algorithms

MDPI

Article

Federated Optimization of ℓ_0-norm Regularized Sparse Learning

Qianqian Tong [1,†], Guannan Liang [1,†], Jiahao Ding [2], Tan Zhu [1], Miao Pan [2] and Jinbo Bi [1,*]

[1] Department of Computer Science and Engineering, University of Connecticut, Storrs, CT 06269, USA
[2] Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77204, USA
* Correspondence: jinbo.bi@uconn.edu
† These authors contributed equally to this work.

Abstract: Regularized sparse learning with the ℓ_0-norm is important in many areas, including statistical learning and signal processing. Iterative hard thresholding (IHT) methods are the state-of-the-art for nonconvex-constrained sparse learning due to their capability of recovering true support and scalability with large datasets. The current theoretical analysis of IHT assumes the use of centralized IID data. In realistic large-scale scenarios, however, data are distributed, seldom IID, and private to edge computing devices at the local level. Consequently, it is required to study the property of IHT in a federated environment, where local devices update the sparse model individually and communicate with a central server for aggregation infrequently without sharing local data. In this paper, we propose the first group of federated IHT methods: Federated Hard Thresholding (Fed-HT) and Federated Iterative Hard Thresholding (FedIter-HT) with theoretical guarantees. We prove that both algorithms have a linear convergence rate and guarantee for recovering the optimal sparse estimator, which is comparable to classic IHT methods, but with decentralized, non-IID, and unbalanced data. Empirical results demonstrate that the Fed-HT and FedIter-HT outperform their competitor—a distributed IHT, in terms of reducing objective values with fewer communication rounds and bandwidth requirements.

Keywords: ℓ_0-norm regularized sparse learning; iterative hard thresholding; federated learning; decentralized non-IID data

check for updates

Citation: Tong, Q.; Liang, G.; Ding, J.; Zhu, T.; Pan, M.; Bi, J. Federated Optimization of ℓ_0-norm Regularized Sparse Learning. *Algorithms* **2022**, *15*, 319. https://doi.org/10.3390/a15090319

Academic Editor: Zebang Shen

Received: 2 July 2022
Accepted: 24 August 2022
Published: 6 September 2022

1. Introduction

Sparse learning has emerged as a central topic of study in a variety of fields that require high-dimensional data analysis. Sparsity-constrained statistical models exploit the fact that high dimensional data arising from real-world applications frequently have low intrinsic complexity and have been shown to perform accurate estimation and inference in a variety of data mining fields, such as bioinformatics [1], image analysis [2,3], graph sparsification [4] and engineering [5]. These models often require solving the following optimization problem with a nonconvex, nonsmooth sparsity constraint:

$$\min_x f(x), \quad \text{subject to} \quad \|x\|_0 \leq \tau, \tag{1}$$

where $f(x)$ is a smooth and convex cost function in terms of a vector of parameters to be optimized x, $\|x\|_0$ denotes the l_0-norm (cardinality) of x, which computes the number of nonzero entries in x, and τ is the sparsity level pre-specified for x. Examples of this model include sparsity-constrained linear/logistic regression problems [6,7] and sparsity-constrained graphical models [8].

Extensive research has been conducted for Problem (1). The methods largely fall into the regimes of either matching pursuit methods [9–12] or iterative hard thresholding (IHT) methods [13–15]. Even though matching pursuit methods achieve remarkable success in minimizing quadratic loss functions (such as the ℓ_0-constrained linear regression problems), they require finding an optimal solution to $min \;\; f(x)$ over the identified support after hard thresholding at each iteration, which lacks analytical solutions for arbitrary losses and can be time-consuming [16]. Hence, gradient-based IHT methods have gained significant interest and become popular for nonconvex sparse learning. IHT methods currently include the gradient descent HT (GD-HT) [14], stochastic gradient descent HT (SGD-HT) [15], hybrid stochastic gradient HT (HSG-HT) [17], and stochastic variance reduced gradient HT (SVRG-HT) [18,19] methods. These methods update the iterate x_t as follows: $x_{t+1} = \mathcal{H}_\tau(x_t - \gamma_t v_t)$, where γ_t is the learning rate, v_t can be the full gradient, stochastic gradient or variance reduced gradient at the t-th iteration, and $\mathcal{H}_\tau(x) : \mathbb{R}^d \to \mathbb{R}^d$ denotes the HT operator that preserves the top τ elements in x and sets other elements to 0. However, finding a solution to Problem (1) is generally NP-hard because of the non-convexity and non-smoothness of the cardinality constraint [20].

Local datasets can be sensitive to sharing during the construction of a sparse inference model when sparse learning becomes distributed and uses data collected by distributed devices. For instance, meta-analyses may integrate genomic data from a large number of labs to identify (a sparse set of) genes contributing to the risk of a disease without sharing data across the labs [21,22]. Smartphone-based healthcare systems may need to learn the most important mobile health indicators from a large number of users; however, the personal health information gathered on the phone is private [23]. Furthermore, communication efficiency can be the main challenge to distributively training a sparse learning model. Due to the power and bandwidth limitations of various sensors, the signal processing community, for instance, has been seeking more communication-efficient methods [24].

Federated learning (FL) is a recently proposed communication-efficient distributed computing paradigm that enables collaborations among a collection of clients while preserving data privacy on each device by avoiding the transmission of local data to the central server [25–27]. Hence, sparse learning can benefit from the setting of federated learning. In this paper, we solve the federated nonconvex sparsity-constrained empirical risk minimization problem with decentralized data as follows:

$$\min_{x \in \mathbb{R}^d} f(x) = \sum_{i=1}^{N} p_i f_i(x), \quad \text{subject to} \;\; \|x\|_0 \leq \tau, \tag{2}$$

where $f(x)$ is a smooth and convex function, $f_i(x) = E_{z \sim \mathcal{D}_i}[f_i(x,z)]$ is the loss function of the i-th client (or device) with weight $p_i \in [0,1)$, $\sum_{i=1}^{N} p_i = 1$, $\mathcal{D}_i$ is the distribution of data located locally on the i-th client, and N is the total number of clients. It is thus desirable to solve Problem (2) in a communication-efficient way and investigate theory and algorithms applicable to a broader class of sparse constrained learning problems in high-dimensional data analyses [6–8,28].

We thus propose federated HT algorithms with lower communication costs and provide the corresponding theoretical analysis under practical federated settings. The analysis of proposed methods is difficult due to the fact that distributions of training data on each client may be non-identical and the data weights can be unbalanced across devices.

Our main contributions are summarized as follows.

(a) We develop two schemes for the federated HT method: the Federated Hard Thresholding (Fed-HT) algorithm and Federated Iterative Hard Thresholding (FedIter-HT) algorithm. In Fed-HT, we apply the HT operator $\mathcal{H}_\tau$ at the central server right before distributing the aggregated model to clients. To further improve the communication efficiency and the ability of sparsity recovery, in FedIter-HT, we consider applying $\mathcal{H}_\tau$ to both local updates and the central server aggregate. Note that this is the first trial to explore IHT algorithms under federated learning settings.

(b) We provide a set of theoretical results for the federated HT method, particularly of Fed-HT and FedIter-HT, under the realistic condition that the distributions of training data over devices can be unbalanced and non-independent and non-identical (non-IID), i.e., for $i \neq j$, $\mathcal{D}_i$ and $\mathcal{D}_j$ are different. We prove that both algorithms enjoy a linear convergence rate and have a strong guarantee for sparsity recovery. In particular, Theorems 1 (for the Fed-HT) and 2 (for the FedIter-HT) show that the estimation error between the algorithm iterate x_T and the optimal solution x^* is upper bounded as: $E\|x_T - x^*\| \leq \theta^T \|x_0 - x^*\|^2 + g(x^*)$, where x_0 is the initial guess of the solution, the convergence rate factor θ is related to the algorithm parameter K (the number of SGD steps on each device before communication), and the closeness between the pre-specified sparsity level τ and the true sparsity τ^*, and $g(x^*)$ determines a statistical bias term related not only to K but also to the gradient of f at the *sparse* solution x^* and the measurement of the non-IIDness of the data across the devices.

The theoretical results allow us to evaluate and compare the proposed methods. For example, greater non-IIDness among clients increases the bias of both algorithms. More local iterations may reduce θ but increase the statistical bias. Due to the utilization of the HT operator on local updates, the statistical bias induced by the FedIter-HT in Theorem 2 matches the best known upper bound for traditional IHT methods [17], which exhibits the powerful capability of sparsity recovery.

(c) When instantiating the general loss function by concrete squared or logistic loss, we arrive at specific sparse learning problems, such as sparse linear regression and sparse logistic regression. We provide statistical analysis of the maximum likelihood estimators (M-estimators) of these problems when using the FedIter-HT to solve them. This result can be regarded as federated HT analysis for generalized linear models.

(d) Extensive experiments in simulations and on real-life datasets demonstrate the effectiveness of the proposed algorithms over standard distributed IHT learning.

Related Work

Distributed sparse learning. Existing IHT algorithms can be extended to their distributed version—Distributed IHT (see Appendix A.1. for details), in which the central server aggregates (averages) the local parameter updates from each client and broadcasts the latest model parameters to individual clients, whereas each client updates the parameters based on the distributed local data with one step of stochastic gradient descend and sends them back to the central server. However, Distributed IHT is communication expensive since it needs to send dense local models to the central server after each step of stochastic gradient updates. Even though variants of the Distributed IHT, such as [29,30], have been developed, information must be exchanged at each iteration, making communication costly and limiting bandwidth. Other distributed methods have also been proposed. For instance, Ref. [31] tries to solve a relaxed l_1-norm regularized problem and thus introduces extra bias to Problem (2); Ref. [32] experimentally studies gradient compression with practical gradient clipping techniques (i.e., the local nodes have to select threshold) in distributed training; Ref. [33] proposed a modified distributed top-k sparsification by choosing the largest absolute gradients before updating the model to reduce communication. The distributed algorithms proposed in [32,33] use some techniques to reduce communication and bandwidth but are not designed for constrained optimization such as sparse model optimization.

Federated learning. FL is a privacy-preserving learning framework for large-scale machine learning on edge computing devices and solves the data-decentralized optimization problem: $\min_{x \in \mathbb{R}^d} f(x) = \sum_{i=1}^N p_i f_i(x)$ (without the sparsity constraint in Problem (2).) The FedAvg algorithm proposed in [25] can significantly reduce the communication cost by running multiple local SGD steps before each communication round and has become the de facto federated learning technique. Later, the client drift problem was observed for FedAvg [34–36], and the FedProx algorithm came to exist [37] in which the individual clients attempt to add a proximal operator to the local subproblem to address the issue of FedAvg. Researchers also study FL in the quantization strategy and the IoT (Internet of Things) systems, and the local updates are sparsified and compressed using signSGD [38–40]. Ref. [41] presents an online gradient sparsification method, which ensures that different clients provide a similar amount of updates and automatically determines the near-optimal communication and computation trade-off that is controlled by the degree of gradient sparsity. Yuan et al. recently studied a federated l_1-regularized logistic regression problem and proposed federated mirror descent algorithm [42] to solve (convex) nonsmooth composite optimization. However, federated optimization of ℓ_0-norm regularized sparse learning (as described in Problem (2)) is still under explored.

Organization of the paper is as follows: Section 2 provides the preliminaries formally, which include the notations used in this study as well as several generally held assumptions and lemmas. In Sections 3 and 4, the Fed-HT and FedIter-HT algorithms are proposed and studied, respectively. In Section 4, we normally perform a statistical analysis for M-estimators in order to emphasize the advantageous property of the FedIter-HT. Experiments simulating numerical performance are presented in Section 5.1, and benchmark datasets are analyzed in Section 5.2. Section 6 summarizes our outcomes. Appendix A contains the proof of our theoretical results and additional experiment details.

