Federated Gradient Matching Pursuit

Halyun Jeong, Deanna Needell, *Member, IEEE*, and Jing Qin, *Member, IEEE*

***Abstract*— Traditional machine learning techniques require centralizing all training data on one server or data hub. However, with the development of communication technologies and a huge amount of decentralized data on many clients, collaborative machine learning has become the main interest while providing privacy-preserving frameworks. Federated learning (FL) provides such a solution to learn a shared model while keeping training data at local clients. On the other hand, in a wide range of machine learning and signal processing applications, the desired solution naturally has a certain structure that can be framed as sparsity with respect to a certain dictionary. This problem can be formulated as an optimization problem with sparsity constraints and solving it efficiently has been one of the primary research topics in the traditional centralized setting. In this paper, we propose a novel algorithmic framework, federated gradient matching pursuit (FedGradMP), to solve the sparsity constrained minimization problem in the FL setting. We also generalize our algorithms to accommodate various practical FL scenarios when only a subset of clients participate per round, when the local model estimation at clients could be inexact, or when the model parameters are sparse with respect to general dictionaries. Our theoretical analysis shows the linear convergence of the proposed algorithms. A variety of numerical experiments are conducted to demonstrate the great potential of the proposed framework – fast convergence both in communication rounds and computation time for many important scenarios without intricate parameter tuning.**

***Index Terms*— Federated learning, sparse recovery, gradient matching pursuit, random algorithm.**

I. Introduction

As technology and science have advanced, machine learning for big data processing has become an emerging field with a wide variety of applications. In general, there are several major considerations in dealing with a large amount of data - data storage and privacy, computation, and communication [1]. To address the limitation of efficiency and scalability of traditional machine learning algorithms for large-scale data, distributed centralized learning allows data and/or model parallelism, where all local data are typically uploaded to a central server but model training is distributed to various clients [2]. Different from the traditional centralized learning, federated learning (FL) [3] is a collaborative learning framework in which many clients work together to solve an optimization problem without sharing local data. In order to preserve privacy, and reduce the communication cost between clients and the server, datasets are only stored at the clients locally and can not be transferred to other clients or the server in FL. In other words, it aims to learn a central model using decentralized datasets. The heterogeneous data distributions among the clients pose an additional challenge in federated learning.

As one of the most popular FL algorithms, Federated Averaging (FedAvg) [4] considers an unconstrained optimization problem where the desired solution has no additional characteristics. FedAvg alternates gradient descent and averaging of distributed solutions from local clients in an iterative way. However, in a lot of applications, the solution of interest has some special structures, such as sparsity and low-rankness by itself or in some transformed domain. This structure, serving as prior information, can be utilized to address the ill-posedness of the problem and improve performance. Thus, recent interest in FL optimization with additional solution structures has grown, which has been shown to be especially effective when only a few data samples are available at each client but the underlying signal dimension is relatively large [5], [6].

In such cases when the solution to an optimization problem from FL applications possesses a certain structure, for example, sparsity and low-rankness, one can use a regularizer to enforce the desired structure [5]. Federated Dual Averaging (FedDualAvg) by Yuan et al. [5], different from FedAvg, uses potentially nonsmooth and convex regularizers to promote the structure of the solution.

When we have more prior information about the solution structure, e.g., the sparsity level, then hard-thresholding based approaches could be often more efficient than the regularization based methods [7], [8]. The hard-thresholding based methods aim to solve nonconvex formulations of the problem, which have been successfully applied to many data processing problems lately, offering enhanced performance [9], [10], [11], [12], [13].

Following this line of research, Tong et al. proposed Federated Hard Thresholding (FedHT) and Federated Iterative Hard Thresholding (FedIterHT) [6], employing hard-thresholding at the aggregation step at the server with potentially additional hard-thresholding after each stochastic gradient step at clients. With a proper choice of step sizes for the stochastic gradients at the clients, these methods guarantee linear convergence up to a neighborhood of the solution to the problem. Despite the

Manuscript received 19 February 2023; revised 19 December 2023; accepted 25 March 2024. Date of publication 9 April 2024; date of current version 21 May 2024. The work of Deanna Needell was supported by NSF under Grant DMS-2011140 and Grant DMS-2108479. The work of Jing Qin was supported by NSF under Grant DMS-1941197. *(Corresponding author: Halyun Jeong.)*

Halyun Jeong and Deanna Needell are with the Department of Mathematics, University of California at Los Angeles, Los Angeles, CA 90095 USA (e-mail: hajeong@math.ucla.edu; deanna@math.ucla.edu).

Jing Qin is with the Department of Mathematics, University of Kentucky, Lexington, KY 40506 USA (e-mail: jing.qin@uky.edu).

Communicated by M. Soltanolkotabi, Associate Editor for Signal Processing and Source Coding.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2024.3386705.

Digital Object Identifier 10.1109/TIT.2024.3386705

fact that these approaches partially inherit the advantages of thresholding based methods over those based on regularization, they necessitate fine-tuning of learning rates at clients, which often have practical limitations and they are not applicable for sparse signals with respect to general dictionaries. Their convergence analysis also requires the mini-batch sizes at the clients to grow exponentially in the number of communication rounds, which further limits the usage in most applications.

Another popular thresholding based method is the gradient matching pursuit (GradMP) [14], an extension of the compressive sampling matching pursuit (CoSaMP) [9]. These methods are known to be more efficient than the others such as regularizer based methods, particularly when the sparsity level of a signal is much smaller than its dimension [13], [15].

A. Contributions

We summarize our contributions below.

- We propose the Federated Gradient Matching Pursuit algorithm, abbreviated as FedGradMP, to overcome the aforementioned drawbacks. More precisely, we show that the proposed FedGradMP enjoys the linear convergence up to a small neighborhood of the solution, without the restrictions for FedHT/FedIterHT to work. Furthermore, our analysis indicates that FedGradMP converges linearly up to a statistical bias term for the recovery of sparse signals under mild conditions.
- The majority of FL algorithm analyses have been carried out either under uniformly bounded gradient or bounded dissimilarity assumptions, which could be problematic in certain scenarios [16]. Only a few recent works for FL in unconstrained setup provide theoretical guarantees without this type of assumptions but instead under the bounded dissimilarity only at optima – a condition that is considered to be the most general type of heterogeneity assumption [17], [18]. Our convergence analysis of FedGradMP has been carried out under this general dissimilarity condition.
- Thanks to the mechanism of GradMP, FedGradMP does not require intensive tuning of learning rates at the clients for the sparse linear regression problem, which could be often still challenging because of data heterogeneity in the FL setting. Approaches based on the local stochastic gradient at clients including FedHT/FedIterHT, as described in the literature [6], [16], [19] and demonstrated in our numerical studies, need tweaking the learning rates (step sizes); otherwise, they diverge or converge slowly, especially when the data at distinct clients are more heterogeneous. In contrast, FedGradMP is based on solving low-dimensional sub-optimization problems at clients that can be often solved efficiently without the need for fine tuning of learning rates.
- Many signals of practical interest are not sparse in the standard basis but rather in a certain dictionary. This observation has led to the development of several sparse recovery methods with general dictionaries in the centralized setting [14], [20], [21]. FedGradMP is a versatile method under a general dictionary framework. One potential problem with using dictionaries in FL methods is the privacy concern if they are correlated with the client datasets. By utilizing dictionaries that are statistically independent with client datasets, such as the random Gaussian dictionary, we demonstrate the effectiveness of FedGradMP as an FL method without such concerns.

B. Further Related Works

There have been numerous extensions and analyses of FedAvg [4], [16], [17], [18], [22], a standard algorithm to train a machine learning model in FL. FedAvg can be considered as a variant of Local SGD, which essentially runs stochastic gradient iterations at each client, and their locally computed model parameters are averaged at a server.

In addition to the considerations of Local SGD [4], [23] for efficient communication in distributed learning, FedAvg aims to handle challenges in the FL settings such as heterogeneous client datasets and partial client participation [16], [17]. Thanks to the recent endeavors of researchers [17], [18], [24], we now have a better understanding of the convergence behavior of FedAvg, especially when the objective function is (strongly) convex. As for the nonconvex case, several works provide the convergence of FedAvg to the stationary points and global convergence under extra assumptions such as Polyak-Lojasiewicz (PL) condition [19], which is a generalization of the strong convexity condition. However, it is worth noting that these assumptions do not imply our main assumptions, the restricted strong convexity/smoothness.

An important research direction in FL algorithm analysis is characterizing the trade-off between convergence speed and accuracy that stems from client data heterogeneity. As the clients run more local iterations, the estimates of the local solution become more accurate at each client (improving the convergence rate) while they tend to drift away from the global solution (making the actual residual error larger), especially in a highly heterogeneous environment [17], [25], [26], [27], [28]. Our analysis and numerical experiments on the convergence behavior of FedGradMP also reflect this phenomenon, which becomes more noticeable when the client datasets are highly heterogeneous.

To further reduce the communication cost between the server and clients, techniques such as sparsifying and reducing the dimensionality of the gradients have been proposed in [29], [30], [31], [32], and [33]. In FedGradMP, the hard-thresholding operation is applied whenever the computed models are sent from a server or clients, so the models are already sparsified with the effective dimension same as the desired sparsity level. This makes FedGradMP more attractive in terms of saving communication resources.

Another active area in FL research is client sampling or partial participation. Because of the limited connection bandwidth or a large population of clients, it is often not possible for every client to participate at each round in FL. Many methods incorporate this by modeling each client to participate randomly per round according to some distribution [16], [34], [35]. There have been recent attempts to employ more elaborate sampling strategies such as importance sampling [36], but this requires

TABLE I
COMPARISON OF OUR WORK WITH RELATED REFERENCES

	Dictionary sparsity and convergence speed-up	Linear convergence to the solution up to optimal statistical bias	No client LR tuning	Bounded heterogeneity only at optima
FedAvg [4]	✗	✗	✗	✓
FedDualAvg [5]	✗	✗	✗	✗
FedHT/FedIterHT [6]	✗	✓*	✗	✓
FedGradMP	✓	✓	✧**	✓

* The convergence analysis of FedHT/FedIterHT, however, requires that the mini-batch sizes increase exponentially in the number of communication rounds, which is generally not practical in many applications.
** FedGradMP does not require learning rate tuning for the sparse linear regression problem that could be still challenging for baseline algorithms based on the stochastic gradient descent, essentially due to heterogeneity in the FL environment as we illustrate Section V-C.

extra care since it could leak private information of client data. We analyze FedGradMP under the more common assumption, i.e., the random client participation model, and show the more client participate at each round, the faster the convergence rate is. This observation is consistent with recent findings [34], [37] on the FL algorithms for the unconstrained problem.

C. Organization

The rest of the paper is structured as follows. In Section II, we introduce the sparse federated learning problem and make important assumptions that will be used for convergence analysis. In Section III, we propose the federated gradient matching pursuit algorithm and discuss the convergence guarantees in detail. Section IV generalizes FedGradMP and its convergence analysis to several practical scenarios such as the partial client participation environment and inexact estimation at the client side. In addition, we provide theoretical justifications of using a shared random Gaussian dictionary at clients to improve the performance of FedGradMP. Section V provides a variety of numerical experiments for sparse signal recovery which demonstrate the effectiveness of the proposed approach. We draw conclusions in Section VI.

D. Notation

We say that a vector is s-sparse if it has at most s nonzero entries. We write $\|\cdot\|_2$ to denote the ℓ_2 norm for a vector. We use $\|\cdot\|_F$ and $\|\cdot\|$ to denote the Frobenius norm and operator norm of a matrix, respectively. For a given positive integer m, $[m]$ denotes the set of integers $\{1,2,\ldots,m\}$. The transpose of matrix A is denoted by $A^\top$. For positive semidefinite matrices A and B, $A \preceq B$ means that $B-A$ is positive semidefinite. For a finite set S, $|S|$ denotes its cardinality.

II. SPARSE FEDERATED LEARNING

Federated learning is a framework for collaboratively solving machine learning problems across multiple clients, potentially coordinated by a server.

While this provides enhanced privacy, it also poses interesting challenges since clients still require to exchange local parameters to other clients or a server in a communication-efficient way. Moreover, in many heterogeneous learning environments, the local data of each client can be non-identically distributed and/or statistically dependent.

To formally describe FL, we begin with introducing the setup. Assume that the number of the clients is N, and the local objective function at the i-th client is denoted by $f_i(x) = \mathbb{E}_{z\sim D_i}[\ell_i(x;z)]$ where D_i is the dataset at the i-th client and $\ell_i(x;z)$ is the loss function about x that depends on the data z.

The optimization problem of interest in FL typically takes the form

$$\min_{x\in\mathbb{R}^n} f(x) := \sum_{i=1}^{N} p_i f_i(x)$$

where $x \in \mathbb{R}^n$ and $p_i \in [0,1]$ is the weight for the i-th client satisfying $\sum_{i=1}^N p_i = 1$. This formulation is sufficiently general to cover the most machine learning settings including, the empirical risk minimization (ERM) by taking the expectation uniformly over the dataset D_i and $p_i = |D_i|/\sum_{i=1}^N |D_i|$ [16], [38].

On the other hand, due to communication efficiency or the nature of many applications, it is natural to assume that the solution we are looking for is sparse with respect to a certain dictionary or an atom set. In order to discuss this general notion of sparsity, we consider a finite set of atoms $\mathscr{A} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_d\}$ where $\mathbf{a}_i \in \mathbb{R}^n$, as defined in [14] and [39]. For example, we recover the standard basis when $\mathscr{A} = \{e_1, \cdots, e_n\}$, where e_is are the standard basis vectors for $\mathbb{R}^n$. We say a vector x is **τ-sparse with respect to $\mathscr{A}$** if x can be represented as

$$x = \sum_{i=1}^{d} \alpha_i \mathbf{a}_i$$

where at most τ number of the coefficients α_i's are nonzero. Then, the support of x with respect to $\mathscr{A}$ is defined in a natural way, $\mathrm{supp}_{\mathscr{A}}(x) = \{i \in [d] : \alpha_i \neq 0\}$.

We define the ℓ_0-norm of x with respect to $\mathscr{A}$ as

$$\|x\|_{0,\mathscr{A}} = \min_{\alpha} \left\{ |\Omega| : x = \sum_{i\in\Omega} \alpha_i \mathbf{a}_i, \Omega \subseteq [d] \right\},$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)^\top$.

With a sparsity constraint, sparse FL aims to solve the followingconstrained problem

$$\min_{x\in\mathbb{R}^n} f(x) = \sum_{i=1}^{N} p_i f_i(x) \quad \text{subject to} \quad \|x\|_{0,\mathscr{A}} \leq \tau, \tag{1}$$

where τ is a preassigned sparsity level. We denote the optimal solution to this problem by x^*. We also assume further that

each local objective function f_i can be expressed by the average of the functions $g_{i,j}:\mathbb{R}^n\to\mathbb{R}$, i.e.,

$$f_i(x)=\frac{1}{M}\sum_{j=1}^{M} g_{i,j}(x), \tag{2}$$

for some integer M. One can interpret $g_{i,j}$ as a loss function associated with the i-th client restricted to the j-th mini-batch, where the index set of all possible mini-batches is $[M]$.

The objective function of (1) usually depends on the data distribution at clients. For example, the least squares problem in FL typically sets

$$f_i(x)=\frac{1}{2|D_i|}\|A_{D_i}x-y_{D_i}\|_2^2,$$

where A_{D_i} is the client i data matrix whose rows consist of the training input data points for client i and y_{D_i} are corresponding observations. Let b be the size of each mini-batch and M be the total number of mini-batches. Then, $M=\binom{|D_i|}{b}$ and $g_{i,j}$ is given by

$$g_{i,j}(x)=\frac{1}{2b}\sum_{(a_k,b_k)\in S_j}(y_k-\langle a_k,x\rangle)^2,$$

where S_j is a subset of D_i associated j-th mini-batch of i-th client. We use $\mathbb{E}_j^{(i)}\varphi_{i,j}(x)$ to denote the expectation of a function $\varphi_{i,j}(x)$ provided that the j-th mini-batch index set is chosen from the set of the all possible mini-batches with size b, uniformly at random. Hence, the function f_i can be expressed as $f_i(x)=\mathbb{E}_j^{(i)}g_{i,j}(x)$.

Since the problem (1) is nonconvex in general, it is difficult to find its solution without additional assumptions on the objective function. In this work, we adopt the following assumptions from [14].

Assumption 1 ($\mathscr{A}$-Restricted Strong Convexity ($\mathscr{A}$-RSC)): The local objective function f_i at the i-th client satisfies the restricted $\rho_\tau^-(i)$-strongly convexity condition: for each $i\in[N]$ and any $x_1,x_2\in\mathbb{R}^n$ with $\|x_1-x_2\|_{0,\mathscr{A}}\le\tau$, we have

$$f_i(x_1)-f_i(x_2)-\langle\nabla f_i(x_2),x_1-x_2\rangle\ge\frac{\rho_\tau^-(i)}{2}\|x_1-x_2\|_2^2. \tag{3}$$

Assumption 2 ($\mathscr{A}$-restricted Strong Smoothness ($\mathscr{A}$-RSS)): The loss function $g_{i,j}$ associated with the j-th mini-batch at the i-th client satisfies the restricted $\rho_\tau^+(i,j)$-strongly smoothness condition: for each i,j and any $x_1,x_2\in\mathbb{R}^n$ with $\|x_1-x_2\|_{0,\mathscr{A}}\le\tau$, we have

$$\|\nabla g_{i,j}(x_1)-\nabla g_{i,j}(x_2)\|_2\le\rho_\tau^+(i,j)\|x_1-x_2\|_2. \tag{4}$$

Remark 1: Assumptions 1 and 2 are widely used in the optimization community for solving the high-dimensional statistical learning or sparse recovery problems. Note that the local function f_i and $g_{i,j}$ may not be convex or smooth in the entire space $\mathbb{R}^n$ since we only need strong convexity and smoothness assumptions for vectors that are sparse with respect to a dictionary. Most convergence analysis for the FL algorithms assumes (strong) convexity of f_i or the Polyak-Lojasiewicz (PL) condition [19]; neither is weaker than Assumption 1.

Several FL algorithms inspired by classical sparse optimization techniques [14], [40], [41] have been proposed, including FedHT and FedIterHT, which are based on IHT [6]. These algorithms use the hard-thresholding operator $\mathscr{H}_\tau(x)$, which keeps the τ largest components of the input vector x in magnitude with respect to the standard basis, whereas our algorithm adopts a more general hard-thresholding operator — the approximate projection operator. To define an approximate projection, we denote by $\mathscr{R}(\mathscr{A}_\Gamma)$ the subspace of $\mathbb{R}^n$ spanned by the atoms in $\mathscr{A}$ whose indices are restricted to $\Gamma\subset[d]$. For $w\in\mathbb{R}^n$, the orthogonal projection of w to $\mathscr{R}(\mathscr{A}_\Gamma)$ is denoted by $P_\Gamma w$. Then, an approximate projection operator with $\eta>0$, denoted by $\text{approx}_\tau(w,\eta)$, constructs an index set Γ of cardinality τ such that

$$\|P_\Gamma w-w\|_2\le\eta\|w-\mathscr{H}_\tau(w)\|_2,$$

where $\mathscr{H}_\tau(w)$ is the best τ-sparse approximation of w with respect to $\mathscr{A}$, i.e.,

$$\mathscr{H}_\tau(w)=\underset{x=\mathscr{A}\alpha,\|\alpha\|_0\le\tau}{\textbf{argmin}}\|w-x\|_2.$$

Here, $\mathscr{A}$ is the matrix whose columns are the atoms by abusing the notation slightly.

The local dissimilarity in FL captures how the data distributions among clients are different, which is typically in the following form, especially in the early works in FL [5], [16], [42]:

$$\mathbb{E}_{i\sim\mathscr{P}}\|\nabla f_i(x)-\nabla f(x)\|_2^2\le\beta^2\|\nabla f(x)\|_2^2+\zeta^2,\quad\forall x\in\mathbb{R}^n. \tag{5}$$

In this work, the assumption of heterogeneity on the client data is much weaker than (5) by assuming heterogeneity only at the solution x^* as follows.

Assumption 3: There is a minimizer for (1), denoted by x^* with a finite ζ_*^2 defined as below:

$$\zeta_*^2=\mathbb{E}_{i\sim\mathscr{P}}\|\nabla f_i(x^*)\|_2^2=\sum_{i=1}^{N}p_i\|\nabla f_i(x^*)\|_2^2.$$

Assumption 3 is the same as the one used for more recent analyses giving sharper convergence guarantees of FL algorithms [17], [18]. This is also a necessary assumption for the FedAvg type of algorithms to converge [18]. But there are a few places where we state the implication of our results under stronger assumptions such as (5), in order to compare the implications of our results to previous works.

Remark 2: When $\beta=0$ in the condition (5), it reduces to the uniform bounded heterogeneity condition that has been used in early convergence analyses of many popular FL algorithms including FedAvg [42] and FedDualAvg [5], [16]. However, it is possible that Assumption 3 (which needs to hold only at the optimal solution x^*) holds but no finite β and ζ exist for (5) as mentioned in [16] and [17] (this is essentially because the bound (5) needs to hold uniformly for all x, for given β and ζ). Below we provide two examples to help understand these relationships further.

- Example 1: Consider the uniformly bounded heterogeneity condition, which is when $\beta=0$ in Eq (5). Now, if in addition $\zeta=0$, this enforces the homogeneous scenario, i.e., all the local objective functions are the same for

all x. In contrast, even when $\zeta^* = 0$ in Assumption 3, this only means that there exists a common minimizer x^* (so, $\nabla f_i(x^*) = 0$ for all i) but the local objective function could still be different, as mentioned in [18].

- Example 2 (Proposition 2.2 in [43]): In general, for the uniformly bounded heterogeneity condition ($\beta = 0$ in (5)), even simple quadratic objective functions do not satisfy the assumption (5) unless all the objective functions have the same Hessian for all x.

Other common assumptions in the analysis of federated learning algorithms are the unbiased and bounded variance conditions of local stochastic gradients.

Assumption 4: The local stochastic gradient $\nabla g_{i,j}$ associated with the randomly selected j-th mini-batch at the i-th client satisfies

$$\mathbb{E}_j^{(i)}[\nabla g_{i,j}(x)] = \nabla f_i(x) \quad \text{for any } \tau\text{-sparse vector } x, \tag{6}$$

and

$$\mathbb{E}_j^{(i)}\|\nabla g_{i,j}(x) - \nabla f_i(x)\|_2^2 \le \sigma_i^2 \quad \text{for any } \tau\text{-sparse vector } x, \tag{7}$$

where $\mathbb{E}_j^{(i)}$ is the expectation taken over the mini-batch index selected from $[M]$ at the client i.

Remark 3: The bounded variance condition of local stochastic gradients associated with mini-batches (7) in Assumption 4 is widely used in FL and stochastic algorithms in general [16], [44], but it may not hold for some settings [17]. This assumption (7) is actually not essential for our main result to hold (See Appendix for the proof of our convergence theorem without this assumption) but we present our work under the assumption for the sake of simplicity.

The following lemma is a well-known consequence of the $\mathscr{A}$-RSS property in Assumption 1.

Lemma 1 (Descent Lemma): Suppose that the function $h(x)$ satisfies the $\mathscr{A}$-RSS property in Assumption 2 with a constant ρ_τ^+. Then for any $x_1, x_2 \in \mathbb{R}^n$ with $\|x_2\|_{0,\mathscr{A}} \le \tau$, it holds

$$\langle \nabla h(x_1), x_2 \rangle \ge h(x_1 + x_2) - h(x_1) - \frac{\rho_\tau^+}{2}\|x_2\|_2^2.$$

III. Federated Gradient Matching Pursuit

In this section, we propose Federated Gradient Matching Pursuit (FedGradMP) in Algorithm 1 and discuss its convergence guarantee.

In the FedGradMP framework, the StoGradMP algorithm [14] is implemented at the client side and the server aggregates the resulting locally computed models projects onto a subspace of dimension at most τ in each round.

