Reduced-Rank Regression with Operator Norm Error

Praneeth Kacham David Woodruff
Carnegie Mellon University

COLT 2021

Problem Statement

- Fast algorithm for

$$\min_{\text{rank-}k\ X} \|AX - B\|_2$$

  $A \in \mathbb{R}^{n\times c}$,$B \in \mathbb{R}^{n\times d}$ and $c, d \gg k$
- $\|M\|_2 = \max_{x:\|x\|_2=1} \|Mx\|_2$
- Finding a good k-dimensional subspace for B inside column space of A
- X is given by $(c + d) \cdot k$ parameters instead of $c \cdot d$
- Our algorithm runs in time $\approx (\text{nnz}(A) + \text{nnz}(B) + c^2)k/\varepsilon^{3/2}$

Why Operator Norm over Frobenius?

- Approximate operator norm solutions maybe better than approximate Frobenius norm solutions
- Consider when the spectrum of B is flat and

$$\varepsilon \sum_{i=k+1}^{d} \sigma_i(B)^2 \approx \sum_{i=1}^{k} \sigma_i(B)^2$$

- Then $X = 0$ is a good approximate solution but isn't very useful
- Standard motivation for preferring Operator Norm LRA

Some Facts and a Lowerbound

- For a matrix B, let $\sigma_1(B) \geq \ldots \geq \sigma_d(B) \geq 0$ be its singular values and $B = \sum_i \sigma_i u_i v_i^{\mathsf{T}}$
- For any k, let $[B]_k = \sum_{i=1}^{k} \sigma_i u_i v_i^{\mathsf{T}}$
- For any matrix X with rank $\leq k$

$$\|B - X\|_2 \geq \|B - [B]_k\|_2 = \sigma_{k+1}(B)$$

$$\|B - X\|_F^2 \geq \|B - [B]_k\|_F^2 = \sum_{i=k+1}^{d} \sigma_i(B)^2$$

- For any matrix X, $\|B - AX\|_2 \geq \|(I - AA^+)B\|_2$
- Both of these imply for any rank k matrix X,

$$\|B - AX\|_2 \geq \max(\sigma_{k+1}(B), \|(I - AA^+)B\|_2)$$

How to solve?

- We use an equivalent condition given by Sou and Rantzer
- Let $\Delta = B^{\mathsf{T}}(I - AA^{+})B = ((I - AA^{+})B)^{\mathsf{T}}((I - AA^{+})B)$
- There is a solution with $\|AX - B\|_2 < \beta$ if and only if

$$\sigma_{k+1}(AA^{+}B(\beta^2 I - \Delta)^{-1/2}) < 1$$

- Has an easy proof!

Main takeaways from Sou and Rantzer

- Surprisingly,

$$\inf_{\text{rank-}k\ X} \|AX - B\|_2 = \max(\sigma_{k+1}(B), \|(I - AA^+)B\|_2)$$

- There is a solution which matches this simple lowerbound!
- Given rank-k X with $\|X - AA^+B(\beta^2 I - \Delta)^{-1/2}\|_2 < 1$, we have

$$\|(AA^+X)(AA^+X)^+B - B\|_2 < \beta$$

  - $\text{rank}((A^+X)(AA^+X)^+B) \leq k$

A slow algorithm

- Let $\beta = (1+\varepsilon)\max(\sigma_{k+1}(B), \|(I - AA^+)B\|_2)$
- Compute SVD of $M := AA^+B(\beta^2 I - \Delta)^{-1/2}$ to obtain $[M]_k$ such that
$$\|M - [M]_k\|_2 \leq \sigma_{k+1}(M) < 1$$
- Obtain $Y = (A^+[M]_k)(AA^+[M]_k)^+B$ with $\|AY - B\|_2 < \beta$
- Issues:
  - Very slow as we have to compute Δ, a negative square root and an SVD
  - Cannot make use of sparsity of the matrices A and B

