Distributed PCA and k-Means Clustering

Yingyu Liang, Maria-Florina Balcan, Vandana Kanchanapally
School of Computer Science
Georgia Institute of Technology
Atlanta, GA 30332
yliang39@gatech.edu,ninamf@cc.gatech.edu,vvandana@gatech.edu

Abstract

This paper proposes a distributed PCA algorithm, with the theoretical guarantee that any good approximation solution on the projected data for k-means clustering is also a good approximation on the original data, while the projected dimension required is independent of the original dimension. When combined with the distributed coreset-based clustering approach in [3], this leads to an algorithm in which the number of vectors communicated is independent of the size and the dimension of the original data. Our experiment results demonstrate the effectiveness of the algorithm.

1 Introduction

Clustering is a classical technique to analyze and summarize data sets, and k-means clustering is probably the mostly widely used one. Most k-means clustering algorithms are designed for the centralized setting, but many modern applications need to cluster large-scale high-dimensional data distributed over different locations, such as distributed databases [17, 5], images and videos over networks [16], surveillance [9] and sensor networks [4, 10].

To address this challenge, a distributed clustering algorithm has been proposed in [3], which is based on distributed coreset construction. A coreset for a data set is a set of weighted points such that its clustering cost on any set of centers approximates the cost of the data, i.e. a summarization of the data with respect to the clustering task. The size of the coreset is independent of the size of the original data, which is useful for large-scale applications. However, it is linear in the dimension of the data, leading to high communication cost for high dimensional data.

In this paper, we propose a distributed PCA algorithm, and show that its output represents the original data in the sense that any good approximation solution of k-means clustering on the output projected data is also a good solution on the original data. When combined with the distributed coreset approach in [3], this leads to an algorithm whose communication cost (in terms of the number of vectors communicated) is independent of the size and the dimension of the original data. Our experiment results demonstrate that this significantly reduces the communication cost while hardly comprising the quality of the k-means clustering solutions.

2 Background and Notations

Distributed Clustering Let $d(p, q)$ denote the Euclidean distance between $p, q \in \mathbf{R}^d$. The goal of k-means clustering is to find a set of k centers $\mathbf{x} = \{x_1, x_2, \ldots, x_k\}$ which minimize the k-means cost of data set $P \subseteq \mathbf{R}^d$. Here the k-means cost is defined as $\mathrm{cost}(P, \mathbf{x}) = \sum_{p \in P} d(p, \mathbf{x})^2$ where $d(p, \mathbf{x}) = \min_{x \in \mathbf{x}} d(p, x)$. For simplicity, we always assume that P is normalized ($\sum_{p \in P} p = 0$).

In the distributed clustering task, we consider a set of n nodes $V = \{v_i, 1 \leq i \leq n\}$ which communicate on an undirected connected graph $G = (V, E)$ with $m = |E|$ edges. More precisely, an edge $(v_i, v_j) \in E$ indicates that v_i and v_j can communicate with each other. On each node v_i, there is a local set of data points $P_i \subseteq \mathbf{R}^d$, and the global data set is $P = \bigcup_{i=1}^{n} P_i$. The goal is to find a set of k centers $\mathbf{x} = \{x_1, x_2, \dots, x_k\}$ which optimize $\mathrm{cost}(P, \mathbf{x})$ while keeping the computation efficient and the communication cost as low as possible. Our focus is to reduce the total communication cost while preserving theoretical guarantees for approximating clustering cost.

PCA Principal Component Analysis is a classical tool for dimension reduction, and has been closely related to k-means [6, 13]. Here we first use PCA on high dimensional data and then do distributed clustering on the projected data, which leads to lower communication cost. We introduce the following notations for PCA. View the local data P_i as a matrix, whose rows are data points. The global data P is then a concatenation of the local data matrix, i.e. $P^T = [P_1^T, P_2^T, \dots, P_n^T]$.

For a matrix $X = [x_{ij}]$, let $||X||_2^2 = \sum_{i,j} x_{i,j}^2$. We say that X has orthonormal columns if its columns are orthogonal unit vectors. Let $L(X)$ denote the linear subspace spanned by the columns of X. For simplicity, for a set of points P, we denote $d^2(P, L(X)) = \sum_{p \in P} d(p, L(X))^2$.