2. Preliminaries

We formalize our problem as Problem (2) and provide the notations (Table 1), assumptions and prepared lemmas used in this paper. We denote vectors by lowercase letters, e.g., x. The model parameters form a vector $x \in \mathbb{R}^d$. The ℓ_0-norm, ℓ_2-norm and the ℓ_∞-norm of a vector are denoted by $\|\cdot\|_0$, $\|\cdot\|$ and $\|\cdot\|_\infty$, respectively. Let $O(\cdot)$ represent the asymptotic upper bound, $[N]$ be the integer set $\{1,\ldots,N\}$. The support $\mathcal{I}^{(i)}_{t,k+1} = supp(x^*) \cup supp(x^{(i)}_{t,k}) \cup supp(x^{(i)}_{t,k+1})$ is associated with the $(k+1)$-th iteration in the t-th round on device i. For simplicity, we use $\mathcal{I}^{(i)} = \mathcal{I}^{(i)}_{t,k+1}$, $\mathcal{I} = \bigcup_{i=1}^{N} \mathcal{I}^{(i)}_{t,k+1}$ throughout the paper without ambiguity, and $\widetilde{\mathcal{I}} = supp(\mathcal{H}_{2N\tau}(\nabla f(\boldsymbol{x}^*))) \cup supp(\boldsymbol{x}^*)$.

Table 1. Brief summary of notations in this paper.

$\mathcal{H}_\tau(x)$	the HT operator that maintains the top τ items of x and sets the remaining elements to 0
N, i	the total number, the index of clients/devices
p_i	the weight of each loss function on client i
T, t	the total number, the index of communication rounds
K, k	the total number, the index of local iterations
$\nabla f_i(\cdot)$	the full gradient
$\nabla f_{I^{(i)}}(\cdot)$	the stochastic gradient over the minibatch $I^{(i)}$
$\nabla f_{i,z}(\cdot)$	the stochastic gradient over a training example indexed by z on the i-th device
γ_t	the stepsize/learning rate of local update
$\mathbb{I}(\cdot)$	an indicator function
$supp(x)$ x^*	the support of x or the index set of nonzero elements in x the optimal solution of Problem (2)
$x^{(i)}_{t,k}$	the local parameter vector on device i at the k-th iteration of the t-th round
τ	the required sparsity level
τ^*	the optimal sparsity level of Problem (2), $\tau^* = \|x^*\|_0$
$\pi_{\mathcal{I}}(x)$	the projector takes only the elements of x indexed in $\mathcal{I}$
$E[\cdot]$, $E^{(i)}[\cdot]$	the expectation over stochasticity across all clients and of client i, respectively

We use the same conditions employed in the theoretical analysis of other IHT methods by assuming that the objective function $f(x)$ satisfies the following conditions:

Assumption 1. *We assume that the loss function $f_i(x)$ on each device i*

1. *is restricted ρ_s-strongly convex (RSC [43]) at the sparsity level s for a given $s \in \mathbb{N}_+$, i.e., there exists a constant $\rho_s > 0$ such that $\forall x_1, x_2 \in \mathbb{R}^d$ with $\|x_1 - x_2\|_0 \leq s$, $i \in [N]$, we have*

$$f_i(x_1) - f_i(x_2) - \langle \nabla f_i(x_2), x_1 - x_2 \rangle \geq \frac{\rho_s}{2} \|x_1 - x_2\|^2;$$

2. *is restricted l_s-strongly smooth (RSS [43]) at the sparsity level s for a given $s \in \mathbb{N}_+$, i.e., there exists a constant $l_s > 0$ such that $\forall x_1, x_2 \in \mathbb{R}^d$ with $\|x_1 - x_2\|_0 \leq s$, $i \in [N]$, we have*

$$f_i(x_1) - f_i(x_2) - \langle \nabla f_i(x_2), x_1 - x_2 \rangle \leq \frac{l_s}{2} \|x_1 - x_2\|^2;$$

3. *has σ_i^2-bounded stochastic gradient variance, i.e.,*

$$E^{(i)}[\|\nabla f_{i,z}(x) - \nabla f_i(x)\|^2] \leq \sigma_i^2.$$

Remark 1. *When $s = d$, the above assumption is no longer restricted to the support at a sparsity level, and f_i is actually ρ_d-strongly convex and l_d-strongly smooth.*

Following the same convention in FL [35,37], we also assume the dissimilarity between the gradients of the local functions f_i and the global function f is bounded as follows.

Assumption 2. *The functions $f_i(x)$ $(i \in [N])$ are $\mathcal{B}$-locally dissimilar, i.e., there exists a constant $\mathcal{B} > 1$, such that*

$$\sum_{i=1}^{N} p_i \|\pi_{\mathcal{I}}(\nabla f_i(x))\|^2 \leq \mathcal{B}^2 \|\pi_{\mathcal{I}} \nabla f(x)\|^2$$

for any $\mathcal{I}$.

From the assumptions mentioned in the main text, we have the following lemmas to prepare for our theorems.

Lemma 1 ([44])**.** *For $\tau > \tau^*$ and for any parameter $x \in \mathbb{R}^d$, we have*

$$\|\mathcal{H}_\tau(x) - x^*\|_2^2 \leq (1 + \alpha)\|x - x^*\|_2^2,$$

where $\alpha = \frac{2\sqrt{\tau^*}}{\sqrt{\tau - \tau^*}}$ and $\tau^* = \|x^*\|_0$.

Lemma 2. *A differentiable convex function $f_i(x) : \mathbb{R}^d \to \mathbb{R}$ is restricted l_s-strongly smooth with parameter s, i.e., there exists a generic constant $l_s > 0$ such that for any x_1, x_2 with $\|x_1 - x_2\|_0 \leq s$ and*

$$f_i(x_1) - f_i(x_2) - \langle \nabla f_i(x_2), x_1 - x_2 \rangle \leq \frac{l_s}{2} \|x_1 - x_2\|^2,$$

then we have:

$$\|\nabla f_i(x_1) - \nabla f_i(x_2)\|^2 \leq 2l_s(f_i(x_1) - f_i(x_2) + \langle \nabla f_i(x_2), x_2 - x_1 \rangle).$$

The above two inequalities also hold for the global smoothness parameter l_d.

The proof of Lemma 2 can be found in Appendix A.3.

3. The Fed-HT Algorithm

In this section, we first describe our new federated ℓ_0-norm regularized sparse learning framework via hard thresholding—Fed-HT, and then discuss the convergence rate of our proposed algorithm.

A high level summary of Fed-HT is described in Algorithm 1. The Fed-HT algorithm generates a sequence of $\tau-$sparse vectors $x_1, x_2, \cdots$, from an initial sparse approximation x_0. At the $(t+1)$-th round, clients receive the global parameter update x_t from the central server, then run K steps of minibatch SGD based on local private data. In each step, the i-th client updates $x_{t,k+1}^{(i)} = argmin_x f_i(x_{t,k}^{(i)}) + \langle g_{t,k}^{(i)}, x - x_{t,k}^{(i)} \rangle + \frac{1}{2\gamma_t}\|x - x_{t,k}^{(i)}\|^2$ for $k \in \{0, \ldots, K-1\}$, i.e., $x_{t,k+1}^{(i)} = x_{t,k}^{(i)} - \gamma_t g_{t,k}^{(i)}$. Clients send $x_{t,K}^{(i)}$ for $i \in [N]$ back to the central server; then, the server averages them to obtain a dense global parameter vector and applies the HT operator to obtain a sparse iterate x_{t+1}. Unlike the commonly used FedAvg, the Fed-HT is designed to solve the family of federated ℓ_0-norm regularized sparse learning problems. It has a strong ability to recover the optimal sparse estimator in decentralized non-IID and unbalanced data settings while at the same time reducing the communication cost by a large margin because the central server broadcasts a sparse iterate for each of the T rounds.

Algorithm 1 Federated Hard Thresholding (Fed-HT)

```
Input: The learning rate $\gamma_t$, the sparsity level $\tau$, and the number of clients $N$.
Initialize $x_0$
for $t = 0$ to $T-1$ do
    for client $i = 1$ to $N$ parallel do
        $x_{t,1}^{(i)} = x_t$
        for $k = 1$ to $K$ do
            Sample uniformly a batch $I_{t,k}^{(i)}$ with batchsize $b_{t,k}^{(i)}$
            $g_{t,k}^{(i)} = \nabla f_{I_{t,k}^{(i)}}(x_{t,k}^{(i)})$
            $x_{t,k+1}^{(i)} = x_{t,k}^{(i)} - \gamma_t g_{t,k}^{(i)}$
        end for
    end for
    $x_{t+1} = \mathcal{H}_\tau(\sum_{i=1}^N p_i x_{t,K}^{(i)})$
end for
```

The following theorem characterizes our theoretical analysis of Fed-HT in terms of its **parameter estimation accuracy** for sparsity-constrained problems. Although this paper is focused on the cardinality constraint, the theoretical result is applicable to other sparsity constraints, such as a constraint based on matrix rank. Then, we have the main theorem and the detailed proof.

Theorem 1. *Let x^* be the optimal solution to Problem (2), $\tau^* = \|x^*\|_0$, and suppose $f(x)$ satisfies Assumptions 1 and 2. The condition number $\kappa_d = \frac{l_d}{\rho_d} \geq 1$. Let stepsize $\gamma_t = \frac{1}{6l_d}$ and the batch size $b_{t,k}^{(i)} = \frac{\Gamma_1}{\omega_1^t}$, $\Gamma_1 \geq \frac{\xi_1 \sum_{i=1}^N p_i \sigma_i^2}{\delta_1 \|x_0 - x^*\|^2}$, $\delta_1 = \alpha(1 - \frac{1}{12\kappa_d})^K$, $\alpha = \frac{2\sqrt{\tau^*}}{\sqrt{\tau - \tau^*}}$, the sparsity level $\tau \geq (16(12\kappa_d - 1)^2 + 1)\tau^*$. Then the following inequality holds for the Fed-HT:*

$$E[\|x_T - x^*\|^2] \leq \theta_1^T \|x_0 - x^*\|^2 + g_1(x^*).$$

where $\theta_1 = \omega_1 = (1 + 2\alpha)(1 - \frac{1}{12\kappa_d})^K \in (0, 1)$, $g_1(x^*) = \frac{\xi_1 \mathcal{B}^2}{1 - \psi_1}\|\nabla f(x^*)\|^2$, $\psi_1 = (1 + \alpha)(1 - \frac{1}{12\kappa_d})^K$, $\xi_1 = \frac{(1+\alpha)(1-(1-\frac{1}{12\kappa_d})^K)\kappa_d}{l_d^2}$.

The proof can be found in Appendix A.4.

Note that if the sparse solution x^* is sufficiently close to an unconstrained minimizer of $f(x)$, then $\|\nabla f(x^*)\|$ is small, so the first exponential term on the right-hand side can be a dominating term, which approaches 0 when T goes to infinity. We further obtain the following corollary that bounds the number of rounds T to obtain a sub-optimal solution, i.e., the difference between the solution and x^* is bounded only by the second term.

