Journal of Machine Learning Research 22 (2021) 1-27 Submitted 3/21; Revised 8/21; Published 8/21

DeEPCA: Decentralized Exact PCA with Linear Convergence Rate

Haishan Ye * YEHAISHAN@XJTU.EDU.CN
Center for Intelligent Decision-Making and Machine Learning
School of Management
Xi'an Jiaotong University
Xi'an, China

Tong Zhang TONGZHANG@TONGZHANG-ML.ORG
Computer Science & Mathematics
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong

Editor: Matthias Seeger

Abstract

Due to the rapid growth of smart agents such as weakly connected computational nodes and sensors, developing decentralized algorithms that can perform computations on local agents becomes a major research direction. This paper considers the problem of decentralized principal components analysis (PCA), which is a statistical method widely used for data analysis. We introduce a technique called subspace tracking to reduce the communication cost, and apply it to power iterations. This leads to a decentralized PCA algorithm called `DeEPCA`, which has a convergence rate similar to that of the centralized PCA, while achieving the best communication complexity among existing decentralized PCA algorithms. `DeEPCA` is the first decentralized PCA algorithm with the number of communication rounds for each power iteration independent of target precision. Compared to existing algorithms, the proposed method is easier to tune in practice, with an improved overall communication cost. Our experiments validate the advantages of `DeEPCA` empirically.

Keywords: Decentralized Algorithm, Principal Component Analysis, Subspace Tracking

1. Introduction

Principal Components Analysis (PCA) is a statistical data analysis method wide applications in machine learning (Moon and Phillips, 2001; Bishop, 2006; Ding and He, 2004; Dhillon et al., 2015), data mining (Cadima et al., 2004; Lee et al., 2010; Qu et al., 2002), and engineering (Bertrand and Moonen, 2014). In recent years, because of the rapid growth of data and quick advances in network technology, developing distributed algorithms has become a more and more important research topic, due to their advantages in privacy preserving, robustness, lower communication cost, etc. (Kairouz et al., 2019; Lian et al., 2017; Nedic and Ozdaglar, 2009). There have been a number of previous studies of decentralized PCA algorithms (Scaglione et al., 2008; Kempe and McSherry, 2008; Suleiman et al., 2016; Wai et al., 2017).

*. Corresponding author.

©2021 Haishan Ye and Tong Zhang.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v22/21-0298.html.

In a typical decentralized PAC setting, we assume that a positive semi-definite matrix A is stored at different agents. Specifically, the matrix A can be decomposed as

$$A = \frac{1}{m}\sum_{j=1}^{m} A_j,$$

where data for A_j is stored in the j-th agent and known only to the agent (This helps to preserve privacy). The agents form a connected and undirected network. Agents can communicate with their neighbors in the network to cooperatively compute the PCA of A.

To obtain the top-k principal components of the positive semi-definite matrix $A \in \mathbb{R}^{d\times d}$, a commonly used centralized algorithm is the power method, which converges fast in practice with a linear convergence rate (Golub and Van Loan, 2012). In the implementation of decentralized PCA, a natural idea is the decentralized power method (`DePM`) which mimics its centralized counterpart. The main procedure of `DePM` can be summarized as a local power iteration plus a multi-consensus step to synchronize the local computations (Kempe and McSherry, 2008; Raja and Bajwa, 2015; Wai et al., 2017; Wu et al., 2018). The multi-consensus step in `DePM` is used to achieve averaging. However, a decentralized PCA algorithm based on `DePM` suffers from a suboptimal communication cost, and is tricky to implement in practice. For each power iteration, theoretically, each agent requires $\Omega\left(\log\frac{1}{\epsilon}\right)$ times communication with its neighbors, where ϵ is the target precision (More specifically, it requires $K = \Omega(\log\frac{1}{\epsilon})$ in Algorithm 3). The communication cost becomes much quite significant when ϵ is small. Although seemingly only a logarithmic factor, in practice, with a data size of merely 10000, this logarithmic factor leads to an order of magnitude more communications. This is clearly prohibitively large for many applications. Moreover, one often has to gradually increase the number of communication rounds in the multi-consensus step to deal with increased precision. However, this strategy makes the tuning of `DePM` difficult for practical applications.

In this paper, we propose a new decentralized PCA algorithm that does not suffer from the weakness of `DePM`. We observe that the communication precision requirement in `DePM` comes from the heterogeneity of data in different agents. Due to the heterogeneity, the local power method will converge to the top-k principal components of the local matrix A_j if no consensus step is conducted to perform averaging. To conquer the weakness of `DePM` whose consensus steps in each power iteration depend on the target precision ϵ, we adapted a technique called gradient tracking in the existing decentralized optimization literature, so that it can be used to track the subspace in power iterations. We call this adapted technique *subspace tracking*. Based on the subspace tracking technique and multi-consensus, we propose Decentralized Exact PCA (`DeEPCA`) which can achieve a linear convergence rate similar to the centralized PCA, but the consensus steps of each power iteration are independent of the target precision ϵ. We summarize our contributions as follows:

1. We propose a novel power-iteration based decentralized PCA called `DeEPCA`, which can achieve the *best* known communication complexity. The communication complexity of `DeEPCA` is much lower than those of existing algorithms, especially when the final error ϵ is small. Furthermore, `DeEPCA` is the first decentralized PCA algorithm whose consensus steps of each power iteration does *not* depend on the target precision ϵ.

2. We show that the 'gradient tracking' technique from the decentralized optimization literature can be adapted to *subspace tracking* for PCA. The resulting `DeEPCA` algorithm can be regarded as a novel decentralized power method. Because the power method is the foundation of many

matrix decomposition problems, subspace tracking and the proof technique of `DeEPCA` can be applied to develop communication efficient decentralized algorithms for spectral analysis, and low rank matrix approximation.

3. The improvement is practically significant. Our experiments show that `DeEPCA` can achieve a linear convergence rate comparable to centralized PCA, even if only a small number of consensus steps are used in each power iteration. In contrast, the conventional decentralized PCA algorithm based on `DePCA` can not converge to the principal components of A when the number of consensus steps is not large.

After the arxiv submission of our paper, Chen et al. (2021) proposed a decentralized Riemannian gradient tracking algorithm (`DRGTA`) which can also be used to obtain the decentralized PCA. However, Chen et al. (2021) can only prove that `DRGTA` converges with a rate $\mathcal{O}(1/T)$, where T is the number of iteration. In contrast, `DeEPCA` converges linearly with a rate almost the same to the one of centralized power method based PCA algorithm. Furthermore, `DRGTA` requires tuning the step size to achieve the best performance. Thus, `DRGTA` is much different from `DeEPCA`.

2. Notation

In this section, we introduce notations and definitions that will be used throughout the paper.

2.1 Notation

Given a matrix $A = [a_{ij}] \in \mathbb{R}^{n\times d}$ and a positive integer $k \le \min\{n, d\}$, its SVD is given as $A = U\Sigma V^T = U_k\Sigma_k V_k^T + U_{\backslash k}\Sigma_{\backslash k}V_{\backslash k}^T$, where U_k and $U_{\backslash k}$ contain the left singular vectors of A, V_k and $V_{\backslash k}$ contain the right singular vectors of A, and $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_\ell)$ with $\sigma_1 \ge \sigma_2 \ge \dots \ge \sigma_{\min\{n,d\}} \ge 0$ are the nonzero singular values of A. Accordingly, we can define the Frobenius norm $\|A\| = \sqrt{\sum_{i=1}^{\min\{n,d\}}\sigma_i^2} = \sqrt{\sum_{i=1,j=1}^{n,d}(A(i,j))^2}$ and the spectral norm $\|A\|_2 = \sigma_1(A)$, where $A(i,j)$ denotes the i,j-th entry of A. We will use $\sigma_{\max}(A)$ to denote the largest singular value and $\sigma_{\min}(A)$ to denote the smallest singular value which may be zero. If A is symmetric positive semi-definite, then it holds that $U = V$ and $\lambda_i(A) = \sigma_i(A)$, where $\lambda_i(A)$ is the i-th largest eigenvalue of A, $\lambda_{\max}(A) = \sigma_{\max}(A)$, and $\lambda_{\min}(A) = \sigma_{\min}(A)$.

Next, we will introduce the angle between two subspaces $U \in \mathbb{R}^{d\times k}$ and $X \in \mathbb{R}^{d\times k}$.

Definition 1 *Let $U \in \mathbb{R}^{d\times k}$ have orthonormal columns and $X \in \mathbb{R}^{d\times k}$ have independent columns. For $V = U^\perp$, then we have*

$$\cos\theta_k(U,X) = \min_{\|w\|=1}\frac{\|U^\top Xw\|}{\|Xw\|},\ \sin\theta_k(U,X) = \max_{\|w\|=1}\frac{\|V^\top Xw\|}{\|Xw\|},\ \textit{and}\ ,\tan\theta_k(U,X) = \max_{\|w\|=1}\frac{\|V^\top Xw\|}{\|U^\top Xw\|}. \tag{1}$$

If X is orthonormal, then it also holds that

$$\cos\theta_k(U,X) = \sigma_{\min}(U^\top X),\ \sin\theta_k(U,X) = \left\|V^\top X\right\|_2,\ \textit{and},\ \tan\theta_k(U,X) = \left\|V^\top X(U^\top X)^{-1}\right\|_2, \tag{2}$$

where $\|\cdot\|_2$ is the spectral norm and $\sigma_{\min}(X)$ is the smallest singular value of matrix X.

The above definitions can be found in the works (Hardt and Price, 2014; Golub and Van Loan, 2012).

2.2 Topology of Networks

Let $\mathbf{L}$ be the weight matrix associated with the network, indicating how agents are connected. The weight matrix $\mathbf{L}$ needs to satisfies the following properties (Shi et al., 2015; Yuan et al., 2016; Nedic and Ozdaglar, 2009):

1. $\mathbf{L}$ is symmetric with $\mathbf{L}_{i,j} \neq 0$ if and if only agents i and j are connected or $i = j$.
2. $\mathbf{0} \preceq \mathbf{L} \preceq I$, $\mathbf{L}\mathbf{1} = \mathbf{1}$, null$(I - \mathbf{L})$ = span$(\mathbf{1})$.

We use I to denote the $m \times m$ identity matrix and $\mathbf{1} = [1, \dots, 1]^\top \in \mathbb{R}^m$ denotes the vector with all ones.

There are many examples of $\mathbf{L}$ satisfying above properties: 1, $\mathbf{L} = \frac{I+P}{2}$, where P is a double stochastic matrix. 2, $\mathbf{L} = I - \frac{\widetilde{L}}{\lambda_{\max}(\widetilde{L})}$ where $\widetilde{L}$ is the Laplacian matrix of the graph. 3, $\mathbf{L} = \frac{I+M}{2}$ where M is the Metropolis weight matrix (Boyd et al., 2004).

The weight matrix has an important property that $\mathbf{L}^\infty = \frac{1}{m}\mathbf{1}\mathbf{1}^\top$ (Xiao and Boyd, 2004). Thus, one can achieve the effect of averaging local variables on different agents by multiple steps of local communications. Recently, Liu and Morse (2011) proposed a more efficient way to achieve averaging described in Algorithm 3 than the one in (Xiao and Boyd, 2004).

Proposition 2 *Let $\mathbf{W}^K \in \mathbb{R}^{d\times k\times m}$ be the output of Algorithm 3 and $\bar{W} = \frac{1}{m}\mathbf{W}^0\mathbf{1} \in \mathbb{R}^{d\times k}$. Then it holds that*

$$\bar{W} = \frac{1}{m}\mathbf{W}^K\mathbf{1}, \quad \textit{and} \quad \left\|\mathbf{W}^K - \bar{W}\otimes\mathbf{1}\right\|^2 \le \left(1 - \sqrt{1-\lambda_2(\mathbf{L})}\right)^{2K}\left\|\mathbf{W}^0 - \bar{W}\otimes\mathbf{1}\right\|^2,$$

where $\lambda_2(\mathbf{L})$ is the second largest eigenvalue of $\mathbf{L}$, and $\otimes$ denotes the tensor outer product.

3. Decentralized Exact PCA

In this section, we propose a novel decentralized exact PCA algorithm with a linear convergence rate. First, we provide the main idea behind our algorithm.