For a point p, let $p_X(p)$ denote its projection to $L(X)$. Note that for an orthogonal matrix X, the projection of a point p to $L(X)$ will be pX using the coordinates with respect to the column space of X, and will be pXX^T using the original coordinates.

Coreset A natural approach for distributed clustering is to generate a summary of the relevant information. The idea of summarization for clustering is formalized by the concept of coresets [11, 7].

Definition 1 (**coreset**). *An ϵ-coreset for a set of points P with respect to a center-based cost function is a set of points S and a set of weights $w : S \to \mathbf{R}$ such that there exists a constant $c_0 \geq 0$, and for any set of centers $\mathbf{x}$, $(1-\epsilon)\mathrm{cost}(P, \mathbf{x}) \leq \sum_{p \in S} w(p)\mathrm{cost}(p, \mathbf{x}) + c_0 \leq (1+\epsilon)\mathrm{cost}(P, \mathbf{x})$.*

A key property of the coreset is that an α-approximation solution for an ϵ-coreset is also a $(1+3\epsilon)\alpha$-approximation for the original data.

3 Distributed PCA

Our distributed PCA algorithm is described in Algorithm 1, where ANNOUNCE is a shorthand for communicating information to all other nodes. The algorithm performs local PCA on each local data set, and communicates the t largest principal components. These are used to estimate the global covariance matrix, which is then used to get the t largest global principal components. Finally, all the local data are projected on these t global principal components. See Figure 1 for an illustration.

Several similar heuristic algorithms have been proposed in [18, 2, 14, 15]. However, they did not provide any theoretical guarantee, or relate distributed PCA to k-means clustering. Here we provide a theoretical analysis, which leads to a way to set the algorithm parameters, so that we will not compromise much on the quality of the clustering obtained on the projected data. Formally,

Theorem 1. *Let X be a $d \times j$ matrix whose columns are orthonormal. Let $\epsilon \in (0, 1]$ and $t \in \mathbf{N}$ with $d - 1 \geq t \geq j + \lceil 4j/\epsilon \rceil - 1$. Then the output of Algorithm 1 satisfies*

$$0 \leq ||PX||_2^2 - ||\hat{P}X||_2^2 \leq \epsilon d^2(P, L(X)) \quad \textit{and} \quad 0 \leq ||PX - \hat{P}X||_2^2 \leq \epsilon d^2(P, L(X))$$

Intuitively, it implies that the squared distances to any low dimension subspace $L(X)$ from the projected data and the original data are approximately equal when the number of principal components used is sufficiently large compared to the dimension of $L(X)$. As shown in the next section, this guarantees that the projected data can act as a proxy for the original data in k-means clustering.

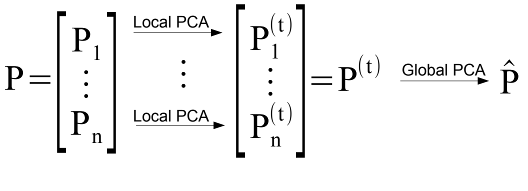

Figure 1: The key points of our distributed PCA algorithm. The local PCA is by SVD on the local data, and the global PCA is by computing eigenvectors on the covariance matrix of $P^{(t)}$.

Algorithm 1 Distributed PCA

```
Input: local data sets $\{P_i, 1 \le i \le n\}$, projected dimension $t$.
▷ Local PCA
Round 1: on each node $v_i \in V$ do
  Compute local SVD: $P_i = U_i D_i (E_i)^T$.
  Let $D_i^{(t)}$ be the matrix that contains the first $t$ diagonal entries of $D_i$ and is 0 otherwise.
  Let $P_i^{(t)} = U_i D_i^{(t)} (E_i)^T$. Let $E_i^{(t)}$ be the matrix that contains the first $t$ columns of $E_i$.
  ANNOUNCE: $D_i^{(t)}, E_i^{(t)}$.
▷ Global PCA
Round 2: on each node $v_i \in V$ do
  Use $D_i^{(t)}, E_i^{(t)}$ to compute $S_i^{(t)} = (P_i^{(t)})^T P_i^{(t)}$. Set $S^{(t)} = \sum_{i=1}^{n} S_i^{(t)}$.
  Compute the eigenvectors for the estimated covariance matrix: $S^{(t)} = E \Lambda E^T$.
  Let $E^{(t)}$ be the matrix that contains the first $t$ columns of $E$.
  Project $P_i^{(t)}$ on $E^{(t)}$, resulting in $\hat{P}_i = P_i^{(t)} E^{(t)} (E^{(t)})^T$.
Output: $\hat{P}^T = [\hat{P}_1^T, \dots, \hat{P}_n^T]$.
```

