Data Encoding for Byzantine-Resilient Distributed Optimization

Deepesh Data, Linqi Song, *Member, IEEE*, and Suhas Diggavi, *Fellow, IEEE*

***Abstract*—We study distributed optimization in the presence of Byzantine adversaries, where both data and computation are distributed among m worker machines, t of which may be corrupt. The compromised nodes may collaboratively and arbitrarily deviate from their pre-specified programs, and a designated (master) node iteratively computes the model/parameter vector for *generalized linear models*. In this work, we primarily focus on two iterative algorithms: *Proximal Gradient Descent* (PGD) and *Coordinate Descent* (CD). Gradient descent (GD) is a special case of these algorithms. PGD is typically used in the data-parallel setting, where data is partitioned across different samples, whereas, CD is used in the model-parallelism setting, where data is partitioned across the parameter space. At the core of our solutions to both these algorithms is a method for Byzantine-resilient matrix-vector (MV) multiplication; and for that, we propose a method based on data encoding and error correction over real numbers to combat adversarial attacks. We can tolerate up to $t \leq \lfloor \frac{m-1}{2} \rfloor$ corrupt worker nodes, which is information-theoretically optimal. We give deterministic guarantees, and our method does not assume any probability distribution on the data. We develop a *sparse* encoding scheme which enables computationally efficient data encoding and decoding. We demonstrate a trade-off between the corruption threshold and the resource requirements (storage, computational, and communication complexity). As an example, for $t \leq \frac{m}{3}$, our scheme incurs only a *constant* overhead on these resources, over that required by the plain distributed PGD/CD algorithms which provide no adversarial protection. To the best of our knowledge, ours is the first paper that connects MV multiplication with CD and designs a specific encoding matrix for MV multiplication whose structure we can leverage to make CD secure against adversarial attacks. Our encoding scheme extends *efficiently* to (i) the data streaming model, in which data samples come in an online fashion and are encoded as they arrive, and (ii) making *stochastic gradient descent* (SGD) Byzantine-resilient. In the end, we give experimental results to show the efficacy of our proposed schemes.**

***Index Terms* — Distributed optimization, (proximal) gradient descent, coordinate descent, Byzantine adversary, data encoding and error correction over reals.**

This paper was presented in parts at the IEEE Allerton 2018 (as an invited talk) [1], and ISIT 2019 [2], [3]. The work of Deepesh Data and Suhas Diggavi was partially supported by the Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196, by the UC-NL grant LFR-18-548554, and by the NSF award 1740047. The work of Linqi Song was partially supported by the NSF awards 1527550, 1514531, by the City University of Hong Kong grant 7200594, and by the Hong Kong RGC ECS 21212419. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

Deepesh Data and Suhas Diggavi are with the University of California, Los Angeles (UCLA), Los Angeles, CA 90095, USA (deepesh.data@gmail.com, suhasdiggavi@ucla.edu).

Linqi Song is with the City University of Hong Kong, Hong Kong. Part of this work was done when Linqi Song was at UCLA (email: linqi.song@cityu.edu.hk).

I. Introduction

Map-reduce architecture [4] is implemented in many distributed learning tasks, where there is one designated machine (called the master) that computes the model iteratively, based on the inputs from the worker machines at each iteration, typically using descent techniques, like (proximal) gradient descent, coordinate descent, stochastic gradient descent, the Newton's method, etc. The worker nodes perform the required computations using local data, distributed to the nodes [5]. Several other architectures, including having no hierarchy among the nodes have been explored [6].

In several applications of distributed learning, including the Internet of Battlefield Things (IoBT) [7], federated optimization [8], the recruited worker nodes might be partially trusted with their computation. Therefore, an important question is whether we can reliably perform distributed computation, taking advantage of partially trusted worker nodes. These Byzantine adversaries can collaborate and arbitrarily deviate from their pre-specified programs. The problem of distributed computation with Byzantine adversaries has a long history [9], and there has been recent interest in applying this computational model to large-scale distributed learning [10]–[12].

In this paper, we study Byzantine-tolerant distributed optimization to learn a regularized *generalized linear model* (GLM) (e.g., linear/ridge regression, logistic regression, Lasso, SVM dual, constrained minimization, etc.). We consider two frameworks for distributed optimization: (i) *data-parallelism* architecture, where data points are distributed across different worker nodes, and in each iteration, they all parallelly compute gradients on their local data and master aggregates them to update the parameter vector using gradient descent (GD) [13]–[15]; and (ii) *model-parallelism* architecture, where data points are partitioned across features, and several worker nodes work in parallel, updating different subsets of coordinates of the model/parameter vector through *coordinate descent* (CD) [16]–[18]. Note that GD requires full gradients to update the parameter vector; and if full gradients are too costly to compute, we can reduce the per-

iteration cost by using CD,[1] which also has been shown to be very effective for solving generalized linear models, and is particularly widely used for sparse logistic regression, SVM, and Lasso [16]. Given its simplicity and effectiveness, CD can be chosen over GD in such applications [19]. Computing gradients in the presence of Byzantine adversaries has been recently studied [10]–[12], [20]–[32], and we discuss them in detail Section III where we also put our work in context. However, as far as we know, making CD robust to Byzantine adversaries has not received much attention, and to the best of our knowledge, ours is the first paper that studies CD against Byzantine attacks and provides an efficient solution for that.

A. Our Contributions

We propose Byzantine-resilient distributed optimization algorithms both for PGD and CD based on data encoding and error correction (over real numbers). As mentioned above, there have been several papers that provide different methods for gradient computation in the presence of Byzantine adversaries, however, our proposed algorithm differs from them in one or more of the following aspects: (i) it does not make statistical assumptions on the data or Byzantine attack patterns; (ii) it can tolerate up to a constant fraction ($< 1/2$) of the worker nodes being Byzantine, which is information-theoretically optimal; and (iii) it enables a trade-off (in terms of storage and computation/communication overhead at the master and the worker nodes) with Byzantine adversary tolerance, without compromising the efficiency at the master node. We give the same guarantees for CD also.

First we design a coding scheme for distributed matrix-vector (MV) multiplication, specifically, for operating in the presence of Byzantine adversaries, and use that in both our algorithms for PGD and CD to learn GLMs. Note that the connection of MV multiplication with gradient computation is straightforward and has been known for some time (see, for example, [33], [34]), however, it is not clear whether we can use MV multiplication methods for CD also. Indeed, since each CD update has a different requirement than that of gradient computation, a general-purpose algorithm for MV multiplication may not be applicable for CD. One distinction is that in gradient computation, we only need to encode the data to compute the MV multiplication, whereas, in CD, in addition to data encoding, since workers update few coordinates of different parts of the parameter vector in parallel, we need to encode the parameter vector as well for master to be able to decode that. In this paper, we design our encoding matrix for MV multiplication in such a way that it is sparse and has a regular structure of non-zero entries (see (11) for the encoding matrix for any worker), which makes it applicable for CD too. This leads to efficient solutions for both PGD and CD, which are our main focus in this paper.

Inspired from the real-error correction (or sparse reconstruction) problem [35], we develop efficient encoding/decoding procedures for MV multiplication, where we encode the data matrix and distribute it to the m worker nodes, and to recover the MV product at the master, we reduce the decoding problem to the sparse reconstruction or real-error correction problem [35]. Note that in PGD, we only need to encode the data, whereas, in CD, we also need to encode the parameter vector, and our coding scheme should facilitate the requirement that the update on a small *fraction* of the encoded parameter vector should affect only a small *fraction* of the original parameter vector. This is a non-trivial requirement, and our coding scheme for MV multiplication is designed in such a way that it supports this requirement in an efficient manner; see Section II-B for a description on plain distributed CD, Section II-E for our approach to making CD robust to Byzantine attacks, and Section V for a complete solution for Byzantine-resilient CD. In the context of PGD/CD, for decoding, the master node processes the inputs from the worker nodes, either to compute the true gradient in the case of PGD or to facilitate the computation at the worker nodes in the case of CD. We take a two-round approach in each iteration of both these algorithms. Our main results are summarized in Theorem 1 (on page 6) for PGD and Theorem 2 (on page 8) for CD, and demonstrate a trade-off between the Byzantine resilience (in terms of the number of adversarial nodes) and the resource requirement (storage, computational, and communication complexity). As an example, for $t \leq \frac{m}{3}$, our scheme incurs only a *constant* overhead on these resources, over that required by the plain distributed PGD and CD algorithms which provide no adversarial protection. Our coding schemes can handle both Byzantine attacks and missing updates (*e.g.,* caused by delay or asynchrony of worker nodes). Our encoding process is also efficient. Though data encoding is a one-time process, it has to be efficient to harness the advantage of distributed computation. We design a sparse encoding process, based on real-error correction, which enables efficient encoding, and the worker nodes encode data using the sparse structure. This allows encoding with storage redundancy[2] of $\frac{2m}{m-2t}$ (which is a constant, *even* if t is a constant ($< \frac{1}{2}$) fraction of m), and a one-time total computation cost for encoding is $O((1+2t)nd)$. Note that the time for data encoding is a factor of $(1+2t)$ (where t is the corruption threshold) more than the time required for plain data distribution which is $O(nd)$, the size of the data matrix.

We extend our encoding scheme in a couple of important ways: first, to make the stochastic gradient descent (SGD) algorithm Byzantine-resilient without compromising much on the resource requirements; and second, to handle streaming data efficiently, where data points arrives one by one (and we encode them as they arrive), rather than being available at the beginning of the computation; we also give few more applications of our method. For the streaming model, more specifically, our encoding requires the same amount of time, irrespective of whether we encode all the data at once, or

[1] Alternatively, we can also use SGD to reduce the per-iteration cost, and we give a method for making SGD Byzantine-resilient in Section VI-A.

[2] Storage redundancy is defined as the ratio of the size of the encoded matrix and the size of the raw data matrix.

we get data points one by one (or in batches) and we encode them as they arrive. This setting encompasses a more realistic scenario, in which we design our coding scheme with the initial set of data points and distribute the encoded data among the workers. Later on, when we get some more samples, we can easily incorporate them into our existing encoded setup. See Section VI for details on these extensions.

B. Paper Organization

We present our problem formulation, description of the plain distributed PGD and CD algorithms, and the high-level ideas of our Byzantine-resilient algorithms for both PGD and CD along-with our main results in Section II. We give detailed related work in Section III. We present our full coding schemes for MV multiplication and also for gradient computation for PGD along-with a complete analysis of their resource requirements in Section IV. In Section V, we provide a complete solution to CD. In Section VI, we show how our method can be extended to SGD and to the data streaming model. We also discuss applicability of our method to a few more important applications in that section. In Section VII, we show numerical results of our method: we show the efficiency of our method for both gradient descent (GD) and coordinate descent (CD) by running them to solve linear regression on two datasets (moderate and large) and plotting the running time with varying number of corrupt worker nodes (up to <1/2 fraction).

C. Notation

We denote vectors by bold small letters (e.g., $\mathbf{x}, \mathbf{y}, \mathbf{z}$, etc.) and matrices by bold capital letters (e.g., $\mathbf{A}, \mathbf{F}, \mathbf{S}, \mathbf{X}$, etc.). We denote the amount of storage required by a matrix $\mathbf{X}$ by $|\mathbf{X}|$. For any positive integer $n \in \mathbb{N}$, we denote the set $\{1, 2, \ldots, n\}$ by $[n]$. For $n_1, n_2 \in \mathbb{N}$, where $n_1 \leq n_2$, we write $[n_1 : n_2]$ to denote the set $\{n_1, n_1 + 1, \ldots, n_2\}$. For any vector $\mathbf{u} \in \mathbb{R}^n$ and any set $\mathcal{S} \subset [n]$, we write $\mathbf{u}_\mathcal{S}$ to denote the $|\mathcal{S}|$-length vector, which is the restriction of $\mathbf{u}$ to the coordinates in the set $\mathcal{S}$. The support of a vector $\mathbf{u} \in \mathbb{R}^n$ is defined by $\mathsf{supp}(\mathbf{u}) := \{i \in [n] : u_i \neq 0\}$. We say that a vector $\mathbf{u} \in \mathbb{R}^n$ is t-sparse if $|\mathsf{supp}(\mathbf{u})| \leq t$. While stating our results, we assume that performing the basic arithmetic operations (addition, subtraction, multiplication, and division) on real numbers takes unit time.

II. Problem Setting and Our Results

Given a dataset consisting of n labelled data points $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i \in [n]$, we want to learn a model/parameter vector $\mathbf{w} \in \mathbb{R}^d$, which is a minimizer of the following *empirical risk minimization* problem:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \left(\left(\frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{w}) \right) + h(\mathbf{w}) \right), \tag{1}$$

where $f_i(\mathbf{w})$, $i = 1, 2, \ldots, n$, denotes the risk associated with the i'th data point with respect to $\mathbf{w}$ and $h(\mathbf{w})$ denotes a regularizer. We call $f(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{w})$ the average empirical risk associated with the n data points with respect to $\mathbf{w}$. Our main focus in this paper is on *generalized linear models* (GLM), where $f_i(\mathbf{w}) = \ell(\langle \mathbf{x}_i, \mathbf{w} \rangle; y_i)$ for some differentiable loss function ℓ. Here, each $f_i : \mathbb{R}^d \to \mathbb{R}$ is differentiable, $h : \mathbb{R}^d \to \mathbb{R}$ is convex but not necessarily differentiable, and $\langle \mathbf{x}_i, \mathbf{w} \rangle$ is the dot product of $\mathbf{x}_i$ and $\mathbf{w}$. We do not necessarily need each f_i to be convex, but we require $f(\mathbf{w})$ to be a convex function. Note that $f(\mathbf{w}) + h(\mathbf{w})$ is a convex function. In the following we study different algorithms for solving (1) to learn a GLM.

A. Proximal Gradient Descent

We can solve (1) using *Proximal Gradient Descent* (PGD). This is an iterative algorithm, in which we choose an arbitrary/random initial $\mathbf{w}_0 \in \mathbb{R}^d$, and then update the parameter vector according to the following update rule:

$$\mathbf{w}_{t+1} = \mathsf{prox}_{h, \alpha_t}(\mathbf{w}_t - \alpha_t \nabla f(\mathbf{w}_t)), \quad t = 1, 2, 3, \ldots \tag{2}$$

where α_t is the step size or the learning rate at the t'th iteration, determining the convergence behaviour. There are standard choices for it; see, for example, [36, Chapter 9]. For any h and α, the proximal operator $\mathsf{prox}_{h,\alpha} : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\mathsf{prox}_{h,\alpha}(\mathbf{w}) = \arg\min_{\mathbf{z} \in \mathbb{R}^d} \frac{1}{2\alpha} \|\mathbf{z} - \mathbf{w}\|_2^2 + h(\mathbf{z}). \tag{3}$$

Observe that if $h = 0$, then $\mathsf{prox}_{h,\alpha}(\mathbf{w}) = \mathbf{w}$ for every $\mathbf{w} \in \mathbb{R}^d$, and PGD reduces to the classical gradient descent (GD). This encompasses several important optimization problems related to learning, for which $prox$ operator has a closed form expression; some of these problems are given below.

- **Lasso.** Here $f_i(\mathbf{w}) = \frac{1}{2}(\langle \mathbf{x}_i, \mathbf{w} \rangle - y_i)^2$ and $h(\mathbf{w}) = \lambda \|\mathbf{w}\|_1$. It turns out that $\mathsf{prox}_{h,\alpha}(\mathbf{z})$ for Lasso is equal to the soft-thresholding operator $S_{\lambda\alpha}(\mathbf{z})$ [37], which, for $j \in [d]$, is defined as

$$(S_{\lambda\alpha}(\mathbf{z}))_j = \begin{cases} z_j + \lambda\alpha & \text{if } z_j < -\lambda\alpha, \\ 0 & \text{if } -\lambda\alpha \leq z_j \leq \lambda\alpha, \\ z_j - \lambda\alpha & \text{if } z_j > \lambda\alpha. \end{cases}$$

- **SVM dual.** Jaggi [38] showed an equivalence between the dual formulation of Support Vector Machines (SVM) and Lasso. Hence, SVM dual is also a special case of (1).
- **Constrained optimization.** We want to solve a constrained minimization problem $\min_{\mathbf{w} \in \mathcal{C}} f(\mathbf{w})$, where $\mathcal{C} \subseteq \mathbb{R}^d$ is a closed, convex set. Define an indicator function $I_\mathcal{C}$ for $\mathcal{C}$ as follows: $I_\mathcal{C}(\mathbf{w}) := 0$, if $\mathbf{w} \in \mathcal{C}$; and $I_\mathcal{C}(\mathbf{w}) := \infty$, otherwise. Now, observe the following equivalence

$$\min_{\mathbf{w} \in \mathcal{C}} f(\mathbf{w}) \iff \min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w}) + I_\mathcal{C}(\mathbf{w}).$$

If we solve the RHS using PGD, then it can be easily verified that the corresponding proximal operator is equal to the projection operator onto the set $\mathcal{C}$ [37]. So, the proximal gradient update step is to compute the usual gradient and then project it back onto the set $\mathcal{C}$.

- **Logistic regression.** Here f_i is the logistic function, defined as

$$f_i(\mathbf{w}) = -y_i \log\left(\frac{1}{1+e^{-u_i}}\right) - (1-y_i)\log\left(\frac{e^{-u_i}}{1+e^{-u_i}}\right),$$

  where $u_i = \langle \mathbf{x}_i, \mathbf{w}\rangle$, and $h = 0$. As noted earlier, since $h = 0$, PGD reduces to GD for logistic regression.
- **Ridge regression.** Here $f_i(\mathbf{w}) = \frac{1}{2}(\langle \mathbf{x}_i, \mathbf{w}\rangle - y_i)^2$ and $h(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|_2^2$. Since f_i's and h are differentiable, we can alternatively solve this simply using GD.

Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ denote the data matrix, whose i'th row is equal to the i'th data point $\mathbf{x}_i$. For simplicity, assume that m divides n, and let $\mathbf{X}_i$ denote the $\frac{n}{m} \times d$ matrix, whose j'th row is equal to $\mathbf{x}_{(i-1)\frac{n}{m}+j}$. In a distributed setup, all the data is distributed among m worker machines (worker i has $\mathbf{X}_i$) and master updates the parameter vector using the update rule (2). At the t'th iteration, master sends $\mathbf{w}_t$ to all the workers; worker i computes the gradient (denoted by $\nabla_i f(\mathbf{w}_t)$) on its local data and sends it to the master; master aggregates all the received m local gradients to obtain the global gradient

$$\nabla f(\mathbf{w}_t) = \frac{1}{m}\sum_{i=1}^{m} \nabla_i f(\mathbf{w}_t). \tag{4}$$

Now, master updates the parameter vector according to (2) and obtains $\mathbf{w}_{t+1}$. Repeat the process until convergence.

If full gradients are too costly to compute. Updating the parameter vector in each iteration of PGD according to (2) requires computing full gradients. This may be prohibitive in large-scale applications, where each machine in a distributed framework has a lot of data, and computing full gradients at local machines may be too expensive and becomes the bottleneck. In such scenarios, there are two alternatives to reduce this per-iteration cost: (i) *Coordinate Descent* (CD), in which we pick a few coordinates (at random), compute the partial gradient along those, and descent along those coordinates only, and (ii) *Stochastic Gradient Descent* (SGD), in which we sample a data point at random, compute the gradient on that point, and descent along that direction. These are discussed in Section II-B and Section VI-A, respectively.

B. Coordinate Descent

For the clear exposition of ideas, we focus on the non-regularized empirical risk minimization from (1) (i.e., taking $h = 0$) for learning a *generalized linear model* (GLM). This can be generalized to objectives with (non-)differentiable regularizers [16], [39]. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ denote the data matrix and $\mathbf{y} \in \mathbb{R}^n$ the corresponding label vector. To make it distinct from the last section, we denote the objective function by ϕ and write it as $\phi(\mathbf{X}\mathbf{w}; \mathbf{y})$ to emphasize that we want to learn a GLM, where the objective function depends on the data points only through their inner products with the parameter vector. Formally, we want to optimize[3]

$$\min_{\mathbf{w}\in\mathbb{R}^d} \left(\phi(\mathbf{X}\mathbf{w};\mathbf{y}) := \sum_{i=1}^{n} \ell(\langle \mathbf{x}_i, \mathbf{w}\rangle; y_i)\right). \tag{5}$$

For $\mathcal{U} \subseteq [d]$, we write $\nabla_{\mathcal{U}}\phi(\mathbf{X}\mathbf{w};\mathbf{y})$ to denote the gradient of $\phi(\mathbf{X}\mathbf{w};\mathbf{y})$ with respect to $\mathbf{w}_{\mathcal{U}}$, where $\mathbf{w}_{\mathcal{U}}$ denotes the $|\mathcal{U}|$-length vector obtained by restricting $\mathbf{w}$ to the coordinates in $\mathcal{U}$. To make the notation less cluttered, let $\phi'(\mathbf{X}\mathbf{w};\mathbf{y})$ denote the n-length vector, whose i'th entry is equal to $\ell'(\langle \mathbf{x}_i, \mathbf{w}\rangle; y_i) := \frac{\partial}{\partial u}\ell(u; y_i)|_{u=\langle \mathbf{x}_i, \mathbf{w}\rangle}$. Note that $\nabla\phi(\mathbf{X}\mathbf{w};\mathbf{y}) = \mathbf{X}^T\phi'(\mathbf{X}\mathbf{w};\mathbf{y})$ and that $\nabla_{\mathcal{U}}\phi(\mathbf{X}\mathbf{w};\mathbf{y}) = \mathbf{X}_{\mathcal{U}}^T\phi'(\mathbf{X}\mathbf{w};\mathbf{y})$, where $\mathbf{X}_{\mathcal{U}}$ denotes the $n \times |\mathcal{U}|$ matrix obtained by restricting the column indices of $\mathbf{X}$ to the elements in $\mathcal{U}$.

Coordinate descent (CD) is an iterative algorithm, where, in each iteration, we choose a set of coordinates and update only those coordinates (while keeping the other coordinates fixed). In distributed CD, we take advantage of the parallel architecture to improve the running time of (centralized) CD. In the distributed setting, we divide the data matrix vertically into m parts and store the i'th part at the i'th worker node. Concretely, assume, for simplicity, that m divides d. Let $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2\ \dots\ \mathbf{X}_m]$ and $\mathbf{w} = [\mathbf{w}_1^T\ \mathbf{w}_2^T\ \dots\ \mathbf{w}_m^T]^T$, where each $\mathbf{X}_i$ is an $n \times \frac{d}{m}$ matrix and each $\mathbf{w}_i$ is a length $\frac{d}{m}$ vector. Each worker i stores $\mathbf{X}_i$ and is responsible for updating (a few coordinates of) $\mathbf{w}_i$ – hence the terminology, model-parallelism. We store the label vector $\mathbf{y}$ at the master node. In coordinate descent, since we update only a few coordinates in each round, there are a few options on how to update these coordinates in a distributed manner:

Subset of workers: Master picks a subset $\mathcal{S} \subset [m]$ of workers and asks them to update their $\mathbf{w}_i$'s [18]. This may not be good in the adversarial setting, because if only a small subset of workers are updating their parameters, the adversary can corrupt those workers and disrupt the computation.

Subset of coordinates for all workers: All the worker nodes update only a subset of the coordinates of their local parameter vector $\mathbf{w}_i$'s. Master can (deterministically or randomly) pick a subset $\mathcal{U}$ (which may or may not be different for all workers) of $f \leq d/m$ coordinates and asks each worker to updates only those coordinates. If master picks $\mathcal{U}$ deterministically, it can cycle through and update all coordinates of the parameter vector in $\lceil d/mf \rceil$ iterations.

In Algorithm 1, we give the distributed CD algorithm with the second approach, where all worker nodes update the coordinates of their local parameter vectors for a single subset $\mathcal{U}$. We will adopt this approach in our method to make the distributed CD Byzantine-resilient. Let $r = \frac{d}{m}$. For any $i \in [m]$, let $\mathbf{w}_i = [w_{i1}\ w_{i2} \dots w_{ir}]^T$ and $\mathbf{X}_i = [\mathbf{X}_{i1}\ \mathbf{X}_{i2} \dots \mathbf{X}_{ir}]$, where $\mathbf{X}_{ij}$ is the j'th column of $\mathbf{X}_i$. For any $i \in [m]$ and $\mathcal{U} \subseteq [r]$, let $\mathbf{w}_{i\mathcal{U}}$ denote the $|\mathcal{U}|$-length vector that is obtained from $\mathbf{w}_i$ by restricting its entries to the coordinates in $\mathcal{U}$;

[3]Here we are not optimizing the *average* of loss functions – since n is a fixed number, this does not affect the solution space.

