Privacy-Preserving Distributed SVD via Federated Power

Xiao Guo Xiaoguo@nwu.edu.cn

Center for Modern Statistics

School of Mathematics, Northwest University, Xi'an, China

Xiang Li SMSLIXIANG@PKU.EDU.CN

 $School\ of\ Mathematical\ Sciences$

Peking University, Beijing, China

Xiangyu Chang* xiangyuchang@xjtu.edu.cn

Center for Intelligent Decision-Making and Machine Learning School of Management, Xi'an Jiaotong University, Xi'an, China

Shusen Wang Shusen.wang@stevens.edu

ZHZHANG@MATH.PKU.EDU.CN

Department of Computer Science

Stevens Institute of Technology, Hoboken, USA

Zhihua Zhang

School of Mathematical Sciences
Peking University, Beijing, China

Editor:

Abstract

Singular value decomposition (SVD) is one of the most fundamental tools in machine learning and statistics. The modern machine learning community usually assumes that data come from and belong to small-scale device users. The low communication and computation power of such devices, and the possible privacy breaches of users' sensitive data make the computation of SVD challenging. Federated learning (FL) is a paradigm enabling a large number of devices to jointly learn a model in a communication-efficient way without data sharing. In the FL framework, we develop a class of algorithms called Fed-Power for the computation of partial SVD in the modern setting. Based on the well-known power method, the local devices alternate between multiple local power iterations and one global aggregation to improve communication efficiency. In the aggregation, we propose to weight each local eigenvector matrix with Orthogonal Procrustes Transformation (OPT). Considering the practical stragglers' effect, the aggregation can be fully participated or partially participated, where for the latter we propose two sampling and aggregation schemes. Further, to ensure strong privacy protection, we add Gaussian noise whenever the communication happens by adopting the notion of differential privacy (DP). We theoretically show the convergence bound for FedPower. The resulting bound is interpretable with each part corresponding to the effect of Gaussian noise, parallelization, and random sampling of devices, respectively. We also conduct experiments to demonstrate the merits of Fed-Power. In particular, the local iterations not only improve communication efficiency but also reduce the chance of privacy breaches.

Keywords: Communication Efficiency, Federated Learning, Power Method, Stragglers' Effect.

^{*.} Xiao Guo and Xiang Li make an equal contribution to this paper. Xiangyu Chang is the corresponding author

1. Introduction

Modern machine learning tasks involve massive data that come from small-scale devices, such as mobile phones, smartwatches, power metering, etc. The computation and communication power of these devices is limited, which makes large-scale data applications challenging. Further, the data from peripheral devices often contain sensitive information, hence, privacy issues become more and more prominent (Bhowmick et al., 2018; Dwork et al., 2014a).

Federated learning (FL) has become a prevalent paradigm of distributed learning for large-scale problems involving user-level data; see Kairouz et al. (2019) and references therein. Typically, in the FL, user devices train the model locally and send updates to the central server whenever communications are required. The server aggregates the updates (maybe randomly) and then sends them back to the devices. These procedures are repeated until convergence or attaining proper conditions. FL confronts the aforementioned challenges coming from small-scale devices, including large-scale data and unreliable communication, autonomy, and privacy issues; see McMahan et al. (2017); Smith et al. (2017); Sattler et al. (2019); Li et al. (2020a), among others. First, the FL algorithms are communication efficient which requires more local computation and fewer communications. Second, it can deal with the scheme when the users' devices are inactive or the users decide not to participate in the following training procedures. Third, since the training data cannot be moved away from its device, privacy is preserved to some extent.

Nevertheless, even if only the updates but the original data are transmitted to the central server, the individuals' privacy could be easily compromised via delicately designed attacks (Dwork et al., 2017; Melis et al., 2018; Zhou and Tang, 2020). Differential privacy (Dwork et al., 2006, 2014a) is a well-adopted notion for private data analysis. A differentially private algorithm pursues that if the data is changed by one row (entry) with pre-specified limits, then the algorithm's output appears similar in probability. Such algorithms protect the users' privacy from any adversary who knows the algorithm's output and even the rest of the data and can resist kind of attack. Commonly, a differentially private algorithm is obtained by adding calibrated noise to the non-differentially private algorithm.

In this paper, we consider the problem of partial singular value decomposition (SVD) in the private-preserving federated learning regime. SVD is the fundamental problem in machine learning and statistics with applications in dimension reduction (Wold et al., 1987), clustering (Von Luxburg, 2007), and matrix completion (Candès and Recht, 2009), among others. The modern computation of SVD can be traced back to 1960s, when the seminal works of Golub and Kahan (1965); Golub and Reinsch (1970) provided the basis for the EISPACK and LAPACK routines. For computing partial SVD of matrices, iterative algorithms such as the power iterations and its variants (Golub and Van Loan, 2012; Hardt and Price, 2014) flourished. Very recently, to solve large-scale problems, distributed learning of SVD or principle components is receiving more and more attention; see Fan et al. (2019b); Chen et al. (2020), among others. However, as far as we are aware, existing works can not meet the challenge that the FL confronts and the privacy concerns simultaneously.

To handle the computation, communication, and privacy challenges that modern data analysis calls for, we propose an algorithm called *Federated Power* method (FedPower). Based on the well-known single-machine power method, the FedPower assumes that the

data is distributed across different devices but never leaves the devices. Each device locally performs the power iterations using its data. After several local steps, the devices send their updates to the central server, and the server aggregates them and sends the result back to local devices. In the aggregation, we use *Orthogonal Procrustes Transformation* (OPT) to post-process the output matrices of the *m* nodes after each iteration so that the *m* matrices are close to each other. Because each device may lose connection to the server actively or passively during the training process, we provide two different protocols, namely, the *full participation* and the *partial participation*. In the partial participation protocol, the server can collect the first few responded devices within a certain time range. Moreover, to alleviate the privacy leakage, we take advantage of the notion of DP to add Gaussian noise to the updates whenever the communications happen. This is based on our assumption that the server is *honest-but-curious* (*semi-honest*), and the devices are *honest*.

Compared to existing works, the FedPower enjoys the following three benefits simultaneously. First, it can handle massive data distributed across local devices and owned by local users. Second, it is communication efficient and can preserve privacy in the sense of DP. Note that the local updates not only improve the communication efficiency but also reduce the possibility of a privacy breach.

With the algorithms at hand, we also study how the FedPower performs theoretically. First, we rigorously prove that the algorithms corresponding to the full and partial participation schemes are differentially private. Second, we analyze the convergence bound of FedPower in terms of the subspace distance between the estimated and the true singular vectors. For the full participation scheme, it turns out that the convergence error consists of two parts, one is induced by the Gaussian noise that is added to preserve privacy, and the other is induced by the parallelization and synchronization. For the partial participation scheme, we consider two random sampling and aggregating schemes. The resulting convergence error bound consists of three parts. Besides the two parts appearing in the error of the full participation scheme, there exists an additional part that comes from the sampling of local devices, which could be regarded as the sampling bias term. The more devices that are sampled, the smaller the bias term would be. As expected, the error bounds can be sufficiently small if the sample size is large enough and the quality of local data is good enough.

The remainder of the paper is organized as follows. Section 2 introduces the basic definitions and some typical algorithms for SVD, distributed power method, and DP. Section 3 includes the proposed algorithms FedPower and the corresponding convergence analysis under two schemes, namely, the full participation and the partial participation. Section 4 reviews and discusses the related works, and also summaries the main contributions of the current work. Section 5 presents the experimental results. Section 6 concludes the paper. Technical proofs and supplementary materials are all included in the Appendix.

2. Preliminaries

In this section, we first present SVD, the power method for computing the partial SVD, and a naive distributed power method designed for the distributed SVD computation. Then, we pose the challenges, namely, the communication and the privacy, that the modern machine learning tasks call for. In particular, we explain why the autonomy for each local device

is needed and propose two adversary models to show how the privacy can be leaked in the current distributed power method. Last, we present the framework of DP (Dwork et al., 2006) and its basic properties.

2.1 SVD and Power Method

Given a data matrix $A \in \mathbb{R}^{n \times d}$ with assumption of $d \leq n$, its full SVD is defined as

$$A = U\Lambda V^{\mathsf{T}} = \sum_{i=1}^{d} \lambda_i u_i v_i^{\mathsf{T}},$$

where $U = [u_1, u_2, \ldots, u_d] \in \mathbb{R}^{n \times d}$ and $V = [v_1, v_2, \ldots, v_d] \in \mathbb{R}^{n \times d}$ are column orthogonal matrices that contain the left and right singular vectors of A, respectively, and $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_d\} \in \mathbb{R}^{d \times d}$ is a diagonal matrix with the singular values in decreasing order on the diagonal. The partial or truncated SVD aims to compute the top k ($k \leq d$) singular vectors $U_k = [u_1, \ldots, u_k]$ and $V_k = [v_1, \ldots, v_k]$, and use the truncated decomposition $U_k \Lambda_k V_k^{\mathsf{T}}$ to approximate A, where $\Lambda_k = \text{diag}\{\lambda_1, \ldots, \lambda_k\} \in \mathbb{R}^{k \times k}$. SVD is one of the most commonly used techniques for various machine learning tasks including dimension reduction (Wold et al., 1987), clustering (Von Luxburg, 2007), ranking (Negahban et al., 2017), matrix completion (Candès and Recht, 2009), multiple testing (Fan et al., 2019a), factor analysis (Bai and Ng, 2013), among others, and it also has applications in many disciplines, such as finance, biology, and neurosciences (Izenman, 2008).

Let $M = \frac{1}{n}A^{\intercal}A \in \mathbb{R}^{d \times d}$. The power method (Golub and Van Loan, 2012) computes V_k , namely, the top k right singular vectors of A and also the top k eigenvectors of M, by iterating

$$Y \leftarrow MZ$$
 and $Z \leftarrow \text{orth}(Y)$, (2.1)

where Y and Z are $d \times k$ matrices, and orth(Y) stands for orthogonalizing the columns of Y via the QR-factorization.

When n is so large that one computer can not preserve all the samples, the power method is blocked. It becomes beneficial to partition the data and compute the power iterations in parallel, which calls for the distributed power methods introduced next.

2.2 Distributed Power Method

Suppose A is partitioned to m blocks by row such that $A^{\intercal} = [A_1^{\intercal}, ..., A_m^{\intercal}]$, where $A_i \in \mathbb{R}^{s_i \times d}$ includes s_i rows of A and $\sum_{i=1}^m s_i = n$. See Figure 1(a) for illustration. Let $M_i = \frac{1}{s_i} A_i^{\intercal} A_i \in \mathbb{R}^{d \times d}$. We can then see that

$$M = \frac{1}{n} A^{\mathsf{T}} A = \sum_{i=1}^{m} \frac{1}{n} A_i^{\mathsf{T}} A_i = \sum_{i=1}^{m} \frac{s_i}{n} M_i = \sum_{i=1}^{m} p_i M_i, \tag{2.2}$$

where $p_i = \frac{s_i}{n}$. Thereby, Y in (2.1) can be written as

$$Y = \sum_{i=1}^{m} \frac{s_i}{n} M_i Z = \sum_{i=1}^{m} p_i M_i Z \in \mathbb{R}^{d \times k}, \tag{2.3}$$

which indicates that the power method can be parallelized. See Figure 1(b) and Algorithm 1, which is called the distributed power method. Note that Algorithm 1 is identical to the power method except that the summations in computing M comes from different workers rather than a single machine. The following theorem is a well-known result on the convergence of the power method as well as the distributed power method (Arbenz, 2012).

Theorem 1 Let σ_k be the k-th largest singular value of M and assume $\sigma_{k+1} > 0$, where $1 \le k < d$. Then for any $\epsilon > 0$, with high probability, after $T = O(\frac{\sigma_k}{\sigma_{k+1}} \log(\frac{d}{\epsilon}))$ iterations, the output Z_T of Algorithm 1 satisfies

$$\sin\theta_k(Z_T, V_k) = \|(\mathbb{I}_d - Z_T Z_T^{\mathsf{T}}) V_k\|_2 \le \epsilon,$$

where $\sin \theta_k$ denotes the k-th principle angles between two subspaces which can be regarded as a subspace distance ¹.

The distributed power method can handle data that are distributed across different workers. Yet, it still can not adapt to the problems or concerns faced in modern data applications. First, Algorithm 1 involves two rounds of communications in every iteration. Hence, simple parallelization of the power method brings large communication costs. In addition, the autonomous effect or the straggler's effect and the privacy issue remain unsolved. We will illustrate the latter two in more detail in the next two subsections, respectively.

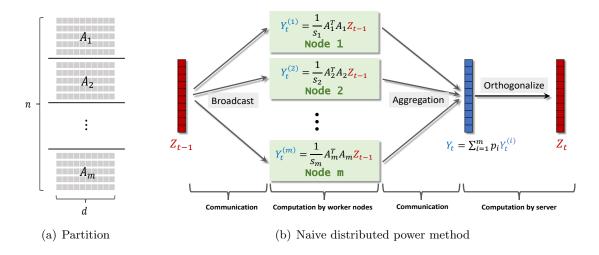


Figure 1: (a) The $n \times d$ data matrix A is partitioned among m worker nodes. (b) In every iteration of the distributed power iteration, there are two rounds of communications. Most of the computations are performed by the worker nodes.

2.3 Straggler's Effect

Different from traditional distributed learning, the FL system often consists of one central server and massive small-scale devices. The local devices are not controlled by the server,

^{1.} The formal definition can be found in Section 2.6.

Algorithm 1 Distributed Power Method

- 1: **Input:** distributed dataset $\{A_i\}_{i=1}^m$, target rank k, iteration rank $r \geq k$, number of iterations
- 2: **Initialization:** orthonormal $Z_0^{(i)}=Z_0\in\mathbb{R}^{d\times r}$ by QR decomposition on a random Gaussian matrix.
- 3: for t = 1 to T do

- The *i*-th worker independently performs $Y_t^{(i)} = M_i Z_{t-1}^{(i)}$ for all $i \in [m]$, where $M_i = \frac{A_i^{\mathsf{T}} A_i}{s_i}$; Each worker *i* sends $Y_t^{(i)}$ to the server and the server performs aggregation: $Y_t = \sum_{i=1}^m p_i Y_t^{(i)}$; The server performs orthogonalization: $Z_t = \mathsf{orth}(Y_t)$ and broadcast Z_t to each worker such that $Z_t^{(i)} = Z_t$;
- 8: Output: approximated eigen-space $Z_T \in \mathbb{R}^{d \times r}$ with orthonormal columns.

instead, the server may lose connection to the devices (Kairouz et al., 2019; Li et al., 2020b). On the one hand, some devices may become stragglers if they are powered off, get broken, or have a poor Internet connection. On the other hand, each device has its autonomy. The owners of some devices may choose not to participate in the following training procedures due to personal reasons. Hence, it is infeasible for the server to wait for all the devices' responses. Instead, the server could just use the updates of the first several responded devices before a pre-specified time. Our proposed algorithms will remedy these issues.

2.4 Adversary Model

Though the data are not shared by all the participants in FL or more general distributed learning systems, privacy threats still exist. To see how the data privacy can be breached, we next consider two types of potential attackers, respectively, as in Bhowmick et al. (2018); Zhou and Tang (2020).

The first is a *curious onlooker* who may eavesdrop on the communication between the server and the devices and know the learning tasks and rules. In particular, the server could be such an onlooker, and it is termed honest-but-curious (semi-honest), meaning that the server does not violate the protocol to attack the raw data but it is curious and will attempt to learn all possible information from its received messages (Goldreich, 2009). For example, the internal employees of an app company who are responsible for fitting models would want to infer the personal information of its users. Meanwhile, we assume that each local device is honest and would not infer information from each other. The following examples show that how the distributed power method (Algorithm 1) may cause privacy threats via curious onlookers.

Example 1 (Privacy breaches via curious onlookers) Recall Algorithm 1 and suppose the server knows the updates $Y_t^{(i)}$ for each $i \in [m]$ and $t \in T$. In addition, the server could deduce $Z_t^{(i)}$ from $Y_t^{(i)}$ because it also knows the learning rule that $Z_t^{(i)}$ is obtained from $Y_t^{(i)}$ via the QR decomposition (Line 6 in Algorithm 1). Then by

$$Y_t^{(i)} = M_i Z_{t-1}^{(i)}$$
 (Line 4 in Algorithm 1), (2.4)

and note that $M_i \in \mathbb{R}^{d \times d}$, one can easily infer all the elements of M_i using enough $Z_t^{(i)}$'s and $Y_t^{(i)}$'s. Specifically, $|T| \geq d^2 + 1$ suffices. Moreover, if the adversary knows external information of M_i , say some entries of M_i , then |T| could be further reduced.

The second kind of attackers that we consider is an *external adversary* who knows the final published results and additional prior information about individuals. For example, people who participate in the data collection and model design may be such kind of attackers. The next example illustrates why an external adversary could lead to privacy leakage.

Example 2 (Privacy breaches via external adversaries) Recall the following SVD approximation,

$$M = \frac{1}{n} V \Lambda^2 V^{\mathsf{T}} \approx \frac{1}{n} V_k \Lambda_k^2 V_k^{\mathsf{T}},\tag{2.5}$$

and suppose the output of Algorithm 1 which is known by the adversary is a good estimator for V_k . In addition, the external adversary often knows additional information in M, say P entries of M. Then, the unknown parameters in (2.5) include unknown entries in $M \in \mathbb{R}^{d \times d}$ and all entries in $\Lambda_k = \text{diag}\{\lambda_1, ..., \lambda_k\}$, with $d^2 - P + k$ parameters in total. Noting that there are d^2 linear (approximated) equations in (2.5), the unknown entries in M can be (approximately) recovered if $d^2 \geq d^2 - P + k$, namely, $P \geq k$, which is easy to attain because the target rank k is often small.

The aforementioned examples indicate that Algorithm 1 can be attacked readily in the two scenarios. Thus, we introduce the differential privacy scheme for aiding the distributed power method to overcome the privacy issue.

2.5 Differential Privacy

We have seen that privacy concern is prevailing in modern data analysis. How to quantitatively describe privacy is the key point to understand and design privacy-preserving algorithms. Differential privacy (DP), first introduced in Dwork et al. (2006), is a rigorous and most widely adopted notion of privacy, which generally guarantees that a randomized algorithm behaves similarly on similar input databases. The (ε, δ) -DP (Dwork et al., 2014a) is defined as follows. With a slight abuse of notation, we use DP to abbreviate "differential privacy" or "differentially private" throughout the paper.

Definition 1 ((ε, δ) -**DP**) A randomized algorithm $\mathcal{M}: \mathcal{X}^n \to \Theta$ is called (ε, δ) -DP if for all pairs of neighboring databases $X, X' \in \mathcal{X}^n$, and for all subsets of range $S \subseteq \Theta$:

$$\mathbb{P}(\mathcal{M}(X) \in S) \le \exp(\varepsilon)\mathbb{P}(\mathcal{M}(X') \in S) + \delta.$$

The definition of neighboring databases varies with contexts. In general, X and X' are called neighboring databases if they are the same except one single entry or one row which may contain the information of one individual. DP achieves the privacy goal that anything can be learned about an individual from the released information can also be learned without that individual's participation. The ε is often called the privacy budget which is a small constant measuring the privacy loss and it should be no larger than 1 typically (Dwork

et al., 2014a). But in some realistic settings where the adversary has limited information (Bhowmick et al., 2018) or the given problem is hard to attack, it could be large. The δ is also a small constant and it can be thought of as a tolerance of the more stringent ε -DP, i.e., (ε, δ) -DP with δ being 0.

DP is a strong notion that can protect against arbitrary risks, including the reconstruction and tracing attacks, among others (Dwork et al., 2014a). Roughly speaking, it introduces more noise than that is required for the success of attacks. Yet, it is not without cost. To achieve DP, the algorithm's accuracy is sacrificed via adding certain noise. The following Gaussian mechanism (Dwork et al., 2014a) provides a concrete example.

Definition 2 (Gaussian mechanism) For any algorithm $\mathcal{M}: \mathcal{X}^n \to \mathbb{R}^d$, the L^p -sensitivity of \mathcal{M} is defined as

$$\triangle_p(\mathcal{M}) = \sup_{X,X' \text{ neighboring}} ||\mathcal{M}(X) - \mathcal{M}(X')||_p, \text{ for } p \ge 1.$$

If $\triangle_2(f) < \infty$, then the Gaussian mechanism given by

$$\mathcal{M}(X,\varepsilon) := \mathcal{M}(X) + (\xi_1,\xi_2,\ldots,\xi_d)^{\mathsf{T}},$$

where the ξ_i are i.i.d. drawn from $\mathcal{N}(0, 2(\triangle_2(\mathcal{M})/\varepsilon)^2\log(1.25/\delta))$, achieves (ε, δ) -differential privacy.

In real applications, the algorithm is often complicated. The common strategy of achieving DP is to divide the algorithm into several parts and manipulate each part respectively. The following two properties of DP are useful (Dwork et al., 2014a). One is the post-processing property, meaning that an (ε, δ) -DP algorithm is still (ε, δ) -DP after any post-processing provided that no additional knowledge about the database is used. The other is the composition property, saying that repeated queries will amplify the privacy leakage.

Proposition 2 (Post-processing) Let $\mathcal{M}: \mathcal{X}^n \to \Theta$ be an (ε, δ) -DP algorithm, and $g: \Theta \to \Theta'$ be an arbitrary randomized mapping. Then $g \circ \mathcal{M}: \mathcal{X}^n \to \Theta'$ is (ε, δ) -DP.

Proposition 3 (Composition) Let $\mathcal{M}_i: \mathcal{X}^n \to \Theta_i$ be an $(\varepsilon_i, \delta_i)$ -DP algorithm for $i \in [k]$. If $\mathcal{M}_{[k]}: \mathcal{X}^n \to \prod_{i=1}^k \Theta_i$ is defined as $\mathcal{M}_{[k]}(x) = (\mathcal{M}_1(x), ..., \mathcal{M}_k(x))$, then $\mathcal{M}_{[k]}$ is $(\sum_{i=1}^k \varepsilon_i, \sum_{i=1}^k \delta_i)$ -DP.

These two propositions will be used throughout this paper.

2.6 Notation

We summarize the notation and notions used in the following parts of this paper. Given a target matrix $A \in \mathbb{R}^{n \times d}$, the k-th largest singular value of A is denoted by λ_k . The matrix A is divided into m partitions by row with the i-th partition (local device) including s_i rows, and $\sum_{i=1}^m s_i = n$. Accordingly, $p_i = \frac{s_i}{n}$ denotes the fraction of rows in the ith partition. For $M = \frac{1}{n}A^{\mathsf{T}}A$ (recall Eq.(2.2)), let $\kappa = \|M\|_2 \|M^{\mathsf{T}}\|_2$ denote its condition number. σ_k denotes the k-th largest singular value of M. It is easy to see that $\sigma_k = \lambda_k^2/n$. Let k and r ($r \geq k$) be the target rank and iteration rank of partial SVD, respectively. The numbers

of total iterations is denoted by T. Let [T] denote the set $\{1,\ldots,T\}$. $\|\cdot\|_2$ denotes the spectral norm of a matrix or the Euclidean norm of a vector, $\|\cdot\|_{\max}$ denotes the entry-wise maximum absolute value of a matrix or a vector, $\|\cdot\|_{\infty}$ denotes the matrix operator ℓ_{∞} norm, and $\|\cdot\|_{\mathrm{m}}$ denotes the minimum singular value of a matrix. \mathcal{O}_r denotes the set of $r \times r$ orthogonal matrices and \mathbb{I}_r denotes the identity matrix with dimension r.

In addition, we use the following standard notation for asymptotics. We write $f(n) \approx g(n)$ if $cg(n) \leq f(n) \leq Cg(n)$ for some constants $0 < c < C < \infty$. $f(n) \lesssim g(n)$ or f(n) = O(g(n)) if $f(n) \leq Cg(n)$ for some constant $C < \infty$. $f(n) = \Omega(g(n))$ if $f(n) \geq cg(n)$ for some constant c > 0. Finally, we provide the definition of projection distance, which measures the distance of two subspaces.

Definition 3 (Projection distance) Given two column-orthonormal matrices $U, \tilde{U} \in \mathbb{R}^{d \times k}$, the projection distance between the two subspaces spanned by their columns is defined as

$$\operatorname{dist}(U, \tilde{U}) := \|UU^{\mathsf{T}} - \tilde{U}\tilde{U}^{\mathsf{T}}\|_{2} = \|\tilde{U}^{\mathsf{T}}U^{\perp}\|_{2} = \|U^{\mathsf{T}}\tilde{U}^{\perp}\|_{2} = \sin\theta_{k}(U, \tilde{U}), \tag{2.6}$$

where $U^{\perp}, \tilde{U}^{\perp}$ denote the complement subspaces of U, \tilde{U} , respectively. And θ_k denotes the k-th principle angle between two subspaces; see Appendix E for the formal definition.

3. Privacy-Preserving Distributed SVD

In this section, we develop a set of power-iteration-based algorithms, called FedPower, for computing SVD which could simultaneously handle the computation, communication, straggler, and privacy issues that distributed machine learning tasks involve. Specifically, we will respectively study two protocols, namely, the full participation and the partial participation for conquering the straggler's effect. This section establishes privacy guarantees and convergence rates.