Proof Sketch of Theorem 1 The algorithm performs two projections: local PCA projecting P_i to $P_i^{(t)}$, and global PCA projecting $P_i^{(t)}$ to $\hat{P}_i$. Let $(P^{(t)})^T = [(P_1^{(t)})^T, \dots, (P_n^{(t)})^T]$. To bound the error between P and $\hat{P}$, we need to first bound the error between P and $P^{(t)}$, and then bound the error between $P^{(t)}$ and $\hat{P}$. The following lemma is useful in bounding these errors.

Lemma 1. *[Lemma 6.1 in [8]] Let $A \in \mathbf{R}^{\ell \times d}$ be an $\ell \times d$ matrix with singular value decomposition $A = UDE^T$, where D has diagonal entries $\sqrt{s_1}, \dots, \sqrt{s_d}$ sorted non-increasingly. Let X be a $d \times j$ matrix whose columns are orthonormal. Let $\epsilon \in (0, 1]$ and $t \in \mathbf{N}$ with $d - 1 \ge t \ge j + \lceil j/\epsilon \rceil - 1$. Let $D^{(t)}$ be the matrix that contains the first t diagonal entries of D and is 0 otherwise, and $A^{(t)} = UD^{(t)}E^T$. Then $0 \le ||(A - A^{(t)})X||_2^2 = ||AX||_2^2 - ||A^{(t)}X||_2^2 \le \epsilon \sum_{i=j+1}^{d} s_i \le \epsilon d^2(A, L(X))$.*

Lemma 1 bounds $||P_i X||_2^2 - ||P_i^{(t)} X||_2^2$ for each local PCA, and thus bounds $||PX||_2^2 - ||P^{(t)}X||_2^2 = \sum_i [||P_i X||_2^2 - ||P_i^{(t)} X||_2^2]$. It also bounds $||P^{(t)}X||_2^2 - ||\hat{P}X||_2^2$ for global PCA. These then lead to the first claim in the theorem. The second claim on $||PX - \hat{P}X||_2^2$ can be proved similarly.

4 Distributed Clustering

In this section, we show that any good approximation solution on the projected data constructed by the distributed PCA algorithm is also a good approximation on the original data.

Theorem 2. *Let $\mathbf{x}$ be a set of k centers in $\mathbf{R}^d$. Let $\epsilon \in (0, 1]$ and $t \in \mathbf{N}$ with $d - 1 \ge t \ge k + \lceil 50k/\epsilon^2 \rceil$. Then there exists a constant $c_0 \ge 0$, such that the output of Algorithm 1 satisfies $(1 - \epsilon)\text{cost}(P, \mathbf{x}) \le \text{cost}(\hat{P}, \mathbf{x}) + c_0 \le (1 + \epsilon)\text{cost}(P, \mathbf{x})$.*

The analysis follows the ideas in [8]. Let $X \in \mathbf{R}^{d \times k}$ has orthonormal columns that span $\mathbf{x}$. The cost of P can be decomposed into two parts: the squared distances $d^2(P, L(X))$ to the subspace spanned by X, and the squared distances $\sum_i d^2(p_X(p_i), \mathbf{x})$ between the projection of the points on $L(X)$ and $\mathbf{x}$. The cost of $\hat{P}$ can be decomposed similarly. Their difference in the first part (compensated by $c_0 = ||P||_2^2 - ||\hat{P}||_2^2$) can be bounded by $||PX||_2^2 - ||\hat{P}X||_2^2$. The difference in the second part can be bounded approximately by $\sum_i d^2(p_X(p_i), p_X(\hat{p}_i))/\epsilon = ||PX - \hat{P}X||_2^2/\epsilon$. Then the theorem follows from Theorem 1. The complete proof is provided in the appendix.