We begin with an overview of StoGradMP which FedGradMP is built upon. First, the method computes the stochastic gradient of the objective function we want to minimize. The computed stochastic gradient will be the proxy of the residual between signal and the previous iterate; For example, for the sparse linear regression case, the proxy is of the form of $A^\top(y - Ax_t) = A^\top A(x^* - x_t)$ when there is no noise in the measurement y. Assume further that the data matrix A satisfies $2s$-RIP with respect to the standard basis for simplicity, and the signal and the thresholding level are s-sparse. Then $x^* - x_t$ approximately equal to the top $2s$ subvector of $A^\top(y - Ax_t)$ upto the RIP constant [13]. Hence, the subspace corresponding to index set of the top $2s$ largest components of $A^\top(y - Ax_t)$ would be a good candidate to explore to minimize the residual $x^* - x_t$ further, along with the previous support estimate of x^* by taking the top s component of x_t. Based on this idea, StoGradMP merges these two subspaces associated with the top $2s$ largest components of $A^\top(y - Ax_t)$ and the top s largest components of x_t and solve a subproblem of minimizing the objective function with respect to this $3s$-dimensional subspace. In the final step, StoGradMP identifies the index of the top-τ largest magnitude entries of the solution of the subproblem that will be used in the next iteration as a support estimate.

Based on this idea of StoGradMP, each iteration of FedGradMP at a client consists of the following five steps:

1) Randomly select a mini-batch from the client batch.
2) Compute the stochastic gradient associated with the selected mini-batch.
3) Merge the subspace associated with the previously estimated local model with the closest subspace of dimension at most 2τ to the stochastic gradient from Step 2.
4) Solve the minimization problem for the local objective function at the client over the merged subspace from Step 3.
5) Identify the closest subspace of dimension at most τ to the solution at Step 4.

Note that in Step 4, the clients are not minimizing the local objective function f_i over the sparsity constraint, but over the subspace associated with the estimated sparsity pattern of the solution in Step 3. This subproblem can be often solved efficiently since f_i are strongly convex/smooth with respect to such subspaces by Assumptions 1 and 2, especially when the dimension of the subspace is small or f_i are quadratic [45], [46]. Nevertheless, it could be expensive to solve this subproblem in general, so we discuss how to obtain its approximate solution by computationally cheap methods in the next section.

A. Linear Convergence of FedGradMP

This subsection is devoted to proving the linear convergence of FedGradMP in the number of communication rounds. The first step of the proof for our main theorem is similar to the one for Theorem 3.1 in [6] but also utilizes several lemmas below from [14] and [39] after some modifications to accommodate the FL setting.

Lemma 2 ([14, Lemma 1]): The approximation error between the $(k+1)$-th local iterate $x_{t,k+1}^{(i)}$ and x^* is bounded by

$$\|x_{t,k+1}^{(i)} - x^*\|_2^2 \le (1 + \eta_2)^2 \|b_{t,k}^{(i)} - x^*\|_2^2.$$

Algorithm 1 FedGradMP

Input: The number of rounds T, the number of clients N, the number of local iterations K, weight vector p, the estimated sparsity level τ, η_1, η_2, η_3.
Output: $\hat{x} = x_T$.
Initialize: $x_0 = 0$, $\Lambda = \emptyset$.
for $t = 0, 1, \ldots, T-1$ **do**
  for client $i = 1, 2, \ldots, N$ **do**
    $x_{t,1}^{(i)} = x_t$
    for $k = 1$ **to** K **do**
      Select a mini-batch index set $j_k := i_{t,k}^{(i)}$ uniformly at random from $\{1, 2, \ldots, M\}$
      Calculate the stochastic gradient $r_{t,k}^{(i)} = \nabla g_{i,j_k}\left(x_{t,k}^{(i)}\right)$
      $\Gamma = \text{approx}_{2\tau}(r_{t,k}^{(i)}, \eta_1)$
      $\widehat{\Gamma} = \Gamma \cup \Lambda$
      $b_{t,k}^{(i)} = \underset{x}{\textbf{argmin}} f_i(x), \quad x \in \mathscr{R}(\mathscr{A}_{\widehat{\Gamma}})$
      $\Lambda = \text{approx}_{\tau}(b_{t,k}^{(i)}, \eta_2)$
      $x_{t,k+1}^{(i)} = P_{\Lambda}(b_{t,k}^{(i)})$
    end for
  end for
  $\Lambda_s = \text{approx}_{\tau}\left(\sum_{i=1}^{N} p_i x_{t,K+1}^{(i)}, \eta_3\right)$
  $x_{t+1} = P_{\Lambda_s}\left(\sum_{i=1}^{N} p_i x_{t,K+1}^{(i)}\right)$
end for

For notational convenience, we define the following two quantities:

$$\rho_{\tau}^{+(i)} = \max_{j} \rho_{\tau}^{+}(i,j), \quad \bar{\rho}_{\tau}^{+(i)} = \frac{1}{M}\sum_{j=1}^{M} \rho_{\tau}^{+}(i,j).$$

Lemma 3 ([39, Lemma 5.7]): Let $\widehat{\Gamma}$ be the set obtained from the k-th iteration at client i. Then, we have

$$\mathbb{E}_{J_k}^{(i)} \|b_{t,k}^{(i)} - x^*\|_2^2 \le \beta_1(i) \mathbb{E}_{J_k}^{(i)} \|P_{\widehat{\Gamma}}^{\perp}(b_{t,k}^{(i)} - x^*)\|_2^2 + \xi_1(i),$$

where

$$\beta_1(i) = \frac{\bar{\rho}_{4\tau}^{+(i)}}{2\rho_{4\tau}^{-}(i) - \bar{\rho}_{4\tau}^{+(i)}}, \quad \xi_1(i) = \frac{2\mathbb{E}_{J_{k,j}}^{(i)} \|P_{\widehat{\Gamma}} \nabla g_{i,j}(x^*)\|_2^2}{\bar{\rho}_{4\tau}^{+(i)}(2\rho_{4\tau}^{-}(i) - \bar{\rho}_{4\tau}^{+(i)})}.$$

Here J_k denotes the set of all previous mini-batch indices $j_1, \ldots, j_k$ randomly selected in or before the k-th step of the local iterations at the i-th client and $\mathbb{E}_{J_k}^{(i)}$ is the expectation taken over J_k.

The following lemma is an extended version of Lemma 3 in [14], whose proof further utilizes Young's inequality to control the trade-off between contraction and residual error due to the noise of the stochastic gradient for the FL setting. It also provides a refinement for the exact projection operator. Since the proof of the lemma is substantially different from the original version due to nontrivial changes to accommodate FL setting, we include its proof.

Lemma 4: Let $\widehat{\Gamma}$ be the set obtained from the k-th iteration at client i. Then, for any $\theta > 0$, we have

$$\mathbb{E}_{j_k}^{(i)} \|P_{\widehat{\Gamma}}^{\perp}(b_{t,k}^{(i)} - x^*)\|_2^2 \le \beta_2(i) \|x_{t,k}^{(i)} - x^*\|_2^2 + \xi_2(i)$$

where, shown in the equation at the bottom of the next page. Here $\mathbb{E}_{j_k}^{(i)}$ is the expectation taken over the randomly selected mini-batch index j_k for the stochastic gradient in the k-th step of the local iterations at the i-th client.

Proof Sketch: Our goal is to estimate the error between the solution to the subproblem in Algorithm 1 and x^* incurred by applying the thresholding operator, in terms of the residual error between the previous iterate and x^*. First, since both solution to the subproblem $b_{t,k}^{(i)}$ and x^* belong to the subspace associated to the support estimate set $\widehat{\Gamma}$, and from the property of the projection operator $P_{\widehat{\Gamma}}^{\perp}$, we have $\|P_{\widehat{\Gamma}}^{\perp}(b_{t,k}^{(i)} - x^*)\| \le \left\|x^* - x_{t,k}^{(i)} - P_{\Gamma}\left(x^* - x_{t,k}^{(i)}\right)\right\|$. We aim to find an upper bound for the right hand side in expectation taken over the randomly selected mini-batch, which is of the form of (a contraction factor) $\cdot \|x^* - x_{t,k}^{(i)}\|$ + $C \cdot \zeta^*$ (ζ^* is the heterogeneity bound at the global optima in Assumption 3). On the other hand, leveraging the $\mathscr{A}$ restricted strong convexity assumption along with the property of the approximate projection operator, $f_i(x^*) - f_i(x_{t,k}^{(i)})$ can be shown to be lower bounded by $\mathbb{E}_{j_k} \|P_{\Gamma} g_{i,j_k}(x_{t,k}^{(i)})\| \|x^* - x_{t,k}^{(i)}\|$ with some additional error term from the approximation projection operator. We apply the $\mathscr{A}$ restricted smoothness assumption to this lower bound, which can be lower bounded further by a quadratic function in the norm of a vector involving $x^* - x_{t,k}^{(i)}$ with additional error term due to stochastic gradient noise. Solving this quadratic function followed by a simple comparison with $\left\|x^* - x_{t,k}^{(i)} - P_{\Gamma}\left(x^* - x_{t,k}^{(i)}\right)\right\|$ shows the claim in the lemma.

Proof: We start with by noticing $P_{\widehat{\Gamma}}^{\perp} b_{t,k}^{(i)} = 0$ and $P_{\widehat{\Gamma}}^{\perp} x_{t,k}^{(i)} = 0$ since both $b_{t,k}^{(i)}$ and $x_{t,k}^{(i)}$ belong to the span of $\mathscr{A}_{\widehat{\Gamma}}$. Let $\Delta := x^* - x_{t,k}^{(i)}$ and the set $\text{supp}_{\mathscr{A}}(\Delta)$ be denoted by R. Note that

$|R| \le 2\tau$. Hence, we have

$$\begin{aligned}\|P_{\widehat{\Gamma}}^{\perp}(b_{t,k}^{(i)}-x^*)\|_2 &= \|P_{\widehat{\Gamma}}^{\perp}(b_{t,k}^{(i)}-x_{t,k}^{(i)}+x_{t,k}^{(i)}-x^*)\|_2\\ &\le \|P_{\widehat{\Gamma}}^{\perp}(b_{t,k}^{(i)}-x_{t,k}^{(i)})\|_2+\|P_{\widehat{\Gamma}}^{\perp}(x_{t,k}^{(i)}-x^*)\|_2\\ &= \|P_{\widehat{\Gamma}}^{\perp}(x_{t,k}^{(i)}-x^*)\|_2\\ &\le \|P_{\Gamma}^{\perp}(x_{t,k}^{(i)}-x^*)\|_2\\ &= \|\Delta-P_{\Gamma}\Delta\|_2.\end{aligned}$$

Here the second inequality follows from the definitions of the sets Γ and $\widehat{\Gamma}$, $\Gamma \subset \widehat{\Gamma}$.

Now we estimate $\|\Delta - P_{\Gamma}\Delta\|_2^2$. With a slight abuse of notation, $\mathbb{E}_{j_k}^{(i)}$ will be denoted by $\mathbb{E}_{j_k}$ throughout the proof. First, from the $\mathscr{A}$-RSC property of f_i, we have

$$\begin{aligned}&f_i(x^*)-f_i(x_{t,k}^{(i)})-\frac{\rho_{4\tau}^{-}(i)}{2}\|x^*-x_{t,k}^{(i)}\|_2^2\\ &\quad\ge \left\langle \nabla f_i(x_{t,k}^{(i)}), x^*-x_{t,k}^{(i)}\right\rangle\\ &\quad= \mathbb{E}_{j_k}\left\langle \nabla g_{i,j_k}(x_{t,k}^{(i)}), x^*-x_{t,k}^{(i)}\right\rangle\\ &\quad= \mathbb{E}_{j_k}\left\langle P_R\nabla g_{i,j_k}(x_{t,k}^{(i)}), x^*-x_{t,k}^{(i)}\right\rangle\\ &\quad\ge -\mathbb{E}_{j_k}\|P_R\nabla g_{i,j_k}(x_{t,k}^{(i)})\|\|\Delta\|_2 \qquad (8)\end{aligned}$$

By applying the inequality (15) in [14] and from the fact that Γ is the support set after the projection of the stochastic gradient $\nabla g_{i,j_k}(x_{t,k}^{(i)})$ in Algorithm 1, we have

$$\begin{aligned}&\|P_R\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2\\ &\quad\le \|P_{\Gamma}\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2+\frac{\sqrt{\eta_1^2-1}}{\eta_1}\|P_{\Gamma}^{\perp}\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2.\end{aligned}$$

To make the notation simple, we define

$$z := -\frac{P_{\Gamma}\nabla g_{i,j_k}(x_{t,k}^{(i)})}{\|P_{\Gamma}g_{i,j_k}(x_{t,k}^{(i)})\|_2}\|\Delta\|_2.$$

Then, the term $-\mathbb{E}_{j_k}\|P_R\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2\|\Delta\|_2$ can be further bounded as follows.

$$\begin{aligned}&-\mathbb{E}_{j_k}\|P_R\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2\|\Delta\|_2\\ &\ge -\mathbb{E}_{j_k}\|P_{\Gamma}\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2\|\Delta\|_2\\ &\quad-\frac{\sqrt{\eta_1^2-1}}{\eta_1}\mathbb{E}_{j_k}\|P_{\Gamma}^{\perp}\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2\|\Delta\|_2\\ &= \mathbb{E}_{j_k}\left\langle P_{\Gamma}\nabla g_{i,j_k}(x_{t,k}^{(i)}), z\right\rangle-\frac{\sqrt{\eta_1^2-1}}{\eta_1}\mathbb{E}_{j_k}\|P_{\Gamma}^{\perp}\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2\|\Delta\|_2\\ &= \mathbb{E}_{j_k}\left\langle \nabla g_{i,j_k}(x_{t,k}^{(i)}), z\right\rangle-\frac{\sqrt{\eta_1^2-1}}{\eta_1}\mathbb{E}_{j_k}\|P_{\Gamma}^{\perp}\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2\|\Delta\|_2\\ &\ge \mathbb{E}_{j_k}\left\langle \nabla g_{i,j_k}(x_{t,k}^{(i)}), z\right\rangle-\frac{\sqrt{\eta_1^2-1}}{2\eta_1}\mathbb{E}_{j_k}\left(\|P_{\Gamma}^{\perp}\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2^2+\|\Delta\|_2^2\right)\\ &= \mathbb{E}_{j_k}\left\langle \nabla f_i\left(x_{t,k}^{(i)}\right), z\right\rangle+\mathbb{E}_{j_k}\left\langle \nabla g_{i,j_k}(x_{t,k}^{(i)})-\nabla f_i\left(x_{t,k}^{(i)}\right), z\right\rangle\\ &\quad-\frac{\sqrt{\eta_1^2-1}}{2\eta_1}\mathbb{E}_{j_k}\left(\|P_{\Gamma}^{\perp}\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2^2+\|\Delta\|_2^2\right),\end{aligned}$$

where the second inequality above follows from the AM-GM inequality, $ab \le (a^2+b^2)/2$ for nonnegative real numbers a, b.

Now, we apply the Young's inequality for the inner product space to the second term in the last line to obtain

$$\begin{aligned}&\left|\left\langle \nabla g_{i,j_k}(x_{t,k}^{(i)})-\nabla f_i\left(x_{t,k}^{(i)}\right), z\right\rangle\right|\\ &\quad\le \frac{\theta^2}{2}\|\nabla g_{i,j_k}(x_{t,k}^{(i)})-\nabla f_i\left(x_{t,k}^{(i)}\right)\|_2^2+\frac{1}{2\theta^2}\|z\|_2^2\end{aligned}$$

for any nonzero θ.
Hence, we have

$$\begin{aligned}&-\mathbb{E}_{j_k}\|P_R\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2\|\Delta\|_2\\ &\ge \mathbb{E}_{j_k}\left\langle \nabla f_i\left(x_{t,k}^{(i)}\right), z\right\rangle-\frac{\theta^2}{2}\mathbb{E}_{j_k}\|\nabla g_{i,j_k}(x_{t,k}^{(i)})-\nabla f_i\left(x_{t,k}^{(i)}\right)\|_2^2\\ &\quad-\frac{1}{2\theta^2}\mathbb{E}_{j_k}\|z\|_2^2-\frac{\sqrt{\eta_1^2-1}}{2\eta_1}\left(\mathbb{E}_{j_k}\|P_{\Gamma}^{\perp}\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2^2+\mathbb{E}_{j_k}\|\Delta\|_2^2\right)\\ &\ge \mathbb{E}_{j_k}\left\langle \nabla f_i\left(x_{t,k}^{(i)}\right), z\right\rangle-\frac{\theta^2}{2}\sigma_i^2-\frac{1}{2\theta^2}\mathbb{E}_{j_k}\|z\|_2^2\\ &\quad-\frac{\sqrt{\eta_1^2-1}}{2\eta_1}\left(\mathbb{E}_{j_k}\|P_{\Gamma}^{\perp}\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2^2+\mathbb{E}_{j_k}\|\Delta\|_2^2\right),\end{aligned}$$

where we have used (7) in Assumption 4 in the last inequality above.

Next, we obtain the upper bound for $\mathbb{E}_{j_k}\|P_{\Gamma}^{\perp}\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2^2$ as follows.

$$\begin{aligned}&\mathbb{E}_{j_k}\|P_{\Gamma}^{\perp}\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2^2\\ &\le \mathbb{E}_{j_k}\|\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2^2\\ &\le 3\mathbb{E}_{j_k}\|\nabla g_{i,j_k}(x_{t,k}^{(i)})-\nabla g_{i,j_k}(x^*)\|_2^2\\ &\quad+3\mathbb{E}_{j_k}\|\nabla g_{i,j_k}(x^*)-\nabla f_i(x^*)\|_2^2+3\mathbb{E}_{j_k}\|\nabla f_i(x^*)\|_2^2\\ &\le 3\mathbb{E}_{j_k}(\rho_{\tau}^{+}(i,j_k))^2\|x_{t,k}^{(i)}-x^*\|_2^2+3\sigma_i^2+3\|\nabla f_i(x^*)\|_2^2\\ &= 3\mathbb{E}_{j_k}(\rho_{\tau}^{+}(i,j_k))^2\|\Delta\|_2^2+3\sigma_i^2+3\|\nabla f_i(x^*)\|_2^2,\end{aligned}$$

where we have used the inequality $\|a+b+c\|_2^2 \le 3\|a\|_2^2+3\|b\|_2^2+3\|c\|_2^2$ in the second inequality, and Assumption 2 and Assumption 4 in the third inequality above.

$$\beta_2(i) = \left(2\frac{\left(\bar{\rho}_{4\tau}^{+(i)}+\frac{1}{\theta^2}\right)-\eta_1^2\rho_{4\tau}^{-}(i)}{\eta_1^2\rho_{4\tau}^{-}(i)}+\frac{2\sqrt{\eta_1^2-1}}{\eta_1\rho_{4\tau}^{-}(i)}(3\mathbb{E}_{j_k}(\rho_{\tau}^{+}(i,j_k))^2+1)\right),$$

$$\xi_2(i) = \begin{cases}\frac{8}{(\rho_{4\tau}^{-}(i))^2}\max\limits_{\substack{\Omega\subset[d]\\|\Omega|=4\tau}}\|P_{\Omega}\nabla f_i(x^*)\|_2^2+\frac{1}{\rho_{4\tau}^{-}(i)}\left[\left(2\theta^2+\frac{6\sqrt{\eta_1^2-1}}{\eta_1}\right)\sigma_i^2+\frac{6\sqrt{\eta_1^2-1}}{\eta_1}\|\nabla f_i(x^*)\|_2^2\right] & \text{if } \eta_1>1,\\ \frac{8}{(\rho_{4\tau}^{-}(i))^2}\max\limits_{\substack{\Omega\subset[d]\\|\Omega|=4\tau}}\|P_{\Omega}\nabla f_i(x^*)\|_2^2+\frac{2\theta^2\sigma_i^2}{\rho_{4\tau}^{-}(i)} & \text{if } \eta_1=1 \text{ (when the projection operator is exact).}\end{cases}$$

Combining this bound with inequality (8) yields

$$
\begin{aligned}
&f_i(x^*) - f_i\left(x_{t,k}^{(i)}\right) - \frac{\rho_{4\tau}^-(i)}{2}\|x^* - x_{t,k}^{(i)}\|_2^2 \\
&\geq \mathbb{E}_{j_k}\left\langle \nabla f_i\left(x_{t,k}^{(i)}\right), z\right\rangle - \frac{\theta^2}{2}\sigma_i^2 - \frac{1}{2\theta^2}\mathbb{E}_{j_k}\|z\|_2^2 \qquad (9)\\
&\quad - \frac{\sqrt{\eta_1^2-1}}{2\eta_1}\left(\mathbb{E}_{j_k}\|P_\Gamma^\perp \nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2^2 + \mathbb{E}_{j_k}\|\Delta\|_2^2\right) \qquad (10)\\
&\geq \mathbb{E}_{j_k}\left\langle \nabla f_i\left(x_{t,k}^{(i)}\right), z\right\rangle - \frac{\theta^2}{2}\sigma_i^2 - \frac{1}{2\theta^2}\mathbb{E}_{j_k}\|z\|_2^2 \\
&\quad - \frac{\sqrt{\eta_1^2-1}}{2\eta_1}\left[(3\mathbb{E}_{j_k}(\rho_\tau^+(i,j_k))^2+1)\|\Delta\|_2^2 + 3\sigma_i^2 + 3\|\nabla f_i(x^*)\|_2^2\right]. \qquad (11)
\end{aligned}
$$

On the other hand, using Lemma 1 for $g_{i,j}$, a consequence of the $\mathscr{A}$-RSS property, we have

$$
\left\langle \nabla g_{i,j}(x_{t,k}^{(i)}), z\right\rangle \geq g_{i,j}\left(x_{t,k}^{(i)} + z\right) - g_{i,j}\left(x_{t,k}^{(i)}\right) - \frac{\rho_{4\tau}^+(i,j)}{2}\|z\|_2^2
$$

for all $j \in [M]$. By taking the average over all $g_{i,j}$ over $j \in [M]$ on both sides of the inequality above and from the definitions of f_i and $\rho_{4\tau}^{+(i)}$, we obtain

$$
\left\langle \nabla f_i(x_{t,k}^{(i)}), z\right\rangle \geq f_i\left(x_{t,k}^{(i)} + z\right) - f_i\left(x_{t,k}^{(i)}\right) - \frac{\bar{\rho}_{4\tau}^{+(i)}}{2}\|z\|_2^2.
$$

Here we have used (6) in Assumption 4. We then take the expectation $\mathbb{E}_{j_k}$ on both sides of the inequality.

$$
\mathbb{E}_{j_k}\left\langle \nabla f_i(x_{t,k}^{(i)}), z\right\rangle \geq \mathbb{E}_{j_k} f_i\left(x_{t,k}^{(i)} + z\right) - f_i\left(x_{t,k}^{(i)}\right) - \frac{\bar{\rho}_{4\tau}^{+(i)}}{2}\mathbb{E}_{j_k}\|z\|_2^2.
$$

