Robust Subspace Estimation in a Stream

Roie Levin, Anish Sevekari, David Woodruff

Carnegie Mellon University

Problem Statement

Least Squares Subspace Estimation

Input: A set of n data points $\{a_i\}_{i=1}^n$ in $\mathbb{R}^d$ and a dimension k

Output: A k-dimensional subspace S such that:

$$\sum_i \mathrm{dist}(S, a_i)^2$$

is minimized, where $\mathrm{dist}(S, x) := \min_{y \in S} \|x - y\|_2$

Problem Statement

Robust Subspace Estimation

Input: A set of n data points $\{a_i\}_{i=1}^n$ in $\mathbb{R}^d$ and a dimension k

Output: A k-dimensional subspace S such that:

$$\sum_i \cancel{\operatorname{dist}(S, a_i)^2} \ \operatorname{dist}(S, a_i)$$

is minimized, where $\operatorname{dist}(S, x) := \min_{y \in S} \|x - y\|_2$

Problem Statement

Robust Subspace Estimation in a Stream

Input: A set of n data points $\{a_i\}_{i=1}^n$ in $\mathbb{R}^d$ and a dimension k

Output: A k-dimensional subspace S such that:

$$\sum_i \cancel{\operatorname{dist}(S, a_i)^2}\ \operatorname{dist}(S, a_i)$$

is minimized, where $\operatorname{dist}(S, x) := \min_{y \in S} \|x - y\|_2$

We are given the data points in a stream:

$$a_1, a_2, a_3, \ldots, a_n$$

and we wish to solve the problem in $\operatorname{poly}(kd \log(nd))$ space.

Problem Statement

Robust Subspace Estimation in a Stream

Input: A set of n data points $\{a_i\}_{i=1}^n$ in $\mathbb{R}^d$ and a dimension k

Output: A k-dimensional subspace S such that:

$$\sum_i \cancel{\mathrm{dist}(S, a_i)^2} \; \mathrm{dist}(S, a_i)$$

is minimized, where $\mathrm{dist}(S, x) := \min_{y \in S} \|x - y\|_2$

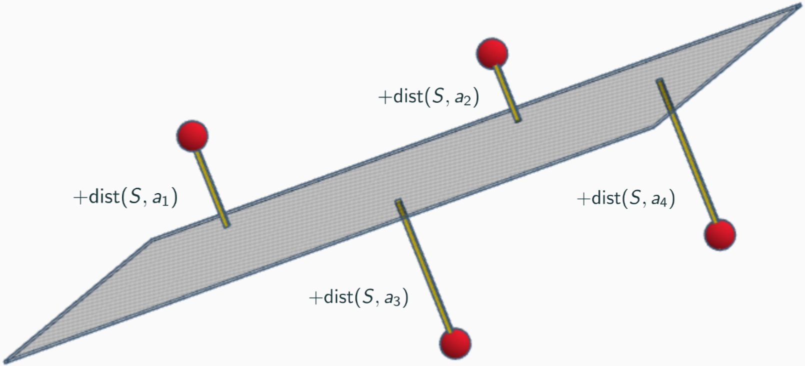
We are given the data points in a stream:

$$a_1, a_2, a_3, \ldots, a_n$$

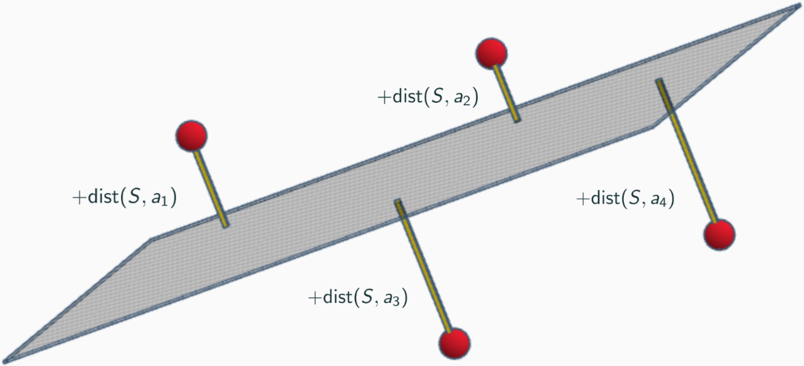
and we wish to solve the problem in $\mathrm{poly}(kd \log(nd))$ space.