By Theorem 2, the distributed coreset construction algorithm in [3] can be applied on the projected data to get a coreset of size independent of the original dimension. Then we get an algorithm with low communication cost for high dimensional data, which is summarized in Theorem 3.

Theorem 3. *Given an α-approximation algorithm for k-means as a subroutine, there exists an algorithm that with probability at least $1 - \delta$ outputs a $(1 + \epsilon)\alpha$-approximation solution for distributed k-means clustering. The total communication cost is $O(m(\frac{k^2}{\epsilon^6} + \frac{1}{\epsilon^4} \log \frac{1}{\delta}) + mnk \log \frac{nk}{\delta})$ vectors.*

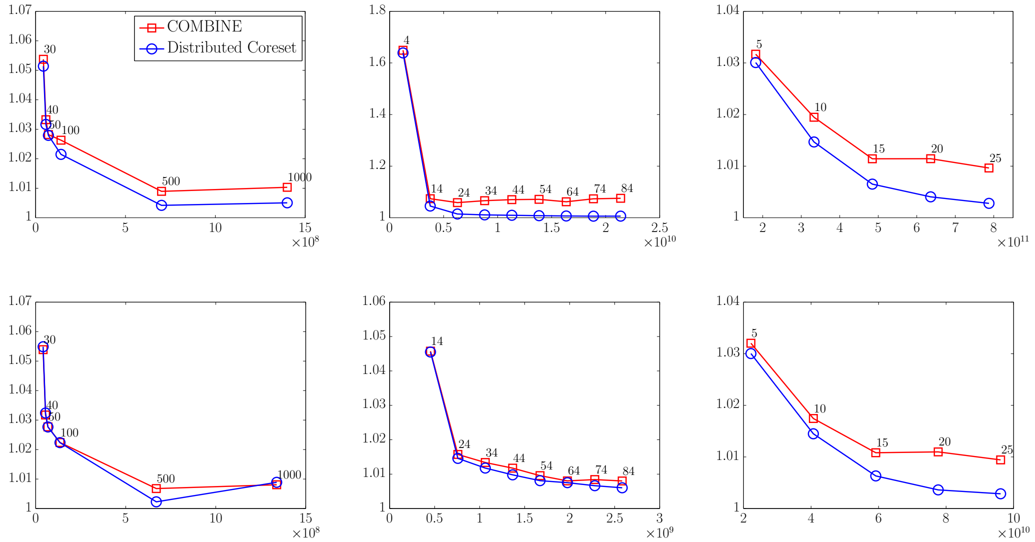

Figure 2: k-means cost (normalized by baseline) v.s. communication cost when the projection dimension varies. Rows: random graphs, and grid graphs. Columns: Daily and Sports Activities, MNIST, and Bag of Words data sets. In each subfigure, the x-axis represents the communication cost, the y-axis represents the k-means cost, and the number labels are the projection dimensions.

5 Experiments

In these experiments we seek to understand how well the projected data approximates the original data, by measuring the k-means costs of the clustering solutions obtained after dimension reduction.

Dataset We choose the following real world data sets from [1]: Daily and Sports Activities data (9210 points in $\mathbf{R}^{5625}$), MNIST handwritten digits ($70,000$ points in $\mathbf{R}^{784}$). We use $k = 10$ for these data sets. We further choose Bag of Words (NYTimes) ($300,000$ points in $\mathbf{R}^{102660}$) and use $k = 20$ for this data set.

Experimental Methodology Following the setup in [3], we first generate a communication graph, which can be a grid graph, or a random graph that includes each edge independently with probability 0.3. For Daily and Sports Activities data set, we use random graphs with 10 nodes and 3×3 grid graphs. For the other data sets, we use random graphs with 100 nodes and 10×10 grid graphs. Then we distribute the data over the graphs using weighted partition, where each data point is distributed to the local sites with probability proportional to the site's weight chosen from $|N(0, 1)|$.