Before going to the details, we here illustrate the basic idea of FedPower, whose structure is shown in Figure 2. For improving the communication efficiency of the distributed power method, FedPower trades more local computations for fewer communications. More specifically, every worker runs

$$Y_t^{(i)} = M_i Z_{t-1}^{(i)}$$
 (Line 4 in Algorithm 1),

multiple times locally between two communications. Let T be the number of local computations performed by every worker. Let \mathcal{I}_T , a subset of [T], index the iterations that perform communications. If $\mathcal{I}_T = [T]$, synchronization happens at every iteration as in the distributed power method (see Figure 1). If $\mathcal{I}_T = \{T\}$, synchronization happens only at the end, and FedPower is similar to the one-shot divide-and-conquer SVD (Fan et al., 2019b). The cardinality $|\mathcal{I}_T|$ is the total number of synchronizations. An important example that we will focus on latter is \mathcal{I}_T^p . It is defined by

$$\mathcal{I}_{T}^{p} = \{ t \in [T] : t \mod p = 0 \} = \{ 0, p, 2p, \cdots, p | T/p | \}, \tag{3.1}$$

where p is a positive integer and $\lfloor T/p \rfloor$ is the largest integer which is smaller than T/p. FedPower with \mathcal{I}_T^p only performs communications every p iterations.

To improve algorithms' performance, when communication happens (i.e. $t \in \mathcal{I}_T^p$), orthogonal transformed $Y_t^{(i)}$'s, namely $Y_t^{(i)}D_t^{(i)}$'s, rather than $Y_t^{(i)}$'s are used before aggregation. The orthogonal matrices $D_t^{(i)}$'s are formed by the following steps. First, we choose a baseline device which has the maximum number of samples. Without loss of generality, we can assume the first device is used (which indicates $1 = \arg\min_{i \in [m]} p_i$). Second, we compute

$$D_t^{(i)} = \underset{D \in \mathcal{F} \cap \mathcal{O}_r}{\operatorname{argmin}} \| Z_{t-1}^{(i)} D - Z_{t-1}^{(1)} \|_F,$$
(3.2)

where recall \mathcal{O}_r denotes the set of $r \times r$ orthogonal matrices. \mathcal{F} can be set differently. When $\mathcal{F} = \{\mathbb{I}_r\}$, (3.2) is invalid. When $\mathcal{F} = \mathcal{O}_r$, (3.2) is the classic matrix approximation problem in linear algebra, named as the *Procrustes problem* (Schönemann, 1966; Cape, 2020). The solution to (3.2) is referred to as *Orthogonal Procrustes Transformation* (OPT) and has a closed form:

$$D_t^{(i)} = W_1 W_2^{\mathsf{T}},$$

where we assume that the SVD of $(Z_{t-1}^{(i)})^\intercal Z_{t-1}^{(1)}$ is $W_1 \Sigma W_2^\intercal$. See more on OPT in Appendix E.

As for privacy, we consider a strong adversary termed honest-but-curious (see Subsection 2.4) which does not violate the rules to peep the raw data but is curious to infer data information from the communicative messages and the training rule. To prevent the potential privacy breaches shown in Example 1, we add Gaussian noise to the transmitted terms whenever communications happen. In particular, the variances of the noise are designed to meet the DP's requirement (see Definitions 1 and 2). In our context, we assume for any $A, A' \in \mathbb{R}^{n \times d}$, A and A' are called neighboring databases if $A^{\mathsf{T}}A$ and $(A')^{\mathsf{T}}A'$ differing in only one entry by at most 1 in absolute value. This assumption can be extended as we discuss later. It can be seen that the local iterations not only save the communication cost but also reduce the amount and scale of added noises under a certain privacy budget, and thus improve the accuracy of the partial SVD.

Finally, to take into account the straggler's effect, we further consider the partial participation protocol, that is, each aggregation only involves the first K responded (not necessarily different) devices before a certain time. Specifically, we assume two random sampling and aggregation schemes with details shown in Subsection 3.2.

3.1 Federated Power Method under Full Participation Protocol

The FedPower under the full participation protocol is shown in Algorithm 2. Specifically, we add two rounds of Gaussian noise. In Line 6 of Algorithm 2, the noise is added to the updates that leave each device, which could be regarded as a kind of *local* protection. And in Line 7 of Algorithm 2, the server adds Gaussian noise to the aggregated updates before sending it to devices, which is *central* protection. Formally, we have the following DP guarantee.

Theorem 4 Algorithm 2 achieves $(2\varepsilon, 2\delta)$ -differential privacy after T iterations.

For a given privacy budget $(2\varepsilon, 2\delta)$, the variance of Gaussian noise is proportional to the number of communications $\lfloor T/p \rfloor$ up to logarithm. Therefore, considering only the error

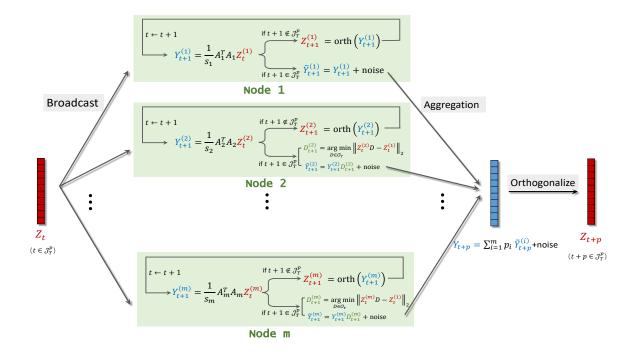


Figure 2: The full participation protocol of FedPower. Each device perform power iterations locally before a communication is required. When communication happens, devices send noise updates to the server in order to prevent privacy breaches, and then the server performs an (partial) aggregation and adds noise before sending it to devices.

that comes from Gaussian noise, local iterations bring benefits to the algorithm's accuracy. However, too many local iterations without synchronization may also cause an error. The next theorem on the convergence of Algorithm 2 reflects this trade-off.

Before going on, we provide the following assumption and definition.

Assumption 1 (Local approximation) For all $i \in [m]$, assume

$$||M_i - M||_2 \le \eta ||M||_2$$
.

The η measures how far the local matrices, M_1, \dots, M_m , are from M. Intuitively, if $s_i = p_i n$ is sufficiently larger than d, then η is sufficiently small.

Definition 4 (Residual Error) Define

$$\rho_t := \max_{i \in [m]} \|Z_t^{(i)} D_{t+1}^{(i)} - Z_t^{(1)}\|_2, \tag{3.3}$$

where if OPT is used, then $D_{t+1}^{(i)}$ is computed via (3.2) with t = t+1 and $\mathcal{F} = \mathcal{O}_r$, and if OPT is not used, then $D_{t+1}^{(i)}$ is computed via (3.2) with t = t+1 and $\mathcal{F} = \{\mathbb{I}_r\}$.

Algorithm 2 FedPower: Full Participation

- 1: **Input:** distributed dataset $\{A_i\}_{i=1}^m$, target rank k, iteration rank $r \geq k$, number of iterations T, synchronous set \mathcal{I}_T^p , the privacy budget (ε, δ) , the variance of noise $\sigma = \frac{\lfloor T/p \rfloor}{\varepsilon \min_i s_i} \sqrt{2\log(\frac{1.25 \lfloor T/p \rfloor}{\delta})}$, and $\sigma' = \frac{\lfloor T/p \rfloor \max_i p_i}{\varepsilon \min_i s_i} \sqrt{2\log(\frac{1.25 \lfloor T/p \rfloor}{\delta})}$. 2: **Initialization** $Z_0^{(i)} = Z_0 \in \mathbb{R}^{d \times r} \sim \mathcal{N}(0, 1)^{d \times r}$.

- The *i*-th worker independently performs $Y_t^{(i)} = M_i Z_{t-1}^{(i)}$ for all $i \in [m]$, where $M_i = \frac{A_i^{\dagger} A_i}{s_i}$; 4:
- 5: if $t \in \mathcal{I}_T^p$ then
- The *i*-th worker adds Gaussian noise: $Y_t^{(i)} = Y_t^{(i)} D_t^{(i)} + \mathcal{N}(0, \|Z_{t-1}^{(i)}\|_{\max}^2 \sigma^2)^{d \times r}$, where $D_t^{(i)}$ is given in (3.2);
- Each worker i sends $Y_t^{(i)}$ to the server and the server performs perturbed aggregation: $Y_t = \sum_{i=1}^m p_i Y_t^{(i)} + \mathcal{N}(0, \max_i \|Z_{t-1}^{(i)} D_t^{(i)}\|_{\max}^2 \sigma'^2)^{d \times r};$ 7:
- Broadcast Y_t to the worker machines and let $Y_t^{(i)} = Y_t$ for all $i \in [m]$; 8:
- 9:
- The *i*-th worker independently performs orthogonalization: $Z_t^{(i)} = \mathsf{orth}(Y_t^{(i)}),$ for all $i \in [m];$ 10:
- 11: **end for**
- 12: **Output:** orthogonalize the approximated eigen-space:

$$\overline{Z}_T := \left\{ \begin{array}{ll} \sum_{i=1}^m p_i Z_T^{(i)} D_{T+1}^{(i)}, & T \notin \mathcal{I}_T^p \\ \sum_{i=1}^m p_i Z_T^{(i)}, & T \in \mathcal{I}_T^p \end{array} \right..$$

The residual error ρ_t measures how the local top-k eigenspace estimator varies across the m workers. Based on the definition, using OPT makes ρ_t smaller than that without using OPT. When $t \in \mathcal{I}_T^p$, $Z_t^{(1)} = \cdots = Z_t^{(m)}$ and thus $\rho_t = 0$. When $t \notin \mathcal{I}_T^p$, each local update would enlarge ρ_t . Hence, intuitively ρ_t depends on p, i.e., the local iterations between two communications. However, later we will show that with OPT ρ_t does not depend on p, while it depends on p without OPT. A residual error is inevitable in previous literature of empirical risk minimization that uses local updates to improve communication efficiency (Stich, 2018; Wang and Joshi, 2018; Yu et al., 2019; Li et al., 2020b, 2019; Li and Zhang, 2021). In our case, it takes the form of ρ_t .

In the next theorem, we establish the convergence of Algorithm 2.

Theorem 5 Let Assumption 1 hold with sufficiently small η , and assume $p_1 = \max_{i \in [m]} p_i$. Recall ρ_t in (3.3) and recall

$$\sigma = \frac{\lfloor T/p \rfloor}{\varepsilon \min_{i} s_{i}} \sqrt{2\log(\frac{1.25 \lfloor T/p \rfloor}{\delta})}.$$

Denote

$$I_1 = \frac{(\sigma_k - \sigma_{k+1})^{-1}}{1 - (1 - p_1) \max_t \rho_t} \cdot \left(\sigma \sqrt{\sum p_i^2} \cdot \sqrt{r} \cdot (\sqrt{d} + \sqrt{\log \lfloor T/p \rfloor})\right), \tag{3.4}$$

and

$$I_2 = \frac{(\sigma_k - \sigma_{k+1})^{-1}}{1 - (1 - \rho_1) \max_t \rho_t} \cdot \sigma_1 \left(\eta + (2 + \eta) \max_t \rho_t \right). \tag{3.5}$$

Let $\epsilon' \approx I_1 + I_2$. If $\epsilon' \lesssim \min\{\frac{1}{2}, \frac{\sqrt{r} - \sqrt{k-1}}{\sqrt{d}}\}$, then after $T = O(\frac{\sigma_k}{\sigma_{k+1}} \log(\frac{d}{\epsilon'}))$ iterations and for some positive constants α and τ , with probability at least $1 - \lfloor T/p \rfloor^{-\alpha} - \tau^{-\Omega(r+1-k)} - e^{-\Omega(d)}$, the output \overline{Z}_T of Algorithm 2 satisfies

$$\sin \theta_k(\overline{Z}_T, V_k) = \|(\mathbb{I}_d - \overline{Z}_T \overline{Z}_T^{\mathsf{T}}) V_k\|_2 \le \epsilon'.$$

From Theorem 5, we see that the convergence bound ϵ' consists of two parts. The term I_1 is induced by the Gaussian noise that DP calls for. Whilst I_2 is induced by the parallelization and synchronization, which is inevitably incurred in the previous literature of empirical risk minimization that uses local updates to improve communication efficiency (Stich, 2018; Wang and Joshi, 2018; Yu et al., 2019; Li et al., 2020b, 2019). Note that without the Gaussian noise, Algorithm 2 reduces to LocalPower introduced in the early version of this paper, and only I_2 remains in the convergence bound (Li et al., 2020c). Note that $\epsilon' \lesssim \min\{\frac{1}{2}, \frac{\sqrt{r} - \sqrt{k-1}}{\sqrt{d}}\}$ is required to make the results valid. We illustrate as follows. For I_1 , we have that $\sigma \sqrt{\sum p_i^2}$ is of order $\frac{\sqrt{m}}{n}$, provided that other parameters are fixed and each local device has the same number of rows. Hence, large n and small m would lead to small ϵ' , which is as expected. As for I_2 , we will show in Theorem 6 that ρ_t is a function of

In principle, the bound of ρ_t depends on whether OPT is used. The next theorem shows that if OPT is used (i.e., $\mathcal{F} = \mathcal{O}_r$), $\rho_t = O(\eta)$, without dependence on p. However, if OPT is not used (i.e., $\mathcal{F} = \{\mathbb{I}_r\}$), then $\rho_t = O(\sqrt{kp\kappa^p\eta})$ has an exponential dependence on p.

 η and would be sufficiently small provided that η is small enough, which in turn results in

Theorem 6 Let $\tau(t) \in \mathcal{I}_T^p$ be the nearest communication time before t and $p = t - \tau(t)$. Let e be the natural constant and $\kappa = \|M\|_2 \|M^{\dagger}\|_2$ be the condition number of M. Suppose $\eta \leq 1/p$ and $\eta \kappa \leq 1/3$. Then ρ_t is a monotone increasing function of η . Moreover, when ϵ' in Theorem 5 is small enough, we have the following upper bound for ρ_t .

• With OPT, ρ_t is bounded by

small I_2 .

$$\min \left\{ 2e^2 \kappa^p p \eta, \frac{\eta \sigma_1}{\delta_k} + 2\gamma_k^{p/4} C_t \right\} = O(\eta), \tag{3.6}$$

where $\gamma_k \in (0,1)$, $\delta_k \simeq (\sigma_k - \sigma_{k+1})$, and $\limsup_t C_t = O(\eta)$.

• Without OPT, ρ_t is bounded by

$$4e\sqrt{k}p\kappa^p\eta = O(\sqrt{k}p\kappa^p\eta). \tag{3.7}$$

Theorem 6 theoretically indicates that why using OPT has such an exponential improvement on dependence on p. This is mainly because of the property of OPT. Let $O^* = \arg\min_{O \in \mathcal{O}_r} \|U - \tilde{U}O\|_F$ for $U, \tilde{U} \in \mathcal{O}_{d \times r}$. Then, up to some universal constant, we have $\|U - \tilde{U}O^*\|_2 \cong \operatorname{dist}(U, \tilde{U})$. See Lemma 25 in Appendix for a formal statement and detailed proof. It implies up to a tractable orthonormal transformation, the difference between the orthonormal bases of two subspaces is no larger than the projection distance between the subspaces. By the Davis-Kahan theorem (see Lemma 22), their projection distance is not larger than $O(\eta)$ up to some problem-dependent constants. However, without OPT, we have to use perturbation theory to bound ρ_t , which inevitably results in exponential dependence on p (see Lemma 15).

3.2 Federated Power Method under Partial Participation Protocol

Full participation is not realistic. The central server can not collect all local devices' output in real-world applications that suffer from the so-called straggler's effect or autonomous effect. Instead, the server could collect the first K ($K \leq m$) responded devices within a certain time range, where the K devices are not necessarily different. Let \mathcal{S}_t ($|\mathcal{S}_t| = K$) be the set of the local devices' indices in the t-th ($t \in \mathcal{I}_T^p$) iteration. Specifically, we consider the following two sampling and aggregating schemes, which have been also used in Li et al. (2020b); McMahan et al. (2017), among others.

Scheme 1 The server generates S_t by i.i.d. sampling with replacement from $\{1, ..., m\}$ for K times. Specifically, index i is selected with probability p_i (i.e., the proportion of the number of samples in local d i over all the samples), and the elements in S_t may occur more than once. In this scheme, the aggregation strategy (before noise addition) is designed as

$$Y_t = \frac{1}{K} \sum_{i \in \mathcal{S}_t} Y_t^{(i)}.$$

Such an aggregation policy could ensure that the partial participation protocol agrees with the full participation protocol in expectation. Indeed, considering only the randomness that comes from S_t , we observe that

$$\mathbb{E}_{\mathcal{S}_t}(Y_t) = \frac{1}{K} \mathbb{E}_{\mathcal{S}_t}(\sum_{k=1}^K Y_t^{(i_k)}) = \mathbb{E}_{\mathcal{S}_t} Y_t^{(i_1)} = \sum_{i=1}^m p_i Y_t^{(i)}.$$

Scheme 2 The server generates S_t by uniformly sampling without replacement from $\{1, \ldots, m\}$ for K times. Hence each index is selected with probability $\frac{1}{m}$ for each time and selected in the final set with probability $\frac{K}{m}$, and each element in S_t occurs once. In such a scheme, we aggregate according to

$$Y_t = \frac{m}{K} \sum_{i \in \mathcal{S}_t} p_i Y_t^{(i)}.$$

Similar to Scheme 1, we have

$$\mathbb{E}_{\mathcal{S}_t}(Y_t) = \frac{m}{K} \mathbb{E}_{\mathcal{S}_t}(\sum_{k=1}^K p_{i_k} Y_t^{(i_k)}) = \sum_{i=1}^m p_i Y_t^{(i)}.$$

The proposed FedPower under the partial participation protocol is described in Algorithm 3, where we use $\tau(t)$ to denote the latest synchronization step before iteration t, and we denote the sampling probability of local devices by $\{q_1, ..., q_m\}$ with $q_i = p_i$ under Scheme 1 and $q_i = 1/m$ under Scheme 2. Note that, when OPT is used, we can use any active device as the baseline device (recall (3.2)), not necessarily the one with the maximum sample size. Similar to the full participation protocol, two rounds of calibrated Gaussian noise are incorporated to ensure local and central privacy protection, respectively. In addition, the OPT is used when aggregation happens. The following theorem provides the formal DP guarantee for Algorithm 3.

Algorithm 3 FedPower: Partial Participation

- 1: **Input:** distributed dataset $\{A_i\}_{i=1}^m$, target rank k, iteration rank $r \geq k$, number of iterations T, synchronous set \mathcal{I}_T^p , the sampling probability of each local device $\{q_1,...,q_m\}$, the number of participated devices K, the privacy budget (ε,δ) , the variance of noise $\sigma = \frac{\lfloor T/p \rfloor}{\varepsilon \min_i s_i} \sqrt{2\log(\frac{1.25 \lfloor T/p \rfloor \max_i q_i}{\delta})}$, $\sigma' = \frac{\lfloor T/p \rfloor}{K\varepsilon \min_i s_i} \sqrt{2\log(\frac{1.25 \lfloor T/p \rfloor}{\delta})}$, and $\sigma'' = \frac{\lfloor T/p \rfloor \max_i p_i}{K\varepsilon \min_i s_i} \sqrt{2\log(\frac{1.25 \lfloor T/p \rfloor}{\delta})}$.
- 2: Initialization: $Z_0^{(i)} = Z_0 \in \mathbb{R}^{d \times r} \sim \mathcal{N}(0, 1)^{d \times r}$.
- 3: for t = 1 to T do
- 4: The *i*-th worker independently performs $Y_t^{(i)} = M_i Z_{t-1}^{(i)}$ for all $i \in [m]$, where $M_i = \frac{A_i^{\dagger} A_i}{s_i}$;
- 5: if $t \in \mathcal{I}_T^p$ then
- 6: The server generates S_t by Scheme 1 or Scheme 2.
- 7: if $i \in \mathcal{S}_t$ then
- 8: The *i*-th worker adds Gaussian noise: $Y_t^{(i)} = Y_t^{(i)} D_t^{(i)} + \mathcal{N}(0, \|Z_{t-1}^{(i)}\|_{\max}^2 \sigma^2)^{d \times r}$ and sends $Y_t^{(i)}$ to the server, where $D_t^{(i)}$ is given in (3.2);
- 9: end if
- 10: The server performs partial aggregation:

$$Y_{t} = \frac{1}{K} \sum_{i \in \mathcal{S}_{t}} Y_{t}^{(i)} + \mathcal{N}(0, \max_{i} \|Z_{t-1}^{(i)} D_{t}^{(i)}\|_{\max}^{2} \sigma'^{2})^{d \times r} \text{ (Scheme 1)},$$

$$Y_{t} = \frac{m}{K} \sum_{i \in \mathcal{S}_{t}} p_{i} Y_{t}^{(i)} + \mathcal{N}(0, \max_{i} ||Z_{t-1}^{(i)} D_{t}^{(i)}||_{\max}^{2} \sigma''^{2})^{d \times r} \text{ (Scheme 2)};$$

- 11: Broadcast Y_t to the worker machines and let $Y_t^{(i)} = Y_t$ for all $i \in [m]$;
- 12: **end if**
- 13: The *i*-th worker independently performs orthogonalization: $Z_t^{(i)} = \operatorname{orth}(Y_t^{(i)})$, for all $i \in [m]$;
- 14: end for
- 15: Output: orthogonalize the approximated eigen-space:

Scheme 1:
$$\overline{Z}_T = \begin{cases} \frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(T)}} Z_T^{(i)} D_{T+1}^{(i)}, & T \notin \mathcal{I}_T^p \\ \frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(T)}} Z_T^{(i)}, & T \in \mathcal{I}_T^p \end{cases}$$

and

Scheme 2:
$$\overline{Z}_T = \begin{cases} \frac{m}{K} \sum_{i \in \mathcal{S}_{\tau(T)}} Z_T^{(i)} D_{T+1}^{(i)}, & T \notin \mathcal{I}_T^p \\ \frac{m}{K} \sum_{i \in \mathcal{S}_{\tau(T)}} Z_T^{(i)}, & T \in \mathcal{I}_T^p \end{cases}$$

Theorem 7 Algorithm 2 achieves $(2\varepsilon, 2\delta)$ -differential privacy after T iterations.

Compared with Algorithm 2, the variance of Gaussian noise in Algorithm 3 is reduced by a factor of $\sqrt{\log(c \max_i q_i)}$ due to the sampling of devices. The next theorem provides the convergence bound of Algorithm 3 under Scheme 1.

Theorem 8 Let Assumption 1 hold with sufficiently small η . Recall

$$\sigma = \frac{\lfloor T/p \rfloor}{\varepsilon \min_{i} s_{i}} \sqrt{2\log(\frac{1.25 \lfloor T/p \rfloor \max_{i} p_{i}}{\delta})},$$

and ρ_t in (3.3), and let

$$I_1 = \frac{(\sigma_k - \sigma_{k+1})^{-1}}{1 - \max_t \rho_t} \cdot \left(K^{-1/2} \sigma \sqrt{r} (\sqrt{d} + \sqrt{\log \lfloor T/p \rfloor}) \right), \tag{3.8}$$

$$I_2 = \frac{(\sigma_k - \sigma_{k+1})^{-1}}{1 - \max_t \rho_t} \cdot \sigma_1 \left(\eta + (2 + \eta) \max_t \rho_t \right), \tag{3.9}$$

and

$$I_3 = \frac{(\sigma_k - \sigma_{k+1})^{-1}}{1 - \max_t \rho_t} \cdot \sigma_1 \phi(K) \left(\log(d+r) + \log\lfloor T/p \rfloor \right), \tag{3.10}$$

where

$$\phi(K) := \frac{1}{K} \sum_{i \neq j, i, j=1}^{m} p_i p_j + \sum_{i=1}^{m} p_i^2 + \frac{1}{K}.$$

Define $\epsilon'' \approx I_1 + I_2 + I_3$. If $\epsilon'' \lesssim \min\{\frac{1}{2}, \frac{\sqrt{r} - \sqrt{k-1}}{\sqrt{d}}\}$, then after $T = O(\frac{\sigma_k}{\sigma_{k+1}}\log(\frac{d}{\epsilon''}))$ iterations and for some positive constants α , β , γ and τ , with probability larger than $1 - \lfloor T/p \rfloor^{-\gamma} \cdot e^{-c\phi(K)}/(d+r)^{\beta} - \lfloor T/p \rfloor^{-\alpha} - \tau^{-\Omega(r+1-k)} - e^{-\Omega(d)}$, the output \overline{Z}_T of Algorithm 3 under Scheme 1 satisfies

$$\sin \theta_k(\overline{Z}_T, V_k) = \|(\mathbb{I}_d - \overline{Z}_T \overline{Z}_T^{\mathsf{T}}) V_k\|_2 \le \epsilon''.$$

The convergence bound of Algorithm 3 under Scheme 1 consists of the following parts. The term I_1 comes from the Gaussian noise, I_2 is incurred by the local iterates, and I_3 can be regarded as the bias that the sampling brings, where note that a larger K yields a smaller $\phi(K)$. All three terms can be sufficiently small in the ideal setting, namely, the total sample size is large, the sample size and quality in each device is large and good, and the number of participated devices K is large. In addition, ρ_t can be upper bounded differently depending on whether OPT is used (see Theorem 6). The following theorem illustrates the convergence of Algorithm 3 under Scheme 2.