Algorithm 1 Distributed Coordinate Descent

1: **Initialize.** Each worker $i \in [m]$ starts with an arbitrary/random $\mathbf{w}_i \in \mathbb{R}^r$, where $r = \frac{d}{m}$ and, for simplicity, we assume that m divides d.
2: **while** (until the stopping criteria at master is not satisfied) **do**
3: **On each worker** $i \in [m]$**, do in parallel:**
4: Worker i computes $\mathbf{X}_i\mathbf{w}_i$ and sends it to the master node.[4]
5: Worker i receives $(\mathcal{U} \subseteq [r], \phi'(\mathbf{X}\mathbf{w};\mathbf{y}))$ from the master node.
6: Worker i updates its local parameter vector as (where $\nabla_{i\mathcal{U}}\phi(\mathbf{X}\mathbf{w};\mathbf{y}) = \mathbf{X}_{i\mathcal{U}}^T\phi'(\mathbf{X}\mathbf{w};\mathbf{y})$)

$$\mathbf{w}_{i\mathcal{U}} \leftarrow \mathbf{w}_{i\mathcal{U}} - \alpha\nabla_{i\mathcal{U}}\phi(\mathbf{X}\mathbf{w};\mathbf{y}) \tag{6}$$

while keeping the other coordinates of $\mathbf{w}_i$ unchanged, and sends the updated $\mathbf{w}_i$ to the master.
7: **At Master:**
8: Master receives $\{\mathbf{X}_i\mathbf{w}_i\}_{i\in[m]}$ from the m workers.
9: Master first computes $\mathbf{X}\mathbf{w} = \sum_{i=1}^m \mathbf{X}_i\mathbf{w}_i$ and then computes $\phi'(\mathbf{X}\mathbf{w};\mathbf{y})$.
10: Master picks $\mathcal{U} \subseteq [r]$ (where $\mathcal{U}$ can be picked either randomly or in a round-robin fashion) and sends $(\mathcal{U} \subseteq [r], \phi'(\mathbf{X}\mathbf{w};\mathbf{y}))$ to all workers.
11: **end while**

similarly, let $\mathbf{X}_{i\mathcal{U}}$ denote the $n \times |\mathcal{U}|$ matrix obtained by restricting the column indices of $\mathbf{X}_i$ to the elements in $\mathcal{U}$.

In Algorithm 1, for each worker i to update $\mathbf{w}_i$ according to (6), where the partial gradient of ϕ with respect to $\mathbf{w}_{i\mathcal{U}}$ is equal to $\nabla_{i\mathcal{U}}\phi(\mathbf{X}\mathbf{w};\mathbf{y}) = \mathbf{X}_{i\mathcal{U}}^T\phi'(\sum_{j=1}^m \mathbf{X}_j\mathbf{w}_j;\mathbf{y})$ and worker i has only $(\mathbf{X}_i, \mathbf{w}_i)$, every other worker j sends $\mathbf{X}_j\mathbf{w}_j$ to the master, who computes $\phi'(\sum_{j=1}^m \mathbf{X}_j\mathbf{w}_j;\mathbf{y})$[5] and sends it back to all the workers. Observe that, even if one worker is corrupt, it can send an adversarially chosen vector to make the computation at the master deviate arbitrarily from the desired computation, which may adversely affect the update at all the worker nodes subsequently.[6] Similarly, corrupt workers can send adversarially chosen information to affect the stopping criterion.

C. Adversary Model

We want to perform the distributed computation described in Section II-A and Section II-B under adversarial attacks, where the corrupt nodes may provide erroneous vectors to the master node. Our adversarial model is described next.

In our adversarial model, the adversary can corrupt at most $t < \frac{m}{2}$ worker nodes[7], and the compromised nodes may collaborate and arbitrarily deviate from their pre-specified programs. If a worker is corrupt, then instead of sending the true vector, it may send an arbitrary vector to disrupt the computation. We refer to the corrupt nodes as erroneous or under the Byzantine attack. We can also handle asynchronous updates, by dropping the straggling nodes beyond a specified delay, and still compute the correct gradient due to encoding. Therefore we treat updates from these nodes as being "erased". We refer to these as erasures/stragglers. For every worker i that sends a message to the master, we can assume, without loss of generality, that the master receives $\mathbf{u}_i + \mathbf{e}_i$, where $\mathbf{u}_i$ is the true vector and $\mathbf{e}_i$ is the error vector, where $\mathbf{e}_i = \mathbf{0}$ if the i'th node is honest, otherwise can be arbitrary. We assume that at most t nodes can be adversarially corrupt and at most s nodes can be stragglers, where s and t are some constants less than $\frac{1}{2}$ that we will decide later. Note that the master node does not know which t worker nodes are corrupted (which makes this problem non-trivial to solve), but knows t. We propose a method that mitigates the effects of both of these anomalies.

Remark 1. *A well-studied problem is that of asynchronous distributed optimization, where the workers can have different delays in updates [40]. One mechanism to deal with this is to wait for a subset of responses, before proceeding to the next iteration, treating the others as missing (or erasures) [41]. Byzantine attacks are quite distinct from such erasures, as the adversary can report wrong local gradients, requiring the master node to create mechanisms to overcome such attacks. If the master node simply aggregates the collected updates as in* (4)*, the computed gradient could be arbitrarily far away from the true one, even with a single adversary [42].*

D. Our Approach to Gradient Computation

Recall that $f_i(\mathbf{w}) = \ell(\langle\mathbf{x}_i, \mathbf{w}\rangle; y_i)$ for some differentiable loss function ℓ, and the gradient of f_i at $\mathbf{w}$ is equal to $\nabla f_i(\mathbf{w}) = (\mathbf{x}_i)^T\ell'(\langle\mathbf{x}_i, \mathbf{w}\rangle; y_i)$, where $\ell'(\langle\mathbf{x}_i, \mathbf{w}\rangle; y_i) := \frac{\partial}{\partial u}\ell(u; y_i)|_{u=\langle\mathbf{x}_i,\mathbf{w}\rangle}$. Note that $\nabla f_i(\mathbf{w}) \in \mathbb{R}^d$ is a column vector. Let $f'(\mathbf{w})$ denote the n-length vector whose i'th entry is equal to $\ell'(\langle\mathbf{x}_i, \mathbf{w}\rangle; y_i)$. With this notation, since $f(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{w})$, we have $\nabla f(\mathbf{w}) = \frac{1}{n}\mathbf{X}^T f'(\mathbf{w})$. Since n is a constant, it is enough to compute $\mathbf{X}^T f'(\mathbf{w})$. So, for simplicity, in the rest of the paper we write

$$\nabla f(\mathbf{w}) = \mathbf{X}^T f'(\mathbf{w}), \quad \forall \mathbf{w} \in \mathbb{R}^d. \tag{7}$$

A natural approach to computing the gradient $\nabla f(\mathbf{w})$ is to compute it in two rounds: (i) compute $f'(\mathbf{w})$ in the 1st round by first multiplying $\mathbf{X}$ with $\mathbf{w}$ and then master

[4] After the 1st iteration, worker i need not multiply $\mathbf{X}_i$ with $\mathbf{w}_i$ to obtain $\mathbf{X}_i\mathbf{w}_i$ in every iteration; as only a few coordinates of $\mathbf{w}_i$ are updated, it only needs to multiply those columns of $\mathbf{X}_i$ that corresponds to the updated coordinates of $\mathbf{w}_i$.

[5] Note that even after computing $\mathbf{X}\mathbf{w}$, master needs access to the labels $y_i, i = 1, 2, \ldots, n$ to compute $\phi'(\mathbf{X}\mathbf{w};\mathbf{y})$. Since $\mathbf{y} \in \mathbb{R}^n$ is just a vector, we can either store that at master, or, alternatively, we can encode $\mathbf{y}$ distributedly at the workers and master can recover that using the method developed in Section IV for Byzantine-resilient distributed matrix-vector multiplication, where the matrix is an identity matrix and vector is equal to $\mathbf{y}$.

[6] Specifically, suppose the i'th worker is corrupt and the adversary wants master to compute $\phi'(\mathbf{X}\mathbf{w} + \mathbf{e};\mathbf{y})$ for any arbitrary vector $\mathbf{e} \in \mathbb{R}^n$ of its choice, then the i'th worker can send $\mathbf{X}_i\mathbf{w}_i + \mathbf{e}$ to the master.

[7] Our results also apply to a slightly *different* adversarial model, where the adversary can adaptively *choose* which of the t worker nodes to attack at each iteration. However, in this model, the adversary cannot modify the local stored data of the attacked node, as otherwise, over time, it can corrupt all the data, making any defense impossible.

locally computes $f'(\mathbf{w})$ from $\mathbf{Xw}$ (master can do this locally, because $\mathbf{Xw}$ is an n-dimensional vector whose i'th entry is equal to $\langle \mathbf{x}_i, \mathbf{w} \rangle$ and $(f'(\mathbf{w}))_i = \ell'(\langle \mathbf{x}_i, \mathbf{w} \rangle; y_i)$);[8] and then (ii) compute $\nabla f(\mathbf{w}) = \mathbf{X}^T f'(\mathbf{w})$ in the 2nd round by multiplying $\mathbf{X}^T$ with $f'(\mathbf{w})$. So, the task of each gradient computation reduces to two matrix-vector (MV) multiplications, where the matrices are fixed and vectors may be different each time. To combat against the adversarial worker nodes, we do both of these MV multiplications using data encoding and real-error correction; see Figure 1 on page 12 for a pictorial description of our approach.

A two-round approach for gradient computation has been proposed for straggler mitigation in [33], but our method for MV multiplication differs from that fundamentally, as we have to provide adversarial protection. Note that in the case of stragglers/erasures we know who the straggling nodes are, but this information is not known in the case of adversarial nodes, and master needs to decode without this information in the context of Byzantine adversaries. This is slightly different from the standard error correcting codes (over finite fields) as the matrix entries in machine learning applications are from reals. In this case, we use ideas from real-error correction (or sparse reconstruction) from the compressive sensing literature [35], and using which we develop an efficient decoding at master, which also gives rise to our sparse encoding matrix; see Section IV for more details. For decoding efficiently, we crucially leverage the block error pattern and design a decoding method at master, which, interestingly, requires *just one* application of the sparse recovery method on a vector of size m, the number of workers, which may be much smaller than the data dimensions n and d, thereby making the decoding computationally efficient. Our encoding matrix (given in (11), designed for MV multiplication) is very sparse and has a regular pattern of non-zero entries, which also makes it applicable for making coordinate-descent (CD) Byzantine-resilient. We emphasize that a general-purpose code for MV multiplication may not be applicable for CD, as each CD iteration requires updating only a few coordinates of the parameter vector, which makes it fundamentally different (and arguably more complicated to robustify) than GD iterations; see Section III-B and Section V for more details. Since iterative algorithms (such as GD and CD) require repeated parameter updates, it is crucial to have a method that has low computational complexity, both at the worker nodes as well as at the master node, and our coding solutions for both GD and CD achieve that, in addition to being highly storage efficient; see Theorem 1 for GD and Theorem 2 for CD.

Coming back to our two-round approach for gradient computations using MV multiplications, for the 1st round, we encode $\mathbf{X}$ using a sparse encoding matrix $\mathbf{S}^{(1)} = [(\mathbf{S}_1^{(1)})^T, \ldots, (\mathbf{S}_m^{(1)})^T]^T$ and store $\mathbf{S}_i^{(1)}\mathbf{X}$ at the i'th worker node; and for the 2nd round, we encode $\mathbf{X}^T$ using another sparse encoding matrix $\mathbf{S}^{(2)} = [(\mathbf{S}_1^{(2)})^T, \ldots, (\mathbf{S}_m^{(2)})^T]^T$, and store $\mathbf{S}_i^{(2)}\mathbf{X}^T$ at the i'th worker node. Now, in the 1st round of the gradient computation at $\mathbf{w}$, the master node broadcasts $\mathbf{w}$ and the i'th worker node replies with $\mathbf{S}_i^{(1)}\mathbf{Xw}$ (a corrupt worker may report an arbitrary vector); upon receiving all the vectors, the master node applies error-correction procedure to recover $\mathbf{Xw}$ and then locally computes $f'(\mathbf{w})$ as described above. In the 2nd round, the master node broadcasts $f'(\mathbf{w})$ and similarly can recover $\mathbf{X}^T f'(\mathbf{w})$ (which is equal to the gradient) at the end of the 2nd round. So, it suffices to devise a method for multiplying a vector $\mathbf{v}$ to a fixed matrix $\mathbf{A}$ in a distributed and adversarial setting. Since this is a linear operation, we can apply error correcting codes over real numbers to perform this task. We describe it briefly below.

A trivial approach. Take a generator matrix $\mathbf{G}$ of any real-error correcting linear code. Encode $\mathbf{A}$ as $\mathbf{A}^T\mathbf{G} =: \mathbf{B}$. Divide the columns of $\mathbf{B}$ into m groups as $\mathbf{B} = [\mathbf{B}_1\ \mathbf{B}_2 \ldots \mathbf{B}_m]$, where worker i stores $\mathbf{B}_i$. Master broadcasts $\mathbf{v}$ and each worker i responds with $\mathbf{v}_T\mathbf{B}_i + \mathbf{e}_i^T$, where $\mathbf{e}_i = \mathbf{0}$ if the i'th worker is honest, otherwise can be arbitrary. Note that at most t of the $\mathbf{e}_i$'s can be non-zero. Responses from the workers can be combined as $\mathbf{v}^T\mathbf{B} + \mathbf{e}^T$. Since every row of $\mathbf{B}$ is a codeword, $\mathbf{v}^T\mathbf{B} = \mathbf{v}^T\mathbf{A}^T\mathbf{G}$ is also a codeword. Therefore, one can take any off-the-shelf decoding algorithm for the code whose generator matrix is $\mathbf{G}$ and obtain $\mathbf{v}^T\mathbf{A}^T$. For example, we can use the Reed-Solomon codes (over real numbers) for this purpose, which only incurs a constant storage overhead and tolerates optimal number of corruptions (up to $< \frac{1}{2}$). Note that we need fast decoding, as it is performed in every iteration of the gradient computation by the master. As far as we know, any off-the-shelf decoding algorithm "over real numbers" requires at least a quadratic computational complexity, which leads to $\Omega(n^2 + d^2)$ decoding complexity per gradient computation, which could be impractical.

The trivial scheme does not exploit the block error pattern which we crucially exploit in our coding scheme to give a $\sim O((n+d)m)$ time decoding per gradient computation, which could be a significant improvement over the trivial scheme, since m typically is much smaller than n and d for large-scale problems. In fact, our coding scheme enables a trade-off (in terms of storage and computation/communication overhead at the master and the worker nodes) with Byzantine adversary tolerance, without compromising the efficiency at the master node. We also want encoding to be efficient (otherwise it defeats the purpose of data encoding) and our sparse encoding matrix achieves that. Our main result for the Byzantine-resilient distributed gradient computation is as follows, which is proved in Section IV:

Theorem 1 (Gradient Computation)**.** *Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote the data matrix. Let m denote the total number of worker nodes. We can compute the gradient exactly in a distributed manner in the presence of t corrupt worker nodes and s stragglers, with the following guarantees, where $\epsilon > 0$ is a free parameter.*

- $(s + t) \leq \left\lfloor \frac{\epsilon}{1+\epsilon} \cdot \frac{m}{2} \right\rfloor$.
- *Total storage requirement is roughly* $2(1 + \epsilon)|\mathbf{X}|$.

[8] Note that even after computing $\mathbf{Xw}$, master needs access to the labels $y_i, i = 1, 2, \ldots, n$ to compute $f'(\mathbf{w})$. See Footnote 5 for a discussion on how master can get access to the labels.

- *Computational complexity for each gradient computation:*
  - *at each worker node is* $O((1+\epsilon)\frac{nd}{m})$.
  - *at the master node is* $O((1+\epsilon)(n+d)m)$.
- *Communication complexity for each gradient computation:*
  - *each worker sends* $\left((1+\epsilon)\frac{n+d}{m}\right)$ *real numbers.*
  - *master broadcasts* $(n+d)$ *real numbers.*
- *Total encoding time is* $O\left(nd\left(\frac{\epsilon}{1+\epsilon}m+1\right)\right)$.

Remark 2. *The statement of Theorem 1 allows for any* s *and* t *as long as* $(s+t) \leq \left\lfloor \frac{\epsilon}{1+\epsilon} \cdot \frac{m}{2} \right\rfloor$. *As we are handling both erasures and errors in the same way*[9] *the corruption threshold does not have to handle* s *and* t *separately. To simplify the discussion, for the rest of the paper, we consider only Byzantine corruption, and denote the corrupted set by* $\mathcal{I} \subset [m]$ *with* $|\mathcal{I}| \leq t$, *with the understanding that this can also work with stragglers.*

In Theorem 1, ϵ *is a design choice and a free parameter that can take any value in the interval* $[0, m-1]$, *where* $\epsilon = 0$ *implies no corruption and* $\epsilon = m-1$ *implies that corruption threshold* t *can be anything up to* $\frac{m-1}{2}$. *If we want to tolerate* t *corrupt workers, then* ϵ *must satisfy* $\epsilon \geq \frac{2t}{m-2t}$.[10]

Remark 3 (Comparison with the plain distributed PGD)**.** *We compare the resource requirements of our method with the plain distributed PGD (which provides* no adversarial protection*), where all the data points are evenly distributed among the* m *workers. In each iteration, master sends the parameter vector* $\mathbf{w}$ *to all the workers; upon receiving* $\mathbf{w}$, *all workers compute the gradients on their local data in* $O(\frac{nd}{m})$ *time (per worker) and send them to the master; master aggregates them in* $O(md)$ *time to obtain the global gradient and then updates the parameter vector using* (2).

In our scheme (i) *the total storage requirement is* $O(1+\epsilon)$ *factor more;*[11] *(see also Remark 4)* (ii) *the amount of computation at each worker node is* $O(1+\epsilon)$ *factor more;* (iii) *the amount of computation at the master node is* $O((1+\epsilon)(1+\frac{n}{d}))$ *factor more, which is comparable in cases where* n *is not much bigger than* d; (iv) *master broadcasts* $(1+\frac{n}{d})$ *factor more data, which is comparable if* n *is not much bigger than* d; *and* (v) *each worker sends* $O\left((1+\epsilon)\frac{1+n/d}{m}\right)$ *factor more data, which is* $O(1+\epsilon)$ *– a constant factor – as long as* $n = O(dm)$.

Remark 4. *Let* m *be an even number. Note that we can get the corruption threshold* t *to be any number less than* $m/2$, *but at the expense of increased storage and computation. For any* $\delta > 0$, *if we want to get* δ *close to m/2, i.e.,* $t = m/2 - \delta$, *then we must have* $(1+\epsilon) \geq m/2\delta$. *In particular, at* $\epsilon = 2$, *we can tolerate up to* $m/3$ *corrupt nodes, with constant overhead in the total storage as well as on the computational complexity.*

Note that when δ *is a constant, i.e.,* t *is close to* $\frac{m-1}{2}$, *then* ϵ *grows linearly with* m; *for example, if* $t = \frac{m-1}{2}$, *then* $\epsilon = m-1$. *In this case, our storage redundancy factor is* $O(m)$. *In contrast, the trivial scheme (see "trivial approach" on page 6) does better in this regime and has only a constant storage overhead, but at the expense of an increased decoding complexity at the master, which is at least quadratic in the problem dimensions* d *and* n, *whereas, our decoding complexity at the master always scales linearly with* d *and* n. *If we always want a constant storage redundancy for all values of the corruption threshold* t, *we can use our coding scheme if* $t \leq c \cdot \frac{m-1}{2}$, *where* $c < 1$ *is a constant, and use the trivial scheme if* t *is close to* $\frac{m-1}{2}$.

Our encoding is also efficient and requires $O\left(nd\left(\frac{\epsilon}{1+\epsilon}m+1\right)\right)$ *time. Note that* $O(nd)$ *is equal to the time required for distributing the data matrix* $\mathbf{X}$ *among* m *workers (for running the distributed gradient descent algorithms* without *the adversary); and the encoding time in our scheme (which results in an encoded matrix that provides Byzantine-resiliency) is a factor of* $(2t+1)$ *more.*

Remark 5. *Our scheme is not only efficient (both in terms of computational complexity and storage requirement), but it can also tolerate up to* $\lfloor \frac{m-1}{2} \rfloor$ *corrupt worker nodes (by taking* $\epsilon = m-1$ *in Theorem 1). It is not hard to prove that this bound is information-theoretically optimal, i.e., no algorithm can tolerate* $\lceil \frac{m}{2} \rceil$ *corrupt worker nodes, and at the same time correctly computes the gradient.*

E. Our Approach to Coordinate Descent

We use data encoding and add redundancy to enlarge the parameter space. Specifically, we encode the data matrix $\mathbf{X}$ using an encoding matrix $\mathbf{R} = [\mathbf{R}_1\ \mathbf{R}_2\ \ldots\ \mathbf{R}_m]$, where each $\mathbf{R}_i$ is a $d \times p$ matrix (with $pm \geq d$), and store $\mathbf{X}\mathbf{R}_i$ at the i'th worker. Define $\widetilde{\mathbf{X}}^R := \mathbf{X}\mathbf{R}$. Now, instead of solving (5), we solve the encoded problem $\arg\min_{\mathbf{v} \in \mathbb{R}^{pm}} \phi(\widetilde{\mathbf{X}}^R\mathbf{v}; \mathbf{y})$ using Algorithm 1 (together with decoding at the master); see Figure 2 on page 18 for a pictorial description of our algorithm. We design the encoding matrix $\mathbf{R}$ such that at every iteration of our algorithm, updating any (small) subset of coordinates of $\mathbf{v}_i$'s (let $\mathbf{v} = [\mathbf{v}_1^T\ \mathbf{v}_2^T\ \ldots\ \mathbf{v}_m^T]$) automatically updates some (small) subset of coordinates of $\mathbf{w}$; and, furthermore, by updating those coordinates of $\mathbf{v}_i$'s, we can efficiently recover the correspondingly updated coordinates of $\mathbf{w}$, despite the errors injected by the adversary. In fact, at any iteration t, the encoded parameter vector $\mathbf{v}_t$ and the original parameter vector $\mathbf{w}_t$ satisfies $\mathbf{v}_t = \mathbf{R}^+\mathbf{w}_t$, where $\mathbf{R}^+ := \mathbf{R}^T(\mathbf{R}\mathbf{R}^T)^{-1}$ is the Moore-Penrose pseudo-inverse of $\mathbf{R}$, and $\mathbf{w}_t$ evolves in the same way as if we are running Algorithm 1 on the original problem.

[9] When there are *only stragglers*, one can design an encoding scheme where *both* the master and the worker nodes operate oblivious to encoding, while solving a slightly altered optimization problem [41], in which gradients are computed *approximately*, leading to more efficient straggler-tolerant GD.

[10] We could have written everything in terms of t, m, n, d, but we chose to introduce another variable ϵ which, in our opinion, clearly brings out the tradeoff between the corruption threshold and the resource requirements without cluttering the expressions.

[11] For example, by taking $\epsilon = 2$, our method can tolerate $m/3$ corrupt worker nodes. So, we can tolerate linear corruption with a *constant* overhead in the resource requirement, compared to the plain distributed gradient computation which does not provide any adversarial protection.

We will be effectively updating the coordinates of the parameter vector $\mathbf{w}$ in chunks of size $(m-2t)$ or its integer multiples (where t is the number of corrupt workers). In particular, if each worker i updates k coordinates of $\mathbf{v}_i$, then $k(m-2t)$ coordinates of $\mathbf{w}$ will get updated. For comparison, Algorithm 1 updates km coordinates of the parameter vector $\mathbf{w}$ in each iteration, if each worker updates k coordinates in that iteration.

As described in Algorithm 1 for the Byzantine-free CD, in order to update its local parameter vector $\mathbf{w}_i$ according to (6), worker i needs access to $\phi'(\mathbf{X}\mathbf{w};\mathbf{y})$, which master computes after receiving $\{\mathbf{X}_j\mathbf{w}_j\}_{j\in[m]}$ from the workers. In our Byzantine-resilient algorithm for CD also master will need to compute $\mathbf{X}\mathbf{w}$ in every CD iteration, and for this purpose, we employ the same encoding-decoding procedure for MV multiplication that we used in the first round of gradient computation, as described in Section II-D. In particular, to make the notation distinct from gradient computation, in order to compute $\mathbf{X}\mathbf{w}$, we encode $\mathbf{X}$ using an encoding matrix $\mathbf{L} = [\mathbf{L}_1^T\ \mathbf{L}_2^T\ \ldots\ \mathbf{L}_m^T]^T$, where each $\mathbf{L}_i$ is a $p' \times n$ matrix (with $p'm \geq n$) and worker i stores $\widetilde{\mathbf{X}}_i^L = \mathbf{L}_i\mathbf{X}$.