For each projection dimension, we first construct a coreset on the projected data, using the COMBINE or distributed coreset algorithm in [3]. The COMBINE algorithm builds a coreset for each local data set and then combines them to get one global coreset, while the distributed coreset algorithm considers the different contributions of the local data to the k-means cost when it builds the global coreset (see [3] for details). After building the coreset, we then run Lloyd's method on it to get a k-means clustering solution. Finally, we compute the ratio of its cost to the k-means cost obtained by running Lloyd's method on the original data. The average results over 10 runs are reported. We lower the projection dimension until there is a significant increase in the k-means costs.

Results Figure 2 shows the results of the data sets. The plots show the increase in k-means cost ratio upon decreasing the dimension of the data. We can observe that there is a slight increase compared to the huge reduction in dimension and thus communication cost. For example, on Daily and Sports Activities data, the k-means cost increases less than 4% when the dimension is reduced from 5625 to as low as 40. This is even more significant on higher dimensional data: on Bag of Words, the dimension can be reduced from 102660 to around 20. Such reduction then lowers the communication cost by magnitudes. The plots also indicate that the distributed coreset algorithm in [3] performs better than the COMBINE algorithm, when applied with our distributed PCA algorithm.

References

[1] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[2] Z.-J. Bai, R. H. Chan, and F. T. Luk. Principal component analysis for distributed data sets with updating. In *Advanced Parallel Processing Technologies*. 2005.

[3] M.-F. Balcan, S. Ehrlich, and Y. Liang. Distributed k-means and k-median clustering on general communication topologies. In *Advances in Neural Information Processing Systems*, 2013.

[4] J. Considine, F. Li, G. Kollios, and J. Byers. Approximate aggregation techniques for sensor databases. In *Proceedings of the International Conference on Data Engineering*, 2004.

[5] J. C. Corbett, J. Dean, M. Epstein, A. Fikes, C. Frost, J. Furman, S. Ghemawat, A. Gubarev, C. Heiser, P. Hochschild, et al. Spanner: Googles globally-distributed database. In *Proceedings of the USENIX Symposium on Operating Systems Design and Implementation*, 2012.

[6] C. Ding and X. He. K-means clustering via principal component analysis. In *Proceedings of the International Conference on Machine learning*, 2004.

[7] D. Feldman and M. Langberg. A unified framework for approximating and clustering data. In *Proceedings of the Annual ACM Symposium on Theory of Computing*, 2011.

[8] D. Feldman, M. Schmidt, and C. Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, 2013.

[9] S. Greenhill and S. Venkatesh. Distributed query processing for mobile surveillance. In *Proceedings of the International Conference on Multimedia*, 2007.

[10] M. Greenwald and S. Khanna. Power-conserving computation of order-statistics over sensor networks. In *Proceedings of the ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 2004.

[11] S. Har-Peled and S. Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the Annual ACM Symposium on Theory of Computing*, 2004.

[12] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. A local search approximation algorithm for k-means clustering. In *Proceedings of the Annual Symposium on Computational Geometry*, 2002.

[13] A. Kumar and R. Kannan. Clustering with spectral norm and the k-means algorithm. In *Proceedings of the Annual IEEE Symposium on Foundations of Computer Science*, 2010.

[14] Y.-A. Le Borgne, S. Raybaud, and G. Bontempi. Distributed principal component analysis for wireless sensor networks. *Sensors*, 2008.

[15] S. V. Macua, P. Belanovic, and S. Zazo. Consensus-based distributed principal component analysis in wireless sensor networks. In *Proceedings of the IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2010.

[16] S. Mitra, M. Agrawal, A. Yadav, N. Carlsson, D. Eager, and A. Mahanti. Characterizing web-based video sharing workloads. *ACM Transactions on the Web*, 2011.

[17] C. Olston, J. Jiang, and J. Widom. Adaptive filters for continuous queries over distributed data streams. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2003.

[18] Y. Qu, G. Ostrouchov, N. Samatova, and A. Geist. Principal component analysis for dimension reduction in massive distributed data sets. In *Proceedings of IEEE International Conference on Data Mining*, 2002.