Theorem 9 Let Assumption 1 hold with sufficiently small η . Recall

$$\sigma = \frac{\lfloor T/p \rfloor}{\varepsilon \min_i s_i} \sqrt{2 \log(\frac{1.25 \lfloor T/p \rfloor m^{-1}}{\delta})},$$

and ρ_t in (3.3), and let

$$I_1 = \frac{(\sigma_k - \sigma_{k+1})^{-1}}{\frac{m}{K} (\underline{\varsigma} - \varsigma \max_t \rho_t)} \cdot \max\{\frac{m}{K} \varsigma, 1\} \cdot \sqrt{\zeta} \frac{m}{K} \sigma \sqrt{r} (\sqrt{d} + \sqrt{\log \lfloor T/p \rfloor}), \tag{3.11}$$

$$I_{2} = \frac{(\sigma_{k} - \sigma_{k+1})^{-1}}{\frac{m}{K}(\underline{\varsigma} - \varsigma \max_{t} \rho_{t})} \cdot \max\{\frac{m}{K}\varsigma, 1\} \cdot \frac{m}{K}\varsigma \cdot \sigma_{1}\left(\eta + (2 + \eta) \max_{t} \rho_{t}\right), \tag{3.12}$$

$$I_3 = \frac{(\sigma_k - \sigma_{k+1})^{-1}}{\frac{m}{K} (\varsigma - \varsigma \max_t \rho_t)} \cdot \max\{\frac{m}{K} \varsigma, 1\} \cdot \sigma_1 \psi(K) \cdot (\log(d+r) + \log\lfloor T/p \rfloor), \tag{3.13}$$

and

$$I_4 = \frac{(\sigma_k - \sigma_{k+1})^{-1}}{\frac{m}{K} (\varsigma - \varsigma \max_t \rho_t)} \cdot \varsigma \sigma_1 \frac{m}{K} \mid \frac{m}{K} \varsigma - 1 \mid,$$
(3.14)

where

$$\psi(K) := \frac{1}{K} \sum_{i \neq j, i, j = 1}^{m} p_i p_j + \sum_{i = 1}^{m} p_i^2 + \frac{m}{K} \sum_{i = 1}^{m} p_i^2,$$

$$\varsigma := \max_{S} \sum_{l \in S \subset [m], |S| = K} p_l, \ \ \underline{\varsigma} := \min_{S} \sum_{l \in S \subset [m], |S| = K} p_l, \ \ \text{and} \ \ \zeta := \max_{S} \sum_{l \in S \subset [m], |S| = K} p_l^2.$$

Define $\epsilon''' \approx I_1 + I_2 + I_3 + I_4$. If $\epsilon''' \lesssim \min\{\frac{1}{2}, \frac{\sqrt{r} - \sqrt{k-1}}{\sqrt{d}}\}$, then after $T = O(\frac{\sigma_k}{\sigma_{k+1}} \log(\frac{d}{\epsilon'''}))$ iterations and for some positive constants α , β , γ and τ , with probability larger than $1 - \lfloor T/p \rfloor^{-\gamma} \cdot e^{-c\psi(K)/(\frac{m}{K}\varsigma+1)}/(d+r)^{\beta} - \lfloor T/p \rfloor^{-\alpha} - \tau^{-\Omega(r+1-k)} - e^{-\Omega(d)}$, the output \overline{Z}_T of Algorithm 3 under Scheme 2 satisfies

$$\sin \theta_k(\overline{Z}_T, V_k) = \|(\mathbb{I}_d - \overline{Z}_T \overline{Z}_T^{\mathsf{T}}) V_k\|_2 \le \epsilon'''.$$

The convergence bound in Theorem 9 shows a similar though slightly different pattern as that in Theorem 8. The term I_1 is induced by the Gaussian noise, I_2 comes from the local iterations in which ρ_t can be upper bounded as in Theorem 6, and I_3 represents the bias that the sampling of K devices rather than m devices brings, where note that a larger K yields a smaller $\psi(K)$ and thus a smaller I_3 . Also note that there is a common multiplicative factor $\max\{\frac{m}{K}\varsigma,1\}$ in all I_1 , I_2 , and I_3 . By recalling the definition of ς , this multiplicative term attains its minimum at 1 and this happens whenever the samples across all devices are in the same size. This is the sampling strategy that each device is selected with the same probability calls for. In addition, there is an additive term I_4 reflecting the heterogeneity of the sample sizes. A large I_4 means that the sample sizes across devices are more heterogeneous than those with a small I_4 .

3.3 Discussion

Bound for η . Assumption 1 is commonly used to guarantee matrix approximation problems. It tries to make sure that each local data set M_i is a typical representative of the whole data matrix M. Prior work (Gittens and Mahoney, 2016; Woodruff, 2014; Wang et al., 2016) showed that uniform sampling and the partition size in Lemma 24 in Appendix suffice for that M_i well approximates M. The proof is based on the matrix Bernstein (Tropp, 2015). Therefore, under uniform sampling, the smallness of η means sufficiently large local dataset size (or equivalently a small number of worker nodes).

Effect of p. Theorems 5, 8 and 9 indicate that for a given final tolerance ϵ , the number of required communications is $\lfloor T/p \rfloor$ with $T = O(\frac{\sigma_k}{\sigma_{k+1}} \log(\frac{d}{\epsilon}))$ being the number of iterations. Thus more local iterations (large p) bring more communication efficiency. In light of this, it is reasonable to think that under the same level of communication strength during the iteration process, larger p will result in larger T that could lead to smaller subspace distance. Moreover, note that for a given total privacy budget ϵ , the privacy leakage of one round of communication is $\epsilon/\lfloor T/p \rfloor$, which monotonously increases with p. Hence it is not hard to imagine that if we redefine the level of noise such that each communication round leaks the same level of privacy for different p's (i.e., reduce the noise level for large p), then under the same privacy leakage (a.k.a., communication strength) during the iteration process, larger p could lead to smaller subspace distance. We verify this intuition empirically in Section 5.

The above benefits of large p is not without price. Recall that in Theorem 6, we show that the residual error ρ_t (with OPT) is bounded by min $\{a_1, a_2\}$ with $a_1 = 2e^2 \kappa^p p \eta$, and $a_2 = \frac{\eta \sigma_1}{\delta_k} + 2\gamma_k^{p/4} C_t$, where $\gamma_k \in (0, 1)$ and $\limsup_t C_t = O(\eta)$. For moderate p, a_1 is smaller than a_2 , and thus larger p would lead to larger final error, though we proved that the residual error does not depend on p provided that the final error $\epsilon', \epsilon'', \epsilon'''$ is small enough.

Overall, larger p would generally lead to the fast decay of error at the beginning but a larger error eventually.

Decay p gradually. We observe that when we use \mathcal{I}_T^p with p = 1, no local power iterations are involved and interestingly we do not require the good-approximation Assumption 1. Therefore, we are inspired to reduce p by one gradually until p = 1. In particular, we set

$$\mathcal{I}_{T}^{p,\text{decay}} = \left\{ t \in [T] : t = \sum_{i=0}^{l} \max(p-i,1), \ l \ge 0 \right\}.$$
 (3.15)

The choice of $\mathcal{I}_T^{p,\text{decay}}$ implies that we decrease p until it reaches 1.

Reduce the computation of OPT with sign-fixing. From our theory, it is important to use OPT. It weakens the assumption on the smallness of a residual error which is incurred by local computation. From our experiments, it stabilizes vanilla FedPower and achieves much smaller errors. While OPT makes FedPower more stable in practice, OPT incurs more local computation. Specifically, it has time complexity $O(dr^2)$ via calling the SVD of $(Z_{t-1}^{(i)})^{\dagger}Z_{t-1}^{(1)}$. To attain both efficiency and stability, we propose to replace the $r \times r$ matrix $D_t^{(i)}$ in (3.2) by

$$D_t^{(i)} = \arg\min_{D \in \mathcal{D}_r} \|Z_{t-1}^{(i)}D - Z_{t-1}^{(1)}\|_F^2, \tag{3.16}$$

where \mathcal{D}_r denotes all the $r \times r$ diagonal matrices with ± 1 diagonal entries. $D_t^{(i)}$ can be computed in O(rd) time by

$$D_t^{(i)}[j,j] = \operatorname{sgn}\left(\left\langle Z_{t-1}^{(i)}[:,j], Z_{t-1}^{(1)}[:,j] \right\rangle\right), \text{ for all } j \in [r].$$

We empirically observe that sign-fixing serves as a good practical surrogate of OPT; it maintains good stability and achieves comparably small errors.

Dependence on $\sigma_k - \sigma_{k+1}$. Our result depends on $\sigma_k - \sigma_{k+1}$ even when r > k where r is the number of columns used in subspace iteration. If we borrow the tool of Balcan et al. (2016a) rather than that of Hardt and Price (2014), we can improve the result to a slightly milder dependency on $\sigma_k - \sigma_{q+1}$, where q is any intermediate integer between k and r.

Further extensions. First, the level of granularity at which privacy is being promised should be ascertained when adopting the framework of DP. In this paper, we protect privacy at the level of elements of M. Suppose M is the original data matrix, say a social network, then its elements refer to binary edges containing social contacts. In many situations, this is sufficient because large groups of social contacts might not contain any sensitive information, say where a person lives and works are considered public information (Dwork et al., 2014a). Yet, our setting could certainly be extended to other settings, for example, A and A' differ in one row and each row has at most unit Euclidean norm.

Second, though we used the notion of (ε, δ) -DP, the proposed FedPower could be extended to incorporate other DP notions, say the Rényi-DP (Mironov, 2017), Gaussian-DP (Dong et al., 2019), among others. Besides, more advanced composition theorem could be also incorporated; see Dwork et al. (2010) for example.

Finally, our proposed FedPower is simple, effective, and well-grounded. While we analyze it in only the centralized setting, FedPower can be extended to broader settings, such as decentralized setting (Gang et al., 2019) and streaming setting (Raja and Bajwa, 2020). To further reduce the communication complexity, we can combine FedPower with sketching techniques (Boutsidis et al., 2016; Balcan et al., 2016b). For example, we could sketch each $Y_t^{(i)}$ and communicate the compressed iterates to a central server in each iteration. We leave the extensions to our future work.

4. Related Work and Contributions

Partial SVD or principal component analysis (PCA) is one of the most important and popular techniques in modern statistics and machine learning. A multitude of researches focus on iterative algorithms such as power iterations or its variants (Golub and Van Loan, 2012; Saad, 2011). These deterministic algorithms inevitably depend on the spectral gap, which can be quite large in large-scale problems. Another branch of algorithm seek alternatives in stochastic and incremental algorithms (Oja and Karhunen, 1985; Arora et al., 2013; Shamir, 2015, 2016; De Sa et al., 2018). Some work could achieve eigengap-free convergence rate and low-iteration-complexity (Musco and Musco, 2015; Shamir, 2016; Allen-Zhu and Li, 2016). Other work seeks to accelerate the SVD via randomization (Halko et al., 2011; Witten and Candès, 2015; Zhang et al., 2020; Guo et al., 2020).

Large-scale problems and large decentralized datasets necessitate cooperation among multiple worker nodes to overcome the obstacles of data storage and heavy computation. For a review of distributed algorithms for PCA, one could refer to (Wu et al., 2018). One feasible approach is divide-and-conquer algorithms which have only one round of communication (Garber et al., 2017; Fan et al., 2019b; Bhaskara and Wijewardena, 2019). Whereas such algorithms often require large local datasets to reach a certain accuracy. Another line of results for distributed eigenspace estimation uses iterative algorithms that perform multiple communication rounds. They require a much smaller sample size and can often achieve arbitrary accuracy. Some works make use of the shift-and-invert framework (S&I) for PCA, which turns the problem of computing the leading eigenvector to that of approximately solving a small system of linear equations (Garber and Hazan, 2015; Garber et al., 2016; Allen-Zhu and Li, 2016; Garber et al., 2017; Gang et al., 2019). However, these works did not consider the potential privacy breaches during communications, and neither the straggler's effect in a realistic setting.

To alleviate the privacy disclosure concern, a few differentially private single-machine algorithms for PCA or SVD have been proposed. Chaudhuri et al. (2012) invoked an exponential mechanism to compute DP principle components and showed its near-optimality for k = 1. Dwork et al. (2014b) analyzed several aspects of the theoretical soundness of the naive method that adds Gaussian noise to the sample covariance matrix. Hardt and Roth (2013); Hardt and Price (2014) studied the power-iteration-based methods to obtain DP singular vectors, and showed their theoretical merits especially when the underlying

matrix is well-structured (like low coherence). For the distributed setting, Ge et al. (2018) proposed the privacy-preserving distributed sparse PCA. Yet, they did not consider the low communication and straggler's effect that the modern FL meets. In addition, the sparsity assumption is required used while our FedPower is model-free. Very recently, Grammenos et al. (2020) proposed a federated, asynchronous, and DP algorithm for PCA. Methodologically, the algorithm is not power-iteration-based. Instead, their algorithm incrementally computes local model updates using streaming procedure and adaptively estimates its leading principal components. In particular, they assume the clients are arranged in a tree-like structure, while we did not make such an assumption. Theoretically, the bounds therein hold in the sense of expectation, while we provide the non-asymptotic bound for the spectral deviation of the estimated singular vectors from true ones.

Note that the technique of local updates emerges as a simple but powerful tool in distributed empirical risk minimization (McMahan et al., 2017; Zhou and Cong, 2017; Stich, 2018; Wang and Joshi, 2018; Yu et al., 2019; Li et al., 2020b, 2019; Khaled et al., 2019). However, our analysis is totally different from the local SGD algorithms (Zhou and Cong, 2017; Stich, 2018; Wang and Joshi, 2018; Yu et al., 2019; Li et al., 2020b, 2019; Khaled et al., 2019). A main challenge in analyzing FedPower is that the local SGD algorithms for empirical risk minimization often involve an explicit form of (stochastic) gradients. For SVD or PCA, the gradient cannot be explicitly expressed, so the existing techniques cannot be applied.

Overall, the main contributions of this work can be summarized as follows. First, we introduce two adversary models to indicate how privacy is leaked in the naive distributed power method. The potential privacy leakage of many algorithms is known to all but to the best of our knowledge, few literatures give explicit illustrations. Second, we develop a set of algorithms called FedPower, which could handle the communication (including disconnection), privacy, computation concern in the modern FL framework simultaneously. Whereas most existing work on SVD only considers handling one or two of the three aspects from different perspectives. Last but not least, we provide a framework to analyze the convergence bound of FedPower. We make a delicate analysis of the error (influenced by local iterates, DP's perturbation, and random straggling of devices) by utilizing the tools designed for single-machine noisy power method (Hardt and Price, 2014) and the tools from random matrix theory (Tropp, 2015), providing convenience for further analysis of distributed and federated power method-based computation of SVD.

5. Experiments

In this section, we numerically evaluate the efficacy of the proposed algorithms. Recall that we add Gaussian noise in each round of communication to meet the DP's requirement. We denote this algorithm by FedPower with noise. For comparison, we also consider a variation of FedPower called FedPower without noise, where DP can not be achieved yet the privacy could also be protected to some extent by the rationality of FL. For these two types of algorithms, we evaluate the effect of number of local iterations p, the effect of the number of devices m, the effect of the number of participated devices K, and the effect of the decaying p strategy. Unless specified, for the sake of efficiency, we use the sign-fixing strategy (see Eq.(3.16)) when aggregating. The used datasets are available on the LIBSVM

Table 1: A summary	of	used	data	sets	from	the	LIBS	VM	website.

Datasets	n	d	Datasets	n	d
Acoustic	78823	50	Aloi	108000	128
A9a	32561	123	Combined	78823	100
Connect-4	67557	126	Covtype	581012	54
Housing	506	13	Ijcnn1	49990	22
MNIST	60000	780	W8a	49749	300

website² and are summarized in Table 1. Specifically, we use the first three typical datasets (i.e., MINIST, A9a and Acoustic) to illustrate our theoretical findings in Section 3 with corresponding results shown in Subsection 5.1 and 5.2. And other datasets in Table 1 will be applied in additional experiments in Subsection 5.3. The n samples are partitioned among m nodes such that each node having $s = \frac{n}{m}$ samples. The features are scaled so that they are in the region [-1,1]. We fix the target rank k=5.

Our experimental design is motivated by the theoretical results; see the discussion on the effect of p in Subsection 3.3. More precisely, the theories indicate that under the same level of communication strength (in the noiseless setting) and the same level of privacy leakage (in the noise setting), more local iterations (larger p) would lead to better accuracy. We empirically verify this theoretical implication in the following more realistic experiments.

5.1 FedPower without Noise

In this regime, the merits of the FedPower lie in communication efficiency. Hence, we evaluate the subspace distance (see Definition 3) between the estimated and the true singular vectors against the number of communications.

Effect of number of local iterations p. Figure 3 shows the curves of projection distance v.s. communications under different settings of p. The FedPower is more communication efficient than the baseline (p=1) in the first 10 rounds of communications. We observe that large p leads to fast convergence in the beginning but a nonvanishing error at the end. Using $p \geq 2$, the error does not converge to zero. This is because that the bias that the local iterations bring is non-negligible under the finite sample setting. In machine learning tasks such as principal component analysis and latent semantic analysis, high-precision solutions are unnecessary (Deerwester et al., 1990). In such tasks, FedPower can solve large-scale truncated SVD using a small number of communications.

Effect of the number of local devices m. Figure 4 shows the performance of FedPower of under different settings of m. It indicates that smaller m could lead to better convergence. It is because when m is small, each device owns more samples, which implies M_i (i = 1, ..., m) approximates the global M better than that in the big m setting.

Effect of the number of participated devices K. Aforementioned experiments considered the full participation protocol. In the partial participation protocol, there is another

^{2.} https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

important parameter K which measures the number of participating devices in each round of communication. The effect of K on the FedPower's performance is displayed in Figure 5. As expected, large K is good for the algorithm. In particular, the sampling and aggregation Scheme 2 leads to better empirical results than Scheme 1 does.

Effect of the decaying p strategy. We observe in Figure 3 that larger p fastens convergence but enlarges the final error. By contrast, p=1 has the lowest error floor but also the lowest convergence rate. Similar phenomena have been previously observed in distributed empirical risk minimization (Wang and Joshi, 2019; Li et al., 2019). To allow for both fast convergence in the beginning and vanishing error at the end, we propose to decay p with iterations according to (3.15). The results are shown in Figure 6. For these three datasets, the shrinkage of p turns out to be useful. Whereas we empirically observe that when the error floor is low, early decaying of p may slow down the convergence.

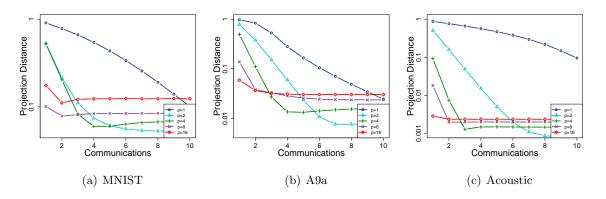


Figure 3: Effect of the number of local iterations p under the *noiseless* regime and the full participation protocol. The number of local devices m for MNIST, A9a, and Acoustic is fixed at 200, 20, and 100, respectively.

5.2 FedPower with Noise

In this regime, the FedPower satisfies the (ε, δ) -DP's requirement, where (ε, δ) measures the privacy leakage. So, we evaluate the accuracy of algorithms against the accumulative privacy leakage. Specifically, for each round of communication, we add the same amount of noise for the FedPower under different parameter settings. In particular, the privacy budget is divided equally among the two times of noise adding within each communication. Note that under such a setting, the same number of communications indicate the same amount of privacy leakage.

Effect of number of local iterations p. Figure 7 displays the curves of projection distance v.s. accumulative privacy leakage (aka. communications) under different settings of p. Under the same level of privacy leakage, more local iterations yield better accuracy than the baseline (p = 1). These empirical results coincides with the theoretical findings; see Subsection 3.3 (effect of p) for details. Slightly different from the noiseless setting, large

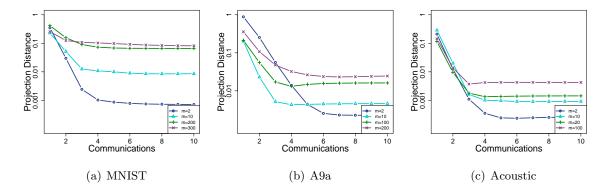


Figure 4: Effect of the number of local devices m under the the noiseless regime and the full participation protocol. The local iterations p is fixed at 4.

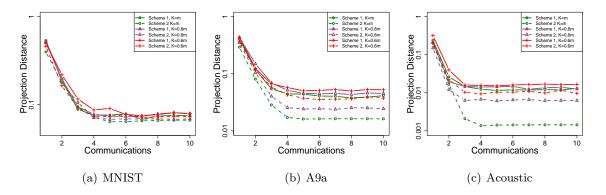


Figure 5: Effect of the number of participated devices K in each round of communication under the *noiseless* regime and the partial participation protocol. The local iterations p is fixed at 4. The number of local devices m for MNIST, A9a, and Acoustic is fixed at 200, 20, and 100, respectively.

p leads to fast improvement of accuracy in the beginning but both large p and small p can not lead to a non-vanishing error eventually. For p=1, this is because more global iterations produce more privacy breaches, and thus more noise must be added. The benefit of more global iterations would be counteracted by the additional noise. While for $p \geq 2$, not only the Gaussian noise but also the local iterations would lead to the final bias of the algorithm.

Effect of the number of local devices m. Similar to the noiseless setting, Figure 8 illustrates that smaller m could generally lead to the better accuracy of FedPower under the same level of privacy leakage. The reason is that local data with a larger sample size could approximate the global data better.

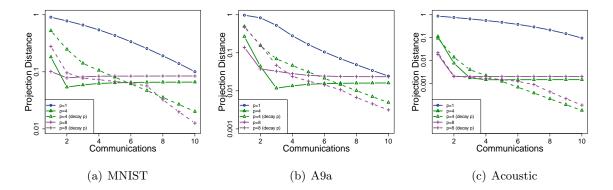


Figure 6: Effect of the decaying p strategy under the *noiseless* regime and the full participation protocol. The number of local devices m for MNIST, A9a, and Acoustic is fixed at 200, 20, and 100, respectively.

Effect of the number of participated devices K. Turning to the partial participation protocol, the parameter K measuring the number of participated devices in each round of communication is meaningful. Figure 9 shows the effect of K on the FedPower's performance. As we can see, large K could lead to slightly better results than small K does. But the overall influence of K on the algorithm's accuracy is not tremendous, which confirms the effectiveness of FedPower in real applications with large numbers of stragglers.

Effect of the decaying p strategy. Recall that in the noiseless setting, we observed that decaying p may turn out to be beneficial if the lowest error floor is high. To test this decaying p strategy in the noisy setting, we plot the curves of projection distance v.s. accumulative privacy leakage under the decaying and non-decaying p settings, shown in Figure 10. It turns out that except for the MNIST dataset, decaying p does not work on the other two datasets. The reason might be that when p decays, the frequency of communication increases, and hence more noise are added to protect the privacy breaches. As a result, the benefits that the decaying p strategy brings may be absorbed by the noise. This is DP algorithm's tradeoff between accuracy and privacy.

5.3 Additional Experiments

We conduct two sets of additional experiments to demonstrate the effectiveness of Fed-Power. First, in the noiseless setting, we compare FedPower with three one-shot baseline methods. Second, in the noise setting, we study how privacy budgets affect the performance of FedPower and discuss when large privacy budgets are permissive.

Comparison with other methods. We evaluate three variants of FedPower: the vanilla version, with OPT, and with sign-fixing. We compare our algorithms with one-shot algorithms, unweighted distributed averaging (UDA) (Fan et al., 2019b), weighted distributed averaging (WDA) (Bhaskara and Wijewardena, 2019) 2019), and distributed randomized SVD (DR-SVD); the details of the algorithms are described in Appendix F. We study

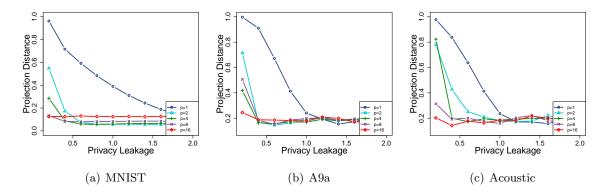


Figure 7: Effect of the number of local iterations p under the *noisy* regime and the full participation protocol. The number of local devices m for MNIST, A9a, and Acoustic is fixed at 200, 20, and 100, respectively.

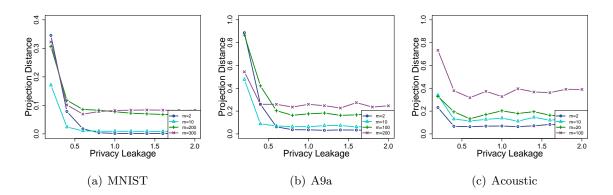


Figure 8: Effect of the number of local devices m under the the noisy regime and the full participation protocol. The local iterations p is fixed at 4.

the precision when the algorithms converge. For three variants of FedPower we fix p=4 (without decaying p). We run each algorithm 10 times and report the mean and standard deviation (std) of the final errors. Table 2 shows the results on ten datasets. The results indicate that one-shot methods do not find high-precision solutions unless the local data size is sufficiently large compared with the FedPower.

Effect of privacy budgets. Note that there are two rounds of Gaussian noise addition (line 6 and line 7 in Algorithm 2) in each communication (a.k.a. once aggregating and broadcasting) of the FedPower. It is shown that the noise scale in Algorithm 2 can ensure the $(\frac{\varepsilon}{2|T/p|}, \frac{\delta}{2|T/p|})$ -DP for every implement of line 6 and line 7, respectively. In this experiment, we redivide the variance of Gaussian noise such that in each communication, line 6 and line 7 of Algorithm 2 meets the requirements of (ε_1, δ) -DP and (ε_2, δ) -DP, respectively. ε_1 and

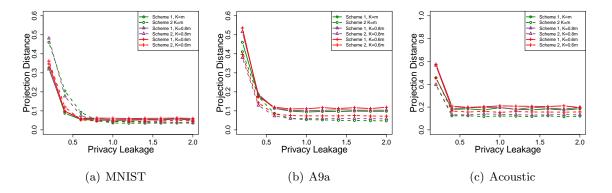


Figure 9: Effect of the number of participated devices K in each round of communication under the *noisy* regime and the partial participation protocol. The local iterations p is fixed at 4. The number of local devices m for MNIST, A9a, and Acoustic is fixed at 200, 20, and 100, respectively.