Note that in order to compute $\mathbf{X}\mathbf{w}$, in the first round of gradient computation as described in Section II-D, master broadcasts $\mathbf{w}$ to all the workers and each worker i computes $\widetilde{\mathbf{X}}_i^L\mathbf{w}$ and sends it the the master (corrupt workers may report arbitrary vectors), who then decodes and obtains $\mathbf{X}\mathbf{w}$. However, in coordinate descent, though master wants to compute $\mathbf{X}\mathbf{w}$ in each CD iteration, we can significantly improve the computation required at each worker: since only a few coordinates of the original parameter vector $\mathbf{w}$ are updated in each CD iteration, master needs to send only those updated coordinates, and workers need to preform MV multiplication with a much smaller matrix, whose number of columns is equal to the number of updated coordinates of $\mathbf{w}$ that they receive from master. Thus, the computational complexity in each CD iteration at worker is proportional to the number of coordinates updated in each CD iteration, as desired.

Our main result for the Byzantine-resilient distributed coordinate descent is stated below, which is proved in Section V.

Theorem 2 (Coordinate Descent)**.** *Under the setting of Theorem 1, our Byzantine-resilient distributed CD algorithm has the following guarantees, where $\epsilon > 0$ is a free parameter.*

- $(s+t) \leq \left\lfloor \frac{\epsilon}{1+\epsilon} \cdot \frac{m}{2} \right\rfloor$.
- *Total storage requirement is roughly* $2(1+\epsilon)|\mathbf{X}|$.
- *If each worker i updates τ coordinates of $\mathbf{v}_i$, then*
  - $\frac{\tau m}{1+\epsilon}$ *coordinates of the corresponding* $\mathbf{w}$ *gets updated.*
  - *the computational complexity in each iteration*
    - *at each worker node is* $O(n\tau)$.
    - *at the master node is* $O((1+\epsilon)nm + \tau m^2)$.
  - *the communication complexity in each iteration*
    - *each worker sends* $\left(\tau + (1+\epsilon)\frac{n}{m}\right)$ *real numbers.*
    - *master broadcasts* $\left(\frac{\tau m}{1+\epsilon} + n\right)$ *real numbers.*
- *Total encoding time is* $O\left(nd\left(\frac{\epsilon}{1+\epsilon}m + 1\right)\right)$.

Remark 6 (Comparison with the plain distributed CD)**.** *We compare the resource requirements of our method with the plain distributed CD described in Algorithm 1 that does not provide any adversarial protection. Let ϵ be any number in the interval $[0, m-1]$ – for illustration, we can take $\epsilon = 2$, which means $t \leq \frac{m}{3}$ workers are corrupt. In Algorithm 1, if each worker i updates $\frac{\tau}{1+\epsilon}$ coordinates of $\mathbf{w}_i$ (in total $\frac{\tau m}{1+\epsilon}$ coordinates of $\mathbf{w}$) in each iteration, then* (i) *each worker requires $O(\frac{n\tau}{1+\epsilon})$ time to multiply $\mathbf{X}_i$ with the updated part of $\mathbf{w}_i$;* (ii) *master requires $O(nm)$ time to compute $\sum_{i=1}^m \mathbf{X}_i\mathbf{w}_i$ from $\{\mathbf{X}_i\mathbf{w}_i\}_{i\in[m]}$;* (iii) *each worker sends n real numbers (required for $\mathbf{X}_i\mathbf{w}_i$) to master; and* (iv) *master broadcasts n real numbers (required for $\phi'(\mathbf{X}\mathbf{w};\mathbf{y})$).*

In our scheme (i) *the total storage requirement is $O(1+\epsilon)$ factor more;* (ii) *the amount of computation at each worker node is $O(1+\epsilon)$ factor more;* (iii) *the amount of computation at the master node is $O((1+\epsilon)+\frac{\tau m}{n})$ factor more – typically, since τ is a constant and number of workers is much less than n, this again could be $O(1+\epsilon)$;* (iv) *master broadcasts $\left(1 + \frac{\tau m}{(1+\epsilon)n}\right)$ factor more data, which could be a constant if τm is smaller than $(1+\epsilon)n$; and* (v) *each worker sends $\left(\frac{\tau}{n} + \frac{(1+\epsilon)}{nm}\right)$ factor more data, where the 1st term is much smaller than 1 as τ is typically a constant, and the 2nd term is close to zero as $(1+\epsilon)$ is always upper-bounded by m.*

Remark 7 (Comparison with the replication-based strategy)**.** *One simple way to make Algorithm 1 Byzantine-resilient is using repetition code, where we first divide the set of m workers into $\frac{m}{2t+1}$ groups of size $(2t+1)$ each and also divide the data matrix as $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2\ \ldots\ \mathbf{X}_{\frac{m}{2t+1}}]$ (assume, for simplicity, that $(2t+1)$ divides m). Now, store the i'th block $\mathbf{X}_i$ at the $(2t+1)$ workers in the i'th group of workers. Let the parameter vector be divided as $\mathbf{w} = [\mathbf{w}_1^T\ \mathbf{w}_2^T\ \ldots\ \mathbf{w}_{\frac{m}{2t+1}}^T]^T$. In each CD iteration, the local parameter updates in any $\mathbf{w}_i$ is replicated at $(2t+1)$ different workers in the i'th group of workers, and since at most t workers are corrupt, master can do a majority vote for decoding. Note that the total storage and the computation at workers in this scheme grow* linearly *by a factor of $(2t+1)$, where t is the number of corruption, which could be significant. In contrast, the method that we propose can tolerate linear corruption, say, $t = \frac{m}{3}$, with a* constant *overhead in storage and computational complexity.*

The Remarks 2, 4, 5 are also applicable for Theorem 2.

III. Related Work

There has been a significant recent interest in using coding-theoretic techniques to mitigate the well-known straggler problem [40], including gradient coding [43]–[46], encoding computation [33], [34], [47], and data encoding [41], [48]. However, one cannot directly apply the methods for straggler mitigation to the Byzantine attacks case, as we do not know which updates are under attack. Distributed computing with Byzantine adversaries is a richly investigated topic since [9], and has received recent attention in the context of large-scale

distributed optimization and learning [10]–[12], [20]–[32]. These can be divided into three categories: (i) One which assume explicit statistical models for data across workers (e.g., data drawn i.i.d. from a probability distribution) and analyze gradient descent [12], [20], [22], [24], [28]. (ii) Other set of works make no probabilistic assumption on data, and optimize through stochastic methods (e.g., stochastic gradient descent) [10], [21], [23], [25]–[27], [30]–[32] and also with deterministic methods (e.g., gradient descent) [30], [31]. Note that none of these two sets of works do data encoding and work with data as it is, and provide Byzantine resilience by applying some robust aggregation procedures (e.g., geometric median, coordinate-wise median, outlier-filtering, etc.) at the master for aggregating gradients. (iii) Another line of work which is most relevant to ours provide Byzantine resiliency using redundant computations, either by encoding the gradients [11] or by encoding the data itself [29]. Note that [26] combines both redundant computations and do a hierarchical robust aggregation and not is directly comparable to ours.

Note that the statistical nature of data/analysis in the first two sets of works leads to a statistical approximation error in the convergence rates, which is also intensified by the inaccuracy of the robust gradient aggregation procedure. One of the main focuses in these works is typically on obtaining faster convergence (where the goal is to match the convergence rate of plain SGD/GD) and as good an approximation error as possible. Note that the approximation error in all these works scales at least as $\Omega(\sqrt{d})$, where d is the dimension of the model parameter vector, which may be significant in high-dimensional settings. Moreover, in all these works, since we are not allowed to pre-process the data (such as, doing data encoding, etc.), we need to make some assumptions on the data, and furthermore, master has to apply a non-trivial decoding for gradient aggregation, which requires significantly more time than what our decoding requires. For example, filtering-based decoding [22], [30], [31], median-based decoding [12], [20], and heuristic approaches [10], all have a *super-linear* complexity in m – in fact, the filtering-based method as in [22], [30], [31] (which is the most effective in terms of the approximation error) requires $O(m^3 d)$ time. In contrast, our decoding has a *linear* dependence on both m and d. Note that, unlike the first two categories, the third line of work (to which ours also belongs) gives deterministic guarantees and work with arbitrary datasets, with no probabilistic assumptions; we elaborate on these and do a detailed comparison with ours below. We skip the comparison with the first two categories, as it would not be a fair comparison because the underlying setting is different – results in the first two categories are based on statistical assumptions on data/algorithm and inaccurate gradient recovery, whereas, results in the third category make no assumption on the data/algorithm and allow exact gradient recovery.

We want to emphasize that all these works use gradient descent (GD) or stochastic gradient descent (SGD) as their optimization algorithm, which is a data-parallelization method; in this paper, additionally, we also use coordinate descent (CD) algorithm for optimization, which is a model-parallelization method and is preferred over GD in some applications; see Section I for more details on this. As will be evident from Section V, making CD secure against Byzantine attacks is arguably more intricate than securing GD.

We divide this section into three categories: first we compare the redundancy-based methods for GD in Section III-A, and then CD in Section III-B. Since we use matrix-vector (MV) multiplication as a core subroutine for both GD and CD, we also compare related work on this in Section III-C.

A. Gradient Descent (GD)

In this section, we do a detailed comparison with [11] and [29], which are the closest related works that also combat Byzantine adversaries using redundant computations.

For the sake of comparison, assume that $t \leq \frac{m-1}{2}$ workers are corrupt. The coding scheme of Chen et al. [11], which they called DRACO, requires repetition of each data point $(2t+1)$ times, storing each copy at different workers. This gives the storage redundancy *factor* of $(2t+1)$ in DRACO, whereas, our coding method requires storage redundancy factor of $2(1+\epsilon) = \frac{2m}{m-2t}$, which is a constant *even* if t is a constant ($< \frac{1}{2}$) fraction of m.[12] Since each worker in DRACO is doing $(2t+1)$-factor more computation for each GD iteration (than simply computing the gradients as in plain distributed GD), the computational cost at workers also grows by the same factor, which is a significant downside of their scheme. In contrast, our scheme only requires $O(\frac{m}{m-2t})$ more computation at worker, which is a constant even if t is a constant ($< \frac{1}{2}$) fraction of m. This significantly reduces the computation time at the worker nodes in our scheme compared to DRACO, without sacrificing much on the computation time required by the master node – the decoding at master in DRACO takes $O(md)$ time, whereas, our scheme requires $O(\frac{m}{m-2t}(n+d)m)$ time, which is a factor of $O(\frac{m}{m-2t}(1+\frac{n}{d}))$ more than DRACO. In high-dimensional settings, where n is not much bigger than d, and t is a constant ($< \frac{1}{2}$) fraction of m, this overhead is constant. Overall, for a constant fraction of corruption, say, $t = \frac{m}{3}$, DRACO requires $\Omega(t)$ times more storage and computation at workers than our scheme (which could be significant in large-scale settings), and requires $\Omega(1+\frac{n}{d})$ times less computation at master. Note that the computation time at workers scales at least as $\Omega(\frac{nd}{m})$, which dominates the time taken by master (since n, d are typically much larger than m), so our scheme

[12]To highlight the storage redundancy gain of our method over that of DRACO, consider the following two concrete scenarios, where the data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ consists of nd real numbers: (i) In a large setup with $m = 1000$ worker nodes, if we want resiliency against $t = 100$ corrupt nodes (1/10 nodes are corrupt), our method requires redundancy of 2.5, whereas DRACO requires redundancy of 201 (i.e., we need to store only $2.5 \times nd$ real numbers, whereas DRACO stores $201 \times nd$ real numbers), a multiplicative-factor of > 80 more than ours. (ii) In a moderate setup with $m = 150$ and $t = 50$ (1/3 nodes are corrupt), the redundancy of our method is 6, whereas DRACO requires redundancy of 101, a multiplicative-factor of ≈ 17 more than ours.

will be faster than DRACO with respect to the overall running time. Note that the coding in DRACO is restricted to data replication redundancy, as they encode the gradient as done in [43], enabling application to (non)-convex problems; in contrast, we encode the data enabling significantly smaller redundancy, and apply it to learn generalized linear models, and is also applicable to MV multiplication.

Yu et al. [29] (which is a concurrent work[13]) proposes Lagrange coded computing in a distributed framework to compute any multivariate polynomial of the input data and simultaneously provides resilience against stragglers, security against adversaries, and privacy of the dataset against collusion of workers. They leverage the Lagrange polynomial to create computation redundancy among workers, and using standard Reed-Solomon decoding, they can tolerate both erasures/stragglers and errors/adversaries. Their method provide privacy by adding random elements from the field (which in the case of gradient computation is the field of all matrices of a certain dimension) while doing the polynomial interpolation. This is a standard method in Shamir secret sharing scheme [49] that is widely used in information-theoretically secure MPC protocols [50] to provide privacy of users' data. For the sake of comparison of the resource requirements of our scheme and the one in [29], consider the task of linear regression (the concrete machine learning application studied in [29]). In the following, we assume that $\frac{m-1}{2} - \delta$ workers are corrupt, which corresponds to $\epsilon = \frac{m}{1+2\delta} - 1$ in our setting; here δ can take any value in $[0 : \frac{m-1}{2}]$. (i) The storage overhead of our scheme is $\frac{m}{\delta+1/2}$, whereas, in [29], it is $\frac{m}{\delta+1}$, which is roughly the same as ours. For example, to tolerate $\frac{m}{3}$ corrupt workers (i.e., $\delta = \frac{m-3}{6}$), the storage overhead of our scheme and of [29] is a multiplicative factor of 6 and $\frac{6}{1+3/m} \approx 6$, respectively. (ii) The encoding time complexity of our scheme is $O(nd(m-2\delta))$, whereas, it is $O(m\log^2(m)\frac{nd}{\delta+1})$ in [29]. Note that for constant δ (i.e., corruption close to $1/2$), the encoding time of our scheme is much less (by a factor of $O(m\log^2(m))$) than that of [29], whereas, for corruption cm, where $c < \frac{1}{2}$, the scheme of [29] takes $O(\frac{m}{\log^2(m)})$-factor less time in encoding than ours. (iii) The computation time at each worker per gradient computation in both our scheme and [29] is roughly the same – ours requires $O(\frac{nd}{1+2\delta})$ time and [29] requires $O(\frac{nd}{1+\delta})$ time. (iv) The decoding time complexity per gradient computation in [29] is $O(m\log^2(m)d)$, whereas, ours requires $O((1+\epsilon)(n+d)m)$ time. Note that when n is not much bigger than d and we want a constant fraction of corruption, say, $\frac{m}{3}$ corruption, then their decoding complexity is worse than ours by a logarithmic factor. Also note that our decoding algorithm is arguably simpler than theirs. (v) For per gradient computation, each worker respectively sends $\frac{n+d}{1+2\delta}$ and d real numbers in ours and the scheme in [29]. Note that if $n \leq dm$ and to tolerate a constant fraction of corruption, say, $\frac{m}{3}$ corruption, each worker sends roughly $O(m)$ less data in our scheme than that of [29]. Overall, if we want tolerance against $\frac{m}{3}$ corrupt worker nodes, then both our scheme and the one in [29] have similar resource requirements, except for that our scheme has a much better communication complexity (by a factor of $O(m)$) from workers to the master, whereas, the encoding time complexity (which is a one-time process) of [29] is better than ours by a factor of $O(\frac{m}{\log^2(m)})$.

[13] Yu et al. [29] is concurrent to our conference versions in Allerton 2018 [1] and ISIT 2019 [2], [3], on which this paper is based.

B. Coordinate Descent (CD)

Even for the straggler problem, we are only aware of one work by Karakus et al. [48] that, in addition to distributed GD, also studies distributed CD, and that for quadratic problems (e.g., linear/ridge regression) only. It also does data encoding and achieves low redundancy and low complexity, by allowing convergence to an approximate rather than exact solution. As far as we know, ours is the first work that studies distributed CD under Byzantine attacks and provides an efficient solution, much better than the replication-based solution (see Remark 7). At the heart of our solution for CD is the matrix-vector (MV) multiplication procedure that we develop in this paper; and it is the specific regular structure of our encoding matrix (given in (11), designed for the MV multiplication) that allows for partially updating the coordinates of the parameter vector in each CD iteration. Note that a general-purpose encoding matrix for MV multiplication may not be applicable for the CD algorithm.

It has been observed earlier in several works (see, for example, [33], [34]) that gradient computation in GD for linear regression can be reduced to MV multiplication, and any general-purpose code for MV multiplication can be used to provide a solution for gradient computation. As far as we know, ours is the first paper that makes the connection of CD and MV multiplication, and provides an efficient solution for CD (which is also resilient to Byzantine attacks) for learning generalized linear models. Note that, unlike GD, not any general-purpose code for MV multiplication can be used for CD: the main challenge in CD comes from the fact that we only update a small number of coordinates of the parameter vector in each CD iteration; when we encode the data and iteratively update some coordinates of the (encoded) parameter vector using the encoded data, we need to make sure that this update in the encoded parameter vector is reconciled with the update in the original parameter vector. This is fundamentally different from GD iterations. See Section V for more details.

C. Matrix-Vector Multiplication

For the task of a more fundamental problem of matrix-vector (MV) multiplication in the presence of Byzantine adversaries, which is at the core of the optimization algorithms in this paper, we are only aware of two concurrent works [29] (see Footnote 13) and [47][14] that provide (coding-theoretic)

[14] The conference version [34] only studies the straggler problem, and the journal version [47] briefly mentions how their results from [34] can be extended to handle adversarial nodes, and we describe that in this section.

solutions to this problem. In the following, we do a detailed comparison of our solution with both of these works and also discuss the (dis)similarities.

We have already done a detailed comparison with Yu et al. [29] (concurrent work, see Footnote 13) with respect to gradient descent in Section III-A. For the problem of MV multiplication, the storage requirement, computation time per worker, and communication complexity to/from workers is the same in both ours and [29]. The comparison of encoding time complexity is same as above; however, for a constant corruption, say, $\frac{m}{3}$ corrupt workers, our method outperforms the one in [29] in terms of the decoding time complexity by a factor of $O(\log^2(m))$. Note that, unlike [29], we make a fundamental connection of handling Byzantine errors with the sparse reconstruction (or the real-error correction) problem from the compressive sensing literature [35].

Dutta et al. [47] (concurrent work, see Footnote 14) focuses on matrix-vector (MV) multiplication. Though their main focus is on providing resilience against stragglers, they also mention that handling stragglers is very different than handling errors, as it requires to correct errors over real numbers, and, unlike stragglers, we do not know which workers are corrupt. Similar to our observation, they also note that since the matrices and vectors have entries from real numbers, the decoding problem reduces to the sparse reconstruction problem from the compressive sensing literature [35] and they also provide such a reduction. Apart from these similarities, our solution for MV multiplication differs from that of [47] in several important ways: (i) [47] provides a detailed solution to the distributed MV multiplication for the straggler problem for the case when the number of rows in the matrix is smaller than the number of workers nodes. As mentioned in [47], this method can be easily generalized to the more general case when the matrix is of arbitrary dimension, in which case, first we can divide the rows of the matrix into several sub-matrices, each having number of rows smaller than the number of workers, and then apply the above method independently to each sub-matrix. This simple extension may work (without losing efficiency) for the straggler/erasure problem, however, leads to a highly inefficient solution for the adversary/error problem. The reason being that, in the presence of Byzantine workers, if we solve the sparse reconstruction problem for each sub-matrix separately, this would be inefficient, as the decoding would then be computationally expensive. To remedy this, we exploit the block error pattern and use a simple idea of linearly combining the response vectors from each worker using coefficients drawn from an absolutely continuous distribution, so that we only need to do *just one* computation for solving the sparse construction problem. This significantly reduces the decoding complexity; see Section IV-A for details. (ii) [47] only shows a connection to the sparse recovery problem, whereas, we provide a complete solution, with a concrete sparse recovery (or real-error correction) matrix and resource (encoding/decoding time, storage, communication) requirement analysis. (iii) Our encoding matrix (given in (11)) to encode data matrices of arbitrary dimensions is very sparse and highly structured which allows us to apply that construction to CD algorithm, which, as far we know, has not been connected with MV multiplication before. Also, ours is the first paper that provides a non-trivial and efficient (data encoding) solution to CD in the presence of a Byzantine adversary. (iv) We also want to mention that the focus in [47] is on making the *encoded* matrix sparse (at the expense of increased computation at workers) so that workers need to compute shorter dot products, whereas, in this paper, we make the *encoding* matrix sparse (much sparser than the *encoded* matrix of [47]) to get efficient encoding/decoding.

IV. Our Solution to Gradient Computation

In this section, we describe the core technical part of our two-round approach for gradient computation described in Section II-D – a method for performing matrix-vector (MV) multiplication in a distributed manner in the presence of a malicious adversary who can corrupt at most t of the m worker nodes. Here, the matrix is fixed and we want to right-multiply a vector with this matrix.

Given a fixed matrix $\mathbf{A} \in \mathbb{R}^{n_r \times n_c}$ and a vector $\mathbf{v} \in \mathbb{R}^{n_c}$, we want to compute $\mathbf{Av}$ in a distributed manner in the presence of at most t corrupt worker nodes; see Section II-C for details on our adversary model. Our method is based on data encoding and error correction over real numbers, where the matrix $\mathbf{A}$ is encoded and distributed among all the worker nodes, and the master node recovers the MV product $\mathbf{Av}$ using real-error correction; see Figure 1. We will think of our encoding matrix as $\mathbf{S} = [\mathbf{S}_1^T \ \mathbf{S}_2^T, \ldots, \mathbf{S}_m^T]$, where each $\mathbf{S}_i$ is a $p \times n_r$ matrix and $pm \geq n_r$. We will derive the matrix $\mathbf{S}$ in Section IV-B. For the value of p, looking ahead, we will set $p = \lceil \frac{n}{m-2t} \rceil$, which is a constant multiple of $\frac{n}{m}$ even if t is a constant ($< \frac{1}{2}$) fraction of m (e.g., if $t = \frac{m}{3}$, we would have $p = \frac{3n}{m}$). For $i \in [m]$, we store the matrix $\mathbf{S}_i\mathbf{A}$ at the i'th worker node. As described in Section II, the computation proceeds as follows: The master sends $\mathbf{v}$ to all the worker nodes and receives $\{\mathbf{S}_i\mathbf{Av} + \mathbf{e}_i\}_{i=1}^m$ back from them. Let $\mathbf{e}_i = [e_{i1}, e_{i2}, \ldots, e_{ip}]^T$ for every $i \in [p]$. Note that $\mathbf{e}_i = \mathbf{0}$ if the i'th node is honest, otherwise can be arbitrary. In order to find the set of corrupt worker nodes, master equivalently writes $\{\mathbf{S}_i\mathbf{Av} + \mathbf{e}_i\}_{i=1}^m$ as p systems of linear equations.

$$\tilde{h}_i(\mathbf{v}) = \tilde{\mathbf{S}}_i\mathbf{Av} + \tilde{\mathbf{e}}_i, \quad i \in [p] \tag{8}$$

where, for every $i \in [p]$, $\tilde{\mathbf{e}}_i = [e_{1i}, e_{2i}, \ldots, e_{mi}]^T$, and $\tilde{\mathbf{S}}_i$ is an $m \times n_r$ matrix whose j'th row is equal to the i'th row of $\mathbf{S}_j$, for every $j \in [m]$. Note that at most t entries in each $\tilde{\mathbf{e}}_i$ are non-zero. Observe that $\{\mathbf{S}_i\mathbf{Av} + \mathbf{e}_i\}_{i=1}^m$ and $\{\tilde{\mathbf{S}}_i\mathbf{Av} + \tilde{\mathbf{e}}_i\}_{i=1}^p$ are equivalent systems of linear equations, and we can get one from the other.