(Our algorithm even works in the **turnstile streaming model** with arbitrary, entry-wise $+/-$ updates.)

Problem Statement



Problem Statement



We can write this objective as a low rank approximation problem:

$$\min_{X \text{ rank } k} \|A - AX\|_{2,1}$$

where $\|X\|_{2,1} := \sum_i \|X_{i,*}\|_2$

Previous Work

Hardness

[Clarkson and Woodruff '15] shows that the offline problem is NP-hard to approximate to within a $\left(1 + \frac{1}{\mathrm{poly}(d)}\right)$ factor.

Previous Work

Hardness

[Clarkson and Woodruff '15] shows that the offline problem is NP-hard to approximate to within a $\left(1 + \frac{1}{\mathrm{poly}(d)}\right)$ factor. $\Rightarrow$ **There cannot be a** $(1+\epsilon)$**-approximation algorithm running in time poly**(k/ϵ)**!**

Previous Work

Hardness

[Clarkson and Woodruff '15] shows that the offline problem is NP-hard to approximate to within a $\left(1 + \frac{1}{\mathrm{poly}(d)}\right)$ factor. $\Rightarrow$ **There cannot be a $(1+\epsilon)$-approximation algorithm running in time poly(k/ϵ)!**

Algorithms

[Clarkson and Woodruff '15] also give a $(1+\epsilon)$-approximation algorithm that runs in time

$$O(\mathrm{nnz}(A)) + (n+d)\mathrm{poly}\left(\frac{k}{\epsilon}\right) + \exp\left(\mathrm{poly}\left(\frac{k}{\epsilon}\right)\right)$$

Our Results

Theorem (Streaming Alg. for Robust Subspace Estimation)

There is a randomized algorithm giving a $(1+\epsilon)$-approximate optimal solution to

$$\min_{X \text{ rank } k} \|A - AX\|_{2,1}$$

with the following guarantees:

Our Results

Theorem (Streaming Alg. for Robust Subspace Estimation)
There is a randomized algorithm giving a $(1+\epsilon)$-approximate optimal solution to

$$\min_{X \text{ rank } k} \|A - AX\|_{2,1}$$

with the following guarantees:

1. *Runs in* **turnstile streaming model** *with space:*

$$O\left(d\, poly\left(\frac{k \log(nd)}{\epsilon}\right)\right)$$

Our Results

Theorem (Streaming Alg. for Robust Subspace Estimation)

There is a randomized algorithm giving a $(1+\epsilon)$-approximate optimal solution to

$$\min_{X \text{ rank } k} \|A - AX\|_{2,1}$$

with the following guarantees:

1. *Runs in* **turnstile streaming model** *with space:*

$$O\left(d\,poly\left(\frac{k\log(nd)}{\epsilon}\right)\right)$$

2. *Runs in time (offline):*

$$O(nnz(A)) + (n+d)poly\left(\frac{k\log(nd)}{\epsilon}\right) + \exp\left(poly\left(\frac{k}{\epsilon}\right)\right)$$

Our Results

Theorem (Streaming Alg. for Robust Subspace Estimation)

There is a randomized algorithm giving a $(1+\epsilon)$-approximate optimal solution to

$$\min_{X \text{ rank } k} \|A - AX\|_{2,1}$$

with the following guarantees:

1. *Runs in* **turnstile streaming model** *with space:*

$$O\left(d\, poly\left(\frac{k \log(nd)}{\epsilon}\right)\right)$$

2. *Runs in time (offline):*

$$O(nnz(A)) + (n+d) poly\left(\frac{k \log(nd)}{\epsilon}\right) + \exp\left(poly\left(\frac{k}{\epsilon}\right)\right)$$

(same as [Clarkson Woodruff '15] in leading order terms)