3.1 Main Idea

In previous works, the common algorithmic frame is to conduct a multi-consensus step to achieve averaging for each local power method (Raja and Bajwa, 2015; Wai et al., 2017; Kempe and McSherry, 2008), that is,

$$\begin{aligned} W_j^{t+1} =& A_j W_j^t, \\ \mathbf{W}^{t+1} =& \mathrm{MultiConsensus}(\mathbf{W}^{t+1}), \\ W_j^{t+1} =& \mathrm{QR}(W_j^{t+1}) \end{aligned} \tag{6}$$

where $\mathrm{QR}(W_j)$ computes the orthonormal basis of W_j by QR decomposition and $\mathbf{W}^t \in \mathbb{R}^{d\times k\times m}$ has its j-th slice $\mathbf{W}^t(:,:,j) = W_j^t$. However, algorithms in this framework will take increasing consensus steps to achieve high precision principal components and the consensus steps of power iterations depend on the target precision ϵ. More specifically, if one uses 'FastMix' (Algorithm 3) to achieve '$\mathrm{MultiConsensus}$', then the parameter K in 'FastMix' should depend on the target precision ϵ. More detailed results can be found in Theorem 18 in Appendix B. The above framework is similar

Algorithm 1 Decentralized Exact PCA (`DeEPCA`)

Input: Proper initial point $W^0 \in \mathbb{R}^{d\times k}$, FastMix parameter K.
Initialize $S_j^0 = W^0$, $W_j^0 = W^0$ and $A_j W_j^{(-1)} = W^0$.
for $t = 0, \dots, T$ **do**
  For each agent j, update

$$S_j^{t+1} = S_j^t + A_j W_j^t - A_j W_j^{t-1} \tag{3}$$

  Communicate S_j^{t+1} with its neighbors K times to achieve averaging, that is

$$\mathbf{S}^{t+1} = \mathrm{FastMix}(\mathbf{S}^{t+1}, K), \text{ with } \mathbf{S}^{t+1}(:,:,j) = S_j^{t+1}. \tag{4}$$

  For each agent j, compute the orthonormal basis of S_j^{t+1} by QR decomposition, that is

$$W_j^{t+1} = \mathrm{QR}(S_j^{t+1}), \quad \text{and} \quad W_j^{t+1} = \mathrm{SignAdjust}(W^{t+1}, W^0). \tag{5}$$

end for
Output: W_j^{T+1}

Algorithm 2 SignAdjust

Input: Matrices W^t and W^0 and column number k.
for $i = 1, \dots, k$ **do**
  if $\left\langle W^t(:,i), W^0(:,i)\right\rangle < 0$ **then**
    Flip the sign, that is, $W^t(:,i) = -W^t(:,i)$
  end if
end for
Output: W^t

to the well-known `DGD` algorithm in decentralized optimization which can not converges to the optima without increasing the number of communications in each multi-consensus step (Yuan et al., 2016; Nedic and Ozdaglar, 2009).

In the decentralized optimization, to overcome the weakness of `DGD`, a novel technique called 'gradient-tracking' was introduced recently (Qu and Li, 2017; Shi et al., 2015). By the advantages of the gradient-tracking, several algorithms have achieved the linear convergence rate without increasing the number of multi-consensus iterations per step. Especially, a recent work `Mudag` showed that gradient tracking can be used to achieve a near optimal communication complexity up to a $\log$ factor (Ye et al., 2020).

To obtain a decentralized exact PCA algorithm with a linear convergence rate without increasing the number of communications per consensus step, we track the subspace in the proposed PCA algorithm by adapting the gradient tracking method to 'subspace tracking'. Compared with previous decentralized PCA (Eqn. (6)) methods, we introduce an extra term S_j to track the space of power iterations. Combining S_j with multi-consensus, we can track the subspace in the power method exactly. We can then obtain the exact principal component W_j after several power iterations. The detailed description of the resulting algorithm `DeEPCA` is in Algorithm 1.

Algorithm 3 FastMix

Input: $\mathbf{W}^0 = \mathbf{W}^{-1} \in \mathbb{R}^{d\times k\times m}$, K, $\mathbf{L}$, step size $\eta_w = \frac{1-\sqrt{1-\lambda_2^2(W)}}{1+\sqrt{1-\lambda_2^2(W)}}$.
for $k = 0, \dots, K$ **do**
  $\mathbf{W}^{k+1} = (1+\eta_w)\mathbf{W}^k\mathbf{L} - \eta_w \mathbf{W}^{k-1}$;
end for
Output: $\mathbf{W}^K$.

Please note that, Algorithm 1 conducts a sign adjustment in Eqn. (5) which is necessary to make `DeEPCA` converge stably. This is because the signs of some columns of W_j^t maybe flip during the local power iterations and the sign flipping does not change the column space of the matrix. However, if some signs are flipped, then the outcome of the aggregation $\bar{W}^t = \frac{1}{m}\sum W_j^t$ will be affected.

The subspace tracking technique in our algorithm is the key to achieving the advantages of `DeEPCA`. The intuition behind the subspace tracking comes from the observation that when W_j^t and W_j^{t-1} are close to the optimal subspace U (where U is the top-k principal components of A), then $A_jW_j^t - A_jW_j^{t-1}$ is close to zero. This implies that different local subspaces S_j^{t+1} in different agents only vary by small perturbations. Thus, we only need a small number of consensus steps to make S_j^{t+1} consistent with each other. In fact, the idea behind subspace tracking has also been used in variance reduction methods for finite sum stochastic optimization algorithms (Johnson and Zhang, 2013; Defazio et al., 2014).

Using the subspace tracking, we can maintain highly consistent subspaces S_j^{t+1} in the power iteration computation $A_jW_j^t$ without increasing the number of communication rounds per consensus step. We can show that the approximation error decreases according to $\mathcal{O}(\epsilon)$ where ϵ is the error precision for the power method.

3.2 Main Result

The following lemma shows how the mean variable $\bar{S}^t = \frac{1}{m}\sum_{j=1}^m S_j^t$ converges to the top-k principal components of A and local variable S_j^t converges to its mean counterpart $\bar{S}^t$.

Lemma 3 *Matrix $A \in \mathbb{R}^{d\times d} = \frac{1}{m}\sum_{j=1}^m A_j$ is positive semi-definite with A_j being stored in j-th agent and $\|A_j\|_2 \le L$. The agents form a undirected connected graph with weighted matrix $\mathbf{L} \in \mathbb{R}^{m\times m}$. Given parameter $k \ge 1$, orthonormal matrix $U \in \mathbb{R}^{d\times k}$ is the top-k principal components of A. λ_k and λ_{k+1} are k-th and $k+1$-th largest eigenvalue of A, respectively. Suppose $\ell(\bar{S}) \triangleq \tan\theta_k(U,\bar{S})$, $\gamma = 1 - \frac{\lambda_k - \lambda_{k+1}}{2\lambda_k}$ and $\ell(\bar{S}^0) < \infty$. If $\rho = \left(1 - \sqrt{1-\lambda_2(\mathbf{L})}\right)^K$ satisfies*

$$\rho \le \min\left\{\frac{\gamma}{2}, \frac{(\lambda_k - \lambda_{k+1})(\lambda_k\lambda_{k+1} + 2L\lambda_{k+1})\cdot\gamma^2}{96kL(\sqrt{k}+1)\left(1+\gamma^{2t}\cdot\ell^2(\bar{S}^0)\right)\left(\lambda_{k+1}+2L+(\lambda_k+2L)\gamma^{t+1}\cdot\ell(\bar{S}^0)\right)^2}, \right.$$
$$\left. \frac{\lambda_k\lambda_{k+1}+2L\lambda_k}{16Lk(\sqrt{k}+1)\sqrt{m}\gamma^{t-1}\cdot\ell(\bar{S}^0)\cdot\sqrt{1+\gamma^{2t}\cdot\ell^2(\bar{S}^0)}\left(\lambda_{k+1}+2L+(\lambda_k+2L)\gamma^{t+1}\cdot\ell(\bar{S}^0)\right)}\right\}, \tag{7}$$

for $t = 1, \dots, T+1$. Letting $\bar{S}^t = \frac{1}{m}\sum_{j=1}^m S_j^t$, then sequence $\{\bar{S}^t\}_{t=0}^{T+1}$ and $\{\mathbf{S}^t\}_{t=0}^{T+1}$ generated by Algorithm 1 satisfy that

$$\ell(\bar{S}^t) \le \gamma^t \cdot \ell(\bar{S}^0) \quad \text{and} \quad \frac{1}{\sqrt{m}}\left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| \le 4\rho L(\sqrt{k+1})\gamma^{t-2}\cdot \ell(\bar{S}^0), \tag{8}$$

and

$$\frac{1}{\sqrt{m}} \cdot \left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| \le \frac{(\lambda_k - \lambda_{k+1})}{24(\lambda_{k+1} + 2L)} \cdot \gamma^t \cdot \ell(\bar{S}^0). \tag{9}$$

Remark 4 *Lemma 3 shows that our `DeEPCA` can achieve a linear convergence rate almost the same to power method, that is, Eqn.* (8) *implies that $\tan\theta_k(U, \bar{S}^t)$ converges with rate $1 - \frac{\lambda_k - \lambda_{k+1}}{2\lambda_k}$ while the power method converges with rate $1 - \frac{\lambda_k - \lambda_{k+1}}{\lambda_k}$ (Golub and Van Loan, 2012). And the number of power iteration to achieve $\tan\theta_k(U, \bar{S}^T) \le \epsilon$ is at most*

$$T = \mathcal{O}\left(\frac{\lambda_k - \lambda_{k+1}}{\lambda_k} \cdot \log\frac{1}{\epsilon}\right), \tag{10}$$

which is the same to the number required in classical power method (Golub and Van Loan, 2012). Furthermore, the difference between local variable S_j^t and its mean variable $\bar{S}^t$ will also converge to zero as iteration goes. This implies that S_j's in different agents will converge to the same subspace. Thus, we can obtain that $W_j^t = \mathrm{QR}(S_j^t)$ will converge to the top-k principal components of A.

Moreover, we can observe that the right hand of Eqn. (7) *decreases as t increases and is independent of ϵ. Hence, `DeEPCA` does* not *require to increase the consensus steps to achieve a high precision solution nor setting consensus steps for each power iteration according to ϵ which is required in previous work (Wai et al., 2017; Kempe and McSherry, 2008). Lemma 3 also reveals such a fact that `DeEPCA` does* not *require A_j to satisfy any special structure but only $A = \frac{1}{m}\sum_{j=1}^m A_j$. Thus, our `DeEPCA` can be widely applied in different settings.*

By Lemma 3, we can easily obtain the iteration and communication complexities of `DeEPCA` to achieve $\tan\theta_k(U, W_j) \le \epsilon$ for each agent-j. The communication complexity depends on the times of local communication which is presented as the product of $\mathbf{W}$ and $\mathbf{L}$ in Algorithm 3. Now we give the detailed iteration complexity and communication complexity of our algorithm in the following theorem.

Theorem 5 *Let A, U, and graph weight matrix $\mathbf{L}$ satisfy the properties in Lemma 3. The initial orthonormal matrix W^0 satisfies that $\tan\theta_k(U, W^0) < \infty$. Let parameter K satisfy*

$$K \le \frac{1}{\sqrt{1 - \lambda_2(\mathbf{L})}} \cdot \log\frac{96kL(\sqrt{k}+1)(\lambda_k + 2L)\left(1 + \tan\theta_k(U, W^0)\right)^4}{\lambda_{k+1}(\lambda_k - \lambda_{k+1}) \cdot \left(1 - \frac{\lambda_k - \lambda_{k+1}}{2\lambda_k}\right)^2}.$$

Given $\epsilon < 1$, to achieve $\tan\theta_k(U, W_j^T) \le \epsilon$ for $j = 1, \dots, m$, the iteration complexity T is at most

$$T = \frac{2\lambda_k}{\lambda_k - \lambda_{k+1}} \cdot \max\left\{\log\frac{4\tan\theta_k(U, W^0)}{\epsilon}, \log\frac{4(\lambda_k + 2L)\tan\theta_k(U, W^0)}{\sqrt{m}(\lambda_k - \lambda_{k+1})\epsilon}\right\}. \tag{11}$$

The communication complexity is at most

$$C = \frac{2\lambda_k}{(\lambda_k - \lambda_{k+1})\sqrt{1-\lambda_2(\mathbf{L})}} \cdot \max\left\{\log\frac{4\tan\theta_k(U,W^0)}{\epsilon}, \log\frac{4(\lambda_k+2L)\tan\theta_k(U,W^0)}{\sqrt{m}(\lambda_k-\lambda_{k+1})\epsilon}\right\}$$
$$\cdot \log\frac{96kL(\sqrt{k}+1)(\lambda_k+2L)\left(1+\tan\theta_k(U,W^0)\right)^4}{\lambda_{k+1}(\lambda_k-\lambda_{k+1})\cdot\left(1-\frac{\lambda_k-\lambda_{k+1}}{2\lambda_k}\right)^2}. \tag{12}$$

Furthermore, it also holds that

$$\left\|W_j^{T+1} - \frac{1}{m}\sum_{j=1}^{m} W_j^{T+1}\right\| \le \frac{\epsilon}{2}, \text{ and } \tan\theta_k(U,\bar{S}^T) \le \frac{\epsilon}{4}. \tag{13}$$

Remark 6 *Theorem 5 shows that for any agent j, W_j takes $T = \mathcal{O}\left(\frac{\lambda_k-\lambda_{k+1}}{\lambda_k}\log\frac{1}{\epsilon}\right)$ iterations to converge to the top-k principal components of A with an ϵ-suboptimality. This iteration complexity is the same to the centralized PCA based on power method (Golub and Van Loan, 2012). Furthermore, each power iteration of `DeEPCA` takes at most*

$$K = \mathcal{O}\left(\frac{1}{\sqrt{1-\lambda_2(\mathbf{L})}}\cdot\log\left(\frac{L^2}{\lambda_k\lambda_{k+1}}\cdot\frac{\lambda_k-\lambda_{k+1}}{\lambda_k}\right)\right) \tag{14}$$

consensus steps. Note that K is independent of the precision parameter ϵ which shows that `DeEPCA` does not *need to tune its consensus parameter K according to ϵ. This also implies that `DeEPCA` does* not *increase its consensus steps gradually to achieve high precision principal components. In contrast, the best known consensus steps for each power iteration of previous decentralized algorithms are (Wai et al., 2017)* [1]

$$K = \mathcal{O}\left(\frac{1}{\sqrt{1-\lambda_2(\mathbf{L})}}\log\left(\frac{L^2}{\lambda_k\lambda_{k+1}}\cdot\frac{\lambda_k-\lambda_{k+1}}{\lambda_k}\cdot\frac{1}{\epsilon}\right)\right). \tag{15}$$

Thus, `DeEPCA` achieves the best communication complexity of decentralized PCA algorithms. Comparing Eqn. (14) *and* (15)*, our result is better than the one of (Wai et al., 2017) up to* $\log\frac{1}{\epsilon}$ *factor. In fact, this advantage will become large even when ϵ is moderate large which can be observed in our experiments. Similar advantage of `EXTRA` over `DGD` in decentralized optimization makes `EXTRA` become one of most important algorithm in decentralized optimization (Shi et al., 2015).*

Furthermore, Eqn. (14) *shows that the consensus steps depend on the ratio $L^2/(\lambda_k\lambda_{k+1})$. In fact, the value $L^2/(\lambda_k\lambda_{k+1})$ reflects the data heterogeneity which can be observed more clearly when $k=1$. Due to the data heterogeneity, the multi-consensus is necessary for `DeEPCA` which will be validated in our experiments.*

Remark 7 *Lemma 3 shows that once ρ satisfies Eqn.* (7)*, $\bar{S}^t$ will converge to the top-k principal components of A linearly. That, any multi-consensus which can satisfy Eqn.* (7)*, `DeEPCA` can achieve linear convergence rate. Thus, though our analysis is based on the undirected graph, the results of `DeEPCA` can be easily extended to directed graphs, gossip models, etc.*

1. Wai et al. (2017) only provided the convergence analysis of `DePCA` for $k = 1$ and we give a detailed convergence analysis of `DePCA` with rank $k \ge 1$ in Appendix B. The result in Eqn. (15) can be obtained directly by Eqn. (40) in Appendix B.