Note that $\tilde{\mathbf{S}}_i$'s constitute the encoding matrix $\mathbf{S}$, which we have to design. In the following, we will design these matrices $\tilde{\mathbf{S}}_i$'s (which in turn will determine the encoding matrix $\mathbf{S}$), with the help of another matrix $\mathbf{F}$, which will be used to find the error locations, i.e., identities of the compromised worker nodes. We will design the matrix $\mathbf{F}$

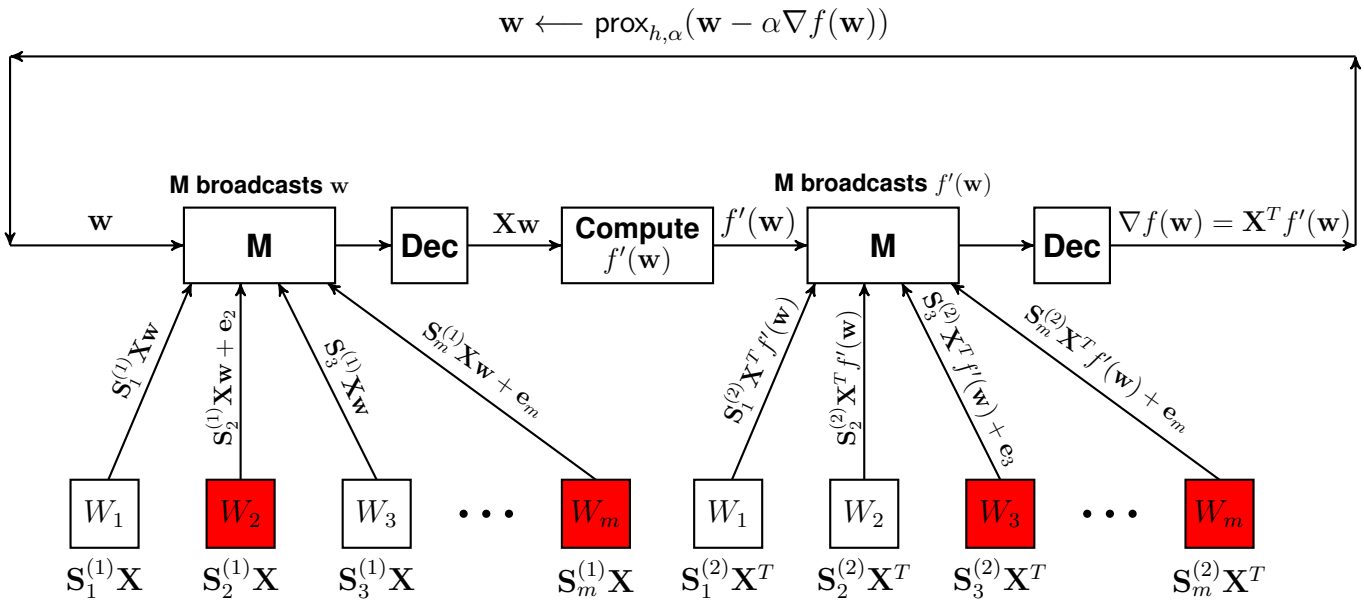

Fig. 1 This figure shows our 2-round approach to the Byzantine-resilient distributed gradient descent to optimize (1) for learning a generalized linear model. Since the gradient at $\mathbf{w}$ is equal to $\nabla f(\mathbf{w}) = \mathbf{X}^T f'(\mathbf{w})$ (see (7)), we compute it in 2 rounds, using a matrix-vector (MV) multiplication as a subroutine in each round. In the 1st round, first we compute $\mathbf{X}\mathbf{w}$, and then compute $f'(\mathbf{w})$ from $\mathbf{X}\mathbf{w}$ – since the j'th entry of $\mathbf{X}\mathbf{w}$ is equal to $\langle \mathbf{x}_j, \mathbf{w}\rangle$, we can compute $f'(\mathbf{w})$ from $\mathbf{X}\mathbf{w}$ (see Section II-D). In the 2nd round we compute $\mathbf{X}^T f'(\mathbf{w})$ – which is equal to $\nabla f(\mathbf{w})$ – using another application of MV multiplication. For a matrix $\mathbf{A}$ and a vector $\mathbf{v}$, to make our distributed MV multiplication $\mathbf{A}\mathbf{v}$ Byzantine-resilient, we encode $\mathbf{A}$ using a sparse matrix $\mathbf{S} = [\mathbf{S}_1^T\ \mathbf{S}_m^T\ \ldots\ \mathbf{S}_m^T]^T$ and distribute $\mathbf{S}_i\mathbf{A}$ to worker i (denoted by W_i). Note that in the first round, we have $\mathbf{A} = \mathbf{X}, \mathbf{v} = \mathbf{w}$, and we encode $\mathbf{X}$ using $\mathbf{S}^{(1)}$, and in the second round, we have $\mathbf{A} = \mathbf{X}^T, \mathbf{v} = f'(\mathbf{w})$, and encode $\mathbf{X}^T$ using $\mathbf{S}^{(2)}$. The adversary can corrupt at most t workers (the compromised ones are denoted in red color), potentially different sets of t workers in different rounds. The master node (denoted by **M**) broadcasts $\mathbf{v}$ to all the workers. Each worker performs the local MV product and sends it back to M. If W_i is corrupt, then it can send an arbitrary vector. Once the master has received all the vectors (out of which t may be erroneous), it sends them to the decoder (denoted by **Dec**), which outputs the correct MV product $\mathbf{A}\mathbf{v}$.

(of dimension $k \times m$, where $k < m$ – here k is determined by the error-correction capability, and we will set $k = 2t$; see Section IV-D for more details) and the matrices $\tilde{\mathbf{S}}_i$'s such that

- **C.1** $\mathbf{F}\tilde{\mathbf{S}}_i = 0$ for every $i \in [p]$.
- **C.2** For any t-sparse $\mathbf{u} \in \mathbb{R}^m$, we can efficiently find all the non-zero locations of $\mathbf{u}$ from $\mathbf{F}\mathbf{u}$.
- **C.3** For any $\mathcal{T} \subset [m]$ such that $|\mathcal{T}| \geq (m-t)$, let $\mathbf{S}_\mathcal{T}$ denote the $|\mathcal{T}|p \times n_r$ matrix obtained from $\mathbf{S}$ by restricting it to all the $\mathbf{S}_i$'s for which $i \in \mathcal{T}$. We want $\mathbf{S}_\mathcal{T}$ to be of full column rank.

If we can find such matrices, then we can recover the desired MV multiplication $\mathbf{A}\mathbf{v}$ exactly: briefly, **C.1** and **C.2** will allow us to locate the corrupt worker nodes; once we have found them, we can discard all the information that the master node had received from them. This will yield $\mathbf{S}_\mathcal{T}\mathbf{A}\mathbf{v}$, where $\mathbf{S}_\mathcal{T}$ is the $|\mathcal{T}|p \times n_r$ matrix obtained from $\mathbf{S}$ by restricting it to $\mathbf{S}_i$'s for all $i \in \mathcal{T}$, where $\mathcal{T}$ is the set of all honest worker nodes. Now, by **C.3**, since $\mathbf{S}_\mathcal{T}$ is of full column rank, we can recover $\mathbf{A}\mathbf{v}$ from $\mathbf{S}_\mathcal{T}\mathbf{A}\mathbf{v}$ exactly. Details follow.

Suppose we have matrices $\mathbf{F}$ and $\tilde{\mathbf{S}}_i$'s such that **C.1** holds. Now, multiplying (8) by $\mathbf{F}$ yields

$$\mathbf{f}_i := \mathbf{F}\tilde{h}_i(\mathbf{v}) = \mathbf{F}\tilde{\mathbf{e}}_i, \tag{9}$$

for every $i \in [p]$, where $\|\tilde{\mathbf{e}}_i\|_0 \leq t$. In Section IV-A, we give our approach for finding all the corrupt worker nodes with the help of any error locator matrix $\mathbf{F}$. Then, in Section IV-B, we give a generic construction for designing $\tilde{\mathbf{S}}_i$'s (and, in turn, our encoding matrix $\mathbf{S}$) such that **C.1** and **C.3** hold. In Section IV-C, we show how to compute the desired matrix-vector product $\mathbf{A}\mathbf{v}$ efficiently, once we have discarded all the data from the corrupt works nodes. Then, in Section IV-D, we will give details of the error locator matrix $\mathbf{F}$ that we use in our construction.

Remark 8. *As we will see in Section IV-B, the structure of our encoding matrix* $\mathbf{S}$ *is independent of our error locator matrix* $\mathbf{F}$*. Specifically, the repetitive structure of the non-zero entries of* $\mathbf{S}$ *as well as their locations will not change irrespective of what the* $\mathbf{F}$ *matrix is. This makes our construction very generic, as we can choose whichever* $\mathbf{F}$ *suits our needs the best (in terms of how many erroneous indices it can locate and with what decoding complexity), and it won't affect the structure of our encoding matrix at all – only the non-zero entries might change, neither their repetitive format, nor their locations!*

A. Finding The Corrupt Worker Nodes

Observe that $\mathsf{supp}(\tilde{\mathbf{e}}_i)$ may not be the same for all $i \in [p]$, but we know, for sure, that the non-zero locations in all these error vectors occur within the same set of t locations. Let $\mathcal{I} = \bigcup_{i=1}^p \mathsf{supp}(\tilde{\mathbf{e}}_i)$, which is the set of all corrupt worker nodes. Note that $|\mathcal{I}| \leq t$. We want to find this set $\mathcal{I}$ efficiently, and for that we note the following crucial observation. Since the non-zero entries of all the error vectors $\tilde{\mathbf{e}}_i$'s occur in the same set $\mathcal{I}$, a random linear combination of $\tilde{\mathbf{e}}_i$'s has support

equal to $\mathcal{I}$ with probability one, if the coefficients of the linear combination are chosen from an *absolutely continuous* probability distribution. This idea has appeared before in [51] in the context of compressed sensing for recovering arbitrary sets of jointly sparse signals that have been measured by the same measurement matrix.

Definition 1. *A probability distribution is called* absolutely continuous*, if every event of measure zero occurs with probability zero.*

It is well-known that a distribution is absolutely continuous if and only if it can be represented as an integral over an integrable density function [52, Theorem 31.8, Chapter 6]. Since Gaussian and uniform distributions have an explicit integrable density function, both are absolutely continuous. Conversely, discrete distributions are not absolutely continuous. Now we state a lemma from [51] that shows that a random linear combination of the error vectors (where coefficients are chosen from an absolutely continuous distribution) preserves the support with probability one.

Lemma 1 ([51])**.** *Let $\mathcal{I} = \bigcup_{i=1}^{p} \mathsf{supp}(\tilde{\mathbf{e}}_i)$, and let $\hat{\mathbf{e}} = \sum_{i=1}^{p} \alpha_i \tilde{\mathbf{e}}_i$, where α_i's are sampled i.i.d. from an absolutely continuous distribution. Then with probability 1, we have $\mathsf{supp}(\hat{\mathbf{e}}) = \mathcal{I}$.*

From (9) we have $\mathbf{f}_i = \mathbf{F}\tilde{\mathbf{e}}_i$ for every $i \in [p]$. Take a random linear combination of $\mathbf{f}_i$'s with coefficients α_i's chosen i.i.d. from an absolutely continuous distribution, for example, the Gaussian distribution. Let $\tilde{\mathbf{f}} = \alpha_i \left(\sum_{i=1}^{p} \mathbf{f}_i\right) = \alpha_i \left(\sum_{i=1}^{p} \mathbf{F}\tilde{\mathbf{e}}_i\right) = \mathbf{F}\left(\sum_{i=1}^{p} \alpha_i \tilde{\mathbf{e}}_i\right) = \mathbf{F}\tilde{\mathbf{e}}$, where $\tilde{\mathbf{e}} = \sum_{i=1}^{p} \alpha_i \tilde{\mathbf{e}}_i$. Note that, with probability 1, $\mathsf{supp}(\tilde{\mathbf{e}})$ is equal to the set of all corrupt worker nodes, and we want to find this set efficiently. In other words, given $\mathbf{F}\tilde{\mathbf{e}}$, we want to find $\mathsf{supp}(\tilde{\mathbf{e}})$ efficiently. For this, we need to design a $k \times m$ matrix $\mathbf{F}$ (where $k < m$) such that for any sparse error vector $\mathbf{e} \in \mathbb{R}^m$, we can efficiently find $\mathsf{supp}(\mathbf{e})$ from $\mathbf{f} = \mathbf{F}\mathbf{e}$. Many such matrices have been known in the literature that can handle different levels of sparsity with varying decoding complexity. We can choose any of these matrices depending on our need, and this will not affect the design of our encoding matrix $\mathbf{S}$. In particular, we will use a $k \times m$ Vandermonde matrix along with the Reed-Solomon type decoding, which can correct up to $k/2$ errors and has decoding complexity of $O(m^2)$; see Section IV-D for details.

Time required in finding the corrupt worker nodes. The time taken in finding the corrupt worker nodes is equal to the sum of the time taken in the following 3 tasks. (i) Computing $\mathbf{F}\tilde{\mathbf{e}}_i$ for every $i \in [p]$: Note that we can get $\mathbf{F}\tilde{\mathbf{e}}_i$ by multiplying (8) with $\mathbf{F}$. Since $\mathbf{F}$ is a $k \times m$ matrix, and we compute $\mathbf{F}\tilde{h}_i(\mathbf{v})$ for p systems, this requires $O(pkm)$ time. (ii) Taking a random linear combination of p vectors each of length m, which takes $O(pm)$ time. (iii) Applying Lemma 2 (in Section IV-D) once to find the error locations, which takes $O(m^2)$ time. Since p is much bigger than m, the total time complexity is $O(pkm)$.

B. Designing The Encoding Matrix $\mathbf{S}$

Now we give a generic construction for designing $\tilde{\mathbf{S}}_i$'s such that **C.1** and **C.3** hold. Fix any $k \times m$ matrix $\mathbf{F}$ such that we can efficiently find $\mathbf{e}$ from $\mathbf{F}\mathbf{e}$, provided $\mathbf{e}$ is sufficiently sparse. We can assume, without loss of generality, that $\mathbf{F}$ has full row-rank; otherwise, there will be redundant observations in $\mathbf{F}\mathbf{e}$ that we can discard and make $\mathbf{F}$ smaller by discarding the redundant rows. Let $\mathcal{N}(\mathbf{F}) \subset \mathbb{R}^m$ denote the null-space of $\mathbf{F}$. Since $\mathsf{rank}(\mathbf{F}) = k$, dimension of $\mathcal{N}(\mathbf{F})$ is $q = (m - k)$. Let $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_q\}$ be a basis of $\mathcal{N}(\mathbf{F})$, and let $\mathbf{b}_i = [b_{i1}\ b_{i2} \ldots b_{im}]^T$, for every $i \in [q]$. We set $\mathbf{b}_i$'s the columns of the following matrix $\mathbf{F}^{\perp}$:

$$\mathbf{F}^{\perp} = \begin{bmatrix} b_{11} & b_{21} & \ldots & b_{q1} \\ b_{12} & b_{22} & \ldots & b_{q2} \\ \vdots & \vdots & \vdots & \vdots \\ b_{1m} & b_{2m} & \ldots & b_{qm} \end{bmatrix}_{m \times q} \tag{10}$$

The following property of $\mathbf{F}^{\perp}$ will be used for recovering the MV product in Section IV-C.

Claim 1. *For any subset $\mathcal{T} \subset [m]$, such that $|\mathcal{T}| \geq (m - t)$, let $\mathbf{F}^{\perp}_{\mathcal{T}}$ be the $|\mathcal{T}| \times q$ matrix, which is equal to the restriction of $\mathbf{F}^{\perp}$ to the rows in $\mathcal{T}$. Then $\mathbf{F}^{\perp}_{\mathcal{T}}$ is of full column rank.*

Proof. Note that $q = m - k$, where $k = 2t$. So, if we show that any q rows of $\mathbf{F}^{\perp}$ are linearly independent, then, this in turn will imply that for every $\mathcal{T} \subset [m]$ with $|\mathcal{T}| \geq (m - t)$, the sub-matrix $\mathbf{F}^{\perp}_{\mathcal{T}}$ will have full column rank. In the following we show that any q rows of $\mathbf{F}^{\perp}$ are linearly independent. To the contrary, suppose not; and let $\mathcal{T}' \subset [m]$ with $|\mathcal{T}'| = q$ be such that the $q \times q$ matrix $\mathbf{F}^{\perp}_{\mathcal{T}'}$ is not a full rank matrix. This implies that there exists a non-zero $\mathbf{c}' \in \mathbb{R}^q$ such that $\mathbf{F}^{\perp}_{\mathcal{T}'}\mathbf{c}' = \mathbf{0}$. Let $\mathbf{b} = \mathbf{F}^{\perp}\mathbf{c}'$. Note that $\mathbf{b} \neq \mathbf{0}$ (because columns of $\mathbf{F}^{\perp}$ are linearly independent) and also that $\|\mathbf{b}\|_0 \leq m - q = k$. Now, since $\mathbf{F}\mathbf{F}^{\perp} = \mathbf{0}$, we have $\mathbf{F}\mathbf{b} = \mathbf{0}$, which contradicts the fact that any k columns of $\mathbf{F}$ are linearly independent. □

Now we design $\tilde{\mathbf{S}}_i$'s. For $i \in [p]$, we set $\tilde{\mathbf{S}}_i$ as follows:

$$\tilde{\mathbf{S}}_i = \begin{bmatrix} 0 & \ldots & 0 & b_{11} & b_{21} & \ldots & b_{l1} & 0 & \ldots & 0 \\ 0 & \ldots & 0 & b_{12} & b_{22} & \ldots & b_{l2} & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \ldots & 0 & b_{1m} & b_{2m} & \ldots & b_{lm} & 0 & \ldots & 0 \end{bmatrix}$$

where $l = q$ if $i < p$; otherwise $l = n_r - (p - 1)q$. The first $(i - 1)q$ and the last $n_r - [(i - 1)q + l]$ columns of $\tilde{\mathbf{S}}_i$ are zero. This also implies that the number of rows in each $\mathbf{S}_i$ is $p = \lceil n_r/q \rceil$.

Claim 2. *For every $i \in [p]$, we have $\mathbf{F}\tilde{\mathbf{S}}_i = 0$.*

Proof. By construction, the null-space of $\mathbf{F}$ is $\mathcal{N}(\mathbf{F}) = \mathsf{span}\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_q\}$, which implies that $\mathbf{F}\mathbf{b}_i = \mathbf{0}$, for every $i \in [q]$. Since all the columns of $\tilde{\mathbf{S}}_i$'s are either $\mathbf{0}$ or $\mathbf{b}_j$ for some $j \in [q]$, the claim follows. □

The above constructed matrices $\tilde{\mathbf{S}}_i$'s give the following encoding matrix $\mathbf{S}_i$ for the i'th worker node:

$$\mathbf{S}_i = \begin{bmatrix} b_{1i} \dots b_{qi} & & & \\ & \ddots & & \\ & & b_{1i} \dots b_{qi} & \\ & & & b_{1i} \dots b_{li} \end{bmatrix}_{p \times n_r} \tag{11}$$

All the unspecified entries of $\mathbf{S}_i$ are zero. The matrix $\mathbf{S}_i$ is for encoding the data for worker i. By stacking up the $\mathbf{S}_i$'s on top of each other gives us our desired encoding matrix $\mathbf{S}$.

To get efficient encoding, we want $\mathbf{S}$ to be as sparse as possible. Since $\mathbf{S}$ is completely determined by $\mathbf{F}^{\perp}$, whose columns are the basis vectors of $\mathcal{N}(\mathbf{F})$, it suffices to find a sparse basis for $\mathcal{N}(\mathbf{F})$. It is known that finding the sparsest basis for the null-space of a matrix is NP-hard [53]. Note that we can always find the basis vectors of $\mathcal{N}(\mathbf{F})$ by reducing $\mathbf{F}$ to its row-reduced-echelon-form (RREF) using the Gaussian elimination [54]. This will result in $\mathbf{F}^{\perp}$ whose last q rows forms a $q \times q$ identity matrix. Note that $q = m - k$, where $k = 2t$. So, if the corruption threshold t is very small as compared to m, the $\mathbf{F}^{\perp}$ that we obtain by the RREF will be very sparse – only the first $2t$ rows may be dense. Since computing $\mathbf{S}$ is equivalent to computing $\mathbf{F}^{\perp}$, and we can compute $\mathbf{F}^{\perp}$ in $O(k^2 m)$ time using the Gaussian elimination, the time complexity of computing $\mathbf{S}$ is also $O(k^2 m)$.

Now we prove an important property of the encoding matrix $\mathbf{S}$ that will be crucial for recovery of the desired matrix-vector product.

Claim 3. *For any $\mathcal{T} \subset [m]$ such that $|\mathcal{T}| \geq (m-t)$, let $\mathbf{S}_{\mathcal{T}}$ denote the $|\mathcal{T}|p \times n_r$ matrix obtained from $\mathbf{S}$ by restricting it to all the blocks $\mathbf{S}_i$'s for which $i \in \mathcal{T}$. Then $\mathbf{S}_{\mathcal{T}}$ is of full column rank.*

Proof. For $i \in [p-1]$, let $\mathcal{B}_i = [(i-1)q + 1 : iq]$ and $\mathcal{B}_p = [(p-1)q + 1 : n_r - (p-1)q]$, where we see $\mathcal{B}_i$'s as a collection of some column indices. Consider any two distinct $i, j \in [p]$. It is clear that for any two vectors $\mathbf{u}_1 \in \mathcal{B}_i, \mathbf{u}_2 \in \mathcal{B}_j$, we have $\mathsf{supp}(\mathbf{u}_1) \cap \mathsf{supp}(\mathbf{u}_2) = \phi$, which means that all the columns in distinct $\mathcal{B}_i$'s are linearly independent. So, to prove the claim, we only need to show that the columns within the same $\mathcal{B}_i$'s are linearly independent. Fix any $i \in [p]$, and consider the $|\mathcal{T}|p \times q$ sub-matrix $\mathbf{S}_{\mathcal{T}}^{(i)}$ of $\mathbf{S}_{\mathcal{T}}$, which is obtained by restricting $\mathbf{S}_{\mathcal{T}}$ to the columns in $\mathcal{B}_i$. There are precisely $|\mathcal{T}|$ non-zero rows in $\mathbf{S}_{\mathcal{T}}^{(i)}$, which are equal to the rows of the matrix $\mathbf{F}_{\mathcal{T}}^{\perp}$ defined in Claim 1. We have already shown in the proof of Claim 1 that $\mathbf{F}_{\mathcal{T}}^{\perp}$ is of full column rank. Therefore, $\mathbf{S}_{\mathcal{T}}^{(i)}$ is also of full column rank. This concludes the proof of Claim 3. □

Since $\mathbf{S}_{\mathcal{T}}$ is of full column rank, in principle, we can recover any vector $\mathbf{u} \in \mathbb{R}^{n_r}$ from $\mathbf{S}_{\mathcal{T}}\mathbf{u}$. In the next section, we show an efficient way for this recovery.

C. Recovering The Matrix-Vector Product $\mathbf{Av}$

Once the master has found the set $\mathcal{I}$ of corrupt worker nodes, it discards all the data received from them. Let $\mathcal{T} = [m] \setminus \mathcal{I} = \{i_1, i_2, \dots, i_f\}$ be the set of all honest worker nodes, where $f = (m - |\mathcal{I}|) \geq (m - t)$. Let $\mathbf{r} = [\mathbf{r}_1^T \mathbf{r}_2^T \dots \mathbf{r}_m^T]$, where $\mathbf{r}_i = \mathbf{S}_i\mathbf{Av} + \mathbf{e}_i$. All the $\mathbf{r}_i$'s from the honest worker nodes can be written as

$$\mathbf{r}_{\mathcal{T}} = \mathbf{S}_{\mathcal{T}}\mathbf{Av}, \tag{12}$$

where $\mathbf{S}_{\mathcal{T}}$ is as defined in Claim 3, and $\mathbf{r}_{\mathcal{T}}$ is also defined analogously and equal to the restriction of $\mathbf{r}$ to all the $\mathbf{r}_i$'s for which $i \in \mathcal{T}$. Since $\mathbf{S}_{\mathcal{T}}$ has full column rank (by Claim 3), in principle, we can recover $\mathbf{Av}$ from (12). Next we show how to recover $\mathbf{Av}$ efficiently, by exploiting the structure of $\mathbf{S}$.