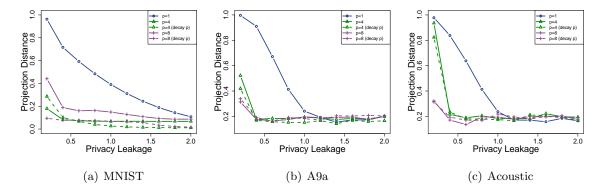


Figure 10: Effect of the decaying p strategy under the *noisy* regime and the full participation protocol. The number of local devices m for MNIST, A9a, and Acoustic is fixed at 200, 20, and 100, respectively.

 ε_2 correspond to the local protection (though not exactly the local DP (Duchi et al., 2018)) and global protection of privacy. It was common sense that the total privacy budget of an algorithm should not be larger than 1, in which case the privacy protection seems meaningless (Wasserman and Zhou, 2010; Bhowmick et al., 2018). However, this requirement might lead to low utility of the algorithm. Very recently, Bhowmick et al. (2018) reconceptualize the protections of local DP (Duchi et al., 2018), rather than providing protection against arbitrary inferences, algorithms that only protects against accurate reconstruction of (functions of) an individual's data can allow a larger privacy budget ($\varepsilon \gg 1$). Based on these findings, in this experiment, we allow ε_1 to be large and ε_2 is fixed to be a small number 0.1

Table 2: Error comparison among three one-shot baseline algorithms and our FedPower. We uniformly distribute n samples into $m = \max(\lfloor \frac{n}{1000} \rfloor, 3)$ devices so that each device has about 1000 samples. We show the mean errors of ten repeated experiments with its standard deviation enclosed in parentheses. Here we use p=4 for all variants of FedPower and sufficiently large T's which guarantee FedPower converges.

Datasets		FedPower with $p=4$		DR-SVD	UDA	WDA	
Datasets	OPT	Sign-fixing	Vanilla	DK-3VD	UDA	WDA	
Acoustic	1.83e-03 (4.40e-04)	2.03e-03 (3.90e-04)	2.38e-03 (8.50e-04)	1.54e-02 (6.59e-03)	7.76e-03 (2.64e-03)	6.67e-03 (2.41e-03)	
Aloi	3.07e-02 (1.10e-02)	6.57e-02 (1.06e-02)	5.24e-02 (1.10e-02)	1.92e-03 (4.30e-04)	4.80e-02 (1.10e-02)	4.37e-02 (4.73e-03)	
A9a	4.09e-03 (4.20e-04)	5.82e-03 (1.41e-03)	8.13e-02 (3.44e-02)	4.63e-02 (9.24e-03)	2.64e-02 (1.58e-02)	2.40e-02 (1.50e-02)	
Combined	6.01e-03 (1.59e-03)	5.57e-03 (1.05e-03)	2.47e-02 (3.40e-02)	5.19e-02 (6.23e-03)	4.63e-02 (2.97e-02)	4.16e-02 (2.76e-02)	
Connect-4	1.27e-02 (4.52e-03)	1.81e-02 (3.79e-03)	1.70e-02 (4.35e-03)	1.61e-02 (2.96e-03)	1.65e-01 (3.48e-02)	1.56e-01 (3.26e-02)	
Covtype	7.38e-03 (8.50e-04)	6.23e-03 (3.30e-04)	1.28e-02 (1.88e-03)	1.82e-01 (8.73e-02)	6.09e-02 (9.70e-03)	5.60e-02 (9.41e-03)	
Housing	1.18e-02 (5.45e-03)	2.76e-02 (1.14e-02)	3.84e-02 (5.11e-02)	5.66e-01 (2.62e-01)	9.16e-02 (5.09e-02)	5.89e-02 (3.25e-02)	
Ijcnn1	1.53e-01 (1.87e-01)	1.95e-01 (2.45e-01)	3.23e-01 (2.24e-01)	1.21e+00 (1.70e-01)	3.85e-01 (7.62e-02)	3.67e-01 (7.59e-02)	
MNIST	2.62e-03 (3.40e-04)	4.85e-03 (8.00e-04)	5.08e-03 (7.90e-04)	5.00e-05 (0.00e+00)	1.08e-02 (3.00e-03)	8.91e-03 (2.53e-03)	
W8a	1.90e-02 (2.46e-03)	1.75e-02 (1.76e-03)	1.68e-02 (1.29e-03)	7.13e-02 (2.06e-02)	1.52e-01 (4.37e-02)	1.51e-01 (4.11e-02)	

Table 3: Error comparison of FedPower with different privacy budget. We use the decaying p strategy with p=4. Other parameters are fixed at $m=100, k=5, r=10, \varepsilon_2=0.1$. The projection distance is the minimum distance between the estimated and the true singular vectors over the first 40 global iterations, and the averaged results over 20 replications are recorded with standard deviations shown in the parentheses.

Datasets	Projection Distance							
	$\varepsilon_1 = \infty$	$\varepsilon_1 = 100$	$\varepsilon_1 = 10$	$\varepsilon_1 = 1$	$\varepsilon_1 = 0.1$			
Acoustic	1.08e-14(1.78e-17)	0.0341(0.0018)	0.0337(0.0013)	0.0467(0.0024)	0.3024(0.0125)			
Aloi	2.83e-13(1.59e-14)	0.0066(0.0001)	0.0066(0.0002)	0.0009(0.0003)	0.0658(0.0011)			
A9a	8.25e-12(4.97e-12)	0.0189(0.0014)	0.0187(0.0017)	0.0247(0.0027)	0.1535(0.0132)			
Combined	2.50e-14(6.96e-17)	0.0353(0.0029)	0.0349(0.0028)	0.0486(0.0044)	0.3179(0.0208)			
Connect-4	1.22e-11(1.15e-11)	0.0038(0.0002)	0.0037(0.0002)	0.005(0.0003)	0.0350(0.0021)			
Covtype	2.07e-9(1.82e-9)	0.0023(0.0001)	0.0023(0.0002)	0.0032(0.0002)	0.0230(0.0016)			
Housing	5.48e-16(5.30e-17)	0.5210(0.0495)	0.5030(0.0678)	0.5027(0.0528)	0.5444(0.0612)			
Ijcnn1	7.15e-11(6.38e-13)	0.0181(0.0017)	0.0182(0.0023)	0.0243(0.0026)	0.1519(0.0101)			
MNIST	7.31e-15(6.92e-15)	8.58e-4(1e-5)	8.53e-4(1e-5)	1.20e-3(1e-5)	0.0086(0.0002)			
W8a	6.16e-6(7.96e-6)	0.0376(0.0026)	0.0378(0.0034)	0.0511(0.0043)	0.3046(0.0200)			

to keep the rationality of DP. When ε_1 is small, the second round of Gaussian noise that corresponds to ε_2 is not needed.

Table 3 shows the minimum projection distance between the estimated and the true singular vectors over 40 iterations under different ε_1 's and the decaying p strategy (p = 4). As expected, enlarge ε_1 could make the algorithm useful. And our above analysis supports this treatment to a certain extent.

6. Conclusion

We have developed a communication efficient, privacy-preserving, stragglers acceptable algorithm that we call FedPower for solving the partial SVD problem in the modern federated machine learning regime. Every worker device performs multiple (say p) local power iterations between two consecutive iterations. The full device or partial device aggregation is performed after every p iterates. The Gaussian noise could be added to the iterates in the communication round to prevent possible privacy leakage. We theoretically proved the convergence bound of FedPower and discussed the effect of local iterations p. Empirically, we showed that the local iterations ($p \ge 2$) of FedPower yield more accurate singular vector solutions than the baseline (p = 1) method does under the same communication rounds and the same amount of accumulated privacy leakage.

Methodologically, our algorithms provide a flexible and general framework for the computation of partial SVD in the modern machine learning setting. In the theoretical part, our analysis gives a new application of noisy power method (Hardt and Price, 2014) by combining the perturbed iterate analysis. Finally, the proposed algorithms can be applied to a wide range of statistical and machine learning tasks, including matrix completion, clustering, and ranking, among others.

Appendix

Subsection $\bf A$ and $\bf B$ includes the proofs (also the proof sketch if necessary) corresponding to the full and partial participation protocols, respectively. Subsection $\bf C$ contains the technical lemmas. Subsection $\bf D$ presents auxiliary lemmas used in the proofs. Subsection $\bf E$ introduces the formal definitions and lemmas on metrics between two subspaces. Subsection $\bf F$ provides the algorithms that we compared in the experiments.

A. Full participation

Proof of Theorem 4

Note that when $t \in \mathcal{I}_T^p$, each worker i adds Gaussian noise as follows,

$$Y_t^{(i)} = \frac{A_i^{\mathsf{T}} A_i}{s_i} Z_{t-1}^{(i)} + \mathcal{N}(0, ||Z_{t-1}^{(i)}||_{\max}^2 \sigma^2)^{d \times r}, \tag{A.1}$$

where

$$\sigma = \frac{\lfloor T/p \rfloor}{\varepsilon \min_{i} s_{i}} \sqrt{2\log(\frac{1.25 \lfloor T/p \rfloor}{\delta})}.$$
 (A.2)

Consider the l-th $(1 \le l \le r)$ column of $Z_{t-1}^{(i)}$, denoted by $[Z_{t-1}^{(i)}]_l$, then the L_2 -sensitivity of $\frac{A_i^\intercal A_i}{s_i}[Z_{t-1}^{(i)}]_l$ is

$$\|\frac{A_i^{\mathsf{T}} A_i}{s_i} [Z_{t-1}^{(i)}]_l - \frac{A_i'^{\mathsf{T}} A_i'}{s_i} [Z_{t-1}^{(i)}]_l \|_2 \leq \frac{1}{s_i} \|Z_{t-1}^{(i)}\|_{\max}.$$

Stacking r such vectors together to obtain a $d \times r$ -dimentional vector and noting the choice of σ , we know equation (A.1) can obtain $(\frac{\varepsilon}{\lfloor T/p \rfloor}, \frac{\delta}{\lfloor T/p \rfloor})$ -differentially private $Y_t^{(i)}$ for each fixed i and t, and thus also obtain $(\frac{\varepsilon}{\lfloor T/p \rfloor}, \frac{\delta}{\lfloor T/p \rfloor})$ -central differential privacy for each $t \in \mathcal{I}_T^p$.

On the other hand, when $t \in \mathcal{I}_T^p$, the server also adds the Gaussian noise after aggregation as follows,

$$Y_t = \sum_{i=1}^m p_i Y_t^{(i)} D_t^{(i)} + \mathcal{N}(0, \max_i ||Z_{t-1}^{(i)}||_{\max}^2 \sigma'^2)^{d \times r},$$

where

$$\sigma' = \frac{\lfloor T/p \rfloor \max_{i} p_i}{\varepsilon \min_{i} s_i} \sqrt{2\log(\frac{1.25 \lfloor T/p \rfloor}{\delta})}.$$

It is easy to see that the L_2 -sensitivity of $\sum_{i=1}^m p_i Y_t^{(i)} D_t^{(i)}$ is bounded by $\frac{\max p_i}{\min s_i} \max_i \|Z_{t-1}^{(i)} D_t^{(i)}\|_{\max}$. Hence by the choice of σ' and the rationality of the Gaussian mechanism, we know that such noise adding procedure obtains $(\frac{\varepsilon}{\lfloor T/p \rfloor}, \frac{\delta}{\lfloor T/p \rfloor})$ -central differential privacy for each $t \in \mathcal{I}_T^p$.

Consequently, considering that [T/p] iterations are required for communication, we finally observe that Algorithm 2 attains $(2\varepsilon, 2\delta)$ -differential privacy via Proposition 3.

Proof of Theorem 5

Proof sketch of Theorem 5: First, we define a virtual sequence

$$\overline{Z}_t = \sum_{i=1}^m p_i Z_t^{(i)} O_t^{(i)}.$$

Here $O_t^{(i)} \in \mathbb{R}^{r \times r}$ is defined as

$$O_t^{(i)} = \begin{cases} \mathbb{I}_r & \text{if } t \in \mathcal{I}_T^p \\ D_{t+1}^{(i)} & \text{if } t \notin \mathcal{I}_T^p. \end{cases}$$

Then, we will write \overline{Z}_t in the following recursive manner,

$$\overline{Z}_t = [M\overline{Z}_{t-1} + \mathcal{G}_t]R_t^{-1},$$

where R_t is a reversible matrix to be defined, and \mathcal{G}_t is some noisy perturbation coming from DP's noise and local iterates. To analyze the convergence of FedPower, we aim to use the analytical framework of noisy power iterates in Hardt and Price (2014). However, their results require \overline{Z}_t to have orthonormal columns, which is not met in our setting. As a remedy, we obtain the following results, which is a modification of Corollary 1.1 (see Lemma 18) in Hardt and Price (2014).

Lemma 10 (Informal version of Lemma 12) Let $\overline{Z}_0 \sim \mathcal{N}(0, I_{d \times r})$. Assume \overline{Z}_t iterates as follows,

$$\overline{Z}_t \leftarrow \frac{1}{n} A^{\mathsf{T}} A \overline{Z}_{t-1} + \mathcal{G}_t.$$

If \mathcal{G}_t satisfies

$$5\|\mathcal{G}_t\|_2 \le \epsilon(\sigma_k - \sigma_{k+1}) \min_t \|\overline{Z}_t\|_{\mathbf{m}} \quad \text{and} \quad 5\|V_k^\mathsf{T} \mathcal{G}_t\|_2 \le (\sigma_k - \sigma_{k+1}) \max_t \|\overline{Z}_t\|_{\mathbf{m}} \frac{\sqrt{r} - \sqrt{k-1}}{\tau \sqrt{d}},$$

for some fixed τ and $\epsilon < 1/2$. Then with high probability, there exists an $T = O(\frac{\sigma_k}{\sigma_k - \sigma_{k+1}} \log(d\tau/\epsilon))$ so that after T steps

$$||(I - \overline{Z}_T \overline{Z}_T^{\mathsf{T}}) V_k||_2 \le \epsilon.$$

The result also holds for the following iterates with any reversible matrix R_t ,

$$\overline{Z}_t \leftarrow [\frac{1}{n} A^{\mathsf{T}} A \overline{Z}_{t-1} + \mathcal{G}_t] R_t^{-1}.$$

In light of this result, the convergence of Algorithm 2 could be established if we could bound the perturbation error induced from un-synchronization and differential privacy.

Proof We provide a proof in three steps.

First step: Perturbed iterate analysis. Recall that we defined a virtual sequence by

$$\overline{Z}_t = \sum_{i=1}^m p_i Z_t^{(i)} O_t^{(i)},$$

where $O_t^{(i)}$ is \mathbb{I}_r if $t \in \mathcal{I}_T^p$, and is $D_{t+1}^{(i)}$ defined by

$$D_{t+1}^{(i)} = \underset{D \in \mathcal{F} \cap \mathcal{O}_r}{\operatorname{argmin}} \| Z_t^{(i)} D - Z_t^{(1)} \|_o,$$

if $t \notin \mathcal{I}_T^p$. For any t, we write $Y_t^{(i)} = Z_t^{(i)} R_t^{(i)}$ which is used repeatedly in the following proofs. Now we discuss the iteration of \overline{Z}_t under $t \notin \mathcal{I}_T^p$ and $t \in \mathcal{I}_T^p$, respectively.

When $t \notin \mathcal{I}_T^p$, we note that $Y_t^{(i)} = M_i Z_{t-1}^{(i)}$. Then, given any invertible R^t (to be specified in Lemma 13), we have

$$\overline{Z}_{t} = \sum_{i=1}^{m} p_{i} Z_{t}^{(i)} O_{t}^{(i)}
= \sum_{i=1}^{m} p_{i} M_{i} Z_{t-1}^{(i)} O_{t-1}^{(i)} R_{t}^{-1} + \sum_{i=1}^{m} p_{i} Z_{t}^{(i)} [O_{t}^{(i)} R_{t} - R_{t}^{(i)} O_{t-1}^{(i)}] R_{t}^{-1}
= (\sum_{i=1}^{m} p_{i} M Z_{t-1}^{(i)} O_{t-1}^{(i)} + H_{t} + W_{t}) R_{t}^{-1}
= (M \overline{Z}_{t-1} + H_{t} + W_{t}) R_{t}^{-1},$$
(A.3)

where

$$H_t = \sum_{i=1}^{m} p_i H_t^{(i)} = \sum_{i=1}^{m} p_i (M_i - M) Z_{t-1}^{(i)} O_{t-1}^{(i)}, \tag{A.4}$$

$$W_t = \sum_{i=1}^m p_i W_t^{(i)} = \sum_{i=1}^m p_i Z_t^{(i)} [O_t^{(i)} R_t - R_t^{(i)} O_{t-1}^{(i)}].$$
 (A.5)

When $t \in \mathcal{I}_T^p$, synchronization happens and Gaussian noise matrices denoted by $N_{t-1}^{(i)}$ whose elements follow $\mathcal{N}(0, \|Z_{t-1}^{(i)}\|_{\max}^2 \sigma^2)$ i.i.d. are added to each local machine before synchronization and Gaussian noise denoted by N_t' whose elements follow $\mathcal{N}(0, \max_i \|Z_{t-1}^{(i)}D_t^{(i)}\|_{\max}^2 \sigma'^2)$ is added the before the server send aggregated output to each local machine. In such cases, $R_t^{(i)}$'s are identical for all $i \in [m]$ and we let them be R_t ; see Lemma 13 for details, and

$$Y_t^{(i)} = \sum_{l=1}^m p_l [M_l Z_{t-1}^{(l)} D_t^{(l)} + N_{t-1}^{(l)}] + N_t' = \sum_{l=1}^m p_l M_l Z_{t-1}^{(l)} D_t^{(l)} + N_t + N_t',$$

for all $i \in [m]$, where

$$N_t = \sum_{i=1}^m p_i N_{t-1}^{(i)}.$$

Hence,

$$\overline{Z}_{t} = \sum_{i=1}^{m} p_{i} Z_{t}^{(i)} O_{t}^{(i)}
= (\sum_{i=1}^{m} p_{i} M_{i} Z_{t-1}^{(i)} D_{t}^{(i)} + N_{t} + N_{t}' + \sum_{i=1}^{m} p_{i} Z_{t}^{(i)} [O_{t}^{(i)} R_{t} - R_{t}^{(i)}]) R_{t}^{-1}
= (\sum_{i=1}^{m} p_{i} M_{i} Z_{t-1}^{(i)} O_{t-1}^{(i)} + N_{t} + N_{t}' + \sum_{i=1}^{m} p_{i} Z_{t}^{(i)} [R_{t} - R_{t}^{(i)}]) R_{t}^{-1}
= (M \overline{Z}_{t-1} + H_{t} + N_{t} + N_{t}') R_{t}^{-1},$$
(A.6)

where we used the fact that $O_t^{(i)} = \mathbb{I}_r$, $O_{t-1}^{(i)} = D_t^{(i)}$, and $R_t = R_t^{(i)}$; H_t is defined in (A.4).

Second step: Bound the noise term. We proceed to bound $\|H_t\|_2$, $\|W_t\|_2$, $\|N_t\|_2$ and $\|N_t'\|_2$, respectively. Note that $\|N_t'\|_2$ is of smaller order than $\|N_t\|_2$. To see this, we only need to observe the following facts $\|Z_{t-1}^{(i)}D_t^{(i)}\|_{\max}$ is not large than $\sqrt{r}\|Z_{t-1}^{(i)}D_t^{(i)}\|_{\max}$, N_t is the summation of m noise matrix compared to just 1 in N_t' , and \sqrt{r} is of smaller order than m.

• For $||H_t||_2$, we have

$$||H_{t}||_{2} = ||\sum_{i=1}^{m} p_{i}H_{t}^{(i)}||_{2} \leq \sum_{i=1}^{m} p_{i}||H_{t}^{(i)}||_{2} = \sum_{i=1}^{m} p_{i}||(M_{i} - M) Z_{t-1}^{(i)} O_{t-1}^{(i)}||_{2}$$

$$\leq \sum_{i=1}^{m} p_{i}||M_{i} - M||_{2}||Z_{t-1}^{(i)} O_{t-1}^{(i)}||_{2} \leq \sum_{i=1}^{m} p_{i}\eta||M||_{2}||Z_{t-1}^{(i)} O_{t-1}^{(i)}||_{2} \leq \eta||M||_{2} = \eta \sigma_{1}.$$
(A.7)

• For $||W_t||_2$, we have

$$||W_t||_2 = ||\sum_{i=1}^m p_i W_t^{(i)}||_2 \le \sum_{i=1}^m p_i ||W_t^{(i)}||_2 = \sum_{i=1}^m p_i ||Z_t^{(i)} \left(O_t^{(i)} R_t - R_t^{(i)} O_{t-1}^{(i)} \right)||_2$$

$$\le \sum_{i=1}^m p_i ||O_t^{(i)} R_t - R_t^{(i)} O_{t-1}^{(i)}||_2.$$

Here, a good choice of R_t should be specified to ensure a tight bound of $||W_t||_2$. Specifically, R_t is chosen in a recursive manner as we show in Lemma 13 in the Appendix C. In particular, we prove in Lemma 13 that for any $i \in [m]$

$$||O_{t}^{(i)}R_{t} - R_{t}^{(i)}O_{t-1}^{(i)}||_{2} \leq \sigma_{1}(M_{1})||Z_{t}^{(i)}O_{t}^{(i)} - Z_{t}^{(1)}||_{2} + ||M_{1} - M_{i}||_{2} + \sigma_{1}(M_{i})||Z_{t-1}^{(i)}O_{t-1}^{(i)} - Z_{t-1}^{(1)}||_{2}$$

$$\leq \sigma_{1}\left(\rho_{t} + \eta + (1 + \eta)\rho_{t-1}\right), \tag{A.8}$$

where we make use of Assumption 1 and define

$$\rho_t = \max_i \|Z_t^{(i)} O_t^{(i)} - Z_t^{(1)}\|_2 = \max_i \|Z_t^{(i)} D_{t+1}^{(i)} - Z_t^{(1)}\|_2, \tag{A.9}$$

with ρ_t upper bounded as in (3.6) and (3.7); see Lemma 14 and 15 for details. As a result, we obtain

$$||W_t||_2 \le \sigma_1 \left(\rho_t + \eta + (1+\eta)\rho_{t-1}\right).$$
 (A.10)

• For $||N_t||_2$, we recall that

$$N_t = \sum_{i=1}^m p_i N_{t-1}^{(i)} \sim \mathcal{N}(0, \sum_i p_i^2 \sigma^2 ||Z_{t-1}^{(i)}||_{\max}^2).$$

Then by the bound of the largest singular value of subgaussian matrices (Rudelson and Vershynin, 2010) (see Lemma 20 in Appendix D), we have for any t, s > 0 and some constants C, c > 0,

$$\mathbb{P}\left(\frac{\|N_t\|_2}{\sqrt{\sum_i p_i^2 \sigma^2 \|Z_{t-1}^{(i)}\|_{\max}^2}} > C(\sqrt{d} + \sqrt{r}) + s\right) \le 2\exp(-cs^2). \tag{A.11}$$

Applying the union bound, we further have,

$$\mathbb{P}(\max_{t} \frac{\|N_{t}\|_{2}}{\sqrt{\sum_{i} p_{i}^{2} \sigma^{2} \|Z_{t-1}^{(i)}\|_{\max}^{2}}} > C\sqrt{d} + s) \le \lfloor \frac{T}{p} \rfloor \exp(-cs^{2}), \tag{A.12}$$

where we used the fact that r < d and note that constants c may be different from place to place. Choosing $s = O(\sqrt{\log \lfloor \frac{T}{p} \rfloor})$, then we have with probability larger than $1 - \lfloor T/p \rfloor^{-\alpha}$ that

$$\max_{t} ||N_{t}||_{2} \leq C \sqrt{\sum_{i} p_{i}^{2} \sigma^{2} \max_{t,i} ||Z_{t-1}^{(i)}||_{\max}^{2} (\sqrt{d} + \sqrt{\log \lfloor T/p \rfloor})}$$

$$\leq C \sqrt{\sum_{i} p_{i}^{2} \sigma \sqrt{r} (\sqrt{d} + \sqrt{\log \lfloor T/p \rfloor})}, \tag{A.13}$$

where α could be any positive constant and the last inequality follows from

$$\max_{t,i} ||Z_{t-1}^{(i)}||_{\max} \le \max_{t,i} ||Z_{t-1}^{(i)}||_{\infty} \le \sqrt{r} \max_{t,i} ||Z_{t-1}^{(i)}||_{2} = \sqrt{r}.$$

Combining (A.7), (A.10) and (A.13) and recalling the expression (A.3) and (A.6), we obtain that the perturbation noise $\mathcal{G}_t := H_t + W_t + N_t + N_t'$ satisfies that

$$\max_{t} \|\mathcal{G}_{t}\|_{2} \leq C \left(\sqrt{\sum_{i} p_{i}^{2} \sigma^{2} r} (\sqrt{d} + \sqrt{\log \lfloor T/p \rfloor}) + \sigma_{1} (\eta + (2 + \eta) \max_{t} \rho_{t}) \right), \quad (A.14)$$

with probability larger than $1 - \lfloor T/p \rfloor^{-\alpha}$. To lighten the notation, we denote the RHS of (A.14) by $\text{Err}(\sigma, d, T, p, k, \eta)$.