High Level Approach

$$\min_{X \text{ rank } k} \|A - AX\|_{2,1}$$

High Level Approach

$$\min_{X \text{ rank } k} \|A - AX\|_{2,1} \; ???$$

High Level Approach

$$\min_{X \text{ rank } k} \|A - AX\|_{2,1}$$

$$\Big\downarrow \quad \begin{aligned} Y &\leftarrow f_1(A) \\ Z &\leftarrow f_2(A) \end{aligned}$$

$$\min_{X \text{ rank } k} \|Y - ZX\|_{2,1}$$

High Level Approach

$$\min_{X \text{ rank } k} \|A - AX\|_{2,1}$$

$$Y \leftarrow f_1(A)$$
$$Z \leftarrow f_2(A)$$

$$\min_{X \text{ rank } k} \|Y - ZX\|_{2,1}$$

Can brute force in time exponential in dimension!

High Level Approach

$$\min_{X \text{ rank } k} \|A - AX\|_{2,1}$$

(i) Solution below is a $(1+\epsilon)$-approx solution to above

$$\begin{aligned} Y &\leftarrow f_1(A) \\ Z &\leftarrow f_2(A) \end{aligned}$$

$$\downarrow$$

$$\min_{X \text{ rank } k} \|Y - ZX\|_{2,1}$$

Can brute force in time exponential in dimension!

High Level Approach

$$\min_{X \text{ rank } k} \|A - AX\|_{2,1}$$

(i) Solution below is a $(1+\epsilon)$-approx solution to above

$$Y \leftarrow f_1(A)$$
$$Z \leftarrow f_2(A)$$

(ii) f_1, f_2 are random, **oblivious**, **linear** sketches

$$\min_{X \text{ rank } k} \|Y - ZX\|_{2,1}$$

Can brute force in time exponential in dimension!

High Level Approach

$\min_{X \text{ rank } k} \|A - AX\|_{2,1}$

(i) Solution below is a $(1+\epsilon)$-approx solution to above

$Y \leftarrow f_1(A)$

$Z \leftarrow f_2(A)$

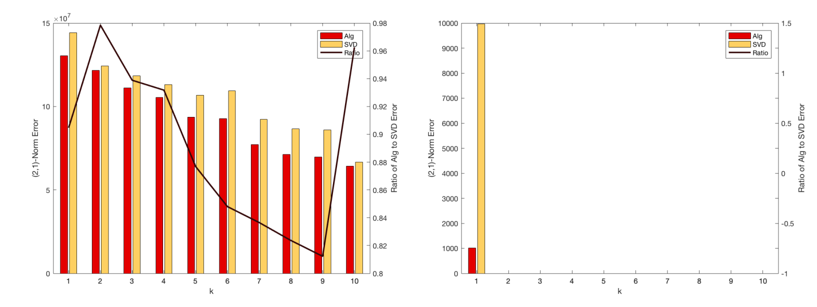
(ii) f_1, f_2 are random, **oblivious**, **linear** sketches

$\min_{X \text{ rank } k} \|Y - ZX\|_{2,1}$

Can brute force in time exponential in dimension!

(iii) All dimensions of Y, Z are small i.e. poly(k/ϵ)

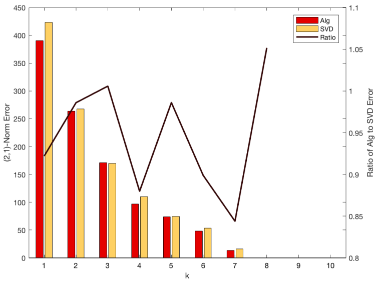
Experiments: Synthetic Data



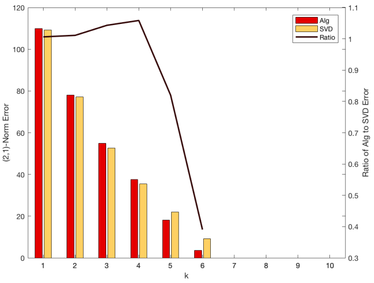
(a) Random matrix + large outliers **(b)** Rank-2 matrix with large outliers

Comparison of Algorithm against SVD on **synthetic** data.

Experiments: Real-World Data



(a) Glass data set



(b) E. Coli. data set

Comparison of Algorithm against SVD on **real-world** data.

Thanks!