Let $\tilde{\mathbf{r}}_j = [r_{i_1 j}, r_{i_2 j}, \dots, r_{i_f j}]^T$, for every $j \in [p]$. The repetitive structure of $\mathbf{S}_i$'s (see (11)) allows us to write (12) equivalently in terms of p smaller systems.

$$\tilde{\mathbf{r}}_j = \mathbf{F}_j(\mathbf{Av})_{\mathcal{B}_j}, \quad \text{for } j \in [p], \tag{13}$$

where, for $j \in [p-1]$, $\mathcal{B}_i = [(i-1)q+1 : iq]$ and $\mathbf{F}_j = \mathbf{F}_{\mathcal{T}}^{\perp}$, and $\mathcal{B}_p = [(p-1)q+1 : n_r - (p-1)q]$ and $\mathbf{F}_p$ is equal to the restriction of $\mathbf{F}_{\mathcal{T}}^{\perp}$ to its first $(n_r - (p-1)q)$ columns. Since $\mathbf{F}_{\mathcal{T}}^{\perp}$ has full column rank (by Claim 1), we can compute $(\mathbf{Av})_{\mathcal{B}_i}$ for all $i \in [p]$, by multiplying (13) by $\mathbf{F}_j^{+} = (\mathbf{F}_j^T\mathbf{F}_j)^{-1}\mathbf{F}_j^T$, which it called the Moore-Penrose inverse of $\mathbf{F}_j$. Since $\mathbf{Av} = [(\mathbf{Av})_{\mathcal{B}_1}^T, (\mathbf{Av})_{\mathcal{B}_2}^T, \dots, (\mathbf{Av})_{\mathcal{B}_p}^T]^T$, we can recover the desired MV product $\mathbf{Av}$.

Time Complexity analysis. The task of obtaining $\mathbf{Av}$ from $\mathbf{S}_{\mathcal{T}}\mathbf{Av}$ reduces to (i) computing $\mathbf{F}_j^{+} = (\mathbf{F}_{\mathcal{T}}^{\perp})^{+}$ once, which takes $O(q^2|\mathcal{T}|)$ time naïvely; (ii) computing $\mathbf{F}_p^{+}$ once, which takes at most $O(q^2|\mathcal{T}|)$ time naïvely; and (iii) computing the MV products $\mathbf{F}_j^{+}\tilde{\mathbf{r}}_j$ for every $j \in [p]$, which takes $O(pq|\mathcal{T}|)$ time in total. Since p is much bigger than q, the total time taken in recovering $\mathbf{Av}$ from $\mathbf{S}_{\mathcal{T}}\mathbf{Av}$ is $O(pq|\mathcal{T}|) = O(pm^2)$.

D. Designing The Error Locator Matrix $\mathbf{F}$

In this section, we design a $k \times m$ matrix $\mathbf{F}$ (where $k < m$) such that for any *sparse* error vector $\mathbf{e} \in \mathbb{R}^m$, we can uniquely and efficiently recover $\mathbf{e}$ (and, therefore, $\mathsf{supp}(\mathbf{e})$) from the under-determined system of linear equations $\mathbf{f} = \mathbf{Fe} \in \mathbb{R}^k$. This is related to the *sparse representation problem*, where one would like to find the sparsest representation of $\mathbf{f}$ in terms of the linear combination of the columns of $\mathbf{F}$, i.e., minimizing $\|\mathbf{e}\|_0$ subject to the constraint that $\mathbf{f} = \mathbf{Fe}$. This problem is of combinatorial nature and is known to be NP-hard [35]. To make this problem computationally tractable, Candes and Tao [35] showed that if $\mathbf{F}$ satisfies a certain regularity condition (which they named the *restricted isometry property* (RIP)), then the sparsest reconstruction problem can be reduced to minimizing $\|\mathbf{e}\|_1 := \sum_{i=1}^{m} |e_i|$ subject to the constraint that $\mathbf{f} = \mathbf{Fe}$, which can be efficiently solved using a linear program. They also showed that a random Gaussian matrix satisfies the RIP condition. A common problem with such random constructions is that they may not work with small block-lengths (in our setting, m is the number of workers which may not be a big number), and can only correct a constant fraction of errors, where the constant is very small. We need a deterministic construction that can

handle a constant fraction (ideally up to 1/2) of errors and that works with small block-lengths.

Akçakaya and Tarokh [55] proposed an efficient solution to the sparse representation problem using *Vandermonde* matrices. To construct them, take m distinct non-zero elements $z_1, z_2, \ldots, z_m$ from $\mathbb{R}$, and consider the following $k \times m$ Vandermonde matrix $\mathbf{F}$.

$$\mathbf{F} = \begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ z_1 & z_2 & z_3 & \ldots & z_m \\ z_1^2 & z_2^2 & z_3^2 & \ldots & z_m^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_1^{k-1} & z_2^{k-1} & z_3^{k-1} & \ldots & z_m^{k-1} \end{bmatrix}_{k \times m} \tag{14}$$

For the above $\mathbf{F}$, it was shown in [55] that, if $|\mathsf{supp}(\mathbf{e})| \leq k/2$, then the Reed-Solomon type decoding can be used for exact reconstruction of $\mathbf{e}$ from $\mathbf{f} = \mathbf{F}\mathbf{e}$.[15] Furthermore, their decoding algorithm is efficient and runs in $O(m^2)$ time. The results in [55] are given for complex vector spaces, and they hold over real numbers also. Below we state the sparse recovery result (specialized to reals) from [55].

Lemma 2 ([55])**.** *Let $\mathbf{F}$ be the $k \times m$ matrix as defined in* (14)*. Let $\mathbf{e} \in \mathbb{R}^m$ be an arbitrary vector with $|\mathsf{supp}(\mathbf{e})| \leq k/2$. We can exactly recover the vector $\mathbf{e}$ from $\mathbf{f} = \mathbf{F}\mathbf{e}$ in $O(m^2)$ time.*

Note that $\mathbf{F}$ is a $k \times m$ matrix, where $k < m$. Choosing k is in our hands, and larger the k, more the number of errors we can correct (but at the expense of increased storage and computation); see Section IV-E for more details.

E. Resource Requirement Analysis

In this section, we analyze the total amount of resources (storage, computation, and communication) required by our method for computing gradients in the presence of t (out of m) adversarial worker nodes and prove Theorem 1. Fix an $\epsilon > 0$. Let the corruption threshold t satisfy $t \leq \lfloor (\epsilon/(1+\epsilon)) \cdot (m/2) \rfloor$.

As described earlier in Section II-D, we compute the gradient $\nabla f(\mathbf{w}) = \mathbf{X}^T f'(\mathbf{w})$ in two-rounds; and in each round we use the Byzantine-tolerant MV multiplication, which we have developed in Section IV, as a subroutine; see Figure 1 for a pictorial representation of our scheme. We encode $\mathbf{X}$ to compute $f'(\mathbf{w})$ in the 1st round: first compute $\mathbf{X}\mathbf{w}$ using MV multiplication and then locally compute $f'(\mathbf{w})$. To compute $\mathbf{X}^T f'(\mathbf{w})$ (which is equal to the gradient) in the 2nd round, we encode $\mathbf{X}^T$ and compute $\mathbf{X}^T f'(\mathbf{w})$. Let $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ be the encoding matrices of dimensions $p_1 m \times n$ and $p_2 m \times d$, respectively, to encode $\mathbf{X}$ and $\mathbf{X}^T$, respectively. Here, $p_1 = \lceil n/q \rceil$ and $p_2 = \lceil d/q \rceil$, where $q = m - k$. Since $k = 2t$ (by Lemma 2), we have $q = (m-k) \geq m/(1+\epsilon)$.

[15]Note that, since any k columns of $\mathbf{F}$ (which is the Vandermonde matrix) are linearly independent, if there exists a vector $\mathbf{e}$ such that $|\mathsf{supp}(\mathbf{e})| \leq k/2$ and $\mathbf{e}$ satisfies $\mathbf{f} = \mathbf{F}\mathbf{e}$ for a fixed $\mathbf{f}$, then $\mathbf{e}$ is unique.

1) Storage Requirement: Each worker node i stores two matrices $\mathbf{S}_i^{(1)}\mathbf{X}$ and $\mathbf{S}_i^{(2)}\mathbf{X}^T$. The first one is a $p_1 \times (d+1)$ matrix, and the second one is a $p_2 \times n$ matrix. So, the total amount of storage at all worker nodes is equal to storing $(p_1(d+1) + p_2 n) \times m$ real numbers. Since $p_1 \leq \lceil (1+\epsilon)\frac{n}{m} \rceil$ and $p_2 \leq \lceil (1+\epsilon)\frac{d}{m} \rceil$, the total storage is

$$\begin{aligned} \big(p_1(d+1) + p_2 n\big)m &= p_1 m(d+1) + p_2 mn \\ &< [(1+\epsilon)n + m](d+1) + [(1+\epsilon)d + m]n \\ &= (1+\epsilon)n(2d+1) + m(n+d+1). \end{aligned}$$

where the first term is roughly equal to a $2(1+\epsilon)$ factor more than the size of $\mathbf{X}$. Note that the second term does not contribute much to the total storage as compared to the first term, because the number of worker nodes m is much smaller than both n and d. In fact, if $m-k$ divides both n and d, then the second term vanishes. Since $|\mathbf{X}|$ is an $n \times d$ matrix, the total storage at each worker node is almost equal to $2(1+\epsilon)\frac{|\mathbf{X}|}{m}$, which is a constant factor of the optimal, that is, $\frac{|\mathbf{X}|}{m}$, and the total storage is roughly equal to $2(1+\epsilon)|\mathbf{X}|$.

2) Computational Complexity: We can divide the computational complexity of our scheme as follows:

- *Encoding the data matrix.* Since, for every $i \leq k$ and $j > k$, the total number of non-zero entries in $\mathbf{S}_i^{(1)}$ and $\mathbf{S}_j^{(1)}$ are at most n and p_1, respectively (see Section IV-B for details), the computational complexity for computing $\mathbf{S}_i^{(1)}\mathbf{X}$ for each $i \leq k$, and $\mathbf{S}_j^{(1)}\mathbf{X}$ for each $j > k$, is $O(nd)$ and $O(p_1 d)$, respectively. So, the encoding time for computing $\mathbf{S}^{(1)}\mathbf{X}$ is equal to $O\left(k(nd) + (m-k)(p_1 d)\right) = O\left((\frac{\epsilon}{1+\epsilon}m + 1)nd\right)$. Similarly, we can show that the encoding time for computing $\mathbf{S}^{(2)}\mathbf{X}^T$ is also equal to $O\left((\frac{\epsilon}{1+\epsilon}m + 1)nd\right)$. Note that computing $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ take $O(k^2 m)$ time each, which is much smaller, as compared to the encoding time. So, the total encoding time is $O\left((\frac{\epsilon}{1+\epsilon}m + 1)nd\right)$. Note that this encoding is to be done only once.
- *Computation at each worker node.* In the first round, upon receiving $\mathbf{w}$ from the master node, each worker i computes $(\mathbf{S}_i^{(1)}\mathbf{X})\mathbf{w}$, and reports back the resulting vector. Similarly, in the second round, upon receiving $f'(\mathbf{w})$ from the master node, each worker i computes $(\mathbf{S}_i^{(2)}\mathbf{X}^T)f'(\mathbf{w})$, and reports back the resulting vector. Since $\mathbf{S}_i^{(1)}\mathbf{X}$ and $\mathbf{S}_i^{(2)}\mathbf{X}^T$ are $p_1 \times (d+1)$ and $p_2 \times n$ matrices, respectively, each worker node i requires $O(p_1 d + p_2 n) = O((1+\epsilon)\frac{nd}{m})$ time.
- *Computation at the master node.* The total time taken by the master node in both the rounds is the sum of the time required in (i) finding the corrupt worker nodes in the 1st and 2nd rounds, which requires $O(p_1 km)$ and $O(p_2 km)$ time, respectively (see Section IV-A), (ii) recovering $\mathbf{X}\mathbf{w}$ from $\mathbf{S}_{\mathcal{T}}^{(1)}\mathbf{X}\mathbf{w}$ in the 1st round, which requires $O(p_1 m^2)$ time, (iii) computing $f'(\mathbf{w})$ from $\mathbf{X}\mathbf{w}$, which takes $O(n)$ time, and (iv) recovering $\mathbf{X}^T f'(\mathbf{w})$ from $\mathbf{S}_{\mathcal{T}}^{(2)}\mathbf{X}^T f'(\mathbf{w})$ in the 2nd round, which requires $O(p_2 m^2)$ time (see Section IV-C). Since $k < m$, the total time is equal to

$O((p_1+p_2)m^2) = O((1+\epsilon)(n+d)m)$.

3) Communication Complexity: In each gradient computation, (i) master broadcasts $(n+d)$ real numbers, d in the first round and n in the second round; and (ii) each worker sends $\big((1+\epsilon)\frac{n+d}{m}\big)$ real numbers to master, $(1+\epsilon)\frac{n}{m}$ in the first round and $(1+\epsilon)\frac{d}{m}$ in the second round.

V. Our Solution to Coordinate Descent

In this section, we give a solution to the distributed coordinate descent (CD) under Byzantine attacks and prove Theorem 2. To make our notation simpler, we remove the dependence on the label vector $\mathbf{y}$ in the problem expression (5) and rewrite it as follows (this is without loss of generality in the light of Footnote 5 and Algorithm 1):

$$\arg\min_{\mathbf{w}\in\mathbb{R}^d} \phi(\mathbf{X}\mathbf{w}) := \sum_{i=1}^{n} \ell(\langle \mathbf{x}_i, \mathbf{w}\rangle). \tag{15}$$

We want to optimize (15) using distributed CD, described in Section II-B. As outlined in Section II-E, we use data encoding and error correction over real numbers for that. To combat the effect of adversary, we add redundancy to enlarge the parameter space. Let $\widetilde{\mathbf{X}}^R = \mathbf{X}\mathbf{R}$, where $\mathbf{R} = [\mathbf{R}_1\ \mathbf{R}_2\ \dots\ \mathbf{R}_m] \in \mathbb{R}^{d\times pm}$ with $pm \geq d$, and each $\mathbf{R}_i$ is a $p\times d$ matrix. We will determine the encoding matrix $\mathbf{R}$ later, after describing what properties we want from it. For the value of p, looking ahead, when t is the number of corrupt workers, we will choose $p = \frac{d}{m-2t}$, which is a constant multiple of $\frac{d}{m}$ even if t is a constant fraction ($<\frac{1}{2}$) of m (e.g., for $t=\frac{m}{3}$, we have $p=\frac{3d}{m}$). We consider $\mathbf{R}$'s which are of full row-rank. Let $\mathbf{R}^+ := \mathbf{R}^T(\mathbf{R}\mathbf{R}^T)^{-1}$ denote its Moore-Penrose inverse such that $\mathbf{R}\mathbf{R}^+ = I_d$, where I_d is the $d\times d$ identity matrix. Note that $\mathbf{R}^+$ is of full column-rank. Let $\mathbf{v} = \mathbf{R}^+\mathbf{w}$ be the transformed vector, which lies in a larger (than d) dimensional space. Let $\mathbf{R}^+ = [(\mathbf{R}_1^+)^T\ (\mathbf{R}_2^+)^T\ \dots\ (\mathbf{R}_m^+)^T]^T$, where each $\mathbf{R}_i^+ := (\mathbf{R}^+)_i$ is a $p\times d$ matrix. With this, by letting $\mathbf{v} = [\mathbf{v}_1^T\ \mathbf{v}_2^T\ \dots\ \mathbf{v}_m^T]^T$, we have that $\mathbf{v}_i = \mathbf{R}_i^+\mathbf{w}$ for every $i\in[m]$. Now, consider the following modified problem over the encoded data.

$$\arg\min_{\mathbf{v}\in\mathbb{R}^{pm}} \phi(\widetilde{\mathbf{X}}^R\mathbf{v}). \tag{16}$$

Observe that, since $\mathbf{R}$ is of full row-rank, $\min_{\mathbf{w}\in\mathbb{R}^d}\phi(\mathbf{X}\mathbf{w})$ is equal to $\min_{\mathbf{v}\in\mathbb{R}^{pm}}\phi(\widetilde{\mathbf{X}}^R\mathbf{v})$; and from an optimal solution to one problem we can obtain an optimal solution to the other problem. We design an encoding/decoding scheme such that when we optimize the encoded problem (16) using Algorithm 1, the vector $\mathbf{v}$ that we get in each iteration is of the form $\mathbf{v} = \mathbf{R}^+\mathbf{w}$ for some vector $\mathbf{w}\in\mathbb{R}^d$.[16] In fact, our encoding/decoding will ensure that the $\mathbf{w}$ for which $\mathbf{v} = \mathbf{R}^+\mathbf{w}$ would be equal to the original parameter vector in that iteration if we had run Algorithm 1 to solve (15). We need this property because in any CD iteration t, we need access to the original parameter vector $\mathbf{w}^t$ (such that $\mathbf{v}^t = \mathbf{R}^+\mathbf{w}^t$) to facilitate the local parameter updates of $\mathbf{v}_1^t,\dots,\mathbf{v}_m^t$ at the workers. See the paragraph after (18) for more details.

Now, instead of solving (15), we solve its encoded form (16) using Algorithm 1 (with decoding at the master), where each worker i stores $\widetilde{\mathbf{X}}_i^R = \mathbf{X}\mathbf{R}_i$ and is responsible for updating (some coordinates of) $\mathbf{v}_i$. In the following, let $\mathcal{U}\subseteq[p]$ be a fixed arbitrary subset of $[p]$. Let $\mathbf{v}^0 := \mathbf{R}^+\mathbf{w}^0$ for some $\mathbf{w}^0$ at time $t=0$. Suppose, at the beginning of the t'th iteration, we have $\mathbf{v}^t = \mathbf{R}^+\mathbf{w}^t$ for some $\mathbf{w}^t$, and each worker i updates $\mathbf{v}_{i\mathcal{U}}^t$ according to

$$\mathbf{v}_{i\mathcal{U}}^{t+1} = \mathbf{v}_{i\mathcal{U}}^t - \alpha_t \nabla_{i\mathcal{U}}\phi(\widetilde{\mathbf{X}}^R\mathbf{v}^t), \tag{17}$$

where $\nabla_{i\mathcal{U}}\phi(\widetilde{\mathbf{X}}^R\mathbf{v}^t) = (\widetilde{\mathbf{X}}_{i\mathcal{U}}^R)^T\phi'(\widetilde{\mathbf{X}}^R\mathbf{v}^t)$. Recall that each $\mathbf{R}_i$ is a $d\times p$ matrix, and each $\mathbf{R}_i^+ := (\mathbf{R}^+)_i$ is a $p\times d$ matrix. We denote by $\mathbf{R}_{i\mathcal{U}}$ the $d\times|\mathcal{U}|$ matrix obtained by restricting the columns of $\mathbf{R}_i$ to the elements of $\mathcal{U}$. Analogously, we denote by $\mathbf{R}_{i\mathcal{U}}^+ := (\mathbf{R}^+)_{i\mathcal{U}}$ the $|\mathcal{U}|\times d$ matrix obtained by restricting the rows of $\mathbf{R}_i^+$ to the elements of $\mathcal{U}$. With this, we can write $\widetilde{\mathbf{X}}_{i\mathcal{U}}^R = \mathbf{X}\mathbf{R}_{i\mathcal{U}}$. Now, (17) can be equivalently written as

$$\mathbf{v}_{i\mathcal{U}}^{t+1} = \mathbf{v}_{i\mathcal{U}}^t - \alpha_t \mathbf{R}_{i\mathcal{U}}^T\mathbf{X}^T\phi'(\widetilde{\mathbf{X}}^R\mathbf{v}^t). \tag{18}$$

In order to update $\mathbf{v}_{i\mathcal{U}}^t$, worker i requires $\phi'(\widetilde{\mathbf{X}}^R\mathbf{v}^t)$, where $\widetilde{\mathbf{X}}^R\mathbf{v}^t = \sum_{j=1}^m \widetilde{\mathbf{X}}_j^R\mathbf{v}_j^t$ and worker i has only $(\widetilde{\mathbf{X}}_i^R, \mathbf{v}_i^t)$. Since $\mathbf{v}^t = \mathbf{R}^+\mathbf{w}^t$, we have $\widetilde{\mathbf{X}}^R\mathbf{v}^t = \mathbf{X}\mathbf{R}\mathbf{v}^t = \mathbf{X}\mathbf{w}^t$. So, it suffices to compute $\mathbf{X}\mathbf{w}^t$ at the master node – once master has $\mathbf{X}\mathbf{w}^t$, it can locally compute $\phi'(\mathbf{X}\mathbf{w}^t)$ and send it to all the workers. Computing $\mathbf{X}\mathbf{w}^t$ is the distributed matrix-vector (MV) multiplication problem, where the matrix $\mathbf{X}$ is fixed and we want to compute $\mathbf{X}\mathbf{w}^t$ for any vector $\mathbf{w}^t$ in the presence of an adversary. In Section IV, we give a method for performing distributed MV multiplication in the presence of an adversary. Now we give an overview, together-with an improvement on its computational complexity.

We encode $\mathbf{X}$ using an encoding matrix $\mathbf{L}\in\mathbb{R}^{(p'm)\times n}$. Let $\mathbf{L} = [\mathbf{L}_1^T\ \mathbf{L}_2^T\ \dots\ \mathbf{L}_m^T]^T$, where each $\mathbf{L}_i$ is a $p'\times n$ matrix with $p' = \lceil\frac{n}{m-2t}\rceil$. Each $\mathbf{L}_i$ has p' rows and n columns, and has the same structure as that of $\mathbf{S}_i$ from (11). Worker i stores $\widetilde{\mathbf{X}}_i^L = \mathbf{L}_i\mathbf{X}$. To compute $\mathbf{X}\mathbf{w}$, master sends $\mathbf{w}$ to all the workers; worker i responds with $\mathbf{L}_i\mathbf{X}\mathbf{w} + \mathbf{e}_i$, where $\mathbf{e}_i = \mathbf{0}$ if the i'th worker is honest, otherwise can be arbitrary; upon receiving $\{\mathbf{L}_i\mathbf{X}\mathbf{w}+\mathbf{e}_i\}_{i=1}^m$, where at most t of the $\mathbf{e}_i$'s can be non-zero, master applies the decoding procedure and recovers $\mathbf{X}\mathbf{w}$ back. We can improve the computational complexity of this method significantly by observing that, in each iteration of our distributed CD algorithm, only a few coordinates of $\mathbf{w}$ get updated and the rest of the coordinates remain unchanged. (Looking ahead, when each worker updates $\mathbf{v}_{i\mathcal{U}}$'s according to (17), it automatically updates $\mathbf{w}_{f(\mathcal{U})}$ according to (6) – for a specific function f as defined in (21) – where $\mathbf{v}$ and $\mathbf{w}$ satisfy $\mathbf{v} = \mathbf{R}^+\mathbf{w}$.) This implies that for computing $\mathbf{X}\mathbf{w}$, master only needs to send the updated coordinates to the workers and keeps the result from the previous MV product with itself. This significantly reduces the local computation

[16] If such a $\mathbf{w}$ exists, then it is unique. This follows from the fact that $\mathbf{R}^+$ is of full column-rank.

at the worker nodes, as now they only need to perform a local MV product of a matrix of size $p' \times |f(\mathcal{U})|$ and a vector of length $|f(\mathcal{U})|$. See Section IV for details.