After applying this bound to inequality (9), we obtain

$$
\begin{aligned}
&f_i(x^*) - f_i\left(x_{t,k}^{(i)}\right) - \frac{\rho_{4\tau}^-(i)}{2}\|\Delta\|_2^2 \\
&\geq \mathbb{E}_{j_k} f_i\left(x_{t,k}^{(i)} + z\right) - f_i\left(x_{t,k}^{(i)}\right) - \frac{\bar{\rho}_{4\tau}^{+(i)}}{2}\mathbb{E}_{j_k}\|z\|_2^2 - \frac{\theta^2}{2}\sigma_i^2 - \frac{1}{2\theta^2}\mathbb{E}_{j_k}\|z\|_2^2 \\
&\quad - \frac{\sqrt{\eta_1^2-1}}{2\eta_1}\left[(3\mathbb{E}_{j_k}(\rho_\tau^+(i,j_k))^2+1)\|\Delta\|_2^2 + 3\sigma_i^2 + 3\|\nabla f_i(x^*)\|_2^2\right].
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
&\left(\frac{\bar{\rho}_{4\tau}^{+(i)}}{2} + \frac{1}{2\theta^2}\right)\mathbb{E}_{j_k}\|z\|_2^2 \qquad (12)\\
&- \frac{1}{2}\left(\rho_{4\tau}^-(i) - \frac{\sqrt{\eta_1^2-1}}{\eta_1}(3\mathbb{E}_{j_k}(\rho_\tau^+(i,j_k))^2+1)\right)\|\Delta\|_2^2 \\
&+ \left(\frac{\theta^2}{2} + \frac{3\sqrt{\eta_1^2-1}}{2\eta_1}\right)\sigma_i^2 + \frac{3\sqrt{\eta_1^2-1}}{2\eta_1}\|\nabla f_i(x^*)\|_2^2 \\
&\geq \mathbb{E}_{j_k} f_i\left(x_{t,k}^{(i)} + z\right) - f_i(x^*) \\
&\geq \frac{\rho_{4\tau}^-(i)}{2}\mathbb{E}_{j_k}\|x_{t,k}^{(i)} + z - x^*\|_2^2 + \mathbb{E}_{j_k}\left\langle \nabla f_i(x^*), x_{t,k}^{(i)} + z - x^*\right\rangle \qquad (13)\\
&= \frac{\rho_{4\tau}^-(i)}{2}\mathbb{E}_{j_k}\|\Delta - z\|_2^2 + \mathbb{E}_{j_k}\langle \nabla f_i(x^*), z - \Delta\rangle \\
&= \frac{\rho_{4\tau}^-(i)}{2}\mathbb{E}_{j_k}\|\Delta - z\|_2^2 + \mathbb{E}_{j_k}\langle \nabla f_i(x^*), P_{\Gamma\cup R}(z-\Delta)\rangle \qquad (\star)\\
&= \frac{\rho_{4\tau}^-(i)}{2}\mathbb{E}_{j_k}\|\Delta - y\|_2^2 + \mathbb{E}_{j_k}\langle P_{\Gamma\cup R}\nabla f_i(x^*), (z-\Delta)\rangle \\
&\geq \frac{\rho_{4\tau}^-(i)}{2}\mathbb{E}_{j_k}\|\Delta - z\|_2^2 - \mathbb{E}_{j_k}\|P_{\Gamma\cup R}\nabla f_i(x^*)\|_2\|z-\Delta\|_2 \\
&\geq \frac{\rho_{4\tau}^-(i)}{2}\|\Delta - z\|_2^2 - \max_{\substack{\Omega\subset[d]\\|\Omega|=4\tau}}\|P_\Omega \nabla f_i(x^*)\|_2\mathbb{E}_{j_k}\|\Delta - z\|_2. \qquad (14)
\end{aligned}
$$

Here, the inequality (12) follows from $\mathscr{A}$-RSC. In $(\star)$ of the above inequality chain, we have used the fact that $z = -\frac{P_\Gamma \nabla g_{i,j_k}(x_{t,k}^{(i)})}{\|P_\Gamma g_{i,j_k}(x_{t,k}^{(i)})\|}\|\Delta\|_2$. Let $u = \mathbb{E}_{j_k}\|\Delta - z\|_2$, $a = \rho_{4\tau}^-(i)$, $b = \max_{\substack{\Omega\subset[d]\\|\Omega|=4\tau}}\|P_\Omega \nabla f_i(x^*)\|_2$, and

$$
\begin{aligned}
c = &\left(\bar{\rho}_{4\tau}^{+(i)} + \frac{1}{\theta^2}\right)\mathbb{E}_{j_k}\|z\|_2^2 \\
&- \left(\rho_{4\tau}^-(i) - \frac{\sqrt{\eta_1^2-1}}{\eta_1}(3\mathbb{E}_{j_k}(\rho_\tau^+(i,j_k))^2+1)\right)\|\Delta\|_2^2 \\
&+ \left(\theta^2 + \frac{3\sqrt{\eta_1^2-1}}{\eta_1}\right)\sigma_i^2 + \frac{3\sqrt{\eta_1^2-1}}{\eta_1}\|\nabla f_i(x^*)\|_2^2.
\end{aligned}
$$

Then the inequality (14) can be rewritten as $au^2 - 2bu - c \leq 0$ which gives

$$
\mathbb{E}_{j_k}\|\Delta - z\|_2 \leq \sqrt{\frac{c}{a}} + \frac{2b}{a}.
$$

Moreover, we have

$$
\|\Delta - P_\Gamma\Delta\|_2^2 \leq \|\Delta - z\|_2^2.
$$

Combining the previous two bounds yields

$$
\mathbb{E}_{j_k}\|\Delta - P_\Gamma\Delta\|_2^2 \leq \left(\sqrt{\frac{c}{a}} + \frac{2b}{a}\right)^2 \leq \frac{2c}{a} + \frac{8b^2}{a^2}.
$$

On the other hand, since

$$
\|z\|_2^2 = \left\|-\frac{P_\Gamma \nabla g_{i,j_k}(x_{t,k}^{(i)})}{\|P_\Gamma g_{i,j_k}(x_{t,k}^{(i)})\|_2}\|\Delta\|_2\right\|_2^2 = \|\Delta\|_2^2,
$$

we have

$$
\begin{aligned}
c \leq &\left(\bar{\rho}_{4\tau}^{+(i)} + \frac{1}{\theta^2} - \rho_{4\tau}^-(i) + \frac{\sqrt{\eta_1^2-1}}{\eta_1}(3\mathbb{E}_{j_k}(\rho_\tau^+(i,j_k))^2+1)\right)\|\Delta\|_2^2 \\
&+ \left(\theta^2 + \frac{3\sqrt{\eta_1^2-1}}{\eta_1}\right)\sigma_i^2 + \frac{3\sqrt{\eta_1^2-1}}{\eta_1}\|\nabla f_i(x^*)\|_2^2.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\mathbb{E}_{j_k}\|\Delta - P_\Gamma\Delta\|_2^2 \\
&\leq \left(2\frac{\left(\bar{\rho}_{4\tau}^{+(i)} + \frac{1}{\theta^2}\right) - \eta_1^2\rho_{4\tau}^-(i)}{\eta_1^2\rho_{4\tau}^-(i)} + \frac{2\sqrt{\eta_1^2-1}}{\eta_1\rho_{4\tau}^-(i)}(3\mathbb{E}_{j_k}(\rho_\tau^+(i,j_k))^2+1)\right)\|\Delta\|_2^2 \\
&+ \frac{8}{(\rho_{4\tau}^-(i))^2}\max_{\substack{\Omega\subset[d]\\|\Omega|=4\tau}}\|P_\Omega \nabla f_i(x^*)\|_2^2 \\
&+ \frac{1}{\rho_{4\tau}^-(i)}\left[\left(2\theta^2 + \frac{6\sqrt{\eta_1^2-1}}{\eta_1}\right)\sigma_i^2 + \frac{6\sqrt{\eta_1^2-1}}{\eta_1}\|\nabla f_i(x^*)\|_2^2\right].
\end{aligned}
$$

□

Equipped with these lemmas, we are ready to prove our main result for the linear convergence of FedGradMP.

Theorem 5: Let x^* be the solution to (1) and x_0 be the initial feasible solution. Assume that the local objective f_i satisfies $\mathscr{A}$-RSC with constant $\rho_{4\tau}^-(i)$ in Assumption 1 and all of the functions $g_{i,j}$ associated with mini-batches satisfy $\mathscr{A}$-RSS with constant $\rho_{4\tau}^+(i,j)$ in Assumption 2. We further assume that $g_{i,j}$ satisfies the bounded variance condition of local stochastic gradients in Assumption 4 with variance bound σ_i^2. Let K be the number of local iterations at each client.

Then, for any $\theta > 0$, the expectation of the recovery error at the $(t+1)$-th round of FedGradMP described in Algorithm 1 obeys

$$\mathbb{E}\|x_{t+1} - x^*\|_2^2 \leq \kappa^{t+1}\|x_0 - x^*\|_2^2 + \frac{(2\eta_3^2+2)\nu}{1-\kappa}\sum_{i=1}^N p_i \frac{1-\mu(i)^K}{1-\mu(i)},$$

where

$$\kappa = (2\eta_3^2+2)\sum_{i=1}^N p_i \left[(1+\eta_2)^2\beta_1(i)\beta_2(i)\right]^K$$

and

$$\mu(i) = (1+\eta_2)^2\beta_1(i)\beta_2(i).$$

Here, shown in the equation at the bottom of the next page.

Remark 4: There are key factors that impact the rate of convergence of FedGradMP in the number of communication rounds and the residual error in Theorem 5 as we discuss below.

1) **Impact of parameters $\beta_1(i)$ and $\beta_2(i)$.** For a fixed number of local iterations K, one can see that as the product of the two parameters $\beta_1(i)$ and $\beta_2(i)$ becomes small, the convergence rate κ improves (decreases). The product $\beta_1(i)\beta_2(i)$ becomes small as the $\mathscr{A}$-RSC constant $\rho_{4\tau}^-(i)$ increases or the $\mathscr{A}$-RSS constant $\bar{\rho}_{4\tau}^{+(i)}$ decreases. Choosing a proper dictionary could often improve the RSC/RSS constants as shown in Section IV-D consequently leading to better convergence, which is also numerically demonstrated in Section V-B.2.c.
2) **Impact of the local iteration number K on the convergence rate.** Increasing the local iteration number K also makes the convergence rate κ decrease when all the terms $(1+\eta_2)^2\beta_1(i)\beta_2(i)$ are less than 1. To see this, recall that the convergence rate is given by
$$\kappa = (2\eta_3^2+2)\sum_{i=1}^N p_i \left[(1+\eta_2)^2\beta_1(i)\beta_2(i)\right]^K.$$
In this case, increasing K leads to the decay of each term in $\sum_{i=1}^N p_i \left[(1+\eta_2)^2\beta_1(i)\beta_2(i)\right]^K$, improving the convergence rate. It is possible, however, some of the terms in the sum $(1+\eta_2)^2\beta_1(i)\beta_2(i)$ exceed 1, while the sum $\sum_{i=1}^N p_i \left[(1+\eta_2)^2\beta_1(i)\beta_2(i)\right]^K$ is still less than 1 for small K, making the convergence rate less than 1. In this case, as we increase the local iteration number K, the largest term starts dominating the sum which could increase the convergence rate (even make it greater than 1 for large K), degrading the performance of FedGradMP.
3) **Impact of the local iteration number K on the residual error.** The residual error $\frac{(2\eta_3^2+2)\nu}{1-\kappa}\sum_{i=1}^N p_i \frac{1-\mu(i)^K}{1-\mu(i)}$ depends on the local iteration number K in a more complicated way. Even when all terms $(1+\eta_2)^2\beta_1(i)\beta_2(i)$ are less than 1, in which case increasing K makes the convergence rate κ decrease (making the factor $\frac{1}{1-\kappa}$ in the residual error decrease), but the factor $\frac{1-\mu(i)^K}{1-\mu(i)}$ increases in K. Hence, the dependence of residual error on the local iteration number K may not be simple and this is actually what we observe in the numerical experiment in Section V-D. This is consistent with commonly accepted knowledge on the effect of the local iteration number on the residual error in the FL literature [5], [6], [16], [17], [28]: taking more local steps at clients makes the local estimates closer to the local solutions while the local estimates could deviate from the global solution in the FL environment in general.

Remark 5 (Interpretation of Theorem 5): Theorem 5 states that the iterates of FedGradMP converge linearly up to the residual error of the solution x^* as long as $\kappa < 1$. The size of the residual error is proportional to ν. In particular, from the expression for ν in Theorem 5, one can see that $\nu = 0$ if the heterogeneity parameter $\zeta_* = 0$ and stochastic gradient noise $\sigma_i = 0$ for all $i \in [N]$. We take a look at the related scenarios in more detail below.

- $\zeta_* = 0$ or $\nabla f_i(x^*) = 0$ for all the client objective function f_i. For example, the function f_i could be the square loss for the noiseless observations of x^* with sparsity level τ.
- $\sigma_i = 0$ for all $1 \leq i \leq N$ holds if and only if $\nabla g_{i,j}(x) = \nabla f_i(x)$ almost surely for all τ-sparse vectors. This happens when the full batch of each client is used instead of mini-batches. In particular, when the projection operator is exact ($\eta_1 = 1$) and under a slightly strong heterogeneity assumption, the residual error is statistically optimal. More precisely, under slightly strong heterogeneity assumptions only at the solution x^* such as
$$\|P_\Omega \nabla f_i(x^*) - P_\Omega \nabla f(x^*)\|_2^2 \leq \beta^2 \|P_\Omega \nabla f(x^*)\|_2^2, \quad (15)$$
where $\beta > 0$ and Ω is any subset of $[d]$ with size 4τ, one can see that we have
$$\max_{\substack{\Omega\subset[d]\\|\Omega|=4\tau}} \|P_\Omega \nabla f_i(x^*)\|_2^2 \leq 2(1+\beta^2) \max_{\substack{\Omega\subset[d]\\|\Omega|=4\tau}} \|P_\Omega \nabla f(x^*)\|_2^2.$$
From Theorem 5, after a sufficient number of rounds, we have
$$\mathbb{E}\|x_{t+1} - x^*\|_2 \leq \left[\mathbb{E}\|x_{t+1} - x^*\|_2^2\right]^{1/2} \leq O\left(\left(\sum_{i=1}^N p_i \max_{\substack{\Omega\subset[d]\\|\Omega|=4\tau}} \|P_\Omega \nabla f_i(x^*)\|_2^2\right)^{1/2}\right),$$
which is bounded from above by $O\left(\max_{\substack{\Omega\subset[d]\\|\Omega|=4\tau}} \|P_\Omega \nabla f(x^*)\|_2\right)$. This is the optimal statistical bias for commonly used FL data including sub-Gaussian datasets of size $|D|$ that are independently generated

for each client, which is of order of $O\left(\sqrt{\frac{\tau \log n}{N|D|}}\right)$ for the sparse linear regression problem (when f_i are the square loss functions). The uniform bounded heterogeneity condition, which is much stronger than (15) is used to show the optimal statistical recovery of Fast FedDualAvg [47]. See [6], [14], and [47] for more details.

- The parameter θ provides a trade-off between the convergence rate and the residual error due to the stochastic gradient. In particular, when the full batch is used ($\sigma_i = 0$), then one can set $\theta = \infty$ giving the fastest convergence.
- When $\sigma_i \neq 0$, the second term of the residual error ν is not vanishing in the number of rounds t. The similar term for FedHT/FedIterHT [6] decreases in t, but this requires that the mini-batch size at each client goes to infinity in t, which could severely restrict the number of communication rounds for the applicability of their theory. The idea of increasing the mini-batch size in the number of iterations is not new and has been used in [7] and [48]. However, the settings for these works are not for FL and the rate of mini-batch size growth is moderate, whereas the growth rates of FedHT/FedIterHT need to be generally much higher – they grow exponentially in the number of local iterations at clients. This potential issue in FL methods based on exponentially increasing mini-batch sizes is also pointed out in [24].

Proof Sketch: By using the definition of approximate projection operator, the property of hard thresholding, and Jensen's inequality, it turns out that the expected residual between the global model parameter of FedGradMP and x^* can be bounded by the average of the expected residual between local model parameter and x^*. Then, the residual for each local model can be further bounded by Lemma 2, 3, and 4. After applying these lemmas, combining all the error terms and bounding them by using the heterogeneity bound at the global optima in Assumption 3 and the variance bound for stochastic gradient descent in Assumption 4 if necessary, we obtain the statement in the theorem.

Proof: [Proof of Theorem 5] Let $\mathscr{F}^{(t)}$ be the filtration by all the randomness up to the t-th communication round, which is all the selected mini-batch indices at all the client up to the t-th round. We begin with analyzing $\mathbb{E}\left[\|x_{t+1} - x^*\|_2^2 | \mathscr{F}^{(t)}\right]$, the expected error of the global iterate x_{t+1} at the $(t+1)$-th round and x^* conditioned on $\mathscr{F}^{(t)}$. Because we will work with this conditional expectation until the very end of the proof, by abusing the notation slightly, $\mathbb{E}\left[\cdot | \mathscr{F}^{(t)}\right]$ will be denoted by $\mathbb{E}[\cdot]$.

$$\begin{aligned}
&\mathbb{E}\|x_{t+1} - x^*\|_2^2 \\
&= \mathbb{E}\left\| P_{\Lambda_s}\left(\sum_{i=1}^N p_i x_{t,K+1}^{(i)}\right) - \sum_{i=1}^N p_i x_{t,K+1}^{(i)} + \sum_{i=1}^N p_i x_{t,K+1}^{(i)} - x^* \right\|_2^2 \qquad (16)\\
&\leq 2\mathbb{E}\left\| P_{\Lambda_s}\left(\sum_{i=1}^N p_i x_{t,K+1}^{(i)}\right) - \sum_{i=1}^N p_i x_{t,K+1}^{(i)} \right\|_2^2 \\
&\quad + 2\mathbb{E}\left\| \sum_{i=1}^N p_i x_{t,K+1}^{(i)} - x^* \right\|_2^2 \\
&\leq 2\eta_3^2 \mathbb{E}\left\| \mathscr{H}_\tau\left(\sum_{i=1}^N p_i x_{t,K+1}^{(i)}\right) - \sum_{i=1}^N p_i x_{t,K+1}^{(i)} \right\|_2^2 \\
&\quad + 2\mathbb{E}\left\| \sum_{i=1}^N p_i x_{t,K+1}^{(i)} - x^* \right\|_2^2 \\
&\leq (2\eta_3^2 + 2)\mathbb{E}\left\| \sum_{i=1}^N p_i x_{t,K+1}^{(i)} - x^* \right\|_2^2 \\
&= (2\eta_3^2 + 2)\mathbb{E}\left\| \sum_{i=1}^N p_i x_{t,K+1}^{(i)} - \sum_{i=1}^N p_i x^* \right\|_2^2 \\
&\leq (2\eta_3^2 + 2)\sum_{i=1}^N p_i \mathbb{E}_{J_K^{(i)}} \left\| x_{t,K+1}^{(i)} - x^* \right\|_2^2 \qquad (17)
\end{aligned}$$

where the second inequality follows from the definition of the approximation projector operator, the third follows from the fact that both x^* and $\mathscr{H}_\tau\left(\sum_{i=1}^N p_i x_{t,K+1}^{(i)}\right)$ are τ-sparse but $\mathscr{H}_\tau\left(\sum_{i=1}^N p_i x_{t,K+1}^{(i)}\right)$ is the best τ-sparse approximation of $\sum_{i=1}^N p_i x_{t,K+1}^{(i)}$ with respect to the dictionary $\mathscr{A}$, and the last one is obtained by applying the Jensen's inequality.

$$\beta_1(i) = \frac{\bar{\rho}_{4\tau}^{+(i)}}{2\rho_{4\tau}^-(i) - \bar{\rho}_{4\tau}^{+(i)}}, \qquad \beta_2(i) = \left(2\frac{\left(\bar{\rho}_{4\tau}^{+(i)} + \frac{1}{\theta^2}\right) - \eta_1^2 \rho_{4\tau}^-(i)}{\eta_1^2 \rho_{4\tau}^-(i)} + \frac{2\sqrt{\eta_1^2 - 1}}{\eta_1 \rho_{4\tau}^-(i)}(3\mathbb{E}_{j_k}(\rho_\tau^+(i, j_k))^2 + 1)\right),$$

$$\nu = \begin{cases} (1+\eta_2)^2 \max_i \left(\frac{8\beta_1(i)}{(\rho_{4\tau}^-(i))^2} + \frac{4}{\bar{\rho}_{4\tau}^{+(i)}(2\rho_{4\tau}^-(i) - \bar{\rho}_{4\tau}^{+(i)})} + \frac{\beta_1(i)}{\rho_{4\tau}^-(i)}\frac{6\sqrt{\eta_1^2-1}}{\eta_1}\right)\zeta_*^2 \\ \quad + (1+\eta_2)^2 \sum_{i=1}^N p_i \left[\frac{\beta_1(i)}{\rho_{4\tau}^-(i)}\left(2\theta^2 + \frac{6\sqrt{\eta_1^2-1}}{\eta_1}\right) + \frac{4}{\bar{\rho}_{4\tau}^{+(i)}(2\rho_{4\tau}^-(i) - \bar{\rho}_{4\tau}^{+(i)})}\right]\sigma_i^2 \qquad \text{if } \eta_1 > 1, \\ (1+\eta_2)^2 \max_i \left(\frac{8\beta_1(i)}{(\rho_{4\tau}^-(i))^2} + \frac{4}{\bar{\rho}_{4\tau}^{+(i)}(2\rho_{4\tau}^-(i) - \bar{\rho}_{4\tau}^{+(i)})}\right) \sum_{i=1}^N p_i \max_{\substack{\Omega \subset [d] \\ |\Omega| = 4\tau}} \|P_\Omega \nabla f_i(x^*)\|_2^2 \\ \quad + (1+\eta_2)^2 \sum_{i=1}^N p_i \left[2\frac{\beta_1(i)}{\rho_{4\tau}^-(i)}\theta^2 + \frac{4}{\bar{\rho}_{4\tau}^{+(i)}(2\rho_{4\tau}^-(i) - \bar{\rho}_{4\tau}^{+(i)})}\right]\sigma_i^2 \qquad \text{if } \eta_1 = 1. \end{cases}$$

After applying Lemma 2, 3, and 4 sequentially to (17), we obtain

$$\begin{aligned}
&\sum_{i=1}^{N} p_i \mathbb{E}_{J_K}^{(i)} \left\| x_{t,K+1}^{(i)} - x^* \right\|_2^2 \qquad (18)\\
&\le (1+\eta_2)^2 \sum_{i=1}^{N} p_i \mathbb{E}_{J_K}^{(i)} \left\| b_{t,K}^{(i)} - x^* \right\|_2^2 \\
&\le (1+\eta_2)^2 \sum_{i=1}^{N} p_i \left[\beta_1(i) \mathbb{E}_{J_K}^{(i)} \| P_{\widehat{\Gamma}}^{\perp} (b_{t,K}^{(i)} - x^*) \|_2^2 + \xi_1(i) \right] \\
&= (1+\eta_2)^2 \sum_{i=1}^{N} p_i \beta_1(i) \mathbb{E}_{J_{K-1}, j_K}^{(i)} \| P_{\widehat{\Gamma}}^{\perp} (b_{t,K}^{(i)} - x^*) \|_2^2 \\
&\quad + (1+\eta_2)^2 \sum_{i=1}^{N} p_i \xi_1(i) \\
&\le (1+\eta_2)^2 \sum_{i=1}^{N} p_i \beta_1(i) \left[\beta_2(i) \mathbb{E}_{J_{K-1}}^{(i)} \| x_{t,K}^{(i)} - x^* \|_2^2 + \xi_2(i) \right] \\
&\quad + (1+\eta_2)^2 \sum_{i=1}^{N} p_i \xi_1(i) \\
&= (1+\eta_2)^2 \sum_{i=1}^{N} p_i \beta_1(i) \beta_2(i) \mathbb{E}_{J_{K-1}}^{(i)} \| x_{t,K}^{(i)} - x^* \|_2^2 \\
&\quad + (1+\eta_2)^2 \sum_{i=1}^{N} p_i \left(\beta_1(i) \xi_2(i) + \xi_1(i) \right). \qquad (19)
\end{aligned}$$

First, the term $\xi_1(i)$ can be bounded as follows:

$$\begin{aligned}
\xi_1(i) &= \frac{2(\mathbb{E}_{J_K,j}^{(i)} \| P_{\widehat{\Gamma}} \nabla g_{i,j}(x^*) \|_2^2)}{\bar{\rho}_{4\tau}^{+(i)} (2\rho_{4\tau}^{-}(i) - \bar{\rho}_{4\tau}^{+(i)})} \\
&\le \frac{4}{\bar{\rho}_{4\tau}^{+(i)} (2\rho_{4\tau}^{-}(i) - \bar{\rho}_{4\tau}^{+(i)})} \left(\max_{\substack{\Omega \subset [d] \\ |\Omega| = 4\tau}} \| P_\Omega \nabla f_i(x^*) \|_2^2 + \sigma_i^2 \right)
\end{aligned}$$