Corollary 1. *If all the conditions in Theorem 1 hold, for a given precision $\epsilon > 0$, we need at most $T \le C_1 \log(\frac{\|x_0 - x^*\|}{\epsilon})$ rounds to obtain*

$$E[\|x_T - x^*\|^2] \le \epsilon + g_1(x^*),$$

where $C_1 = -(\log(\theta_1))^{-1}$, $\theta_1 = (1+2\alpha)(1 - \frac{1}{12\kappa_d})^K \in (0,1)$, and $g_1(x^*) = \frac{\xi_1 \mathcal{B}^2}{1-\psi_1}\|\nabla f(x^*)\|^2$.

Remark 2. *Corollary 1 indicates that under proper conditions and with sufficient rounds, the estimation error of the Fed-HT is determined by the second term—the statistical bias term—which we denote as $g_1(x^*)$. The term $g_1(x^*)$ can become small if x^* is sufficiently close to an unconstrained minimizer of $f(x)$, so it represents the sparsity-induced bias to the solution of the unconstrained optimization problem. The upper bound result guarantees that the Fed-HT can closely approach x^* arbitrarily under a sparsity-induced bias, and the speed of approaching the biased solution is linear (or geometric) and determined by θ_1. In Theorem 1 and Corollary 1, θ_1 is closely related to the number of local updates K. The condition number $\kappa_d > 1$, so $(1 - \frac{1}{12\kappa_d}) < 1$. When K is larger, θ_1 is smaller, so is the number of rounds T required for reaching a target ϵ. In other words, the Fed-HT converges faster with fewer communication rounds. However, the bias term $g_1(x^*)$ will increase when K increases. Therefore, K should be chosen to balance the convergence rate and statistical bias.*

We further investigate how the objective function $f(x)$ approaches the optimal $f(x^*)$.

Corollary 2. *If all the conditions in Theorem 1 hold, let $\Delta_1 = l_d\|x_0 - x^*\|^2$, and $g_2(x^*) = O(\|\nabla f(x^*)\|^2)$, we have*

$$E[f(x_T) - f(x^*)] \le \theta_1^T \Delta_1 + g_2(x^*).$$

The proof details can be found in Appendix A.5.

Because the local updates on each device are based on SGD with dense parameters, without the HT operator, l_d-smoothness and ρ_d-strongly convexity are required, which depend on dimension d and are stronger requirements for f. Furthermore, $\|\nabla f(x^*)\| \le d\|f(x^*)\|_\infty$, i.e., $g_1(x^*)$ and $g_2(x^*)$ are $O(d^2\|f(x^*)\|_\infty^2)$, which are suboptimal compared with the results for traditional IHT methods in terms of dimension d. In order to solve such drawbacks, we develop a new algorithm in the next section.

4. The FedIter-HT Algorithm

If we apply the HT operator to each local update as well, we obtain the FedIter-HT algorithm, as described in Algorithm 2. Hence, the local update on each device performs multiple SGD-HT steps, which further reduces the communication cost because model parameters sent back from clients to the central server are also sparse. If a client has a communication bandwidth so small that it can not effectively pass the full set of parameters, the FedIter-HT provides a good solution and also can relax the strict requirements for the objective function f and reduce the statistical bias. In this section, we first present a more communication-efficient federated ℓ_0-norm regularized sparse learning framework—FedIter-HT; then, we theoretically show it enjoys a better convergence rate compared with Fed-HT, and we further provide statistical analysis for M-estimators under the framework of FedIter-HT.

We again examine the convergence of the FedIter-HT by developing an upper bound on the distance between the estimator x_T and the optimal x^*, i.e., $E[\|x_T - x^*\|^2]$ in the following theorem.

Algorithm 2 Federated Iterative Hard Thresholding (FedIter-HT)

```
Input: The learning rate $\gamma_t$, the sparsity level $\tau$, and the number of clients $N$.
Initialize $x_0$
for $t = 0$ to $T - 1$ do
    for client $i = 1$ to $N$ parallel do
        $x_{t,1}^{(i)} = x_t$
        for $k = 1$ to $K$ do
            Sample uniformly a batch $I_{t,k}^{(i)}$ with batchsize $b_{t,k}^{(i)}$
            $g_{t,k}^{(i)} = \nabla f_{I_{t,k}^{(i)}}(x_{t,k}^{(i)})$
            $x_{t,k+1}^{(i)} = \mathcal{H}_\tau(x_{t,k}^{(i)} - \gamma_t g_{t,k}^{(i)})$
        end for
    end for
    $x_{t+1} = \mathcal{H}_\tau(\sum_{i=1}^N p_i x_{t,K}^{(i)})$
end for
```

Theorem 2. *Let x^* be the optimal solution to (2), $\tau^* = \|x^*\|_0$, and suppose $f(x)$ satisfies Assumptions 1 and 2. The condition number $\kappa_s = \frac{l_s}{\rho_s} \geq 1$. Let stepsize $\gamma_t = \frac{1}{6l_s}$ and the batch size $b_{t,k}^{(i)} = \frac{\Gamma_2}{\omega_2^t}$, $\Gamma_2 \geq \frac{\xi_2 \sum_{i=1}^N p_i \sigma_i^2}{\delta_2 \|x_0 - x^*\|^2}$, $\delta_2 = (2\alpha + 3\alpha^2)(1 - \frac{1}{12\kappa_s})^K$, $\alpha = \frac{2\sqrt{\tau^*}}{\sqrt{\tau - \tau^*}}$, the sparsity level $\tau \geq (\frac{16}{(\sqrt{\frac{12\kappa_s}{12\kappa_s - 1}} - 1)^2} + 1)\tau^*$. Then, the following inequality holds for the FedIter-HT:*

$$E[\|x_T - x^*\|^2] \leq \theta_2^T \|x_0 - x^*\|^2 + g_3(x^*).$$

where $\theta_2 = \omega_2 = (1 + 2\alpha)^2(1 - \frac{1}{12\kappa_s})^K \in (0, 1)$, $g_3(x^*) = \frac{\xi_2 \mathcal{B}^2}{1 - \psi_2} \|\pi_{\tilde{\mathcal{I}}}(\nabla f(x^*))\|^2$, $\xi_2 = \frac{(1+\alpha)^2(1-(1-\frac{1}{12\kappa_s})^K)\kappa_s}{l_s^2}$, $\psi_2 = (1 + \alpha)^2(1 - \frac{1}{12\kappa_s})^K$, $\alpha = \frac{2\sqrt{\tau^*}}{\sqrt{\tau - \tau^*}}$, $\widetilde{\mathcal{I}}^i = supp(\mathcal{H}_{2\tau}(\nabla f_i(\boldsymbol{x}^*))) \cup supp(\boldsymbol{x}^*)$ and $\widetilde{\mathcal{I}} = supp(\mathcal{H}_{2N\tau}(\nabla f(\boldsymbol{x}^*))) \cup supp(\boldsymbol{x}^*)$.

The proof details can be found in Appendix A.6.

Remark 3. *The factor θ_2, compared with θ_1 in Theorem 1, is smaller if $2\alpha = \frac{4\sqrt{\tau^*}}{\sqrt{\tau - \tau^*}} \leq (\frac{1 - 1/12\kappa_d}{1 - 1/12\kappa_s})^K - 1$, which means that the FedIter-HT converges faster than the Fed-HT when the beforehand-guessed sparsity τ is much larger than the true sparsity. Both θ_2 and θ_1 will decrease when the number of internal iterations K increases, but θ_2 decreases faster than θ_1 because $1 - \frac{1}{12\kappa_s}$ is smaller than $1 - \frac{1}{12\kappa_d}$. Thus, the FedIter-HT is more likely to benefit by increasing K than the Fed-HT. The statistical bias term $g_3(x^*)$ can be much smaller than $g_1(x^*)$ in Theorem 1 because $g_3(x^*)$ only depends on the norm of $\nabla f(x^*)$ restricted to the support $\tilde{\mathcal{I}}$ of size $2N\tau + \tau^*$. Because the norm of the gradient is a dominating term in g_1 and g_3, slightly increasing K does not significantly vary the statistical bias terms (when $d \gg 2N\tau + \tau^*$).*

Using the results in Theorem 2, we can further derive Corollary 3 to specify the number of rounds required to achieve a given estimation precision.

Corollary 3. *If all the conditions in Theorem 2 hold, for a given $\epsilon > 0$, the FedIter-HT requires the most $T \leq C_2 \log(\frac{\|x_0 - x^*\|}{\epsilon})$ rounds to obtain*

$$E[\|x_T - x^*\|^2] \leq \epsilon + g_3(x^*),$$

where $C_2 = -(\log(\theta_2))^{-1}$.

Because $g_3(x^*) = O(\|\pi_{\tilde{\mathcal{I}}}(\nabla f(x^*))\|^2)$, and we also know $\|\pi_{\tilde{\mathcal{I}}}(\nabla f(x^*))\|^2 \leq (2N\tau + \tau^*)^2 \|\nabla f(x^*)\|_\infty^2$ and $2N\tau + \tau^* \ll d$ in high dimensional statistical problems, the result in Corollary 3 gives a tighter bound than the one obtained in Corollary 1. Similarly, we also obtain a tighter upper bound for the convergence performance of the objective function $f(x)$.

Corollary 4. *If all the conditions in Theorem 2 hold, let* $\Delta_2 = l_s\|x_0 - x^*\|^2$, *and* $g_4(x^*) = O(\|\pi_{\tilde{\mathcal{I}}}(\nabla f(x^*))\|^2)$, *we have*

$$E[f(x_T) - f(x^*)] \le \theta_2^T \Delta_2 + g_4(x^*).$$

The proof details can be found in Appendix A.7.

The theorem and corollaries developed in this section only depend on the l_s-restricted smoothness and ρ_s-restricted strong convexity, where $s = 2\tau + \tau^*$, which are the same conditions used in the analysis of existing IHT methods. Moreover, $\|\pi_{\tilde{\mathcal{I}}}(\nabla f(x^*))\| \le (2N\tau + \tau^*)\|\nabla f(x^*)\|_\infty$, which means $g_3(x^*)$ and $g_4(x^*)$ are $O((2N\tau + \tau^*)^2\|\nabla f(x^*)\|_\infty^2)$, where $2N\tau + \tau^*$ is the size of support $\tilde{\mathcal{I}}$. Therefore, our results match the current best-known upper bound for the statistic bias term compared with the results for traditional IHT methods.

Statistical Analysis for M-Estimators

Due to the good property of the FedIter-HT, we also study its constrained M-estimators derived from more concrete learning formulations. Although we focus on the sparse linear regression and sparse logistic regression in this paper, our method can be used to analyze other statistical learning problems as well.

Sparse Linear Regression can be formulated as follows:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{B}\|Y^{(i)} - Z^{(i)}x\|_2^2,$$
$$\text{subject to } \ \|x\|_0 \le \tau,$$

where $Z^{(i)} \in \mathbb{R}^{B \times d}$ is a design matrix associated with client i. For each row of matrix $Z^{(i)}$, we further assume that they are independently drawn from a sub-Gaussian distribution with parameter $\beta^{(i)}$, $Y^{(i)} = Z^{(i)}x^* + \epsilon^{(i)}$ denotes the response vector, and $\epsilon^{(i)} \in \mathbb{R}^B$ is a noise vector following Normal distribution $N(0, \sigma^2 I)$, $x^* \in \mathbb{R}^d$ with $\|x^*\|_0 = \tau^*$ is the underlying sparse regression coefficient vector.

