Dimensionality Reduction for Sum-of-Distances Metric

Zhili Feng, Praneeth Kacham and David Woodruff
Carnegie Mellon University

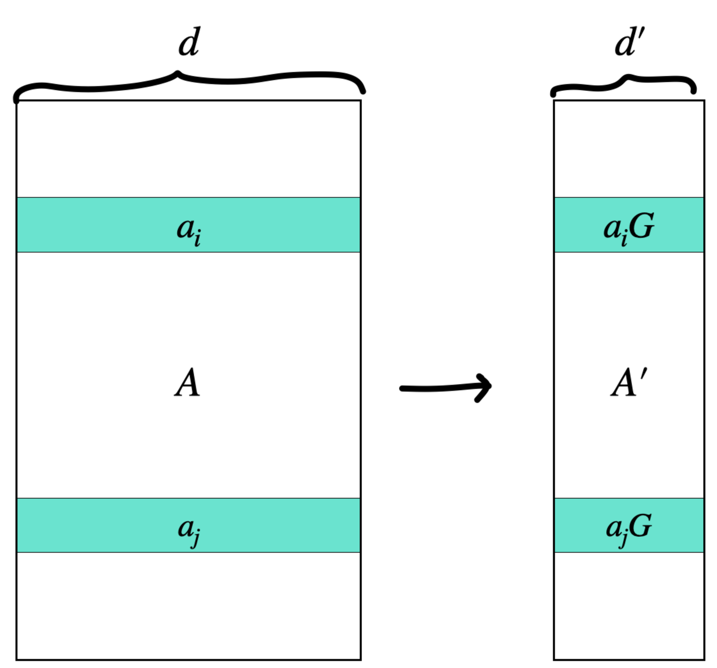
Introduction

- Datasets $A \in \mathbb{R}^{n \times d}$ these days are huge and high-dimensional, where n is the number of data and d is the data dimension.
- Crucial to decrease size of the data to save on storage and computation.
- Two ways to reduce datasets:
  - Dimensionality reduction – reducing d.
  - Coresets – decreasing n (typically a weighted subset of the dataset).
- This work
  - Introduces a novel dimensionality reduction technique for shape fitting problems with the sum of distance metric.
  - Gives a coreset construction for k-median and k-subspace approximation using our dimensionality reduction.

Background

Dimensionality Reduction

- Let $d' \ll d$ to attain significant size reduction



- A is task dependent. E.g. if we want to preserve pairwise ℓ_2 distances, G can be a random Gaussian of size $d \times O(\log(n)/\varepsilon^2)$ by JL lemma.

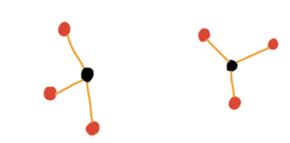
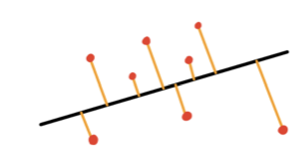
Shape Fitting

- Given data set A and a set of "shapes" $\mathcal{S}$, we want to find $S \in \mathcal{S}$ that minimizes

$$d(A, S) = \sum_i d(a_i, S) = \sum_i \min_{s \in S} d(a_i, s).$$

Here "shape" is just any set of points.

Examples of Shapes

- S is a set of k points → k-median 
- S is a k-dimensional subspace → k subspace approximation 

Related Work

- Sohler and Woodruff [1] give an algorithm for dimensionality reduction with a running time involving an $\exp(\mathrm{poly}(k/\varepsilon))$ factor. We remove the $\exp(\mathrm{poly}(k/\varepsilon))$ term in our results.
- Huang and Vishnoi [2] gave an efficient coreset construction for k-median problem. But their algorithm works solely for coreset construction, whereas our dimension reduction can be used for more tasks.