This is from the property of the projection operator and (7) in Assumption 4 implying

$$\begin{aligned}
&\mathbb{E}_j^{(i)} \| P_{\widehat{\Gamma}} \nabla g_{i,j}(x) - P_{\widehat{\Gamma}} \nabla f_i(x) \|_2^2 \\
&\quad \le \mathbb{E}_j^{(i)} \| \nabla g_{i,j}(x) - \nabla f_i(x) \|_2^2 \le \sigma_i^2,
\end{aligned}$$

so we have

$$\begin{aligned}
&\mathbb{E}_j^{(i)} \| P_{\widehat{\Gamma}} \nabla g_{i,j}(x^*) \|_2^2 \\
&\quad \le 2(\mathbb{E}_j^{(i)} \| P_{\widehat{\Gamma}} \nabla f_i(x^*) \|_2^2 + \sigma_i^2) \\
&\quad \le 2 \left(\max_{\substack{\Omega \subset [d] \\ |\Omega| = 4\tau}} \| P_\Omega \nabla f_i(x^*) \|_2^2 + \sigma_i^2 \right).
\end{aligned}$$

Hence, each term $\beta_1(i)\xi_2(i) + \xi_1(i)$ in (19) can be bounded as below.

$$\begin{aligned}
&\beta_1(i)\xi_2(i) + \xi_1(i) \\
&\le \beta_1(i) \left(\frac{8}{(\rho_{4\tau}^{-}(i))^2} \max_{\substack{\Omega \subset [d] \\ |\Omega| = 4\tau}} \| P_\Omega \nabla f_i(x^*) \|_2^2 \right. \\
&\quad \left. + \frac{1}{\rho_{4\tau}^{-}(i)} \left[\left(2\theta^2 + \frac{6\sqrt{\eta_1^2 - 1}}{\eta_1} \right) \sigma_i^2 + \frac{6\sqrt{\eta_1^2 - 1}}{\eta_1} \| \nabla f_i(x^*) \|_2^2 \right] \right) \\
&\quad + \frac{4}{\bar{\rho}_{4\tau}^{+(i)} (2\rho_{4\tau}^{-}(i) - \bar{\rho}_{4\tau}^{+(i)})} \left(\max_{\substack{\Omega \subset [d] \\ |\Omega| = 4\tau}} \| P_\Omega \nabla f_i(x^*) \|_2^2 + \sigma_i^2 \right) \\
&\le \left(\frac{8\beta_1(i)}{(\rho_{4\tau}^{-}(i))^2} + \frac{4}{\bar{\rho}_{4\tau}^{+(i)} (2\rho_{4\tau}^{-}(i) - \bar{\rho}_{4\tau}^{+(i)})} \right) \max_{\substack{\Omega \subset [d] \\ |\Omega| = 4\tau}} \| P_\Omega \nabla f_i(x^*) \|_2^2 \\
&\quad + \frac{\beta_1(i)}{\rho_{4\tau}^{-}(i)} \frac{6\sqrt{\eta_1^2 - 1}}{\eta_1} \| \nabla f_i(x^*) \|_2^2 \\
&\quad + \left[\frac{\beta_1(i)}{\rho_{4\tau}^{-}(i)} \left(2\theta^2 + \frac{6\sqrt{\eta_1^2 - 1}}{\eta_1} \right) + \frac{4}{\bar{\rho}_{4\tau}^{+(i)} (2\rho_{4\tau}^{-}(i) - \bar{\rho}_{4\tau}^{+(i)})} \right] \sigma_i^2.
\end{aligned}$$

By plugging the above bound to (19), we have

$$\begin{aligned}
&\sum_{i=1}^{N} p_i \mathbb{E}_{J_K}^{(i)} \left\| x_{t,K+1}^{(i)} - x^* \right\|_2^2 \\
&\le (1+\eta_2)^2 \sum_{i=1}^{N} p_i \beta_1(i) \beta_2(i) \mathbb{E}_{J_{K-1}}^{(i)} \| x_{t,K}^{(i)} - x^* \|_2^2 \qquad (20)\\
&+ (1+\eta_2)^2 \sum_{i=1}^{N} p_i \left(\left(\frac{8\beta_1(i)}{(\rho_{4\tau}^{-}(i))^2} + \frac{4}{\bar{\rho}_{4\tau}^{+(i)} (2\rho_{4\tau}^{-}(i) - \bar{\rho}_{4\tau}^{+(i)})} \right) \max_{\substack{\Omega \subset [d] \\ |\Omega| = 4\tau}} \| P_\Omega \nabla f_i(x^*) \|_2^2 \right. \\
&\quad \left. + \frac{\beta_1(i)}{\rho_{4\tau}^{-}(i)} \frac{6\sqrt{\eta_1^2 - 1}}{\eta_1} \| \nabla f_i(x^*) \|_2^2 \right) \\
&+ (1+\eta_2)^2 \sum_{i=1}^{N} p_i \left[\frac{\beta_1(i)}{\rho_{4\tau}^{-}(i)} \left(2\theta^2 + \frac{6\sqrt{\eta_1^2 - 1}}{\eta_1} \right) + \frac{4}{\bar{\rho}_{4\tau}^{+(i)} (2\rho_{4\tau}^{-}(i) - \bar{\rho}_{4\tau}^{+(i)})} \right] \sigma_i^2.
\end{aligned}$$

Now consider first the case when $\eta_1 > 1$. Then, since $\max_{\substack{\Omega \subset [d] \\ |\Omega| = 4\tau}} \| P_\Omega \nabla f_i(x^*) \|_2^2 \le \| \nabla f_i(x^*) \|_2^2$, we have

$$\begin{aligned}
&\sum_{i=1}^{N} p_i \mathbb{E}_{J_K}^{(i)} \left\| x_{t,K+1}^{(i)} - x^* \right\|_2^2 \\
&\le (1+\eta_2)^2 \sum_{i=1}^{N} p_i \beta_1(i) \beta_2(i) \mathbb{E}_{J_{K-1}}^{(i)} \| x_{t,K}^{(i)} - x^* \|_2^2 \\
&+ (1+\eta_2)^2 \sum_{i=1}^{N} p_i \left(\frac{8\beta_1(i)}{(\rho_{4\tau}^{-}(i))^2} + \frac{4}{\bar{\rho}_{4\tau}^{+(i)} (2\rho_{4\tau}^{-}(i) - \bar{\rho}_{4\tau}^{+(i)})} \right. \\
&\quad \left. + \frac{\beta_1(i)}{\rho_{4\tau}^{-}(i)} \frac{6\sqrt{\eta_1^2 - 1}}{\eta_1} \right) \| \nabla f_i(x^*) \|_2^2 \\
&+ (1+\eta_2)^2 \sum_{i=1}^{N} p_i \left[\frac{\beta_1(i)}{\rho_{4\tau}^{-}(i)} \left(2\theta^2 + \frac{6\sqrt{\eta_1^2 - 1}}{\eta_1} \right) + \frac{4}{\bar{\rho}_{4\tau}^{+(i)} (2\rho_{4\tau}^{-}(i) - \bar{\rho}_{4\tau}^{+(i)})} \right] \sigma_i^2 \\
&\le (1+\eta_2)^2 \sum_{i=1}^{N} p_i \beta_1(i) \beta_2(i) \mathbb{E}_{J_{K-1}}^{(i)} \| x_{t,K}^{(i)} - x^* \|_2^2 \\
&+ (1+\eta_2)^2 \max_i \left(\frac{8\beta_1(i)}{(\rho_{4\tau}^{-}(i))^2} + \frac{4}{\bar{\rho}_{4\tau}^{+(i)} (2\rho_{4\tau}^{-}(i) - \bar{\rho}_{4\tau}^{+(i)})} \right. \\
&\quad \left. + \frac{\beta_1(i)}{\rho_{4\tau}^{-}(i)} \frac{6\sqrt{\eta_1^2 - 1}}{\eta_1} \right) \sum_{i=1}^{N} p_i \| \nabla f_i(x^*) \|_2^2 \\
&+ (1+\eta_2)^2 \sum_{i=1}^{N} p_i \left[\frac{\beta_1(i)}{\rho_{4\tau}^{-}(i)} \left(2\theta^2 + \frac{6\sqrt{\eta_1^2 - 1}}{\eta_1} \right) + \frac{4}{\bar{\rho}_{4\tau}^{+(i)} (2\rho_{4\tau}^{-}(i) - \bar{\rho}_{4\tau}^{+(i)})} \right] \sigma_i^2.
\end{aligned}$$

Let $\mu(i) = (1+\eta_2)^2 \beta_1(i) \beta_2(i)$ and

$$\begin{aligned}
\nu &= (1+\eta_2)^2 \max_i \left(\frac{8\beta_1(i)}{(\rho_{4\tau}^{-}(i))^2} + \frac{4}{\bar{\rho}_{4\tau}^{+(i)} (2\rho_{4\tau}^{-}(i) - \bar{\rho}_{4\tau}^{+(i)})} + \frac{\beta_1(i)}{\rho_{4\tau}^{-}(i)} \frac{6\sqrt{\eta_1^2 - 1}}{\eta_1} \right) \zeta_*^2 \\
&+ (1+\eta_2)^2 \sum_{i=1}^{N} p_i \left[\frac{\beta_1(i)}{\rho_{4\tau}^{-}(i)} \left(2\theta^2 + \frac{6\sqrt{\eta_1^2 - 1}}{\eta_1} \right) + \frac{4}{\bar{\rho}_{4\tau}^{+(i)} (2\rho_{4\tau}^{-}(i) - \bar{\rho}_{4\tau}^{+(i)})} \right] \sigma_i^2.
\end{aligned}$$

Hence, when $\eta_1 > 1$, we have

$$\sum_{i=1}^{N} p_i \mathbb{E}_{J_K}^{(i)} \left\| x_{t,K+1}^{(i)} - x^* \right\|_2^2 \le \sum_{i=1}^{N} p_i \mu(i) \mathbb{E}_{J_{K-1}}^{(i)} \| x_{t,K}^{(i)} - x^* \|_2^2 + \nu. \qquad (21)$$

The case when the projection operator for the gradient is exact ($\eta_1 = 1$) follows the same argument. Setting $\eta_1 = 1$ in the inequality (20) reduces the inequality to

$$\begin{aligned}
&\sum_{i=1}^{N} p_i \mathbb{E}_{J_K}^{(i)} \left\| x_{t,K+1}^{(i)} - x^* \right\|_2^2 \\
&\le (1+\eta_2)^2 \sum_{i=1}^{N} p_i \beta_1(i)\beta_2(i) \mathbb{E}_{J_{K-1}}^{(i)} \|x_{t,K}^{(i)} - x^*\|_2^2 \\
&+ (1+\eta_2)^2 \sum_{i=1}^{N} p_i \left(\frac{8\beta_1(i)}{(\rho_{4\tau}^-(i))^2} \right. \\
&\qquad \left. + \frac{4}{\bar{\rho}_{4\tau}^{+(i)}(2\rho_{4\tau}^-(i) - \bar{\rho}_{4\tau}^{+(i)})} \right) \max_{\substack{\Omega \subset [d] \\ |\Omega| = 4\tau}} \|P_\Omega \nabla f_i(x^*)\|_2^2 \\
&+ (1+\eta_2)^2 \sum_{i=1}^{N} p_i \left[2\frac{\beta_1(i)}{\rho_{4\tau}^-(i)}\theta^2 + \frac{4}{\bar{\rho}_{4\tau}^{+(i)}(2\rho_{4\tau}^-(i) - \bar{\rho}_{4\tau}^{+(i)})} \right] \sigma_i^2 \\
&\le (1+\eta_2)^2 \sum_{i=1}^{N} p_i \beta_1(i)\beta_2(i) \mathbb{E}_{J_{K-1}}^{(i)} \|x_{t,K}^{(i)} - x^*\|_2^2 \\
&+ (1+\eta_2)^2 \max_i \left(\frac{8\beta_1(i)}{(\rho_{4\tau}^-(i))^2} \right. \\
&\qquad \left. + \frac{4}{\bar{\rho}_{4\tau}^{+(i)}(2\rho_{4\tau}^-(i) - \bar{\rho}_{4\tau}^{+(i)})} \right) \sum_{i=1}^{N} p_i \max_{\substack{\Omega \subset [d] \\ |\Omega| = 4\tau}} \|P_\Omega \nabla f_i(x^*)\|_2^2 \\
&+ (1+\eta_2)^2 \sum_{i=1}^{N} p_i \left[2\frac{\beta_1(i)}{\rho_{4\tau}^-(i)}\theta^2 + \frac{4}{\bar{\rho}_{4\tau}^{+(i)}(2\rho_{4\tau}^-(i) - \bar{\rho}_{4\tau}^{+(i)})} \right] \sigma_i^2.
\end{aligned}$$

The bound for ν for the exact projection case ($\eta_1 = 1$) is given as below.

$$\begin{aligned}
\nu &= (1+\eta_2)^2 \max_i \left(\frac{8\beta_1(i)}{(\rho_{4\tau}^-(i))^2} \right. \\
&\qquad \left. + \frac{4}{\bar{\rho}_{4\tau}^{+(i)}(2\rho_{4\tau}^-(i) - \bar{\rho}_{4\tau}^{+(i)})} \right) \sum_{i=1}^{N} p_i \max_{\substack{\Omega \subset [d] \\ |\Omega| = 4\tau}} \|P_\Omega \nabla f_i(x^*)\|_2^2 \\
&+ (1+\eta_2)^2 \sum_{i=1}^{N} p_i \left[2\frac{\beta_1(i)}{\rho_{4\tau}^-(i)}\theta^2 + \frac{4}{\bar{\rho}_{4\tau}^{+(i)}(2\rho_{4\tau}^-(i) - \bar{\rho}_{4\tau}^{+(i)})} \right] \sigma_i^2.
\end{aligned}$$

Then, applying the bound (21) to (17) repeatedly using the induction argument on K, we have

$$\begin{aligned}
&\mathbb{E}\|x_{t+1} - x^*\|_2^2 \\
&\le (2\eta_3^2 + 2) \left(\sum_{i=1}^{N} p_i \left[\mathbb{E}^{(i)} \mu(i)^K \|x_{t,1}^{(i)} - x^*\|_2^2 \right] + \sum_{i=1}^{N} p_i \frac{\nu(1-\mu(i)^K)}{1-\mu(i)} \right) \\
&= (2\eta_3^2 + 2) \mathbb{E} \left(\sum_{i=1}^{N} p_i \left[\mu(i)^K \|x_t - x^*\|_2^2 \right] + \sum_{i=1}^{N} p_i \frac{\nu(1-\mu(i)^K)}{1-\mu(i)} \right) \\
&= \kappa \|x_t - x^*\|_2^2 + (2\eta_3^2 + 2)\nu \sum_{i=1}^{N} p_i \frac{(1-\mu(i)^K)}{1-\mu(i)},
\end{aligned}$$

where the first equality follows from $x_{t,1}^{(i)} = x_t$ and the second follows from

$$\kappa = (2\eta_3^2 + 2) \sum_{i=1}^{N} p_i \mu(i)^K = (2\eta_3^2 + 2) \sum_{i=1}^{N} p_i \left[(1+\eta_2)^2 \beta_1(i)\beta_2(i) \right]^K.$$

Now, taking the unconditional expectation on both sides of the above yields

$$\begin{aligned}
&\mathbb{E}\left[\|x_{t+1} - x^*\|_2^2 \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\|x_{t+1} - x^*\|_2^2 | \mathscr{F}^{(t)} \right] \right] \\
&\le \kappa \mathbb{E}\left[\mathbb{E}\left[\|x_t - x^*\|_2^2 | \mathscr{F}^{(t-1)} \right] \right] + (2\eta_3^2 + 2)\nu \sum_{i=1}^{N} p_i \frac{(1-\mu(i)^K)}{1-\mu(i)}.
\end{aligned}$$

By applying this result repeatedly, we obtain

$$\mathbb{E}\|x_{t+1} - x^*\|_2^2 \le \kappa^{t+1} \|x_0 - x^*\|_2^2 + \frac{(2\eta_3^2 + 2)\nu}{1-\kappa} \sum_{i=1}^{N} p_i \frac{1-\mu(i)^K}{1-\mu(i)}.$$

□

Corollary 6: Under the same conditions and notation in Theorem 5, we have

$$\begin{aligned}
\mathbb{E} f(x_{t+1}) &\le f(x^*) + \frac{1}{2\rho} \|\nabla f(x^*)\|_2^2 \\
&\quad + \rho \left[\kappa^{t+1} \|x_0 - x^*\|_2^2 + \frac{(2\eta_3^2 + 2)\nu}{1-\kappa} \sum_{i=1}^{N} p_i \frac{1-\mu(i)^K}{1-\mu(i)} \right],
\end{aligned}$$

where $\rho = \sum_{i=1}^{N} p_i \bar{\rho}_\tau^{+(i)}$.
See Appendix for the proof of Lemma 6.

IV. Discussion and Extensions

A. Inexact FedGradMP

In FedGradMP, each client solves the minimization problem $\underset{x}{\mathbf{argmin}} f_i(x)$ over a subspace $\mathscr{R}(\mathscr{A}_{\widehat{\Gamma}})$ to update the support estimate of the solution x^*. When a closed-form solution exists to the minimization problem such as the least squares problem and the sparsity level τ is relatively small compared with the signal dimension, an exact minimizer can be obtained efficiently. This can be achieved, for example, by computing the pseudo-inverse with respect to a τ-dimensional subspace $\mathscr{R}(\mathscr{A}_{\widehat{\Gamma}})$ or by algorithms based on Cholesky, QR factorizations, or SVD for the least squares problem [49], [50].

But for the other cases, although the sub-optimization problem is typically convex due to the $\mathscr{A}$-RSC assumption, one may still want to reduce the computational cost in the optimization. By solving it only approximately but with a desired accuracy, we can save computational resources further. Because the local loss function f_i for the i-th client satisfies the $\mathscr{A}$-RSC and $\mathscr{A}$-RSS properties with the respective constants $\rho_\tau^-(i)$ and $\bar{\rho}_\tau^{+(i)}$, f_i is strongly convex/smooth with the same constants on the domain of the minimization problem, the linear subspace $\mathscr{R}(\mathscr{A}_{\widehat{\Gamma}})$. Recall that $|D_i|$ is the number of data points at the i-th client. We define a δ-approximate solution to $\underset{x}{\mathbf{argmin}} f_i(x)$ with $x \in \mathscr{R}(\mathscr{A}_{\widehat{\Gamma}})$ as a vector b such that $\|b - b_{\text{opt}}\|_2^2 \le \delta^2$ where b_{opt} is its exact solution. The number of steps required to achieve a δ-approximate solution at client i using popular standard algorithms is shown as follows:

- Gradient descent (GD): $O\left(|D_i| \left(\frac{\bar{\rho}_\tau^{+(i)}}{\rho_\tau^-(i)} \right) \log\left(\frac{1}{\delta} \right) \right)$ [46].
- Stochastic gradient descent with variance reduction such as SAG or SVRG: $O\left(|D_i| + \frac{\bar{\rho}_\tau^{+(i)}}{\rho_\tau^-(i)} \log\left(\frac{1}{\delta} \right) \right)$ [51], [52].

Since the domain is a τ-dimensional space, the computational complexity per data point of the above algorithms is $O(\tau)$ for the squared or logistic loss, so the overall complexity of the local step to compute a δ-approximate solution is $O\left(|D_i| \tau \left(\frac{\bar{\rho}_\tau^{+(i)}}{\rho_\tau^-(i)} \log\left(\frac{1}{\delta} \right) \right) \right)$ for GD. For Newton's method, the

Algorithm 2 Inexact FedGradMP With Partial Participation

Input: The number of rounds T, the number of clients N, the cohort size L, the number of local iterations K, weight vector p, the estimated sparsity level τ, $\eta_1, \eta_2, \eta_3, \delta$.
Output: $\hat{x} = x_T$.
Initialize: $x_0 = 0$, $\Lambda = \emptyset$.
for $t = 0, 1, \ldots, T-1$ **do**
  Randomly select a subset S_t **of clients with size** L
  for each client i **in** S_t**, do**
    $x_{t,1}^{(i)} = x_t$
    for $k = 1$ **to** K **do**
      Select a mini-batch index set $j_k := i_{t,k}^{(i)}$ uniformly at random from $\{1, 2, \ldots, M\}$
      Calculate the stochastic gradient $r_{t,k}^{(i)} = \nabla g_{i,j_k}\left(x_{t,k}^{(i)}\right)$
      $\Gamma = \text{approx}_{2\tau}(r_{t,k}^{(i)}, \eta_1)$
      $\widehat{\Gamma} = \Gamma \cup \Lambda$
      Solve $b_{t,k}^{(i)} = \underset{x}{\textbf{argmin}} f_i(x), \quad x \in \mathscr{R}(\mathscr{A}_{\widehat{\Gamma}})$ up to accuracy δ
      $\Lambda = \text{approx}_{\tau}(b_{t,k}^{(i)}, \eta_2)$
      $x_{t,k+1}^{(i)} = P_{\Lambda}(b_{t,k}^{(i)})$
      $\left(x_{t,k+1}^{(i)} \leftarrow \Pi_R\left(x_{t,k+1}^{(i)}\right)\right)$ [Optional projection onto a ball]
    end for
  end for
  $\Lambda_s = \text{approx}_{\tau}\left(\sum_{i=1}^{N} p_i x_{t,K+1}^{(i)}, \eta_3\right)$
  $x_{t+1} = P_{\Lambda_s}\left(\sum_{i=1}^{N} p_i x_{t,K+1}^{(i)}\right)$
  $\left(x_{t+1} \leftarrow \Pi_R\left(x_{t+1}\right)\right)$ [Optional projection onto a ball]
end for

total computational cost to achieve a δ-approximate solution is roughly $O\left(\left(|D_i|\tau^2 + \tau^3\right)\log\left(\frac{1}{\delta}\right)\right)$ [46]. Hence, if the sparsity level τ is much smaller than the signal dimension n, the subproblem in FedGradMP can be solved efficiently up to accuracy δ. As a comparison, most FL methods run (stochastic) gradient descent to solve $\underset{x}{\textbf{argmin}} f_i(x)$ over the whole space $\mathbb{R}^n$, which would cost computationally more to acquire its δ-approximate solution.

Theorem 7: Under the same notation and assumptions as in Theorem 5, for any $\theta > 0$, the expectation of the recovery error at the $(t+1)$-th round of inexact FedGradMP described in Algorithm 2 obeys

$$\mathbb{E}\|x_{t+1} - x^*\|_2^2 \leq \kappa^{t+1}\|x_0 - x^*\|_2^2 + \frac{(2\eta_3^2+2)(\nu+\delta^2)}{1-\kappa}\sum_{i=1}^{N} p_i \frac{1-\mu(i)^K}{1-\mu(i)},$$

where

$$\kappa = (2\eta_3^2+2)\sum_{i=1}^{N} p_i \left[2(1+\eta_2)^2\beta_1(i)\beta_2(i)\right]^K \quad \text{and}$$
$$\mu(i) = 2(1+\eta_2)^2\beta_1(i)\beta_2(i).$$

Here, the parameters $\beta_1(i), \beta_2(i)$, and ν are the same as in Theorem 5.

See Appendix for the proof of Theorem 7.

B. Client Sampling and the Impact of Cohort Size

In practical FL scenarios, it may not be possible for all of the clients to participate in each communication round. This could particularly stand out when there are a large population of clients or the communication bandwidth of connections between the server and clients is limited. A common theoretical assumption to capture this partial client participation is that participating clients for each communication round are drawn randomly according to some probability distribution, independent with other rounds. It could be considered as client sampling as noted in [16]. One could also consider more sophisticated sampling strategies such as importance sampling, but it seems to be not easy to implement such sampling techniques for FL since it could leak the private information of the client datasets [36]. Furthermore, in many real-world scenarios, client availability (which is usually random) solely controls participation rather than the server, ruling out the potential of using such methods [53].