Remark 8 *`DeEPCA` is a novel decentralized exact power method. Because the power method is the key tool in eigenvector computation and low rank approximation (SVD decomposition) (Golub and Van Loan, 2012), `DeEPCA` provides a solid foundation for developing decentralized eigenvalue decomposition, decentralized SVD, decentralized spectral analysis, etc.*

4. Convergence Analysis

In this section, we will give the detailed convergence analysis of `DeEPCA`. For notation convenience, we first introduce local and aggregate variables.

4.1 Local and Aggregate Variables

Matrix $W_j^t \in \mathbb{R}^{d\times k}$ is the local copy of the variable of W for agent j at t-th power iteration and we introduce its aggregate variable $\mathbf{W}^t \in \mathbb{R}^{d\times k\times m}$ whose j-th slice $\mathbf{W}^t(:,:,j)$ is W_j^t, that is,

$$\mathbf{W}^t(:,:,j) = W_j^t. \tag{16}$$

Furthermore, we introduce $G_j^t = A_j W_j^{t-1} \in \mathbb{R}^{d\times k}$ and tracking variable $S_j \in \mathbb{R}^{d\times k}$. We also introduce the aggregate variables $\mathbf{G}^t \in \mathbb{R}^{d\times k\times m}$ and $\mathbf{S}^t \in \mathbb{R}^{d\times k\times m}$ of G_i^t and S_i^t, respectively which satisfy

$$\mathbf{G}^t(:,:,j) = G_j^t \text{ and } \mathbf{S}^t(:,:,j) = S_j^t. \tag{17}$$

Using the local and aggregate variables, we can represent Algorithm 1 as

$$\mathbf{S}^{t+1} = \mathrm{FastMix}\left(\mathbf{S}^t + \mathbf{G}^{t+1} - \mathbf{G}^t, K\right) \tag{18}$$
$$W_j^{t+1} = \mathrm{QR}(S_j^{t+1}). \tag{19}$$

For the convergence analysis, we further introduce the mean values

$$\bar{W}^t = \frac{1}{m}\sum_{j=1}^m W_j^t, \quad \bar{G}^t = \frac{1}{m}\sum_{j=1}^m G_j^t, \quad \bar{S}^t = \frac{1}{m}\sum_{j=1}^m S_j^t, \quad \bar{H}^t = \frac{1}{m}\sum_{j=1}^m A_i\bar{W}^{t-1}, \quad \tilde{W}^t = \mathrm{QR}(\bar{S}^t). \tag{20}$$

4.2 Sketch of Proof

First, we give the relationship between $\bar{S}^t$, $\bar{G}^t$, and $\bar{H}^t$ in Lemma 9 and Lemma 10. These two lemmas show that $\bar{S}^t$ and $\bar{H}^t$ are close to each other but perturbed by $\frac{L}{\sqrt{m}}\left\|\mathbf{W}^{t-1} - \bar{W}^{t-1}\otimes\mathbf{1}\right\|$. Furthermore, by the definition of $\bar{H}^t$, we can obtain that

$$\bar{S}^{t+1} \approx \bar{H}^{t+1} = A\bar{W}^t. \tag{21}$$

If $\bar{S}^t$ is also close to S_j^t, then we can obtain that

$$\bar{W}^{t+1} \approx \mathrm{QR}(\bar{S}^{t+1}). \tag{22}$$

We can observe that Eqn. (21) and (22) are the two steps of a power iteration but with some perturbation. Based on $\bar{S}^t$, $\bar{G}^t$, and $\bar{H}^t$, we can observe that `DeEPCA` can fit into the framework of power method but with some perturbation. This is the reason why DeEPCA will converge to the top-k principal components of A.

Next, we will bound the error between local and mean variables (Defined in Section 4.1) such as $\left\|\mathbf{S}^{t+1} - \bar{S}^{t+1} \otimes \mathbf{1}\right\|$ (in Lemma 11) and $\left\|\mathbf{W}^{t+1} - \bar{W}^{t+1} \otimes \mathbf{1}\right\|$ (in Lemma 13). Lemma 11 shows that $\left\|\mathbf{S}^{t+1} - \bar{S}^{t+1} \otimes \mathbf{1}\right\|$ will decay with a rate $\rho < 1$ for each iteration but adding an extra error term $L\rho \left\|\mathbf{W}^{t} - \mathbf{W}^{t-1}\right\|$. When DeEPCA converges, then $\mathbf{W}^t$ and $\mathbf{W}^{t-1}$ will both converge to the top-k principal components, that is, $\left\|\mathbf{W}^{t} - \mathbf{W}^{t-1}\right\|$ will converge to zero (in Lemma 15). Thus, $\left\|\mathbf{S}^{t+1} - \bar{S}^{t+1} \otimes \mathbf{1}\right\|$ will also converge to zero. This implies that $\left\|\mathbf{W}^{t} - \bar{W}^{t} \otimes \mathbf{1}\right\|$ goes to zero as t increases by Lemma 13. Hence, the noisy power method described in Eqn. (21) and (22) becomes the exact power method gradually.

Finally, Lemma 14 shows that $\tan\theta_k(U, \bar{S}^t)$ converges with rate $\gamma = 1 - \frac{\lambda_k - \lambda_{k+1}}{2\lambda_k}$ when the perturbation term $\left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^{t} - \bar{S}^{t} \otimes \mathbf{1}\right\|$ is upper bounded as Eqn. (28). Combining Lemma 11, Lemma 12 and Lemma 14, we use induction in the proof of Lemma 3 to show that the assumption (28) and Eqn. (29) hold for $t = 1, \dots, T+1$ when ρ is properly chosen. This leads to the results of Lemma 3.

4.3 Main Lemmas

In our analysis, we aim to show $\tan\theta_k(U, \bar{S}^{T+1})$ and $\left\|\mathbf{S}^{T+1} - \bar{S}^{T} \otimes \mathbf{1}\right\|$ will converge to ϵ. First, we give the relationship between $\bar{S}^t$, $\bar{G}^t$, and $\bar{H}^t$. Based on $\bar{S}^t$, $\bar{G}^t$, and $\bar{H}^t$, we can observe that DeEPCA can fit into the framework of power method but with some perturbation.

Lemma 9 *Let $\bar{W}^0$, $\bar{G}^0$, and $\bar{S}^0$ be initialized as W^0. Supposing $\bar{G}^t$, and $\bar{S}^t$ be defined in Eqn.* (20) *and $\mathbf{S}^t$ update as Eqn.* (18)*, it holds that*

$$\bar{S}^{t+1} = \bar{S}^t + \bar{G}^{t+1} - \bar{G}^t = \bar{G}^{t+1}.$$

Lemma 10 *Letting $\bar{G}^t$ and $\bar{H}^t$ be defined in Eqn.* (20) *and $\|A_j\|_2 \le L$ for $j = 1, \dots, m$, they have the following properties*

$$\left\|\bar{G}^t - \bar{H}^t\right\| \le \frac{L}{\sqrt{m}} \left\|\mathbf{W}^{t-1} - \bar{W}^{t-1} \otimes \mathbf{1}\right\|. \tag{23}$$

In the next lemmas, we will bound the error between local and mean variables (Defined in Section 4.1). First, we upper bound the error $\left\|\mathbf{S}^{t+1} - \bar{S}^{t+1} \otimes \mathbf{1}\right\|$ recursively.

Lemma 11 *Letting $\mathbf{S}^t$ be updated as Eqn.* (18) *and $\|A_j\| \le L$, then $\mathbf{S}^{t+1}$ and $\bar{S}^{t+1}$ have the following properties*

$$\left\|\mathbf{S}^{t+1} - \bar{S}^{t+1} \otimes \mathbf{1}\right\| \le \rho \left\|\mathbf{S}^{t} - \bar{S}^{t} \otimes \mathbf{1}\right\| + L\rho \left\|\mathbf{W}^{t} - \mathbf{W}^{t-1}\right\|, \text{ with } \rho \triangleq \left(1 - \sqrt{1 - \lambda_2(\mathbf{L})}\right)^K. \tag{24}$$

Lemma 12 *If for $t = 0, 1, \dots, t$, it holds that $\sigma_{\min}(U^\top \tilde{W}^t) > 0$ with $\tilde{W}$ defined in Eqn.* (20) *and U being top principal components of A, then we can obtain that*

$$\sigma_{\min}(\bar{S}^{t+1}) \ge \lambda_k \cdot \frac{1}{\sqrt{1 + \ell^2(\bar{S}^t)}} - \frac{24L}{\sqrt{m}} \left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^{t} - \bar{S}^{t} \otimes \mathbf{1}\right\|. \tag{25}$$

Now, we will bound the error $\left\|\mathbf{W}^t - \bar{W}^t \otimes \mathbf{1}\right\|$.

Lemma 13 *Assuming that* $\left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\bar{S}^t - S_j^t\right\| \le \frac{1}{4}$ *for* $j = 1, \dots, m$, *where* $\left[\bar{S}^t\right]^\dagger$ *is the pseudo inverse of* $\bar{S}^t$, *then it holds that*

$$\left\|\mathbf{W}^t - \bar{W}^t \otimes \mathbf{1}\right\| \le 12 \left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\|. \tag{26}$$

Letting $\bar{S}^t = \tilde{W}^t \tilde{R}^t$ *be the QR decomposition of* $\bar{S}^t$, *then it holds that*

$$\left\|\tilde{W}^t - \bar{W}^t\right\| \le \frac{12}{\sqrt{m}} \left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\|. \tag{27}$$

Next, we will give the convergence rate of $\bar{S}^t$ under the assumption that the error between local variable S_j^t and its mean counterpart $\bar{S}^t$ is upper bounded.

Lemma 14 *Letting* $\ell(\bar{S}) \triangleq \tan\theta_k(U, \bar{S})$, $\gamma \triangleq 1 - \frac{\lambda_k - \lambda_{k+1}}{2\lambda_k}$ *and* $\frac{1}{\sqrt{m}} \cdot \left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\|$ *satisfy*

$$\frac{1}{\sqrt{m}} \cdot \left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| \le \frac{(\lambda_k - \lambda_{k+1}) \cdot \gamma^t \cdot \ell(\bar{S}^0)}{24\sqrt{1 + \gamma^{2t} \cdot \ell^2(\bar{S}^0)}\left(\lambda_{k+1} + 2L + (\lambda_k + 2L)\gamma^{t+1}\ell(\bar{S}^0)\right)}, \tag{28}$$

for $t = 0, 1, \dots, T$, *sequence* $\{\bar{S}^t\}$ *generated by Algorithm 1 satisfies*

$$\ell(\bar{S}^t) \le \gamma^{t+1} \cdot \ell(\bar{S}^0). \tag{29}$$

Finally, we will bound the difference between $\mathbf{W}^t$ and $\mathbf{W}^{t-1}$.

Lemma 15 *Letting* $\mathbf{W}$ *be defined in Eqn.* (20) *and* $\ell(\bar{S}) \triangleq \tan\theta_k(U, \bar{S})$, *then it holds that*

$$\begin{aligned}\left\|\mathbf{W}^t - \mathbf{W}^{t-1}\right\| \le & 24\left(\left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| + \left\|\left[\bar{S}^{t-1}\right]^\dagger\right\| \left\|\mathbf{S}^{t-1} - \bar{S}^{t-1} \otimes \mathbf{1}\right\|\right) \\ & + \sqrt{mk} \cdot \left(\ell(\bar{S}^t) + \ell(\bar{S}^{t-1})\right).\end{aligned} \tag{30}$$

4.4 Proof of Main Results

Using lemmas in previous subsection, we can prove Lemma 3 and Theorem 5 as follows.