Our goal in each iteration of CD is to update some coordinates of the original parameter vector $\mathbf{w}$; instead, by solving the encoded problem, we are updating coordinates of the transformed vector $\mathbf{v}$. We would like to design an algorithm/encoding such that it has exactly the same convergence properties as if we are running the distributed CD on the original problem without any adversary. For this, naturally, we would like our algorithm to satisfy the following property:

Update on any (small) subset of coordinates of $\mathbf{w}$ *should be achieved by updating some (small) subset of coordinates of* $\mathbf{v}_i$*'s; and, by updating those coordinates of* $\mathbf{v}_i$*'s, we should be able to efficiently recover the correspondingly updated coordinates of* $\mathbf{w}$*. Furthermore, this should be doable despite the errors injected by the adversary in every iteration of the algorithm.*

Note that if each coordinate of $\mathbf{v}$ depends on too many coordinates of $\mathbf{w}$, then updating a few coordinates of $\mathbf{v}$ may affect many coordinates of $\mathbf{w}$, and it becomes information-theoretically impossible to satisfy the above property (even without the presence of an adversary).[17] This imposes a restriction that each row of $\mathbf{R}^+$ must have few non-zero entries, in such a way that updating $\mathbf{v}^t_{i\mathcal{U}}$'s, for any choice of $\mathcal{U} \subseteq [p]$, will collectively update only a subset (which may potentially depend on $\mathcal{U}$) of coordinates of the original parameter vector $\mathbf{w}^t$, and we can uniquely and efficiently recover those updated coordinates of $\mathbf{w}^t$, even from the *erroneous* vectors $\{\mathbf{v}^{t+1}_{i\mathcal{U}} + \mathbf{e}_{i\mathcal{U}}\}_{i=1}^m$, where at most t out of m error vectors $\{\mathbf{e}_{i\mathcal{U}}\}_{i=1}^m$ are non-zero and may have arbitrary entries. In order to achieve this, we will design a sparse encoding matrix $\mathbf{R}^+$ (which in turn determines $\mathbf{R}$), that satisfies the following properties:

P.1 $\mathbf{R}^+$ has structured sparsity, which induces a map $f : [p] \to \mathcal{P}([d])$ (where $\mathcal{P}([d])$ denotes the power set of $[d]$) such that

a) $\{f(i) : i \in [p]\}$ partitions $\{1, 2, \ldots, d\}$, i.e., for every $i, j \in [p]$, such that $i \neq j$, we have $f(i) \cap f(j) = \emptyset$ and that $\bigcup_{i=1}^p f(i) = [d]$.

b) $|f(i)| = |f(j)|$ for every $i, j \in [p-1]$, and $|f(p)| \leq |f(i)|$, for any $i \in [p-1]$.

c) For any $\mathcal{U} \subseteq [p]$, define $f(\mathcal{U}) := \cup_{j \in \mathcal{U}} f(j)$. If we update $\mathbf{v}^t_{i\mathcal{U}}$, $\forall i \in [m]$, according to (18), it automatically updates $\mathbf{w}^t_{f(\mathcal{U})}$ according to

$$\mathbf{w}^{t+1}_{f(\mathcal{U})} = \mathbf{w}^t_{f(\mathcal{U})} - \alpha_t \mathbf{X}^T_{f(\mathcal{U})} \phi'(\mathbf{X}\mathbf{w}^t). \tag{19}$$

If we set $\mathbf{v}^{t+1}_{i\overline{\mathcal{U}}} := \mathbf{v}^t_{i\overline{\mathcal{U}}}$ and $\mathbf{w}^{t+1}_{\overline{f(\mathcal{U})}} := \mathbf{w}^t_{\overline{f(\mathcal{U})}}$, then $\mathbf{v}^{t+1} = \mathbf{R}^+ \mathbf{w}^{t+1}$, i.e., our invariant holds.

Note that (19) is the same update rule if we run the plain CD algorithm to update $\mathbf{w}_{f(\mathcal{U})}$. In fact, our encoding matrix satisfies a stronger property, that $\mathbf{v}^{t+1}_{i\mathcal{U}} = \mathbf{R}^+_{i\mathcal{U}, f(\mathcal{U})} \mathbf{w}^{t+1}_{f(\mathcal{U})}$ holds for every $i \in [m]$, $\mathcal{U} \subseteq [p]$, where $\mathbf{R}^+_{i\mathcal{U}, f(\mathcal{U})}$ denotes the $|\mathcal{U}| \times |f(\mathcal{U})|$ matrix obtained from $\mathbf{R}^+_{i\mathcal{U}}$ by restricting its column indices to the elements in $f(\mathcal{U})$.

P.2 We can efficiently recover $\mathbf{w}^{t+1}_{f(\mathcal{U})}$ from the erroneous vectors $\{\mathbf{v}^{t+1}_{i\mathcal{U}} + \mathbf{e}_{i\mathcal{U}}\}_{i=1}^m$, where at most t of $\mathbf{e}_{i\mathcal{U}}$'s are non-zero and may have arbitrary entries. Since $\mathbf{v}^{t+1}_{i\mathcal{U}} = \mathbf{R}^+_{i\mathcal{U}, f(\mathcal{U})} \mathbf{w}^{t+1}_{f(\mathcal{U})}$, for every $i \in [m], \mathcal{U} \subseteq [p]$, this property requires that not only $\mathbf{R}^+$, but its sub-matrices also have error correcting capabilities.

Remark 9. *Note that* **P.1** *implies that for every* $i \in [p]$*, we have* $|f(i)| \leq d/p$*. As we see later, this will be equal to* $m/(1+\epsilon)$ *for some* $\epsilon > 0$ *which is determined by the corruption threshold. This means that in each iteration of the CD algorithm running on the modified encoded problem, we will be effectively updating the coordinates of the parameter vector* $\mathbf{w}$ *in chunks of size* $m/(1+\epsilon)$ *or its integer multiples. In particular, if each worker* i *updates* k *coordinates of* $\mathbf{v}_i$*, then* $km/(1+\epsilon)$ *coordinates of* $\mathbf{w}$ *will get updated. For comparison, Algorithm 1 updates* km *coordinates of the parameter vector* $\mathbf{w}$ *in each iteration, if each worker updates* k *coordinates in that iteration.*

Now we design an encoding matrix $\mathbf{R}^+$ and a decoding method that satisfy **P.1** and **P.2**.

A. Encoding and Decoding

In this section, we first design an encoding matrix $\mathbf{R}^+$ that satisfies **P.1**. $\mathbf{R}^+$ will be such that it has orthonormal rows, so, $\mathbf{R}$ is easy to compute, $\mathbf{R} = (\mathbf{R}^+)^T$. For simplicity, we denote $\mathbf{R}^+$ by $\mathbf{S}$. We show that the encoding matrix that we design for the MV multiplication in Section IV satisfies all the properties that we want.[18] In the MV multiplication, we had a fixed matrix $\mathbf{A}$ and the master node wants to compute $\mathbf{A}\mathbf{w}$ for any vector $\mathbf{w}$ of its choice. In the solution presented in Section IV, we encode $\mathbf{A}$ and store $\mathbf{S}_i\mathbf{A}$ at the i'th worker node. Now, the master sends $\mathbf{w}$ to all the worker nodes, and each worker i responds with $\mathbf{S}_i\mathbf{A}\mathbf{w} + \mathbf{e}_i$, where $\mathbf{e}_i = \mathbf{0}$ if worker i is honest, otherwise can be arbitrary. Once master receives $\{\mathbf{S}_i\mathbf{A}\mathbf{w} + \mathbf{e}_i\}_{i=1}^m$, it can run the error correcting procedure to recover $\mathbf{A}\mathbf{w}$. To apply this in our setting, we take $\mathbf{A}$ to be the identity matrix, such that $\mathbf{S}_i\mathbf{A} = \mathbf{S}_i$, and the master can recover $\mathbf{w}$ from $\{\mathbf{r}_i = \mathbf{S}_i\mathbf{w} + \mathbf{e}_i\}_{i=1}^m$, if at most t of the $\mathbf{e}_i$'s are non-zero. For convenience, we rewrite the encoding matrix $\mathbf{S}_i$ for the i'th worker node from

[17] To see this, consider the case when each worker i updates only the first coordinate of $\mathbf{v}_i$ and no worker is corrupt. Master receives m linear equations $\mathbf{v}_{i1} = \mathbf{R}^+_{i1}\mathbf{w}$, $i = 1, 2, \ldots, m$, where $\mathbf{R}^+_{i1}$ is the first row of $\mathbf{R}^+_i$ for every $i \in [m]$. Assume, for simplicity, that these m equations are linearly independent. When m is smaller than d (which is always the case), there are infinite solutions to this system of linear equations, unless at most m elements of $\mathbf{w}$ are involved in the m linear equations (i.e., the number of unknowns are at most the number of equations), which is equivalent to saying that the rows $\mathbf{R}^+_{i1}$ for $i = 1, 2, \ldots, m$ are sparse. Our encoding matrix will satisfy this property; see Section V-A for more detail.

[18] The encoding and decoding of this section is based on the corresponding algorithms from Section IV.

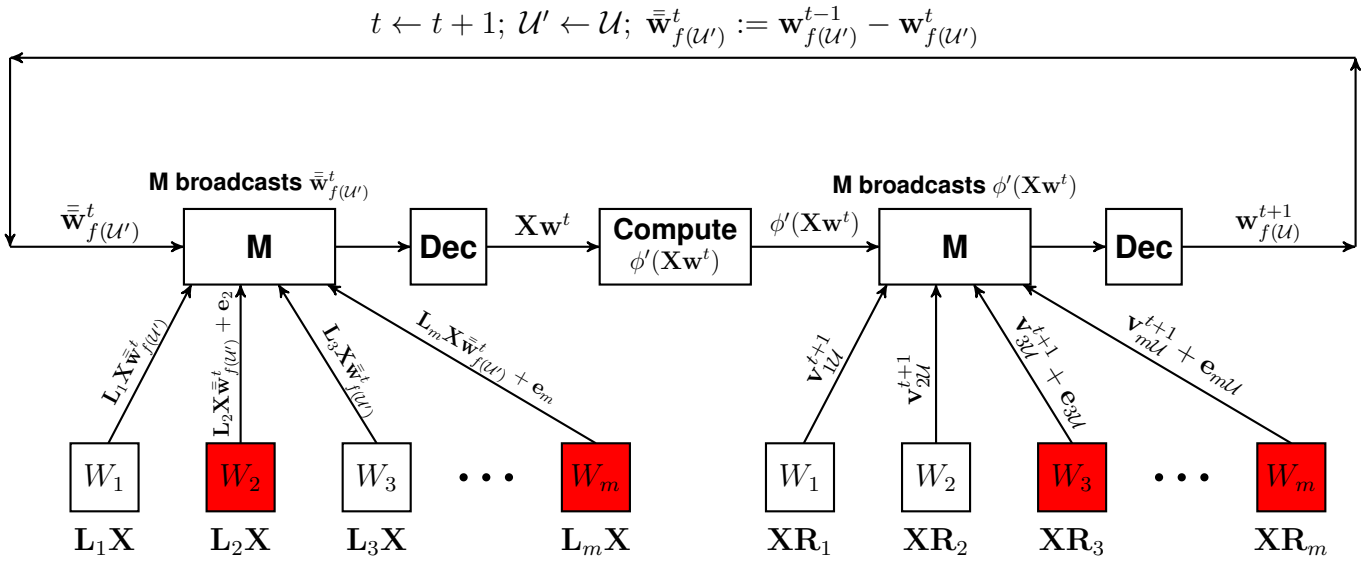

Fig. 2 This figure shows our 2-round approach to the Byzantine-resilient distributed coordinate descent (CD) for solving (15) using data encoding and real-error correction. We encode $\mathbf{X}$ with the encoding matrix $[\mathbf{R}_1\ \ldots\ \mathbf{R}_m] \in \mathbb{R}^{d\times p_2 m}$ and store $\widetilde{\mathbf{X}}^R_i := \mathbf{X}\mathbf{R}_i$ at the i'th worker and solve (16) over an enlarged parameter vector $\mathbf{v} \in \mathbb{R}^{p_2 m}$. At the t'th iteration, for some $\mathcal{U} \subseteq [p_2]$, the update at the i'th worker is $\mathbf{v}^{t+1}_{i\mathcal{U}} = \mathbf{v}^t_{i\mathcal{U}} - \alpha_t \mathbf{R}^T_{i\mathcal{U}}\mathbf{X}^T\phi'(\widetilde{\mathbf{X}}^R\mathbf{v}^t)$, which requires $\phi'(\widetilde{\mathbf{X}}^R\mathbf{v}^t)$, where $\widetilde{\mathbf{X}}^R\mathbf{v}^t = \mathbf{X}\mathbf{w}^t$. The first part of the figure is for providing $\phi'(\mathbf{X}\mathbf{w}^t)$ to every worker in each iteration so that they can update $\mathbf{v}^t_{i\mathcal{U}}$'s. For this, we encode $\mathbf{X}$ using the encoding matrix $[\mathbf{L}^T_1\ \ldots\ \mathbf{L}^T_m]^T \in \mathbb{R}^{p_1 m\times n}$ and store $\widetilde{\mathbf{X}}^L_i := \mathbf{L}_i\mathbf{X}$ at worker i. The encoding has the property that we can recover $\mathbf{X}\mathbf{w}^t$ from the erroneous vectors $\{\widetilde{\mathbf{X}}^L_i\mathbf{w}^t + \mathbf{e}_i\}_{i=1}^m$, where at most t of the $\mathbf{e}_i$'s are non-zero and can be arbitrary. We can make it computationally more efficient at the workers' side by observing that, in each iteration, only a subset of coordinates of $\mathbf{w}$ are being updated: suppose we updated $\mathbf{v}^t_{i\mathcal{U}'}$'s in the t'th iteration, which automatically updated $\mathbf{w}^t_{f(\mathcal{U}')}$. Since $\mathbf{w}^t_{[d]\setminus f(\mathcal{U}')}$ remain unchanged, we need to send only $\mathbf{w}^t_{f(\mathcal{U}')}$ to the workers – in the figure, to take care of a technicality, we let master broadcast $\bar{\bar{\mathbf{w}}}^t_{f(\mathcal{U}')} := \mathbf{w}^{t-1}_{f(\mathcal{U}')} - \mathbf{w}^t_{f(\mathcal{U}')}$, each worker i computes $\widetilde{\mathbf{X}}_i\bar{\bar{\mathbf{w}}}^t_{f(\mathcal{U}')}$ and sends it backs to the master. Since master keeps $\mathbf{X}\mathbf{w}^{t-1}$ from the previous iteration with itself, it can compute $\mathbf{X}\mathbf{w}^t$. The set of corrupt workers may be different in different rounds – the corrupt ones are shown in red color and they can send arbitrary outcomes to master. Once master has recovered $\mathbf{X}\mathbf{w}^t$, it computes $\phi'(\mathbf{X}\mathbf{w}^t)$ and broadcasts it; upon receiving it worker i updates $\mathbf{v}^{t+1}_{i\mathcal{U}}$ and sends it back. By **P.1**, this reflects an update on $\mathbf{w}^{t+1}_{f(\mathcal{U})}$ according to (19); and by **P.2**, the master can recover $\mathbf{w}^{t+1}_{f(\mathcal{U})}$.

Section IV-B below:

$$\mathbf{S}_i = \begin{bmatrix} b_{1i}\ldots b_{qi} & & & \\ & \ddots & & \\ & & b_{1i}\ldots b_{qi} & \\ & & & b_{1i}\ldots b_{li} \end{bmatrix}_{p\times d} \tag{20}$$

Here $q = (m-2t)$ and $l = d-(p-1)q$, where $p = \lceil \frac{d}{q} \rceil$. Note that $1 \le l < q$, and if q divides d, then $l = q$. All the unspecified entries of $\mathbf{S}_i$ are zero. By stacking up the $\mathbf{S}_i$'s gives us our desired encoding matrix $\mathbf{S} = [\mathbf{S}^T_1\ \mathbf{S}^T_2\ \ldots\ \mathbf{S}^T_m]^T$. Note that $b_{1i}, b_{2i}, \ldots, b_{qi}$ are such that if we let $\mathbf{b}_i = [b_{i1}\ b_{i2}\ldots b_{im}]^T$ for every $i \in [q]$, then $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_q\}$ is a set of orthonormal vectors. This implies that $\mathbf{S}$ is orthonormal, and, therefore, $\mathbf{S}^+ = \mathbf{S}^T$. By taking $\mathbf{R} = \mathbf{S}^T$, we have $\mathbf{R}^+ = \mathbf{S}$. Now we show that $\mathbf{S}$ satisfies **P.1**-**P.2**.

Our Encoding Satisfies P.1. We need to show a map $f : [p] \to \mathcal{P}([d])$ that satisfies **P.1**. Let us define the function f as follows, where $(q = m-2t)$ and $p = \lceil \frac{d}{q} \rceil$:

$$f(i) := \begin{cases} [(i-1)*q+1 : i*q] & \text{if } 1 \le i < p, \\ [(p-1)*q+1 : d] & \text{if } i = p, \end{cases} \tag{21}$$

and for any $\mathcal{U} \subseteq [p]$, we define $f(\mathcal{U}) := \cup_{i\in\mathcal{U}} f(i)$. It is clear from the definition of f that (i) $\{f(i) : i \in [p]\}$ partitions $[d]$; (ii) for every $i \in [p-1]$ we have $|f(i)| = q$, and that $|f(p)| \le q$. Recall that $q = m-2t$. For the 3rd property, note that, for any $\mathcal{U} \subseteq [p]$, all the columns of $\mathbf{S}_{i\mathcal{U}}$ whose indices belong to $[d] \setminus f(\mathcal{U})$ are identically zero, which implies that we have

$$\mathbf{S}_{i\mathcal{U}}\mathbf{w} = \mathbf{S}_{i\mathcal{U},f(\mathcal{U})}\mathbf{w}_{f(\mathcal{U})}, \qquad \text{for every } \mathbf{w} \in \mathbb{R}^d, \tag{22}$$

which in turn implies that

$$\mathbf{S}_{i\mathcal{U}}\mathbf{X}^T = \mathbf{S}_{i\mathcal{U},f(\mathcal{U})}\mathbf{X}^T_{f(\mathcal{U})}. \tag{23}$$

Since $\mathbf{S}^+ = \mathbf{S}^T$, we have $\mathbf{S}^+_{i\mathcal{U}} = \mathbf{S}^T_{i\mathcal{U}}$ for every $i \in [m]$ and every $\mathcal{U} \subseteq [p]$. With these, our update rule $\mathbf{v}^{t+1}_{i\mathcal{U}} = \mathbf{S}_{i\mathcal{U}}\mathbf{w}^t - \alpha_t\mathbf{S}_{i\mathcal{U}}\mathbf{X}^T\phi'(\mathbf{X}\mathbf{w}^t)$[19] can equivalently be written as

$$\mathbf{v}^{t+1}_{i\mathcal{U}} = \mathbf{S}_{i\mathcal{U},f(\mathcal{U})}\mathbf{w}^{t+1}_{f(\mathcal{U})}, \tag{24}$$

where

$$\mathbf{w}^{t+1}_{f(\mathcal{U})} = \mathbf{w}^t_{f(\mathcal{U})} - \alpha_t\mathbf{X}^T_{f(\mathcal{U})}\phi'(\mathbf{X}\mathbf{w}^t). \tag{25}$$

[19]We emphasize that we used $\mathbf{S}^+ = \mathbf{S}^T$ crucially to equivalently write our update rule $\mathbf{v}^{t+1}_{i\mathcal{U}} = \mathbf{R}^+_{i\mathcal{U}}\mathbf{w}^t - \alpha\mathbf{R}^T_{i\mathcal{U}}\mathbf{X}^T\phi'(\mathbf{X}\mathbf{w}^t)$ from (18) as $\mathbf{v}^{t+1}_{i\mathcal{U}} = \mathbf{S}_{i\mathcal{U}}\mathbf{w}^t - \alpha_t\mathbf{S}_{i\mathcal{U}}\mathbf{X}^T\phi'(\mathbf{X}\mathbf{w}^t)$. This follows because $\mathbf{S}^+ = \mathbf{S}^T$ and we take $\mathbf{R}^+ = \mathbf{S}$, which together imply that $\mathbf{R}^+_{i\mathcal{U}} = \mathbf{R}^T_{i\mathcal{U}} = \mathbf{S}_{i\mathcal{U}}$.

Observe that (25) is the same update rule as (19), which implies that if each worker i updates $\mathbf{v}_{i\mathcal{U}}$ according to the CD update rule, then the collective update at all the worker nodes automatically updates $\mathbf{w}_{f(\mathcal{U})}$ according the CD update rule. Now we show that our invariant $\mathbf{v}^{t+1} = \mathbf{S}\mathbf{w}^{t+1}$ is maintained. We show this by induction. Base case $\mathbf{v}^0 = \mathbf{S}\mathbf{w}^0$ holds by construction. For the inductive case, assume that $\mathbf{v}^t = \mathbf{S}\mathbf{w}^t$ holds at time t and we show $\mathbf{v}^{t+1} = \mathbf{S}\mathbf{w}^{t+1}$ holds at time $t+1$.

Define $\overline{\mathcal{U}} := [p] \setminus \mathcal{U}$ and $\overline{f(\mathcal{U})} := [d] \setminus f(\mathcal{U})$. Since we did not update $\mathbf{v}^t_{i\overline{\mathcal{U}}}$'s, we have $\mathbf{v}^{t+1}_{i\overline{\mathcal{U}}} = \mathbf{v}^t_{i\overline{\mathcal{U}}}$ for every $i \in [m]$. This, together with the inductive hypothesis (i.e., $\mathbf{v}^t = \mathbf{S}\mathbf{w}^t$), implies that

$$\mathbf{v}^{t+1}_{i\overline{\mathcal{U}}} = \mathbf{S}_{i\overline{\mathcal{U}}}\mathbf{w}^t. \tag{26}$$

Since $f(\overline{\mathcal{U}}) = \overline{f(\mathcal{U})}$, we have from (22) that

$$\mathbf{S}_{i\overline{\mathcal{U}}}\mathbf{w}^t = \mathbf{S}_{i\overline{\mathcal{U}},\overline{f(\mathcal{U})}}\mathbf{w}^t_{\overline{f(\mathcal{U})}}. \tag{27}$$

It is clear from (25) that $\mathbf{w}^t_{\overline{f(\mathcal{U})}}$ did not get an update when we updated $\mathbf{v}^t_{i\mathcal{U}}$'s, which implies that $\mathbf{w}^{t+1}_{\overline{f(\mathcal{U})}} = \mathbf{w}^t_{\overline{f(\mathcal{U})}}$. Substituting this in (27) gives $\mathbf{S}_{i\overline{\mathcal{U}}}\mathbf{w}^t = \mathbf{S}_{i\overline{\mathcal{U}},\overline{f(\mathcal{U})}}\mathbf{w}^{t+1}_{\overline{f(\mathcal{U})}}$, which, by (22), yields $\mathbf{S}_{i\overline{\mathcal{U}}}\mathbf{w}^t = \mathbf{S}_{i\overline{\mathcal{U}}}\mathbf{w}^{t+1}$. This, together with (26), implies

$$\mathbf{v}^{t+1}_{i\overline{\mathcal{U}}} = \mathbf{S}_{i\overline{\mathcal{U}}}\mathbf{w}^{t+1}. \tag{28}$$

We already have from (22) and (24) that

$$\mathbf{v}^{t+1}_{i\mathcal{U}} = \mathbf{S}_{i\mathcal{U}}\mathbf{w}^{t+1}. \tag{29}$$

Since (28) and (29) hold for every $i \in [m]$, we have $\mathbf{v}^{t+1} = \mathbf{S}\mathbf{w}^{t+1}$. Hence, the invariant is maintained.