Corollary 5. *If all the conditions in Theorem 2 hold, with* $B \ge C_1\tau\log(d)\max_i\{(\beta^{(i)})^2\}$ *and a sufficiently large number of communication rounds* T, *we have*

$$E[\|x_T - x^*\|^2] \le O(\frac{(2N\tau + \tau^*)\sigma^2\mathcal{B}^2(\sum_{i=1}^N \beta^{(i)})^2\log(d)}{NB})$$

with a probability of at least $(1 - \exp(-C_5 NB))$, *where* C_5 *is a universal constant.*

Proof. Let $Z = [Z^{(1)}; \dots; Z^{(N)}] \in \mathbb{R}^{NB \times d}$ be the overall design matrix of the linear regression problem, and each row of Z can be treated as drawn IID from a sub-Gaussian distribution with parameter $\sum_{i=1}^N \beta^{(i)}$. $\epsilon = [\epsilon^{(1)}; \dots; \epsilon^{(N)}] \in \mathbb{R}^{NB \times 1}$ is the random Gaussian noise. Then Lemma C.1 in [45] immediately implies that f_i is restricted ρ_s-strongly convex and restricted l_s-strongly smooth with $\rho_s = \frac{4}{5}$ and $l_s = \frac{6}{5}$, respectively, with a probability of at least $(1 - \exp(-C_2 B))$ if the total sample size $B \ge C_1\tau\log(d)\max_i\{(\beta^{(i)})^2\}$, where C_1 and C_2 are universal constants. Furthermore, we know that $\|\nabla f(x^*)\|_\infty = \|\frac{Z^T\epsilon}{NB}\|_\infty \le C_3\sigma\sum_{i=1}^N \beta^{(i)}\sqrt{\frac{\log(d)}{NB}}$, with a probability of at least $(1 - \exp(-C_4 NB))$, where C_3, C_4 are universal constants. Gathering everything together yields the following bound with a high probability. □

Sparse Logistic Regression can be formulated as follows:

$$\underset{x}{min} f(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{B}\sum_{j=1}^{B}(\log(1+\exp(z_{i,j}^T x)) - y_{i,j}z_{i,j}^T x)$$
$$\text{subject to} \quad \|x\|_0 \le \tau,$$

where $z_{i,j} \in \mathbb{R}^d$ for $j \in [B]$ is a predictive vector and drawn from a sub-Gaussian distribution associated with client i, each observation $y_{i,j}$ on client i is drawn from the Bernoulli distribution $\mathbb{P}(y_{i,j}|z_{i,j}, x^*) = \frac{\exp(z_{i,j}^T x^*)}{1+\exp(z_{i,j}^T x^*)}$, and $x^* \in \mathbb{R}^d$ with $\|x^*\|_0 = \tau^*$ is the underlying true parameter that we want to recover.

Corollary 6. *If all the conditions in Theorem 2 hold, $\|z_{i,j}\| \le \mathcal{K}$, $C_{lower} \le \exp(z_{i,j}^T x)/(1+\exp(z_{i,j}^T x))^2 \le C_{upper}$ for $i \in [N]$ and $j \in [B]$ and $B \ge C_7\tau\mathcal{K}^2 log(d)$ and with a sufficiently large number of communication rounds T, we have*

$$E[\|x_T - x^*\|^2] \le O(\frac{(2N\tau + \tau^*)\mathcal{B}^2\mathcal{K}^2\log(d)}{NB})$$

with a probability of at least $(1 - \exp(-C_6 NB) - C_9\exp(-C_{10}log(d)) + \frac{C_9}{\exp(C_6 NB)\exp(C_{10}log(d))})$, where C_6, C_9 and C_{10} are constants.

Proof. If we further assume $\|z_{i,j}\| \le \mathcal{K}$ and $C_{lower} \le \exp(z_{i,j}^T x)/(1+\exp(z_{i,j}^T x))^2 \le C_{upper}$ for $i \in [N]$ and $j \in [B]$, the sparse logistic regression objective function is restricted ρ_s-strongly convex and restricted l_s-strongly smooth with $\rho_s = \frac{4}{5}C_{lower}$ and $l_s = \frac{6}{5}C_{upper}$, respectively, with a probability of at least $(1-\exp(-C_6B))$ if $B \ge C_7\tau\mathcal{K}^2 log(d)$, where C_{lower}, C_{upper}, C_6 and C_7 are constants. Furthermore, according to Corollary 2 in [46], we have $\|\nabla f(x^*)\|_\infty \le C_8\mathcal{K}\sqrt{\log(d)/NB}$ with a probability of at least $(1 - C_9 exp(-C_{10}log(d))$, where C_8, C_9 and C_{10} are universal constants. Therefore, we can obtain the following corollary. Based on the above result, the estimation error specified in terms of the distances x_T and x^* decreases when the total sample size NB is large or the dissimilarity level $\mathcal{B}$ and the dimension d are small. □

5. Experiments

We empirically evaluate our methods in both simulations and on three real-world datasets: E2006-tfidf, RCV1 and MNIST (Table 2, which are downloaded from the LibSVM website (https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/, accessed on 1 July 2022)), and compare them against a baseline method. The baseline method is a standard Distributed IHT and communicates every local update to the central server, which then aggregates and broadcasts back to clients (see Appendix A.1 for more details). Specifically, experiments for simulation I and on the E2006-tfidf dataset are conducted for sparse linear regression. We solve the sparse logistic regression problem in simulation II and for the RCV1 data set. The last experiment uses MNIST data in a multi-class softmax regression problem. The exact loss functions for the various problems are available in the Appendix A.2.

Following the convention in the federated learning literature, we use the number of communication rounds to measure the communication cost. For a comprehensive comparison, we also include the number of iterations. For both synthetic and real-world datasets, algorithm parameters are determined by the following criteria. The number of local iterations K is searched from $\{3, 5, 8, 10\}$. We have tested the performance of our proposed algorithms under different K conditions (see Figure 1). The stepsize γ for each algorithm is set by a grid search from $\{10, 1, 0.6, 0.3, 0.1, 0.06, 0.03, 0.01, 0.001\}$. All the algorithms are initialized with $x^{(0)} = 0$. The sparsity τ is 500 for the MNIST dataset and 200 for the other two datasets.

Table 2. Statistics of three real-world datasets in the federated setting.

Dataset	Samples	Dimension	Samples Per Device	
			Mean	**Stdev**
E2006-tfidf	3308	150,360	33.8	9.1
RCV1	20,242	47,236	202.4	114.5
MNIST	60,000	784	600	–

5.1. Simulations

To generate synthetic data, we follow a similar setup to that in [37]. In simulation I, for each device $i \in [100]$, we generate samples $(z_{i,j}, y_{i,j})$ for $j \in [100]$ according to $y_{i,j} = z_{i,j}^T x_i + b_{i,j}$, where $z_{i,j} \in \mathbb{R}^{1000}$, $x_i \in \mathbb{R}^{1000}$. The first 100 elements of x_i are drawn from $\mathcal{N}(u_i, 1)$ and the remaining elements in x_i are zeros, $b_{i,j} \sim \mathcal{N}(u_i, 1)$, $u_i \sim \mathcal{N}(0.1, \alpha)$, $z_{i,j} \sim \mathcal{N}(v_i, \Sigma)$, where Σ is a diagonal matrix with the i-th diagonal element equal to $\frac{1}{i^{1.2}}$. Each element in the mean vector v_i is drawn from $\mathcal{N}(B_i, 1)$, $B_i \sim \mathcal{N}(0, \beta)$. Therefore, α controls how much the local models differ from each other, and β controls how much the local on-device data differ between one another; hence, we have simulated Non-IID federated data. In simulation I, $(\alpha, \beta) \in \{(0.1, 0.1), (0.5, 0.5), (1, 1)\}$. The data generation procedure for simulation II is the same as the procedure of simulation I, except that $y'_{i,j} = \exp(z_{i,j}^T x_i + b_{i,j}) / (1 + \exp(z_{i,j}^T x_i + b_{i,j}))$; then, for the i-th client, we set $y_{i,j} = 1$ corresponding to the top 100 of $y'_{i,j}$ for $j \in [1000]$; otherwise, $y_{i,j} = 0$. In simulation II, we also set $(\alpha, \beta) \in \{(0.1, 0.1), (0.5, 0.5), (1, 1)\}$.

The results in Figure 2 show that, with a higher degree of Non-IID, both Fed-HT and FedIter-HT tend to converge slower. We also compare the proposed methods with the baseline method—Distributed IHT. In Figure 3, we observe that in simulation I, FedIter-HT only needs 20 (∼5× less) communication rounds to reach the same objective value that the Distributed-IHT obtains with more than 100 communication rounds; in simulation II, the FedIter-HT needs 50 communication rounds (∼4× less) to achieve the same objective value that the Distributed-IHT obtains with 200 communication rounds.

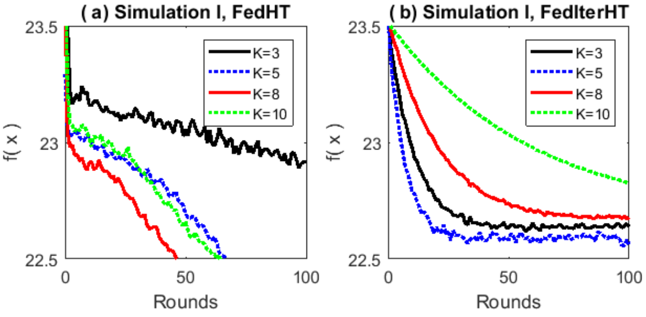

Figure 1. The comparison of proposed algorithms for different K values in terms of the objective function value vs. communication rounds (**a**,**b**).

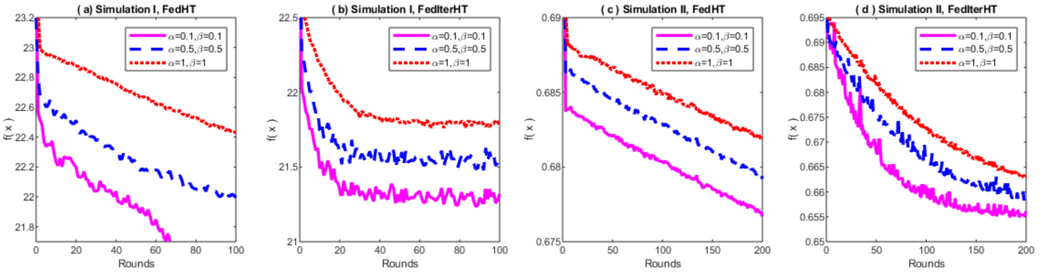

Figure 2. The objective function value vs. communication rounds for regression (**a**,**b**) and classification (**c**,**d**), and for Fed-HT (**a**,**c**) and FedIter-HT (**b**,**d**) with varying degrees of non-IID data.

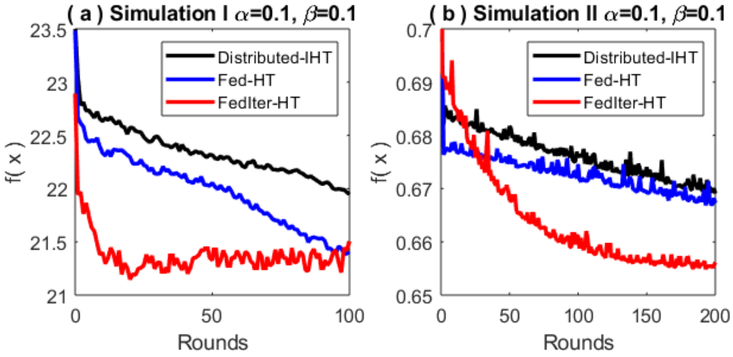

Figure 3. The comparison of different algorithms in terms of the objective function value vs. communication rounds (**a**,**b**) and for regression (**a**) and classification (**b**). Note that the Distributed-IHT is the baseline method that communicates every local update (so the number of rounds equals the number of iterations) and may be the best scenario for reducing the objective value.