Proof [Proof of Lemma 3] We prove the result by induction. When $t = 0$, Eqn. (28) holds since each agent shares the same initialization. This implies that $\ell(\bar{S}^1) \le \gamma \cdot \ell(\bar{S}^0)$.

Now, we assume that Eqn. (28) and (29) hold for $t = 0, \dots, T$. In this case, for $t = 1, \dots, T$, it holds that

$$\ell(\bar{S}^t) \le \gamma^t \cdot \ell(\bar{S}^0),$$

and

$$\frac{1}{\sqrt{m}} \cdot \left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| \le \frac{(\lambda_k - \lambda_{k+1}) \cdot \gamma^t \cdot \ell(\bar{S}^0)}{24\sqrt{1 + \gamma^{2t} \cdot \ell^2(\bar{S}^0)}\left(\lambda_{k+1} + 2L + (\lambda_k + 2L)\gamma^{t+1}\ell(\bar{S}^0)\right)}$$

$$\leq \frac{(\lambda_k - \lambda_{k+1})}{24(\lambda_{k+1} + 2L)} \cdot \gamma^t \cdot \ell(\bar{S}^0). \tag{31}$$

We will show that the result holds for $t = T+1$ and we only need to prove Eqn. (28) will hold for $t = T+1$. First, by Eqn. (24), we have

$$\begin{aligned}
&\left\|\mathbf{S}^{t+1} - \bar{S}^{t+1} \otimes \mathbf{1}\right\| \\
\overset{(24)}{\leq}& \rho \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| + \rho L \left\|\mathbf{W}^t - \mathbf{W}^{t-1}\right\| \\
\overset{(30)}{\leq}& \rho \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| + \rho L\sqrt{mk} \cdot \left(\ell(\bar{S}^t) + \ell(\bar{S}^{t-1})\right) \\
&+ 24\rho L \left(\left\| \left[\bar{S}^t\right]^\dagger \right\| \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| + \left\| \left[\bar{S}^{t-1}\right]^\dagger \right\| \left\|\mathbf{S}^{t-1} - \bar{S}^{t-1} \otimes \mathbf{1}\right\| \right) \\
\overset{(31)}{\leq}& \rho \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| + \rho L\sqrt{mk} \cdot \left(\ell(\bar{S}^t) + \ell(\bar{S}^{t-1})\right) + 24\rho L\sqrt{m} \cdot \frac{\lambda_k - \lambda_{k+1}}{24(\lambda_{k+1} + 2L)} \left(\gamma^t + \gamma^{t-1}\right) \cdot \ell(\bar{S}^0) \\
\leq& \rho \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| + 2\rho L\sqrt{m}(\sqrt{k} + 1)\gamma^{t-1} \cdot \ell(\bar{S}^0),
\end{aligned}$$

which implies that

$$\frac{1}{\sqrt{m}} \left\|\mathbf{S}^{t+1} - \bar{S}^{t+1} \otimes \mathbf{1}\right\| \leq \rho \cdot \frac{1}{\sqrt{m}} \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| + 2\rho L(\sqrt{k} + 1)\gamma^{t-1} \cdot \ell(\bar{S}^0).$$

Using above equation recursively, we can obtain that

$$\begin{aligned}
\frac{1}{\sqrt{m}} \left\|\mathbf{S}^{T+1} - \bar{S}^{T+1} \otimes \mathbf{1}\right\| \leq& \rho^{T+1} \cdot \frac{1}{\sqrt{m}} \left\|\mathbf{S}^0 - \bar{S}^0 \otimes \mathbf{1}\right\| + 2\rho L(\sqrt{k} + 1) \cdot \ell(\bar{S}^0) \sum_{i=1}^{T} \rho^{T-i}\gamma^i \\
=& 2\rho L(\sqrt{k} + 1) \cdot \ell(\bar{S}^0) \cdot \frac{\gamma^T - \gamma\rho^T}{\gamma - \rho} \\
\leq& 4\rho L(\sqrt{k} + 1)\gamma^{T-1} \cdot \ell(\bar{S}^0),
\end{aligned}$$

where the first equality is because each agent shares the same initialization and last inequality is because of the assumption that $\rho \leq \frac{\gamma}{2}$.

Furthermore, we have

$$\begin{aligned}
&\sigma_{\min}(\bar{S}^{T+1}) \\
\overset{(25)}{\geq}& \frac{\lambda_k}{\sqrt{1 + \ell^2(\bar{S}^T)}} - \frac{24L}{\sqrt{m}} \left\| \left[\bar{S}^T\right]^\dagger \right\| \left\|\mathbf{S}^T - \bar{S}^T \otimes \mathbf{1}\right\| \\
\overset{(28)}{\geq}& \frac{\lambda_k}{\sqrt{1 + \gamma^{2T} \cdot \ell^2(\bar{S}^0)}} - \frac{L(\lambda_k - \lambda_{k+1})\gamma^T \cdot \ell(\bar{S}^0)}{\sqrt{1 + \gamma^{2T} \cdot \ell^2(\bar{S}^0)}(\lambda_{k+1} + 2L + (\lambda_k + 2L)\gamma^{T+1} \cdot \ell(\bar{S}^0))} \\
=& \frac{\lambda_k\lambda_{k+1} + 2L\lambda_k + \left(\frac{\lambda_k(\lambda_k + \lambda_{k+1})}{2} + 2L\lambda_{k+1}\right)\gamma^T \cdot \ell(\bar{S}^0)}{\sqrt{1 + \gamma^{2T} \cdot \ell^2(\bar{S}^0)}(\lambda_{k+1} + 2L + (\lambda_k + 2L)\gamma^{T+1} \cdot \ell(\bar{S}^0))}.
\end{aligned} \tag{32}$$

Therefore, we can obtain that

$$\frac{1}{\sqrt{m}} \left\| \left[\bar{S}^{T+1}\right]^\dagger \right\| \left\|\mathbf{S}^{T+1} - \bar{S}^{T+1} \otimes \mathbf{1}\right\|$$

$$\leq k \cdot \frac{\sqrt{1+\gamma^{2T}\cdot\ell^2(\bar{S}^0)}(\lambda_{k+1}+2L+(\lambda_k+2L)\gamma^{T+1}\cdot\ell(\bar{S}^0))}{\lambda_k\lambda_{k+1}+2L\lambda_k+\left(\frac{\lambda_k(\lambda_k+\lambda_{k+1})}{2}+2L\lambda_{k+1}\right)\gamma^T\cdot\ell(\bar{S}^0)}\cdot 4\rho L(\sqrt{k}+1)\gamma^{T-1}\cdot\ell(\bar{S}^0).$$

First, we need to satisfy the condition in Lemma 13, that is,

$$\left\|\left[\bar{S}^{T+1}\right]^\dagger\right\|\left\|S_j^{T+1}-\bar{S}^{T+1}\right\| \leq \left\|\left[\bar{S}^{T+1}\right]^\dagger\right\|\left\|\mathbf{S}^{T+1}-\bar{S}^{T+1}\otimes\mathbf{1}\right\| \leq \frac{1}{4}.$$

Therefore, ρ only needs

$$\rho \leq \frac{1}{16Lk(\sqrt{k}+1)\sqrt{m}\gamma^{T-1}\cdot\ell(\bar{S}^0)}\cdot\frac{\lambda_k\lambda_{k+1}+2L\lambda_k+\left(\frac{\lambda_k(\lambda_k+\lambda_{k+1})}{2}+2L\lambda_{k+1}\right)\gamma^T\cdot\ell(\bar{S}^0)}{\sqrt{1+\gamma^{2T}\cdot\ell^2(\bar{S}^0)}(\lambda_{k+1}+2L+(\lambda_k+2L)\gamma^{T+1}\cdot\ell(\bar{S}^0))}.$$

In fact, we only require that

$$\rho \leq \frac{\lambda_k\lambda_{k+1}+2L\lambda_k}{16Lk(\sqrt{k}+1)\sqrt{m}\gamma^{T-1}\cdot\ell(\bar{S}^0)\cdot\sqrt{1+\gamma^{2T}\cdot\ell^2(\bar{S}^0)}\left(\lambda_{k+1}+2L+(\lambda_k+2L)\gamma^{T+1}\cdot\ell(\bar{S}^0)\right)}.$$

To satisfy Eqn. (28) for $t=T+1$, ρ only needs

$$\begin{aligned}\rho \leq & \frac{(\lambda_k-\lambda_{k+1})\cdot\gamma^2}{96kL(\sqrt{k}+1)\sqrt{1+\gamma^{2(T+1)}\cdot\ell^2(\bar{S}^0)}\left(\lambda_{k+1}+2L+(\lambda_k+2L)\gamma^{T+2}\ell(\bar{S}^0)\right)}\\ &\cdot\frac{\lambda_k\lambda_{k+1}+2L\lambda_k+\left(\frac{\lambda_k(\lambda_k+\lambda_{k+1})}{2}+2L\lambda_{k+1}\right)\gamma^T\cdot\ell(\bar{S}^0)}{\sqrt{1+\gamma^{2T}\cdot\ell^2(\bar{S}^0)}(\lambda_{k+1}+2L+(\lambda_k+2L)\gamma^{T+1}\cdot\ell(\bar{S}^0))}.\end{aligned}$$

We only require that

$$\rho \leq \frac{(\lambda_k-\lambda_{k+1})(\lambda_k\lambda_{k+1}+2L\lambda_{k+1})\cdot\gamma^2}{96kL(\sqrt{k}+1)\left(1+\gamma^{2T}\cdot\ell^2(\bar{S}^0)\right)\left(\lambda_{k+1}+2L+(\lambda_k+2L)\gamma^{T+1}\cdot\ell(\bar{S}^0)\right)^2}.$$

Since Eqn. (28) holds for $t=T+1$ when ρ satisfies the condition, then Eqn. (29) also holds for $t=T+1$. This concludes the proof. ■

Using the results of Lemma 3, we can prove Theorem 5 as follows.

Proof [Proof of Theorem 5] First, by Eqn. (26), Eqn. (9), and the condition that $T \geq \frac{2\lambda_k}{\lambda_k-\lambda_{k+1}}\log\frac{4(\lambda_k+2L)\tan\theta_k(U,W^0)}{\sqrt{m}(\lambda_k-\lambda_{k+1})\epsilon}$, we can obtain that

$$\left\|\mathbf{W}^T-\bar{W}^T\otimes\mathbf{1}\right\| \leq \sqrt{m}\cdot\frac{\lambda_k-\lambda_{k+1}}{2(\lambda_{k+1}+2L)}\cdot\gamma^T\cdot\ell(\bar{S}^0) \leq \frac{\epsilon}{2}.$$

Similarly, we can obtain that $\tan\theta_k(U,\bar{S}^T) \leq \frac{\epsilon}{4}$. Thus, we can obtain the results in Eqn. (13).

Furthermore, by the definition of angels between two subspaces, we have

$$\tan\theta_k(U,W_j^t) \overset{(1)}{=} \max_{\|w\|=1}\frac{\left\|V^\top W_j^T w\right\|}{\left\|U^\top W_j^T w\right\|}$$

$$
\begin{aligned}
&\leq \max_{\|w\|=1} \frac{\left\|V^\top \bar{W}^T w\right\| + \left\|W_j^T - \bar{W}^T\right\|}{\left\|U^\top \bar{W}^T w\right\| - \left\|W_j^T - \bar{W}^T\right\|} \\
&\leq \max_{\|w\|=1} \frac{\left\|V^\top \tilde{W}^T w\right\| + \left\|\tilde{W}^T - \bar{W}^T\right\| + \left\|W_j^T - \bar{W}^T\right\|}{\left\|U^\top \tilde{W}^T w\right\| - \left\|\tilde{W}^T - \bar{W}^T\right\| - \left\|W_j^T - \bar{W}^T\right\|} \\
&\overset{(26),(27)}{\leq} \max_{\|w\|=1} \frac{\left\|V^\top \tilde{W}^T w\right\| + 24\left\|\left[\bar{S}^T\right]^\dagger\right\| \left\|\mathbf{S}^T - \bar{S}^T \otimes \mathbf{1}\right\|}{\left\|U^\top \tilde{W}^T w\right\| - 24\left\|\left[\bar{S}^T\right]^\dagger\right\| \left\|\mathbf{S}^T - \bar{S}^T \otimes \mathbf{1}\right\|} \\
&= \frac{\tan\theta_k(U, \tilde{W}^T) + 24\left\|\left[\bar{S}^T\right]^\dagger\right\| \left\|\mathbf{S}^T - \bar{S}^T \otimes \mathbf{1}\right\| / \cos\theta_k(U, \tilde{W}^T)}{1 - 24\left\|\left[\bar{S}^T\right]^\dagger\right\| \left\|\mathbf{S}^T - \bar{S}^T \otimes \mathbf{1}\right\| / \cos\theta_k(U, \tilde{W}^T)} \\
&\overset{(8),(9)}{\leq} \frac{\gamma^T \cdot \ell(\bar{S}^0) + \sqrt{m} \cdot \frac{\lambda_k - \lambda_{k+1}}{\lambda_k + 2L} \cdot \gamma^T \cdot \ell(\bar{S}^0) \cdot \sqrt{1 + \gamma^{2T} \cdot \ell^2(\bar{S}^0)}}{1 - \sqrt{m} \cdot \frac{\lambda_k - \lambda_{k+1}}{\lambda_k + 2L} \cdot \gamma^T \cdot \ell(\bar{S}^0) \cdot \sqrt{1 + \gamma^{2T} \cdot \ell^2(\bar{S}^0)}} \\
&= \frac{\gamma^T \cdot \tan\theta_k(U, W^0) + \sqrt{m} \cdot \frac{\lambda_k - \lambda_{k+1}}{\lambda_k + 2L} \cdot \gamma^T \cdot \tan\theta_k(U, W^0) \cdot \sqrt{1 + \gamma^{2T} \cdot \tan^2\theta_k(U, W^0)}}{1 - \sqrt{m} \cdot \frac{\lambda_k - \lambda_{k+1}}{\lambda_k + 2L} \cdot \gamma^T \cdot \ell(\bar{S}^0) \cdot \sqrt{1 + \gamma^{2T} \cdot \ell^2(\bar{S}^0)}}
\end{aligned}
$$