Our Encoding Satisfies P.2. If we let

$$\mathbf{v}_{[m]\mathcal{U}} := [\mathbf{v}^T_{1\mathcal{U}}\ \mathbf{v}^T_{2\mathcal{U}} \dots \mathbf{v}^T_{m\mathcal{U}}]^T,$$
$$\mathbf{S}_{[m]\mathcal{U},f(\mathcal{U})} := [\mathbf{S}^T_{1\mathcal{U},f(\mathcal{U})}\ \mathbf{S}^T_{2\mathcal{U},f(\mathcal{U})} \dots \mathbf{S}^T_{m\mathcal{U},f(\mathcal{U})}]^T,$$

then the collective update (24) from all the workers can be written as

$$\mathbf{v}^{t+1}_{[m]\mathcal{U}} = \mathbf{S}_{[m]\mathcal{U},f(\mathcal{U})}\mathbf{w}^{t+1}_{f(\mathcal{U})}. \tag{30}$$

It is easy to verify that for every choice of $\mathcal{U} \subseteq [p]$, $\mathbf{S}_{[m]\mathcal{U},f(\mathcal{U})}$ is a full column-rank matrix, which implies that we can in principle recover the updated $\mathbf{w}^{t+1}_{f(\mathcal{U})}$ from $\mathbf{v}^{t+1}_{[m]\mathcal{U}} = \mathbf{S}_{[m]\mathcal{U},f(\mathcal{U})}\mathbf{w}^{t+1}_{f(\mathcal{U})}$. Now we show that not only can we recover $\mathbf{w}^{t+1}_{f(\mathcal{U})}$ from $\{\mathbf{S}_{i\mathcal{U},f(\mathcal{U})}\mathbf{w}^{t+1}_{f(\mathcal{U})}\}_{i=1}^m$, but also efficiently recover $\mathbf{w}^{t+1}_{f(\mathcal{U})}$ from the *erroneous* vectors $\{\mathbf{S}_{i\mathcal{U},f(\mathcal{U})}\mathbf{w}^{t+1}_{f(\mathcal{U})} + \mathbf{e}_{i\mathcal{U}}\}_{i=1}^m$, where at most t out of m error vectors $\{\mathbf{e}_{i\mathcal{U}}\}_{i=1}^m$ are non-zero and may have arbitrary entries. Let $\mathcal{U} = \{j_1, j_2, \dots, j_{|\mathcal{U}|}\}$, and for every $i \in [m]$, let $\mathbf{e}_{i\mathcal{U}} = [e_{ij_1}e_{ij_2}\dots e_{ij_{|\mathcal{U}|}}]^T$. Master equivalently writes $\{\mathbf{S}_{i\mathcal{U},f(\mathcal{U})}\mathbf{w}^{t+1}_{f(\mathcal{U})} + \mathbf{e}_{i\mathcal{U}}\}_{i=1}^m$ as $|\mathcal{U}|$ systems of linear equations.

$$\tilde{h}_i(\mathbf{w}^{t+1}_{f(\mathcal{U})}) = \tilde{\mathbf{S}}_{i,f(\mathcal{U})}\mathbf{w}^{t+1}_{f(\mathcal{U})} + \tilde{\mathbf{e}}_i, \quad i \in \mathcal{U}, \tag{31}$$

where, for every $i \in \mathcal{U}$, $\tilde{\mathbf{e}}_i = [e_{1i}, e_{2i}, \dots, e_{mi}]^T$ and $\tilde{\mathbf{S}}_{i,f(\mathcal{U})}$ is an $m \times |f(\mathcal{U})|$ matrix whose j'th row is equal to the i'th row of $\mathbf{S}_{j\mathcal{U}}$, for every $j \in [m]$. Note that at most t entries in each $\tilde{\mathbf{e}}_i$ are non-zero. Observe that $\{\mathbf{S}_{i\mathcal{U},f(\mathcal{U})}\mathbf{w}^{t+1}_{f(\mathcal{U})} + \mathbf{e}_{i\mathcal{U}}\}_{i=1}^m$ and $\{\tilde{\mathbf{S}}_{i,f(\mathcal{U})}\mathbf{w}^{t+1}_{f(\mathcal{U})} + \tilde{\mathbf{e}}_i\}_{i\in\mathcal{U}}$ are equivalent systems of linear equations, and we can get one from the other. Observe that (31) is similar to (8): $\tilde{\mathbf{S}}_{i,f(\mathcal{U})}$ is equal to $\tilde{\mathbf{S}}_i$ (for the same i) with some of its zero columns removed; and adding zero columns to $\tilde{\mathbf{S}}_{i,f(\mathcal{U})}$ will not change the value of $\tilde{h}_i(\mathbf{w}^{t+1}_{f(\mathcal{U})})$. Now, using the machinery developed in Section IV we can recover $\mathbf{w}^{t+1}_{f(\mathcal{U})}$ from (31) in $O(|\mathcal{U}|m^2)$ time.

B. Resource Requirement Analysis

In this section, first we give our algorithm developed for distributed coordinate descent in the presence of t (out of m) adversarial worker nodes, whose pictorial description is given in Figure 2.

We use two encoding matrices $\mathbf{L} \in \mathbb{R}^{(p_1 m)\times n}$ and $\mathbf{R} \in \mathbb{R}^{d\times(p_2 m)}$. Let $\mathbf{L} = [\mathbf{L}_1^T\ \mathbf{L}_2^T\ \dots\ \mathbf{L}_m^T]^T$ and $\mathbf{R} = [\mathbf{R}_1\ \mathbf{R}_2\ \dots\ \mathbf{R}_m]$, where each $\mathbf{L}_i$ is a $p_1 \times n$ matrix with $p_1 = \lceil \frac{n}{m-2t} \rceil$ and each $\mathbf{R}_i$ is a $d \times p_2$ matrix with $p_2 = \lceil \frac{d}{m-2t} \rceil$. Worker i stores both $\widetilde{\mathbf{X}}_i^L = \mathbf{L}_i\mathbf{X}$ and $\widetilde{\mathbf{X}}_i^R = \mathbf{X}\mathbf{R}_i$. Roughly, $\mathbf{L}$ is used to recover $\mathbf{X}\mathbf{w}$ from the erroneous $\{\mathbf{L}_i\mathbf{X}\mathbf{w} + \mathbf{e}_i\}_{i=1}^m$, and $\mathbf{R}$ is used to update the parameter vector reliably despite errors. Here $\mathbf{L}$ is a full column-rank matrix and $\mathbf{R}$ is a full row-rank matrix. Initialize with an arbitrary $\mathbf{w}^0$ and let $\mathbf{v}^0 = \mathbf{R}^+\mathbf{w}^0$. Repeat the following until convergence:

1) At iteration t, master sends $(\mathbf{w}^{t-1}_{f(\mathcal{U})} - \mathbf{w}^t_{f(\mathcal{U})})$[20] to all the workers (at $t = 0$, master sends $\mathbf{w}_0$), where $\mathcal{U} \subseteq [p_2]$ is the set of indices used for updating $\mathbf{v}^{t-1}_{i\mathcal{U}}$'s in the previous iteration, which in turn updated $\mathbf{w}^{t-1}_{f(\mathcal{U})}$; see (24) and (25) in Section V-A.
2) Worker i computes $\widetilde{\mathbf{X}}_i^L(\mathbf{w}^{t-1}_{f(\mathcal{U})} - \mathbf{w}^t_{f(\mathcal{U})}) = \mathbf{L}_i\mathbf{X}(\mathbf{w}^{t-1}_{f(\mathcal{U})} - \mathbf{w}^t_{f(\mathcal{U})})$ and sends it to the master.[21] Upon receiving $\{\widetilde{\mathbf{X}}_i^L(\mathbf{w}^{t-1}_{f(\mathcal{U})} - \mathbf{w}^t_{f(\mathcal{U})}) + \mathbf{e}_i\}_{i=1}^m$, where at most t of the $\mathbf{e}_i$'s are non-zero and may have arbitrary entries, the master applies the decoding procedure of Section IV and recovers $\mathbf{X}(\mathbf{w}^{t-1}_{f(\mathcal{U})} - \mathbf{w}^t_{f(\mathcal{U})})$. We assume that master keeps $\mathbf{X}\mathbf{w}^{t-1}$ from the previous iteration (which is equal to $\mathbf{0}$ if $t = 0$), it can compute $\mathbf{X}\mathbf{w}^t = \mathbf{X}\mathbf{w}^{t-1} - \mathbf{X}(\mathbf{w}^{t-1}_{f(\mathcal{U})} - \mathbf{w}^t_{f(\mathcal{U})})$. Note that if $t = 0$, this is equal to $\mathbf{X}\mathbf{w}^0$.
3) After obtaining $\mathbf{X}\mathbf{w}^t$, master computes $\phi'(\mathbf{X}\mathbf{w}^t)$, picks a subset $\mathcal{U} \subseteq [p_2]$ (randomly or in a round robin fashion to cover $[p_2]$ in a few iterations), and sends $(\phi'(\mathbf{X}\mathbf{w}^t), \mathcal{U})$ to all the workers.
4) Each worker node $i \in [m]$ updates $\mathbf{v}^{t+1}_{i\mathcal{U}} \leftarrow \mathbf{v}^t_{i\mathcal{U}} - \alpha_t\nabla_{i\mathcal{U}}\phi(\widetilde{\mathbf{X}}\mathbf{v}^t) = \mathbf{v}^t_{i\mathcal{U}} - \alpha_t(\widetilde{\mathbf{X}}^R_{i\mathcal{U}})^T\phi'(\mathbf{X}\mathbf{w}^t)$, while keeping the other coordinates of $\mathbf{v}^t_i$ unchanged. Worker i sends $\mathbf{v}^{t+1}_{i\mathcal{U}}$ to the master. Note that $\mathbf{v}^{t+1}_{i\mathcal{U}} = \mathbf{R}^+_{i\mathcal{U},f(\mathcal{U})}\mathbf{w}^{t+1}_{f(\mathcal{U})}$, where $\mathbf{w}^{t+1}_{f(\mathcal{U})} = [\mathbf{w}^t_{f(\mathcal{U})} - \alpha\mathbf{X}^T_{f(\mathcal{U})}\phi'(\mathbf{X}\mathbf{w}^t)]$; see (24) and (25) in Section V-A.

[20] Observe that master need not send the locations $f(\mathcal{U})$, because workers can compute those by themselves, as they know both $\mathcal{U}$ and the function f.

[21] With some abuse of notation, when we write $\mathbf{X}\mathbf{w}_{f(\mathcal{U})}$, we implicitly assume that $\mathbf{w}_{f(\mathcal{U})}$ is a length d vector, which has 0's in the indices that lie in $\overline{f(\mathcal{U})}$.

5) Upon receiving $\{\mathbf{v}_{i\mathcal{U}}^{t+1} + \mathbf{e}_{i\mathcal{U}}\}_{i=1}^{m}$, where at most t of the $\{\mathbf{e}_{i\mathcal{U}}\}_{i=1}^{m}$'s are non-zero and may have arbitrary entries, master applies the decoding procedure (since our encoding satisfies **P.2**) and recovers $\mathbf{w}_{f(\mathcal{U})}^{t+1}$.

Now we analyze the total amount of resources (storage, computation, and communication) required by the above algorithm and prove Theorem 2. Fix an $\epsilon > 0$. Let the corruption threshold t satisfy $t \leq \lfloor(\epsilon/(1+\epsilon)) \cdot (m/2)\rfloor$.

1) Storage Requirement:: By a similar analysis done in Section IV-E, we can show that the total storage at all worker nodes is roughly equal to $2(1+\epsilon)|\mathbf{X}|$.

2) Computational Complexity:: We can divide the computational complexity of our scheme as follows:

- *Encoding the data matrix.* By a similar analysis done in Section IV-E, we can show that the total encoding time is $O\left((\frac{\epsilon}{1+\epsilon}m+1)nd\right)$. Note that this encoding is to be done only once.
- *Computation at each worker node.* Suppose that in each iteration of our algorithm, all the workers update τ coordinates of $\mathbf{v}_i$'s. Fix an iteration t and assume that at iteration $(t-1)$, workers updated the coordinates in the set $\mathcal{U} \subseteq [p_2]$, where $|\mathcal{U}| = \tau$. Recall from **P.1** that updating $\tau = |\mathcal{U}|$ coordinates of each $\mathbf{v}_i^{t-1}$ automatically updates $\mathbf{w}_{f(\mathcal{U})}^{t-1}$. Upon receiving $(\mathbf{w}_{f(\mathcal{U})}^{t-1} - \mathbf{w}_{f(\mathcal{U})}^{t})$ from the master node, each worker i computes $\widetilde{\mathbf{X}}_i^{\mathbf{L}}(\mathbf{w}_{f(\mathcal{U})}^{t-1} - \mathbf{w}_{f(\mathcal{U})}^{t})$, and reports back the resulting vector. Note that $(\mathbf{w}_{f(\mathcal{U})}^{t-1} - \mathbf{w}_{f(\mathcal{U})}^{t})$ has at most $|f(\mathcal{U})| = \frac{\tau m}{1+\epsilon}$ non-zero elements, which together with that $\widetilde{\mathbf{X}}_i^{\mathbf{L}}$ is a $p_1 \times d$ matrix, implies that computing $\widetilde{\mathbf{X}}_i^{\mathbf{L}}(\mathbf{w}_{f(\mathcal{U})}^{t-1} - \mathbf{w}_{f(\mathcal{U})}^{t})$ takes $O(p_1 \cdot |f(\mathcal{U})|) = O(n\tau)$ time.[22] In the second round, given $\phi'(\mathbf{X}\mathbf{w}^t)$, since $(\widetilde{\mathbf{X}}_{i\mathcal{U}}^{R})^T$ is of dimension $n \times \tau$, updating $\mathbf{v}_{i\mathcal{U}}^{t}$ requires $O(n\tau)$ time, where $\tau = |\mathcal{U}|$. So, the total time taken by each worker is $O(n\tau)$.
- *Computation at the master node.* Once master receives $\{\mathbf{L}_i\mathbf{X}(\mathbf{w}_{f(\mathcal{U})}^{t-1} - \mathbf{w}_{f(\mathcal{U})}^{t}) + \mathbf{e}_i\}_{i=1}^{m}$, applying the decoding procedure of Section IV to obtain $\mathbf{X}(\mathbf{w}_{f(\mathcal{U})}^{t-1} - \mathbf{w}_{f(\mathcal{U})}^{t})$ from these erroneous vectors requires $O(p_1 m^2) = O((1+\epsilon)nm)$ time. After that obtaining $\mathbf{X}\mathbf{w}^t$ takes another $O(n)$ time. Given $\mathbf{X}\mathbf{w}^t$, computing $\phi'(\mathbf{X}\mathbf{w}^t)$ takes $O(n)$ time, assuming that computing $\ell'(\langle \mathbf{x}_i, \mathbf{w}^t\rangle; y_i)$ requires unit time, where $\langle \mathbf{x}_i, \mathbf{w}^t\rangle$ is equal to the i'th entry of $\mathbf{X}\mathbf{w}^t$. Upon receiving $\{\mathbf{v}_{i\mathcal{U}}^{t+1} + \mathbf{e}_{i\mathcal{U}}\}_{i=1}^{m}$, where $\mathbf{v}_{i\mathcal{U}}^{t+1} = \mathbf{R}_{i\mathcal{U},f(\mathcal{U})}^{+}\mathbf{w}_{f(\mathcal{U})}^{t+1}$, for all $i \in [m]$, recovering $\mathbf{w}_{f(\mathcal{U})}^{t+1}$ requires $O(\tau m^2)$ time. So, the total time taken by the master node is $O((1+\epsilon)nm + \tau m^2)$.

3) Communication Complexity:: Suppose workers update τ coordinates of $\mathbf{v}_i$'s in each iteration. Then (i) master broadcasts $\left(\frac{\tau m}{1+\epsilon} + n\right)$ real numbers, $\frac{\tau m}{1+\epsilon}$ in the first round to represent $\mathbf{w}_{f(\mathcal{U})}^{t}$ and n in the second round to represent $\phi'(\mathbf{X}\mathbf{w}^t)$; and (ii) each worker sends $\left(\tau + (1+\epsilon)\frac{n}{m}\right)$ real numbers, $(1+\epsilon)\frac{n}{m}$ in the first round for computing $\mathbf{X}\mathbf{w}^t$ at the master node and τ in the second iteration to represent $\mathbf{v}_{i\mathcal{U}}^{t}$.

[22] Note that in the very first iteration, master sends $\mathbf{w}^0$, which may be a dense length d vector, and computing $\widetilde{\mathbf{X}}_i\mathbf{L}\mathbf{w}^0$ at the i'th worker can take $O(p_1 d) = O((1+\epsilon)\frac{nd}{m})$ time. This is only for the first iteration.

VI. Extensions

In this section, we give a few important extensions of our coding scheme developed earlier in Section IV. First we give a Byzantine-resilient and communication-efficient method for stochastic gradient descent (SGD). Second we show how to exploit the specific structure of our encoding matrix to efficiently extend our coding technique to the streaming data model. In the end, we give a few more important applications, where our method can be applied constructively.

A. Stochastic Gradient Descent

Stochastic gradient descent (SGD) [56] is another alternative if full gradients are too costly to compute. In each iteration of SGD, we sample a data point uniformly at random, compute a gradient on that sample, and update the parameter vector based on that. We start with an arbitrary/random parameter vector $\mathbf{w}_0 \in \mathbb{R}^d$ and update it according the following update rule:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha_t \nabla f_{r_t}(\mathbf{w}_t), \quad t = 1, 2, 3, \ldots \tag{32}$$

where r_t is sampled uniformly at random from $\{1, 2, \ldots, n\}$. This ensures that the expected value of the gradient is equal to the true gradient. Due to its simplicity and remarkable empirical performance, SGD has become arguably the most widely-used optimization algorithm in many large-scale applications, especially in deep learning [14], [15], [57]. We want to run SGD in a distributed setup, where data is distributed among m worker nodes and at most t of them can be corrupt; see Section II-C for details on our adversary model.

Our solution. In the plain SGD, we sample a data point randomly and compute its gradient. So, we give a method in which, at any iteration t, master picks a random number r_t in $\{1, 2, \ldots, n\}$, broadcasts it, and recovers the r_t'th data point $\mathbf{x}_{r_t}$. Once the master has obtained $\mathbf{x}_{r_t}$, it can compute a gradient on it and updates the parameter vector. Since master recovers the data points, we can optimize for non-convex problems also; essentially, we could optimize anything that the plain SGD can. Our method is described below.

We encode $\mathbf{X}^T$ using the $\lceil d/(m-2t)\rceil \times d$ encoding matrix $\mathbf{S}^{(2)}$, which has been defined in Section IV-E. For simplicity, we denote $\mathbf{S}^{(2)}$ by $\mathbf{S}$. Let $\mathbf{S} = [\mathbf{S}_1^T\ \mathbf{S}_2^T\ \ldots\ \mathbf{S}_m^T]^T$. Note that the j'th worker stores $\mathbf{S}_j\mathbf{X}^T$. Let $\widetilde{\mathbf{X}} := \mathbf{S}\mathbf{X}^T$, which is a $\lceil d/(m-2t)\rceil \times n$ matrix, whose i'th column is the encoding $\widetilde{\mathbf{x}}_i := \mathbf{S}\mathbf{x}_i$ of the i'th data point $\mathbf{x}_i$. Using the method developed in Section IV, given $\{\mathbf{S}_j\mathbf{x}_i + \mathbf{e}_j\}_{j=1}^{m}$, where $\mathbf{e}_j = \mathbf{0}$ if the j'th worker is honest, otherwise can be arbitrary, master can recover $\mathbf{x}_i$ exactly in $O((1+\epsilon)md)$ time. Our main theorem is stated below, a proof of which trivially follows from Section IV.

Theorem 3 (Stochastic Gradient Descent)**.** *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ denote the data matrix. Let m denote the total number of worker nodes. We can compute a stochastic gradient in a distributed manner in the presence of t corrupt worker nodes and s stragglers, with the following guarantees, where $\epsilon > 0$ is a free parameter.*

- $(s+t) \leq \left\lfloor \frac{\epsilon}{1+\epsilon} \cdot \frac{m}{2} \right\rfloor$.
- *Total storage requirement is roughly* $(1+\epsilon)|\mathbf{X}|$.
- *Computational complexity for each stochastic gradient computation:*
  - *at each worker node is* $O((1+\epsilon)\frac{d}{m})$.
  - *at the master node is* $O((1+\epsilon)dm)$.
- *Communication complexity for each stochastic gradient computation:*
  - *each worker sends* $\left((1+\epsilon)\frac{d}{m}\right)$ *real numbers.*
  - *master broadcasts* $\lceil \log n \rceil$ bits.
- *Total encoding time is* $O\left(nd\left(\frac{\epsilon}{1+\epsilon}m+1\right)\right)$.

Observe the distributed gain of our method in the communication exchanged between the workers and the master: (i) master only broadcasts an index in $\{1,2,\ldots,n\}$, which only takes $\lceil \log n \rceil$ *bits*; and (ii) each worker sends roughly $\frac{1+\epsilon}{m}$ fraction of the total dimension d. Hence, this method is particularly useful in distributed settings with communication-constrained and band-limited links. The Remarks 2, 4, 5 are also applicable for Theorem 3.

Remark 10 (One-round vs. two-round approach)**.** *Unlike the two-round approach taken for gradient computation in PGD and also for CD, we give a one-round approach for each iteration of SGD. This is because in each SGD iteration we need to compute the gradient on* only one *data point (not the entire dataset, as in the case for each PGD iteration). Because of this, recovering a (random) data-point itself at the master and then computing a gradient on it locally (which is what we do) would be far more efficient than computing gradient on a single data point in a distributed manner. This is in contrast to each gradient computation for PGD, which requires computation of the full gradient (which is the summation of gradients on all n data points). In principle, we can use the one-round for each PGD iteration also in which first we recover all the n data points at master and then compute the full gradient locally, but this approach would defeat the purpose of distributed computation both in terms of storage and computational complexity. Note that our two-round approach for PGD is significantly more efficient than this.*

The reason behind taking the two-round approach for CD is because in order to update the local parameter vectors in the t'th iteration, workers need access to the MV multiplication $\widetilde{\mathbf{X}}^R\mathbf{v}^t = \mathbf{X}\mathbf{w}^t$ (see the paragraph after (18) *for more details), and in order to provide that we use an extra round – the first round is used for computing $\mathbf{X}\mathbf{w}$ and the second round is used for updating the local parameter vectors. Again, for CD also, we could adopt a one-round approach where master recovers all the n data points and then do the parameter update, but that would be highly inefficient and defeat the purpose of distributed computation.*

One of the main advantages of the one-round approach for SGD is that since we are recovering the data point itself at the master, we can use it to optimize any function, both convex and non-convex. This is in contrast to the two-round approach, which can only be used for generalized linear models only.

B. Encoding in The Streaming Data Model

An attractive property of our encoding scheme is that it is very easy to update with new data points. More specifically, our encoding requires the same amount of time, irrespective of whether we get all the data at once, or we get each sample point one by one, as in the online/streaming model. This setting encompasses a more realistic scenario, in which we design our coding scheme with the initial set of data points and distribute the encoded data among the workers. Later on, when we get some more samples, we can easily incorporate them into our existing encoded data. We show that updating $(m-2t)$ new data points in $\mathbb{R}^d$ requires $O\left((m-2t)\left((2t+1)d\right)\right)$ time in total, i.e., $O\left((2t+1)d\right)$ amortized-time per data point. This is the best one can hope for, since the offline encoding of n data points requires $O\left((2t+1)nd\right)$ total time. At the end of the update, the final encoded matrix that we get is the same as the one we would have got had we had all the $n+1$ data points in the beginning. Therefore, the decoding is not affected by this method at all. Note that we use the same encoding matrices both for gradient computation as well as for coordinate descent. So, it suffices to prove our result in the streaming model for any one of them, and we show it for gradient computation below.

Theorem 4. *The total time complexity in encoding all the data points at once, i.e., when encoding is done offline, is the same as the total time complexity in encoding the data points one by one as they come in the streaming model, i.e., when encoding is done online.*

Proof. Let $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ denote the encoding matrices for encoding $\mathbf{X}$ and $\mathbf{X}^T$, respectively; see Section IV-B. For convenience, we copy over the corresponding encoding matrices $\mathbf{S}_i^{(1)}$ and $\mathbf{S}_i^{(2)}$ from (11) for the i'th worker node in Figure 3.

Suppose at some point of time we have encoded n data points each lying in $\mathbb{R}^d$ and distributed the encoded data among the m worker nodes. Now a new data sample $\mathbf{x} \in \mathbb{R}^d$ comes in. We will show how to incorporate it in the existing scheme in $O\left((2t+1)d\right)$ time on average.

Updating the encoding matrices. Fix an arbitrary worker $i \in [m]$. Note that the new data matrix $\mathbf{X}$ has dimension $(n+1)\times d$. So, the new encoding matrix $\mathbf{S}_i^{(1)}$ should have $(n+1)$ columns, and we have to add one more column to $\mathbf{S}_i^{(1)}$. By examining the repetitive structure of $\mathbf{S}_i^{(1)}$, it is obvious which column to add: if $l_1 < q$, then we add the p_1-dimensional vector $[0,0,\ldots,0,b_{(l_1+1)i}]^T$ as the last column; otherwise, if $l_1 = q$, then we add the

$$\mathbf{S}_i^{(1)} = \begin{bmatrix} b_{1i} \dots b_{qi} & & & \\ & \ddots & & \\ & & b_{1i} \dots b_{qi} & \\ & & & b_{1i} \dots b_{l_1 i} \end{bmatrix}_{p_1 \times n}$$

(a)

$$\mathbf{S}_i^{(2)} = \begin{bmatrix} b_{1i} \dots b_{qi} & & & \\ & \ddots & & \\ & & b_{1i} \dots b_{qi} & \\ & & & b_{1i} \dots b_{l_2 i} \end{bmatrix}_{p_2 \times d}$$

(b)

Fig. 3 Figure 3a depicts the encoding matrix for the i'th worker node for encoding $\mathbf{X}$, which is used in the first round of the gradient computation. Here $p_1 = \lceil n/q \rceil$, where $q = (m-k)$ and k is equal to the number of rows in the error recovery matrix $\mathbf{F}$ in (14), and $l_1 = n - (p_1 - 1)q$. Figure 3b depicts the encoding matrix for the i'th worker node for encoding $\mathbf{X}^T$, which is used in the second round of the gradient computation. Here $p_2 = \lceil d/q \rceil$ and $l_2 = d - (p_2 - 1)q$. All the unspecified entries in both the matrices are zero.