Third step: Establish convergence. Now we make use of the result in Lemma 12 to establish convergence. Note that in Lemma 12, there still exists an unknown term $\|\overline{Z}_t\|_{\mathbf{m}}$. We prove in Lemma 16 that

$$\|\overline{Z}_t\|_{\mathrm{m}} \ge 1 - (1 - p_1) \max_t \rho_t.$$

Denote

$$\epsilon' := c \frac{\operatorname{Err}(\sigma, d, T, p, k, \eta)}{(\sigma_k - \sigma_{k+1})(1 - (1 - p_1) \operatorname{max}_t \rho_t)},$$

then by (A.14),

$$5\max_{t} \|\mathcal{G}_{t}\|_{2} \le \epsilon'(\sigma_{k} - \sigma_{k+1}) \|\overline{Z}_{t}\|_{m}. \tag{A.15}$$

Hence the first condition in Lemma 12 is satisfied. For the second condition, we have that

$$5\max_t ||V_k^{\mathsf{T}} \mathcal{G}_t||_2 \le 5\max_t ||\mathcal{G}_t||_2$$

which implies that the second condition would be met automatically if $\epsilon' < \frac{\sqrt{r} - \sqrt{k-1}}{r\sqrt{d}}$, which is our condition. Consequently, by Lemma 12, we have after $T = O(\frac{\sigma_k}{\sigma_{k+1}} \log(\frac{d}{\epsilon'}))$ iterations,

$$\|(\mathbb{I}_d - \overline{Z}_T \overline{Z}_T^{\mathsf{T}}) V_k\|_2 \le \epsilon',$$

with probability larger than $1 - \lfloor T/p \rfloor^{-\alpha} - \tau^{-\Omega(r+1-k)} - e^{-\Omega(d)}$.

Proof of Theorem 6

By Lemma 14 and 15, the following results hold.

• If $\mathcal{F} = \mathcal{O}_r$, then

$$\rho_t \le \sqrt{2} \min \left\{ \frac{2\kappa^p \eta (1+\eta)^{p-1}}{(1-\eta)^p}, \frac{\eta \sigma_1}{\delta_k} + 2\gamma_k^{p/4} \max_{i \in [m]} \tan \theta_k(Z_{\tau(t)}, V_k^{(i)}) \right\}.$$

with the parameters $\delta_k = \min_{i \in [m]} \delta_k^{(i)}$ with $\delta_k^{(i)} = \min\{|\sigma_j(M_i) - \sigma_k(M)| : j \ge k + 1\}$ and $\gamma_k = \max\{\max_{i \in [m]} \frac{\sigma_{k+1}(M_i)}{\sigma_k(M_i)}, \frac{\sigma_{k+1}(M)}{\sigma_k(M)}\}$. By requiring $\eta \kappa \le 1/3$, and Wely's inequality, we have

$$\sigma_{j+1}(M) - \sigma_k(M) - \frac{1}{3}\sigma_d(M) \le \sigma_j(M_i) - \sigma_k(M) \le \sigma_{j+1}(M) - \sigma_k(M) + \frac{1}{3}\sigma_d(M).$$

Hence, $\delta_k \asymp \sigma_{k+1}(M) - \sigma_k(M)$. By requiring $\eta \leq 1/p$, we have $\frac{(1+\eta)^{p-1}}{(1-\eta)^p} \leq \frac{(1+1/p)^{p-1}}{(1-1/p)^p} \leq$ e^2 . Define $C_t = \max_{i \in [m]} \tan \theta_k(Z_{\tau(t)}, V_k^{(i)})$. Actually, we have shown in Theorem 5 that $\sin \theta_k(Z_{\tau(t)}, V_k) \leq \epsilon'$ for sufficiently large t. By the condition that ϵ' is small enough, we can obtain $\lim_{t \to \infty} \theta_k(Z_{\tau(t)}, V_k) = 0$. Then, we have

$$\limsup_{t\to\infty} C_t = \limsup_{t\to\infty} \max_{i\in[m]} \tan\theta_k(Z_{\tau(t)}, V_k^{(i)}) \leq \max_{i\in[m]} \tan\theta_k(V_k, V_k^{(i)}) \leq \max_{i\in[m]} \tan\arg\sin\frac{\eta\sigma_1}{\delta_k} = O(\eta),$$

where the last inequality follows from the Davis-Kahan theorem (see Lemma 22).

• If $\mathcal{F} = {\mathbb{I}_r}$, then

$$\rho_t \le 4\sqrt{2k}p\kappa^p\eta(1+\eta)^{p-1} \le 4e\sqrt{2k}p\kappa^p\eta,$$

where the last inequality requires $\eta \leq 1/p$.

Simply put together, we confirm that the bounds of ρ_t in Theorem 6 hold.

B. Partial participation

Proof of Theorem 7

We use the definition of (ε, δ) -DP to prove. To simplyfy the notation, we use \mathcal{E}_i^t to denote that event that the *i*th local machine is selected in the *t*th iteration. Consider $t \in \mathcal{I}_T^p$ and $i \in [m]$, and two neighboring databases M_i and M_i' . By the choice of σ , we observe that

$$\mathbb{P}(Y_t^{(i)} \mid M_i) = \mathbb{P}(Y_t^{(i)} \mid M_i, \mathcal{E}_i^t) \cdot q_i + \mathbb{P}(Y_t^{(i)} \mid M_i, (\mathcal{E}_i^t)^c) \cdot (1 - q_i) \\
\leq \left(\exp(\frac{\varepsilon}{\lfloor T/p \rfloor}) \mathbb{P}(Y_t^{(i)} \mid M_i', \mathcal{E}_i^t) + \frac{\delta / \max_i q_i}{\lfloor T/p \rfloor} \right) \cdot q_i + 0 \\
\leq \exp(\frac{\varepsilon}{\lfloor T/p \rfloor}) \mathbb{P}(Y_t^{(i)} \mid M_i') + \frac{\delta}{\lfloor T/p \rfloor}, \tag{B.1}$$

where the first inequality used the Gaussian mechanism and the assumption that when ith machine is not selected, the machine do not output anything. Hence, each iteration produces $(\frac{\varepsilon}{|T/p|}, \frac{\delta}{|T/p|})$ -differentially private $Y_t^{(i)}$ for each fixed i and t, and hence arriving $(\frac{\varepsilon}{|T/p|}, \frac{\delta}{|T/p|})$ -central differential privacy for fixed t.

On the other hand, it is easy to see that the noise added to the aggregated $Y_t^{(i)}$'s by the server can ensure $(\frac{\varepsilon}{|T/p|}, \frac{\delta}{|T/p|})$ -differential privacy when $t \in \mathcal{I}_T^p$.

Finally, considering that $\lceil T/p \rceil$ iterations are required for communication, Algorithm 3 can obtain $(2\varepsilon, 2\delta)$ -differential privacy by Proposition 3.

Proof of Theorem 8

Proof sketch of Theorem 8: Similar to the proof under the full participation setting, we first define a virtual sequence

$$\overline{Z}_t := \frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} Z_t^{(i)} O_t^{(i)}.$$

Here $O_t^{(i)} \in \mathbb{R}^{r \times r}$ is defined as

$$O_t^{(i)} = \begin{cases} \mathbb{I}_r & \text{if } t \in \mathcal{I}_T^p \\ D_{t+1}^{(i)} & \text{if } t \notin \mathcal{I}_T^p. \end{cases}$$

Then, we will write \overline{Z}_t in the following recursive manner,

$$\overline{Z}_t = [M\overline{Z}_{t-1} + \mathcal{G}_t]R_t^{-1},$$

where R_t^{-1} is the a reversible matrix to be specified, and \mathcal{G}_t is some noisy perturbation. It turns out that \mathcal{G}_t comes from three sources. Except for the DP's noise and the local iterates' perturbation, the random sampling of local devices also contributes to \mathcal{G}_t . To be specific, when $\tau(t) \neq \tau(t-1)$, it holds with high probability that $\mathcal{S}_{\tau(t)} \neq \mathcal{S}_{\tau(t-1)}$. As a result, we need to bound the bias term that the random sampling of local machines brings. With \mathcal{G}_t properly bounded, the convergence of would be established using Lemma 12.

First step: Perturbed iterate analysis. Recall that we defined a virtual sequence by

$$\overline{Z}_t := \frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} Z_t^{(i)} O_t^{(i)}.$$

where $O_t^{(i)}$ is \mathbb{I}_r if $t \in \mathcal{I}_T^p$, and is $D_{t+1}^{(i)}$ defined by

$$D_{t+1}^{(i)} = \underset{D \in \mathcal{F} \cap \mathcal{O}_n}{\operatorname{argmin}} \| Z_t^{(i)} D - Z_t^{(1)} \|_o,$$

if $t \notin \mathcal{I}_T^p$. For any t, we write $Y_t^{(i)} = Z_t^{(i)} R_t^{(i)}$ which should be kept in mind in the following proofs. Now we proceed to derive the iteration of \overline{Z}_t under $t \notin \mathcal{I}_T^p$ and $t \in \mathcal{I}_T^p$, respectively.

When $t \notin \mathcal{I}_T^p$, we have $Y_t^{(i)} = M_i Z_{t-1}^{(i)}$. Thus, given any invertible R^t (to be specified in Lemma 13), we have

$$\overline{Z}_{t} = \frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} Z_{t}^{(i)} O_{t}^{(i)}
= \frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} M_{i} Z_{t-1}^{(i)} O_{t-1}^{(i)} R_{t}^{-1} + \frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} Z_{t}^{(i)} [O_{t}^{(i)} R_{t} - R_{t}^{(i)} O_{t-1}^{(i)}] R_{t}^{-1}
= \frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} M Z_{t-1}^{(i)} O_{t-1}^{(i)} R_{t}^{-1} + \frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} (M_{i} - M) Z_{t-1}^{(i)} O_{t-1}^{(i)} R_{t}^{-1}
+ \frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} Z_{t}^{(i)} [O_{t}^{(i)} R_{t} - R_{t}^{(i)} O_{t-1}^{(i)}] R_{t}^{-1}
:= (\mathcal{J}_{t} + H_{t} + W_{t}) R_{t}^{-1},$$
(B.2)

for which \mathcal{J}_t could be further expressed as

$$\begin{split} \mathcal{J}_{t} &= \frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(t-1)}} MZ_{t-1}^{(i)} O_{t-1}^{(i)} + \frac{1}{K} \left(\sum_{i \in \mathcal{S}_{\tau(t)}} MZ_{t-1}^{(i)} O_{t-1}^{(i)} - \sum_{i \in \mathcal{S}_{\tau(t-1)}} MZ_{t-1}^{(i)} O_{t-1}^{(i)} \right) \\ &= M\overline{Z}_{t-1} + \left(\sum_{i \in \mathcal{S}_{\tau(t)}} MZ_{t-1}^{(i)} O_{t-1}^{(i)} / K - \mathbb{E}_{\mathcal{S}_{\tau(t)}} \left[\sum_{i \in \mathcal{S}_{\tau(t)}} MZ_{t-1}^{(i)} O_{t-1}^{(i)} / K \right] \right) \\ &+ \left(\mathbb{E}_{\mathcal{S}_{\tau(t)}} \left[\sum_{i \in \mathcal{S}_{\tau(t)}} MZ_{t-1}^{(i)} O_{t-1}^{(i)} / K \right] - \sum_{i \in \mathcal{S}_{\tau(t-1)}} MZ_{t-1}^{(i)} O_{t-1}^{(i)} / K \right) \\ &= M\overline{Z}_{t-1} + \left(\sum_{i \in \mathcal{S}_{\tau(t)}} MZ_{t-1}^{(i)} O_{t-1}^{(i)} / K - \sum_{i=1}^{m} p_{i} MZ_{t-1}^{(i)} O_{t-1}^{(i)} \right) + \left(\sum_{i=1}^{m} p_{i} MZ_{t-1}^{(i)} O_{t-1}^{(i)} - \sum_{i \in \mathcal{S}_{\tau(t-1)}} MZ_{t-1}^{(i)} O_{t-1}^{(i)} / K \right) \\ &:= M\overline{Z}_{t-1} + E_{t} + F_{t}. \end{split}$$

$$(B.3)$$

Combining (B.3) with (B.2), we obtain when $t \notin \mathcal{I}_T^p$ that,

$$\overline{Z}_t = (M\overline{Z}_{t-1} + E_t + F_t + H_t + W_t)R_t^{-1},$$
(B.4)

where note that when $\tau(t) = \tau(t-1)$, we have $E_t + F_t \equiv 0$.

On the other hand, when $t \in \mathcal{I}_T^p$, the synchronization happens and two round of Gaussian noise is added. Thereby, for all $i \in \mathcal{S}_{\tau(t)}$,

$$Y_{t}^{(i)} = \frac{1}{K} \sum_{l \in \mathcal{S}_{\tau(t)}} M_{l} Z_{t-1}^{(l)} D_{t}^{(l)} + \frac{1}{K} \sum_{l \in \mathcal{S}_{\tau(t)}} N_{t-1}^{(l)} + N_{t}'$$

$$:= \frac{1}{K} \sum_{l \in \mathcal{S}_{\tau(t)}} M_{l} Z_{t-1}^{(l)} D_{t}^{(l)} + N_{t} + N_{t}', \tag{B.5}$$

where

$$N_{t-1}^{(l)} \sim \mathcal{N}(0, \|Z_{t-1}^{(l)}\|_{\max}^2 \sigma^2)^{d \times r} \text{ and } N_t' \sim \mathcal{N}(0, \max_i \|Z_{t-1}^{(i)} D_t^{(i)}\|_{\max}^2 \sigma'^2),$$

with σ and σ' being defined in Algorithm 3. Using similar treatments as in (B.2) and (B.3), we have when $t \in \mathcal{I}_T^p$ that,

$$\overline{Z}_t = (M\overline{Z}_{t-1} + E_t + F_t + H_t + N_t + N_t')R_t^{-1},$$
(B.6)

where note that similar to the full participation protocol, W_t does not appear when $t \in \mathcal{I}_T^p$.

Second step: Bound the noise term. We proceed to bound E_t , F_t , H_t , W_t , N_t and N'_t , respectively. Note that, similar to the full participation scheme, $||N'_t||_2$ is of smaller order than $||N_t||_2$, hence we only deal with N_t . Also note that E_t and F_t behave similarly, so we only bound one of them.

• For $||E_t||_2$, we denote

$$E_{t} := \sum_{i \in \mathcal{S}_{\tau(t)}} M Z_{t-1}^{(i)} O_{t-1}^{(i)} / K - \sum_{i=1}^{m} p_{i} M Z_{t-1}^{(i)} O_{t-1}^{(i)}$$

$$= M \left(\sum_{i \in \mathcal{S}_{\tau(t)}} Z_{t-1}^{(i)} O_{t-1}^{(i)} / K - \sum_{i=1}^{m} p_{i} Z_{t-1}^{(i)} O_{t-1}^{(i)} \right) := M S_{t}$$
(B.7)

We will make use of the matrix Bernstein inequality (Tropp (2015), restated in Lemma 21). Consider only the randomness that the $S_{\tau(t)}$ brings, we have

$$\mathbb{E}_{S_{\tau(t)}}(S_t) = 0$$
 and $||S_t||_2 \le 2$.

Define

$$\nu(S_t) = \max\{\|\mathbb{E}(S_t S_t^{\mathsf{T}})\|_2, \|\mathbb{E}(S_t^{\mathsf{T}} S_t)\|_2\}.$$
(B.8)

Then applying Lemma 21 with $Z = S_t$ and L = 2, for any $t \ge \nu(S_t)/2$, we have

$$\mathbb{P}(\|S_t\|_2 \ge t) \le (d+r) \cdot e^{-3t/16}. \tag{B.9}$$

Choosing $t = O(\nu(S_t)(\log(d+r) + \log\lfloor T/p \rfloor))$, then (B.9) implies

$$||S_t||_2 \le C\nu(S_t)(\log(d+r) + \log|T/p|),$$
 (B.10)

with probability larger than $1 - \lfloor T/p \rfloor^{-\gamma'} \cdot e^{-c\nu(S_t)}/(d+r)^{\beta}$, where $\beta, \gamma' > 0$ is some positive constant. Now we turn to bound $\nu(S_t)$. Denote $\xi = \sum_{i=1}^m p_i Z_{t-1}^{(i)} O_{t-1}^{(i)}$, then

$$S_{t}S_{t}^{\mathsf{T}} = \frac{1}{K^{2}} \sum_{i \in \mathcal{S}_{\tau(t)}} Z_{t-1}^{(i)} O_{t-1}^{(i)} \sum_{i \in \mathcal{S}_{\tau(t)}} (Z_{t-1}^{(i)} O_{t-1}^{(i)})^{\mathsf{T}} - \xi \frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} (Z_{t-1}^{(i)} O_{t-1}^{(i)})^{\mathsf{T}} - \frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} Z_{t-1}^{(i)} O_{t-1}^{(i)} \xi^{\mathsf{T}} + \xi \xi^{\mathsf{T}}.$$

$$(B.11)$$

To shorten the notation, we denote $Z_{t-1}^{o(i)} := Z_{t-1}^{(i)} O_{t-1}^{(i)}$. Taking expectation with respect to $S_{\tau(t)}$, we then have

$$\mathbb{E}_{\mathcal{S}_{\tau(t)}} S_{t} S_{t}^{\mathsf{T}} = \mathbb{E}_{\mathcal{S}_{\tau(t)}} \left[\frac{1}{K^{2}} \sum_{i \in \mathcal{S}_{\tau(t)}} Z_{t-1}^{o(i)} \sum_{i \in \mathcal{S}_{\tau(t)}} (Z_{t-1}^{o(i)})^{\mathsf{T}} \right] - \xi \xi^{\mathsf{T}} \\
= \frac{1}{K^{2}} \mathbb{E}_{\mathcal{S}_{\tau(t)}} \left[\sum_{i \neq j, i, j \in \mathcal{S}_{\tau(t)}} Z_{t-1}^{o(i)} (Z_{t-1}^{o(j)})^{\mathsf{T}} + \sum_{i = j, i, j \in \mathcal{S}_{\tau(t)}} Z_{t-1}^{o(i)} (Z_{t-1}^{o(j)})^{\mathsf{T}} \right] - \xi \xi^{\mathsf{T}} \\
= \frac{K(K-1)}{K^{2}} \sum_{i \neq j, i, j = 1}^{m} p_{i} p_{j} Z_{t-1}^{o(i)} (Z_{t-1}^{o(j)})^{\mathsf{T}} + \frac{K}{K^{2}} \sum_{i = j = 1}^{m} p_{i} Z_{t-1}^{o(i)} (Z_{t-1}^{o(j)})^{\mathsf{T}} - \sum_{i, j = 1}^{m} p_{i} p_{j} Z_{t-1}^{o(i)} (Z_{t-1}^{o(j)})^{\mathsf{T}} \\
= -\frac{1}{K} \sum_{i \neq i, i, i = 1}^{m} p_{i} p_{j} Z_{t-1}^{o(i)} (Z_{t-1}^{o(j)})^{\mathsf{T}} + \sum_{i = 1}^{m} p_{i} (\frac{1}{K} - p_{i}) Z_{t-1}^{o(i)} (Z_{t-1}^{o(i)})^{\mathsf{T}}. \tag{B.12}$$

As a result,

$$\|\mathbb{E}_{\mathcal{S}_{\tau(t)}}(S_t S_t^{\mathsf{T}})\|_2 \le \frac{1}{K} \sum_{i \neq i, i, j=1}^m p_i p_j + \sum_{i=1}^m p_i^2 + \frac{1}{K} := \phi(K).$$
 (B.13)

Similarly, we could obtain

$$\|\mathbb{E}_{\mathcal{S}_{\tau(t)}}(S_t^{\mathsf{T}}S_t)\|_2 \le \phi(K).$$

Hence we have $\nu(S_t) \leq \phi(K)$ by recalling the definition of $\nu(S_t)$. Consequently, by (B.10), (B.7), and the union bound, we have

$$\max_{t} ||E_{t} + F_{t}||_{2} \leq \max_{t \in \mathcal{I}_{T}^{p}} 2||E_{t}||_{2} \leq C\sigma_{1} \left(\frac{1}{K} \sum_{i \neq j, i, j=1}^{m} p_{i} p_{j} + \sum_{i=1}^{m} p_{i}^{2} + \frac{1}{K}\right) (\log(d+r) + \log\lfloor T/p \rfloor),$$
(B.14)

with probability larger than $1 - \lfloor T/p \rfloor^{-\gamma} e^{-c\phi(K)}/(d+r)^{\beta}$, where $\beta, \gamma > 0$ is some positive constant.

• For $||H_t||_2$, recall that

$$H_t = \frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} (M_i - M) Z_{t-1}^{(i)} O_{t-1}^{(i)}.$$

It is easy to see that for any $t, \mathcal{S}_{\tau(t)}$,

$$||H_t||_2 \le \frac{1}{K} \cdot K \cdot \eta ||M||_2 = \eta \sigma_1.$$
 (B.15)

• For $||W_t||_2$, we have

$$W_t := \frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} Z_{t-1}^{(i)}(O_t^{(i)} R_t - R_t^{(i)} O_{t-1}^{(i)}) \le \max_i \|O_t^{(i)} R_t - R_t^{(i)} O_{t-1}^{(i)}\|_2.$$

We specify in Lemma 13 the choice of R^t and prove that for any $i \in [m]$,

$$||O_t^{(i)}R_t - R_t^{(i)}O_{t-1}^{(i)}||_2 \le \sigma_1(M_1)||Z_t^{(i)}O_t^{(i)} - Z_t^{(1)}||_2 + ||M_1 - M_i||_2 + \sigma_1(M_i)||Z_{t-1}^{(i)}O_{t-1}^{(i)} - Z_{t-1}^{(1)}||_2$$

$$\le \sigma_1\left(\rho_t + \eta + (1+\eta)\rho_{t-1}\right),$$

where ρ_t is defined as

$$\rho_t = \max_i \|Z_t^{(i)} O_t^{(i)} - Z_t^{(1)}\|_2 = \max_i \|Z_t^{(i)} D_{t+1}^{(i)} - Z_t^{(1)}\|_2,$$

and upper bounded as in (3.6) and (3.7); see Lemma 14 and 15 for details. As a result,

$$||W_t||_2 \le \sigma_1 \left(\eta + (2 + \eta) \max_t \rho_t \right).$$
 (B.16)

• For $||N_t||_2$, conditioning on $\mathcal{S}_{\tau(t)}$, we have

$$N_t \sim \mathcal{N}(0, \frac{1}{K^2} \sigma^2 \sum_{l \in \mathcal{S}_{\tau(t)}} \|Z_{t-1}^{(l)}\|_{\max}^2).$$

Applying the bound of the largest singular value of subgaussian matrices (Rudelson and Vershynin, 2010) (see Lemma 20), we have for any C, c > 0,

$$\mathbb{P}\left(\frac{\|N_t\|_2}{\sigma_{\max_{t,l}}\|Z_{t-1}^{(l)}\|_{\max}/\sqrt{K}} > C(\sqrt{d} + \sqrt{r}) + s\right) \le 2\exp(-cs^2),\tag{B.17}$$

where we enlarged the variance of Gaussian noise to lighten the notation. Moreover, using the union bound and choosing $s = O(\sqrt{\log \lfloor \frac{T}{p} \rfloor})$, we have with probability larger than $1 - \lfloor T/p \rfloor^{-\alpha}$ that

$$\max_{t} ||N_{t}||_{2} \le CK^{-1/2} \sigma \max_{t,l} ||Z_{t-1}^{(l)}||_{\max} (\sqrt{d} + \sqrt{\log|T/p|}). \tag{B.18}$$

Putting (B.14), (B.15), (B.16) and (B.18) together, and recalling (B.4) and (B.6), then the perturbation noise $\mathcal{G}_t := E_t + F_t + H_t + W_t + N_t + N_t'$ satisfies that

$$\max_t ||G_t||_2 \lesssim K^{-1/2} \sigma \sqrt{r} (\sqrt{d} + \sqrt{\log \lfloor T/p \rfloor})$$

$$+ \sigma_{1} \left(\frac{1}{K} \sum_{i \neq j, i, j=1}^{m} p_{i} p_{j} + \sum_{i=1}^{m} p_{i}^{2} + \frac{1}{K} \right) \left(\log(d+r) + \log \lfloor T/p \rfloor \right) + \sigma_{1} \left(\eta + (2+\eta) \max_{t} \rho_{t} \right),$$
(B.19)

with probability larger than $1 - \lfloor T/p \rfloor^{-\gamma} \cdot e^{-c\phi(K)}/(d+r)^{\beta} - \lfloor T/p \rfloor^{-\alpha}$ for positive constants α, β, γ . For notational simplicity, we denote the RHS of (B.19) as $\operatorname{Err}(\sigma, K, d, T, p, \eta, k, r)$.

Third step: Establish convergence. At last, we use Lemma 12 to establish convergence. Denote

$$\epsilon'' := c \frac{\operatorname{Err}(\sigma, K, d, T, p, \eta, k, r)}{(\sigma_k - \sigma_{k+1})(1 - \max_t \rho_t)},$$

then by (B.19),

$$5\max_t \|\mathcal{G}_t\|_2 \le \epsilon''(\sigma_k - \sigma_{k+1}) \|\overline{Z}_t\|_{\mathrm{m}},$$

where we used the fact in Lemma 16 that

$$\|\overline{Z}_t\|_{\mathrm{m}} \geq 1 - \max_t \rho_t.$$

The first condition in Lemma 12 is thus satisfied. For the second condition, we have that

$$5\max_t ||V_k^{\mathsf{T}} \mathcal{G}_t||_2 \le 5\max_t ||\mathcal{G}_t||_2,$$

which implies that the second condition would be met automatically if $\epsilon'' < \frac{\sqrt{r} - \sqrt{k-1}}{\tau \sqrt{d}}$, which is our condition. Overall, by Lemma 12, we have after $T = O(\frac{\sigma_k}{\sigma_{k+1}} \log(\frac{d}{\epsilon''}))$ iterations,

$$\|(\mathbb{I}_d - \overline{Z}_T \overline{Z}_T^{\mathsf{T}}) V_k\|_2 \le \epsilon''$$

with probability larger than $1 - \lfloor T/p \rfloor^{-\gamma} \cdot e^{-c\phi(K)}/(d+r)^{\beta} - \lfloor T/p \rfloor^{-\alpha} - \tau^{-\Omega(r+1-k)} - e^{-\Omega(d)}$.