5.2. Benchmark Datasets

We use the E2006-tfidf dataset [47] to predict the volatility of stock returns based on the SEC-mandated financial text report, represented by tf-idf. It was collected from thousands of publicly traded U.S. companies, for which data from different companies are inherently non-identical and the privacy consideration for financial data demands federated learning. The RCV1 dataset [48] is used to predict categories of newswire stories recently collected by Reuters, Ltd. The RCV1 can be naturally partitioned based on the news category and used for federated learning experiments since readers may only be interested in one or two categories of news. Our model training process mimics the personalized privacy-preserving news recommender system where we use the K-means method to partition the datasets, respectively, into 10 clusters. Each device randomly selects two of the clusters for use in the learning. We run t-SNE to visualize the hidden structures found by K-means as shown in Figures 4 and 5, respectively, for the E2006-tfidf dataset (sparse linear regression) and the RCV1 dataset (sparse logistic regression). For the MNIST images, there are 10 digits that automatically serve as the clusters.

For all datasets, the data in each cluster are evenly partitioned into 20 parts, and each client randomly picks two clusters and selects one part of data from each of the clusters. Because the MNIST images are evenly collected for each digit, the partitioned decentralized MNIST data are balanced in terms of categories, whereas the other two datasets are unbalanced.

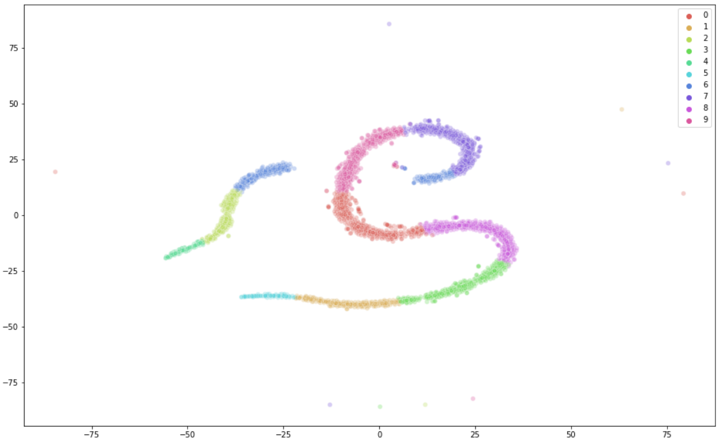

Figure 4. Visualization of 10 K-means clusters for E2006-dfidf using t-SNE.

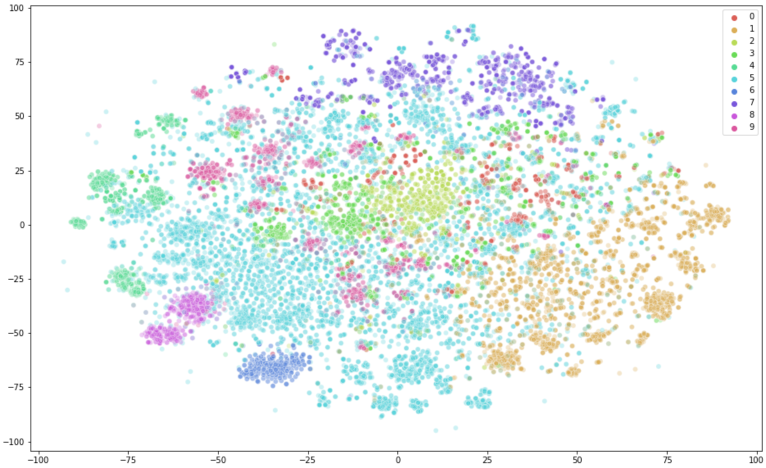

Figure 5. Visualization of 10 K-means clusters for RCV1 using t-SNE.

Figure 6 shows that our proposed Fed-HT and FedIter-HT can significantly reduce the communication rounds required to achieve a given accuracy. In Figure 6a,c, we further notice that federated learning displays more randomness when approaching the optimal solution. This may be caused by dissimilarity across clients. For instance, the three different algorithms in Figure 6c reach the neighborhood of different solutions at the end, where the proposed FedIter-HT obtains the lowest objective value. These behaviors may be worth exploring further in the future.

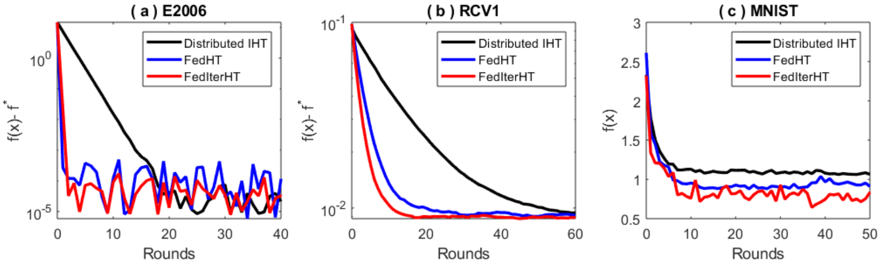

Figure 6. Comparison of the algorithms on different datasets in terms of the objective function value vs. communication rounds. f^* is a lower bound of $f(x)$. FedIter-HT performs consistently better across all datasets, which confirms our theoretical result.

6. Conclusions

In this paper, we propose two communication-efficient federated IHT methods—Fed-HT and FedIter-HT—to deal with ℓ_0-norm regularized sparse learning with decentralized non-IID data. The Fed-HT algorithm is designed to impose a hard thresholding operator at a central server, whereas the FedIter-HT applies this operator at each update regardless of local clients or a central server. Both methods reduce communication costs—in both the communication rounds and the communication load at each round. Theoretical analyses show a linear convergence rate for both algorithms where the Fed-HT has a better convergence rate θ, but the FedIter-HT has a better statistical estimation bias. Similar to the conventional IHT methods with IID data, there is still a guarantee to recover the best sparse estimator even with decentralized non-IID data. They outperform the traditional Distributed-IHT in simulations and on benchmark datasets, according to empirical findings.

Author Contributions: Conceptualization, Q.T. and G.L.; methodology, Q.T. and J.B.; software, G.L.; validation, G.L., J.D. and T.Z.; formal analysis, Q.T. and G.L.; investigation, Q.T.; resources, Q.T., G.L. and J.B.; writing—original draft preparation, Q.T., G.L. and J.B.; writing—review and editing, J.B. and M.P.; visualization, Q.T., G.L., J.D. and T.Z.; supervision, J.B.; project administration, J.B. All authors have read and agreed to the published version of the manuscript.

Funding: The work of Q.T. and G.L. was partially funded by the U.S. National Science Foundation (NSF) under grants IIS-1718738 to J.B. T.Z. was partially funded by a grant of National Institutes of Health (NIH) 5K02DA043063-03 to J.B. J.B. was also funded by an NIH R01-DA051922-01 grant. The work of J.D. and M.P. was funded in part by the NSF under grants US CNS-1801925 and CNS-2029569 to M.P.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

IHT	iterative hard thresholding
SGD	stochastic gradient descent
HT	hard thresholding
FL	federated learning
IID	independent and identically distributed

Appendix A

Appendix A.1. Distributed IHT Algorithm

Here we describe the distributed implementation of the IHT method in Algorithm A1, and we use it as a baseline to compare with the two federated IHT methods proposed in the present paper.

Algorithm A1 Distributed-IHT

```
Input: Learning rate $\gamma_t$, number of workers $N$.
Initialize $x_0$
for $t = 0$ to $T - 1$ do
    for worker $i = 1$ to $N$ parallel do
        Receive $x_t^{(i)} = x_t$ from the central server
        Calculate unbiased stochastic gradient direction $v_t^{(i)}$ on worker $i$
        Locally update: $x_{t+1}^{(i)} = x_t^{(i)} - \gamma_t v_t^{(i)}$
        Send $x_{t+1}^{(i)}$ to the central server
    end for
    Receive all local updates and average on a remote server: $x_{t+1} = \mathcal{H}_\tau(\sum_{i=1}^N p_i x_{t+1}^{(i)})$
end for
```

Appendix A.2. More Experimental Details

In more detail, experiments in simulation I and on the real-life dataset E2006-tfidf were conducted with sparse linear regression,

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{B^{(i)}} \|Y^{(i)} - Z^{(i)} x\|_2^2, \quad \text{subject to} \quad \|x\|_0 \leq \tau.$$

Experiments in simulation II and on the RCV1 dataset were conducted with sparse logistic regression

$$\min_x f(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{B^{(i)}}\sum_{j=1}^{B^{(i)}}(\log(1+exp(y_{i,j}z_{i,j}^T x)) + \frac{\lambda}{2}\|x\|^2), \quad \text{subject to} \quad \|x\|_0 \le \tau.$$

The last experiment solves a multi-class softmax regression problem on the MNIST dataset as follows:

$$\min_x \{f(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{B^{(i)}}\sum_{j=1}^{B^{(i)}}(\sum_{r=1}^{c}(-\mathbb{I}(y_{i,j}=r)\log(\frac{\exp(z_{i,j}^T x_r)}{\sum_{l=1}^{c}\exp(z_{i,j}^T x_l)}) + \frac{\lambda}{2}\|x_r\|^2))\},$$
$$\text{subject to} \quad \|x_r\|_0 \le \tau, \quad \forall r \in \{1,2,\dots,c\}.$$

Appendix A.3. Proof of Lemma 2

Results of Lemma 2 are used particularly in the proof of the Corollary 2, we provide a brief proof of this lemma.

Proof. Let $\phi(v) = f_i(v) - \langle \nabla f_i(x), v\rangle$, then $\phi(y)$ is restricted l_s-strongly smooth with parameter s too. Because f_i is convex, $\phi(v)$ is also convex, and x is a minimizer of $\phi(v)$ due to $\nabla\phi(x) = 0$. We define

$$\begin{aligned}\phi(x) &= \min_v \phi(v) && \text{(A1)}\\ &\le \min_v\{\phi(y) + \langle\nabla\phi(y), v-y\rangle + \frac{l_s}{2}\|v-y\|^2\} && \text{(A2)}\\ &= \phi(y) - \frac{1}{2l_s}\|\nabla\phi(y)\|^2\end{aligned}$$

where the equality (A1) is due to $\nabla\phi(x) = 0$; inequality (A2) is due to restricted l_s-strongly smoothness.

Let $y = x_1$ and $x = x_2$ and reorganize, we have

$$\|\nabla f_i(x_1) - \nabla f_i(x_2)\|^2 \le 2l_s(f_i(x_1) - f_i(x_2) + \langle\nabla f_i(x_2), x_2 - x_1\rangle).$$

Furthermore, for the global smoothness parameter l_d, we have

$$\|\nabla f_i(x_1) - \nabla f_i(x_2)\|^2 \le 2l_d(f_i(x_1) - f_i(x_2) + \langle\nabla f_i(x_2), x_2 - x_1\rangle).$$

□

Appendix A.4. Proof of Theorem 1

Proof. For the Fed-HT algorithm:

$$\begin{aligned}E[\|x_{t+1} - x^*\|^2] &= E[\|\mathcal{H}_\tau(\sum_{i=1}^{N} p_i x_{t,K}^{(i)}) - x^*\|^2]\\ &\le (1+\alpha)E[\|\sum_{i=1}^{N} p_i x_{t,K}^{(i)} - x^*\|^2] && \text{(A3)}\\ &= (1+\alpha)E[\|\sum_{i=1}^{N} p_i x_{t,K}^{(i)} - \sum_{i=1}^{N} p_i x^*\|^2] && \text{(A4)}\\ &\le (1+\alpha)\sum_{i=1}^{N} p_i E^{(i)}[\|x_{t,K}^{(i)} - x^*\|^2]. && \text{(A5)}\end{aligned}$$

Equation (A3) holds due to Lemma 1, Equation (A4) holds because $\sum_{i=1}^N p_i = 1$, Equation (A5) holds due to Jensen's Inequality, and the sampling procedures across different clients are independent of each other.