Since $T = \frac{2\lambda_k}{\lambda_k - \lambda_{k+1}} \log \frac{4\tan\theta_k(U, W^0)}{\epsilon}$, it holds that $\gamma^T \cdot \tan\theta_k(U, W^0) \leq \frac{\epsilon}{4}$. Furthermore, when $T = \frac{2\lambda_k}{\lambda_k - \lambda_{k+1}} \log \frac{4(\lambda_k + 2L)\tan\theta_k(U, W^0)}{\sqrt{m}(\lambda_k - \lambda_{k+1})\epsilon}$, it holds that $\sqrt{m} \cdot \frac{\lambda_k - \lambda_{k+1}}{\lambda_k + 2L} \cdot \gamma^T \cdot \tan\theta_k(U, W^0) \leq \frac{\epsilon}{4}$. Thus, when $\epsilon < 1$, we can obtain that

$$
\tan\theta_k(U, W_j^T) \leq \frac{\epsilon/4 + \epsilon/4 \cdot \sqrt{1 + 1/4^2}}{1 - 1/4 \cdot \sqrt{1 + 1/4^2}} < \epsilon.
$$

Since the right hand of Eqn. 7 is monotone deceasing as t increases, ρ only satisfies that

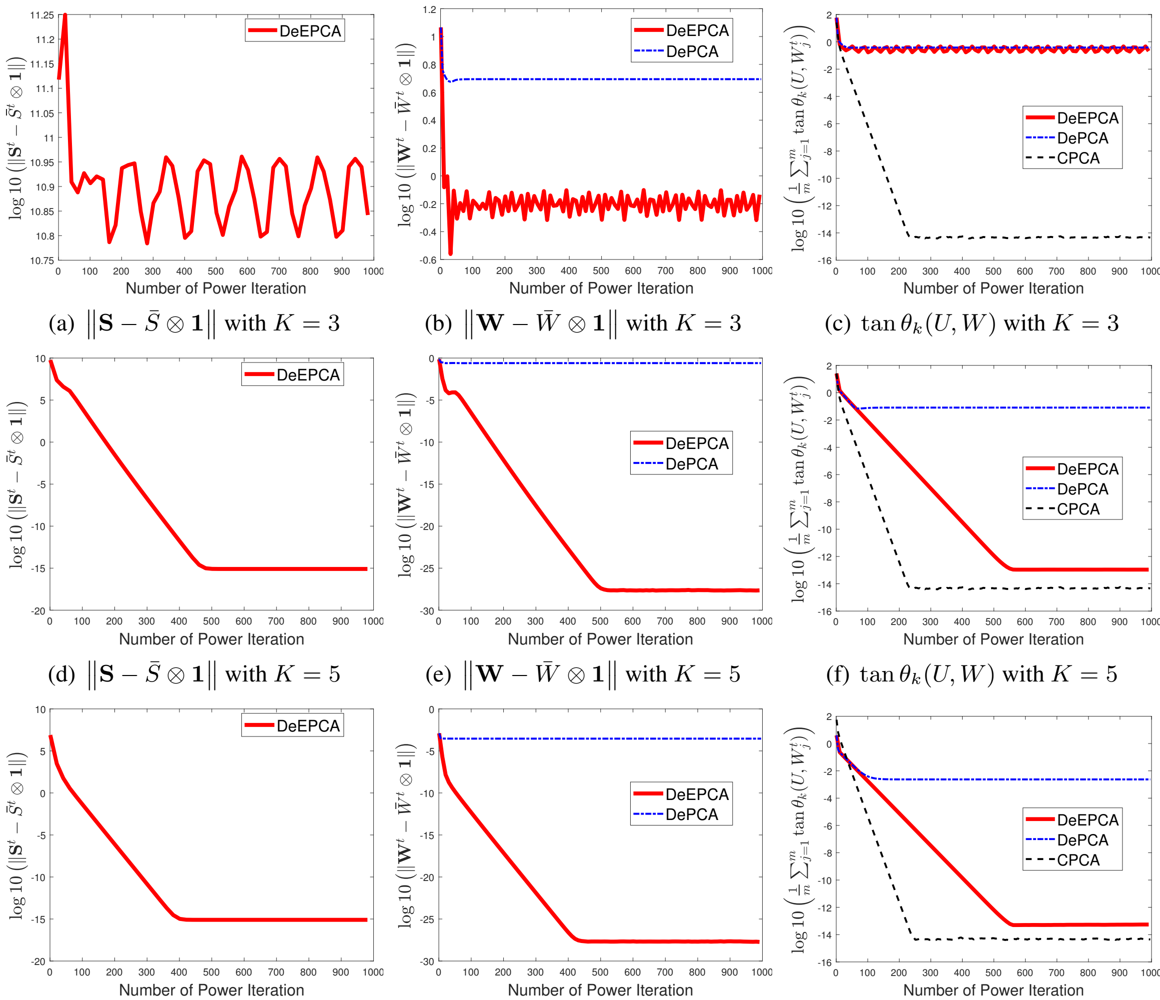
$$
\begin{aligned}
\rho \leq \min \Bigg\{ &\frac{(\lambda_k - \lambda_{k+1})(\lambda_k \lambda_{k+1} + 2L\lambda_{k+1}) \cdot \gamma^2}{96kL(\sqrt{k} + 1)\left(1 + \ell^2(\bar{S}^0)\right)\left(\lambda_{k+1} + 2L + (\lambda_k + 2L) \cdot \ell(\bar{S}^0)\right)^2}, \\
&\frac{\lambda_k \lambda_{k+1} + 2L\lambda_k}{16Lk(\sqrt{k} + 1)\sqrt{m} \cdot \ell(\bar{S}^0) \cdot \sqrt{1 + \ell^2(\bar{S}^0)}\left(\lambda_{k+1} + 2L + (\lambda_k + 2L) \cdot \ell(\bar{S}^0)\right)} \Bigg\}.
\end{aligned}
$$

Furthermore, ρ only requires to satisfy

$$
\rho \leq \frac{(\lambda_k - \lambda_{k+1})(\lambda_k \lambda_{k+1} + 2L\lambda_{k+1}) \cdot \gamma^2}{96kL(\sqrt{k} + 1)\left(1 + \ell^2(\bar{S}^0)\right)\left(\lambda_k + 2L + (\lambda_k + 2L) \cdot \ell(\bar{S}^0)\right)^2} \tag{33}
$$

Replacing the definition of $\ell(\bar{S}^0)$ and Proposition 2, we can obtain if K satisfies that

$$
K \leq \frac{1}{\sqrt{1 - \lambda_2(\mathbf{L})}} \cdot \log \frac{96kL(\sqrt{k} + 1)(\lambda_k + 2L)\left(1 + \tan\theta_k(U, W^0)\right)^4}{\lambda_{k+1}(\lambda_k - \lambda_{k+1}) \cdot \left(1 - \frac{\lambda_k - \lambda_{k+1}}{2\lambda_k}\right)^2},
$$



(a) $\|\mathbf{S} - \bar{S} \otimes \mathbf{1}\|$ with $K = 3$ (b) $\|\mathbf{W} - \bar{W} \otimes \mathbf{1}\|$ with $K = 3$ (c) $\tan\theta_k(U, W)$ with $K = 3$

(d) $\|\mathbf{S} - \bar{S} \otimes \mathbf{1}\|$ with $K = 5$ (e) $\|\mathbf{W} - \bar{W} \otimes \mathbf{1}\|$ with $K = 5$ (f) $\tan\theta_k(U, W)$ with $K = 5$

(g) $\|\mathbf{S} - \bar{S} \otimes \mathbf{1}\|$ with $K = 10$ (h) $\|\mathbf{W} - \bar{W} \otimes \mathbf{1}\|$ with $K = 10$ (i) $\tan\theta_k(U, W)$ with $K = 10$

Figure 1: Experiment on 'w8a'.

the requirement of ρ in Eqn. (33) is satisfied. Combining with iteration complexity, we can obtain the total communication complexity

$$C = T \times K = \frac{2\lambda_k}{(\lambda_k - \lambda_{k+1})\sqrt{1 - \lambda_2(\mathbf{L})}} \cdot \max\left\{\log \frac{4\tan\theta_k(U, W^0)}{\epsilon}, \log \frac{4(\lambda_k + 2L)\tan\theta_k(U, W^0)}{\sqrt{m}(\lambda_k - \lambda_{k+1})\epsilon}\right\} \cdot \log \frac{96kL(\sqrt{k}+1)(\lambda_k + 2L)\left(1 + \tan\theta_k(U, W^0)\right)^4}{\lambda_{k+1}(\lambda_k - \lambda_{k+1}) \cdot \left(1 - \frac{\lambda_k - \lambda_{k+1}}{2\lambda_k}\right)^2}.$$

■

5. Experiments

In the previous sections, we presented a theoretical analysis of our algorithm. In this section, we will provide empirical studies.

Experiment Setting In our experiments, we consider random networks where each pair of agents has a connection with a probability of $p = 0.5$. We set $\mathbf{L} = I - \frac{M}{\lambda_{\max}(M)}$ where M is the Laplacian matrix associated with a weighted graph. We set $m = 50$, that is, there exists 50 agents in this network. In our experiments, the gossip matrix $\mathbf{L}$ satisfies $1 - \lambda_2(\mathbf{L}) = 0.4563$.

We conduct experiments on the datasets 'w8a' and 'a9a' which can be downloaded in libsvm datasets. For 'w8a', we set $n = 800$ and $d = 300$. For 'a9a', we set $n = 600$ and $d = 123$. For each agent, A_j has the following form

$$A = \frac{1}{m}\sum_{j=1}^{m} A_j, \text{ and } A_j = \sum_{i=1}^{n} v_i v_i^\top, \text{ with } v_i = a_{(j-1)*n+i}, \tag{34}$$

where $a_{(j-1)*n+i} \in \mathbb{R}^d$ is the $((j-1)*n+i)$-th input vector of the dataset.

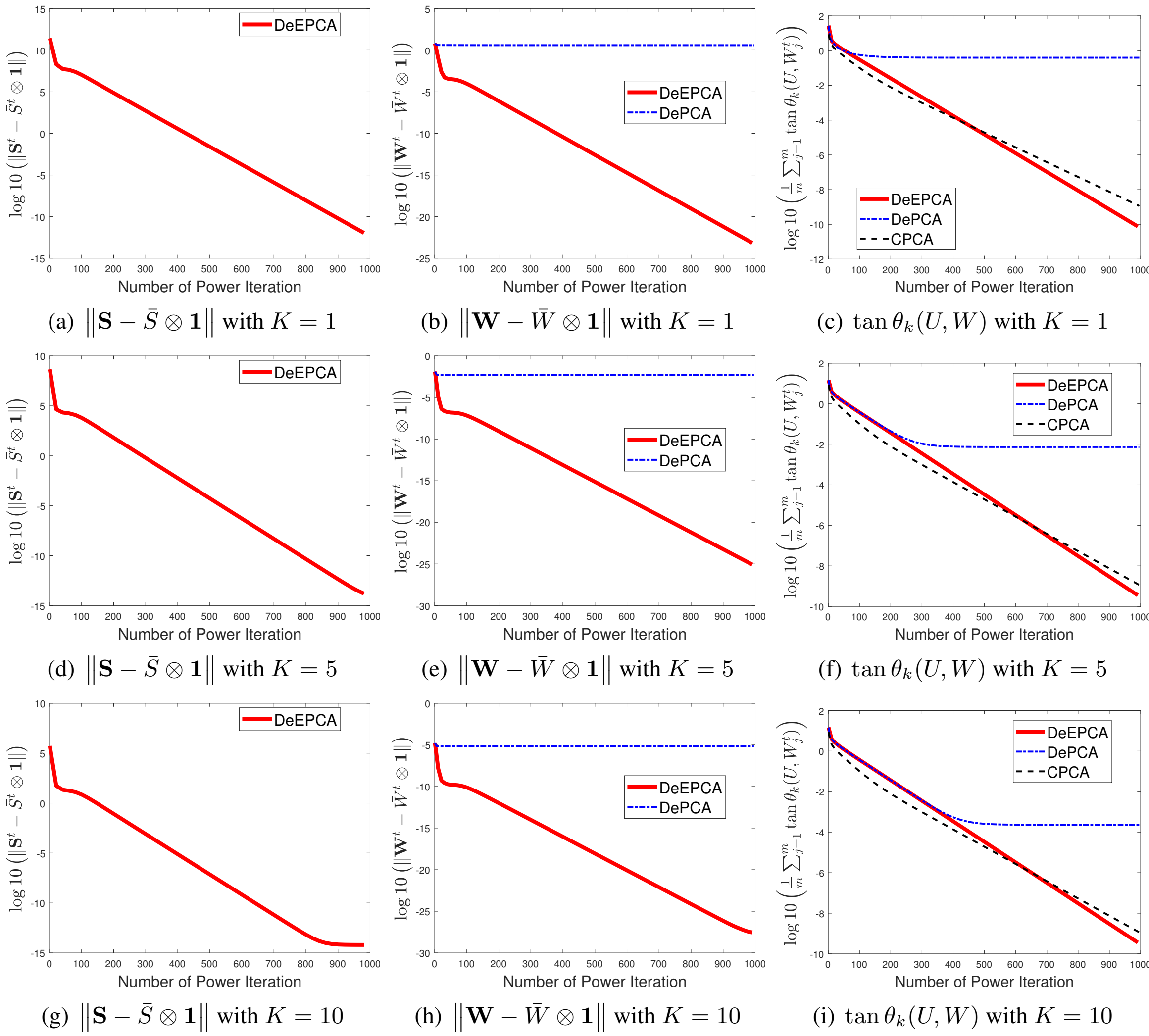
Experiment Results In our experiments, we compare `DeEPCA` with decentralized PCA (`DePCA`) (Wai et al., 2017), and centralized PCA (`CPCA`). We will study how consensus steps affect the convergence rate of `DeEPCA` empirically. Thus, we set different K's in our experiment. We will report the convergence rate of $\left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\|$, $\left\|\mathbf{W}^t - \bar{W}^t \otimes \mathbf{1}\right\|$ and $\frac{1}{m}\sum_{j=1}^{m} \tan\theta_k(U, W_j^t)$. We report experiment results in Figure 1 and Figure 2.

