Sketching as a Tool for Numerical Linear Algebra

David P. Woodruff

presented by Sepehr Assadi

$O(n)$ Big Data Reading Group
University of Pennsylvania

February, 2015
Goal

- New survey by David Woodruff:
  - Sketching as a Tool for Numerical Linear Algebra

- Topics:
  - Subspace Embeddings
  - Least Squares Regression
  - Least Absolute Deviation Regression
  - Low Rank Approximation
  - Graph Sparsification
  - Sketching Lower Bounds
Goal

- New survey by David Woodruff:
  - Sketching as a Tool for Numerical Linear Algebra

- Topics:
  - Subspace Embeddings
  - Least Squares Regression
  - Least Absolute Deviation Regression
  - Low Rank Approximation
  - Graph Sparsification
  - Sketching Lower Bounds
Introduction

- You have “Big” data!
  - Computationally expensive to deal with
  - Excessive storage requirement
  - Hard to communicate
  - ...

Introduction

- You have “Big” data!
  - Computationally expensive to deal with
  - Excessive storage requirement
  - Hard to communicate
  - ...

- Summarize your data
  - Sampling
    - A representative subset of the data
  - Sketching
    - An aggregate summary of the whole data
Model

- **Input:**
  - matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$
  - vector $\mathbf{b} \in \mathbb{R}^n$.

- **Output:** function $F(\mathbf{A}, \mathbf{b}, \ldots)$
  - e.g. least square regression

- **Different goals:**
  - Faster algorithms
  - Streaming
  - Distributed
Linear Sketching

- **Input:**
  - matrix $A \in \mathbb{R}^{n \times d}$

- Let $r \ll n$ and $S \in \mathbb{R}^{r \times n}$ be a random matrix

- Let $S \cdot A$ be the sketch

- Compute $F(S \cdot A)$ instead of $F(A)$
Linear Sketching (cont.)

- **Pros:**
  - Compute on a $r \times d$ matrix instead of $n \times d$
  - Smaller representation and faster computation
  - **Linearity:**
    - $S \cdot (A + B) = S \cdot A + S \cdot B$
    - We can *compose* linear sketches!

- **Cons:**
  - $F(S \cdot A)$ is an approximation of $F(A)$
Least Square Regression ($\ell_2$-regression)

- **Input:**
  - matrix $A \in \mathbb{R}^{n \times d}$ (full column rank)
  - vector $b \in \mathbb{R}^{n}$

- **Output** $x^* \in \mathbb{R}^{d}$:

  $$x^* = \arg \min_x \|Ax - b\|_2$$

- **Closed form solution:**

  $$x^* = (A^T A)^{-1} A^T b$$

- $\Theta(nd^2)$-time algorithm using naive matrix multiplication
Approximate $\ell_2$-regression

- **Input:**
  - matrix $A \in \mathbb{R}^{n \times d}$ (full column rank)
  - vector $b \in \mathbb{R}^n$
  - parameter $0 < \varepsilon < 1$

- **Output** $\hat{x} \in \mathbb{R}^d$:

\[
\|A\hat{x} - b\|_2 \leq (1 + \varepsilon) \arg\min_x \|Ax - b\|_2
\]
Approximate $\ell_2$-regression (cont.)

- A sketching algorithm:
  - Sample a random matrix $S \in \mathbb{R}^{r \times n}$
  - Compute $S \cdot A$ and $S \cdot b$
  - Output $\hat{x} = \arg\min_x \| (SA)x - (Sb) \|_2$

- Which randomized family of matrices $S$ and what value of $r$?
Approximate $\ell_2$-regression (cont.)

- An introductory construction:
  - Let $r = \Theta(d/\varepsilon^2)$
  - Let $S \in \mathbb{R}^{r \times n}$ be a matrix of i.i.d normal random variables with mean zero and variance $1/r$

Proof Sketch.
On the board
Approximate $\ell_2$-regression (cont.)