Proof of Theorem 9

This proof follows a similar strategy as that of the Theorem 8 but there exists some differences. Define a virtual sequence

$$\overline{Z}_t := \frac{m}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} p_i Z_t^{(i)} O_t^{(i)}.$$

Here $O_t^{(i)} \in \mathbb{R}^{r \times r}$ is defined as

$$O_t^{(i)} = \begin{cases} \mathbb{I}_r & \text{if } t \in \mathcal{I}_T^p \\ D_{t+1}^{(i)} & \text{if } t \notin \mathcal{I}_T^p. \end{cases}$$

We aim to use Lemma 12 to establish the convergence. In particular, we prove by the following three steps.

First step: Perturbed iterate analysis. For any t, we write $Y_t^{(i)} = Z_t^{(i)} R_t^{(i)}$. We proceed to derive the iteration of \overline{Z}_t under $t \notin \mathcal{I}_T^p$ and $t \in \mathcal{I}_T^p$, respectively.

When $t \notin \mathcal{I}_T^p$, we have $Y_t^{(i)} = M_i Z_{t-1}^{(i)}$. So, given any invertible R^t (to be specified in Lemma 13), we have

$$\overline{Z}_{t} = \frac{m}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} p_{i} Z_{t}^{(i)} O_{t}^{(i)}
= \frac{m}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} p_{i} M_{i} Z_{t-1}^{(i)} O_{t-1}^{(i)} R_{t}^{-1} + \frac{m}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} p_{i} Z_{t}^{(i)} [O_{t}^{(i)} R_{t} - R_{t}^{(i)} O_{t-1}^{(i)}] R_{t}^{-1}
= \frac{m}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} p_{i} M Z_{t-1}^{(i)} O_{t-1}^{(i)} R_{t}^{-1} + \frac{m}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} p_{i} (M_{i} - M) Z_{t-1}^{(i)} O_{t-1}^{(i)} R_{t}^{-1}
+ \frac{m}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} p_{i} Z_{t}^{(i)} [O_{t}^{(i)} R_{t} - R_{t}^{(i)} O_{t-1}^{(i)}] R_{t}^{-1}
:= (\mathcal{J}_{t} + H_{t} + W_{t}) R_{t}^{-1},$$
(B.20)

for which \mathcal{J}_t could be further written as

$$\begin{split} \mathcal{J}_{t} &= \frac{m}{K} \sum_{i \in \mathcal{S}_{\tau(t-1)}} p_{i} M Z_{t-1}^{(i)} O_{t-1}^{(i)} + \frac{m}{K} \left(\sum_{i \in \mathcal{S}_{\tau(t)}} p_{i} M Z_{t-1}^{(i)} O_{t-1}^{(i)} - \sum_{i \in \mathcal{S}_{\tau(t-1)}} p_{i} M Z_{t-1}^{(i)} O_{t-1}^{(i)} \right) \\ &= M \overline{Z}_{t-1} + \left(\sum_{i \in \mathcal{S}_{\tau(t)}} p_{i} M Z_{t-1}^{(i)} O_{t-1}^{(i)} \cdot \frac{m}{K} - \mathbb{E}_{\mathcal{S}_{\tau(t)}} \left[\sum_{i \in \mathcal{S}_{\tau(t)}} p_{i} M Z_{t-1}^{(i)} O_{t-1}^{(i)} \cdot \frac{m}{K} \right] \right) \\ &+ \left(\mathbb{E}_{\mathcal{S}_{\tau(t)}} \left[\sum_{i \in \mathcal{S}_{\tau(t)}} p_{i} M Z_{t-1}^{(i)} O_{t-1}^{(i)} \cdot \frac{m}{K} \right] - \sum_{i \in \mathcal{S}_{\tau(t-1)}} p_{i} M Z_{t-1}^{(i)} O_{t-1}^{(i)} \cdot \frac{m}{K} \right) \\ &= M \overline{Z}_{t-1} + \left(\sum_{i \in \mathcal{S}_{\tau(t)}} p_{i} M Z_{t-1}^{(i)} O_{t-1}^{(i)} \cdot \frac{m}{K} - \sum_{i=1}^{m} p_{i} M Z_{t-1}^{(i)} O_{t-1}^{(i)} \right) \\ &+ \left(\sum_{i=1}^{m} p_{i} M Z_{t-1}^{(i)} O_{t-1}^{(i)} - \sum_{i \in \mathcal{S}_{\tau(t-1)}} p_{i} M Z_{t-1}^{(i)} O_{t-1}^{(i)} \cdot \frac{m}{K} \right) \\ &:= M \overline{Z}_{t-1} + E_{t} + F_{t}. \end{split} \tag{B.21}$$

Combining (B.21) with (B.20), when $t \notin \mathcal{I}_T^p$, we have,

$$\overline{Z}_t = (M\overline{Z}_{t-1} + E_t + F_t + H_t + W_t)R_t^{-1},$$
(B.22)

where note that when $\tau(t) = \tau(t-1)$, $E_t + F_t \equiv 0$.

On the other side, when $t \in \mathcal{I}_T^p$, the synchronization happens and two round of Gaussian noise is added. Consequently, for all $i \in \mathcal{S}_{\tau(t)}$,

$$Y_{t}^{(i)} = \frac{m}{K} \sum_{l \in \mathcal{S}_{\tau(t)}} p_{l} M_{l} Z_{t-1}^{(l)} D_{t}^{(l)} + \frac{m}{K} \sum_{l \in \mathcal{S}_{\tau(t)}} p_{l} N_{t-1}^{(l)} + N_{t}'$$

$$:= \frac{m}{K} \sum_{l \in \mathcal{S}_{\tau(t)}} p_{l} M_{l} Z_{t-1}^{(l)} + N_{t} + N_{t}', \tag{B.23}$$

where

$$N_{t-1}^{(l)} \sim \mathcal{N}(0, \|Z_{t-1}^{(l)}\|_{\max}^2 \sigma^2)^{d \times r} \text{ and } N_t' \sim \mathcal{N}(0, \max_i \|Z_{t-1}^{(i)} D_t^{(i)}\|_{\max}^2 \sigma''^2),$$

with σ and σ'' being defined in Algorithm 3. Using similar calculations as in (B.20) and (B.21), we obtain,

$$\overline{Z}_{t} = \left(M\overline{Z}_{t-1} + P_{t} + \frac{m}{K} \cdot \left(\sum_{i \in \mathcal{S}_{\tau(t)}} p_{i}\right) \cdot \left(E_{t} + F_{t} + H_{t} + N_{t} + N_{t}'\right)\right) R_{t}^{-1}.$$
(B.24)

where

$$P_{t} := \frac{m^{2}}{K^{2}} \sum_{l \in \mathcal{S}_{\tau(t)}} p_{l} \sum_{i \in \mathcal{S}_{\tau(t)}} p_{i} M Z_{t-1}^{(i)} O_{t-1}^{(i)} - \frac{m}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} p_{i} M Z_{t-1}^{(i)} O_{t-1}^{(i)},$$
(B.25)

and W_t does not appear because of our definition of R_t in Lemma 13. Note that when p_i 's are all the same, then $m/K \sum_{l \in \mathcal{S}_{\tau(t)}} p_l = 1$ and thus $P_t \equiv 0$. Hence P_t can be regarded as bias that comes from the heterogeneity in the sample size of each local machine. Note that this term did not appear under Scheme 1.

Second step: Bound the noise term. In view of (B.22) and (B.24), we analyze $P_t, E_t, F_t, H_t, W_t, N_t, N_t'$, respectively. It is easy to see that $||N_t'||_2$ is of smaller order than $||N_t||_2$, hence we only deal with N_t . In addition, E_t and F_t are formulated similarly, so we only need to bound one of them.

• For $||P_t||_2$, we have

$$||P_t||_2 \le \left| \frac{m}{K} \sum_{l \in \mathcal{S}_{\tau(t)}} p_l - 1 \right| \cdot \frac{m}{K} ||M||_2 \cdot \sum_{l \in \mathcal{S}_{\tau(t)}} p_l \le \varsigma \sigma_1 \frac{m}{K} \mid \frac{m}{K} \varsigma - 1 \mid,$$
 (B.26)

where we denote

$$\varsigma := \max_{S} \sum_{l \in S \subset [m], |S| = K} p_l.$$

• For $||E_t||_2$, we denote

$$E_t := M(\frac{m}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} p_i Z_{t-1}^{(i)} O_{t-1}^{(i)} - \sum_{i=1}^m p_i Z_{t-1}^{(i)} O_{t-1}^{(i)}) := MS_t.$$

Specifically, we apply the matrix Bernstein inequality (Tropp, 2015) (see Lemma 21) to bound $||S_t||_2$. Consider merely the randomness that comes from $S_{\tau(t)}$, we have

$$\mathbb{E}_{S_{\tau(t)}}(S_t) = 0$$
 and $||S_t||_2 \le \frac{m}{K}\varsigma + 1$.

Define

$$\nu(S_t) = \max\{\|\mathbb{E}(S_t S_t^{\mathsf{T}})\|_2, \|\mathbb{E}(S_t^{\mathsf{T}} S_t)\|_2\}.$$

Then employing Lemma 21 with $Z = S_t$ and $L = \frac{m}{K}\varsigma + 1$, for any $t \ge \nu(S_t)/(\frac{m}{K}\varsigma + 1)$, we can obtain

$$\mathbb{P}(\|S_t\|_2 \ge t) \le (d+r) \cdot e^{-3t/16}. \tag{B.27}$$

Choosing $t = O(\nu(S_t)(\log(d+r) + \log|T/p|))$, then (B.27) yields

$$||S_t||_2 \le C\nu(S_t)(\log(d+r) + \log|T/p|),$$
 (B.28)

with probability higher than $1 - \lfloor T/p \rfloor^{-\gamma'} \cdot e^{-c\nu(S_t)/(\frac{m}{K}\varsigma+1)}/(d+r)^{\beta}$, where $\beta, \gamma' > 0$ is some positive constant. Now we bound $\nu(S_t)$. To lighten the notation, denote $\xi = \sum_{i=1}^m p_i Z_{t-1}^{o(i)}$ with $Z_{t-1}^{o(i)} := Z_{t-1}^{(i)} O_{t-1}^{(i)}$, then

$$S_{t}S_{t}^{\mathsf{T}} = \frac{m^{2}}{K^{2}} \sum_{i \in \mathcal{S}_{\tau(t)}} p_{i}Z_{t-1}^{o(i)} \sum_{i \in \mathcal{S}_{\tau(t)}} p_{i}(Z_{t-1}^{o(i)})^{\mathsf{T}} - \xi \frac{m}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} p_{i}(Z_{t-1}^{o(i)})^{\mathsf{T}} - \frac{m}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} p_{i}Z_{t-1}^{o(i)}\xi^{\mathsf{T}} + \xi\xi^{\mathsf{T}}.$$
(B.29)

Taking expectation with respect to $S_{\tau(t)}$, we then have

$$\begin{split} \mathbb{E}_{\mathcal{S}_{\tau(t)}} S_{t} S_{t}^{\intercal} &= \mathbb{E}_{\mathcal{S}_{\tau(t)}} \big[\frac{m^{2}}{K^{2}} \sum_{i \in \mathcal{S}_{\tau(t)}} p_{i} Z_{t-1}^{o(i)} \sum_{i \in \mathcal{S}_{\tau(t)}} p_{i} (Z_{t-1}^{o(i)})^{\intercal} \big] - \xi \xi^{\intercal} \\ &= \frac{m^{2}}{K^{2}} \mathbb{E}_{\mathcal{S}_{\tau(t)}} \big[\sum_{i \neq j, i, j \in \mathcal{S}_{\tau(t)}} p_{i} p_{j} Z_{t-1}^{o(i)} (Z_{t-1}^{o(j)})^{\intercal} + \sum_{i = j, i, j \in \mathcal{S}_{\tau(t)}} p_{i} p_{j} Z_{t-1}^{o(i)} (Z_{t-1}^{o(j)})^{\intercal} \big] - \xi \xi^{\intercal} \\ &= \frac{K(K-1)}{K^{2}} \sum_{i \neq j, i, j = 1}^{m} p_{i} p_{j} Z_{t-1}^{o(i)} (Z_{t-1}^{o(j)})^{\intercal} + \frac{m}{K} \sum_{i = j = 1}^{m} p_{i} p_{j} Z_{t-1}^{o(i)} (Z_{t-1}^{o(j)})^{\intercal} - \sum_{i, j = 1}^{m} p_{i} p_{j} Z_{t-1}^{o(i)} (Z_{t-1}^{o(j)})^{\intercal} \\ &= -\frac{1}{K} \sum_{i \neq i, i, j = 1}^{m} p_{i} p_{j} Z_{t-1}^{o(i)} (Z_{t-1}^{o(j)})^{\intercal} + (\frac{m}{K} - 1) \sum_{i = 1}^{m} p_{i}^{2} Z_{t-1}^{o(i)} (Z_{t-1}^{o(i)})^{\intercal}. \end{split} \tag{B.30}$$

Further,

$$\|\mathbb{E}_{\mathcal{S}_{\tau(t)}}(S_t S_t^{\mathsf{T}})\|_2 \le \frac{1}{K} \sum_{i \neq j, i, j=1}^m p_i p_j + \sum_{i=1}^m p_i^2 + \frac{m}{K} \sum_{i=1}^m p_i^2 := \psi(K).$$
 (B.31)

Analogously, we could prove

$$\|\mathbb{E}_{\mathcal{S}_{\tau(t)}}(S_t^{\mathsf{T}}S_t)\|_2 \leq \psi(K).$$

So $\nu(S_t) \leq \psi(K)$ by recalling the definition of $\nu(S_t)$. Consequently, by (B.28), the definition of E_t , and the union bound, we have

$$\max_{t \in \mathcal{I}_{T}^{p}} \|E_{t}\|_{2} \leq 2 \max_{t \in \mathcal{I}_{T}^{p}} \|E_{t}\|_{2} \leq C \sigma_{1} \left(\frac{1}{K} \sum_{i \neq j, i, j = 1}^{m} p_{i} p_{j} + \sum_{i = 1}^{m} p_{i}^{2} + \frac{m}{K} \sum_{i = 1}^{m} p_{i}^{2} \right) \times \left(\log(d + r) + \log|T/p| \right), \tag{B.32}$$

with probability larger than $1 - \lfloor T/p \rfloor^{-\gamma} \cdot e^{-c\psi(K)/(\frac{m}{K}\varsigma+1)}/(d+r)^{\beta}$, where $\beta, \gamma > 0$ is some positive constant.

• For $||H_t||_2$, recall that

$$H_t := \frac{m}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} p_i(M_i - M) Z_{t-1}^{(i)} O_{t-1}^{(i)}.$$

Recall the definition of ς , then it is easy to obtain that for any $t, \mathcal{S}_{\tau(t)}$,

$$||H_t||_2 \le \frac{m}{K} \cdot \varsigma \cdot \eta \sigma_1. \tag{B.33}$$

• For $||W_t||_2$, recall that

$$W_t := \frac{m}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} p_i Z_t^{(i)} [O_t^{(i)} R_t - R_t^{(i)} O_{t-1}^{(i)}].$$

Applying Lemma 13, we have,

$$||W_t||_2 \le \frac{m}{K} \varsigma \cdot \max ||O_t^{(i)} R_t - R_t^{(i)} O_{t-1}^{(i)}||_2 \le \frac{m}{K} \varsigma \cdot \sigma_1 \left(\eta + (2+\eta) \max_t \rho_t\right), \tag{B.34}$$

where ρ_t is defined as

$$\rho_t = \max_i \|Z_t^{(i)} O_t^{(i)} - Z_t^{(1)}\|_2 = \max_i \|Z_t^{(i)} D_{t+1}^{(i)} - Z_t^{(1)}\|_2,$$

and upper bounded as in (3.6) and (3.7); see Lemma 14 and 15 for details.

• For $||N_t||_2$, recall that

$$N_t := m/K \sum_{l \in \mathcal{S}_{\tau(t)}} p_l N_{t-1}^{(l)}.$$

Conditioning on $\mathcal{S}_{\tau(t)}$, we have

$$N_t \sim \mathcal{N}(0, \frac{m^2}{K^2} \sigma^2 \sum_{l \in \mathcal{S}_{\tau(t)}} p_l^2 ||Z_{t-1}^{(l)}||_{\max}^2).$$

To facilitate further notation, we denote

$$\zeta := \max_S \sum_{l \in S \subset [m], |S| = K} p_l^2.$$

Using the concentration inequality of the largest singular value of subgaussian matrices (Rudelson and Vershynin, 2010) (see Lemma 13), we have for any C, c > 0,

$$\mathbb{P}(\frac{\|N_t\|_2}{\sqrt{\zeta}\sigma\max_{t,l}\|Z_{t-1}^{(l)}\|_{\max}m/K} > C(\sqrt{d} + \sqrt{r}) + s) \le 2\exp(-cs^2).$$
 (B.35)

Further, applying the union bound and choosing $s = O(\sqrt{\log \lfloor \frac{T}{p} \rfloor})$, we have with probability larger than $1 - \lfloor T/p \rfloor^{-\alpha}$ that

$$\max_{t} ||N_{t}||_{2} \le C\sqrt{\zeta} \frac{m}{K} \sigma \max_{t,l} ||Z_{t-1}^{(l)}||_{\max} (\sqrt{d} + \sqrt{\log \lfloor T/p \rfloor}).$$
 (B.36)

Combining (B.26), (B.32), (B.33), (B.34) and (B.36) together, we obtain that the perturbation noise $\mathcal{G}_t := P_t + \max\{\frac{m}{K}\varsigma, 1\} \cdot \left(E_t + F_t + H_t + W_t + N_t + N_t'\right)$ satisfies that

$$\max_{t} \|G_{t}\|_{2} \lesssim \varsigma \sigma_{1} \frac{m}{K} \mid \frac{m}{K} \varsigma - 1 \mid + \max\{\frac{m}{K} \varsigma, 1\} \left\{ \sqrt{\zeta} \frac{m}{K} \sigma \sqrt{r} (\sqrt{d} + \sqrt{\log \lfloor T/p \rfloor}) + \sigma_{1} \left(\frac{1}{K} \sum_{i \neq i, i, i = 1}^{m} p_{i} p_{j} + \sum_{i = 1}^{m} p_{i}^{2} + \frac{m}{K} \sum_{i = 1}^{m} p_{i}^{2} \right) \cdot (\log(d+r) + \log \lfloor T/p \rfloor) + \frac{m}{K} \varsigma \cdot \sigma_{1} \left(\eta + (2+\eta) \max_{t} \rho_{t} \right) \right\}, \quad (B.37)$$

with probability larger than $1 - \lfloor T/p \rfloor^{-\gamma} \cdot e^{-c\psi(K)/(\frac{m}{K}\varsigma+1)}/(d+r)^{\beta} - \lfloor T/p \rfloor^{-\alpha}$ for positive constants α, β, γ . For simplicity, we denote the RHS of (B.37) as $\operatorname{Err}(\sigma, K, \zeta, \varsigma, d, T, p, \eta, k, r)$.

Third step: Establish convergence. Denote

$$\epsilon''' = c \frac{\operatorname{Err}(\sigma, K, \zeta, \varsigma, d, T, p, \eta, k, r)}{(\sigma_k - \sigma_{k+1}) \cdot \frac{m}{K} (\underline{\varsigma} - \varsigma \max_t \rho_t)},$$

and apply Lemma 12 and note Lemma 16 in the same way as what we did in proving Theorem 8, we arrive the results of Theorem 9.

C. Technical lemmas

The following lemma is a variant of Lemma 2.2 in Hardt and Price (2014) (see Lemma 17). Given the relation $\overline{Z}_t = M\overline{Z}_{t-1} + \mathcal{G}_t$, they require \overline{Z}_t to have orthonormal columns, i.e., $\overline{Z}_t^{\mathsf{T}}\overline{Z}_t = \mathbb{I}_r$. However, it is impossible in our analysis. As a remedy, we slightly change the lemma to allow arbitrary \overline{Z}_t . This will also change the condition on \mathcal{G}_t .

Lemma 11 Let $V_k \in \mathbb{R}^{d \times k}$ denote the top k eigenvectors of $M := \frac{1}{n} A^{\mathsf{T}} A$ and let $\sigma_1 \leq ... \leq \sigma_d$ denote its singular values. Let $\overline{Z}_t \in \mathbb{R}^{d \times r}$ for some $r \geq k$. Let \mathcal{G}_t satisfy

$$4\|V_k^{\mathsf{T}}\mathcal{G}_t\|_2 \leq (\sigma_k - \sigma_{k+1})\cos\theta_k(V_k, \overline{Z}_t)\|\overline{Z}_t\|_{\mathsf{m}} \quad \text{and} \quad 4\|\mathcal{G}\|_2 \leq (\sigma_k - \sigma_{k+1})\|\overline{Z}_t\|_{\mathsf{m}}\epsilon,$$

for some $\epsilon \leq 1$, where $\|\overline{Z}_t\|_{\mathrm{m}}$ denotes the minimum singular value of \overline{Z}_t . Then

$$\tan \theta_k(V_k, M\overline{Z}_t + \mathcal{G}_t) \le \max \left(\epsilon, \max \left(\epsilon, \left(\frac{\sigma_{k+1}}{\sigma_k}\right)^{1/4}\right) \tan \theta_k(V_k, \overline{Z}_t)\right),$$

where the LHS can be replaced by $\tan \theta_k(V_k, (M\overline{Z}_t + \mathcal{G}_t)R_t^{-1})$ with any reversible matrix R_t .

Proof The proof actually follows closely from that of Hardt and Price (2014). Hence, we here only show the main steps.

First, by the definition of angles between subspaces, the Lemma 2.2 in Hardt and Price (2014) obtain that,

$$\tan \theta_k(V_k, \frac{1}{n} A^{\mathsf{T}} A \overline{Z}_t + \mathcal{G}_t) \leq \max_{\|w\|_2 = 1, \Pi^* w = w} \frac{1}{\|V_k^{\mathsf{T}} \overline{Z}_t w\|_2} \cdot \frac{\sigma_{k+1} \|(V_k^{\perp})^{\mathsf{T}} \overline{Z}_t w\|_2 + \|(V_k^{\perp})^{\mathsf{T}} \mathcal{G}_t w\|_2}{\sigma_k - \|V_k^{\mathsf{T}} \mathcal{G}_t w\|_2 / \|V_k^{\mathsf{T}} \overline{Z}_t w\|_2},$$

where Π^* is the matrix projecting onto the smallest k principal angles of \overline{Z}_t . Define $\Delta = (\sigma_k - \sigma_{k+1})/4$. Then, by the assumption on \mathcal{G}_t ,

$$\max_{\|w\|_2=1,\Pi^*w=w} \frac{\|V_k^\intercal \mathcal{G}_t w\|_2}{\|V_k^\intercal \overline{Z}_t w\|_2} \leq \frac{\|V_k^\intercal \mathcal{G}_t\|_2}{\cos \theta_k(V_k, \overline{Z}_t) \|\overline{Z}_t\|_{\mathrm{m}}} \leq \Delta,$$

where we used Fact 2 on the principle angle in Appendix E. Similarly, using the fact that $\cos \theta \le 1 + \tan \theta$ for any angle θ , we have

$$\max_{\|w\|_2 = 1, \Pi^* w = w} \frac{\|(V_k^{\perp})^{\intercal} \mathcal{G}_t w\|_2}{\|V_k^{\intercal} \overline{Z}_t w\|_2} \leq \frac{\|\mathcal{G}_t\|_2}{\cos \theta_k(V_k, \overline{Z}_t) \|\overline{Z}_t\|_{\mathrm{m}}} \leq \epsilon \Delta (1 + \tan \theta_k(V_k, \overline{Z}_t)).$$

Given the above two inequalities, the remaining proofing strategy is the same with that of Hardt and Price (2014). Hence we here omit it. In addition, noting the Fact 1 in Appendix

E, we know that the result can be generalized to $\tan \theta_k(V_k, (M\overline{Z}_t + \mathcal{G}_t)R_t^{-1})$ with any reversible matrix R_t .

With Lemma 11 at hand, it is easy to derive an analogue of Corollary 1.1 in Hardt and Price (2014) (see Lemma 18). We summarize the results as the following lemma.

Lemma 12 Let k and r ($k \le r$) be the target rank and iteration rank, respectively. Let $V_k \in \mathbb{R}^{d \times k}$ denote the top k eigenvectors of $\frac{1}{n}A^{\mathsf{T}}A$ and let $\sigma_1 \le ... \le \sigma_d$ denote its singular values. Suppose $\overline{Z}_0 \sim \mathcal{N}(0, I_{d \times r})$. Assume the noisy power method iterates as follows,

$$\overline{Z}_t \leftarrow \frac{1}{n} A^{\mathsf{T}} A \overline{Z}_{t-1} + \mathcal{G}_t,$$

where \overline{Z}_t does not necessarily have orthonormal columns and \mathcal{G}_t is some noisy perturbation that satisfies

$$5\|\mathcal{G}_t\|_2 \le \epsilon(\sigma_k - \sigma_{k+1})\min_t \|\overline{Z}_t\|_{\mathbf{m}} \quad \text{and} \quad 5\|V_k^\mathsf{T}\mathcal{G}_t\|_2 \le (\sigma_k - \sigma_{k+1})\min_t \|\overline{Z}_t\|_{\mathbf{m}} \frac{\sqrt{r} - \sqrt{k-1}}{\tau\sqrt{d}},$$

for some fixed τ and $\epsilon < 1/2$. Then with all but $\tau^{-\Omega(r+1-k)} + e^{-\Omega(d)}$ probability, there exists an $T = O(\frac{\sigma_k}{\sigma_k - \sigma_{k+1}} \log(d\tau/\epsilon))$ so that after T steps

$$||(I - \overline{Z}_T \overline{Z}_T^{\mathsf{T}}) V_k||_2 \le \epsilon.$$

The result also holds for the sequence

$$\overline{Z}_t \leftarrow [\frac{1}{n}A^{\mathsf{T}}A\overline{Z}_{t-1} + \mathcal{G}_t]R_t^{-1},$$

with any reversible R_t .

