If $q \geq 1$

$$\mathbb{E}\|f(\boldsymbol{A}) - f(\widehat{\boldsymbol{A}})\|_* \leq \left(1 + \gamma^{2(q-1)} \frac{k}{p-1}\right) \|f(\mathbf{\Lambda}_2)\|_*$$

If $q \geq 1$

$$\mathbb{E}\|f(\boldsymbol{A}) - f(\widehat{\boldsymbol{A}})\|_2 \leq \mathbb{E} f(\|\boldsymbol{A} - \widehat{\boldsymbol{A}}\|_2) \leq f(\mathbb{E}\|\boldsymbol{A} - \widehat{\boldsymbol{A}}\|_2) \leq$$

$$\|f(\mathbf{\Lambda}_2)\|_2 + \left\| f\left(\gamma^{2(q-1)} \frac{2k}{p-1} \mathbf{\Lambda}_2\right) \right\|_2 + \left\| f\left(\gamma^{2(q-1)} \frac{2e^2(k+p)}{p^2-1} \mathbf{\Lambda}_2\right) \right\|_*.$$

Other remarks

Application to trace estimation

funNyström + Hutch++ $\Rightarrow$ low rank approx. phase cheaper

Other remarks

Application to trace estimation

funNyström + Hutch++ $\Rightarrow$ low rank approx. phase cheaper

Operator monotonicity?

- Empirically, the bounds do not hold for any arbitrary monotone functions.
- $f(x) = x^3$ is an example...
- ... but funNyström still good provided you set $q = 3$!