Figure 1 shows that multi-consensus step is required in our `DeEPCA`. When $K = 3$, `DeEPCA` can not converge to the top-k principal components of A. The number of consensus steps of `DeECPA` in each power iteration should be determined by the heterogeneity of the data just as discussed in Remark 6. Furthermore, once consensus steps of `DeECPA` are sufficient, then `DeEPCA` can achieve a fast convergence rate comparable to centralized PCA which can be observed from Figure 1 and Figure 2. This validates our convergence analysis of `DeEPCA` in Theorem 5.

Figure 1 and Figure 2 show that without increasing consensus steps, `DePCA` can not converge to the top-k principal components of A. Because of lacking subspace tracking, to achieve a high precision solution, `DePCA` can only depend on increasing consensus steps which can be observed from third columns of Figure 1 and Figure 2. Comparing `DeEPCA` and `DePCA`, we can conclude that `DeEPCA` has great advantages in communication cost.

6. Conclusion

This paper proposed a novel decentralized PCA algorithm `DeEPCA` that can achieve a linear convergence rate similar to the centralized PCA method, and the number of communications per multi-consensus step does *not* depend on the target precision ϵ. In this way, `DeEPCA` can achieve the best known communication complexity for decentralized PCA. Our experiments also verify the communication efficiency of `DeEPCA`. Although the analysis of `DeEPCA` is based on undirected graph and 'FastMix', it can be easily extended to handle directed graphs because our analysis of `DeEPCA` only requires averaging. As a final remark, we note that `DeEPCA` employs the power method, which can be applied to eigenvector finding, low rank matrix approximation, spectral analysis, etc. Therefore `DeEPCA` can be used to design communication efficient decentralized algorithms for these problems as well.



(a) $\left\|\mathbf{S} - \bar{S} \otimes \mathbf{1}\right\|$ with $K = 1$ (b) $\left\|\mathbf{W} - \bar{W} \otimes \mathbf{1}\right\|$ with $K = 1$ (c) $\tan\theta_k(U, W)$ with $K = 1$

(d) $\left\|\mathbf{S} - \bar{S} \otimes \mathbf{1}\right\|$ with $K = 5$ (e) $\left\|\mathbf{W} - \bar{W} \otimes \mathbf{1}\right\|$ with $K = 5$ (f) $\tan\theta_k(U, W)$ with $K = 5$

(g) $\left\|\mathbf{S} - \bar{S} \otimes \mathbf{1}\right\|$ with $K = 10$ (h) $\left\|\mathbf{W} - \bar{W} \otimes \mathbf{1}\right\|$ with $K = 10$ (i) $\tan\theta_k(U, W)$ with $K = 10$

Figure 2: Experiment on ‘a9a’.

Acknowledgments

We thank the anonymous reviewers for their helpful suggestions. Haishan Ye thanks to Shenzhen Research Institute of Big Data where most of this work was done. Haishan Ye has been supported by the National Natural Science Foundation of China (No. 12101491). Tong Zhang has been supported by Hong Kong GRF 16201320.

Appendix A. Proof of Lemmas in Section 4.3

We will prove our lemmas in the order of their appearance.

A.1 Proof of Lemma 9

Proof [Proof of Lemma 9] First, because the operation 'FastMix' is linear, we can obtain that

$$\bar{S}^{t+1} = \bar{S}^t + \bar{G}^{t+1} - \bar{G}^t.$$

We prove the result by induction. When $t = 0$, it holds that $\bar{S}^0 = \bar{G}^0 = W^0$. Supposing it holds that $\bar{S}^t = \bar{G}^t$, then we have

$$\bar{S}^{t+1} = \bar{S}^t + \bar{G}^{t+1} - \bar{G}^t = \bar{G}^{t+1}.$$

Thus, for each $t = 0, 1, \dots,$ it holds that $\bar{S}^t = \bar{G}^t$. ■

A.2 Proof of Lemma 10

Proof [Proof of Lemma 10] By the definition of $\bar{G}^t$ and $\bar{H}^t$ in Eqn. (20), we have

$$\begin{aligned}
\left\| \frac{1}{m}\sum_{j=1}^m A_j W_j^{t-1} - \frac{1}{m}\sum_{j=1}^m A_j \bar{W}^{t-1} \right\|^2 \leq & \frac{1}{m}\sum_{j=1}^m \left\| A_j (W_j^{t-1} - \bar{W}^{t-1}) \right\|^2 \\
\leq & \frac{1}{m}\sum_{j=1}^m \|A_j\|_2^2 \cdot \left\| W_j^{t-1} - \bar{W}^{t-1} \right\|^2 \\
\leq & \frac{L^2}{m}\sum_{j=1}^m \left\| W_j^{t-1} - \bar{W}^{t-1} \right\|^2 \\
= & \frac{L^2}{m} \left\| \mathbf{W}^{t-1} - \bar{W}^{t-1} \otimes \mathbf{1} \right\|^2,
\end{aligned}$$

where the last inequality is because of the assumption $\|A_j\|_2 \leq L$ for $j = 1, \dots, m$. ■

A.3 Proof of Lemma 11

Proof [Proof of Lemma 11] For notation convenience, we use $\mathbb{T}(\mathbf{W})$ to denote the 'FastMix' operation on $\mathbf{W}$, which is used in Algorithm 1. That is,

$$\mathbb{T}(\mathbf{W}) \triangleq \text{FastMix}(\mathbf{W}, K).$$

Then for $\mathbf{W}$, it holds that

$$\left\| \mathbb{T}(\mathbf{W}) - \bar{W} \otimes \mathbf{1} \right\| \leq \rho \cdot \left\| \mathbf{W} - \bar{W} \otimes \mathbf{1} \right\|. \tag{35}$$

It is obvious that the 'FastMix' operation $\mathbb{T}(\cdot)$ is linear. By the update rule of $\mathbf{S}^t$, we have

$$\left\| \mathbf{S}^{t+1} - \bar{S}^{t+1} \otimes \mathbf{1} \right\| \overset{(18)}{=} \left\| \mathbb{T}(\mathbf{S}^t + \mathbf{G}^{t+1} - \mathbf{G}^t) - \left(\bar{S}^{t+1} + \bar{G}^{t+1} - \bar{G}^t \right) \otimes \mathbf{1} \right\|$$

$$
\begin{aligned}
&\overset{(35)}{\leq} \rho \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| + \rho \left\|\mathbf{G}^{t+1} - \mathbf{G}^t - (\bar{G}^{t+1} - \bar{G}^t) \otimes \mathbf{1}\right\| \\
&\leq \rho \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| + \rho \left\|\mathbf{G}^{t+1} - \mathbf{G}^t\right\| \\
&= \rho \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| + \rho \sqrt{\sum_j^m \left\|A_j(W_j^t - W_j^{t-1})\right\|^2} \\
&\leq \rho \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| + L\rho \left\|\mathbf{W}^t - \mathbf{W}^{t-1}\right\|.
\end{aligned}
$$

where the second inequality is because the fact that for any $\mathbf{W} \in \mathbb{R}^{d\times k\times m}$, it holds that

$$
\begin{aligned}
\left\|\mathbf{W} - \bar{W} \otimes \mathbf{1}\right\|^2 &= \sum_{j=1}^m \left\|W_j - \frac{1}{m}\sum_{i=1}^m W_i\right\|^2 \\
&= \sum_{j=1}^m \|W_j\|^2 + \left\|\frac{1}{m}\sum_{i=1}^m W_i\right\|^2 - 2\sum_{j=1}^m \left\langle W_j, \frac{1}{m}\sum_{i=1}^m W_i\right\rangle \\
&= \sum_{j=1}^m \|W_j\|^2 - \left\|\frac{1}{m}\sum_{i=1}^m W_i\right\|^2 \\
&\leq \sum_{j=1}^m \|W_j\|^2 \\
&= \|\mathbf{W}\|^2.
\end{aligned}
$$

The last inequality is because of

$$
\sum_j^m \left\|A_j(W_j^t - W_j^{t-1})\right\|^2 \leq \sum_{j=1}^m \|A_j\|_2^2 \cdot \left\|W_j^t - W_j^{t-1}\right\|^2 \leq L^2 \sum_{j=1}^m \left\|W_j^t - W_j^{t-1}\right\| = L^2 \left\|\mathbf{W}^t - \mathbf{W}^{t-1}\right\|^2.
$$

■

A.4 Proof of Lemma 12

Proof [Proof of Lemma 12] By the definition of $\sigma_{\min}(\bar{S}^{t+1})$ and Lemma 9, we can obtain

$$
\begin{aligned}
\sigma_{\min}(\bar{S}^{t+1}) &= \sigma_{\min}(\bar{G}^{t+1}) \geq \sigma_{\min}(\bar{H}^{t+1}) - \left\|\bar{H}^{t+1} - \bar{G}^{t+1}\right\| \\
&= \sigma_{\min}(A\bar{W}^t) - \left\|\bar{H}^{t+1} - \bar{G}^{t+1}\right\| \\
&\geq \sigma_{\min}(A\tilde{W}^t) - \left\|A(\tilde{W}^t - \bar{W}^t)\right\| - \left\|\bar{H}^{t+1} - \bar{G}^{t+1}\right\| \\
&\geq \sigma_{\min}(A\tilde{W}^t) - L\left\|\tilde{W}^t - \bar{W}^t\right\| - \left\|\bar{H}^{t+1} - \bar{G}^{t+1}\right\| \\
&\overset{(23),(26),(27)}{\geq} \sigma_{\min}(A\tilde{W}^t) - \frac{24L}{\sqrt{m}}\left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\|.
\end{aligned}
$$

Furthermore, we have

$$
\sigma_{\min}(A\tilde{W}^t) = \sigma_{\min}\left(\begin{bmatrix}\Sigma_k U^\top \tilde{W}^t \\ \Sigma_{\backslash k} V^\top \tilde{W}^t\end{bmatrix}\right) \geq \sigma_{\min}\left(\begin{bmatrix}\Sigma_k U^\top \tilde{W}^t \\ \mathbf{0}\end{bmatrix}\right)
$$

$$
\begin{aligned}
&\geq \lambda_k \cdot \sigma_{\min}(U^\top \tilde{W}^t) \overset{(2)}{=} \lambda_k \cdot \cos\theta_k(U, \tilde{W}^t) \\
&= \lambda_k \cdot \frac{1}{\sqrt{1 + \ell^2(\bar{S}^t)}},
\end{aligned}
$$

where the first inequality is because of Corollary 7.3.6 of Horn and Johnson (2012) and matrix $\Sigma_k U^\top \tilde{W}^t$ is non-singular.

Therefore, we can obtain

$$
\sigma_{\min}(\bar{S}^{t+1}) \geq \lambda_k \cdot \frac{1}{\sqrt{1 + \ell^2(\bar{S}^t)}} - \frac{24L}{\sqrt{m}} \left\| \left[\bar{S}^t\right]^\dagger \right\| \left\| \mathbf{S}^t - \bar{S}^t \otimes \mathbf{1} \right\|.
$$

■

A.5 Proof of Lemma 13

First, we give a important lemma that will be used in our proof.