A Proof of Lemma 1

The proof comes from [8]. We first have

$$\begin{aligned}||AX||_2^2 - ||A^{(t)}X||_2^2 &= ||UDE^TX||_2^2 - ||UD^{(t)}E^TX||_2^2 \\ &= ||DE^TX||_2^2 - ||D^{(t)}E^TX||_2^2 = ||(D - D^{(t)})E^TX||_2^2.\end{aligned}$$

where the second equality follows since U has orthonormal columns, the third equality follows since for $M = E^TX$ we have

$$||DM||_2^2 - ||D^{(t)}M||_2^2 = \sum_{i=1}^{d}\sum_{j=1}^{d} s_i m_{ij}^2 - \sum_{i=1}^{t}\sum_{j=1}^{d} s_i m_{ij}^2 = \sum_{i=t+1}^{d}\sum_{j=1}^{d} s_i m_{ij}^2 = ||(D - D^{(t)})M||_2^2.$$

Then $||AX||_2^2 - ||A^{(t)}X||_2^2 = ||AX - A^{(t)}X||_2^2 \geq 0$ since U has orthonormal columns. Also,

$$||AX||_2^2 - ||A^{(t)}X||_2^2 = ||(D - D^{(t)})E^TX||_2^2 \leq ||(D - D^{(t)})||_S^2||X||_2^2 = js_{t+1}$$

where the inequality follows because the spectral norm is consistent with the Euclidean norm. It follows for our choice of t that

$$js_{t+1} \leq \epsilon(t - j + 1)s_{t+1} \leq \epsilon \sum_{i=j+1}^{t+1} s_i \leq \epsilon \sum_{i=j+1}^{d} s_i.$$

By the property of singular value decomposition, we have $\sum_{i=j+1}^{d} s_i \leq d^2(A, L(X))$, which completes the proof.

B Proof of Theorem 1

Besides Lemma 1, the following fact is useful for our analysis.

Fact 1. $||A||_2^2 = ||A^{(t)}||_2^2 + \sum_{i=t+1}^{d} s_i$.

We now bound the errors of the local PCA and global PCA in the algorithm.

Lemma 2. *Let X be a $d \times j$ matrix whose columns are orthonormal. Let $\epsilon \in (0, 1]$ and $t \in \mathbf{N}$ with $d - 1 \geq t \geq j + \lceil j/\epsilon \rceil - 1$. Then*

$$0 \leq ||(P - P^{(t)})X||_2^2 = ||PX||_2^2 - ||P^{(t)}X||_2^2 \leq \epsilon d^2(P, L(X)).$$

Proof. We first decompose the error on the global data into errors on the local data:

$$||PX||_2^2 - ||P^{(t)}X||_2^2 = \sum_{i=1}^{n} \left[||P_iX||_2^2 - ||P_i^{(t)}X||_2^2\right].$$

Note that for each P_i, by Lemma 1 we have

$$0 \leq ||P_iX||_2^2 - ||P_i^{(t)}X||_2^2 = ||(P_i - P_i^{(t)})X||_2^2 \leq \epsilon d^2(P_i, L(X)).$$

This holds for any $0 \leq i \leq n$. Then the lemma follows from $||(P - P^{(t)})X||_2^2 = \sum_{i=1}^{n} ||(P_i - P_i^{(t)})X||_2^2$ and $\sum_{i=1}^{n} d^2(P_i, L(X)) = d^2(P, L(X))$. □

Lemma 3. *Let X be a $d \times j$ matrix whose columns are orthonormal. Let $\epsilon \in (0, 1]$ and $t \in \mathbf{N}$ with $d - 1 \geq t \geq j + \lceil j/\epsilon \rceil - 1$. Then*

$$0 \quad \leq \quad ||P^{(t)}X||_2^2 - ||\hat{P}X||_2^2 = ||(P^{(t)} - \hat{P})X||_2^2 \leq \epsilon(1 + \epsilon)d^2(P, L(X)).$$

Proof. By Lemma 1 we have

$$0 \leq ||P^{(t)}X||_2^2 - ||\hat{P}X||_2^2 = ||(P^{(t)} - \hat{P})X||_2^2 \leq \epsilon d^2(P^{(t)}, L(X)).$$