For simplicity, we assume that the weight p_i is $1/N$ in the global objective function f and a fixed number of clients (the cohort size) participate per round as in [54] to study the impact of the cohort size.

Theorem 8: Assume the uniform weights $p_i = 1/N$ and L participating clients are drawn uniformly at random over the client set without replacement per round. Then, under the same assumptions and notation in Theorem 5, for any $\theta > 0$, the

expectation of the recovery error is bounded from above by

$$\mathbb{E}\|x_{t+1}-x^*\|_2^2 \le \kappa^{t+1}\|x_0-x^*\|_2^2 + \frac{(2\eta_3^2+2)\tilde{\nu}(1-\mu^K)}{(1-\mu)(1-\kappa)},$$

where

$$\kappa = (2\eta_3^2+2)\max_{\substack{S\subset[N]\\|S|=L}} \frac{1}{L}\sum_{i\in S}\left[(1+\eta_2)^2\beta_1(i)\beta_2(i)\right]^K,$$

$$\mu = \max_{i\in[N]}\left[(1+\eta_2)^2\beta_1(i)\beta_2(i)\right]^K,$$

and, shown in the equation at the bottom of the next page.

See Appendix for the proof of Theorem 8.

Remark 6: Note that the convergence rate κ of FedGradMP improves (decreases) as the cohort size L increases in Theorem 8, aligning with some of the previous works about the impact of cohort size on the convergence speed [34]. Our numerical experiments in Section V-E also validate our theory about the impact of cohort size on the convergence rates. On the other hand, it appears that the residual error in Theorem 8 is pessimistic and the actual behavior of FL algorithms depends on the cohort size in a more complicated way. See Section V-E for the numerical experiment and discussion.

C. FedGradMP With a Constraint

Many machine learning problems can be formulated as an ℓ_2-norm constrained optimization problem [5], [15], [55]. Since we focus on the FL setting with a sparse structure, our goal is to solve the following problem:

$$\min_{x\in\mathbb{R}^n} f(x) = \sum_{i=1}^N p_i f_i(x) \ \text{ subject to } \ \|x\|_{0,\mathscr{A}} \le \tau,\ \|x\|_2 \le R, \tag{22}$$

for some $R>0$, which is our main optimization problem (1) with the additional ℓ_2 constraint $\|x\|_2 \le R$. The ℓ_2 constraint ensures the global minimum exists in the domain. Another advantage of using the ℓ_2-norm constraint is that its orthogonal projection computationally is cheaper than projections of other constraints such as the ℓ_1-norm [15]. We denote by Π_R the orthogonal projection of a vector to the set $\{\|x\|_2 \le R\}$, which is implemented as follows. For any vector $u\in\mathbb{R}^N$,

$$\Pi_R(u) = \begin{cases} u, & \text{if } \|u\|_2 \le R;\\ Ru/\|u\|_2, & \text{otherwise.}\end{cases}$$

Let x^* be a minimizer of the problem (22) and the heterogeneity at the solution x^* is defined as in Assumption 3. By executing additional steps, the projection to a ℓ_2-norm ball in Algorithm 2, FedGradMP converges to the solution x^* under the same conditions in Theorem 5, 7, and 8. The proof follows a simple modification of the proofs of the theorems due to the fact that the orthogonal projection of a vector u to a ball with radius R does not increase the ℓ_2-norm distance between u and v for a vector v in the ball. For instance, we replace (18) in the proof of Theorem 5 as follows.

$$\sum_{i=1}^N p_i \mathbb{E}_{J_K}^{(i)} \left\| x_{t,K+1}^{(i)} - x^* \right\|_2^2$$

$$= \sum_{i=1}^N p_i \mathbb{E}_{J_K}^{(i)} \left\| \Pi_R\left(P_{\Lambda_s}\left(b_{t,K}^{(i)}\right)\right) - x^* \right\|_2^2$$

$$\le \sum_{i=1}^N p_i \mathbb{E}_{J_K}^{(i)} \left\| P_{\Lambda_s}\left(b_{t,K}^{(i)}\right) - x^* \right\|_2^2$$

$$\le (1+\eta_2)^2 \sum_{i=1}^N p_i \mathbb{E}_{J_K}^{(i)} \left\| b_{t,K}^{(i)} - x^* \right\|_2^2,$$

where we have used the fact that x^* belongs the ℓ_2-norm ball with radius R and the aforementioned property of Π_R in the first inequality above.

Similarly, note that all the local iterates satisfy $\|x_{t,K+1}^{(i)}\|_2 \le R$ because of the projection to the ball in Algorithm 2. Thus, their convex combination $\sum_{i=1}^N p_i x_{t,K+1}^{(i)}$ also belongs to the ball.

Now we apply the same argument to the first step of the proof of Theorem 5

$$\begin{aligned}
&\mathbb{E}\|x_{t+1}-x^*\|_2^2\\
&= \mathbb{E}\left\| \Pi_R\left(P_{\Lambda_s}\left(\sum_{i=1}^N p_i x_{t,K+1}^{(i)}\right)\right) - \sum_{i=1}^N p_i x_{t,K+1}^{(i)} + \sum_{i=1}^N p_i x_{t,K+1}^{(i)} - x^* \right\|_2^2\\
&\le 2\mathbb{E}\left\| \Pi_R\left(P_{\Lambda_s}\left(\sum_{i=1}^N p_i x_{t,K+1}^{(i)}\right)\right) - \sum_{i=1}^N p_i x_{t,K+1}^{(i)} \right\|_2^2\\
&\quad + 2\mathbb{E}\left\| \sum_{i=1}^N p_i x_{t,K+1}^{(i)} - x^* \right\|_2^2\\
&\le 2\mathbb{E}\left\| P_{\Lambda_s}\left(\sum_{i=1}^N p_i x_{t,K+1}^{(i)}\right) - \sum_{i=1}^N p_i x_{t,K+1}^{(i)} \right\|_2^2 + 2\mathbb{E}\left\| \sum_{i=1}^N p_i x_{t,K+1}^{(i)} - x^* \right\|_2^2.
\end{aligned}$$

After these modifications, we proceed as in the rest of the proof of Theorem 5.

D. Impact of Dictionary Choice

Recall that our convergence guarantees depend on the restricted convexity/smoothness ($\mathscr{A}$-RSC, $\mathscr{A}$-RSS) constants $\rho_{4\tau}^-(i)$ and $\bar{\rho}_{4\tau}^{+(i)}$ as many works for sparse recovery [6], [9], [13], [14]. In particular, the product $\beta_1(i)\beta_2(i)$ in Theorems 5, 7, and 8 critically impact the convergence rate κ; for faster convergence, $\beta_1(i)$ and $\beta_2(i)$ should be small as stated in Remark 4. This can be achieved especially when the $\mathscr{A}$-RSS/$\mathscr{A}$-RSC constants $\rho_{4\tau}^-(i)$ and $\bar{\rho}_{4\tau}^{+(i)}$ are close to each other or their ratio (the restricted condition number) is close to 1.

1) Sparse Linear Regression: When the local objective function is the square loss function associated with the local dataset at the client, the $\mathscr{A}$-RSC and $\mathscr{A}$-RSS constants essentially reduce to the restricted isometry property ($\mathscr{A}$-RIP) [20], [21]. Indeed, let the square loss function be given by $h(x) = \frac{1}{2l}\|Bx-y\|_2^2$ where the rows of matrix $B\in\mathbb{R}^{l\times m}$ are the input data vectors denoted by b_i and y is the observation vector. Assume that $\|b_i\|_2 = 1$ for all $1\le i\le l$, which can be done by normalizing the data vector b_i and corresponding y_i. Since the function h is the square loss function, by looking into its Hessian, we study the restricted strong convexity (RSC) and smoothness (RSS) properties. The Hessian $\nabla^2 h$ of h is given by

$$\frac{1}{l}B^\top B.$$

The RSC and RSS constants are the largest $c \geq 0$ and the smallest $d \geq 0$ such that $c\|w-z\|_2^2 \leq (w-z)^\top \left(\frac{1}{l}B^\top B\right)(w-z) \leq d\|w-z\|_2^2$ for all vectors w and z such that $\|w-z\|_0 \leq \tau$.

This observation and the definition of RIP [13] imply that if the RIP constant of $\frac{1}{\sqrt{l}}B$ is at least δ then $(1-\delta)\|z\|_2^2 \leq z^\top \left(\frac{1}{l}B^\top B\right) z \leq (1+\delta)\|z\|_2^2$ for all τ-sparse vectors z, making it satisfy the RSC/RSS with constants $1-\delta$ and $1+\delta$ respectively.

It could be possible, however, that the data matrix B whose rows consist of the local data at each client may not satisfy RSC with respect to the standard basis, but RSC with respect to a certain dictionary $\mathscr{A}$. Put it differently, if h is not restricted strong convex for τ-sparse vectors, then $B(w-z)=0$ for some vectors w and z such that $\|w-z\|_0 \leq \tau$ or $Bu=0$ for some τ-sparse vector u, i.e., B is not τ-RIP.

We present our idea of simply using a random Gaussian dictionary A to improve the ratio RSS to RSC constants of the associated new loss function $\frac{1}{2l}\|BAx-y\|_2^2$ (or improve the $\mathscr{A}$-RIP constant of B with respect to a dictionary A) with high probability.

Our idea to improve the RIP with a random dictionary is based on a recent development in high dimensional geometry. More specifically, we use the following theorem from [56].

Theorem 9 (Theorem 1.1 in [56]): Let $B \in \mathbb{R}^{l\times m}$ be a fixed matrix, let $A \in \mathbb{R}^{m\times n}$ be a mean zero, isotropic and sub-Gaussian matrix with sub-Gaussian parameter K and let $T \subset \mathbb{R}^n$ be a bounded set. Then

$$\mathbb{E}\sup_{x\in T}\left|\|BAx\|_2 - \|B\|_F\|x\|_2\right| \leq CK\sqrt{\log K}\,\|B\|\,[w(T)+\mathrm{rad}(T)],$$

and with probability at least $1-3e^{-u^2}$,

$$\sup_{x\in T}\left|\|BAx\|_2 - \|B\|_F\|x\|_2\right| \leq CK\sqrt{\log K}\,\|B\|\,[w(T)+u\cdot\mathrm{rad}(T)].$$

Here $w(T)$ is the Gaussian width for the set T, $\mathrm{rad}(T) = \sup_{y\in T}\|y\|_2$, and C is an absolute constant.

The following is an immediate consequence of the above theorem and the well-known fact that $w(T) \leq Cr\sqrt{\tau\log(n/\tau)}$ for the set T of all τ-sparse vectors x with $\|x\| \leq r$ for some universal constant $C>0$.

Corollary 10: Let $r>0$ and $\mathbb{B}$ be the closed unit ball in $\mathbb{R}^n$. For the set T of all τ-sparse vectors in $r\mathbb{B}$ and Gaussian random matrix A, we have

$$\mathbb{E}\sup_{x\in T}|\|BAx\|_2 - \|B\|_F\|x\|_2| \leq C\|B\|\left[r\sqrt{\tau\log(n/\tau)}+r\right],$$

and with probability at least $1-3e^{-u^2}$,

$$\sup_{x\in T}|\|BAx\|_2 - \|B\|_F\|x\|_2| \leq C\|B\|\left[r\sqrt{\tau\log(n/\tau)}+ru\right].$$

Since both terms in the bounds in Corollary 10 are homogeneous in r for all x in T with $\|x\|_2 = r$, the corollary implies that matrix $\frac{1}{\|B\|_F}BA$ satisfies the RIP with a constant $\delta_\tau = C_1\frac{\|B\|^2}{\|B\|_F^2}\tau\log(n/\tau)$ with high probability, whenever the stable rank of B

$$\mathrm{sr}(B) := \frac{\|B\|_F^2}{\|B\|^2} \geq C_2\tau\log(n/\tau)$$

for a sufficiently large constant $C_2>0$.

The above corollary can be readily applied to the data matrix B_{D_i} in the local objective function $f_i = \frac{1}{2|D_i|}\|B_{D_i}Ax-y\|_2^2$ for client i. First, recall that $\|b_i\|_2=1$ and by the definition of the Frobenious norm, $\|B_{D_i}\|_F = \sqrt{|D_i|}$. The data matrix B_{D_i} may not satisfy the RIP in general but $\frac{1}{\|B_{D_i}\|_F}B_{D_i}A = \frac{1}{\sqrt{|D_i|}}B_{D_i}A$ does with RIP constant

$$\delta_\tau = C\frac{\|B_{D_i}\|^2}{\|B_{D_i}\|_F^2}\tau\log(n/\tau) = C\frac{\|B_{D_i}\|^2}{|D_i|}\tau\log(n/\tau).$$

Thus, with high probability, $\frac{1}{2|D_i|}\|B_{D_i}Ax-y\|_2^2$ is $\mathscr{A}$-RSC and $\mathscr{A}$-RSS with the constant ratio $\frac{1+\delta_\tau}{1-\delta_\tau}$ with respect to the Gaussian random dictionary A, under the stable rank condition for B_{D_i} (which could be a mild condition for many data matrices). Since $\frac{1+\delta_\tau}{1-\delta_\tau}$ is close to 1 whenever the RIP constant δ_τ is close to 0, this makes $\beta_1(i)$ and $\beta_2(i)$ small, improving the convergence rate in Theorems 5 7, and 8 as we discussed before. Furthermore, note that since the Gaussian random matrix is statistically independent of the client datasets, there is no privacy leakage.

2) Sparse Binary Logistic Regression: The previous analysis for the square loss can be extended to the logistic losses. First, we consider the binary logistic loss function $h(x) = \frac{1}{l}\sum_{i=1}^l \log(1+\exp(-2y_jb_j^\top x))$ with input data vector b_j and labels $y_j \in \{-1,1\}$. Assume that $\|b_i\|_2 = 1$ for all $1 \leq i \leq l$ and x is a τ-sparse vector with $\|x\| \leq r$. Since the function h is twice-differentiable, we can study the RSC and RSS by investigating its Hessian. We denote the sigmoid function by $s(z) = \frac{1}{1+\exp(z)}$. By a direct computation or from the lecture notes https://www.cs.mcgill.ca/ dprecup/courses/ML/Lectures/ml-lecture05.pdf, one can verify that Hessian $\nabla^2 h$ of the logistic

$$\tilde{\nu} = \begin{cases} (1+\eta_2)^2\max_i\left(\frac{8\beta_1(i)}{(\rho_{4\tau}^-(i))^2} + \frac{4}{\bar{\rho}_{4\tau}^{+(i)}(2\rho_{4\tau}^-(i)-\bar{\rho}_{4\tau}^{+(i)})} + \frac{\beta_1(i)}{\rho_{4\tau}^-(i)}\frac{6\sqrt{\eta_1^2-1}}{\eta_1}\right)\zeta_*^2 \\ \quad +(1+\eta_2)^2\frac{1}{L}\sum_{i=1}^N\left[\frac{\beta_1(i)}{\rho_{4\tau}^-(i)}\left(2\theta^2+\frac{6\sqrt{\eta_1^2-1}}{\eta_1}\right)+\frac{4}{\bar{\rho}_{4\tau}^{+(i)}(2\rho_{4\tau}^-(i)-\bar{\rho}_{4\tau}^{+(i)})}\right]\sigma_i^2, \quad \text{if } \eta_1>1, \\ (1+\eta_2)^2\max_i\left(\frac{8\beta_1(i)}{(\rho_{4\tau}^-(i))^2}+\frac{4}{\bar{\rho}_{4\tau}^{+(i)}(2\rho_{4\tau}^-(i)-\bar{\rho}_{4\tau}^{+(i)})}\right)\left(\frac{1}{N}\sum_{i=1}^N\max_{\substack{\Omega\subset[d]\\|\Omega|=4\tau}}\|P_\Omega\nabla f_i(x^*)\|_2^2\right) \\ \quad +(1+\eta_2)^2\frac{1}{L}\sum_{i=1}^N\left[\beta_1(i)\frac{2\theta^2}{\rho_{4\tau}^-(i)}+\frac{4}{\bar{\rho}_{4\tau}^{+(i)}(2\rho_{4\tau}^-(i)-\bar{\rho}_{4\tau}^{+(i)})}\right]\sigma_i^2 \quad \text{if } \eta_1=1. \end{cases}$$

loss function h is given by

$$\nabla^2 h(x) = \frac{1}{l} B^\top \Lambda(x) B,$$

where B is the matrix whose rows b_i consist of a client dataset and $\Lambda(x)$ is the diagonal matrix whose j-th diagonal entry is $4s(2b_j^\top x)(1-s(2b_j^\top x))$.

First, it is easy to check that h is L-smooth [57] with

$$L \le \frac{1}{l}\sum_{i=1}^{l} \max_x 4s(2b_j^\top x)(1-s(2b_j^\top x)) \le 1.$$

Next, since x is a τ-sparse vector with $\|x\|_2 \le r$, from the definition of the sigmoid function τ, we deduce $[\Lambda(x)]_{jj} \ge \frac{4}{(1+\exp(r))^2}$. Then, we have

$$\frac{4}{(1+\exp(r))^2}\cdot\frac{B^\top B}{l} \preceq \nabla^2 h = \frac{1}{l}B^\top\Lambda(x)B \preceq \frac{B^\top B}{l}.$$

Note that the above bound does not imply that h is RSC since it is possible that $Bx = 0$ for some τ-sparse vector x. However, if we use a random Gaussian dictionary A, then a similar derivation gives

$$\frac{4}{(1+\exp(r))^2}\cdot\frac{A^\top B^\top BA}{l} \preceq \nabla^2 h = \frac{1}{l}A^\top B^\top\Lambda(x)BA \preceq \frac{A^\top B^\top BA}{l}.$$

Collorary 10 implies that

$$\begin{aligned}&\|B\|_F\|x\| - C\|B\|\|x\|\left[\sqrt{\tau\log(n/\tau)}+u\right]\\ &\le \sqrt{x^\top A^\top B^\top BAx} \le \|B\|_F\|x\| + C\|B\|\|x\|\left[\sqrt{\tau\log(n/\tau)}+u\right]\end{aligned}$$

for all τ-sparse vectors with probability at least $1-3e^{-u^2}$. Finally, it is easy to check that applying this bound to the previous bound on $\nabla^2 h$ yields that with high probability, h is $\mathscr{A}$-RSS and $\mathscr{A}$-RSC with the constant ratio

$$\frac{(1+\exp(r))^2}{4}\cdot\left(\frac{\|B\|_F + C\|B\|\left[\sqrt{\tau\log(n/\tau)}+u\right]}{\|B\|_F - C\|B\|\left[\sqrt{\tau\log(n/\tau)}+u\right]}\right)^2,$$

which is close to $\frac{(1+\exp(r))^2}{4}$ as long as the stable rank

$$\mathrm{sr}(B) = \frac{\|B\|_F^2}{\|B\|^2} \gg \tau\log(n/\tau).$$

This implies that for any τ-sparse vector x with $\|x\|_2 \le r$, the logistic loss function is $\mathscr{A}$-RSC/$\mathscr{A}$-RSS with respect to a random Gaussian dictionary with constant $\approx \frac{(1+\exp(r))^2}{4}$ under a mild condition, even if the function is not RSC/RSS in the standard basis (for example, the ratio is infinite if the RSC constant in the standard basis is 0). We apply the above argument to each binary logistic loss function f_i. Note that since the RSC/RSS ratio can be understood as a restricted condition number that controls the convergence rates by Theorems 5, 7, and 8, a random Gaussian dictionary is appropriate for FedGradMP with an ℓ_2-norm constraint that is discussed in Section IV-C.

3) Sparse Multiclass Logistic Regression: We only highlight the difference between the multiclass and binary logistic regression cases since the arguments are very similar to each other. Consider the multinomial logistic regression function with K classes. The label y_{ij} is 1 if the j-th training input belongs to the class i and 0 otherwise, b_j are normalized data vectors (i.e., $\|b_i\|_2 = 1$), and $x^{(i)}$ are τ-sparse classifier vectors with $\|x^{(i)}\|_2 \le r$.

The corresponding loss function is given as

$$\begin{aligned}&h(x^{(1)}, x^{(2)}, \ldots, x^{(K)})\\ &= \sum_{j=1}^{l}\left[\sum_{i=1}^{K} -y_{ij} b_j^\top x^{(i)} + \ln\left(\exp\left(\sum_{i=1}^{K} b_j^\top x^{(i)}\right)\right)\right].\end{aligned}$$

Similar to the binary logistic regression case, the direct computation of the Hessian of h gives

$$\nabla^2_{x^{(i)}} h = \frac{1}{l}B^\top\Lambda(x^{(i)})B.$$

Here $\Lambda(x)$ is a diagonal matrix whose diagonal entries are defined as $[\Lambda(x)]_{jj} = s(b_j^\top x)(1-s(b_j^\top x))$, where

$$s(b_j^\top x) = \frac{\exp(b_j^\top x)}{1+\sum_{i=1}^{K}\exp(b_i^\top x)}.$$

By the same argument used for the sparse binary logistic regression, h is $\mathscr{A}$-RSS and $\mathscr{A}$-RSC with a constant ratio

$$(1+K\exp(2r))^2\cdot\left(\frac{\|B\|_F + C\|B\|\left[\sqrt{\tau\log(n/\tau)}+u\right]}{\|B\|_F - C\|B\|\left[\sqrt{\tau\log(n/\tau)}+u\right]}\right)^2.$$

This again indicates that for any τ-sparse vector x with $\|x\| \le r$, the multiclass logistic loss function is $\mathscr{A}$-RSC/$\mathscr{A}$-RSS with respect to a random Gaussian dictionary even if it may not be RSC/RSS in the standard basis. As we saw in the binary logistic regression, this shows that it is beneficial to use a random Gaussian dictionary in logistic regression for ℓ_2-norm constrained FedGradMP, which is also verified in our numerical experiments in Section V-B.2.c.

Remark 7 (Random dictionary): The idea of using a Gaussian random dictionary to improve the restricted condition number should be distinguished from the sketching in the FL literature [30], [33], [58]. Our formulation and analysis are fundamentally different from those for sketching schemes that focus on compressing the gradient to save communication cost between a server and clients. In these work [30], [33], and [58], the sketching mappings (commonly random matrices) developed for numerical linear algebra [59] are applied after the clients computed the gradients to compress the information, whereas our Gaussian random mappings are used to transform the domain of the solution space to improve the restricted condition number.

Remark 8 (Sharing the dictionary among clients): The server either broadcasts the dictionary to clients or the shared memory can be used to share the dictionary among clients as suggested in [60] and [61]. When the latter option is available, the server does not need to send the dictionary to the clients.

V. Numerical Experiments

In this section, we provide numerical experiments validating our theory and showing the effectiveness of the proposed algorithm.

A. FedGradMP for Sparse Linear Regression

1) Synthetic Dataset:

a) Experiment settings: The first numerical experiment uses synthetic datasets. We run FedGradMP (Algorithm 1) with the square loss function. More precisely, we consider the component function of the form $f_i = \frac{1}{2\|D_i\|}\|A_{D_i}x - y_{D_i}\|_2^2$ where A_{D_i} is the client i data matrix in $\mathbb{R}^{100\times 1000}$ whose elements are synthetically generated according to the normal distribution $\mathcal{N}(\mu_i, 1/i^{1.1})$ with the mean value μ_i that is randomly generated from the mean-zero Gaussian with variance α. Here, y_{D_i} are observations with $y_{D_i} = A_{D_i}x^{\#}$ and $x^{\#} \in \mathbb{R}^{1000}$ is a randomly generated vector that is 10-sparse with respect to the standard basis whose 10 nonzero components are drawn from the unit sphere $\mathbb{S} \subset \mathbb{R}^{10}$. Since the random mean μ_i obeys the normal distribution $\mathcal{N}(\alpha, 0)$, the parameter α modulates the degree of client data heterogeneity: as α increases, the more likely μ_i vary wildly which in turn makes the client dataset distributions more different. This type of model is commonly used in FL numerical experiments to generate synthetic datasets [5], [6], [16] since randomly generated mean μ_i and decreasing variance $1/i^{1.1}$ make the client dataset heterogeneous.