Lemma 16 (Theorem 3.1 of Stewart (1977)) *Let $A = QR$, where $A \in \mathbb{R}^{d\times k}$ has rank k and $Q^\top Q = I$ with I being the identity matrix. Let E satisfy $\|A^\dagger\| \|E\| < \frac{1}{2}$ where $A^\dagger$ is the pseudo inverse of A. Moreover $A + E = (Q + \Delta_Q)(R + \Delta_R)$, where $Q + \Delta_Q$ has orthogonal columns. Then it holds that*

$$
\|\Delta_Q\| \leq \frac{3\|A^\dagger\| \|E\|}{1 - 2\|A^\dagger\| \|E\|}. \tag{36}
$$

Proof [Proof of Lemma 13] For notation convenience, we will omit the superscript. Let $S_j = W_j R_j$ and $\bar{S} = \tilde{W}\tilde{R}$ be the QR decomposition of S_j and $\bar{S}$, respectively . Then we have

$$
\begin{aligned}
&\left\| \mathbf{W} - \bar{W} \otimes \mathbf{1} \right\|^2 \\
=& \sum_{j=1}^m \left\| W_j - \frac{1}{m}\sum_{i=1}^m W_i \right\|^2 \leq 2\sum_{j=1}^m \left\| W_j - \tilde{W} \right\|^2 + 2m \left\| \tilde{W} - \frac{1}{m}\sum_{i=1}^m W_i \right\|^2 \\
\leq& 4\sum_{j=1}^m \left\| W_j - \tilde{W} \right\|^2 \\
\overset{(36)}{\leq}& 4\sum_{j=1}^m \left(\frac{3\left\|\bar{S}^\dagger\right\| \left\|\bar{S} - S_j\right\|}{1 - 2\left\|\bar{S}^\dagger\right\| \left\|\bar{S} - S_j\right\|} \right)^2 \\
\leq& (12)^2 \cdot \left\| \bar{S}^\dagger \right\|^2 \left\| \mathbf{S} - \bar{S} \otimes \mathbf{1} \right\|^2,
\end{aligned}
$$

where the last inequality is because of the assumption $\left\|\bar{S}^\dagger\right\| \left\|\bar{S} - S_j\right\| \leq \frac{1}{4}$. Hence, we can obtain that

$$
\left\| \mathbf{W} - \bar{W} \otimes \mathbf{1} \right\| \leq 12 \left\| \bar{S}^\dagger \right\| \left\| \mathbf{S} - \bar{S} \otimes \mathbf{1} \right\|.
$$

■

A.6 Proof of Lemma 14

Proof [Proof of Lemma 14] By the update rule of Algorithm, we can obtain that

$$
\begin{aligned}
\ell(\bar{S}^{t+1}) &= \tan\theta_k(U, \bar{S}^{t+1}) \\
&= \max_{\|w\|=1} \frac{\left\|V^\top \bar{S}^{t+1} w\right\|}{\left\|U^\top \bar{S}^{t+1} w\right\|} = \max_{\|w\|=1} \frac{\left\|V^\top \bar{G}^{t+1} w\right\|}{\left\|U^\top \bar{G}^{t+1} w\right\|} \\
&\le \max_{\|w\|=1} \frac{\left\|V^\top \bar{H}^{t+1} w\right\| + \left\|\bar{G}^{t+1} - \bar{H}^{t+1}\right\|}{\left\|U^\top \bar{H}^{t+1} w\right\| - \left\|\bar{G}^{t+1} - \bar{H}^{t+1}\right\|} \\
&\overset{(23)}{\le} \max_{\|w\|=1} \frac{\left\|V^\top \bar{H}^{t+1} w\right\| + \frac{L}{\sqrt{m}} \left\|\mathbf{W}^t - \bar{W}^t \otimes \mathbf{1}\right\|}{\left\|U^\top \bar{H}^{t+1} w\right\| - \frac{L}{\sqrt{m}} \left\|\mathbf{W}^t - \bar{W}^t \otimes \mathbf{1}\right\|} \\
&= \max_{\|w\|=1} \frac{\left\|V^\top A \bar{W}^t w\right\| + \frac{L}{\sqrt{m}} \left\|\mathbf{W}^t - \bar{W}^t \otimes \mathbf{1}\right\|}{\left\|U^\top A \bar{W}^t w\right\| - \frac{L}{\sqrt{m}} \left\|\mathbf{W}^t - \bar{W}^t \otimes \mathbf{1}\right\|} \\
&\le \max_{\|w\|=1} \frac{\lambda_{k+1}\left\|V^\top \bar{W}^t w\right\| + \frac{L}{\sqrt{m}} \left\|\mathbf{W}^t - \bar{W}^t \otimes \mathbf{1}\right\|}{\lambda_k\left\|U^\top \bar{W}^t w\right\| - \frac{L}{\sqrt{m}} \left\|\mathbf{W}^t - \bar{W}^t \otimes \mathbf{1}\right\|} \\
&\le \max_{\|w\|=1} \frac{\lambda_{k+1}\left\|V^\top \tilde{W}^t w\right\| + \lambda_{k+1}\left\|\tilde{W}^t - \bar{W}^t\right\| + \frac{L}{\sqrt{m}} \left\|\mathbf{W}^t - \bar{W}^t \otimes \mathbf{1}\right\|}{\lambda_k\left\|U^\top \tilde{W}^t w\right\| - \lambda_k\left\|\tilde{W}^t - \bar{W}^t\right\| - \frac{L}{\sqrt{m}} \left\|\mathbf{W}^t - \bar{W}^t \otimes \mathbf{1}\right\|} \\
&\overset{(27),(26)}{\le} \max_{\|w\|=1} \frac{\lambda_{k+1}\left\|V^\top \tilde{W}^t w\right\| + \frac{12(\lambda_{k+1}+L)}{\sqrt{m}} \left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\|}{\lambda_k\left\|U^\top \tilde{W}^t w\right\| - \frac{12(\lambda_k+L)}{\sqrt{m}} \left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\|} \\
&= \max_{\|w\|=1} \frac{\lambda_{k+1}\left\|V^\top \tilde{W}^t w\right\| / \left\|U^\top \tilde{W}^t w\right\| + \frac{12(\lambda_{k+1}+L)}{\sqrt{m}} \left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| / \left\|U^\top \tilde{W}^t w\right\|}{\lambda_k - \frac{12(\lambda_k+L)}{\sqrt{m}} \left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| / \left\|U^\top \tilde{W}^t w\right\|}.
\end{aligned}
$$

Furthermore, we have

$$
\frac{1}{\left\|U^\top \tilde{W}^t w\right\|} \le \max_{\|w\|=1} \frac{1}{\left\|U^\top \tilde{W}^t w\right\|} = \frac{1}{\cos\theta_k(U, \tilde{W}^t)}.
$$

Thus, we can obtain that

$$
\begin{aligned}
\ell(\bar{S}^{t+1}) &\le \max_{\|w\|=1} \frac{\lambda_{k+1}\left\|V^\top \tilde{W}^t w\right\| / \left\|U^\top \tilde{W}^t w\right\| + \frac{12(\lambda_{k+1}+L)}{\sqrt{m}} \left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| / \cos\theta_k(U, \tilde{W}^t)}{\lambda_k - \frac{12(\lambda_k+L)}{\sqrt{m}} \left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| / \cos\theta_k(U, \tilde{W}^t)} \\
&= \frac{\lambda_{k+1}\ell(\bar{S}^t) + \frac{12(\lambda_{k+1}+L)}{\sqrt{m}} \left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| \cdot \sqrt{1+\ell^2(\bar{S}^t)}}{\lambda_k - \frac{12(\lambda_k+L)}{\sqrt{m}} \left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| \cdot \sqrt{1+\ell^2(\bar{S}^t)}},
\end{aligned}
\tag{37}
$$

where the last equality is because of the fact $1 + \tan^2\theta = \frac{1}{\cos^2\theta}$.

Now we will prove the result by induction. When $t = 0$, it holds that S_j^0's are equal to each other, that is, $\left\|\mathbf{S}^0 - \bar{S}^0 \otimes \mathbf{1}\right\| = 0$. Hence, we can obtain that

$$\ell(\bar{S}^1) \le \frac{\lambda_{k+1}}{\lambda_k}\ell(\bar{S}^0) < \left(1 - \frac{\lambda_k - \lambda_{k+1}}{2\lambda_k}\right) \cdot \ell(\bar{S}^0).$$

We assume that $\ell(\bar{S}^t) \le \gamma^t \cdot \ell(\bar{S}^0)$ and Eqn. (28) hold. Replacing the assumptions to Eqn. (37), we can obtain that

$$\ell(\bar{S}^{t+1}) \le \left(1 - \frac{\lambda_k - \lambda_{k+1}}{2\lambda_k}\right)^{t+1} \cdot \ell(\bar{S}^0) = \gamma^{t+1} \cdot \ell(\bar{S}^0).$$

This concludes the proof. ■

A.7 Proof of Lemma 15

Proof [Proof of Lemma 15] First, by triangle inequality, we can obtain

$$\begin{aligned}\left\|\mathbf{W}^t - \mathbf{W}^{t-1}\right\| \le& \left\|\mathbf{W}^t - \bar{W}^t \otimes \mathbf{1}\right\| + \left\|\mathbf{W}^{t-1} - \bar{W}^{t-1} \otimes \mathbf{1}\right\| + \left\|\bar{W}^t \otimes \mathbf{1} - \bar{W}^{t-1} \otimes \mathbf{1}\right\| \\ \overset{(26)}{\le}& 12\left(\left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| + \left\|\left[\bar{S}^{t-1}\right]^\dagger\right\| \left\|\mathbf{S}^{t-1} - \bar{S}^{t-1} \otimes \mathbf{1}\right\|\right) + \sqrt{m}\left\|\bar{W}^t - \bar{W}^{t-1}\right\|.\end{aligned}$$

Furthermore, we have

$$\begin{aligned}\left\|\bar{W}^t - \bar{W}^{t-1}\right\| \le& \left\|\bar{W}^t - U\right\| + \left\|\bar{W}^{t-1} - U\right\| \\ \le& \left\|\tilde{W}^t - U\right\| + \left\|\tilde{W}^t - \bar{W}^t\right\| + \left\|\tilde{W}^{t-1} - U\right\| + \left\|\tilde{W}^{t-1} - \bar{W}^{t-1}\right\| \\ \overset{(27)}{\le}& \left\|\tilde{W}^t - U\right\| + \left\|\tilde{W}^{t-1} - U\right\| \\ &+ \frac{12}{\sqrt{m}}\left(\left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| + \left\|\left[\bar{S}^{t-1}\right]^\dagger\right\| \left\|\mathbf{S}^{t-1} - \bar{S}^{t-1} \otimes \mathbf{1}\right\|\right).\end{aligned}$$

Now we begin to bound the value of $\left\|\tilde{W}^t - U\right\|$. Note that due to sign adjustment in Eqn. (5) in Algorithm 1, then $\mathbf{W}^t$ and $\mathbf{W}^{t-1}$ share the same direction, that is the dot product of columns of $\mathbf{W}^t$ and $\mathbf{W}^{t-1}$ are positive. Thus, we can choose such U that shares the same direction with $\mathbf{W}^t$ and $\mathbf{W}^{t-1}$. In this case, $\tilde{W}^t$ and $\tilde{W}^{t-1}$ can also share the same direction with U. Combining with the definition of $\tilde{W}$ in Eqn. (20), we have

$$\begin{aligned}\left\|\tilde{W}^t - U\right\|^2 =& \left\|\tilde{W}^t\right\|^2 + \|U\|^2 - 2\left\langle\tilde{W}^t, U\right\rangle \le 2k - 2k \cdot \sigma_{\min}(U^\top \tilde{W}^t) \\ =& 2k(1 - \cos\theta_k(\tilde{W}^t, U)) = 2k\left(1 - \frac{1}{\sqrt{1 + \ell^2(\bar{S}^t)}}\right) \\ =& 2k \cdot \frac{\sqrt{1 + \ell^2(\bar{S}^t)} - 1}{\sqrt{1 + \ell^2(\bar{S}^t)}} = 2k \cdot \frac{\ell^2(\bar{S}^t)}{\sqrt{1 + \ell^2(\bar{S}^t)}(\sqrt{1 + \ell^2(\bar{S}^t)} + 1)} \\ \le& k \cdot \ell^2(\bar{S}^t),\end{aligned}$$

where the first inequality is because $U^\top(:,i)\tilde{W}^t(:,i) > 0$ and

$$\left\langle \tilde{W}^t, U \right\rangle = \sum_{i=1}^{k} U^\top(:,i)\tilde{W}^t(:,i) \geq k \cdot \sigma_{\min}(U^\top \tilde{W}^t).$$

Therefore, we can obtain that

$$\begin{aligned}
&\left\|\mathbf{W}^t - \mathbf{W}^{t-1}\right\| \\
\leq& 12\left(\left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| + \left\|\left[\bar{S}^{t-1}\right]^\dagger\right\| \left\|\mathbf{S}^{t-1} - \bar{S}^{t-1} \otimes \mathbf{1}\right\|\right) \\
&+ 12\left(\left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| + \left\|\left[\bar{S}^{t-1}\right]^\dagger\right\| \left\|\mathbf{S}^{t-1} - \bar{S}^{t-1} \otimes \mathbf{1}\right\|\right) \\
&+ \sqrt{mk} \cdot \left(\ell(\bar{S}^t) + \ell(\bar{S}^{t-1})\right) \\
=& 24\left(\left\|\left[\bar{S}^t\right]^\dagger\right\| \left\|\mathbf{S}^t - \bar{S}^t \otimes \mathbf{1}\right\| + \left\|\left[\bar{S}^{t-1}\right]^\dagger\right\| \left\|\mathbf{S}^{t-1} - \bar{S}^{t-1} \otimes \mathbf{1}\right\|\right) + \sqrt{mk} \cdot \left(\ell(\bar{S}^t) + \ell(\bar{S}^{t-1})\right).
\end{aligned}$$

■

Appendix B. Convergence Analysis of `DePCA`

We can represent the algorithmic procedure of `DePCA` as follows:

$$\begin{aligned}
G_j^{t+1} =& A_j W_j^t \\
\mathbf{S}^{t+1} =& \text{FastMix}\left(\mathbf{G}^{t+1}, K\right) \\
W_j^{t+1} =& \text{QR}(S_j^{t+1}).
\end{aligned} \tag{38}$$

Using above representation, it is easy to check that the results in Lemma 9, Lemma 10, Lemma 12, Lemma 13, and Lemma 14 still hold for the `DePCA`. On the other hand, $\left\|\mathbf{S}^{t+1} - \bar{S}^{t+1} \otimes \mathbf{1}\right\|$ of `DePCA` has the following property. The main difference of convergence analysis between `DeEPCA` and `DePCA` lies the difference of Lemma 11 and Lemma 17.