- Problems:
  - Computing $S \cdot A$ takes $\Theta(nrd)$ time
  - Constructing $S$ requires $\Theta(nr)$ space

- Different constructions for $S$:
  - Fast Johnson-Lindenstrauss transforms:
    $O(nd \log d) + \text{poly}(d/\varepsilon)$ time [Sarlos, FOCS ’06]
  - Optimal $O(\text{nnz}(A)) + \text{poly}(d/\varepsilon)$ time algorithm [Clarkson, Woodruff, STOC ’13]
  - Random sign matrices with $\Theta(d)$-wise independent entries:
    $O(d^2/\varepsilon \log (nd))$-space streaming algorithm [Clarkson, Woodruff, STOC ’09]
Subspace Embedding

**Definition (ℓ₂-subspace embedding)**

A \((1 \pm \varepsilon)\) ℓ₂-subspace embedding for a matrix \(A \in \mathbb{R}^{n \times d}\) is a matrix \(S\) for which for all \(x \in \mathbb{R}^{n}\)

\[
\|SAx\|_2^2 = (1 \pm \varepsilon) \|Ax\|_2^2
\]

- Actually subspace embedding for column space of \(A\)
- Oblivious ℓ₂-subspace embedding
  - The distribution from which \(S\) is chosen is oblivious to \(A\)
- One very common tool for (oblivious) ℓ₂-subspace embedding is Johnson-Lindenstrauss transform (JLT)
Johnson-Lindenstrauss transform

**Definition (JLT(\(\epsilon, \delta, f\)))**

A random matrix \(S \in \mathbb{R}^{r \times d}\) forms a JLT(\(\epsilon, \delta, f\)), if with probability at least \(1 - \delta\), for any \(f\)-element subset \(V \subseteq \mathbb{R}^n\), it holds that:

\[
\forall \ v, v' \in V \quad |\langle Sv, Sv' \rangle - \langle v, v' \rangle| \leq \epsilon \|v\|_2 \|v'\|_2
\]
Johnson-Lindenstrauss transform

**Definition (JLT(ε, δ, f))**

A random matrix $S \in \mathbb{R}^{r \times d}$ forms a JLT(ε, δ, f), if with probability at least $1 - δ$, for any $f$-element subset $V \subseteq \mathbb{R}^n$, it holds that:

$$\forall v, v' \in V \quad |\langle Sv, Sv' \rangle - \langle v, v' \rangle| \leq \varepsilon \|v\|_2 \|v'\|_2$$

**Usual statement (i.e. original Johnson-Lindenstrauss Lemma)**

**Lemma (JLL)**

Given $N$ points $q_1, \ldots, q_N \in \mathbb{R}^n$, there exists a matrix $S \in \mathbb{R}^{t \times n}$ (linear map) for $t = \Theta(\log N/\varepsilon^2)$ such that with high probability, simultaneously for all pairs $q_i$ and $q_j$,

$$\|S(q_i - q_j)\|_2 = (1 \pm \varepsilon) \|(q_i - q_j)\|_2$$
Johnson-Lindenstrauss transform (cont.)

- A simple construction of JLT\((\varepsilon, \delta, f)\):

**Theorem**

\[
\text{Let } 0 < \varepsilon, \delta < 1 \text{ and } S = \frac{1}{\sqrt{r}} \mathbf{R} \in \mathbb{R}^{r \times n} \text{ where the entries } R_{i,j} \text{ are independent standard normal random variables. Assuming } \]
\[
r = \Omega(\varepsilon^{-2} \log (f/\delta)) \text{ then } S \text{ is a JLT}(\varepsilon, \delta, f).
\]

- Other constructions:
  - Random sign matrices
    
    [Achlioptas, ’03], [Clarkson, Woodruff, STOC ’09]
  - Random sparse matrices
    
    [Dasgupta, Kumar, Sarlos, STOC ’10], [Kane, Nelson, J. ACM ’14]
  - Fast Johnson-Lindenstrauss transforms
    
    [Ailon, Chazelle, STOC ’06]
JLT results in $\ell_2$-subspace embedding

**Claim**

$S = JLT(\varepsilon, \delta, f)$ is an oblivious $\ell_2$-subspace embedding for $A \in \mathbb{R}^{n \times d}$

**Challenge:**

- JLT $(\varepsilon, \delta, f)$ provides a guarantee for a **single finite set** in $\mathbb{R}^n$
- $\ell_2$-subspace embedding requires the guarantee for an **infinite set**, i.e. the column space of $A$
JLT results in $\ell_2$-subspace embedding (cont.)