The number of clients is 50, the number of data points of each client is 100, and the mini-batch size of each client for FedGradMP is 40.

b) Simulation results: Figure 1 shows that FedGradMP converges linearly for various heterogeneity levels α, validating Theorem 5. Note that the higher α is, the larger the variance of random mean shift μ_i or the higher the degree of heterogeneity is. The curves on the top panel are the relative error of FedGradMP for the noiseless case and the curves on the bottom are for the Gaussian noise case. We observe that FedGradMP still converges for highly heterogeneous datasets but with slower convergence rates in both cases.

2) Real Dataset: Sparse Video Recovery: In this experiment, we test FedGradMP on video frame recovery from a real-world dataset. Our dataset is a xylophone video consisting of 120 frames from YouTube https://www.youtube.com/watch?v=ORipY6OXltY, which can be also downloaded from the MathWorks website https://www.mathworks.com/help/matlab/ref/videoreader.html. Each frame is of size 240×320 after the conversion to gray-scale frames. We reshape the 82-th frame as a vector in $\mathbb{R}^{76800}$ and our goal is to recover this frame.

For this experiment, we use the K-SVD algorithm [62] to generate a dictionary $\Psi \in \mathbb{R}^{76800\times 50}$ consisting of 50 atoms that are trained over the first 80 frames.

The number of clients to reconstruct this video frame is 50 and non i.i.d. random matrix of size 30×76800 is used for each client. More specifically, it is generated according to the normal distribution $\mathcal{N}(\mu_i, 1/i^{0.9})$ where $\mu_i \sim \mathcal{N}(0, \alpha = 0.5)$, similar to the one in Sections V-A and V-B.1.

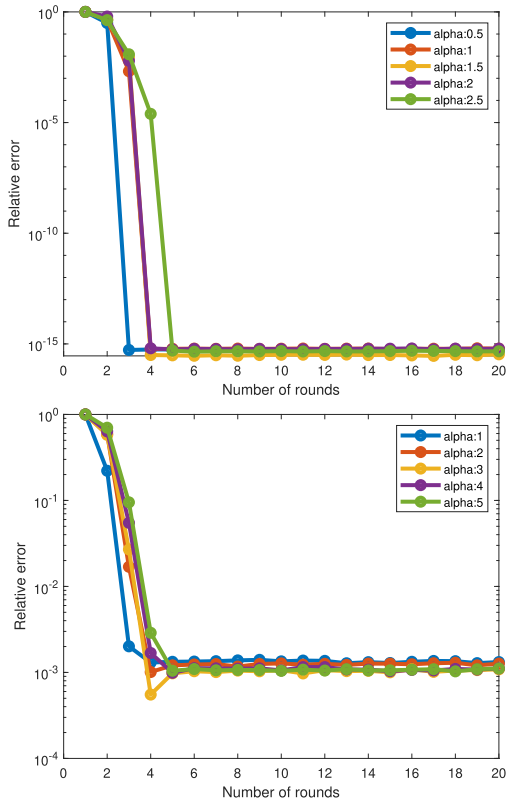

Fig. 1. Linear convergence of FedGradMP with for datasets with various heterogeneity levels.

Figure 2 shows one frame of the input image sequence on the top, the image recovered by FedGradMP + K-SVD in the middle, and the difference on the bottom. Considering that the sensing matrices for clients are highly heterogeneous, the recovered image quality is reasonably satisfactory.

B. Comparison of FedGradMP With Other FL Algorithms

1) Federeated Sparse Linear Regression: The next experiments illustrate FedGradMP outperforms other FL algorithms in both low and highly heterogeneous data environments.

a) Experiment settings: We compare FedGradMP with FedIterHT, FedMid, and FedDualAvg for the sparse linear regression or compressed sensing. To make a fair comparison, we apply FedMid and FedDualAvg with sparsity constraints, making all the methods have the same objective function. In this case, the corresponding regularizer is the characteristic function on the set of τ-sparse vectors and it is easy to check that the proximal steps in FedMid and FedDualAvg in [5] are the gradient decent with the hard-thresholding $\mathcal{H}_\tau(x)$ (projected gradient descent).

In the low-heterogeneity data experiments for Figure 3, the 100×1000 data matrices A_{D_i} are generated by the randomly shifted mean Gaussian model used for the experiments for Figure 1 with whose elements are synthetically generated according to $\mathcal{N}(\mu_i, 1/i^{0.2})$ where $\mu_i \sim \mathcal{N}(0, \alpha = 0.2)$.

On the other hand, under the same setting as before but a higher value of the parameter $\alpha = 0.5$ is used to generate the data matrices A_{D_i} to obtain a more heterogeneous client dataset for the experiment for Figure 4.

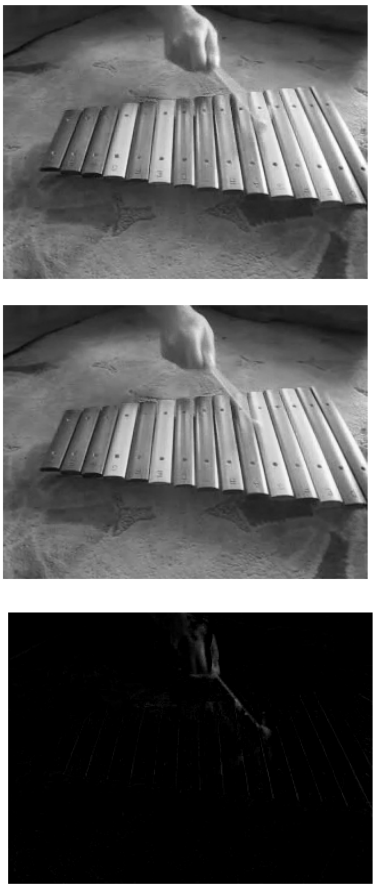

Fig. 2. Input image on the top: 82-th frame of the xylophone video. The output image of FedGradMP with K-SVD dictionary in the middle. The difference between the two images is displayed on the bottom.

The previous two experiments for Figures 3 and 4 are conducted for a signal with sparsity level 15. The relative error curves in Figure 5 are obtained for a signal $x^{\#}$ that 400-sparse under the same heterogeneous model as in Figure 4. Because of the high sparsity level (about the same order as the ambient dimension 1000), we run the Inexact-FedGradMP (Algorithm 2) with gradient descent to solve the sub-optimization problem more efficiently as we have discussed in Section IV.

b) Simulation results: The plots for Figure 3 demonstrate FedGradMP converges faster than other methods in the number of communication rounds for a low heterogeneous environment both in the number of rounds and wall-clock time. FedIterHT converges linearly as shown in [6], but with a slower convergence rate than FedGradMP. We also notice that FedGradMP offers the smallest residual error evidencing our theory that FedGradMP guarantees the optimal statistical bias in Remark 5.

In the highly heterogenous environment setting, FedGradMP still performs well whereas other algorithms start degrading significantly, as we observe in the plots in Figure 4.

As for the signals with higher sparsity levels, from the plots in Figure 5 show, we see that FedGradMP performs better than other baseline algorithms in terms of both criteria.

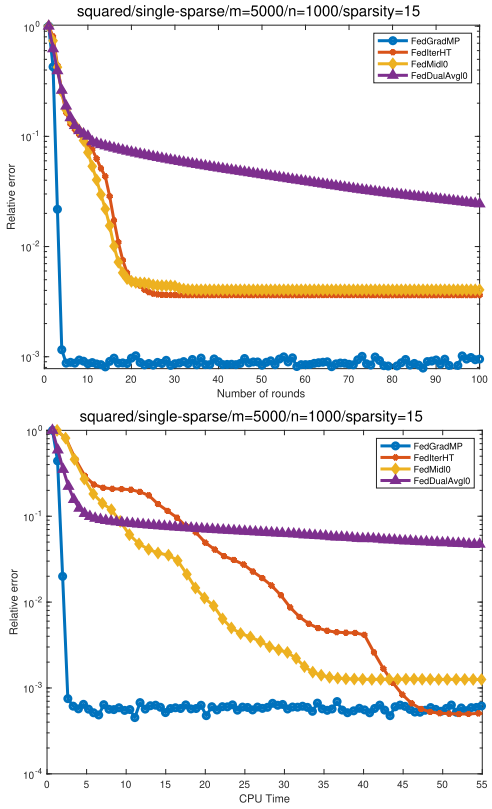

Fig. 3. FedGradMP outperforms other methods in a low data heterogeneous environment.

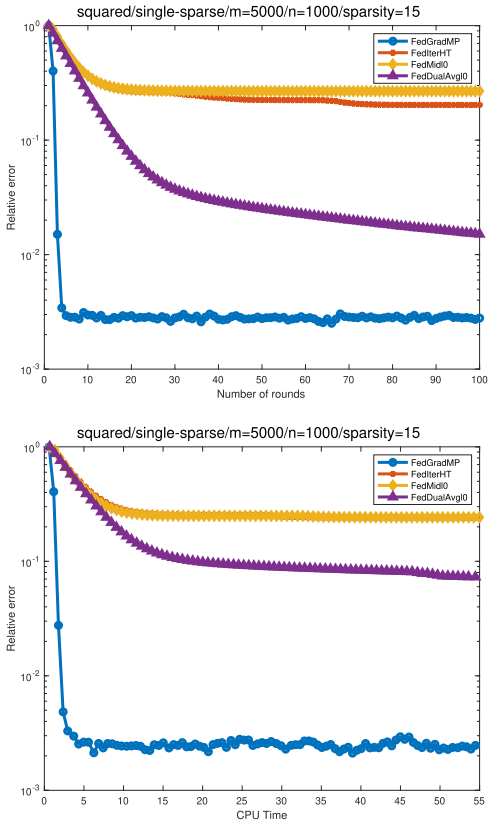

Fig. 4. FedGradMP outperforms other methods in a high data heterogeneous environment.

2) Logistic Regression for Federated EMNIST Dataset:

a) Experiment settings: The dataset we use is the Federated EMNIST-10 dataset (FEMNIST-10), a commonly used

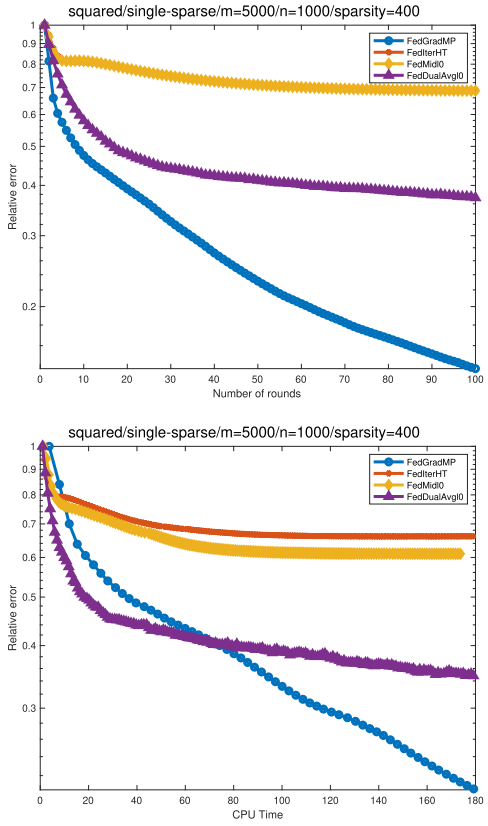

Fig. 5. FedGradMP outperforms other methods in a high data heterogeneous environment for high sparsity level signals.

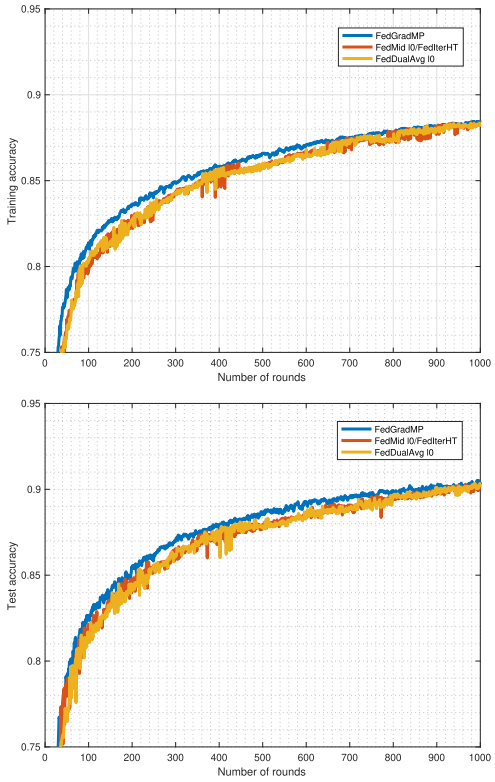

Fig. 6. In both experiments using the training dataset (on the top) and the test dataset (on the bottom), the performance of FedGradMP is better than other baseline methods.

dataset to test FL algorithms. FEMNIST-10 is a collection of handwritten digits and 10 labels, grouped by writers. Each data point of FEMNIST-10 consists of a 28×28 gray-scale image and its label belongs to one of the 10 classes. Note that the dimension of solution space is $28 \times 28 = 784$.

In the experiment, we use 350 clients, which is about 10% of the original dataset with 100 examples each. We split the data into a training dataset with 300 clients and a test dataset with 50 clients. The number of participating clients per round is 10 and the mini-batch size is 50. This is similar to the standard settings used for FL algorithm benchmark [5], [47]. We run the Inexact-FedGradMP with an ℓ_2 norm constraint with 20 local iterations, in which we solve the sub-optimization problem in FedGradMP by SGD with 2 iterations. The number of local iterations for FedIterHT, FedAvg, FedMid, FedDualAvg is 40. Note that the total number of the effective number of local iterations for all the algorithms is the same, 40 iterations. The number of communication rounds is 1000.

The local objective function is $f_i(x^{(1)}, x^{(2)}, \ldots, x^{(N)}) = \frac{1}{|D_i|} \sum_{j=1}^{|D_i|} \left[\sum_{i=1}^{10} -y_{ij} b_j^\top x^{(i)} + \ln \left(\exp \left(\sum_{i=1}^{10} b_j^\top x^{(i)} \right) \right) \right]$, the multi-class logistic regression function with the sparsity constraint $\|x\|_0 \leq 90$ and the ℓ_2 ball constraint $\|x\| \leq 10^5$ throughout all the methods.

b) Simulation results: Figure 6 demonstrates that FedGradMP outperforms the baseline algorithms in terms of prediction accuracy on training and test datasets.

c) Improving FedGradMP performance using random dictionaries: In this section, we show that FedGradMP combined with a random Gaussian dictionary empirically outperforms the one with the standard basis. The experiment settings are the same as the ones in Section V-B.2 except we use the random Gaussian dictionary of size 200×784. As a comparison, we have also included the prediction accuracy curves of FedGradMP in Figure 6.

The plot in Figure 7 indicates that FedGradMP + random Gaussian dictionary outperforms FedGradMP + the standard basis, supporting our theory in Section IV-D.

C. Difficulties of Tuning Learning Rates for FL Methods

As we saw in the numerical experiments, Section V-B.1, other FL methods suffer especially in a highly heterogeneous environment. This can be alleviated by tuning hyperparameters individually for each client such as learning rates, but it could be challenging or at least time-consuming. To showcase the difficulties of tuning the learning rates of FL methods, we study FedIterHT but we empirically observed the same phenomenon for other baseline algorithms.

The convergence of FedIterHT in [6] strongly depends on the learning rates. Although they provide the learning rates that depend on the dissimilarity parameter and restricted strong convexity/smoothness parameters at the clients, they are quite often not available and difficult to estimate in practice since the data at clients are non i.i.d.. FedGradMP is free from this issue at least for sparse linear regression and is often still computationally efficient since clients only solve optimization problems over smaller spaces after the support estimation.

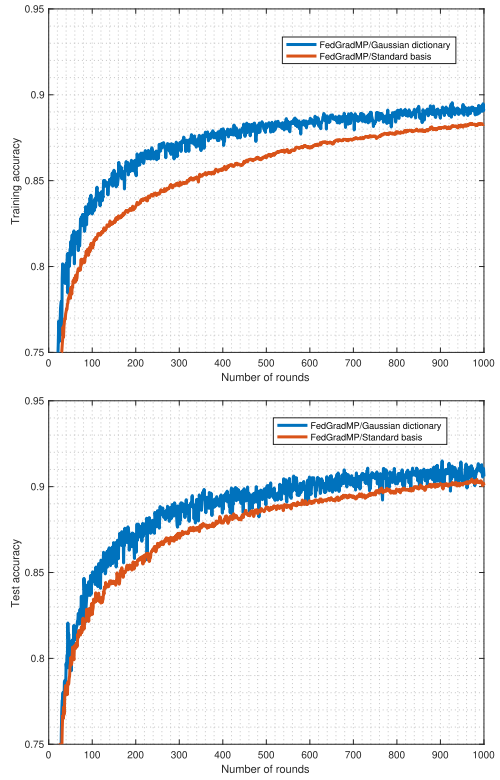

Fig. 7. Training accuracy curves of FedGradMP for the FEMINST dataset with respect to the random Gaussian dictionary and the standard basis.

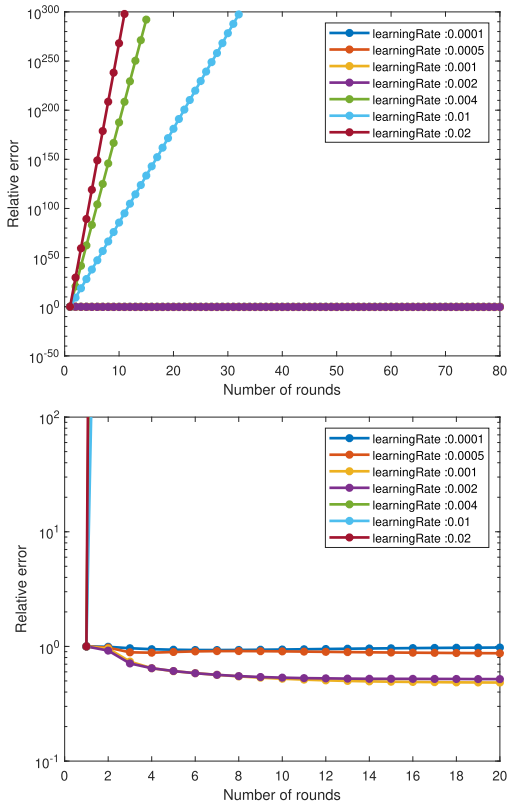

Fig. 8. FedIterHT with learning rates $\{0.0001, 0.0005, 0.001, 0.002, 0.004, 0.01, 0.02\}$ for non i.i.d. datasets.

1) Experiment Settings: We run FedIterHT for the squared loss function with a randomly generated 10-sparse vector as ground truth. The local loss function $f_i = \frac{1}{2\|D_i\|}\|A_{D_i}x - y_{D_i}\|_2^2$ where A_{D_i} is the client i data matrix in $\mathbb{R}^{100\times 1000}$ whose elements are synthetically generated according to $\mathcal{N}(\mu_i, 1/i^{1.1})$ with randomly generated mean μ_i from the mean-zero Gaussian with variance $\alpha = 1.0$. This setting is similar to the synthetic dataset in [6] except we have common sparse ground truth. The number of clients is 30 with mini-batch size 40. The number of total data points $m = 3000$ and the dimension of solution space $n = 1000$. The client learning rate combinations for the experiment are $\{0.0001, 0.0005, 0.001, 0.002, 0.004, 0.01, 0.02\}$.

2) Simulation Results: If the learning rates are chosen from $\{0.004, 0.01, 0.02\}$, then the top plot in Figure 8 shows that they quickly diverge from the optimal solution.

On the other hand, the bottom panel in Figure 8 shows the relative error and squared loss curves for FedIterHT when the learning rate is in $\{0.0001, 0.0005, 0.001, 0.002\}$. For these smaller learning rates, the iterates of FedIterHT tend to converge to a highly suboptimal local solution. It has been observed in the literature [63] that approaches based on stochastic gradient descents combined with hard-thresholding (such as FedIterHT) suffer from such phenomena when the learning rates are chosen to be too small.

Hence, our numerical experiments indicate that the learning rates should be chosen very carefully for each client. Working learning rates should depend on the statistics and heterogeneity of the local dataset at the client. Obtaining this information could be challenging because it might not be available in general, so usually, a grid search is performed to find learning rates.

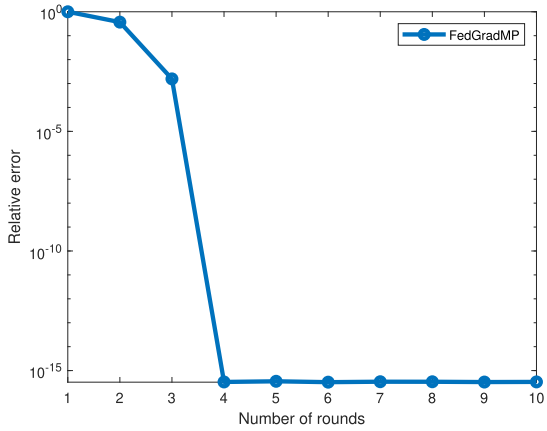

Fig. 9. FedGradMP for non i.i.d. datasets.

On the other hand, the iterates of FedGradMP converge to the ground truth up to (almost) machine precision as shown in Figure 9 under the same setting, only in four rounds with three local iterations at the clients. Unlike FedIterHT, FedGradMP does not require fine tuning of learning rates per client.

D. Impact of the Number of Local Iterations

We provide numerical evidence supporting Theorem 5 about how the number of local iterations at clients affects the convergence rate and the residual error of FedGradMP.

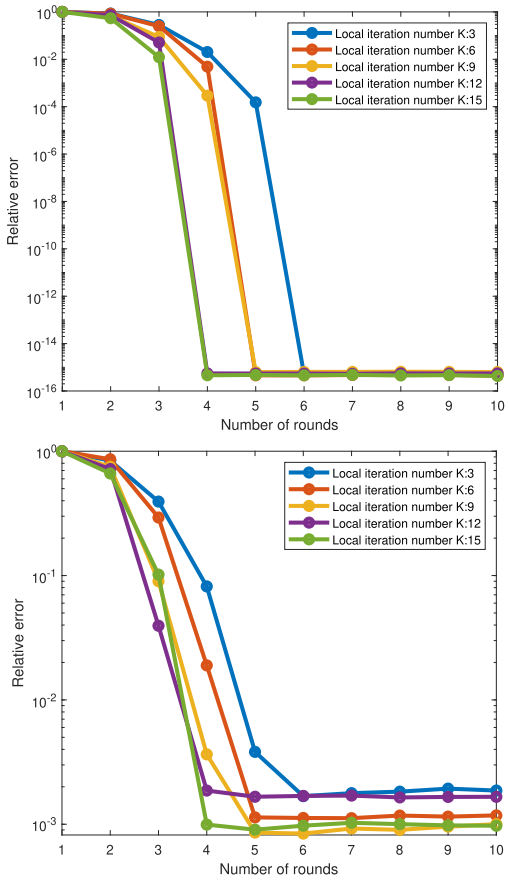

Fig. 10. The convergence rate improves as the local iterations at clients increase in Theorem 5.

