Matrix Sketching as a Tool in Numerical Algebra

Chengtao Li

Massachusetts Institute of Technology

ctli@mit.edu

December 1, 2015

Overview

Linear Regression

- Have: $A \in \mathbb{R}^{n \times d}$, $b \in R^n$.
- Want: Find $x \in \mathbb{R}^d$ s.t. Ax and b as close as possible

Linear Regression

- Have: $A \in \mathbb{R}^{n \times d}$, $b \in R^n$.
- Want: Find $x \in \mathbb{R}^d$ s.t. Ax and b as close as possible
- Objective: $\min_x \|Ax - b\|_p$

Linear Regression cont.

- For $p = 2$, minimize Euclidean distances btw Ax and b.
- Solution: $(A^\top A)x^* = A^\top b$.

Linear Regression cont.

- For $p = 2$, minimize Euclidean distances btw Ax and b.
- Solution: $(A^\top A)x^* = A^\top b$.
- *Problem*: Overly constrained, Massive datasets at least $O(nd^2)$

To Circumvent

- Use *Sketching Techniques* to improve upon the time complexities.
- Relaxation: find x s.t. $\|Ax - b\|_p \leq (1+\varepsilon)\|Ax^* - b\|_p$

Other Examples

- Matrix Low-rank Approximation
- Matrix Product Approximation
- Kernel Methods
- Social Networks

Matrix Sketching Framework

1. Specify $r \ll n$
2. Sample a random matrix $S \in \mathbb{R}^{r \times n}$
3. Solve the problem with the sketched matrices.

Matrix Sketching Framework

1. Specify $r \ll n$
2. Sample a random matrix $S \in \mathbb{R}^{r \times n}$
3. Solve the problem with the sketched matrices.

In linear regression, the sketched regression problem becomes

$$\min_x \|(SA)x - (Sb)\|_2 \tag{1.1}$$

Questions to ask

1. How to choose $r \ll n$?

Questions to ask

1. How to choose $r \ll n$?
2. From what distribution should we sample S?

Questions to ask

1. How to choose $r \ll n$?
2. From what distribution should we sample S?
3. Approximation bounds?

Questions to ask

1. How to choose $r \ll n$?
2. From what distribution should we sample S?
3. Approximation bounds?
4. Efficiency?

Questions to ask

1. How to choose $r \ll n$?
2. From what distribution should we sample S?
3. Approximation bounds?
4. Efficiency?
5. (Optional) Optimality?

Different forms of Sketching Matrix S:

- Projection Matrix (JL-style)
- Selection Matrix (Our focus)

Different forms of Sketching Matrix S:

- Projection Matrix (JL-style)
- Selection Matrix (Our focus)
- Why Selection?
  1. Interpretability
  2. Preserves Sparsity/Structure

The Problem

- Given a matrix A, select a small number of columns from A so that the selected columns serve as a good "snapshot" or "summarization" of A.

The Problem

- Given a matrix A, select a small number of columns from A so that the selected columns serve as a good "snapshot" or "summarization" of A.
- Writing in math language: Let $A \in \mathbb{R}^{n \times m}$ be a huge matrix, sample $|C| = c$ columns of A so as to approximate A as $A_C A_C^{\dagger} A$.

The Problem

- Given a matrix A, select a small number of columns from A so that the selected columns serve as a good "snapshot" or "summarization" of A.
- Writing in math language: Let $A \in \mathbb{R}^{n \times m}$ be a huge matrix, sample $|C| = c$ columns of A so as to approximate A as $A_C A_C^{\dagger} A$.
- Intuition

Objective

The objective function is to choose columns $C \subseteq [m]$ to minimize matrix norm:

$$\min_{C \subseteq [m]} \|A - A_C A_C^{\dagger} A\|_{\xi} \tag{2.1}$$

where $\xi \in \{2, F, *\}$.

Types of Error Bounds:

- Additive Bounds:

$$\|A - A_C A_C^{\dagger} A\|_{\xi} \leq \|A - A_k\|_{\xi} + \varepsilon \|A\|_{\xi} \tag{2.2}$$

Types of Error Bounds:

- Additive Bounds:

$$\|A - A_C A_C^{\dagger} A\|_{\xi} \leq \|A - A_k\|_{\xi} + \varepsilon \|A\|_{\xi} \qquad (2.2)$$

- Relative Bounds:

$$\|A - A_C A_C^{\dagger} A\|_{\xi} \leq (1 + \varepsilon)\|A - A_k\|_{\xi} \qquad (2.3)$$

Types of Error Bounds:

- Additive Bounds:

$$\|A - A_C A_C^{\dagger} A\|_{\xi} \leq \|A - A_k\|_{\xi} + \varepsilon \|A\|_{\xi} \tag{2.2}$$

- Relative Bounds:

$$\|A - A_C A_C^{\dagger} A\|_{\xi} \leq (1 + \varepsilon) \|A - A_k\|_{\xi} \tag{2.3}$$

Required to hold either *in expectation* or *with high probability*.

