Optimal Deterministic Coresets for Ridge Regression

Praneeth Kacham David P. Woodruff

Computer Science Department
Carnegie Mellon University

AISTATS 2020

Introduction

- Ridge Regression

$$\min_X \|AX - B\|_F^2 + \lambda \|X\|_F^2$$

Introduction

- Ridge Regression

$$\min_X \|AX - B\|_F^2 + \lambda\|X\|_F^2$$

- Coresets

$$\min_X \|SAX - SB\|_F^2 + \lambda\|X\|_F^2$$

Introduction

- Ridge Regression

$$\min_X \|AX - B\|_F^2 + \lambda \|X\|_F^2$$

- Coresets

$$\min_X \|SAX - SB\|_F^2 + \lambda \|X\|_F^2$$

- Why Deterministic?
    - Composability

Introduction

- Ridge Regression

$$\min_X \|AX - B\|_F^2 + \lambda \|X\|_F^2$$

- Coresets

$$\min_X \|SAX - SB\|_F^2 + \lambda \|X\|_F^2$$

- Why Deterministic?
  - Composability
- Optimality
  - Matching Lower Bound

Statistical Dimension

Given a matrix A and $\lambda \geq 0$, we define

$$\mathrm{sd}_\lambda(A) = \sum_{i=1}^{\mathrm{rank}(A)} \frac{1}{1 + \lambda/\sigma_i^2}$$

Statistical Dimension

Given a matrix A and $\lambda \geq 0$, we define

$$\mathrm{sd}_\lambda(A) = \sum_{i=1}^{\mathrm{rank}(A)} \frac{1}{1 + \lambda/\sigma_i^2}$$

- Clearly, $\mathrm{sd}_\lambda(A) \leq d_A$
- Captures the fact that importance of A decreases as λ increases
- Would like coreset sizes to depend on $\mathrm{sd}_\lambda(A)$ instead of d_A

Subspace Embeddings

- Maps high-dimensional vectors to low-dimensional vectors

$$y \to Sy$$
$$\|Sy\|^2 \in (1 \pm \varepsilon)\|y\|^2$$

Subspace Embeddings

- Maps high-dimensional vectors to low-dimensional vectors

$$y \to Sy$$
$$\|Sy\|^2 \in (1 \pm \varepsilon)\|y\|^2$$

- Preserves lengths of all vectors in the subspace

Subspace Embeddings

If S is ε subspace embedding for column span of $[A, B]$ then

$$\|SAX - SB\|_F^2 \in (1 \pm \varepsilon)\|AX - B\|_F^2$$

Can obtain $1 + O(\varepsilon)$ approximation

Subspace Embeddings - Upshot

Theorem

Given a $\sqrt{\varepsilon/4}$ subspace embedding S for column span of $[A, B]$, solution for

$$\min_X \|SAX - SB\|_F^2$$

is a $(1+\varepsilon)$ approximation for

$$\min_X \|AX - B\|_F^2.$$

Subspace Embeddings - Various Techniques

There are mainly three different types of subspace embeddings

- Randomized and Oblivious - Gaussians, SRHT, CountSketch, etc.
- Randomized and Non-Oblivious - Leverage Score Sampling, Length Squared Sampling etc.
- Deterministic - Spectral Sparsification (BSS)

Approximate Matrix Multiplication

Theorem

Given matrices A, B and $k > 0$, we can deterministically obtain a matrix S that selects and scales $O(k/\varepsilon^2)$ rows of A, B such that

$$\|A^T S^T S B - A^T B\|_2 \leq \varepsilon \sqrt{\|A\|_2^2 + \frac{\|A\|_F^2}{k}} \sqrt{\|B\|_2^2 + \frac{\|B\|_F^2}{k}}$$

We can use this result to get smaller coresets for Ridge Regression.

Construction of Coreset

$$\|AX - B\|_F^2 + \lambda\|X\|_F^2 = \| \underbrace{\begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix}}_{\widehat{A}} X - \underbrace{\begin{bmatrix} B \\ 0 \end{bmatrix}}_{\widehat{B}} \|_F^2$$

- We need $\sqrt{\varepsilon/4}$ subspace embeddings for column span of $[\widehat{A}\ \widehat{B}]$.
- Orthogonalizing $[\widehat{A}\ \widehat{B}]$, we obtain

$$\mathcal{U} = \begin{bmatrix} \mathcal{U}_1 \\ \mathcal{U}_2 \end{bmatrix} = \begin{bmatrix} U_1 & U_1' \\ U_2 & U_2' \end{bmatrix}$$

where $\mathcal{U}_1$ is a basis for $[A\ B]$ and $\begin{bmatrix} U_1 \\ U_2 \end{bmatrix}$ is an *orthonormal* basis for $\widehat{A}$.