So it suffices to show that

$$d^2(P^{(t)}, L(X)) \leq (1+\epsilon)d^2(P, L(X)).$$

In fact, by Pythagorean Theorem, $d^2(P^{(t)}, L(X)) - d^2(P, L(X)) = ||P^{(t)}||_2^2 - ||P^{(t)}X||_2^2 - (||P||_2^2 - ||PX||_2^2)$. By Fact 1, we have $||P_i^{(t)}||_2^2 \leq ||P_i||_2^2$, and thus $||P^{(t)}||_2^2 = \sum_{i=1}^n ||P_i^{(t)}||_2^2 \leq \sum_{i=1}^n ||P_i||_2^2 = ||P||_2^2$. Then

$$\begin{aligned} d^2(P^{(t)}, L(X)) - d^2(P, L(X)) &= ||P^{(t)}||_2^2 - ||P^{(t)}X||_2^2 - (||P||_2^2 - ||PX||_2^2) \\ &\leq ||PX||_2^2 - ||P^{(t)}X||_2^2 \leq \epsilon d^2(P, L(X)) \end{aligned}$$

where the last inequality follows from Lemma 2. □

Applying Lemma 2 and 3 with accuracy $\epsilon/4$, we have the theorem.

C Proof of Theorem 2

The analysis follows the ideas in [8], but is tailored for the distributed setting. We first begin with the following lemma, showing that the cost of the projected data to any low dimension subspace approximates that of the original data, compensated by a positive constant.

Lemma 4. *Let X be a $d \times j$ matrix whose columns are orthonormal. Let $\epsilon \in (0, 1]$ and $t \in \mathbf{N}$ with $d - 1 \geq t \geq j + \lceil 4j/\epsilon \rceil - 1$. Then there exists $c_1 \geq 0$ such that*

$$0 \leq d^2(\hat{P}, L(X)) + c_1 - d^2(P, L(X)) \leq \epsilon d^2(P, L(X)).$$

Proof. We have from Pythagorean Theorem

$$\begin{aligned} d^2(\hat{P}, L(X)) - d^2(P, L(X)) &= ||\hat{P}||_2^2 - ||\hat{P}X||_2^2 - (||P||_2^2 - ||PX||_2^2) \\ &= ||PX||_2^2 - ||\hat{P}X||_2^2 - c_1 \end{aligned}$$

where $c_1 = ||P||_2^2 - ||\hat{P}||_2^2$. Note that by Fact 1,

$$c_1 = \sum_{i=1}^n \left[||P_i||_2^2 - ||P_i^{(t)}||_2^2 \right] + ||P^{(t)}||_2^2 - ||\hat{P}||_2^2 \geq 0.$$

Then the lemma follows from Theorem 1. □

The next lemma shows that the projection of the projected data to any low dimension subspace approximates the projection of the projected data in the sense that their distances are small.

Let p_i denote the ith row of the data P, and let $\hat{p}_i$ denote the ith row of $\hat{P}$.

Lemma 5. *Let X be a $d \times j$ matrix whose columns are orthonormal. Let $\epsilon \in (0, 1]$ and $t \in \mathbf{N}$ with $d - 1 \geq t \geq j + \lceil 4j/\epsilon \rceil - 1$. Then*

$$\sum_{i=1}^{|P|} d(p_X(p_i), p_X(\hat{p}_i))^2 \leq \epsilon d^2(P, L(X)).$$

Proof. Since X is orthogonal, $p_X(p) = pXX^T$. Then

$$\sum_{i=1}^{|P|} d(p_X(p_i), p_X(\hat{p}_i))^2 = \sum_{i=1}^{|P|} ||p_i XX^T - \hat{p}_i XX^T||_2^2 = ||PXX^T - \hat{P}XX^T||_2^2.$$

This can be simplified to $||(P - \hat{P})X||_2^2$ since

$$\begin{aligned} ||PXX^T - \hat{P}XX^T||_2^2 &= ||(P - \hat{P})XX^T||_2^2 = \text{trace}[(P - \hat{P})XX^TXX^T(P - \hat{P})^T)] \\ &= \text{trace}[(P - \hat{P})XX^T(P - \hat{P})^T)] = ||(P - \hat{P})X||_2^2. \end{aligned}$$

The lemma then follows from Theorem 1. □

The above two lemmas are the key elements needed to show our final theorem. Before proving the theorem, we further need the following "weak triangle inequality", which is well known in the coreset literature. The proof is included in the appendix for completeness.