We calculate the stochastic gradient, which is essential in a local update, and we split the stochastic gradient into three terms. Note that the last inequality holds due to bounded variance on support assumption and the inequality $\|\nabla f_i(x_t) - \nabla f_i(x^*)\|^2 \le 2l_d(f_i(x_t) - f_i(x^*) + \langle \nabla f_i(x^*), x_t - x^*\rangle)$.

$$
\begin{aligned}
&\sum_{i=1}^N p_i E^{(i)}[\|g_{t,K-1}^{(i)}\|^2] = \sum_{i=1}^N p_i E^{(i)}[\|g_{t,K-1}^{(i)} - \nabla f_i(x_{t,K-1}^{(i)}) + \nabla f_i(x_{t,K-1}^{(i)}) - \nabla f_i(x^*) + \nabla f_i(x^*)\|^2] \\
&\le 3\sum_{i=1}^N p_i E^{(i)}[\|g_{t,K-1}^{(i)} - \nabla f_i(x_{t,K-1}^{(i)})\|^2] + 3\sum_{i=1}^N p_i E^{(i)}[\|\nabla f_i(x_{t,K-1}^{(i)}) - \nabla f_i(x^*)\|^2] \\
&+ 3\sum_{i=1}^N p_i \|\nabla f_i(x^*)\|^2 \\
&\le 3\sum_{i=1}^N p_i \frac{\sigma_i^2}{b_t} + 3\sum_{i=1}^N p_i \|\nabla f_i(x^*)\|^2 + 6l_d \sum_{i=1}^N p_i E^{(i)}[(f_i(x_{t,K-1}^{(i)}) - f_i(x^*) + \langle \nabla f_i(x^*), x_{t,K-1}^{(i)} - x^*\rangle)].
\end{aligned}
\tag{A6}
$$

Next, we want to build the connection of $\sum_{i=1}^N p_i E^{(i)}[\|x_{t,K}^{(i)} - x^*\|^2]$ and $\sum_{i=1}^N p_i E^{(i)}[\|x_{t,K-1}^{(i)} - x^*\|^2]$. Let $\gamma_t = \frac{1}{6l_d}$. Consider the inner loop iteration:

$$
\begin{aligned}
&\sum_{i=1}^N p_i E^{(i)}[\|x_{t,K}^{(i)} - x^*\|^2] = \sum_{i=1}^N p_i E^{(i)}[\|x_{t,K-1}^{(i)} - \frac{1}{6l_d} g_{t,K-1}^{(i)} - x^*\|^2] \\
&= \sum_{i=1}^N p_i E^{(i)}[\|x_{t,K-1}^{(i)} - x^*\|^2] + \frac{1}{36l_d^2}\sum_{i=1}^N p_i E^{(i)}[\|g_{t,K-1}^{(i)}\|^2] - \frac{1}{3l_d}\sum_{i=1}^N p_i E^{(i)}[\langle x_{t,K-1}^{(i)} - x^*, g_{t,K-1}^{(i)}\rangle] \\
&\le \sum_{i=1}^N p_i E^{(i)}[\|x_{t,K-1}^{(i)} - x^*\|^2] + \frac{1}{36l_d^2}\sum_{i=1}^N p_i E^{(i)}[\|g_{t,K-1}^{(i)}\|^2] - \frac{1}{3l_d}\sum_{i=1}^N p_i E^{(i)}[f_i(x_{t,K-1}^{(i)}) - f_i(x^*)].
\end{aligned}
$$

Plug in (A6), and we further derive

$$
\begin{aligned}
&\sum_{i=1}^N p_i E^{(i)}[\|x_{t,K}^{(i)} - x^*\|^2] \\
&\le \sum_{i=1}^N p_i E^{(i)}[\|x_{t,K-1}^{(i)} - x^*\|^2] + \frac{1}{36l_d^2}(3\sum_{i=1}^N p_i \frac{\sigma_i^2}{b_t} + 6l_d \sum_{i=1}^N p_i E^{(i)}[f_i(x_{t,K-1}^{(i)}) - f_i(x^*) \\
&+ \langle \nabla f_i(x^*), x_{t,K-1}^{(i)} - x^*\rangle] + 3\sum_{i=1}^N p_i \|\nabla f_i(x^*)\|^2) - \frac{1}{3l_d}\sum_{i=1}^N p_i E^{(i)}[f_i(x_{t,K-1}^{(i)}) - f_i(x^*)] \\
&= \sum_{i=1}^N p_i E^{(i)}[\|x_{t,K-1}^{(i)} - x^*\|^2] + \frac{1}{12l_d^2}\sum_{i=1}^N p_i \frac{\sigma_i^2}{b_t} - \frac{1}{6l_d}\sum_{i=1}^N p_i E^{(i)}[f_i(x_{t,K-1}^{(i)}) - f_i(x^*)] \\
&+ \frac{1}{6l_d}\sum_{i=1}^N p_i E^{(i)}[\langle \pi_I(\nabla f_i(x^*)), x_{t,K-1}^{(i)} - x^*\rangle] + \frac{1}{12l_d^2}\sum_{i=1}^N p_i \|\nabla f_i(x^*)\|^2 \\
&\le \sum_{i=1}^N p_i E^{(i)}[\|x_{t,K-1}^{(i)} - x^*\|^2] + \frac{1}{12l_d^2}\sum_{i=1}^N p_i \frac{\sigma_i^2}{b_t} - \frac{1}{6l_d}\sum_{i=1}^N p_i E^{(i)}[\langle \pi_I(\nabla f_i(x^*)), x_{t,K-1}^{(i)} - x^*\rangle \\
&+ \frac{\rho_d}{2}\|x_{t,K-1}^{(i)} - x^*\|^2] + \frac{1}{12l_d^2}\sum_{i=1}^N p_i \|\nabla f_i(x^*)\|^2 + \frac{1}{6l_d}\sum_{i=1}^N p_i E^{(i)}[\langle \pi_I(\nabla f_i(x^*)), x_{t,K-1}^{(i)} - x^*\rangle] \\
&= (1 - \frac{1}{12\kappa_d})\sum_{i=1}^N p_i E^{(i)}[\|x_{t,K-1}^{(i)} - x^*\|^2] + \frac{1}{12l_d^2}\sum_{i=1}^N p_i \frac{\sigma_i^2}{b_t} + \frac{1}{12l_d^2}\sum_{i=1}^N p_i \|\nabla f_i(x^*)\|^2,
\end{aligned}
$$

where the last inequality holds due to strongly restricted convexity and $\kappa_d = \frac{l_d}{\rho_d}$. Then, iteratively, we have

$$\begin{aligned}\sum_{i=1}^{N} p_i E^{(i)}[\|x_{t,K}^{(i)} - x^*\|^2] \le (1 - \frac{1}{12\kappa_d})^K \sum_{i=1}^{N} p_i E^{(i)}[\|x_{t,0}^{(i)} - x^*\|^2] + \sum_{k=0}^{K-1} (1 - \frac{1}{12\kappa_d})^k \frac{1}{12 l_d^2} \sum_{i=1}^{N} p_i \frac{\sigma_i^2}{b_t} \\ + \sum_{k=0}^{K-1} (1 - \frac{1}{12\kappa_d})^k \frac{1}{12 l_d^2} \sum_{i=1}^{N} p_i \|\nabla f_i(x^*)\|^2 \\ \le (1 - \frac{1}{12\kappa_d})^K \sum_{i=1}^{N} p_i E^{(i)}[\|x_{t,0}^{(i)} - x^*\|^2] + \sum_{k=0}^{K-1} (1 - \frac{1}{12\kappa_d})^k \frac{1}{12 l_d^2} \sum_{i=1}^{N} p_i (\frac{\sigma_i^2}{b_t} + \|\nabla f_i(x^*)\|^2).\end{aligned}$$

Let $\psi_1 = (1+\alpha)(1 - \frac{1}{12\kappa_d})^K$ and $\xi_1 = \frac{(1+\alpha)(1-(1-\frac{1}{12\kappa_d})^K)\kappa_d}{l_d^2}$. Then, we have

$$\begin{aligned}E[\|x_{t+1} - x^*\|^2] &\le (1+\alpha)(1 - \frac{1}{12\kappa_d})^K \sum_{i=1}^{N} p_i E^{(i)}[\|x_{t,0}^{(i)} - x^*\|^2] \\ &+ \frac{(1+\alpha)(1-(1-\frac{1}{12\kappa_d})^K)\kappa_d}{l_d^2} \sum_{i=1}^{N} p_i (\frac{\sigma_i^2}{b_t} + \|\nabla f_i(x^*)\|^2) \\ &= \psi_1 \sum_{i=1}^{N} p_i E^{(i)}[\|x_{t,0}^{(i)} - x^*\|^2] + \frac{\xi_1 \sum_{i=1}^{N} p_i \sigma_i^2}{b_t} + \xi_1 \sum_{i=1}^{N} p_i \|\nabla f_i(x^*)\|^2.\end{aligned}$$

Since $x_{t,0} = x_t$, we derive the relation between $\|x_{t+1} - x^*\|^2$ and $\|x_t - x^*\|^2$,

$$E[\|x_{t+1} - x^*\|^2] \le \psi_1 E[\|x_t - x^*\|^2] + \frac{\xi_1 \sum_{i=1}^{N} p_i \sigma_i^2}{b_t} + \xi_1 \sum_{i=1}^{N} p_i \|\nabla f_i(x^*)\|^2.$$

We further set $b_t = \frac{\Gamma_1}{\omega_1^t}$ and assume Γ_1 is large enough such that

$$v := \frac{\xi_1 \sum_{i=1}^{N} p_i \sigma_i^2}{\Gamma_1} \le \delta_1 \|x_0 - x^*\|^2,$$

where δ_1 is a positive constant and will be set later.