Construction of Coreset

- We can show that $\|U_1\|_F^2 = \mathrm{sd}_\lambda(A)$ and therefore
$$\|\mathcal{U}_1\|_F^2 \leq \mathrm{sd}_\lambda(A) + d_B$$
- Clearly, $\|\mathcal{U}_1\|_2^2 \leq 1$

Using the AMM theorem, we can determinisitically obtain a matrix S with $O((\mathrm{sd}_\lambda(A) + d_B)/\varepsilon)$ rows such that

$$\begin{aligned}\|\mathcal{U}_1^T S^T S \mathcal{U}_1 - \mathcal{U}_1^T \mathcal{U}_1\|_2 &\leq \sqrt{\varepsilon/64}\left(\|\mathcal{U}_1\|_2^2 + \frac{\|\mathcal{U}_1\|_F^2}{(\mathrm{sd}_\lambda(A) + d_B)}\right)^2 \\ &\leq \sqrt{\varepsilon/64}(4) = \sqrt{\varepsilon/4}\end{aligned}$$

Construction of Coreset

Now we can show that $\mathcal{S} = \begin{bmatrix} S & 0 \\ 0 & I \end{bmatrix}$ is a $\sqrt{\varepsilon/4}$ Subspace Embedding.

$$\begin{aligned} \|\mathcal{U}^T\mathcal{S}^T\mathcal{S}\mathcal{U} - \mathcal{U}^T\mathcal{U}\|_2 &= \|(\mathcal{U}_1^T S^T S\mathcal{U}_1 + \mathcal{U}_2^T\mathcal{U}_2) - (\mathcal{U}_1^T\mathcal{U}_1 + \mathcal{U}_2^T\mathcal{U}_2)\|_2 \\ &= \|\mathcal{U}_1^T S^T S\mathcal{U}_1 - \mathcal{U}_1^T\mathcal{U}_1\|_2 \\ &\leq \sqrt{\varepsilon/4}. \end{aligned}$$

As $\mathcal{U}$ is orthonormal basis for $[\widehat{A}\ \widehat{B}]$, we obtain that $\mathcal{S}$ is a $\sqrt{\varepsilon/4}$ subspace embedding

Back to Ridge Regression

We now have that solution to $\min_X \|\mathcal{S}\widehat{A}X - \mathcal{S}\widehat{B}\|_F^2$ is a $(1+\varepsilon)$ approximation to ridge regression. But this problem is equivalent to

$$\min_X \|SAX - SB\|_F^2 + \lambda\|X\|_F^2$$

So we have $O((\mathrm{sd}_\lambda(A) + d_B)/\varepsilon)$ size *deterministic* coresets for Ridge Regression.

Communication Model

- Rows of A and B are partitioned among s servers
- There is a central server through which other servers communicate
- Want to solve Ridge Regression on the union of matrices at all the servers
- t is the maximum number of non-zero entries in a row of $[A, B]$

Results

- We show that there is a Communication Protocol which computes a $(1+\varepsilon)$ approximation by using

$$O\left(\frac{s \cdot t \cdot (\mathrm{sd}_{\lambda/s}(A) + d_B)}{\varepsilon}\right) \text{ words}$$

- Protocol computes a coreset for λ/s at each server and send to central server
- We show that union of these coresets give a $1+\varepsilon$ approximation

Lower Bounds

Theorem

For all ε such that $1 \leq 1/100\varepsilon \leq 1/d_A$ and $\lambda \leq 1/4\varepsilon$, there are matrices A, B with $d_B = O(sd_\lambda(A))$ such that any deterministic coreset that selects and scales rows of A, B must select $\Omega(sd_\lambda(A)/\varepsilon)$ rows.

- Coreset sizes given by our algorithm matches the lower bound
- Optimal upto constant factor

Proof Sketch

$$A = \begin{bmatrix} \underbrace{\begin{matrix} 1 \\ \vdots \\ 1 \end{matrix}}_{1/100\varepsilon} & 0 & \cdots & 0 \\ 0 & \underbrace{\begin{matrix} 1 \\ \vdots \\ 1 \end{matrix}}_{1/100\varepsilon} & \cdots & 0 \\ \vdots & & \ddots & \end{bmatrix} \qquad B = \begin{bmatrix} e_1^T \\ \vdots \\ e_{1/100\varepsilon}^T \\ e_1^T \\ \vdots \\ e_{1/100\varepsilon}^T \\ \vdots \end{bmatrix}$$

Proof Sketch

- Solving d different instances of

$$\min_x \|1x^T - \begin{bmatrix} e_1^T \\ \vdots \\ e_{1/100\varepsilon}^T \end{bmatrix}\|_F^2 + \lambda\|x\|^2$$

- Can show that optimum for each instance is $1/100\varepsilon - 1/(1+100\lambda\varepsilon)$
- Cost of any solution computed only on k rows is

$$\geq 1/100\varepsilon - \frac{k}{\lambda + 1/100\varepsilon}$$

- This is $1+2\varepsilon$ approximation only if

$$k \geq 1/400\varepsilon$$

Proof Sketch

- To get a $1 + \varepsilon$ approximation, we need to solve at least $d/2$ sub problems upto $1 + 2\varepsilon$ approximation
- Implies that any coreset needs to select at least $d/800\varepsilon$ rows
- Shows a lower bound of $\Omega(\mathrm{sd}_\lambda(A)/\varepsilon)$ lower bound.

Proof Sketch

- To get a $1+\varepsilon$ approximation, we need to solve at least $d/2$ sub problems upto $1+2\varepsilon$ approximation
- Implies that any coreset needs to select at least $d/800\varepsilon$ rows
- Shows a lower bound of $\Omega(\mathrm{sd}_\lambda(A)/\varepsilon)$ lower bound.

Thank You!