Lemma 6. *[Lemma 7.1 in [8]] For any $0 \le \epsilon \le 1$, a compact set $C \subseteq \mathbf{R}^d$, and $p, q \in \mathbf{R}^d$,*

$$|d(p,C)^2 - d(q,C)^2| \le \frac{12d(p,q)^2}{\epsilon} + \frac{\epsilon}{2}d(p,C)^2.$$

Proof. Using the triangle inequality,

$$\begin{aligned} d(p,C)^2 - d(q,C)^2| &= |d(p,C) - d(q,C)| \cdot (d(p,C) + d(q,C)) \\ &\le d(p,q)(2d(p,C) + d(p,q)) \\ &\le d(p,q)^2 + 2d(p,C)d(p,q). \end{aligned} \tag{1}$$

Either $d(p,C) \le d(p,q)/\epsilon$ or $d(p,q) < \epsilon d(p,C)$. Hence,

$$d(p,C)d(p,q) \le \frac{d(p,q)^2}{\epsilon} + \epsilon d(p,C)^2.$$

Combining the last inequality with (1) yields

$$|d(p,C)^2 - d(q,C)^2| \le d(p,q)^2 + \frac{2d(p,q)^2}{\epsilon} + 2\epsilon d(p,C)^2 \le \frac{3d(p,q)^2}{\epsilon} + 2\epsilon d(p,C)^2.$$

Finally, the lemma follows by replacing ϵ with $\epsilon/4$. □

We are now ready to prove the theorem, which guarantees that a coreset for the output of the distributed PCA algorithm is also a coreset for the original data.

Theorem 2. *Let $\mathbf{x}$ be a set of k centers in $\mathbf{R}^d$. Let $\epsilon \in (0,1]$ and $t \in \mathbf{N}$ with $d - 1 \ge t \ge k + \lceil 50k/\epsilon^2 \rceil$. Then there exists a constant $c_0 \ge 0$, such that the output of Algorithm 1 satisfies*

$$(1-\epsilon)d^2(P,\mathbf{x}) \le d^2(\hat{P},\mathbf{x}) + c_0 \le (1+\epsilon)d^2(P,\mathbf{x}).$$

Proof. Let $X \in \mathbf{R}^{d\times k}$ has orthonormal columns that span $\mathbf{x}$. Let c_0 be the constant c_1 in Lemma 4. Then by Pythagorean theorem we have

$$|d^2(\hat{P},\mathbf{x}) + c_0 - d^2(P,\mathbf{x})| = \left| d^2(\hat{P},L(X)) + c_0 - d^2(P,L(X)) + \sum_{i=1}^{|P|}\left[d(p_X(p_i),\mathbf{x})^2 - d(p_X(\hat{p}_i),\mathbf{x})^2\right]\right|.$$

By Lemma 4 we have

$$\left| d^2(\hat{P},L(X)) + c_0 - d^2(P,L(X))\right| \le \frac{\epsilon^2}{4}d^2(P,L(X)). \tag{2}$$

By Lemma 5 and Lemma 6 we have

$$\begin{aligned} \sum_{i=1}^{|P|}\left|d(p_X(p_i),\mathbf{x})^2 - d(p_X(\hat{p}_i),\mathbf{x})^2\right| &\le \sum_{i=1}^{|P|}\left[\frac{12d(p_X(p_i),p_X(\hat{p}_i))^2}{\epsilon} + \frac{\epsilon}{2}d(p_X(p_i),\mathbf{x})^2\right] \\ &\le \frac{\epsilon}{4}d^2(P,\mathbf{x}) + \frac{\epsilon}{2}\sum_{i=1}^{|P|}d(p_X(p_i),\mathbf{x})^2. \end{aligned} \tag{3}$$

Since $d^2(P,L(X)) \le d^2(P,\mathbf{x})$ and $d(p_X(p_i),\mathbf{x}) \le d(p_i,\mathbf{x})$, the theorem follows from (2)(3). □