We now use mathematical induction to prove that there exists a $\theta_1 \in (0,1)$ such that the following inequality holds.

$$E[\|x_t - x^*\|^2] \le \theta_1^t E[\|x_0 - x^*\|^2] + \frac{\xi_1}{1-\psi_1} \sum_{i=1}^{N} p_i \|\nabla f_i(x^*)\|^2.$$

When $t = 0$, the above inequality is true. Now we assume that for $k = t$, it holds. Then, for $k = t+1$, we have

$$\begin{aligned}E[\|x_{t+1} - x^*\|^2] &\le \psi_1 E[\|x_t - x^*\|^2] + \frac{\xi_1 \sum_{i=1}^{N} p_i \sigma_i^2}{b_t} + \xi_1 \sum_{i=1}^{N} p_i \|\nabla f_i(x^*)\|^2 \\ &\le \psi_1 E[\|x_t - x^*\|^2] + \omega_1^t \delta_1 \|x_0 - x^*\|^2 + \xi_1 \sum_{i=1}^{N} p_i \|\nabla f_i(x^*)\|^2 \\ &\le (\psi_1 \theta_1^t + \delta_1 \omega_1^t) E[\|x_0 - x^*\|^2] + (\frac{\psi_1}{1-\psi_1} + 1)\xi_1 \sum_{i=1}^{N} p_i \|\nabla f_i(x^*)\|^2 \\ &\le (\psi_1 \theta_1^t + \delta_1 \omega_1^t) E[\|x_0 - x^*\|^2] + \frac{\xi_1}{1-\psi_1} \sum_{i=1}^{N} p_i \|\nabla f_i(x^*)\|^2.\end{aligned}$$

We now find an appropriate value for θ_1. Let $\theta_1 = \omega_1 = \psi_1 + \delta_1$, we have $\psi_1\theta_1^t + \delta_1\omega_1^t = \theta_1^{t+1}$, and then further obtain

$$E[\|x_{t+1} - x^*\|^2] \le \theta_1^{t+1} E[\|x_0 - x^*\|^2] + \frac{\xi_1}{1-\psi_1}\sum_{i=1}^{N} p_i\|\nabla f_i(x^*)\|^2$$
$$\le \theta_1^{t+1} E[\|x_0 - x^*\|^2] + \frac{\xi_1\mathcal{B}^2}{1-\psi_1}\|\nabla f(x^*)\|^2.$$

Furthermore, there exists a large $\Gamma_1 \ge \frac{\xi_1\sum_{i=1}^{N} p_i\sigma_i^2}{\delta_1\|x_0-x^*\|^2}$, such that $\delta_1 = \alpha(1 - \frac{1}{12\kappa_d})^K$. Then we have $\theta_1 = \psi_1 + \delta_1 = (1+2\alpha)(1 - \frac{1}{12\kappa_d})^K$. If we require $\theta_1 < 1$, (and we also set $\alpha = 2\sqrt{\tau^*}/\sqrt{\tau - \tau^*}$), we can derive the restriction on sparse parameter $\tau \ge (16(12\kappa_d - 1)^2 + 1)\tau^*$. □

Appendix A.5. Proof of Corollary 2

Proof of Corollary 2. In the next stage, we use a previous upper bound for $E[\|x_T - x^*\|^2]$ and l_d-restricted strongly smooth conditions to establish epoch-based convergence of $f(x_T) - f(x^*)$.

We first use l_s-restricted strongly smooth conditions and $\langle a, b\rangle \le \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$ and obtain:

$$f(x_T) \le f(x^*) + \langle\nabla f(x^*), x_T - x^*\rangle + \frac{l_d}{2}\|x_T - x^*\|^2$$
$$= f(x^*) + (\langle\nabla f(x^*), x_T - x^*\rangle) + \frac{l_d}{2}\|x_T - x^*\|^2$$
$$\le f(x^*) + \frac{1}{2l_d}\|(\nabla f(x^*))\|^2 + \frac{l_d}{2}\|x_T - x^*\|^2 + \frac{l_d}{2}\|x_T - x^*\|^2$$
$$= f(x^*) + \frac{1}{2l_d}\|(\nabla f(x^*))\|^2 + l_d\|x_T - x^*\|^2.$$

Take the expectation on both sides,

$$E[f(x_T) - f(x^*)] = \frac{1}{2l_d}\|(\nabla f(x^*))\|^2 + l_d E[\|x_T - x^*\|^2].$$

From the upper bound of $E[\|x_T - x^*\|^2]$,

$$E[\|x_T - x^*\|^2] \le \theta_1^T\|x_0 - x^*\|^2 + \frac{\xi_1\mathcal{B}^2}{1-\psi_1}\|\nabla f(x^*)\|^2.$$

We can obtain the final convergence result:

$$E[f(x_T) - f(x^*)] \le \frac{1}{2l_d}\|(\nabla f(x^*))\|^2 + l_d E[\|x_T - x^*\|^2]$$
$$\le \theta_1^T l_d\|x_0 - x^*\|^2 + (\frac{\xi_1\mathcal{B}^2 l_d}{1-\psi_1} + \frac{1}{2l_d})\|\nabla f(x^*)\|^2 = \theta_1^T\Delta_1 + g_2(x^*),$$

where $\Delta_1 = l_d\|x_0 - x^*\|^2$, $g_2(x^*) = (\frac{\xi_1\mathcal{B}^2 l_d}{1-\psi_1} + \frac{1}{2l_d})\|\nabla f(x^*)\|^2 = O(\|\nabla f(x^*)\|^2)$. □

Appendix A.6. Proof of Theorem 2

Proof. For the FedIter-HT Algorithm, we also begin with

$$E[\|x_{t+1} - x^*\|^2] = E[\|\mathcal{H}_\tau(\sum_{i=1}^{N} p_i x_{t,K}^{(i)}) - x^*\|^2] \le (1+\alpha)\sum_{i=1}^{N} p_i E^{(i)}[\|x_{t,K}^{(i)} - x^*\|^2].$$

This time we calculate the stochastic gradient on support, which is different from the analysis of the Fed-HT Algorithm. We also split the stochastic gradient on support into three terms,

$$
\begin{aligned}
&\sum_{i=1}^{N} p_i E^{(i)}[\|\pi_{\mathcal{I}^{(i)}}(g_{t,K-1}^{(i)})\|^2] \\
&= \sum_{i=1}^{N} p_i E^{(i)}[\|\pi_{\mathcal{I}^{(i)}}(g_{t,K-1}^{(i)} - \nabla f_i(x_{t,K-1}^{(i)}) + \nabla f_i(x_{t,K-1}^{(i)}) - \nabla f_i(x^*) + \nabla f_i(x^*))\|^2] \\
&\leq 3\sum_{i=1}^{N} p_i E^{(i)}[\|\pi_{\mathcal{I}^{(i)}}(g_{t,K-1}^{(i)} - \nabla f_i(x_{t,K-1}^{(i)}))\|^2] + 3\sum_{i=1}^{N} p_i E^{(i)}[\|\pi_{\mathcal{I}^{(i)}}(\nabla f_i(x_{t,K-1}^{(i)}) - \nabla f_i(x^*))\|^2] \\
&+ 3\sum_{i=1}^{N} p_i \|\pi_{\mathcal{I}^{(i)}}(\nabla f_i(x^*))\|^2 \\
&\leq 3\sum_{i=1}^{N} p_i \frac{\sigma_i^2}{b_t} + 6l_s \sum_{i=1}^{N} p_i E^{(i)}[(f_i(x_{t,K-1}^{(i)}) - f_i(x^*) + \langle \pi_{\mathcal{I}^{(i)}}(\nabla f_i(x^*)), x_{t,K-1}^{(i)} - x^* \rangle)] \\
&+ 3\sum_{i=1}^{N} p_i \|\pi_{\mathcal{I}^{(i)}}(\nabla f_i(x^*))\|^2, \qquad \text{(A7)}
\end{aligned}
$$

where the last inequality holds due to bounded variance on the support assumption and the inequality $\|\pi_{\mathcal{I}^{(i)}}(\nabla f_i(x_t) - \nabla f_i(x^*))\|^2 \leq 2l_s(f_i(x_t) - f_i(x^*) + \langle \pi_{\mathcal{I}^{(i)}}(\nabla f_i(x^*)), x_t - x^* \rangle)$.

Next, we want to build the connection of $\sum_{i=1}^{N} p_i E^{(i)}[\|x_{t,K}^{(i)} - x^*\|^2]$ and $\sum_{i=1}^{N} p_i E^{(i)}$ $[\|x_{t,K-1}^{(i)} - x^*\|^2]$. Let $\gamma_t = \frac{1}{6l_s}$. Consider the inner loop iteration,

$$
\begin{aligned}
\sum_{i=1}^{N} p_i E^{(i)}[\|x_{t,K}^{(i)} - x^*\|^2] &= \sum_{i=1}^{N} p_i E^{(i)}[\|\mathcal{H}_\tau(x_{t,K-1}^{(i)} - \frac{1}{6l_s}\pi_{\mathcal{I}^{(i)}}(g_{t,K-1}^{(i)})) - x^*\|^2] \\
&\leq (1+\alpha)\sum_{i=1}^{N} p_i E^{(i)}[\|x_{t,K-1}^{(i)} - \frac{1}{6l_s}\pi_{\mathcal{I}^{(i)}}(g_{t,K-1}^{(i)}) - x^*\|^2].
\end{aligned}
$$

Further deriving from the above result yields

$$
\begin{aligned}
&\sum_{i=1}^{N} p_i E^{(i)}[\|x_{t,K-1}^{(i)} - \frac{1}{6l_s}\pi_{\mathcal{I}^{(i)}}(g_{t,K-1}^{(i)}) - x^*\|^2] \\
&\leq (1 - \frac{1}{12\kappa_s})\sum_{i=1}^{N} p_i E^{(i)}[\|x_{t,K-1}^{(i)} - x^*\|^2] + \frac{1}{12l_s^2}\sum_{i=1}^{N} p_i \frac{\sigma_i^2}{b_t} + \frac{1}{12l_s^2}\sum_{i=1}^{N} p_i \|\pi_{\mathcal{I}^{(i)}}(\nabla f_i(x^*))\|^2,
\end{aligned}
$$

and then we can have

$$
\begin{aligned}
\sum_{i=1}^{N} p_i E^{(i)}[\|x_{t,K}^{(i)} - x^*\|^2] &\leq (1+\alpha)(1 - \frac{1}{12\kappa_s})\sum_{i=1}^{N} p_i E^{(i)}[\|x_{t,K-1}^{(i)} - x^*\|^2] \\
&+ \frac{(1+\alpha)(1 - (1 - \frac{1}{12\kappa_s})^K)\kappa_s}{l_s^2}(\sum_{i=1}^{N} p_i \frac{\sigma_i^2}{b_t} + \|\pi_{\mathcal{I}^{(i)}}(\nabla f_i(x^*))\|^2)
\end{aligned}
$$

$$
\begin{aligned}
E[\|x_{t+1} - x^*\|^2] &\leq (1+\alpha)^2(1 - \frac{1}{12\kappa_s})^K \sum_{i=1}^{N} p_i E^{(i)}[\|x_{t,0}^{(i)} - x^*\|^2] \\
&+ \frac{(1+\alpha)^2(1 - (1 - \frac{1}{12\kappa_s})^K)\kappa_s}{l_s^2}(\sum_{i=1}^{N} p_i \frac{\sigma_i^2}{b} + \|\pi_{\mathcal{I}^{(i)}}(\nabla f_i(x^*))\|^2).
\end{aligned}
$$

Similarly, we have the following result:

$$E[\|x_{t+1} - x^*\|^2] \leq \theta_2^{t+1} E[\|x_0 - x^*\|^2] + \frac{\xi_2 \mathcal{B}^2}{1 - \psi_2} \sum_{i=1}^{N} p_i \|\pi_{\mathcal{I}^{(i)}}(\nabla f_i(x^*))\|^2,$$

where $\theta_2 = (1+2\alpha)^2(1 - \frac{1}{12\kappa_s})^K$, $\xi_2 = \frac{(1+\alpha)^2(1-(1-\frac{1}{12\kappa_s})^K)\kappa_s}{l_s^2}$, $\psi_2 = (1+\alpha)^2(1 - \frac{1}{12\kappa_s})^K$ and $b_t = \frac{\Gamma_2}{\omega_2^t}$. Furthermore, there exists a large $\Gamma_2 \geq \frac{\xi_2 \mathcal{B}^2 \sum_{i=1}^{N} p_i \sigma_i^2}{\delta_2 \|x_0 - x^*\|^2}$, such that $\delta_2 = (2\alpha + 3\alpha^2)(1 - \frac{1}{12\kappa_s})^K$. Therefore, we have $\omega_2 = \theta_2 = \psi_2 + \delta_2 = (1+2\alpha)^2(1 - \frac{1}{12\kappa_s})^K < 1$. Then, we can derive the restriction on sparse parameter $\tau \geq (\frac{16}{(\sqrt{\frac{12\kappa_s}{12\kappa_s - 1}} - 1)^2} + 1)\tau^*$. □

Appendix A.7. Proof of Corollary 4

Proof. In the next stage, we use the previous upper bound for $E[\|x_T - x^*\|^2]$ and l_s-restricted strongly smooth conditions to establish epoch-based convergence of $f(x_T) - f(x^*)$.