Our Results

Given a dataset $A \in \mathbb{R}^{n \times d}$, there exists a poly$(k/\varepsilon)$-dimensional subspace P such that projections of each point on P and distance of each point to the subspace P are sufficient to approximate $d(A, S)$ for any shape S that lies in a k dimensional subspace. Such a subspace can be found in time $O(\mathrm{nnz}(A)/\varepsilon^2 + (n + d)\,\mathrm{poly}(k/\varepsilon))$

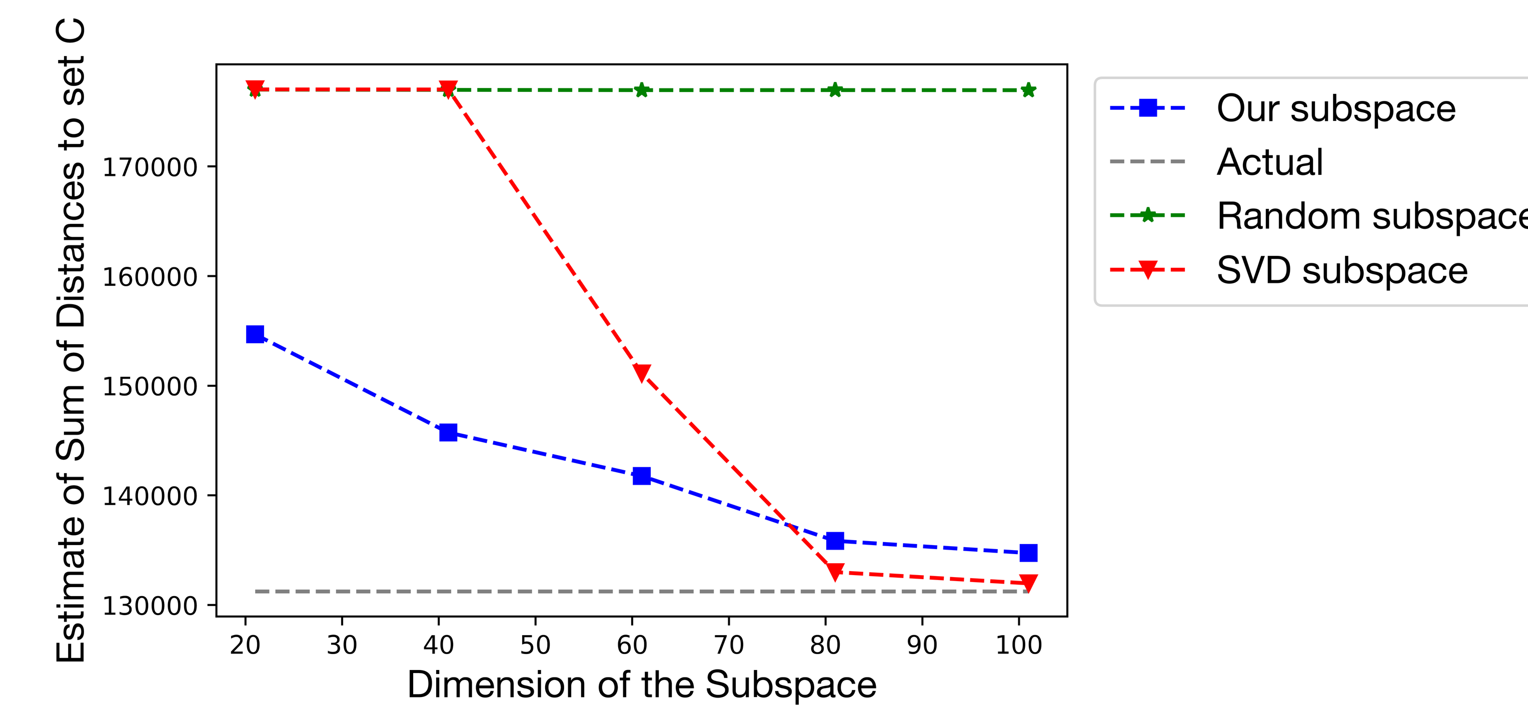
- We also give an algorithm that runs in time $nd \log(nd) + (n + d)\,\mathrm{poly}(k/\varepsilon)$ which is faster when $\mathrm{nnz}(A) \approx nd$.
- Using our dimensionality reduction, small coresets can be constructed for several problems.
- We also show that the coreset construction of [2] can be implemented in $O(\mathrm{nnz}(A) + (n + d)\,\mathrm{poly}(k/\varepsilon))$ time. This does not need our main result.