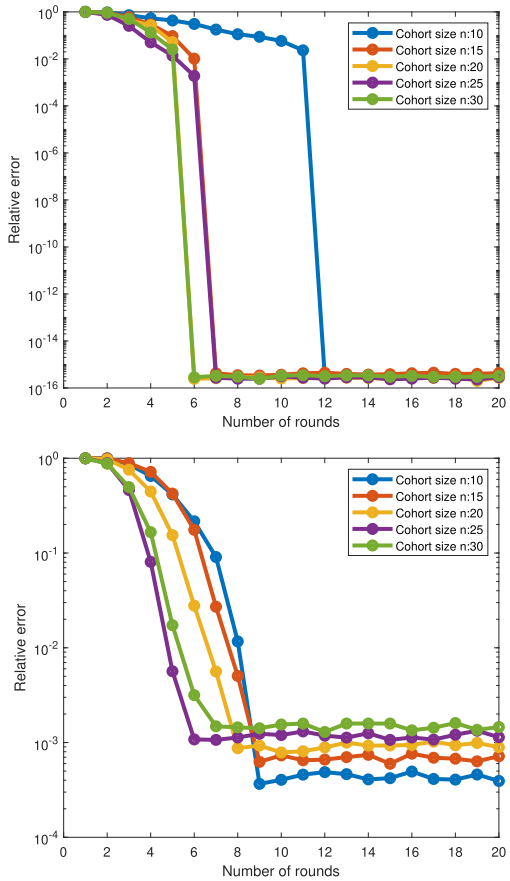

Fig. 11. The convergence rates improve as the cohort size increases as predicted in Theorem 8. Note that the residual errors in the right panel decay to zero (up to the machine precision) since all the non i.i.d. measurements are noiseless with the squared loss function.

1) Experiment Settings: The number of clients is 50, the dimension of solution space is 1000, the number of data points of each client is 100, the mini-batch size of each client is 30, and the cohort size is 50. The local objective function f_i is the squared loss with associated data matrix A_{D_i} for client i, similar to the one used for the heterogeneous case with $\alpha = 2.5$ in Section V-A. We run FedGradMP with local iterations $3, 6, 9, 12, 15$ for noiseless and noisy setup ($y_{D_i} = A_{D_i} x^{\#} + e$, where e is a Gaussian noise where each component are independently generated according to $\mathcal{N}(0, 4 \times 10^{-6})$).

2) Simulation Results: We display the relative error curves of iterates of FedGradMP on the top and bottom panels of Figure 10 for noiseless and noisy case respectively.

The error decay curves in the top plot for the noiseless case demonstrate that as we increase the number of local iterations at clients, FedGradMP converges faster or the convergence rates improve. The plot on the bottom for the noisy case also exhibits a similar pattern but with a few exceptions probably due to the noise. This supports our theory about the dependence of convergence rate κ on the number of local iterations in Theorem 5 as explained in Remark 4.

As for the residual error of FedGradMP, we observe a general trend in the right panel that increasing the local iterations decreases the residual error, but this effect is not as noticeable as the convergence rate. This is somewhat expected since the residual error term in 5 depends on the local iteration numbers in a complicated way as explained in Remark 4.

E. Impact of Cohort Size

The next experiment illustrates how well FedGradMP performs when cohort size (the number of participating clients per round) varies. We notice that Figure 11 provides numerical evidence supporting Theorem 8 about how the cohort size affects the convergence rate and the residual error of FedGradMP.

1) Experiment Settings: The number of clients is 50, the dimension of solution space is 50 and we set the mini-batch size 30. The local objective function f_i is the squared loss with associated non iid data matrix A_{D_i} for client i, similar to the one used for the heterogeneous case with $\alpha = 2.5$ in Section V-A. We run FedGradMP with cohort size $10, 15, 20, 25, 30$ for noiseless and noisy setup.

2) Simulation Results: The relative error curves of iterates of FedGradMP are given on the top panel (noiseless case) and bottom panels (noisy case) of Figure 11. These error plots indicate that the convergence rate improves as we increase the cohort size, for both noiseless and noisy cases as predicted in Theorem 8. On the other hand, a careful reader might have noticed that the residual error actually slightly increases as the cohort size increases. This implies that the dependence of our residual error bound on the cohort size in Theorem 8 is pessimistic and may not capture the true dependence as most of the other works in FL algorithm analysis. For more details, see the discussion and criticism on the gap between

the current theoretical analyses of the impact of cohort size in FL algorithms and their empirical performance [35].

VI. Conclusion

In this paper, we propose a novel federated stochastic gradient matching pursuit algorithm framework and show the linear convergence in expectation under certain assumptions of the objective function, including the dictionary restricted-RSS/RSC conditions and the bounded heterogeneity only at optima assumption. For the sparse linear regression problem, our method does not require learning rate tuning at the client side, which could be challenging for existing baseline algorithms in highly heterogeneous data environments. Numerical experiments on large scale heterogeneous datasets such as EFMINIST and videos have shown the effectiveness of the proposed approach over the state-of-the-art federated learning algorithms. Our analysis reveals the benefits of adopting random dictionaries such as Gaussian random dictionary, which is also confirmed by our numerical experiments.

Appendix A
Proofs

Proof: [Proof of Corollary 6] First, we recall that the global objective function $f(x) = \sum_{i=1}^N p_i f_i(x)$ and $f_i(x) = \frac{1}{M}\sum_{j=1}^M g_{i,j}(x)$. From Assumption 2 on the $\mathscr{A}$-RSS property of $g_{i,j}$ with constant $\rho_\tau^+(i,j)$, we have

$$\|\nabla g_{i,j}(x_1) - \nabla g_{i,j}(x_2)\|_2 \le \rho_\tau^+(i,j)\|x_1 - x_2\|_2$$

for all $x_1, x_2 \in \mathbb{R}^n$ with $\|x_1 - x_2\|_{0,\mathscr{A}} \le \tau$. By Lemma 1, we have

$$\left\langle \nabla g_{i,j}(x_1), x_2 \right\rangle \ge g_{i,j}(x_1 + x_2) - g_{i,j}(x_1) - \frac{\rho_\tau^+(i,j)}{2}\|x_2\|_2^2.$$

Taking average $g_{i,j}$ over j to recover f_i and over i with probability p_i to recover f, the above inequality implies that

$$\langle \nabla f(x_1), x_2 \rangle \ge f(x_1 + x_2) - f(x_1) - \frac{1}{2}\sum_{i=1}^N p_i \bar{\rho}_\tau^{+(i)}\|x_2\|_2^2.$$

Denote $\sum_{i=1}^N p_i \bar{\rho}_\tau^{+(i)}$ by ρ. Setting $x_2 = x_{t+1} - x^*$ and $x_1 = x^*$ in the above inequality yields

$$\begin{aligned}
&f(x_{t+1}) \\
&\le f(x^*) + \langle \nabla f(x^*), x_{t+1} - x^* \rangle + \frac{\rho}{2}\|x_{t+1} - x^*\|_2^2 \\
&\le f(x^*) + \|\nabla f(x^*)\|_2 \|x_{t+1} - x^*\|_2 + \frac{\rho}{2}\|x_{t+1} - x^*\|_2^2 \\
&\le f(x^*) + \frac{1}{2\rho}\|\nabla f(x^*)\|_2^2 + \frac{\rho}{2}\|x_{t+1} - x^*\|_2^2 + \frac{\rho}{2}\|x_{t+1} - x^*\|_2^2 \\
&\le f(x^*) + \frac{1}{2\rho}\|\nabla f(x^*)\|_2^2 + \rho\|x_{t+1} - x^*\|_2^2.
\end{aligned}$$

Here the third inequality follows from the AM-GM inequality. Taking the expectation to the last inequality, we have

$$\mathbb{E} f(x_{t+1}) \le f(x^*) + \frac{1}{2\rho}\|\nabla f(x^*)\|_2^2 + \rho\mathbb{E}\|x_{t+1} - x^*\|_2^2.$$

Finally, we apply Theorem 5 to the above inequality to establish the statement in the corollary. □

Proof: [Proof of Theorem 7] We follow the same arguments used in the first few steps of the proof of Theorem 5 and obtain the following inequality.

$$\mathbb{E}\|x_{t+1} - x^*\|_2^2 \le (2\eta_3^2 + 2)\sum_{i=1}^N p_i \mathbb{E}_{J_K}^{(i)} \left\| x_{t,K+1}^{(i)} - x^* \right\|_2^2.$$

Because we are solving $b_{t,k}^{(i)} = \underset{x}{\mathbf{argmin}} f_i(x)$ for $x \in R(D_{\hat{\Gamma}})$ with an accuracy δ, we have

$$\begin{aligned}
&\sum_{i=1}^N p_i \mathbb{E}_{J_K}^{(i)} \left\| x_{t,K+1}^{(i)} - x^* \right\|_2^2 \\
&\le (1+\eta_2)^2 \sum_{i=1}^N p_i \mathbb{E}_{J_K}^{(i)} \left\| b_{t,K}^{(i)} - x^* \right\|_2^2 \\
&\le (1+\eta_2)^2 \sum_{i=1}^N p_i \left[2\mathbb{E}_{J_K}^{(i)} \left\| b_{t,K}^{(i,\mathrm{opt})} - x^* \right\|_2^2 + 2\mathbb{E}_{J_K}^{(i)} \left\| b_{t,K-1}^{(i,\mathrm{opt})} - b_{t,K}^{(i)} \right\|_2^2 \right] \\
&\le (1+\eta_2)^2 \sum_{i=1}^N p_i \left[2\mathbb{E}_{J_K}^{(i)} \left\| b_{t,K}^{(i,\mathrm{opt})} - x^* \right\|_2^2 + 2\delta^2 \right] \\
&\le 2(1+\eta_2)^2 \sum_{i=1}^N p_i \left[\beta_1(i)\mathbb{E}_{J_K}^{(i)} \|P_{\hat{\Gamma}}^{\perp}(b_{t,K}^{(i)} - x^*)\|_2^2 + \xi_1(i) + \delta^2 \right].
\end{aligned}$$

The rest of the proof is similar to that of Theorem 5. □

Proof: [Proof of Theorem 8] As in the proof of Theorem 5, let $\mathscr{F}^{(t)}$ be the filtration by all the randomness up to the t-th communication round, but in this case, it is all the selected participating clients and the selected mini-batch indices at all these clients up to the t-th round. Let us denote the client subset selected at round t by I_t. Note that I_t is chosen uniformly at random over all possible subsets of cardinality L whose elements belong to $[N]$, so $|I_t| = L$. Again, as we did in the proof of Theorem 5, by abusing the notation slightly, $\mathbb{E}\left[\cdot|\mathscr{F}^{(t)}\right]$ will be denoted $\mathbb{E}_{(I_t)}[\mathbb{E}[\cdot]]$, where $\mathbb{E}_{(I_t)}$ is the expectation taken over the randomly selected participating clients at round t.

We first consider the case for $\eta_1 > 1$. By following the same argument for the first step of the proof for Theorem 5, we have

$$\begin{aligned}
&\mathbb{E}_{(I_t)}\mathbb{E}\|x_{t+1} - x^*\|_2^2 \\
&= \mathbb{E}_{(I_t)}\mathbb{E}\left\| P_{\Lambda_s}\left(\sum_{i\in I_t} \frac{1}{L} x_{t,K+1}^{(i)} \right) - \sum_{i\in I_t} \frac{1}{L} x_{t,K+1}^{(i)} + \sum_{i\in I_t} \frac{1}{L} x_{t,K+1}^{(i)} - x^* \right\|_2^2 \\
&\le 2\mathbb{E}_{(I_t)}\mathbb{E}\left\| P_{\Lambda_s}\left(\sum_{i\in I_t} \frac{1}{L} x_{t,K+1}^{(i)} \right) - \sum_{i\in I_t} \frac{1}{L} x_{t,K+1}^{(i)} \right\|_2^2 \\
&\quad + 2\mathbb{E}_{(I_t)}\mathbb{E}\left\| \sum_{i\in I_t} \frac{1}{L} x_{t,K+1}^{(i)} - x^* \right\|_2^2 \\
&= (2\eta_3^2 + 2)\mathbb{E}_{(I_t)}\mathbb{E}\left\| \sum_{i\in I_t} \frac{1}{L} x_{t,K+1}^{(i)} - \sum_{i\in I_t} \frac{1}{L} x^* \right\|_2^2 \\
&\le (2\eta_3^2 + 2)\mathbb{E}_{(I_t)}\left[\sum_{i\in I_t} \frac{1}{L}\mathbb{E}\left\| x_{t,K+1}^{(i)} - x^* \right\|_2^2 \right] \qquad (23) \\
&\le (2\eta_3^2 + 2)\mathbb{E}_{(I_t)}\left[\sum_{i\in I_t} \frac{1}{L}\mathbb{E}_{J_K}^{(i)}\left\| x_{t,K+1}^{(i)} - x^* \right\|_2^2 \right]. \qquad (24)
\end{aligned}$$

Moreover, the argument used in the proof of Theorem 5 yields

$$\begin{aligned}
&\sum_{i\in I_t} \frac{1}{L}\mathbb{E}_{J_K}^{(i)}\left\| x_{t,K+1}^{(i)} - x^* \right\|_2^2 \\
&\le (1+\eta_2)^2 \sum_{i\in I_t} \frac{1}{L}\beta_1(i)\beta_2(i)\mathbb{E}_{J_{K-1}}^{(i)}\|x_{t,K}^{(i)} - x^*\|_2^2
\end{aligned}$$

$$+(1+\eta_2)^2\max_i\left(\frac{8\beta_1(i)}{(\rho_{4\tau}^-(i))^2}+\frac{4}{\bar\rho_{4\tau}^{+(i)}(2\rho_{4\tau}^-(i)-\bar\rho_{4\tau}^{+(i)})}+\frac{\beta_1(i)}{\rho_{4\tau}^-(i)}\frac{6\sqrt{\eta_1^2-1}}{\eta_1}\right)\zeta_*^2$$
$$+(1+\eta_2)^2\sum_{i\in I_t}\frac{1}{L}\left[\frac{\beta_1(i)}{\rho_{4\tau}^-(i)}\left(2\theta^2+\frac{6\sqrt{\eta_1^2-1}}{\eta_1}\right)+\frac{4}{\bar\rho_{4\tau}^{+(i)}(2\rho_{4\tau}^-(i)-\bar\rho_{4\tau}^{+(i)})}\right]\sigma_i^2.$$

We define $\nu(I_t)$ that depends on the random index set I_t as follows:

$$\begin{aligned}&\nu(I_t)\\&=(1+\eta_2)^2\max_i\left(\frac{8\beta_1(i)}{(\rho_{4\tau}^-(i))^2}+\frac{4}{\bar\rho_{4\tau}^{+(i)}(2\rho_{4\tau}^-(i)-\bar\rho_{4\tau}^{+(i)})}+\frac{\beta_1(i)}{\rho_{4\tau}^-(i)}\frac{6\sqrt{\eta_1^2-1}}{\eta_1}\right)\zeta_*^2\\&\quad+(1+\eta_2)^2\sum_{i\in I_t}\frac{1}{L}\left[\frac{\beta_1(i)}{\rho_{4\tau}^-(i)}\left(2\theta^2+\frac{6\sqrt{\eta_1^2-1}}{\eta_1}\right)+\frac{4}{\bar\rho_{4\tau}^{+(i)}(2\rho_{4\tau}^-(i)-\bar\rho_{4\tau}^{+(i)})}\right]\sigma_i^2.\end{aligned}$$

After rewriting the previous inequality, we obtain

$$\sum_{i\in I_t}\frac{1}{L}\mathbb{E}_{J_K}^{(i)}\left\|x_{t,K+1}^{(i)}-x^*\right\|_2^2\le\sum_{i\in I_t}\frac{1}{L}\mu(i)\mathbb{E}_{J_{K-1}}^{(i)}\|x_{t,K}^{(i)}-x^*\|_2^2+\nu(I_t).\tag{25}$$

Hence, by the induction on K and using the fact that the cohort set I_t is fixed while the local iterations are running, we obtain a similar upper bound on $\mathbb{E}\|x_{t+1}-x^*\|_2^2$ as follows.

$$\begin{aligned}&\mathbb{E}\|x_{t+1}-x^*\|_2^2\\&\le(2\eta_3^2+2)\mathbb{E}_{(I_t)}\sum_{i\in I_t}\frac{1}{L}\left(\mu(i)^K\left[\mathbb{E}^{(i)}\|x_{t,1}^{(i)}-x^*\|_2^2\right]+\frac{\nu(I_t)(1-\mu(i)^K)}{1-\mu(i)}\right)\\&=(2\eta_3^2+2)\mathbb{E}_{(I_t)}\left(\left(\sum_{i\in I_t}\frac{1}{L}\mu(i)^K\right)\mathbb{E}\|x_t-x^*\|_2^2+\nu(I_t)\sum_{i\in I_t}\frac{1}{L}\frac{(1-\mu(i)^K)}{1-\mu(i)}\right)\\&\le(2\eta_3^2+2)\mathbb{E}_{(I_t)}\left(\left(\sum_{i\in I_t}\frac{1}{L}\mu(i)^K\right)\mathbb{E}\|x_t-x^*\|_2^2+\nu(I_t)\frac{(1-\mu^K)}{1-\mu}\right).\end{aligned}$$

Recall that the index set I_t is a subset of $[N]$, uniformly selected at random, for the communication round t. By taking the maximum of $\sum_{i\in I_t}\frac{1}{L}\mu(i)^K$ over all possible subsets, we have

$$\begin{aligned}&\mathbb{E}\|x_{t+1}-x^*\|_2^2\\&\le\kappa\mathbb{E}\|x_t-x^*\|_2^2+\frac{(2\eta_3^2+2)(1-\mu^K)}{1-\mu}\mathbb{E}_{(I_t)}[\nu(I_t)]\\&\le\kappa\mathbb{E}\|x_t-x^*\|_2^2+\frac{(2\eta_3^2+2)\tilde\nu(1-\mu^K)}{1-\mu}\end{aligned}$$

where

$$\kappa=(2\eta_3^2+2)\max_{\substack{S\subset[N]\\|S|=L}}\frac{1}{L}\sum_{z\in S}\left[(1+\eta_2)^2\beta_1(z)\beta_2(z)\right]^K,$$

and

$$\tilde\nu=(1+\eta_2)^2\max_i\left(\frac{8\beta_1(i)}{(\rho_{4\tau}^-(i))^2}+\frac{4}{\bar\rho_{4\tau}^{+(i)}(2\rho_{4\tau}^-(i)-\bar\rho_{4\tau}^{+(i)})}+\frac{\beta_1(i)}{\rho_{4\tau}^-(i)}\frac{6\sqrt{\eta_1^2-1}}{\eta_1}\right)\zeta_*^2$$
$$+(1+\eta_2)^2\frac{1}{L}\sum_{i=1}^N\left[\frac{\beta_1(i)}{\rho_{4\tau}^-(i)}\left(2\theta^2+\frac{6\sqrt{\eta_1^2-1}}{\eta_1}\right)+\frac{4}{\bar\rho_{4\tau}^{+(i)}(2\rho_{4\tau}^-(i)-\bar\rho_{4\tau}^{+(i)})}\right]\sigma_i^2.$$

Hence, by the induction on t, we have

$$\mathbb{E}\|x_{t+1}-x^*\|_2^2\le\kappa^{t+1}\mathbb{E}\|x_0-x^*\|_2^2+\frac{(2\eta_3^2+2)\tilde\nu(1-\mu^K)}{(1-\kappa)(1-\mu)}.$$

The case for $\eta_1=1$ follows from a similar argument. □

APPENDIX B
FEDGRADMP CONVERGENCE WITHOUT THE BOUNDED VARIANCE CONDITION OF STOCHASTIC GRADIENTS

We start with the following lemma replacing the bounded variance condition of stochastic gradients (7) in Assumption 4 only under the $\mathscr{A}$-RSS condition.

Lemma 11: Let $\mathbb{E}_j$ be the expectation over the uniform distribution on all possible mini-batches. Then, for all τ-sparse vectors x, we have

$$\begin{aligned}&\mathbb{E}_j\|\nabla g_{i,j}(x)-\nabla f_i(x)\|_2^2\\&\le3\mathbb{E}_j((\rho_\tau^+(i,j))^2+\bar\rho_{4\tau}^{+(i)})\|\Delta\|_2^2+12\mathbb{E}_j\|\nabla g_{i,j}(x^*)\|_2^2\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}_j\|P_\Gamma(\nabla g_{i,j}(x)-\nabla f_i(x))\|_2^2\\&\le3\mathbb{E}_j((\rho_\tau^+(i,j))^2+\bar\rho_{4\tau}^{+(i)})\|\Delta\|_2^2+12\mathbb{E}_j\|P_\Gamma\nabla g_{i,j}(x^*)\|_2^2,\end{aligned}$$

where x^* is a solution to (1) and $\Delta=x-x^*$.

Proof: [Proof of Lemma]

$$\begin{aligned}&\mathbb{E}_j\|\nabla g_{i,j}(x)-\nabla f_i(x)\|_2^2\\&\le3\mathbb{E}_j\|\nabla g_{i,j}(x)-\nabla g_{i,j}(x^*)\|_2^2+3\mathbb{E}_j\|\nabla g_{i,j}(x^*)-\nabla f_i(x^*)\|_2^2\\&\quad+3\mathbb{E}_j\|\nabla f_i(x^*)-\nabla f_i(x)\|_2^2\\&\le3\mathbb{E}_j(\rho_\tau^+(i,j))^2\|x-x^*\|_2^2+6\mathbb{E}_j\|\nabla g_{i,j}(x^*)\|_2^2+6\|\nabla f_i(x^*)\|_2^2\\&\quad+3\mathbb{E}_j\bar\rho_{4\tau}^{+(i)}\|x-x^*\|_2^2\\&=3\mathbb{E}_j(\rho_\tau^+(i,j))^2\|\Delta\|_2^2+6\mathbb{E}_j\|\nabla g_{i,j}(x^*)\|_2^2+6\|\nabla f_i(x^*)\|_2^2\\&\quad+3\mathbb{E}_j\bar\rho_{4\tau}^{+(i)}\|\Delta\|_2^2\\&\le3\mathbb{E}_j((\rho_\tau^+(i,j))^2+\bar\rho_{4\tau}^{+(i)})\|\Delta\|_2^2+6\mathbb{E}_j\|\nabla g_{i,j}(x^*)\|_2^2\\&\quad+6\|\nabla f_i(x^*)\|_2^2\\&\le3\mathbb{E}_j((\rho_\tau^+(i,j))^2+\bar\rho_{4\tau}^{+(i)})\|\Delta\|_2^2+12\mathbb{E}_j\|\nabla g_{i,j}(x^*)\|_2^2.\end{aligned}$$

The second inequality follows from the $\mathscr{A}$-RSS condition for $\nabla g_{i,j}$ with constant $\rho_\tau^+(i,j)$, and the fact that ∇f_i is the average of $\nabla g_{i,j}$. The last inequality is from the Jensen's inequality. This proves the first part of the lemma and the second part follows from a similar argument. □

This lemma allows us to prove a similar statement as in Lemma 4 without the bounded variance condition (7). Since the underlying argument of the proof of the following lemma is the same, we only point out the difference from the proof for Lemma 4.

Lemma 12: Let $\widehat{\Gamma}$ be the set obtained from the k-th iteration at client i. Then, for any $\theta > 0$, we have

$$\mathbb{E}_{j_k}^{(i)}\|P_{\widehat{\Gamma}}^{\perp}(b_{t,k}^{(i)}-x^*)\|_2^2 \le \beta_2(i)\|x_{t,k}^{(i)}-x^*\|_2^2+\xi_2(i),$$

where

$$\beta_2(i)=\left(4\frac{(2\eta_1^2-1)\left(\bar{\rho}_{4\tau}^{+(i)}+\frac{1}{\theta^2}\right)-\eta_1^2\rho_{4\tau}^{-}(i)}{\eta_1^2\rho_{4\tau}^{-}(i)} + \left(\frac{3\theta^2\mathbb{E}_{j_k}(\rho_\tau^+(i,j_k)+\bar{\rho}_{4\tau}^{+(i)})}{\rho_{4\tau}^{-}(i)}\right)+\frac{2(\eta_1^2-1)}{\eta_1^2}\right)$$

$$\xi_2(i)=\frac{8}{(\rho_{4\tau}^{-}(i))^2}\max_{\substack{\Omega\subset[d]\\|\Omega|=4\tau}}\|P_\Omega\nabla f_i(x^*)\|_2^2 + 2\left(6\theta^2+\frac{15\sqrt{\eta_1^2-1}}{2\eta_1}\right)\mathbb{E}_{j_k}\|\nabla g_{i,j_k}(x^*)\|_2^2.$$

Note that if $\eta_1=1$, then the projection operator is exact. Here $\mathbb{E}_{j_k}^{(i)}$ is the expectation taken over the randomly selected index j_k at the k-th step of the local iterations of the i-th client.

