funNystrom 2
or
Algorithm-agnostic low-rank approximation of operator monotone matrix functions

By: Raphael A. Meyer, David Persson, Christopher Musco
NYU EPFL NYU

Motivation

Given time series data, fit Gaussian Process to data with kernel function $K(\underline{x}, \underline{x}')$ and noise of variance σ

$$A = K + \sigma^2 I \quad , \quad [K]_{ij} = K(\underline{x}_i, \underline{x}_j)$$

To learn σ, maximize log-likeliood

$$\log \Pr[\underline{y} \mid \underline{x}_1, \cdots, \underline{x}_n, \sigma] = \qquad + \frac{1}{2} \underbrace{\log \det(K + \sigma^2 I)}_{\text{Need to compute this}}$$

How to compute $\log\det(K+\sigma^2 I)$?

$$\log\det(K+\sigma^2 I) = \mathrm{tr}(\log(A))$$

Naïve Algo:

1. Use Lanczos to approx $\underline{x} \to \log(A)\underline{x}$
2. Run Hutch++

Lanczos is slow. Can we do better?

Hutch++

1. Find $B \approx (\log(A))_K$
2. Compute $\underline{x}^T(\log(A)-B)\underline{x}$ many times

We focus on 1

$f(A)$ **Low-Rank Approximation**

Given $\overset{\text{PSD}}{A} \in \mathbb{R}^{n \times n}$, k, $\varepsilon > 0$, $f: \mathbb{R} \to \mathbb{R}$ find matrix $C \in \mathbb{R}^{n \times k}$ with

$$\|f(A) - C\|_{F,2} \leq (1+\varepsilon)\|f(A) - (f(A))_k\|_{F,2}$$

Naively: Lanczos to approximate $f(A)\underline{x}$, run Krylov

Ideally, what should C be?

Low-Rank Approx and Matrix Functions

Eigendecomposition $A = U \Lambda U^T$

$$\boxed{A} = \boxed{U} \boxed{\begin{matrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_k \end{matrix}} \boxed{U^T}$$

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$

Best rank-k approx = A_k = zero out all except k largest eigs

Matrix Function = $f(A) = U\, f(\Lambda) U^T$

$$\boxed{A} = \boxed{U} \boxed{\begin{matrix} f(\lambda_1) & & \\ & \ddots & \\ & & f(\lambda_k) \end{matrix}} \boxed{U^T}$$

Best rank-k approx $= (f(A))_k$ = zero all except k largest of $f(A)$

$f(x) \geq 0$, f increasing = zero all except k largest of A

$f(0) = 0$ $= f(A_k)$

$f(A_k)$ takes $O(n^2 k)$ time to compute!

Meta-algo

(our result, informally:
this is always a good idea)

Since $(f(A))_k = f(A_k)$

1. Find low-rank $B \approx A_k$
2. Compute $C = f(B)$ — $O(n^2 k)$ time
3. Return C_k — allows $f(0) \neq 0$, $\text{rank}(B) > k$

Avoids Lanczos! $C_k \approx (f(A))_k$ Without accessing $f(A)$!

Problem: B must be symmetric PSD for $f(B)$ to exist

funNystrom [Persson Kresner]

Nystrom Approximation

Given ortho $Q \in \mathbb{R}^{n \times k}$, let $B = AQ(Q^T AQ)^{-1} Q^T A$

Then,

B is the best rank-k PSD approx to A in $\text{range}(AQ)$

1. Run subpsace iteration to find good Q
2. Build $B = AQ(Q^T AQ)^{-1} Q^T A$
3. Return $(f(B))_k$

Thm: If f is "nice" then $O\left(\frac{n^2}{g_{k \to k+1}} \log \frac{1}{\varepsilon}\right)$ time suffices

Eigengap $= \frac{\lambda_{k+1} - \lambda_k}{\lambda_{k+1}}$

funNystrom Proof Sketch

f is Operator Monotone if

- for $A \preceq B$ we have $f(A) \preceq f(B)$
- implies f is concave (unintuitive)

Proof Sketch

1. f is concave, so eigengaps of $f(A)$ are $\leq$ those of A
2. Subspace iter convergest fast when eigengaps large
3. A would need $O\left(\frac{1}{g_{K\to K+1}} \log\left(\frac{1}{\varepsilon}\right)\right)$ steps of subspace iter

4. A needs $O\left(\frac{1}{g_{K\to K+1}} \log\left(\frac{1}{\varepsilon}\right)\right)$ steps for funNystrom to converge

$= O\left(\frac{n^2}{g_{K\to K+1}} \log\left(\frac{1}{\varepsilon}\right)\right)$ time

Our Contribution

Their proof is specific to subspace iteration

We often use other algos (Krylov, ID, Sketching, ...)

$\boxed{\text{Thms}}$

Let ortho $Q \in \mathbb{R}^{n \times k+p}$ s.t. $\|A - (QQ^TA)_k\|_{F,2,*} \leq (1+\varepsilon)\|A - A_k\|_{F,2,*}$

Then $B = (AQ(Q^TAQ)^{-1}Q^TA)_k$ has $\|A - B\|_{F,2,*} \leq (1+\varepsilon)\|A - A_k\|_{F,2,*}$

And $C = (f(B))_k$ has $\|f(A) - C\|_{F,2,*} \leq (1+\varepsilon)\|f(A) - (f(A))_k\|_{F,2,*}$

Key points: algorithm agnostic, same value of ε!

(some nuances hidden, see paper)

Takeaways

If $\mathcal{A}lgo$ produces ε-good Q in time T,
then funNystrom is ε-good in time $T + O(n^2 k)$

Proofs: boring & short

Ready for implimentation!

Secondary motivation:
If $f(A)$ has quickly decaying eigenvalues,
Then funNystrom(A) preserves "all" structure of $f(A)$ fast!

Future Directions: Beyond operator monotone (log-concave?)