Optimal Column Selection (Guruswami & Sinop'12)

- Lower bounds: there exists a matrix M for which the best error achieved by a low rank matrix, whose columns are restricted to belong to the span of $r \geq k/\varepsilon$ columns of M, is at least $1 + \varepsilon - o(1)$ times the best rank-k approximation

Optimal Column Selection (Guruswami & Sinop'12)

- Lower bounds: there exists a matrix M for which the best error achieved by a low rank matrix, whose columns are restricted to belong to the span of $r \geq k/\varepsilon$ columns of M, is at least $1 + \varepsilon - o(1)$ times the best rank-k approximation
- Optimal Result: $k/\varepsilon + k - 1$ columns is sufficient for achieving $(1 + \varepsilon)$ bound (match the lower bound up to lower order terms).

Optimal Column-based Reconstruction cont.

Volume Sampling on A

- Ground set $G = [n]$
- Select $S \subseteq G$ where $|S| = k$:

$$\Pr(S \subseteq G) \propto det(A_S A_S^\top) \tag{2.4}$$

Optimal Column-based Reconstruction cont.

Algorithm:

1. Sample S where $|S| = c$ according to volume sampling
2. Approximate A by $A_S A_S^\dagger A$

Optimal Column-based Reconstruction cont.

Algorithm:

1. Sample S where $|S| = c$ according to volume sampling
2. Approximate A by $A_S A_S^{\dagger} A$

Optimality:

- For $c \geq k$, we have

$$E_S[\|A - A_S A_S^{\dagger} A\|_F^2] \leq \frac{c+1}{c+1-k} \|A - A_k\|_F^2 \tag{2.5}$$

Optimal Column-based Reconstruction Proof

Optimal Column-based Reconstruction cont.

Practical Issues:

Optimal Column-based Reconstruction cont.

Practical Issues:

- Running Time: $O(cm^2n)$

Optimal Column-based Reconstruction cont.

Practical Issues:

- Running Time: $O(cm^2n)$;
- Derandomization

Optimal Column-based Reconstruction cont.

Practical Issues:

- Running Time: $O(cm^2n)$;
- Derandomization

Another notable method:
Near-Optimal Column Selection (Boutsidis et.al.), selects $2k\varepsilon^{-1}(1+o(1))$ columns for reconstruction, runs in $O(mnk+nk^2)/\varepsilon^{2/3}$

The Problem

- Given a matrix A, select a small number of columns and rows from A so that selected columns and rows serve as a good "snapshot" or "summarization" of A.
- Writing in math language: Let $A \in \mathbb{R}^{n \times m}$ sample $|C| = c$ columns and $|R| = r$ rows and calculate U so that we approximate A with $A_C U A^R$.
- Intuition

Objective

The objective is to choose columns $C \subseteq [m]$, $R \subseteq [n]$ and $U \in \mathbb{R}^{c \times r}$ so as to minimize matrix norm:

$$\min_{C,U,R} \|A - A_C U A^R\|_\xi \tag{3.1}$$

where $\xi \in \{2, F, *\}$.

Observation: Fixing C and R, and for $\xi = F$, have

$$\arg\min_{U} \|A - A_C U A^R\| = A_C^{\dagger} A (A^R)^{\dagger} \tag{3.2}$$

Thus, usually the objective becomes:

$$\min_{C,R} \|A - A_C A_C^{\dagger} A (A^R)^{\dagger} A^R\|_F \tag{3.3}$$

CUR with Adaptive Sampling (Wang & Zhang'13)

Adaptive Sampling:
Given matrix $A \in \mathbb{R}^{n \times m}$ and let A_C be already selected columns of A. Define the residual $B = A - A_C A_C^{\dagger} A$. The adaptive sampling is to sample from distribution defined as

$$p_i = \|b_i\|_2^2 / \|B\|_F^2 \tag{3.4}$$

CUR with Adaptive Sampling (Wang & Zhang'13) cont.

Theorem 5 (The Adaptive Sampling Algorithm) *Given a matrix* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *and a matrix* $\mathbf{C} \in \mathbb{R}^{m\times c}$ *such that* $\mathrm{rank}(\mathbf{C}) = \mathrm{rank}(\mathbf{C}\mathbf{C}^\dagger\mathbf{A}) = \rho$ $(\rho \le c \le n)$. *We let* $\mathbf{R}_1 \in \mathbb{R}^{r_1\times n}$ *consist of* r_1 *rows of* $\mathbf{A}$, *and define the residual* $\mathbf{B} = \mathbf{A} - \mathbf{A}\mathbf{R}_1^\dagger\mathbf{R}_1$. *Additionally, for* $i = 1, \cdots, m$, *we define*