We first use l_s-restricted strongly smooth conditions and $\langle a, b\rangle \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$ and obtain:

$$\begin{aligned} f(x_T) &\leq f(x^*) + \langle \nabla f(x^*), x_T - x^*\rangle + \frac{l_s}{2}\|x_T - x^*\|^2 \\ &= f(x^*) + \pi_{\tilde{\mathcal{I}}}(\langle \nabla f(x^*), x_T - x^*\rangle) + \frac{l_s}{2}\|x_T - x^*\|^2 \\ &\leq f(x^*) + \frac{1}{2l_s}\|\pi_{\tilde{\mathcal{I}}}(\nabla f(x^*))\|^2 + \frac{l_s}{2}\|x_T - x^*\|^2 + \frac{l_s}{2}\|x_T - x^*\|^2 \\ &= f(x^*) + \frac{1}{2l_s}\|\pi_{\tilde{\mathcal{I}}}(\nabla f(x^*))\|^2 + l_s\|x_T - x^*\|^2. \end{aligned}$$

Take the expectation on both sides,

$$E[f(x_T) - f(x^*)] \leq \frac{1}{2l_s}\|\pi_{\tilde{\mathcal{I}}}(\nabla f(x^*))\|^2 + l_s E[\|x_T - x^*\|^2].$$

From the upper bound of $E[\|x_T - x^*\|^2]$,

$$E[\|x_T - x^*\|^2] \leq \theta_2^T \|x_0 - x^*\|^2 + \frac{\xi_2 \mathcal{B}^2}{1 - \psi_2}\|\pi_{\tilde{\mathcal{I}}}(\nabla f(x^*))\|^2.$$

Then, we can obtain the final convergence result:

$$\begin{aligned} E[f(x_T) - f(x^*)] &\leq \theta_2^T l_s \|x_0 - x^*\|^2 + (\frac{\xi_2 \mathcal{B}^2 l_s}{1 - \psi_2} + \frac{1}{2l_s})\|\nabla f(x^*)\|^2 \\ &= \theta_2^T \Delta_2 + g_4(x^*) \end{aligned}$$

where $\Delta_2 = l_s\|x_0 - x^*\|^2$, $g_4(x^*) = (\frac{\xi_2 \mathcal{B}^2 l_s}{1-\psi_2} + \frac{1}{2l_s})\|\nabla f(x^*)\|^2 = O(\pi_{\tilde{\mathcal{I}}}(\|\nabla f(x^*)\|^2))$. □

References

1. Mohamed, S.; Heller, K.; Ghahramani, Z. Bayesian and l1 approaches to sparse unsupervised learning. *arXiv* **2011**, arXiv:1106.1157.
2. Quattoni, A.; Collins, M.; Darrell, T. Transfer learning for image classification with sparse prototype representations. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
3. Lu, X.; Huang, Z.; Yuan, Y. MR image super-resolution via manifold regularized sparse learning. *Neurocomputing* **2015**, *162*, 96–104. [CrossRef]
4. Chen, K.; Che, H.; Li, X.; Leung, M.F. Graph non-negative matrix factorization with alternative smoothed L_0 regularizations. *Neural Comput. Appl.* **2022**, 1–15 . [CrossRef]

5. Ravishankar, S.; Bresler, Y. Learning sparsifying transforms. *IEEE Trans. Signal Process.* **2012**, *61*, 1072–1086. [CrossRef]
6. Tropp, J.A.; Gilbert, A.C. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inf. Theory* **2007**, *53*, 4655–4666. [CrossRef]
7. Boufounos, S.; Raj, P.; Bahmani, S.; Boufounos, P.; Raj, B. Greedy Sparsity-Constrained Optimization. Available online: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.365.3874&rep=rep1&type=pdf (accessed on 1 July 2022).
8. Jalali, A.; Johnson, C.; Ravikumar, P. On learning discrete graphical models using greedy methods. *Adv. Neural Inf. Process. Syst.* **2011**, *24*, 1935–1943.
9. Mallat, S.G.; Zhang, Z. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.* **1993**, *41*, 3397–3415. [CrossRef]
10. Pati, Y.C.; Rezaiifar, R.; Krishnaprasad, P.S. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In Proceedings of the 27th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 1–3 November 1993; pp. 40–44.
11. Needell, D.; Tropp, J.A. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal.* **2009**, *26*, 301–321. [CrossRef]
12. Foucart, S. Hard thresholding pursuit: An algorithm for compressive sensing. *SIAM J. Numer. Anal.* **2011**, *49*, 2543–2563. [CrossRef]
13. Blumensath, T.; Davies, M.E. Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.* **2009**, *27*, 265–274. [CrossRef]
14. Jain, P.; Tewari, A.; Kar, P. On iterative hard thresholding methods for high-dimensional m-estimation. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 685–693.
15. Nguyen, N.; Needell, D.; Woolf, T. Linear convergence of stochastic iterative greedy algorithms with sparse constraints. *IEEE Trans. Inf. Theory* **2017**, *63*, 6869–6895. [CrossRef]
16. Bahmani, S.; Raj, B.; Boufounos, P.T. Greedy sparsity-constrained optimization. *J. Mach. Learn. Res.* **2013**, *14*, 807–841.
17. Zhou, P.; Yuan, X.; Feng, J. Efficient stochastic gradient hard thresholding. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 1988–1997.
18. Li, X.; Zhao, T.; Arora, R.; Liu, H.; Haupt, J. Stochastic variance reduced optimization for nonconvex sparse learning. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 917–925.
19. Shen, J.; Li, P. A tight bound of hard thresholding. *J. Mach. Learn. Res.* **2017**, *18*, 7650–7691.
20. Natarajan, B.K. Sparse approximate solutions to linear systems. *SIAM J. Comput.* **1995**, *24*, 227–234. [CrossRef]
21. Wahlsten, D.; Metten, P.; Phillips, T.J.; Boehm, S.L.; Burkhart-Kasch, S.; Dorow, J.; Doerksen, S.; Downing, C.; Fogarty, J.; Rodd-Henricks, K.; et al. Different data from different labs: Lessons from studies of gene–environment interaction. *J. Neurobiol.* **2003**, *54*, 283–311.
22. Kavvoura, F.K.; Ioannidis, J.P. Methods for meta-analysis in genetic association studies: A review of their potential and pitfalls. *Hum. Genet.* **2008**, *123*, 1–14.
23. Lee, Y.G.; Jeong, W.S.; Yoon, G. Smartphone-based mobile health monitoring. *Telemed. E-Health* **2012**, *18*, 585–590. [CrossRef]
24. Qin, Z.; Fan, J.; Liu, Y.; Gao, Y.; Li, G.Y. Sparse representation for wireless communications: A compressive sensing approach. *IEEE Signal Process. Mag.* **2018**, *35*, 40–58. [CrossRef]
25. McMahan, H.B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. *arXiv* **2016**, arXiv:1602.05629.
26. Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol. (TIST)* **2019**, *10*, 1–19. [CrossRef]
27. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A.N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. Advances and open problems in federated learning. *Found. Trends® Mach. Learn.* **2021**, *14*, 1–210. [CrossRef]
28. Donoho, D.L. Compressed sensing. *IEEE Trans. Inf. Theory* **2006**, *52*, 1289–1306. [CrossRef]
29. Patterson, S.; Eldar, Y.C.; Keidar, I. Distributed compressed sensing for static and time-varying networks. *IEEE Trans. Signal Process.* **2014**, *62*, 4931–4946. [CrossRef]
30. Lafond, J.; Wai, H.T.; Moulines, E. D-FW: Communication efficient distributed algorithms for high-dimensional sparse optimization. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 4144–4148.
31. Wang, J.; Kolar, M.; Srebro, N.; Zhang, T. Efficient distributed learning with sparsity. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 3636–3645.
32. Lin, Y.; Han, S.; Mao, H.; Wang, Y.; Dally, W.J. Deep gradient compression: Reducing the communication bandwidth for distributed training. *arXiv* **2017**, arXiv:1712.01887.
33. Shi, S.; Wang, Q.; Zhao, K.; Tang, Z.; Wang, Y.; Huang, X.; Chu, X. A distributed synchronous SGD algorithm with global top-k sparsification for low bandwidth networks. In Proceedings of the 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), Dallas, TX, USA, 7–10 July 2019; pp. 2238–2247.
34. Hsu, T.M.H.; Qi, H.; Brown, M. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv* **2019**, arXiv:1909.06335.

35. Karimireddy, S.P.; Kale, S.; Mohri, M.; Reddi, S.J.; Stich, S.U.; Suresh, A.T. SCAFFOLD: Stochastic controlled averaging for on-device federated learning. *arXiv* **2019**, arXiv:1910.06378.
36. Reddi, S.; Charles, Z.; Zaheer, M.; Garrett, Z.; Rush, K.; Konečnỳ, J.; Kumar, S.; McMahan, H.B. Adaptive Federated Optimization. *arXiv* **2020**, arXiv:2003.00295.
37. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smith, V. Federated optimization in heterogeneous networks. *arXiv* **2018**, arXiv:1812.06127.
38. Bernstein, J.; Zhao, J.; Azizzadenesheli, K.; Anandkumar, A. signSGD with majority vote is communication efficient and fault tolerant. *arXiv* **2018**, arXiv:1810.05291.
39. Sattler, F.; Wiedemann, S.; Müller, K.R.; Samek, W. Robust and communication-efficient federated learning from non-iid data. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *31*, 3400–3413. [CrossRef]
40. Li, C.; Li, G.; Varshney, P.K. Communication-efficient federated learning based on compressed sensing. *IEEE Internet Things J.* **2021**, *8*, 15531–15541. [CrossRef]
41. Han, P.; Wang, S.; Leung, K.K. Adaptive gradient sparsification for efficient federated learning: An online learning approach. In Proceedings of the 2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS), Singapore, 29 November–1 December 2020; pp. 300–310.
42. Yuan, H.; Zaheer, M.; Reddi, S. Federated composite optimization. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 18–24 July 2021; pp. 12253–12266.
43. Agarwal, A.; Negahban, S.; Wainwright, M.J. Fast Global Convergence Rates of Gradient Methods for High-Dimensional Statistical Recovery. Available online: https://proceedings.neurips.cc/paper/2010/file/7cce53cf90577442771720a370c3c723-Paper.pdf (accessed on 1 July 2022).
44. Li, X.; Arora, R.; Liu, H.; Haupt, J.; Zhao, T. Nonconvex sparse learning via stochastic optimization with progressive variance reduction. *arXiv* **2016**, arXiv:1605.02711.
45. Wang, L.; Gu, Q. Differentially Private Iterative Gradient Hard Thresholding for Sparse Learning. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019.
46. Loh, P.L.; Wainwright, M.J. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.* **2015**, *16*, 559–616.
47. Kogan, S.; Levin, D.; Routledge, B.R.; Sagi, J.S.; Smith, N.A. Predicting risk from financial reports with regression. In Proceedings of the Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Boulder, CO, USA, 31 May–5 June 2009; Association for Computational Linguistics: Boulder, CO, USA, 2009; pp. 272–280.
48. Lewis, D.D.; Yang, Y.; Rose, T.G.; Li, F. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.* **2004**, *5*, 361–397.