Lemma 17 (`DePCA` version of Lemma 11) *Letting $\mathbf{S}^t$ be updated as Eqn. (38) and $\|A_j\| \leq L$, then $\mathbf{S}^{t+1}$ and $\bar{S}^{t+1}$ have the following properties*

$$\left\|\mathbf{S}^{t+1} - \bar{S}^{t+1} \otimes \mathbf{1}\right\| \leq \rho \cdot \sqrt{mk} L, \text{ with } \rho \triangleq \left(1 - \sqrt{1 - \lambda_2(\mathbf{L})}\right)^K. \tag{39}$$

Proof By the linearity of 'FastMix' and Eqn. (35), it holds that

$$\left\|\mathbf{S}^{t+1} - \bar{S}^{t+1} \otimes \mathbf{1}\right\| \leq \rho \cdot \left\|\mathbf{G}^{t+1} - \bar{G}^{t+1} \otimes \mathbf{1}\right\|.$$

Furthermore,

$$\left\|\mathbf{G}^{t+1} - \bar{G}^{t+1} \otimes \mathbf{1}\right\|^2 = \sum_{j=1}^{m} \left\|G_j^{t+1} - \frac{1}{m}\sum_{i=1}^{m} G_i^{t+1}\right\|^2$$

$$\leq \sum_{j=1}^{m} \left\| G_j^{t+1} \right\|^2 = \sum_{j=1}^{m} \left\| A_j W_j^t \right\|^2$$
$$\leq \sum_{j=1}^{m} \left\| A_j \right\|_2^2 \left\| W_j^t \right\|^2 \leq mkL^2,$$

where the last inequality is because of $\left\| A_j \right\|_2 \leq L$ and W_j^t is a othonormal matrix. Therefore,we have

$$\left\| \mathbf{S}^{t+1} - \bar{S}^{t+1} \otimes \mathbf{1} \right\| \leq \rho \cdot \sqrt{mk} L.$$

■

Theorem 18 *Matrix $A \in \mathbb{R}^{d\times d} = \frac{1}{m}\sum_{j=1}^{m} A_j$ is positive semi-definite with A_j being stored in j-th agent and $\left\| A_j \right\|_2 \leq L$. The agents form a undirected connected graph with weighted matrix $\mathbf{L} \in \mathbb{R}^{m\times m}$. Given parameter $k \geq 1$, orthonormal matrix $U \in \mathbb{R}^{d\times k}$ is the top-k principal components of A. λ_k and λ_{k+1} are k-th and $k+1$-th largest eigenvalue of A, respectively. Suppose $\ell(\bar{S}) \triangleq \tan\theta_k(U, \bar{S})$, $\gamma = 1 - \frac{\lambda_k - \lambda_{k+1}}{2\lambda_k}$ and $\ell(\bar{S}^0) < \infty$. If $\rho = \left(1 - \sqrt{1 - \lambda_2(\mathbf{L})}\right)^K$ satisfies*

$$\begin{aligned}\rho \leq \min\Bigg\{ &\frac{\gamma}{2}, \frac{\lambda_k\lambda_{k+1} + 2L\lambda_k}{4Lk\sqrt{mk}\cdot\sqrt{1+\gamma^{2T}\cdot\ell^2(\bar{S}^0)}\left(\lambda_{k+1} + 2L + (\lambda_k + 2L)\gamma^{T+1}\cdot\ell(\bar{S}^0)\right)}, \\ &\frac{(\lambda_k - \lambda_{k+1})(\lambda_k\lambda_{k+1} + 2L\lambda_{k+1})\cdot\gamma^{T+1}\cdot\ell(\bar{S}^0)}{24k\sqrt{k}L\left(1+\gamma^{2T}\cdot\ell^2(\bar{S}^0)\right)\left(\lambda_{k+1} + 2L + (\lambda_k + 2L)\gamma^{T+1}\cdot\ell(\bar{S}^0)\right)^2}\Bigg\},\end{aligned} \tag{40}$$

for $t = 1, \dots, T+1$. Letting $\bar{S}^t = \frac{1}{m}\sum_{j=1}^{m} S_j^t$, then sequence $\{\bar{S}^t\}_{t=0}^{T+1}$ generated by `DePCA` satisfies that

$$\ell(\bar{S}^{T+1}) \leq \gamma^{T+1}\cdot\ell(\bar{S}^0). \tag{41}$$

Proof We prove the result by induction. When $t = 0$, Eqn. (28) holds since each agent shares the same initialization. This implies that $\ell(\bar{S}^1) \leq \gamma\cdot\ell(\bar{S}^0)$.

Now, we assume that Eqn. (28) and (29) hold for $t = 0, \dots, T$. In this case, for $t = 1, \dots, T$, it holds that

$$\ell(\bar{S}^t) \leq \gamma^t\cdot\ell(\bar{S}^0).$$

We are going to prove Eqn. (28) and (29) hold for $t = T+1$. It is easy to check that Eqn. (32) still holds for the `DePCA` update. Combining with Eqn. (39), we have

$$\begin{aligned}&\frac{1}{\sqrt{m}}\left\| \left[\bar{S}^{T+1}\right]^\dagger \right\| \left\| \mathbf{S}^{T+1} - \bar{S}^{T+1}\otimes\mathbf{1} \right\| \\ \leq& k\cdot\frac{\sqrt{1+\gamma^{2T}\cdot\ell^2(\bar{S}^0)}(\lambda_{k+1} + 2L + (\lambda_k + 2L)\gamma^{T+1}\cdot\ell(\bar{S}^0))}{\lambda_k\lambda_{k+1} + 2L\lambda_k + \left(\frac{\lambda_k(\lambda_k+\lambda_{k+1})}{2} + 2L\lambda_{k+1}\right)\gamma^T\cdot\ell(\bar{S}^0)}\cdot\rho\sqrt{k}L.\end{aligned}$$

First, we need to satisfy the condition in Lemma 13, that is,

$$\left\| \left[\bar{S}^{T+1}\right]^{\dagger} \right\| \left\| S_j^{T+1} - \bar{S}^{T+1} \right\| \le \left\| \left[\bar{S}^{T+1}\right]^{\dagger} \right\| \left\| \mathbf{S}^{T+1} - \bar{S}^{T+1} \otimes \mathbf{1} \right\| \le \frac{1}{4}.$$

Therefore, ρ only needs

$$\rho \le \frac{1}{4Lk\sqrt{mk}} \cdot \frac{\lambda_k\lambda_{k+1} + 2L\lambda_k + \left(\frac{\lambda_k(\lambda_k+\lambda_{k+1})}{2} + 2L\lambda_{k+1}\right)\gamma^T \cdot \ell(\bar{S}^0)}{\sqrt{1+\gamma^{2T}\cdot\ell^2(\bar{S}^0)}(\lambda_{k+1} + 2L + (\lambda_k + 2L)\gamma^{T+1}\cdot\ell(\bar{S}^0))}.$$

In fact, we only require that

$$\rho \le \frac{\lambda_k\lambda_{k+1} + 2L\lambda_k}{4Lk\sqrt{mk}\cdot\sqrt{1+\gamma^{2T}\cdot\ell^2(\bar{S}^0)}\left(\lambda_{k+1} + 2L + (\lambda_k + 2L)\gamma^{T+1}\cdot\ell(\bar{S}^0)\right)}.$$

To satisfy Eqn. (28) for $t = T+1$, ρ only needs

$$\begin{aligned}\rho \le & \frac{(\lambda_k - \lambda_{k+1})\cdot\gamma^{T+1}\cdot\ell(\bar{S}^0)}{24k\sqrt{k}L\sqrt{1+\gamma^{2(T+1)}\cdot\ell^2(\bar{S}^0)}\left(\lambda_{k+1} + 2L + (\lambda_k + 2L)\gamma^{T+2}\ell(\bar{S}^0)\right)} \\ & \cdot \frac{\lambda_k\lambda_{k+1} + 2L\lambda_k + \left(\frac{\lambda_k(\lambda_k+\lambda_{k+1})}{2} + 2L\lambda_{k+1}\right)\gamma^T\cdot\ell(\bar{S}^0)}{\sqrt{1+\gamma^{2T}\cdot\ell^2(\bar{S}^0)}(\lambda_{k+1} + 2L + (\lambda_k + 2L)\gamma^{T+1}\cdot\ell(\bar{S}^0))}.\end{aligned}$$

We only require that

$$\rho \le \frac{(\lambda_k - \lambda_{k+1})(\lambda_k\lambda_{k+1} + 2L\lambda_{k+1})\cdot\gamma^{T+1}\cdot\ell(\bar{S}^0)}{24k\sqrt{k}L\left(1+\gamma^{2T}\cdot\ell^2(\bar{S}^0)\right)\left(\lambda_{k+1} + 2L + (\lambda_k + 2L)\gamma^{T+1}\cdot\ell(\bar{S}^0)\right)^2}.$$

Since Eqn. (28) holds for $t = T+1$ when ρ satisfies the condition, then Eqn. (29) also holds for $t = T+1$. This concludes the proof. ∎

References

Alexander Bertrand and Marc Moonen. Distributed adaptive estimation of covariance matrix eigenvectors in wireless sensor networks with application to distributed pca. *Signal Processing*, 104:120–135, 2014.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

Stephen Boyd, Persi Diaconis, and Lin Xiao. Fastest mixing markov chain on a graph. *SIAM review*, 46(4):667–689, 2004.

Jorge Cadima, J Orestes Cerdeira, and Manuel Minhoto. Computational aspects of algorithms for variable selection in the context of principal components. *Computational statistics & data analysis*, 47(2):225–236, 2004.

Shixiang Chen, Alfredo Garcia, Mingyi Hong, and Shahin Shahrampour. Decentralized riemannian gradient descent on the stiefel manifold. *arXiv preprint arXiv:2102.07091*, 2021.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: a fast incremental gradient method with support for non-strongly convex composite objectives. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 1*, pages 1646–1654, 2014.

Paramveer S Dhillon, Dean P Foster, and Lyle H Ungar. Eigenwords: Spectral word embeddings. *The Journal of Machine Learning Research*, 16(1):3035–3078, 2015.

Chris Ding and Xiaofeng He. K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, page 29, 2004.

Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. *Advances in Neural Information Processing Systems*, 27:2861–2869, 2014.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26:315–323, 2013.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

David Kempe and Frank McSherry. A decentralized algorithm for spectral analysis. *Journal of Computer and System Sciences*, 74(1):70–83, 2008.

Donghwan Lee, Woojoo Lee, Youngjo Lee, and Yudi Pawitan. Super-sparse principal component analyses for high-throughput genomic data. *BMC bioinformatics*, 11(1):296, 2010.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 5330–5340, 2017.

Ji Liu and A Stephen Morse. Accelerated linear iterations for distributed averaging. *Annual Reviews in Control*, 35(2):160–165, 2011.

Hyeonjoon Moon and P Jonathon Phillips. Computational and performance aspects of pca-based face-recognition algorithms. *Perception*, 30(3):303–321, 2001.

Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017.

Yongming Qu, George Ostrouchov, Nagiza Samatova, and Al Geist. Principal component analysis for dimension reduction in massive distributed data sets. In *Proceedings of IEEE International Conference on Data Mining (ICDM)*, volume 1318, page 1788, 2002.

Haroon Raja and Waheed U Bajwa. Cloud k-svd: A collaborative dictionary learning algorithm for big, distributed data. *IEEE Transactions on Signal Processing*, 64(1):173–188, 2015.

Anna Scaglione, Roberto Pagliari, and Hamid Krim. The decentralized estimation of the sample covariance. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pages 1722–1726. IEEE, 2008.

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

GW Stewart. Perturbation bounds for the qr factorization of a matrix. *SIAM Journal on Numerical Analysis*, 14(3):509–518, 1977.

Wassim Suleiman, Marius Pesavento, and Abdelhak M Zoubir. Performance analysis of the decentralized eigendecomposition and esprit algorithm. *IEEE Transactions on Signal Processing*, 64(9): 2375–2386, 2016.

Hoi-To Wai, Anna Scaglione, Jean Lafond, and Eric Moulines. Fast and privacy preserving distributed low-rank regression. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4451–4455. IEEE, 2017.

Sissi Xiaoxiao Wu, Hoi-To Wai, Lin Li, and Anna Scaglione. A review of distributed algorithms for principal component analysis. *Proceedings of the IEEE*, 106(8):1321–1340, 2018.

Lin Xiao and Stephen Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004.

Haishan Ye, Luo Luo, Ziang Zhou, and Tong Zhang. Multi-consensus decentralized accelerated gradient descent. *arXiv preprint arXiv:2005.00797*, 2020.

Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.