Our Idea

- For feasible β, $\sigma_{k+1}(M) < 1$
- Previous idea was to just compute a rank k approximation given by SVD that satisfies

$$\|[M]_k - M\|_2 \leq \sigma_{k+1}(M) < 1$$

- We show that even if $\|X - M\|_2 < 1 + c\varepsilon$ for some small c

$$\|(AA^+X)(AA^+X)^+B - B\|_2 < (1+\varepsilon)\beta$$

- We can use fast algorithms for computing approximate low rank approximations

Block Krylov Iteration

Theorem

Given any vectors x and y, if we can compute Mx and $M^{\mathsf{T}}y$ in time T, then in time

$$\approx Tqk$$

for $q \approx 1/\sqrt{\varepsilon}$, we can compute rank-$k$ matrix M' with

$$\|M - M'\|_2 \leq (1+\varepsilon)\sigma_{k+1}(M)$$

Krylov Subspace $K = [MG, (MM^{\mathsf{T}})MG, \ldots, (MM^{\mathsf{T}})^{O(q)}MG]$

Issues

- The algorithm needs exact matrix-vector products with

$$M = AA^{+}B(\beta^{2}I - \Delta)^{-1/2}$$

- Very expensive to compute :(

Key Technical Contribution

- Runs even with approximate matrix-vector products!

Theorem

Given any vectors x and y, if we can compute x' and y' with

$$\|x' - Mx\|_2 \leq \alpha \|M\|_2 \|x\|_2$$
$$\|y' - M^{\mathsf{T}} y\|_2 \leq \alpha \|M^{\mathsf{T}}\|_2 \|y\|_2$$

in time $T(\alpha)$, then in time

$$\approx T\left(\frac{\varepsilon}{\kappa^q \mathrm{poly}(k)}\right) qk$$

for $q \approx 1/\sqrt{\varepsilon}$, we can compute rank-$k$ matrix M' with

$$\|M - M'\|_2 \leq (1 + \varepsilon)\sigma_{k+1}(M)$$

- $\kappa = \sigma_1(M)/\sigma_{k+1}(M)$

Techniques

- First analysis of Block Krylov Iteration with approximate products
- Analysis follows along the lines of Musco and Musco
- We use properties of Gaussian Matrices to conclude that even approximate Krylov subspace K' spans a good approximation
- We then show that approximate products are enough to compute a good approximation inside K'

Wrapping Up

- We only need a way to compute approximate matrix-vector products with $M = AA^{+}B(\beta^2 I - \Delta)^{-1/2}$
- We replace $(\beta^2 I - \Delta)^{-1/2}$ with $r(\Delta)$ where $r(x)$ is a polynomial

$$r(\Delta) = r_0 I + r_1 \Delta + r_2 \Delta^2 + \cdots + r_t \Delta^t$$

- $\Delta y = B^{\mathsf{T}} B y - B^{\mathsf{T}} A A^{+} B y$
- For any y, the vector By can be computed in $\mathtt{nnz}(B)$ time
- Only have to approximate $AA^{+} z$ for arbitrary z
- High Precision Regression to obtain α approximations in time proportional to $\log(1/\alpha)$

Overall ε dependence

$$\underbrace{\frac{1}{\sqrt{\varepsilon}}}_{\text{Block Krylov Iterations}} \times \underbrace{\frac{1}{\sqrt{\varepsilon}}}_{\text{Degree of } r(x)} \times \underbrace{\frac{1}{\sqrt{\varepsilon}}}_{\text{Approximating } \Delta x}$$

Overall ε dependence

$$\underbrace{\frac{1}{\sqrt{\varepsilon}}}_{\text{Block Krylov Iterations}} \times \underbrace{\frac{1}{\sqrt{\varepsilon}}}_{\text{Degree of } r(x)} \times \underbrace{\frac{1}{\sqrt{\varepsilon}}}_{\text{Approximating } \Delta x}$$

- Optimizing ε dependence is an interesting open problem