- Let $S$ be the unit sphere in column space of $A$

$$S = \{ y \in \mathbb{R}^n \mid y = Ax \text{ for some } x \in \mathbb{R}^d \text{ and } \|y\|_2 = 1 \}$$

- We seek a finite subset $\mathcal{N} \subseteq S$ so that if

$$\forall \ w, w' \in \mathcal{N} \quad \langle Sw, Sw' \rangle = \langle w, w' \rangle \pm \varepsilon$$

then

$$\forall \ y \in S \quad \|Sy\| = (1 \pm \varepsilon) \|y\|$$
Lemma ($\frac{1}{2}$-net for $S$)

**Suffices to choose any $N$ such that**

$$\forall y \in S \exists w \in N \text{ s.t. } \|y - w\|_2 \leq 1/2$$

**Proof.**

1. **Decompose** $y$:

$$y = y^{(0)} + y^{(1)} + y^{(2)} + \ldots$$

where $\|y^{(i)}\|_2 \leq \frac{1}{2^i}$ and $\frac{y^i}{\|y^{(i)}\|} \in N$

2. $\|Sy\|_2^2 = \|S(y^{(0)} + y^{(1)} + y^{(2)} + \ldots)\| = 1 \pm O(\varepsilon)$
There exists a \( \frac{1}{2} \)-net \( \mathcal{N} \) of \( S \) for which \( |\mathcal{N}| \leq 5^d \)

**Proof.**

1. Find a set \( \mathcal{N}' \) of maximal number of points in \( \mathbb{R}^d \) so that no two points are within \( 1/2 \) distance from each other.
2. Let \( U \) be the orthonormal matrix of column space of \( A \).
3. \( \mathcal{N} = \{ y \in \mathbb{R}^n \mid y = Ux \text{ for some } x \in \mathcal{N}' \text{ and } \|y\|_2 = 1 \} \)
Subspace Embedding via JLT

**Theorem**

Let $0 < \varepsilon, \delta < 1$ and $S = JLT(\varepsilon, \delta, 5^d)$. For any fixed matrix $A \in \mathbb{R}^{n \times d}$, with probability $1 - \delta$, $S$ is a $(1 \pm \varepsilon)$ $\ell_2$-subspace embedding for $A$, i.e.

$$\forall x \in \mathbb{R}^d, \|SAx\|_2 = (1 \pm \varepsilon) \|Ax\|_2$$

- Results in
  - $O(\text{nnz}(A) \cdot \varepsilon^{-1} \log d)$ time algorithm using column-sparsity transform of Kane and Nelson [Kane, Nelson, J. ACM ’14]
  - $O(nd \log n)$ time algorithm using Fast Johnson-Lindenstrauss transform of Ailon and Chazelle [Ailon, Chazelle, STOC ’06]
Other Subspace Embedding Algorithms

- Not JLT-based subspace embedding
  - $O(\text{nnz}(A)) + \text{poly}(d/\varepsilon)$ time algorithm [Clarkson, Woodruff, STOC '13]

- None oblivious subspace embeddings
  - Based on Leverage Score Sampling [Drineas, Mahoney, Muthukrishnan, SODA '06]
Theorem

Let \( S \in \mathbb{R}^{r \times n} \) be any oblivious subspace embedding matrix and \( \hat{x} = \arg \min_x \|SAx - Sb\|_2 \); then,

\[
\|SA\hat{x} - Sb\|_2 \leq (1 + \varepsilon) \arg \min_x \|Ax - b\|_2
\]

Proof.

1. Let matrix \( U \in \mathbb{R}^{n \times (d+1)} \) be the orthonormal basis of columns of \( A \) together with vector \( b \)
2. Suppose \( S \) is a \( \ell_2 \)-subspace embedding for \( U \)
Questions?