Proof By Lemma 11 and the proofing techniques of Corollary 1.1 in Hardt and Price (2014) (see Lemma 18), the result follows.

In the next lemma, we specify the choice of R_t and analyze the residual error bound $\|O_t^{(i)}R_t - R_t^{(i)}O_{t-1}^{(i)}\|_2$ when $t \notin \mathcal{I}_T^p$. In particular, given a baseline data matrix M_o , R_t is the shadow matrix that depicts what the upper triangle matrix ought to be, if we start from the nearest synchronized matrix and perform QR factorization using the matrix M_o . We will set $M_o = M_1$ (by assuming $1 = \operatorname{argmax}_{i \in [m]} p_i$).

Lemma 13 (Choice of R_t) Fix any t and let $t_0 = \tau(t) \in \mathcal{I}_T^p$ be the latest synchronization step before t, then $t \geq \tau(t)$.

- If $t = t_0$, we define $R_t = R_t^{(i)}$ for any $i \in [m]$ since all $R_t^{(i)}$'s are equal.
- If $t > t_0$, given a baseline data matrix M_o , we define $R_t \in \mathbb{R}^{r \times r}$ recursively as the following. Let $Z_{t_0} = \overline{Z}_{t_0}$. For $l = t_0 + 1, \dots, t$, we use the following QR factorization to define R_t 's:

$$M_{o}Z_{l} = Z_{l+1}R_{l+1}.$$

With such choice of R_t 's, for any $i \in [m]$, we have

$$||O_{t}^{(i)}R_{t} - R_{t}^{(i)}O_{t-1}^{(i)}||_{2} \leq \sigma_{1}(M_{o})||Z_{t}^{(i)}O_{t}^{(i)} - Z_{t}||_{2} + ||M_{o} - M_{i}||_{2} + \sigma_{1}(M_{i})||Z_{t-1}^{(i)}O_{t-1}^{(i)} - Z_{t-1}||_{2}.$$
(C.1)

Proof Note that $t \notin \mathcal{I}_T^p$ and thus $t > t_0$. Let's fix some $i \in [m]$ and denote $\Delta M = M_i - M_o$. Based on FedPower, we have for $l = t_0 + 1, \dots, t$,

$$M_i Z_l^{(i)} = Z_{l+1}^{(i)} R_{t+1}^{(i)}.$$

Then,

$$Z_{l}^{(i)}R_{l}^{(i)}O_{l-1}^{(i)} = M_{i}Z_{l-1}^{(i)}O_{l-1}^{(i)}$$

$$= (M_{o} + \Delta M)(Z_{l-1} + \Delta Z_{l-1})$$

$$= M_{o}Z_{l-1} + \Delta M \cdot Z_{l-1} + M_{i} \cdot \Delta Z_{l-1}$$

$$:= M_{o}Z_{l-1} + E_{l-1} = Z_{l}R_{l} + E_{l-1}$$

where $E_{l-1} = \Delta M \cdot Z_{l-1} + M_i \cdot \Delta Z_{l-1}$ and $\Delta Z_{l-1} = Z_{l-1}^{(i)} O_{l-1}^{(i)} - Z_{l-1}$. Note that

$$Z_t^{(i)} R_t^{(i)} O_{t-1}^{(i)} = Z_t R_t + E_{t-1}.$$

Then we have

$$\begin{split} \|O_t^{(i)}R_t - R_t^{(i)}O_{t-1}^{(i)}\|_2 &= \|Z_t^{(i)}O_t^{(i)}R_t - Z_t^{(i)}R_t^{(i)}O_{t-1}^{(i)}\|_2 \\ &\stackrel{(a)}{=} \|Z_t^{(i)}O_t^{(i)}R_t - Z_tR_t - E_{t-1}\|_2 \\ &\leq \|(Z_t^{(i)}O_t^{(i)} - Z_t)R_t\|_2 + \|E_{t-1}\|_2 \\ &\stackrel{(b)}{\leq} \|Z_t^{(i)}O_t^{(i)} - Z_t\|_2 \|R_t\|_2 + \|\Delta M\|_2 + \|M_i\|_2 \|Z_{t-1}^{(i)}O_{t-1}^{(i)} - Z_{t-1}\|_2 \\ &\stackrel{(c)}{\leq} \sigma_1(M_o) \|Z_t^{(i)}O_t^{(i)} - Z_t\|_2 + \|M_o - M_i\|_2 + \sigma_1(M_i) \|Z_{t-1}^{(i)}O_{t-1}^{(i)} - Z_{t-1}\|_2 \end{split}$$

where (a) uses the equality of $Z_t^{(i)} R_t^{(i)} O_{t-1}^{(i)}$; (b) uses the definition of E_{t-1} and $O_t^{(i)} = D_t^{(i)}$; and (c) uses $||R_t||_2 \le ||M_o||_2 = \sigma_1(M_o)$.

Note that the bound (C.1) in the above lemma depends on the following unknown terms

$$\rho_t = \max_i \|Z_t^{(i)} O_t^{(i)} - Z_t\|_2 = \max_i \|Z_t^{(i)} D_{t+1}^{(i)} - Z_t\|_2.$$

Hence, in the next two lemmas, we provide the upper bounds for ρ_t in two cases, namely, $\mathcal{F} = \mathcal{O}_r$ and $\mathcal{F} = \{\mathbb{I}_r\}$. Before going on we note that Z_t in ρ_t is actually $Z_t^{(1)}$ as M_o is chose to be M_1 in Lemma 13.

Lemma 14 (Bound for ρ_t **when** $\mathcal{F} = \mathcal{O}_r$ **)** Let Assumption 1 hold with sufficiently small η . If $D_t^{(i)}$ is solved from

$$D_t^{(i)} = \underset{D \in \mathcal{F} \cap \mathcal{O}_r}{\operatorname{argmin}} \| Z_{t-1}^{(i)} D - Z_{t-1}^{(1)} \|_o$$

with $\mathcal{F} = \mathcal{O}_r$, where $\|\cdot\|_o$ can be either the Frobenius norm or the spectral norm though in the body text we use only the Frobenius norm, then

$$\rho_t \le \min \sqrt{2} \left\{ \frac{2\kappa^p p \eta (1+\eta)^{p-1}}{(1-\eta)^p}, \frac{\eta \sigma_1}{\delta_k} + 2\gamma_k^{p/4} \max_{i \in [m]} \tan \theta_k(Z_{\tau(t)}, V_k^{(i)}) \right\}.$$

where

- $\delta_k = \min_{i \in [m]} \delta_k^{(i)}$ with $\delta_k^{(i)} = \min\{ |\sigma_j(M_i) \sigma_k(M)| : j \ge k + 1 \};$
- $\gamma_k = \max\{\max_{i \in [m]} \frac{\sigma_{k+1}(M_i)}{\sigma_k(M_i)}, \frac{\sigma_{k+1}(M)}{\sigma_k(M)}\},$
- $\kappa = ||M||_2 ||M^{\dagger}||_2$ is the condition number of M;
- $p = t \tau(u), \ \tau(t) \in \mathcal{I}_T^p$ is defined as the nearest synchronization time before t.

Proof By Lemma 25, we have

$$||Z_{t-1}^i D_t^{(i)} - Z_{t-1}^{(1)}||_2 \le \sqrt{2} \operatorname{dist}(Z_{t-1}^{(i)}, Z_{t-1}^{(1)}),$$

so we only need to bound $\max_{i \in [m]} \operatorname{dist}(Z_t^{(i)}, Z_t^{(1)})$. We will bound each $\operatorname{dist}(Z_t^{(i)}, Z_t^{(1)})$ uniformly in two ways. Then the minimum of the two upper bounds holds for their maximum that is exactly ρ_t . Fix any $i \in [m]$ and $t \in [T]$. Let $\tau(t)$ be the latest synchronization step before t and $p = t - \tau(t)$ be the number of nearest local updates.

• For small p, by Lemma 23, it follows that

$$\begin{aligned} \operatorname{dist}(Z_{t}^{(i)}, Z_{t}^{(1)}) &= \operatorname{dist}(M_{i}^{p} Z_{\tau(t)}, M_{1}^{p} Z_{\tau(t)}) \\ &\leq \operatorname{dist}(M_{i}^{p} Z_{\tau(t)}, M^{p} Z_{\tau(t)}) + \operatorname{dist}(M^{p} Z_{\tau(t)}, M_{1}^{p} Z_{\tau(t)}) \\ &\leq \min\{\|(M_{i}^{p} Z_{\tau(t)})^{\dagger}\|_{2}, \|(M^{p} Z_{\tau(t)})^{\dagger}\|_{2}\}\|(M_{i}^{p} - M^{p}) Z_{\tau(t)}\|_{2} \\ &\quad + \min\{\|(M^{p} Z_{\tau(t)})^{\dagger}\|_{2}, \|(M_{1}^{p} Z_{\tau(t)})^{\dagger}\|_{2}\}\|(M^{p} - M_{1}^{p}) Z_{\tau(t)}\|_{2} \\ &\leq 2\kappa^{p} (1 + \eta)^{p} - 1 \leq 2\kappa^{p} \frac{(1 + \eta)^{p} - 1}{(1 - \eta)^{p}} \\ &\leq \frac{2\kappa^{p} p \eta (1 + \eta)^{p-1}}{(1 - \eta)^{p}} \end{aligned}$$

where $\kappa = ||M||_2 ||M^{\dagger}||_2$ is the condition number of M.

• For large p, let the top-k eigenspace of M_1 and M_i be respectively $V_k^{(1)}$ and $V_k^{(i)}$ (both of which are orthonormal). The k-largest eigenvalue of M is denoted by $\sigma_k(M_1)$ and similarly for $\sigma_k(M_i)$. Then by Lemma 22, we have

$$\operatorname{dist}(V_k, V_k^{(i)}) \le \frac{\|M_i - M\|_2}{\delta_k^{(i)}} \le \frac{\eta \sigma_1}{\delta_k^{(i)}}.$$

where $\sigma_1 = \sigma_1(M)$ and $\delta_k^{(i)} = \min\{|\sigma_j(M_i) - \sigma_k(M)| : j \ge k+1\}.$

Note that local updates are equivalent to noiseless power method. Then, using Lemma 17 and setting $\epsilon = 0$ and $\mathcal{G} = 0$ therein, we have

$$\tan \theta_k(Z_t^{(i)}, V_k^{(i)}) \le \left(\frac{\sigma_{k+1}(M_i)}{\sigma_k(M_i)}\right)^{1/4} \tan \theta_k(Z_{t-1}^{(i)}, V_k^{(i)}).$$

Hence,

$$\begin{split} \operatorname{dist}(Z_t^{(i)}, Z_t^{(1)}) & \leq \operatorname{dist}(Z_t^{(i)}, V_k^{(i)}) + \operatorname{dist}(V_k^{(i)}, V_k^{(1)}) + \operatorname{dist}(V_k^{(1)}, Z_t^{(1)}) \\ & \leq \frac{\eta \sigma_1}{\delta_k^{(i)}} + \left(\frac{\sigma_{k+1}(M_i)}{\sigma_k(M_i)}\right)^{p/4} \tan\theta_k(Z_{\tau(t)}, V_k^{(i)}) + \left(\frac{\sigma_{k+1}(M)}{\sigma_k(M)}\right)^{p/4} \tan\theta_k(Z_{\tau(t)}, V_k^{(1)}) \\ & \leq \frac{\eta \sigma_1}{\min_{i \in [m]} \delta_k^{(i)}} + 2\gamma_k^{p/4} \max_{i \in [m]} \tan\theta_k(Z_{\tau(t)}, V_k^{(i)}). \end{split}$$

Combining the two cases, we have

$$\rho_t \le \sqrt{2} \min \left\{ \frac{2\kappa^p p \eta (1+\eta)^{p-1}}{(1-\eta)^p}, \frac{\eta \sigma_1}{\delta_k} + 2\gamma_k^{p/4} \max_{i \in [m]} \tan \theta_k(Z_{\tau(t)}, V_k^{(i)}) \right\}.$$

Lemma 15 (Bound for ρ_t when $\mathcal{F} = \{I_r\}$) Let Assumption 1 hold with sufficiently small η . If $D_t^{(i)}$ is solved from

$$D_t^{(i)} = \underset{D \in \mathcal{F} \cap \mathcal{O}_r}{\operatorname{argmin}} \| Z_{t-1}^{(i)} D - Z_{t-1}^{(1)} \|_o,$$

with $\mathcal{F} = \{\mathbb{I}_r\}$, then

$$\rho_t \le 4\sqrt{2k}p\kappa^p\eta(1+\eta)^{p-1},$$

where $\kappa = ||M||_2 ||M^{\dagger}||_2$ is the condition number of M, $p = t - \tau(u)$, $\tau(t) \in \mathcal{I}_T^p$ is defined as the nearest synchronization time before t.

Proof In this case, we are going to bound $\rho_t = \max_{i \in [m]} ||Z^{(i)} - Z_t^{(1)}||_2$. Fix any $i \in [m]$ and $t \in [T]$. We will bound $||Z^{(i)} - Z_t^{(1)}||_2$ uniformly so that the bound holds for their maximum.

Fix any $i \in [m]$ and $t \in [T]$. Let $\tau(t)$ be the latest synchronization step before t and $p = t - \tau(t)$ be the number of nearest local updates. Note that $Z_t^{(i)}$ and $Z_t^{(1)}$ are the Q-factor of the QR factorization of $M_i^p Z_{\tau(t)}$ and $M_1^p Z_{\tau(t)}$. Let \tilde{Z}_t be the Q-factor of the QR factorization of $M^p Z_{\tau(t)}$. Then Lemma 19 yields

$$||Z_t^{(i)} - \tilde{Z}_t||_2 \le \sqrt{2k} \frac{||(M^p Z_{\tau(t)})^{\dagger}||_2 ||(M_i^p - M^p) Z_{\tau(t)}||_2}{1 - ||(M^p Z_{\tau(t)})^{\dagger}||_2 ||(M_i^p - M^p) Z_{\tau(t)}||_2} := \sqrt{2k} \frac{\omega}{1 - \omega}$$

where $\omega = \|(M^p Z_{\tau(t)})^{\dagger}\|_2 \|(M_i^p - M^p) Z_{\tau(t)}\|_2$ for short. If $\omega \leq 1/2$, then we have $\|Z_t^{(i)} - Z_t^{(i)}\|_2$ $\tilde{Z}_t\|_2 \leq 2\sqrt{2k}\omega$. Otherwise, we have $\omega \geq 1/2$ and $\|Z_t^{(i)} - \tilde{Z}_t\|_2 \leq 2 \leq \sqrt{2k} \leq 2\sqrt{2k}\omega$. Then we have for all $i \in [m]$,

$$||Z_t^{(i)} - \tilde{Z}_t||_2 \le 2\sqrt{2k} ||(M^p Z_{\tau(t)})^{\dagger}||_2 ||(M_i^p - M^p) Z_{\tau(t)}||_2.$$

Hence,

$$\rho_{t} = \|Z_{t}^{(i)} - Z_{t}^{(1)}\|_{2}
\leq \|Z_{t}^{(i)} - \tilde{Z}_{t}\|_{2} + \|\tilde{Z}_{t} - Z_{t}^{(1)}\|_{2}
\leq 2\sqrt{2k} \left[\|(M^{p}Z_{\tau(t)})^{\dagger}\|_{2} \|(M_{i}^{p} - M^{p})Z_{\tau(t)}\|_{2} + \|(M^{p}Z_{\tau(t)})^{\dagger}\|_{2} \|(M_{1}^{p} - M^{p})Z_{\tau(t)}\|_{2} \right]
\leq 4\sqrt{2k}\kappa^{p} \left[(1+\eta)^{p} - 1 \right]
\leq 4\sqrt{2k}p\kappa^{p}\eta(1+\eta)^{p-1},$$

where $\kappa = ||M||_2 ||M^{\dagger}||_2$ is the condition number of M.

The next lemma provide a lower bound for $\|\overline{Z}_t\|_{\mathrm{m}}$, which is needed when using Lemma 11 to carry out the convergence analysis of FedPower.

Lemma 16 (Bound for $\|\overline{Z}_t\|_{\rm m}$) Recall that

$$\rho_t = \max_i \|Z_t^{(i)} O_t^{(i)} - Z_t^{(1)}\|_2 = \max_i \|Z_t^{(i)} D_{t+1}^{(i)} - Z_t^{(1)}\|_2.$$

Then the following holds

(a) If
$$\overline{Z}_t := \sum_{i=1}^m p_i Z_t^{(i)} O_t^{(i)}$$
, then

$$\|\overline{Z}_t\|_{\rm m} \ge 1 - (1 - p_1) \max_t \rho_t := \mu_{t1};$$

(b) If
$$\overline{Z}_t := \frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} Z_t^{(i)} O_t^{(i)}$$
, then

$$\|\overline{Z}_t\|_{\mathbf{m}} \ge 1 - \max_t \rho_t := \mu_{t2};$$

(c) If
$$\overline{Z}_t := \frac{m}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} p_i Z_t^{(i)} O_t^{(i)}$$
, then

$$\|\overline{Z}_t\|_{\mathbf{m}} \ge \frac{m}{K} \left(\min_{S_{\tau(t)}} \sum_{l \in S_{\tau(t)}} p_l - \max_{S_{\tau(t)}} \sum_{l \in S_{\tau(t)}} p_l \right) \max_t \rho_t := \mu_{t3},$$

where

Proof It suffices to show $\|\overline{Z}_t^{\dagger}\|_2 \leq 1/\mu_t$ ($\mu_t = \mu_{t1}, \mu_{t2}, \mu_{t3}$) by noting $\|\overline{Z}_t\|_{\mathrm{m}} \|\overline{Z}_t^{\dagger}\|_2 = 1$. Next we show (a), (b) and (c),respectively. (a) For $\overline{Z}_t = \sum_{i=1}^m p_i Z_t^{(i)} O_t^{(i)}$, we have

(a) For
$$\overline{Z}_t = \sum_{i=1}^m p_i Z_t^{(i)} O_t^{(i)}$$
, we have

$$\|\overline{Z}_t^{\dagger}\|_2 = \max\{\|w\|_2 : \|\sum_{i=1}^m p_i Z_t^{(i)} O_t^{(i)} w\|_2 \le 1\}.$$

When $t \in \mathcal{I}_T^p$, $O_t^{(i)} = \mathbb{I}$ and $Z_t^{(i)}$'s are equal, hence $\|\overline{Z}_t^{\dagger}\|_2 = 1$ and the result holds naturally. When $t \notin \mathcal{I}_T^p$, we have

$$\begin{split} \| \sum_{i=1}^{m} p_{i} Z_{t}^{(i)} O_{t}^{(i)} w \|_{2} &= \| \sum_{i=1}^{m} p_{i} Z_{t}^{(i)} D_{t+1}^{(i)} w \|_{2} \\ &\geq \| \sum_{i=1}^{m} p_{i} Z_{t}^{(1)} w \|_{2} - \| \sum_{i=1}^{m} p_{i} (Z_{t}^{(i)} D_{t+1}^{(i)} - Z_{t}^{(1)}) w \|_{2} \\ &= \| w \|_{2} (1 - \sum_{i=1}^{m} p_{i} \| Z_{t}^{(i)} D_{t+1}^{(i)} - Z_{t}^{(1)} \|_{2}) \\ &= \| w \|_{2} (1 - \sum_{i \neq 1} p_{i} \| Z_{t}^{(i)} D_{t+1}^{(i)} - Z_{t}^{(1)} \|_{2}) \geq \| w \|_{2} \mu_{t1}. \end{split}$$

Hence,
$$\|\overline{Z}_t^{\dagger}\|_2 \leq 1/\mu_{t1}$$
.
(b) For $\overline{Z}_t = \frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} Z_t^{(i)} O_t^{(i)}$, we have

$$\begin{split} \|\frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} Z_{t}^{(i)} O_{t}^{(i)} w \|_{2} &= \|\frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} Z_{t}^{(i)} D_{t+1}^{(i)} w \|_{2} \\ &\geq \|\frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} Z_{t}^{(1)} w \|_{2} - \|\frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} (Z_{t}^{(i)} D_{t+1}^{(i)} - Z_{t}^{(1)}) w \|_{2} \\ &= \|w\|_{2} (1 - \frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} \|Z_{t}^{(i)} D_{t+1}^{(i)} - Z_{t}^{(1)} \|_{2}) \\ &\geq \|w\|_{2} \mu_{t2}, \end{split}$$

then the result follows.

(c) For
$$\overline{Z}_t = \frac{m}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} p_i Z_t^{(i)} O_t^{(i)}$$
, we have

$$\|\frac{m}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} p_i Z_t^{(i)} O_t^{(i)} w\|_2 = \|\frac{1}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} p_i Z_t^{(i)} D_{t+1}^{(i)} w\|_2$$

$$\geq \|\frac{m}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} p_i Z_t^{(1)} w\|_2 - \|\frac{m}{K} \sum_{i \in \mathcal{S}_{\tau(t)}} p_i (Z_t^{(i)} D_{t+1}^{(i)} - Z_t^{(1)}) w\|_2$$

$$\geq \|w\|_2 \mu_{t2},$$

and the result follows.

D. Auxiliary lemmas

Lemma 17 (Lemma 2.2 of Hardt and Price (2014)) Let $V_k \in \mathbb{R}^{d \times k}$ denote the top k eigenvectors of $\frac{1}{n}A^{\intercal}A$ and let $\sigma_1 \leq ... \leq \sigma_d$ denote its singular values. Let $Z \in \mathbb{R}^{d \times r}$ with $Z^{\intercal}Z = \mathbb{I}_r$ for some $r \geq k$. Let \mathcal{G} satisfy

$$4\|V_k^{\mathsf{T}}\mathcal{G}\|_2 \leq (\sigma_k - \sigma_{k+1}) \cos \theta_k(V_k, Z) \quad \text{and} \quad 4\|\mathcal{G}\|_2 \leq (\sigma_k - \sigma_{k+1})\epsilon,$$

for some $\epsilon \leq 1$. Then

$$\tan \theta_k(V_k, \frac{1}{n}A^{\mathsf{T}}AZ + \mathcal{G}) \leq \max \left(\epsilon, \max \left(\epsilon, \left(\frac{\sigma_{k+1}}{\sigma_k}\right)^{1/4}\right) \tan \theta_k(V_k, Z)\right).$$

Lemma 18 (Corollary 1.1 of Hardt and Price (2014)) Let k and r ($k \leq r$) be the target rank and iteration rank, respectively. Let $V_k \in \mathbb{R}^{d \times k}$ denote the top k eigenvectors of $\frac{1}{n}A^{\mathsf{T}}A$ and let $\sigma_1 \leq ... \leq \sigma_d$ denote its singular values. Suppose $Z_0 \sim \mathcal{N}(0, I_{d \times r})$. Assume the noisy power method iterates as follows,

$$Y_t \leftarrow \frac{1}{n} A^{\mathsf{T}} A Z_{t-1} + \mathcal{G}_t \quad \text{and} \quad Z_t \leftarrow \operatorname{orth}(Y_t),$$

where $Z_t \in \mathbb{R}^{d \times r}$ with $Z_t^\intercal Z_t = \mathbb{I}_r$ and \mathcal{G}_t is some noisy perturbation that satisfies

$$5\|\mathcal{G}_t\|_2 \le \epsilon(\sigma_k - \sigma_{k+1})$$
 and $5\|V_k^\mathsf{T}\mathcal{G}_t\|_2 \le (\sigma_k - \sigma_{k+1}) \frac{\sqrt{r} - \sqrt{k-1}}{\tau\sqrt{d}}$,

for some fixed τ and $\epsilon < 1/2$. Then with all but $\tau^{-\Omega(r+1-k)} + e^{-\Omega(d)}$ probability, there exists an $T = O(\frac{\sigma_k}{\sigma_k - \sigma_{k+1}} \log(d\tau/\epsilon))$ so that after T steps

$$||(I - Z_T Z_T^{\mathsf{T}}) V_k||_2 \le \epsilon.$$

Lemma 19 Let $A \in \mathbb{R}^{d \times k}$ with $d \geq k$ be any matrix with full rank. Denote by its QR factorization as A = QR where Q is an orthogonal metrix. Let E be some perturbation matrix and $A + E = \tilde{Q}\tilde{R}$ the resulting QR factorization of A + E. When $\|E\|_2 \|A^{\dagger}\|_2 < 1$, A + E is of full rank. What's more, it follows that

$$\|\tilde{Q} - Q\|_2 \le \sqrt{2k} \frac{\|A^{\dagger}\|_2 \|E\|_2}{1 - \|A^{\dagger}\|_2 \|E\|_2}.$$

Proof Actually, we have

$$\|\tilde{Q} - Q\|_F \stackrel{(a)}{\leq} \frac{\sqrt{2}\|E\|_F}{\|E\|_2} \ln \frac{1}{1 - \|A^{\dagger}\|_2 \|E\|_2} \stackrel{(b)}{\leq} \sqrt{2} \frac{\|A^{\dagger}\|_2 \|E\|_F}{1 - \|A^{\dagger}\|_2 \|E\|_2} \stackrel{(c)}{\leq} \sqrt{2k} \frac{\|A^{\dagger}\|_2 \|E\|_2}{1 - \|A^{\dagger}\|_2 \|E\|_2}$$

where (a) comes from Theorem 5.1 in Sun (1995); (b) uses $\ln(1+x) \le x$ for all x > -1; and (c) uses $||E||_F \le \sqrt{k}||E||_2$.