Proof: We follow the same steps in the proof of Lemma 4 for the bound $f_i(x^*)-f_i\left(x_{t,k}^{(i)}\right)-\frac{\rho_{4\tau}^{-}(i)}{2}\|x^*-x_{t,k}^{(i)}\|_2^2$ but apply Lemma 12 to the inequality 26 as follows.

$$\begin{aligned}
&f_i(x^*)-f_i\left(x_{t,k}^{(i)}\right)-\frac{\rho_{4\tau}^{-}(i)}{2}\|x^*-x_{t,k}^{(i)}\|_2^2\\
&\ge \mathbb{E}_{j_k}\left\langle\nabla f_i\left(x_{t,k}^{(i)}\right),z\right\rangle-\frac{\theta^2}{2}\|\nabla g_{i,j_k}(x_{t,k}^{(i)})-\nabla f_i\left(x_{t,k}^{(i)}\right)\|_2^2\\
&\quad-\frac{1}{2\theta^2}\mathbb{E}_{j_k}\|z\|_2^2-\frac{\bar{\rho}_{4\tau}^{+(i)}}{2}\mathbb{E}_{j_k}\|z\|_2^2\\
&\quad-\frac{\sqrt{\eta_1^2-1}}{2\eta_1}\left(\mathbb{E}_{j_k}\|P_\Gamma^{\perp}\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2^2+\mathbb{E}_{j_k}\|\Delta\|_2^2\right)\\
&\ge \mathbb{E}_{j_k}\left\langle\nabla f_i\left(x_{t,k}^{(i)}\right),z\right\rangle-\frac{\theta^2}{2}(3\mathbb{E}_j((\rho_\tau^+(i,j))^2+\bar{\rho}_{4\tau}^{+(i)})\|\Delta\|_2^2\\
&\qquad+12\mathbb{E}_j\|\nabla g_{i,j}(x^*)\|_2^2)\\
&\quad-\frac{1}{2\theta^2}\mathbb{E}_{j_k}\|z\|_2^2-\frac{\bar{\rho}_{4\tau}^{+(i)}}{2}\mathbb{E}_{j_k}\|z\|_2^2\\
&\quad-\frac{\sqrt{\eta_1^2-1}}{2\eta_1}\left(\mathbb{E}_{j_k}\|P_\Gamma^{\perp}\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2^2+\mathbb{E}_{j_k}\|\Delta\|_2^2\right).
\end{aligned}\tag{26}$$

Similarly, we obtain the upper bound for $\mathbb{E}_{j_k}\|P_\Gamma^{\perp}\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2^2$ as follows.

$$\begin{aligned}
&\mathbb{E}_{j_k}\|P_\Gamma^{\perp}\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2^2\\
&\le \mathbb{E}_{j_k}\|\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2^2\\
&\le 3\mathbb{E}_{j_k}\|\nabla g_{i,j_k}(x_{t,k}^{(i)})-\nabla g_{i,j_k}(x^*)\|_2^2\\
&\qquad+3\mathbb{E}_{j_k}\|\nabla g_{i,j_k}(x^*)-\nabla f_i(x^*)\|_2^2+3\mathbb{E}_{j_k}\|\nabla f_i(x^*)\|_2^2\\
&\le 3\mathbb{E}_{j_k}(\rho_\tau^+(i,j_k))^2\|x_{t,k}^{(i)}-x^*\|_2^2\\
&\qquad+6\mathbb{E}_{j_k}\|\nabla g_{i,j_k}(x^*)\|_2^2+6\|\nabla f_i(x^*)\|_2^2+3\|\nabla f_i(x^*)\|_2^2\\
&= 3\mathbb{E}_{j_k}(\rho_\tau^+(i,j_k))^2\|\Delta\|_2^2+6\mathbb{E}_{j_k}\|\nabla g_{i,j_k}(x^*)\|_2^2+9\|\nabla f_i(x^*)\|_2^2\\
&= 3\mathbb{E}_{j_k}(\rho_\tau^+(i,j_k))^2\|\Delta\|_2^2+15\mathbb{E}_{j_k}\|\nabla g_{i,j_k}(x^*)\|_2^2,
\end{aligned}$$

where the last inequality is from the Jensen's inequality.

Applying this bound for $\mathbb{E}_{j_k}\|P_\Gamma^{\perp}\nabla g_{i,j_k}(x_{t,k}^{(i)})\|_2^2$ yields

$$\begin{aligned}
&f_i(x^*)-f_i\left(x_{t,k}^{(i)}\right)-\frac{\rho_{4\tau}^{-}(i)}{2}\|\Delta\|_2^2\\
&\ge \mathbb{E}_{j_k}\left\langle\nabla f_i\left(x_{t,k}^{(i)}\right),z\right\rangle-\frac{\theta^2}{2}(3\mathbb{E}_j((\rho_\tau^+(i,j))^2+\bar{\rho}_{4\tau}^{+(i)})\|\Delta\|_2^2\\
&\qquad+12\mathbb{E}_j\|\nabla g_{i,j}(x^*)\|_2^2)-\frac{\bar{\rho}_{4\tau}^{+(i)}}{2}\mathbb{E}_{j_k}\|z\|_2^2-\frac{1}{2\theta^2}\mathbb{E}_{j_k}\|z\|_2^2\\
&\qquad-\frac{\sqrt{\eta_1^2-1}}{2\eta_1}\Big(3\mathbb{E}_{j_k}(\rho_\tau^+(i,j_k))^2\|\Delta\|_2^2\\
&\qquad\qquad+15\mathbb{E}_{j_k}\|\nabla g_{i,j_k}(x^*)\|_2^2+\mathbb{E}_{j_k}\|\Delta\|_2^2\Big).
\end{aligned}$$

Following the same argument in the proof of Lemma 4, we have

$$\begin{aligned}
&\left(\frac{\bar{\rho}_{4\tau}^{+(i)}}{2}+\frac{1}{2\theta^2}\right)\mathbb{E}_{j_k}\|z\|_2^2\\
&\quad-\frac{1}{2}\left(\rho_{4\tau}^{-}(i)-3\theta^2\mathbb{E}_{j_k}((\rho_\tau^+(i,j_k))^2+\bar{\rho}_{4\tau}^{+(i)})\right.\\
&\qquad\left.-\frac{\sqrt{\eta_1^2-1}}{\eta_1}(3\mathbb{E}_{j_k}\rho_\tau^+(i,j_k))^2+1)\right)\|\Delta\|_2^2\\
&\quad+\left(6\theta^2+\frac{15\sqrt{\eta_1^2-1}}{2\eta_1}\right)\mathbb{E}_{j_k}\|\nabla g_{i,j_k}(x^*)\|_2^2\\
&\ge \mathbb{E}_{j_k}f_i\left(x_{t,k}^{(i)}+z\right)-f_i(x^*)\\
&\ge \frac{\rho_{4\tau}^{-}(i)}{2}\|\Delta-z\|_2^2-\max_{\substack{\Omega\subset[d]\\|\Omega|=4\tau}}\|P_\Omega\nabla f_i(x^*)\|_2\mathbb{E}_{j_k}\|\Delta-z\|_2.
\end{aligned}$$

Let $u=\mathbb{E}_{j_k}\|\Delta-y\|_2$, $a=\rho_{4\tau}^{-}(i)$, $b=\max_{\substack{\Omega\subset[d]\\|\Omega|=4\tau}}\|P_\Omega\nabla f_i(x^*)\|_2$, and

$$\begin{aligned}
c=&\left(\frac{\bar{\rho}_{4\tau}^{+(i)}}{2}+\frac{1}{2\theta^2}\right)\mathbb{E}_{j_k}\|z\|_2^2\\
&-\frac{1}{2}\left(\rho_{4\tau}^{-}(i)-3\theta^2\mathbb{E}_{j_k}((\rho_\tau^+(i,j_k))^2+\bar{\rho}_{4\tau}^{+(i)})\right.\\
&\qquad\left.-\frac{\sqrt{\eta_1^2-1}}{\eta_1}(3\mathbb{E}_{j_k}\rho_\tau^+(i,j_k))^2+1)\right)\|\Delta\|_2^2\\
&+\left(6\theta^2+\frac{15\sqrt{\eta_1^2-1}}{2\eta_1}\right)\mathbb{E}_{j_k}\|\nabla g_{i,j_k}(x^*)\|_2^2.
\end{aligned}$$

Then above inequality can be rewritten in $au^2-2bu-c\le 0$ and solving it gives

$$\mathbb{E}_{j_k}\|\Delta-y\|_2\le\sqrt{\frac{c}{a}}+\frac{2b}{a}.$$

Again, following the same argument for the proof of Lemma 4, we get

$$\mathbb{E}_{j_k}^{(i)}\|\Delta-P_\Gamma\Delta\|_2^2\le\frac{2c}{a}+\frac{8b^2}{a^2}.$$

Thus,

$$\mathbb{E}_{j_k}^{(i)}\|\Delta-P_\Gamma\Delta\|_2^2$$

$$\beta_1(i)=\frac{\bar{\rho}_{4\tau}^{+(i)}}{2\rho_{4\tau}^{-}(i)-\bar{\rho}_{4\tau}^{+(i)}},$$

$$\beta_2(i)=\left(2\frac{\left(\bar{\rho}_{4\tau}^{+(i)}+\frac{1}{\theta^2}\right)-\eta_1^2\rho_{4\tau}^{-}(i)}{\eta_1^2\rho_{4\tau}^{-}(i)}+\left(\frac{3\theta^2\mathbb{E}_{j_k}((\rho_\tau^+(i,j_k))^2+\bar{\rho}_{4\tau}^{+(i)})}{\rho_{4\tau}^{-}(i)}\right)+\frac{\sqrt{\eta_1^2-1}}{\eta_1}(3\mathbb{E}_{j_k}\rho_\tau^+(i,j_k))^2+1)\right),$$

$$\nu=(1+\eta_2)^2\max_i\left(\frac{8\beta_1(i)}{(\rho_{4\tau}^{-}(i))^2}\right)\sum_{i=1}^{N}p_i\max_{\substack{\Omega\subset[d]\\|\Omega|=4\tau}}\|P_\Omega\nabla f_i(x^*)\|_2^2$$
$$+(1+\eta_2)^2\sum_{i=1}^{N}p_i\left[\frac{\beta_1(i)}{\rho_{4\tau}^{-}(i)}\left(2\theta^2+\frac{6\sqrt{\eta_1^2-1}}{\eta_1}\right)+\frac{4}{\bar{\rho}_{4\tau}^{+(i)}(2\rho_{4\tau}^{-}(i)-\bar{\rho}_{4\tau}^{+(i)})}\right]\sum_{i=1}^{N}p_i\mathbb{E}_j\|\nabla g_{i,j}(x^*)\|_2^2.$$

$$\le\left(2\frac{\left(\bar{\rho}_{4\tau}^{+(i)}+\frac{1}{\theta^2}\right)-\eta_1^2\rho_{4\tau}^{-}(i)}{\eta_1^2\rho_{4\tau}^{-}(i)}+\left(\frac{3\theta^2\mathbb{E}_{j_k}((\rho_\tau^+(i,j_k))^2+\bar{\rho}_{4\tau}^{+(i)})}{\rho_{4\tau}^{-}(i)}\right)\right.$$
$$\left.+\frac{\sqrt{\eta_1^2-1}}{\eta_1}(3\mathbb{E}_{j_k}\rho_\tau^+(i,j_k))^2+1)\right)\|\Delta\|_2^2$$
$$+\frac{8}{(\rho_{4\tau}^{-}(i))^2}\max_{\substack{\Omega\subset[d]\\|\Omega|=4\tau}}\|P_\Omega\nabla f_i(x^*)\|_2^2$$
$$+2\left(6\theta^2+\frac{15\sqrt{\eta_1^2-1}}{2\eta_1}\right)\mathbb{E}_{j_k}\|\nabla g_{i,j_k}(x^*)\|_2^2.$$

□

We follow the idea of the proof for Theorem 5 but use Lemma 12 to show the following convergence theorem of FedGradMP.

Theorem 13: Under the same notations and assumptions but without the bounded variance condition (7), the expectation of the recovery error at the $(t+1)$-th round of FedGradMP described in Algorithm 1 is upper bounded by

$$\mathbb{E}\|x_{t+1}-x^*\|_2^2\le\kappa^{t+1}\|x_0-x^*\|_2^2+\frac{(2\eta_3^2+2)\nu}{1-\kappa}\sum_{i=1}^{N}p_i\frac{1-\mu(i)^K}{1-\mu(i)},$$

where

$$\kappa=(2\eta_3^2+2)\sum_{i=1}^{N}p_i\left[(1+\eta_2)^2\beta_1(i)\beta_2(i)\right]^K.$$

Here, shown in the equation at the top of the page.

REFERENCES

[1] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.

[2] J. Verbraeken, "A survey on distributed machine learning," *ACM Comput. Surv.*, vol. 53, no. 2, pp. 1–33, 2020.

[3] B. McMahan and D. Ramage, "Federated learning: Collaborative machine learning without centralized training data," *Google AI Blog*, Apr. 2017. [Online]. Available: http://research.google/blog/federated-learning-collaborative-machine-learning-without-centralized-training-data/

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.

[5] H. Yuan, M. Zaheer, and S. Reddi, "Federated composite optimization," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 12253–12266.

[6] Q. Tong, G. Liang, T. Zhu, and J. Bi, "Federated nonconvex sparse learning," 2021, *arXiv:2101.00052*.

[7] P. Zhou, X. Yuan, and J. Feng, "Efficient stochastic gradient hard thresholding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1988–1997.

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[9] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301–321, May 2009.

[10] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 265–274, 2009.

[11] H. Rauhut, R. Schneider, and Ž. Stojanac, "Low rank tensor recovery via iterative hard thresholding," *Linear Algebra Appl.*, vol. 523, pp. 220–262, Jun. 2017.

[12] R. Grotheer, S. Li, A. Ma, D. Needell, and J. Qin, "Iterative hard thresholding for low CP-rank tensor models," *Linear Multilinear Algebra*, vol. 70, no. 22, pp. 7452–7468, Dec. 2022.

[13] S. Foucart and H. Rauhut, "An invitation to compressive sensing," in *A Mathematical Introduction to Compressive Sensing*. Cham, Switzerland: Springer, 2013, pp. 1–39.

[14] N. Nguyen, D. Needell, and T. Woolf, "Linear convergence of stochastic iterative greedy algorithms with sparse constraints," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 6869–6895, Nov. 2017.

[15] J. Shen and P. Li, "A tight bound of hard thresholding," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 7650–7691, 2017.

[16] J. Wang et al., "A field guide to federated optimization," 2021, *arXiv:2107.06917*.

[17] A. Khaled, K. Mishchenko, and P. Richtárik, "Tighter theory for local SGD on identical and heterogeneous data," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2020, pp. 4519–4529.

[18] B. E. Woodworth, K. K. Patel, and N. Srebro, "Minibatch vs local SGD for heterogeneous distributed learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6281–6292.

[19] F. Haddadpour and M. Mahdavi, "On the convergence of local descent methods in federated learning," 2019, *arXiv:1910.14425*.

[20] M. A. Davenport, D. Needell, and M. B. Wakin, "Signal space CoSaMP for sparse recovery with redundant dictionaries," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6820–6829, Oct. 2013.

[21] R. Baraniuk, S. Foucart, D. Needell, Y. Plan, and M. Wootters, "One-bit compressive sensing of dictionary-sparse signals," *Inf. Inference, A J. IMA*, vol. 7, no. 1, pp. 83–104, Mar. 2018.

[22] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," 2016, *arXiv:1610.02527*.

[23] L. O. Mangasarian, "Parallel gradient distribution in unconstrained optimization," *SIAM J. Control Optim.*, vol. 33, no. 6, pp. 1916–1925, Nov. 1995.

[24] F. Haddadpour, M. M. Kamani, M. Mahdavi, and V. Cadambe, "Local SGD with periodic averaging: Tighter analysis and adaptive synchronization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 11082–11094.

[25] T.-M. Harry Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," 2019, *arXiv:1909.06335*.

[26] J. Wang, Q. Liu, H. Liang, G. Joshi, and V. H. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in *Proc. NIPS*, 2020, pp. 7611–7623.

[27] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5132–5143.

[28] Z. Charles and J. Konečnỳ, "Convergence and accuracy trade-offs in federated learning and meta-learning," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 2575–2583.

[29] D. Basu, D. Data, C. Karakus, and S. Diggavi, "Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 14695–14706.

[30] D. Rothchild et al., "FetchSGD: Communication-efficient federated learning with sketching," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 8253–8265.

[31] A. Mitra, R. Jaafar, G. J. Pappas, and H. Hassani, "Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 14606–14619.

[32] F. Haddadpour, M. M. Kamani, A. Mokhtari, and M. Mahdavi, "Federated learning with compression: Unified analysis and sharp guarantees," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2021, pp. 2350–2358.

[33] Z. Song, Y. Wang, Z. Yu, and L. Zhang, "Sketching for first order method: Efficient algorithm for low-bandwidth channel and vulnerability," 2022, *arXiv:2210.08371*.

[34] H. Yang, M. Fang, and J. Liu, "Achieving linear speedup with partial worker participation in non-IID federated learning," 2021, *arXiv:2101.11203*.

[35] Z. Charles, Z. Garrett, Z. Huo, S. Shmulyian, and V. Smith, "On large-cohort training for federated learning," in *Proc. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 20461–20475.

[36] W. Chen, S. Horvath, and P. Richtarik, "Optimal client sampling for federated learning," 2020, *arXiv:2010.13723*.

[37] H. Yang, X. Zhang, P. Khanduri, and J. Liu, "Anarchic federated learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2022, pp. 25331–25363.

[38] A. Koloskova, S. U. Stich, and M. Jaggi, "Sharper convergence guarantees for asynchronous SGD for distributed and federated learning," 2022, *arXiv:2206.08307*.

[39] J. Qin et al., "Stochastic greedy algorithms for multiple measurement vectors," *Inverse Problems Imag.*, vol. 15, no. 1, pp. 79–107, 2021.

[40] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, 2004.

[41] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari, "Composite objective mirror descent," in *Proc. COLT*, vol. 10, 2010, pp. 14–26.

[42] T. Li, A. Kumar Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," 2018, *arXiv:1812.06127*.

[43] K. K. Patel, M. Glasgow, L. Wang, N. Joshi, and N. Srebro, "On the still unreasonable effectiveness of federated averaging for heterogeneous distributed learning," in *Proc. Federated Learn. Anal. Pract., Algorithms, Syst., Appl., Opportunities Workshop Int. Conf. Mach. Learn.*, 2023.

[44] Y. J. Cho, J. Wang, and G. Joshi, "Towards understanding biased client selection in federated learning," in *Proc. 25th Int. Conf. Artif. Intell. Statist. (AISTATS)*, Mar. 2022, pp. 10351–10375.

[45] A. Beck, *First-order Methods in Optimization*. Philadelphia, PA, USA: SIAM, 2017.

[46] Y. Nesterov, *Lectures on Convex Optimization*, vol. 137. Cham, Switzerland: Springer, 2018.

[47] Y. Bao, M. Crawshaw, S. Luo, and M. Liu, "Fast composite optimization and statistical recovery in federated learning," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 1508–1536.

[48] M. P. Friedlander and M. Schmidt, "Hybrid deterministic-stochastic methods for data fitting," *SIAM J. Scientific Comput.*, vol. 34, no. 3, pp. A1380–A1405, Jan. 2012.

[49] L. N. Trefethen and D. Bau, *Numerical Linear Algebra*, vol. 50. Philadelphia, PA, USA: SIAM, 1997.

[50] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD, USA: Johns Hopkins Univ. Press, 2013.

[51] N. Roux, M. Schmidt, and F. Bach, "A stochastic gradient method with an exponential convergence _rate for finite training sets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2663–2671.

[52] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 315–323.

[53] K. Bonawitz et al., "Towards federated learning at scale: System design," in *Proc. Mach. Learn. Syst.*, vol. 1, 2019, pp. 374–388.

[54] R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik, "SGD: General analysis and improved rates," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5200–5209.

[55] P.-L. Loh and M. J. Wainwright, "High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 2726–2734.

[56] H. Jeong, X. Li, Y. Plan, and O. Yilmaz, "Sub-Gaussian matrices on sets: Optimal tail dependence and applications," *Commun. Pure Appl. Math.*, vol. 75, no. 8, pp. 1713–1754, Aug. 2022.

[57] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[58] F. Haddadpour, B. Karimi, P. Li, and X. Li, "FedSKETCH: Communication-efficient and private federated learning via sketching," 2020, *arXiv:2008.04975*.

[59] D. P. Woodruff, "Sketching as a tool for numerical linear algebra," *Found. Trends Theor. Comput. Sci.*, vol. 10, nos. 1–2, pp. 1–157, 2014.

[60] X. Gu, K. Huang, J. Zhang, and L. Huang, "Fast federated learning in the presence of arbitrary device unavailability," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12052–12064.

[61] F. Mo, H. Haddadi, K. Katevas, E. Marin, D. Perino, and N. Kourtellis, "PPFL: Privacy-preserving federated learning with trusted execution environments," in *Proc. 19th Annu. Int. Conf. Mobile Syst., Appl., Services*, Jun. 2021, pp. 94–108.

[62] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[63] A. Aghazadeh et al., "Mission: Ultra large-scale feature selection using count-sketches," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 80–88.

Halyun Jeong received the bachelor's degree (Hons.) in mathematics, computer science, and electrical engineering from Pohang University of Science and Technology (POSTECH), the master's degree in electrical and computer engineering from the University of California, San Diego, and the Ph.D. degree in mathematics from the Courant Institute of Mathematical Sciences, New York University, in 2017. From August 2017 to June 2021, he was a Pacific Institute for the Mathematical Sciences (PIMS) Post-Doctoral Fellow with The University of British Columbia. Since July 2021, he has been an Assistant Adjunct Professor with the University of California, Los Angeles. He was a recipient of the Kwanjeong Scholarship Foundation for the Ph.D. Program.

Deanna Needell (Member, IEEE) received the Ph.D. degree from the University of California, Davis. She is currently a Full Professor of mathematics with the University of California, Los Angeles (UCLA), the Dunn Family Endowed Chair in Data Theory, and the Executive Director for the Institute for Digital Research and Education, UCLA. She is also a Post-Doctoral Fellow with Stanford University. She has been a Research Professor Fellow at several top research institutes including the SLMath (formerly MSRI) and Simons Institute, Berkeley, CA, USA. She has earned many awards, including the Alfred P. Sloan Fellowship, the NSF CAREER, the IMA Prize in Applied Mathematics, the 2022 American Mathematical Society (AMS) Fellow, and the 2024 Society for Industrial and Applied Mathematics (SIAM) Fellow. She also serves as an Associate Editor for several journals, including *Linear Algebra and its Applications* and *SIAM Journal on Imaging Sciences*, as well as on the organizing committee for SIAM sessions and the Association for Women in Mathematics.

Jing Qin (Member, IEEE) received the B.S. degree in mathematics from Xuzhou Normal University, Jiangsu, China, in 2005, the M.S. degree in mathematics from East China Normal University, Shanghai, China, in 2008, and the Ph.D. degree in applied mathematics from Case Western Reserve University, Cleveland, OH, USA, in 2013. From 2013 to 2016, she was an Assistant Adjunct Professor with the Department of Mathematics, University of California at Los Angeles, Los Angeles, CA, USA. From 2016 to 2019, she was an Assistant Professor with the Department of Mathematical Sciences, Montana State University, Bozeman, MT, USA. She is currently an Assistant Professor with the Department of Mathematics, University of Kentucky, Lexington, KY, USA. Her research interests include mathematical image processing, numerical optimization, and high-dimensional data analysis.