$$p_i = \|\mathbf{b}^{(i)}\|_2^2 / \|\mathbf{B}\|_F^2.$$

We further sample r_2 *rows i.i.d. from* $\mathbf{A}$, *in each trial of which the i-th row is chosen with probability* p_i. *Let* $\mathbf{R}_2 \in \mathbb{R}^{r_2\times n}$ *contain the* r_2 *sampled rows and let* $\mathbf{R} = [\mathbf{R}_1^T, \mathbf{R}_2^T]^T \in \mathbb{R}^{(r_1+r_2)\times n}$. *Then we have*

$$\mathbb{E}\|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger\mathbf{A}\mathbf{R}^\dagger\mathbf{R}\|_F^2 \le \|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger\mathbf{A}\|_F^2 + \frac{\rho}{r_2}\|\mathbf{A} - \mathbf{A}\mathbf{R}_1^\dagger\mathbf{R}_1\|_F^2,$$

where the expectation is taken w.r.t. $\mathbf{R}_2$.

CUR with Adaptive Sampling (Wang & Zhang'13) cont.

A general framework

- Select $c \geq C(k, \varepsilon)$ columns (A_C) of A and $r_1 = c$ rows (A^{R_1}) with some CSSP method
- Select $r_2 = c/\varepsilon$ additional rows (A^{R_2}) with adaptive sampling with respect to $A - A - A(A^{R_1})^\dagger A^{R_1}$
- Let $R = R_1 \cup R_2$, construct $U = (A_C)^\dagger A (A_R)^\dagger$

CUR with Adaptive Sampling (Wang & Zhang'13) cont.

Plug in and Play!

CUR with Adaptive Sampling (Wang & Zhang'13) cont.

Plug in and Play!

- Plug in near-optimal CSSP
  - $c = \frac{2k}{\varepsilon}(1 + o(1))$ columns
  - $r = \frac{c}{\varepsilon}(1 + \varepsilon)$ rows
  - Running time: $O((m+n)k^3\varepsilon^{-2/3} + mk^2\varepsilon^{-2} + nk^2\varepsilon^{-4}) + T_M(mnk/\varepsilon)$.

CUR with Adaptive Sampling (Wang & Zhang'13) cont.

Plug in and Play!

- Plug in near-optimal CSSP
  - $c = \frac{2k}{\varepsilon}(1 + o(1))$ columns
  - $r = \frac{c}{\varepsilon}(1 + \varepsilon)$ rows
  - Running time: $O((m+n)k^3\varepsilon^{-2/3} + mk^2\varepsilon^{-2} + nk^2\varepsilon^{-4}) + T_M(mnk/\varepsilon)$.
- Plug in optimal CSSP:
  - $c = k\varepsilon^{-1}(1 + o(1))$ columns
  - $r = c\varepsilon^{-1}(1 + \varepsilon)$ rows

Problem

Let $A \in \mathbb{R}^{n\times m}$, we want to sample a set of re-weighted columns of A (with selection rescaled matrix S such that the spectrum of AS is similar to that of A. Specifically, we want

$$(1-\varepsilon)AA^\top \preceq AS(AS)^\top \preceq (1+\varepsilon)AA^\top \tag{4.1}$$

Sampling with Statistical Leverage Score

- Statistical leverage score for the i-th column A_i is

$$p_i = A_i^\top (AA^\top)^\dagger A_i \tag{4.2}$$

Sampling with Statistical Leverage Score

- Statistical leverage score for the i-th column A_i is

$$p_i = A_i^\top (AA^\top)^\dagger A_i \tag{4.2}$$

- Theorem (Spielman & Srivastava): If we independently sample $O(n\varepsilon^{-2}\log n)$ columns of A with probability proportional to p_i and rescale with $1/p_i$, with probability $1 - 1/n$ we have

$$(1-\varepsilon)AA^\top \preceq AS(AS)^\top \preceq (1+\varepsilon)AA^\top \tag{4.3}$$

Sampling with Statistical Leverage Score

- Statistical leverage score for the i-th column A_i is

$$p_i = A_i^\top (AA^\top)^\dagger A_i \tag{4.2}$$

- Theorem (Spielman & Srivastava): If we independently sample $O(n\varepsilon^{-2}\log n)$ columns of A with probability proportional to p_i and rescale with $1/p_i$, with probability $1 - 1/n$ we have

$$(1-\varepsilon)AA^\top \preceq AS(AS)^\top \preceq (1+\varepsilon)AA^\top \tag{4.3}$$

- Running time: $\tilde{O}(m(\log n)\varepsilon^{-2})$.

Remarks: Optimality

- Optimal CSSP: $O(\frac{k}{\varepsilon})$ columns
- Optimal CUR: $O(\frac{k}{\varepsilon})$ columns and $O(\frac{k}{\varepsilon})$ rows
- Optimal Spectral Sparsification: $O(\frac{k}{\varepsilon^2})$ columns

Thanks! Questions?