Techniques

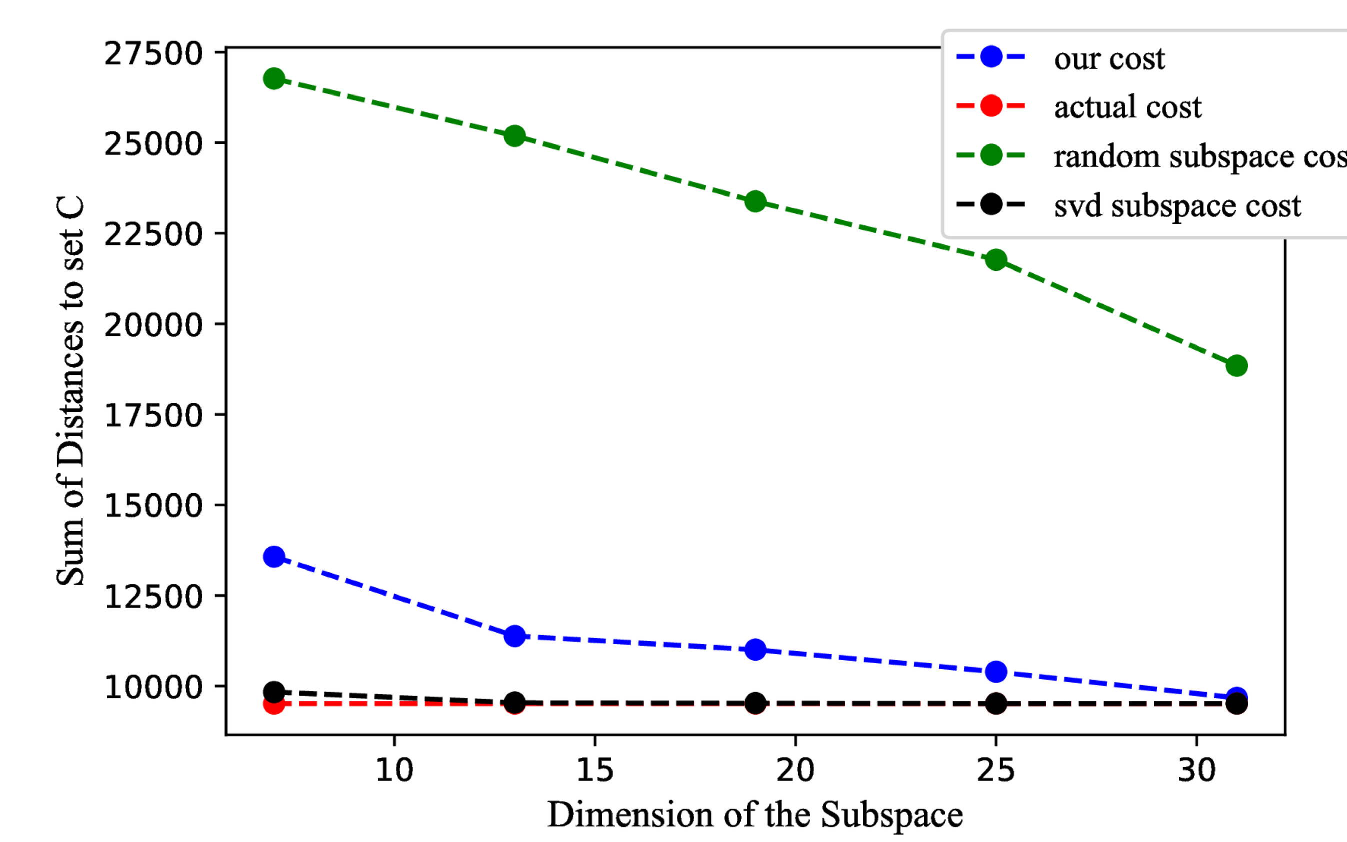
- We adaptively compute $1 + \varepsilon$ approximate bicriteria solutions for subspace approximation with sum-of-distances cost and show that the sum of the bicritera subspaces after $O(1/\varepsilon^2)$ iterations has the desired properties.
- For computing $1 + \varepsilon$ approximate solutions, we use lopsided embeddings, Lewis weight sampling and residual sampling.

Experiments

- We generate a random k-median dataset with 10000 points in $\mathbb{R}^{10000}$ and compute a 100 dimensional subspace using our algorithm. We then compute approximate cost of a center set using our subspace, SVD subspace and a random subspace.



- We run the same experiment on a randomly sampled subset of the CoverType dataset.



References

[1] Sohler, Christian, and David P. Woodruff. "Strong coresets for k-median and subspace approximation: Goodbye dimension." 2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS). IEEE, 2018.
[2] Huang, Lingxiao, and Nisheeth K. Vishnoi. "Coresets for clustering in euclidean spaces: Importance sampling is nearly optimal." Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing. 2020.