Lemma 20 (Proposition 2.4 of Rudelson and Vershynin (2010)) Let A be an $N \times n$ random matrix whose entries are independent mean zero sub-gaussian random variables and whose subgaussian moments are bounded by 1. Then

$$\mathbb{P}(\|A\|_2 > C(\sqrt{N} + \sqrt{n}) + t) \le 2\exp(-ct^2), \quad t \ge 0,$$

where c and C denote positive absolute constants.

Lemma 21 (Matrix Bernstein inequality (Tropp (2015), Chapter 6)) Consider a finite sequence $\{S_k\}$ of independent, random matrices with common dimension $d_1 \times d_2$. Assume that

$$\mathbb{E}S_k = 0$$
 and $||S_k||_2 \le L$ for each index k .

Introduce the random matrix

$$Z = \sum_{k} S_k.$$

Let $\nu(Z)$ be the matrix variance statistics of the sum:

$$\nu(Z) = \max\{\|\mathbb{E}(ZZ^{\mathsf{T}})\|_2, \|\mathbb{E}(Z^{\mathsf{T}}Z)\|_2\}.$$

Then

$$\mathbb{E}||Z||_2 \le \sqrt{2\nu(Z)\log(d_1 + d_2)} + \frac{1}{3}L\log(d_1 + d_2).$$

Moreover, for all $t \geq 0$,

$$\mathbb{P}(\|Z\|_2 \ge t) \le (d_1 + d_2) \exp\left(\frac{-t^2/2}{\nu(Z) + Lt/3}\right),$$

for which it is helpful to make a further estimate:

$$\mathbb{P}(\|Z\|_2 \ge t) \le \begin{cases} (d_1 + d_2) \cdot e^{-3t^2/8\nu(Z)}, & t \le \nu(Z)/L \\ (d_1 + d_2) \cdot e^{-3t/8L}, & t \ge \nu(Z)/L. \end{cases}$$

Lemma 22 (Davis-Kahan $\sin(\theta)$ **theorem)** Let the top-k eigenspace of M and \tilde{M} be respectively U_k and \tilde{U}_k (both of which are orthonormal). The k-largest eigenvalue of M is denoted by $\sigma_k(M)$ and similarly for $\sigma_k(\tilde{M})$. Define $\delta_k = \min\{|\sigma_k(M) - \sigma_j(\tilde{M})| : j \ge k+1\}$, then

$$\operatorname{dist}(U_k, \tilde{U}_k) = \sin \theta_k(U_k, \tilde{U}_k) \le \frac{\|M - \tilde{M}\|_2}{\delta_k}.$$

Lemma 23 (Perturbation theorem of projection distance) $Let \operatorname{rank}(X) = \operatorname{rank}(Y),$ then

$$dist(X, Y) \le \min\{\|X^{\dagger}\|_2, \|Y^{\dagger}\|_2\} \|X - Y\|_2.$$

Proof See Theorem 2.3 of Ji-Guang (1987).

Lemma 24 (Uniform sampling) Let $\eta, \zeta \in (0,1)$. Assume the rows of A_i are sampled from the rows of A uniformly at random. Assume each node has sufficiently many samples, that is, for all $i \in m$,

$$s_i \geq \frac{3\mu\rho}{\eta^2}\log\left(\frac{\rho m}{\zeta}\right),$$

where ρ is the rank of A and μ is the row coherence of A. Then with probability greater than $1-\zeta$, Assumption 1 holds.

E. Definitions on subspace distance

In this subsection, we introduce additional definitions and lemmas on metrics between two subspaces. Let $\mathcal{O}_{d\times k}$ be the set of all $d\times k$ orthonormal matrices and \mathcal{O}_k short for $\mathcal{O}_{k\times k}$ denote the set of $k\times k$ orthogonal matrices.

Principle Angles. Given two matrices $U, \tilde{U} \in \mathcal{O}_{d \times k}$ which are both full rank with $1 \le k \le d$, we define the *i*-th $(1 \le i \le k)$ between U and \tilde{U} in a recursive manner:

$$\theta_i(U, \tilde{U}) = \min \left\{ \arccos \left(\frac{x^{\mathsf{T}} y}{\|x\|_2 \|y\|_2} \right) : x \in \mathcal{R}(U), y \in \mathcal{R}(\tilde{U}), x \perp x_j, y \perp y_j, \forall j < i \right\},$$

where $\mathcal{R}(U)$ denotes by the space spanned by all columns of U. In this definition, we require that $0 \leq \theta_1 \leq \cdots \leq \theta_k \leq \frac{\pi}{2}$ and that $\{x_1, \cdots, x_k\}$ and $\{y_1, \cdots, y_k\}$ are the associated principal vectors. Principle angles can be used to quantify the similarity between two given subspaces.

We have following facts about the k-th principle angle between U and \tilde{U} :

Fact 1 Let U^{\perp} denote by the complement subspace of U (so that $[U, U^{\perp}] \in \mathbb{R}^{d \times d}$ forms an orthonormal basis of \mathbb{R}^d) and so does \tilde{U}^{\perp} ,

- 1. $\sin \theta_k(U, \tilde{U}) = ||U^{\mathsf{T}} \tilde{U}^{\perp}||_2 = ||\tilde{U}^{\mathsf{T}} U^{\perp}||_2;$
- 2. $\tan \theta_k(U, \tilde{U}) = \| \left[(U^{\perp})^{\dagger} \tilde{U} \right] (U^{\dagger} \tilde{U})^{\dagger} \|_2$ where \dagger denotes by the Moore-Penrose inverse.
- 3. For any reversible matrix $R \in \mathbb{R}^{k \times k}$, $\tan \theta_k(U, \tilde{U}) = \tan \theta_k(U, \tilde{U}R)$.

Fact 2 Let U^{\perp} denote by the complement subspace of U (so that $[U, U^{\perp}] \in \mathbb{R}^{d \times d}$ forms an orthonormal basis of \mathbb{R}^d) and so does \tilde{U}^{\perp} ,

- 1. $\sin \theta_k(U, \tilde{U}) = ||U^{\mathsf{T}} \tilde{U}^{\perp}||_2 = ||\tilde{U}^{\mathsf{T}} U^{\perp}||_2;$
- 2. $\tan \theta_k(U, \tilde{U}) = \| \left[(U^{\perp})^{\dagger} \tilde{U} \right] (U^{\dagger} \tilde{U})^{\dagger} \|_2$ where \dagger denotes by the Moore-Penrose inverse.
- 3. For any reversible matrix $R \in \mathbb{R}^{k \times k}$, $\tan \theta_k(U, \tilde{U}) = \tan \theta_k(U, \tilde{U}R)$.

Projection Distance. Define the projection distance³ between two subspaces by

$$\operatorname{dist}(U, \tilde{U}) = \|UU^{\mathsf{T}} - \tilde{U}\tilde{U}^{\mathsf{T}}\|_{2}.$$

This metric has several equivalent expressions:

$$\operatorname{dist}(U, \tilde{U}) = \|U^{\mathsf{T}} \tilde{U}^{\perp}\|_{2} = \|\tilde{U}^{\mathsf{T}} U^{\perp}\|_{2} = \sin \theta_{k}(U, \tilde{U}).$$

More generally, for any two matrix $A, B \in \mathbb{R}^{d \times k}$, we define the projection distance between them as

$$\operatorname{dist}(A,B) = \|U_A U_A^{\mathsf{T}} - U_B U_B^{\mathsf{T}}\|_2,$$

where U_A, U_B are the orthogonal basis of $\mathcal{R}(A)$ and $\mathcal{R}(B)$ respectively.

^{3.} Unlike the spectral norm or the Frobenius norm, the projection norm will not fall short of accounting for global orthonormal transformation. Check Ye and Lim (2016) to find more information about distance between two spaces.

Orthogonal Procrustes. Let $U, \tilde{U} \in \mathbb{R}^{d \times k}$ be two orthonormal matrices. $\mathbb{R}(U)$ is close to $\mathbb{R}(\tilde{U})$ does not necessarily imply U is close to \tilde{U} , since any orthonormal invariant of U forms a base of $\mathcal{R}(U)$. However, the converse is true. If we try to map \tilde{U} to U using an orthogonal transformation, we arrive at the following optimization

$$O^* = \operatorname{argmin}_{O \in \mathcal{O}_x} \|U - \tilde{U}O\|_F,$$

where \mathcal{O}_k denotes the set of $k \times k$ orthogonal matrices. The following lemma shows there is an interesting relationship between the subspace distance and their corresponding basis matrices. It implies that as a metric on linear space, $\operatorname{dist}(U,\tilde{U})$ is equivalent to $\|U-\tilde{U}O^*\|_2$ (or $\min_{O\in\mathcal{O}_k}\|U-\tilde{U}O\|_2$) up to some universal constant. The optimization problem involved in is named as the orthogonal procrustes problem and has been well studied (Schönemann, 1966; Cape, 2020).

Lemma 25 Let $U, \tilde{U} \in \mathcal{O}_{d \times k}$ and O^* is the solution of the following optimization,

$$O^* = \operatorname{argmin}_{O \in \mathcal{O}_h} \|U - \tilde{U}O\|_F,$$

Then we have

- 1. O^* has a closed form given by $O^* = W_1W_2$ where $\tilde{U}^{\dagger}U = W_1\Sigma W_2$ is the singular value decomposition of $\tilde{U}^{\dagger}U$.
- 2. Define $d(U, \tilde{U}) := \|U \tilde{U}O^*\|_2$ where $\|\cdot\|_2$ is the spectral norm. Then we have

$$d(U, \tilde{U}) = \sqrt{2 - 2\sqrt{1 - \operatorname{dist}(U, \tilde{U})^2}} = 2\sin\frac{\theta_k(U, \tilde{U})}{2}.$$

- 3. $d(U_1, U_2) = d(U_2, U_1)$ for any $U_1, U_2 \in \mathcal{O}_{d \times k}$.
- 4. $\operatorname{dist}(U, \tilde{U}) \le d(U, \tilde{U}) \le \sqrt{2} \operatorname{dist}(U, \tilde{U})$.
- 5. Define

$$\ell(U, \tilde{U}) := \min_{O \in \mathcal{O}_k} ||U - \tilde{U}O||_2.$$

Then $\ell(U, \tilde{U})$ is a metric satisfying

- $\ell(U, \tilde{U}) \geq 0$ for all $U, \tilde{U} \in \mathcal{O}_{d \times k}$. $\ell(U, \tilde{U}) = 0$ if and only if $\mathcal{R}(U) = \mathcal{R}(\tilde{U})$.
- $\ell(U, \tilde{U}) = \ell(\tilde{U}, U)$ for all $U, \tilde{U} \in \mathcal{O}_{d \times k}$.
- $\ell(U_1, U_2) \le \ell(U_1, U_3) + \ell(U_3, U_2)$ for any U_1, U_2 and $U_3 \in \mathcal{O}_{d \times k}$.
- 6. $\frac{1}{\sqrt{r}} \operatorname{dist}(U, \tilde{U}) \le \ell(U, \tilde{U}) \le d(U, \tilde{U}) \le \sqrt{2} \operatorname{dist}(U, \tilde{U}).$

Proof The first item comes from Schönemann (1966). The second item comes from Cape (2020). The third and forth items follow from the second one. The fifth item follows directly from definition. For the rightest two \leq of the last item, we use $\ell(U, \tilde{U}) \leq d(U, \tilde{U})$ and the forth item. For the leftest \leq , we use $\min_{O \in \mathcal{O}_k} ||U - \tilde{U}O||_2 \geq \frac{1}{\sqrt{k}} \min_{O \in \mathcal{O}_k} ||U - \tilde{U}O||_F$ and $\min_{O \in \mathcal{O}_k} ||U - \tilde{U}O||_F \geq \operatorname{dist}(U, \tilde{U})$ (which is referred from Proposition 2.2 of Vu et al. (2013)).

F. One-shot baseline algorithms

In this subsection, we provide the three algorithms that we compared in the experiments.

Algorithm 4 Unweighted Distributed Averaging (UDA) (Fan et al., 2019b)

- 1: **Input:** distributed dataset $\overline{\{A_i\}_{i=1}^m}$ with $A_i \in \mathbb{R}^{s_i \times d}$, target rank k.
- 2: **Local:** Each device computes the rank-k SVD of $M_i = \frac{1}{s_i} A_i^{\mathsf{T}} A_i$ as $\hat{V}_i \Sigma_i \hat{V}_i^{\mathsf{T}}$ with $\Sigma_i \in \mathbb{R}^{k \times k}$ and $\hat{V}_i \in \mathbb{R}^{d \times k}$.
- 3: **Server:** The central server computes $\tilde{M} = \frac{1}{m} \sum_{i=1}^{n} \hat{V}_{i} \hat{V}_{i}^{\top}$, then output the top k eigenvalues and the corresponding eigenvectors of \tilde{M} .

Algorithm 5 Weighted Distributed Averaging (WDA) (Bhaskara and Wijewardena, 2019)

- 1: **Input:** distributed dataset $\{A_i\}_{i=1}^m$ with $A_i \in \mathbb{R}^{s_i \times d}$, target rank k.
- 2: **Local:** Each device computes the rank-k SVD of $M_i = \frac{1}{s_i} A_i^{\mathsf{T}} A_i$ as $\hat{V}_i \Sigma_i \hat{V}_i^{\mathsf{T}}$ with $\Sigma_i \in \mathbb{R}^{k \times k}$ and $\hat{V}_i \in \mathbb{R}^{d \times k}$.
- 3: **Server:** The central server computes $\tilde{M} = \frac{1}{m} \sum_{i=1}^{n} \hat{V}_{i} \Sigma_{i} \hat{V}_{i}^{\top}$, then output the top k eigenvalues and the corresponding eigenvectors of \tilde{M} .

Algorithm 6 Distributed Randomized SVD (DR-SVD)

- 1: **Input:** distributed dataset $\{A_i\}_{i=1}^m$, $A = [A_1^{\mathsf{T}}, \cdots, A_m^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{R}^{n \times d}$ with target rank k, $A_i \in \mathbb{R}^{s_i \times d}$ and $r = k + \lfloor \frac{d-k}{4} \rfloor$.
- 2: The server generates a $d \times r$ random Gaussian matrix Ω ;
- 3: The server learns $Y = AA^{\dagger}A\Omega$ and obtains an orthonormal $Q \in \mathbb{R}^{n \times r}$ by QR decomposition on Y;
- 4: Let $Q = [Q_1^\intercal, \cdots, Q_m^\intercal]^\intercal$ with $Q_i \in \mathbb{R}^{s_i \times r}$ and each worker receives Q_i ;
- 5: The *i*-th worker computes $B_i = Q_i^\intercal A_i \in \mathbb{R}^{r \times d}$ for all $i \in [m]$;
- 6: The server aggregate $B = \sum_{i=1}^{m} B_i = Q^{\mathsf{T}} A$ and perform SVD: $B = \tilde{U} \hat{\Sigma} \hat{V}^{\mathsf{T}}$;
- 7: Set $\hat{U} = Q\hat{U}$;
- 8: **Output:** the first k columns of $(\hat{U}, \hat{\Sigma}, \hat{V})$.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Lazysvd: even faster svd decomposition yet without agonizing pain. In *Advances in Neural Information Processing Systems*, pages 974–982, 2016.

Peter Arbenz. Lecture notes on solving large scale eigenvalue problems. 2012.

Raman Arora, Andy Cotter, and Nati Srebro. Stochastic optimization of pca with capped msg. In *Advances in Neural Information Processing Systems*, pages 1815–1823, 2013.

Jushan Bai and Serena Ng. Principal components estimation and identification of static factors. *Journal of Econometrics*, 176(1):18–29, 2013.

- Maria-Florina Balcan, Simon Shaolei Du, Yining Wang, and Adams Wei Yu. An improved gap-dependency analysis of the noisy power method. In *Conference on Learning Theory*, pages 284–309, 2016a.
- Maria Florina Balcan, Yingyu Liang, Le Song, David Woodruff, and Bo Xie. Communication efficient distributed kernel principal component analysis. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 725–734, 2016b.
- Aditya Bhaskara and Pruthuvi Maheshakya Wijewardena. On distributed averaging for stochastic k-PCA. In *Advances in Neural Information Processing Systems*, pages 11024–11033, 2019.
- Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. Protection against reconstruction and its applications in private federated learning. arXiv preprint arXiv:1812.00984, 2018.
- Christos Boutsidis, David P Woodruff, and Peilin Zhong. Optimal principal component analysis in distributed and streaming models. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 236–249. ACM, 2016.
- Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. Foundations of Computational mathematics, 9(6):717, 2009.
- Joshua Cape. Orthogonal procrustes and norm-dependent optimality. The Electronic Journal of Linear Algebra, 36(36):158–168, 2020.
- Kamalika Chaudhuri, Anand Sarwate, and Kaushik Sinha. Near-optimal differentially private principal components. In *Advances in Neural Information Processing Systems*, pages 989–997, 2012.
- Xi Chen, Jason D Lee, He Li, and Yun Yang. Distributed estimation for principal component analysis: a gap-free approach. arXiv preprint arXiv:2004.02336, 2020.
- Christopher De Sa, Bryan He, Ioannis Mitliagkas, Christopher Ré, and Peng Xu. Accelerated stochastic power iteration. *Proceedings of machine learning research*, 84:58, 2018.
- Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.
- Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. arXiv preprint arXiv:1905.02383, 2019.
- John C Duchi, Michael I Jordan, and Martin J Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521): 182–201, 2018.
- Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

- Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In 2010 IEEE 51st Annual Symposium on Foundations of Computer Science, pages 51–60. IEEE, 2010.
- Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3-4):211–407, 2014a.
- Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 11–20, 2014b.
- Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed! a survey of attacks on private data. *Annal Review of Statistics and its Applications*, 2017.
- Jianqing Fan, Yuan Ke, Qiang Sun, and Wen-Xin Zhou. Farmtest: Factor-adjusted robust multiple testing with approximate false discovery control. *Journal of the American Statistical Association*, 114(528):1880–1893, 2019a.
- Jianqing Fan, Dong Wang, Kaizheng Wang, Ziwei Zhu, et al. Distributed estimation of principal eigenspaces. *The Annals of Statistics*, 47(6):3009–3031, 2019b.
- Arpita Gang, Haroon Raja, and Waheed U Bajwa. Fast and communication-efficient distributed PCA. In *ICASSP 2019-2019 IEEE International Conference on Acoustics*, Speech and Signal Processing (ICASSP), pages 7450–7454. IEEE, 2019.
- Dan Garber and Elad Hazan. Fast and simple PCA via convex optimization. arXiv preprint arXiv:1509.05647, 2015.
- Dan Garber, Elad Hazan, Chi Jin, Sham M Kakade, Cameron Musco, Praneeth Netrapalli, and Aaron Sidford. Faster eigenvector computation via shift-and-invert preconditioning. In ICML, pages 2626–2634, 2016.
- Dan Garber, Ohad Shamir, and Nathan Srebro. Communication-efficient algorithms for distributed stochastic principal component analysis. arXiv preprint arXiv:1702.08169, 2017.
- Jason Ge, Zhaoran Wang, Mengdi Wang, and Han Liu. Minimax-optimal privacy-preserving sparse PCA in distributed systems. In *International Conference on Artificial Intelligence and Statistics*, AISTATS 2018, 2018.
- Alex Gittens and Michael W Mahoney. Revisiting the nyström method for improved large-scale machine learning. *The Journal of Machine Learning Research*, 17(1):3977–4041, 2016.
- Oded Goldreich. The Foundations of Cryptography, volume 2. Cambridge University Press, 2009.
- Gene H Golub and William Kahan. Calculating the singular values and pseudo-inverse of a matrix. Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis, 2(2):205–224, 1965.

- Gene H Golub and C Reinsch. Singular value decomposition and least squares solutions. Numerische Mathematik, 14:403–420, 1970.
- Gene H Golub and Charles Francis. Van Loan. *Matrix computations, volume 3.* JHU Press, 2012.
- Andreas Grammenos, Rodrigo Mendoza-Smith, Jon Crowcroft, and Cecilia Mascolo. Federated principal component analysis. In 34th Conference on Neural Information Processing Systems (NeurIPS 2020), 2020.
- Xiao Guo, Yixuan Qiu, Hai Zhang, and Xiangyu Chang. Randomized spectral co-clustering for large-scale directed networks. arXiv preprint arXiv:2004.12164, 2020.
- Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.
- Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. In *Advances in Neural Information Processing Systems*, pages 2861–2869, 2014.
- Moritz Hardt and Aaron Roth. Beyond worst-case analysis in private singular vector computation. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 331–340, 2013.
- Alan Julian Izenman. Modern multivariate statistical techniques. Regression, classification and manifold learning, 10:978–0, 2008.
- Sun Ji-Guang. Perturbation of angles between linear subspaces. *Journal of Computational Mathematics*, pages 58–61, 1987.
- 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.
- Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. First analysis of local GD on heterogeneous data. arXiv preprint arXiv:1909.04715, 2019.
- Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3): 50–60, 2020a.
- Xiang Li and Zhihua Zhang. Delayed projection techniques for linearly constrained problems: Convergence rates, acceleration, and applications. arXiv preprint arXiv:2101.01505, 2021.
- Xiang Li, Wenhao Yang, Shusen Wang, and Zhihua Zhang. Communication efficient decentralized training with multiple local updates. arXiv preprint arXiv:1910.09126, 2019.

- Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of FedAvg on non-iid data. In *International Conference on Learning Representations*, 2020b.
- Xiang Li, Shusen Wang, Kun Chen, and Zhihua Zhang. Communication-efficient distributed SVD via local power iterations. arXiv preprint arXiv:2002.08014, 2020c.
- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.
- Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Inference attacks against collaborative learning. arXiv preprint arXiv:1805.04049, 13, 2018.
- Ilya Mironov. Rényi differential privacy. In 2017 IEEE 30th Computer Security Foundations Symposium (CSF), pages 263–275. IEEE, 2017.
- Cameron Musco and Christopher Musco. Randomized block Krylov methods for stronger and faster approximate singular value decomposition. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.
- Sahand Negahban, Sewoong Oh, and Devavrat Shah. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287, 2017.
- Erkki Oja and Juha Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of mathematical analysis and applications*, 106(1):69–84, 1985.
- Haroon Raja and Waheed U Bajwa. Distributed stochastic algorithms for high-rate streaming principal component analysis. arXiv preprint arXiv:2001.01017, 2020.
- Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pages 1576–1602. World Scientific, 2010.
- Yousef Saad. Numerical methods for large eigenvalue problems. preparation. Available from: http://www-users. cs. umn. edu/saad/books. html, 2011.
- Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems*, 31(9):3400–3413, 2019.
- Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.
- Ohad Shamir. A stochastic pca and svd algorithm with an exponential convergence rate. In *International Conference on Machine Learning*, pages 144–152, 2015.
- Ohad Shamir. Convergence of stochastic gradient descent for pca. In *International Conference on Machine Learning*, pages 257–265, 2016.

- Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, pages 4424–4434, 2017.
- Sebastian U Stich. Local SGD converges fast and communicates little. arXiv preprint arXiv:1805.09767, 2018.
- Ji-Guang Sun. On perturbation bounds for the QR factorization. *Linear algebra and its* applications, 215:95–111, 1995.
- Joel A Tropp. An introduction to matrix concentration inequalities. arXiv preprint arXiv:1501.01571, 2015.
- Ulrike Von Luxburg. A tutorial on spectral clustering. Statistics and computing, 17(4): 395–416, 2007.
- Vincent Q Vu, Jing Lei, et al. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, 2013.
- Jianyu Wang and Gauri Joshi. Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms. arXiv preprint arXiv:1808.07576, 2018.
- Jianyu Wang and Gauri Joshi. Adaptive communication strategies to achieve the best errorruntime trade-off in local-update sgd. *Proceedings of Machine Learning and Systems*, 1: 212–229, 2019.
- Shusen Wang, Luo Luo, and Zhihua Zhang. SPSD matrix approximation vis column selection: Theories, algorithms, and extensions. *The Journal of Machine Learning Research*, 17(1):1697–1745, 2016.
- Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.
- Rafi Witten and Emmanuel Candès. Randomized algorithms for low-rank matrix factorizations: sharp performance bounds. *Algorithmica*, 72(1):264–281, 2015.
- Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics* and intelligent laboratory systems, 2(1-3):37–52, 1987.
- David P Woodruff. Sketching as a tool for numerical linear algebra. arXiv preprint arXiv:1411.4357, 2014.
- 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.
- Ke Ye and Lek-Heng Lim. Schubert varieties and distances between subspaces of different dimensions. SIAM Journal on Matrix Analysis and Applications, 37(3):1176–1197, 2016.

- Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In AAAI Conference on Artificial Intelligence, 2019.
- Hai Zhang, Xiao Guo, and Xiangyu Chang. Randomized spectral clustering in large-scale stochastic block models. arXiv preprint arXiv:2002.00839, 2020.
- Fan Zhou and Guojing Cong. On the convergence properties of a k-step averaging stochastic gradient descent algorithm for nonconvex optimization. arXiv preprint arXiv:1708.01012, 2017.
- Yaqin Zhou and Shaojie Tang. Differentially private distributed learning. *INFORMS Journal on Computing*, 32(3):779–789, 2020.