$(p_1 + 1)$-dimensional vector $[0, 0, \dots, 0, b_{1i}]^T$ as the last column. In the second case, the number of rows of $\mathbf{S}_i^{(1)}$ increases by one – the last row has all zeros, except for the last element, which is equal to b_{1i}. Note that $\mathbf{S}_i^{(2)}$ does not change at all. Observe that if the i'th worker performs this update, then it does not have to store its entire encoding matrix $\mathbf{S}_i^{(1)}$, it only needs to store n, $q = (m-k)$, and the q real numbers $b_{1i}, b_{2i}, \dots, b_{qi}$, where $q = m - k$, which could be much smaller as compared to n and d, and are enough to define $\mathbf{S}_i^{(1)}$ and $\mathbf{S}_i^{(2)}$.

Updating the encoded data. Now we show how to update the encoded data with the new sample $\mathbf{x}$. We need to update both $\mathbf{S}_i^{(1)}\mathbf{X}$ as well as $\mathbf{S}_i^{(2)}\mathbf{X}^T$ for every worker $i \in [m]$.

- *Updating* $\mathbf{S}_i^{(1)}\mathbf{X}$. If $l_1 < q$, then we add $b_{(l_1+1)i}\mathbf{x}^T$ to the last row of $\mathbf{S}_i^{(1)}\mathbf{X}$; otherwise, if $l_1 = q$, then we add $b_{1i}\mathbf{x}$ as a new row in $\mathbf{S}_i^{(1)}\mathbf{X}$. In the first case, the resulting matrix still has p_1 rows, whose first $p_1 - 1$ rows are same as before, and the last row is the sum of the previous row and $b_{(l_1+1)i}\mathbf{x}^T$. In the second case, the resulting matrix has $(p_1 + 1)$ rows, whose first p_1 rows are the same as before and the last row is equal to $b_{1i}\mathbf{x}^T$. Note that each row of $\mathbf{S}_i^{(1)}$ for $i \leq 2t$, has at most $(m - 2t)$ non-zero elements; whereas, for $i > 2t$, each row of $\mathbf{S}_i^{(1)}$ has exactly one non-zero entry. Since there are $p_1 = \lceil n/(m-2t) \rceil$ rows in each $\mathbf{S}_i^{(1)}$, updating $\mathbf{S}_i^{(1)}\mathbf{X}$ for every $i \leq 2t$ takes $O(d)$ time; and for $i > 2t$, update in $\mathbf{S}_i^{(1)}\mathbf{X}$ happens only once in $(m - 2t)$ new data points (whenever the second case occurs and the resulting $\mathbf{S}_i^{(1)}$ has $(p_1 + 1)$ rows). So, updating $(m-2t)$ data points at all m worker nodes require $O\left(2t * (m-2t)d + (m-2t) * d\right) = O((m-2t)(2t+1)d)$ time, i.e., $O\left((2t+1)d\right)$ time per data point.
- *Updating* $\mathbf{S}_i^{(2)}\mathbf{X}^T$. Note that $\mathbf{X}^T$ is a $d \times (n+1)$ matrix whose last column is equal to the new data sample $\mathbf{x}$. Now, to update $\mathbf{S}_i^{(2)}\mathbf{X}^T$, we add $\mathbf{S}_i^{(2)}\mathbf{x}$ as an extra column. The resulting matrix is of size $p_2 \times (n+1)$, whose first n columns are the same as before and the last column is equal to $\mathbf{S}_i^{(2)}\mathbf{x}$. Since total number of non-zero entries in $\mathbf{S}_i^{(2)}$ is equal to d if $i \leq 2t$ and equal to $p_2 = \lceil d/(m-2t) \rceil$ if $i > 2t$, the total time required to update a new data point is $O(2t * d + (m-2t) * p_2) = O\left((2t+1)d\right)$.

Observe that at the end of this local update at each worker node, the final encoded matrix that we get is the same as the one we would have got had we had all the $n+1$ data points in the beginning. The decoding is not affected by this method at all. This completes the proof of Theorem 4. □

Remark 11 (Updating the encoded data efficiently with new features)**.** *Observe that since we encode both X and X^T in an analogous fashion, it follows by symmetry that we can not only update efficiently upon receiving a new data sample, but can also update efficiently if we decide to enlarge the dimension d of each of the n data samples at some point of time – maybe we figure out some new features of the data to get a more accurate model to overcome under-fitting. In these situations, we don't need to encode the entire dataset all over again, just a simple update is enough to incorporate the changes.*

Remark 12 (What allows our encoding to be efficient for streaming data?)**.** *The efficient update property of our coding scheme is made possible by the repetitive structure of our encoding matrix (see Figure 3), together with the fact that this structure is independent of the number of data points n and the dimension d – it only depends on the number of worker nodes m and the corruption threshold t. We remark that other data encoding methods in literature, even for weaker models, do not support efficient update. For example, the encoding of [41], which was designed for mitigating stragglers, depends on the dimensions n and d of the data matrix. So, it may not efficiently update if a new data point comes in.*

C. More Applications.

There are many iterative algorithms, other than the gradient descent for learning GLMs, which use repeated MV multiplication. Some of them include (i) the power method for computing the largest eigenvalue of a diagonalizable matrix, which is used in Google's PageRank algorithm [58], Twitter's recommendation system [59], etc.; (ii) iterative methods for solving sparse linear systems [60]; (iii) many graph algorithms, where the graph is represented by a fixed adjacency matrix, [61]. In large-scale implementation of these systems, where Byzantine faults are inevitable, the method described in this paper can be of interest.

In most of these applications, the underlying matrix $\mathbf{A}$ is generally sparse, which is exploited to gain computational efficiency. So, it is desired not to lose sparsity even if we want resiliency against Byzantine attacks. Fortunately, our encoding matrix $\mathbf{S}$ is sparse (see (11)), which ensures that the

encoded matrix $\mathbf{SA}$ will not lose the sparsity of $\mathbf{A}$: Each of the first pk rows of $\mathbf{S}$ has at most $(m-k)$ (where $k = 2t$) non-zero elements, and each of the remaining rows has exactly one 1. Since m is the number of worker nodes, which may be small, and we can take t to be up to $\lfloor \frac{m-1}{2} \rfloor$, we may have a few non-zero entries in each row of $\mathbf{S}$ (in the extreme case when $2t = m-1$, each row of $\mathbf{S}$ has only one non-zero entry). In a sense, we are getting Byzantine-resiliency almost for free without compromising the computational efficiency that is made possible by the sparsity of the matrix.

VII. Numerical Experiments

In this section, we validate the efficacy of our proposed methods by numerical experiments. We run distributed gradient descent (GD) and coordinate descent (CD) for linear regression $\arg\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2$. As mentioned in Section II-A, for linear regression (which is equal to ridge regression when $h = 0$), the projected gradient descent (PGD) reduces to gradient descent (GD). Since we are doing exact computation (computing the gradients exactly in the case of GD and updating the coordinates exactly in the case of CD), (i) there is no need to check the convergence, and (ii) our algorithm will perform exactly the same whether we are working with synthetic datasets or real datasets, hence, we will work with a synthetic dataset. We run our algorithms[23] with $m = 15$ worker nodes on two datasets: $(n = 10,000, d = 250)$ and $(n = 20,000, d = 22,000)$. For both the datasets, we generate $(\mathbf{X}, \mathbf{y})$ by sampling $\mathbf{X} \leftarrow \mathcal{N}(0, I)$ and $\mathbf{y} = \mathbf{X}\theta + \mathbf{z}$, where $\theta \in \mathbb{R}^d$ has $d/3$ non-zero entries, all of them are i.i.d. according to $\mathcal{N}(0, 4)$, and each entry of $\mathbf{z} \in \mathbb{R}^n$ is sampled from $\mathcal{N}(0, 1)$ i.i.d. In each round of the gradient computation, the adversary picks t worker nodes uniformly at random, and adds independent random vectors of appropriate length as errors, whose entries are sampled from $\mathcal{N}(0, \sigma^2)$ i.i.d. with $\sigma = 100$, to the true vectors.

A. $n = 10,000, d = 250, m = 15$

In Figure 4, we plot the total time taken (which is the sum of the maximum time taken by any single worker node and the time taken by the master node in both rounds) for updating different number of coordinates in one CD iteration, with varying number of corrupt worker nodes from $t = 1$ to $t = 7$. We plot the time needed for updating γ-fraction of d coordinates for four different values of γ (i.e., $\gamma = 0.1, 0.25, 0.5, 1$) and we denote it by CD(γd) for $\gamma = 0.1, 0.25, 0.5, 1$. Recall that CD($d$) is equivalent to full gradient computation as in the case of GD. Note that, when $t = 7$, we have $\epsilon = m - 1$, which is the main cause behind the significant increment in time for $t = 7$.

[23]We implement our algorithm in Python, and run it on an iMac machine with 3.8 GHz Quad-Core Intel Core i5 processor and 16 GB 2400 MHz DDR4 memory.

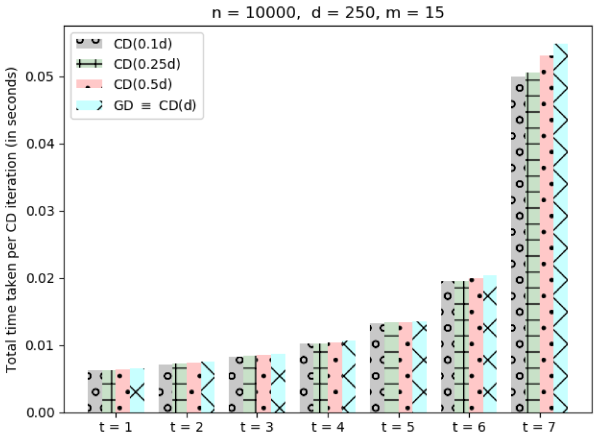

Fig. 4 We run our algorithms (CD and GD) with 15 worker nodes on a dataset with $n = 10,000, d = 250$. This plot reports how the total time taken (in seconds) for updating different number of coordinates in each CD iteration changes with varying number of corrupt worker nodes from $t = 1$ to $t = 7$. In the figure, we plot the total time taken (per iteration) for updating the γ-fraction of d coordinates for $\gamma = 0.1, 0.25, 0.5, 1$. Note that CD with $\gamma = 1$ is equivalent to full gradient computation as in GD.

B. $n = 20,000, d = 22,000, m = 15$

In Figure 5, we report separately, the maximum time taken by any single worker node and the time taken by the master node (together in both the rounds) in one CD iteration for updating different number of coordinates and also for GD, with varying number of corrupt worker nodes from $t = 1$ to $t = 6$. As in the above case, we report the time needed for updating γ-fraction of d coordinates for four different values of γ. Observe that the time taken by the master node is orders of magnitude less than the time taken by the worker nodes. We can also observe that with the running time in a worker node per iteration for CD($0.1d$) is 95% less than that for GD, while this time saving in the master node is more than 40%.

References

[1] D. Data, L. Song, and S. N. Diggavi, "Data encoding for byzantine-resilient distributed gradient descent," in *Allerton Conference on Communication, Control, and Computing, Allerton 2018*, 2018, pp. 863–870.

[2] ——, "Data encoding methods for byzantine-resilient distributed optimization," in *IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 2719–2723.

[3] D. Data and S. N. Diggavi, "Byzantine-tolerant distributed coordinate descent," in *IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 2724–2728.

[4] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008.

[5] M. Zinkevich, M. Weimer, L. Li, and A. J. Smola, "Parallelized stochastic gradient descent," in *Advances in neural information processing systems*, 2010, pp. 2595–2603.

[6] X. Lian, C. Zhang, H. Zhang, C. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent," in *Advances in Neural Information Processing Systems, NIPS 2017, 4-9 December 2017, Long Beach, CA, USA*, 2017, pp. 5336–5346.

[7] T. F. Abdelzaher *et al.*, "Will distributed computing revolutionize peace? the emergence of battlefield iot," in *ICDCS 2018*, 2018, pp. 1129–1138.

[8] J. Konecný, "Stochastic, distributed and federated optimization for machine learning," Ph.D. dissertation, University of Edinburgh, 2017.

	CD($0.1d$)		CD($0.25d$)		CD($0.5d$)		GD $\equiv$ CD(d)	
	Worker	Master	Worker	Master	Worker	Master	Worker	Master
$t = 1$	0.0020	0.0120	0.0073	0.0182	0.0122	0.0199	0.0493	0.0214
$t = 2$	0.0044	0.0187	0.0092	0.0212	0.0188	0.0277	0.0953	0.0393
$t = 3$	0.0054	0.0201	0.0118	0.0242	0.0269	0.0324	0.1213	0.0561
$t = 4$	0.0063	0.0253	0.0159	0.0327	0.0488	0.0468	0.1602	0.0610
$t = 5$	0.0107	0.0342	0.0328	0.0460	0.0776	0.0738	0.2943	0.0826
$t = 6$	0.0205	0.0717	0.0764	0.0833	0.1330	0.1088	0.8929	0.1227

Fig. 5 We run our algorithms (CD and GD) with 15 worker nodes on a dataset with $n = 20,000, d = 22,000$, and separately report the maximum time taken by any single worker and the master per iteration against varying number of corrupt worker nodes from $t = 1$ to 6. For CD, we run our algorithm for updating different number of coordinates. The first two columns correspond to the case when updating 0.1-fraction of d coordinates, the next two columns for 0.25-fraction, and so on. The last two columns correspond to updating all the coordinates, which is equivalent to full gradient computation as in GD.

[9] L. Lamport, R. Shostak, and M. Pease, "The byzantine generals problem," *ACM Trans. Program. Lang. Syst.*, vol. 4, no. 3, pp. 382–401, Jul. 1982.

[10] P. Blanchard, E. M. E. Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Advances in Neural Information Processing Systems, NIPS 2017, 4-9 December 2017, Long Beach, CA, USA*, 2017, pp. 118–128.

[11] L. Chen, H. Wang, Z. B. Charles, and D. S. Papailiopoulos, "DRACO: byzantine-resilient distributed training via redundant gradients," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, 2018, pp. 902–911.

[12] Y. Chen, L. Su, and J. Xu, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," *POMACS*, vol. 1, no. 2, pp. 44:1–44:25, 2017.

[13] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1989.

[14] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010. Physica-Verlag HD*, 2010.

[15] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, M. Z. Mao, M. Ranzato, A. W. Senior, P. A. Tucker, K. Yang, and A. Y. Ng, "Large scale distributed deep networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1232–1240.

[16] J. K. Bradley, A. Kyrola, D. Bickson, and C. Guestrin, "Parallel coordinate descent for l1-regularized loss minimization," in *ICML*, 2011, pp. 321–328.

[17] S. J. Wright, "Coordinate descent algorithms," *Math. Program.*, vol. 151, no. 1, pp. 3–34, 2015.

[18] P. Richtárik and M. Takáč, "Parallel coordinate descent methods for big data optimization," *Mathematical Programming*, vol. 156, no. 1, pp. 433–484, Mar 2016.

[19] Y. Nesterov, "Efficiency of coordinate descent methods on huge-scale optimization problems," *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 341–362, 2012.

[20] D. Yin, Y. Chen, K. Ramchandran, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, 2018, pp. 5636–5645.

[21] D. Alistarh, Z. Allen-Zhu, and J. Li, "Byzantine stochastic gradient descent," in *Neural Information Processing Systems (NeurIPS)*, 2018, pp. 4618–4628.

[22] L. Su and J. Xu, "Securing distributed gradient descent in high dimensional statistical learning," *POMACS*, vol. 3, no. 1, pp. 12:1–12:41, 2019.

[23] C. Xie, S. Koyejo, and I. Gupta, "Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance," in *International Conference on Machine Learning (ICML)*, 2019, pp. 6893–6901.

[24] D. Yin, Y. Chen, K. Ramchandran, and P. L. Bartlett, "Defending against saddle point attack in byzantine-robust distributed learning," in *ICML*, 2019, pp. 7074–7084.

[25] N. Gupta and N. H. Vaidya, "Byzantine fault-tolerant parallelized stochastic gradient descent for linear regression," in *57th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2019, Monticello, IL, USA, September 24-27, 2019*. IEEE, 2019, pp. 415–420.

[26] S. Rajput, H. Wang, Z. B. Charles, and D. S. Papailiopoulos, "DETOX: A redundancy-based framework for faster and more robust gradient aggregation," in *NeurIPS*, 2019, pp. 10 320–10 330.

[27] L. Li, W. Xu, T. Chen, G. B. Giannakis, and Q. Ling, "RSA: byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets," in *Conference on Artificial Intelligence (AAAI)*, 2019, pp. 1544–1551.

[28] A. Ghosh, J. Hong, D. Yin, and K. Ramchandran, "Robust federated learning in a heterogeneous environment," *CoRR*, vol. abs/1906.06629, 2019. [Online]. Available: http://arxiv.org/abs/1906.06629

[29] Q. Yu, S. Li, N. Raviv, S. M. M. Kalan, M. Soltanolkotabi, and A. S. Avestimehr, "Lagrange coded computing: Optimal design for resiliency, security, and privacy," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019, pp. 1215–1225.

[30] D. Data and S. N. Diggavi, "Byzantine-resilient SGD in high dimensions on heterogeneous data," *CoRR*, vol. abs/2005.07866, 2020. [Online]. Available: https://arxiv.org/abs/2005.07866

[31] ——, "Byzantine-resilient high-dimensional SGD with local iterations on heterogeneous data," *CoRR*, vol. abs/2006.13041, 2020. [Online]. Available: https://arxiv.org/abs/2006.13041

[32] L. He, S. P. Karimireddy, and M. Jaggi, "Byzantine-robust learning on heterogeneous datasets via resampling," *CoRR*, vol. abs/2006.09365, 2020. [Online]. Available: https://arxiv.org/abs/2006.09365

[33] K. Lee, M. Lam, R. Pedarsani, D. S. Papailiopoulos, and K. Ramchandran, "Speeding up distributed machine learning using codes," *IEEE Trans. Information Theory*, vol. 64, no. 3, pp. 1514–1529, 2018.

[34] S. Dutta, V. R. Cadambe, and P. Grover, "Short-dot: Computing large linear transforms distributedly using coded short dot products," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016, pp. 2092–2100.

[35] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[36] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[37] R. Tibshirani, "Convex optimization lecture notes," http://www.stat.cmu.edu/~ryantibs/convexopt-S15/scribes/08-prox-grad-scribed.pdf, 2015.

[38] M. Jaggi, "An equivalence between the lasso and support vector machines," *CoRR*, vol. abs/1303.1152, 2013. [Online]. Available: http://arxiv.org/abs/1303.1152

[39] S. Shalev-Shwartz and A. Tewari, "Stochastic methods for l_1-regularized loss minimization," *Journal of Machine Learning Research*, vol. 12, pp. 1865–1892, 2011.

[40] J. Dean and L. A. Barroso, "The tail at scale," *Commun. ACM*, vol. 56, no. 2, pp. 74–80, Feb. 2013.

[41] C. Karakus, Y. Sun, S. N. Diggavi, and W. Yin, "Straggler mitigation in distributed optimization through data encoding," in *In Advances in Neural Information Processing Systems, NIPS 2017, 4-9 December 2017, Long Beach, CA, USA*, 2017, pp. 5440–5448.

[42] E. M. E. Mhamdi, R. Guerraoui, and S. Rouault, "The hidden vulnerability of distributed learning in byzantium," in *Proceedings of*

the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, 2018, pp. 3518–3527.

[43] R. Tandon, Q. Lei, A. G. Dimakis, and N. Karampatziakis, "Gradient coding: Avoiding stragglers in distributed learning," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, 2017, pp. 3368–3376.

[44] N. Raviv, R. Tandon, A. Dimakis, and I. Tamo, "Gradient coding from cyclic MDS codes and expander graphs," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, 2018, pp. 4302–4310.

[45] Z. B. Charles and D. S. Papailiopoulos, "Gradient coding using the stochastic block model," in *2018 IEEE International Symposium on Information Theory, ISIT 2018, Vail, CO, USA, June 17-22, 2018*, 2018, pp. 1998–2002.

[46] W. Halbawi, N. A. Ruhi, F. Salehi, and B. Hassibi, "Improving distributed gradient descent using reed-solomon codes," in *2018 IEEE International Symposium on Information Theory, ISIT 2018, Vail, CO, USA, June 17-22, 2018*, 2018, pp. 2027–2031.

[47] S. Dutta, V. R. Cadambe, and P. Grover, ""short-dot": Computing large linear transforms distributedly using coded short dot products," *IEEE Trans. Inf. Theory*, vol. 65, no. 10, pp. 6171–6193, 2019.

[48] C. Karakus, Y. Sun, S. N. Diggavi, and W. Yin, "Redundancy techniques for straggler mitigation in distributed optimization and learning," *J. Mach. Learn. Res.*, vol. 20, pp. 72:1–72:47, 2019.

[49] A. Shamir, "How to share a secret," *Commun. ACM*, vol. 22, no. 11, pp. 612–613, 1979.

[50] R. Cramer, I. Damgård, and J. B. Nielsen, *Secure Multiparty Computation and Secret Sharing*. Cambridge University Press, 2015.

[51] M. Mishali and Y. C. Eldar, "Reduce and boost: Recovering arbitrary sets of jointly sparse vectors," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 4692–4702, Oct 2008.

[52] P. Billingsley, *Probability and Measure*, ser. Wiley Series in Probability and Statistics. Wiley, 1995.

[53] T. F. Coleman and A. Pothen, "The null space problem I. complexity," *SIAM Journal on Algebraic Discrete Methods*, vol. 7, no. 4, pp. 527–537, 1986.

[54] K. M. Hoffman and R. Kunze, *Linear algebra*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[55] M. Akçakaya and V. Tarokh, "A frame construction and a universal distortion bound for sparse representations," *IEEE Trans. Signal Processing*, vol. 56, no. 6, pp. 2443–2450, 2008.

[56] R. Herbert and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics. JSTOR, www.jstor.org/stable/2236626.*, vol. vol. 22, no. 3, pp. 400–407, 1951.

[57] A. Rakhlin, O. Shamir, and K. Sridharan, "Making gradient descent optimal for strongly convex stochastic optimization," in *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012.

[58] I. Ipsen and R. S. Wills, "Mathematical properties and analysis of google's pagerank," *Boletín de la Sociedad Española de Matemática Aplicada*, vol. 34, pp. 191–196, 01 2006.

[59] P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh, "Wtf: The who to follow service at twitter," in *Proceedings of the 22Nd International Conference on World Wide Web*, ser. WWW '13. New York, NY, USA: ACM, 2013, pp. 505–514.

[60] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2003.

[61] J. Kepner and J. Gilbert, *Graph Algorithms in the Language of Linear Algebra*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2011.

Deepesh Data Deepesh Data graduated from the International Institute of Information Technology, Hyderabad, India, with a B. Tech. degree in Computer Science and Engineering, in 2011. He received M.Sc. and Ph.D. degrees from the School of Technology and Computer Science at the Tata Institute of Fundamental Research, Mumbai, India, in 2017.

After that, he joined the Indian Institute of Technology Bombay as a post-doctoral fellow. Since March 2018, he has been with the University of California, Los Angeles, as a post-doctoral scholar. His research interests are in distributed optimization, machine learning, differential privacy, cryptography, algorithms, and information theory. He has received the Microsoft Research India Ph.D. Fellowship, the ACM India Doctoral Dissertation Award (Honorable Mention), and the TIFR-Sasken Best Ph.D. Thesis Award in Technology and Computer Sciences.

Linqi Song Linqi Song is an Assistant Professor in the Computer Science Department at the City University of Hong Kong. Prior to that, he was a Postdoctoral Scholar in the Department of Electrical and Computer Engineering at the University of California, Los Angeles (UCLA). He received the Ph.D. degree in Electrical Engineering from UCLA, and the B.S. and M.S. degrees from Tsinghua University. His research interests encompass information theory and coding theory, communications, machine learning, and big data. He has received the Hong Kong RGC Early Career Scheme in 2019 and the Best Paper Award at IEEE MIPR 2020.

Suhas N. Diggavi Suhas N. Diggavi received the B. Tech. degree in electrical engineering from the Indian Institute of Technology, Delhi, India, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 1998. After completing his Ph.D., he was a Principal Member Technical Staff in the Information Sciences Center, AT&T Shannon Laboratories, Florham Park, NJ. After that he was on the faculty of the School of Computer and Communication Sciences, EPFL, where he directed the Laboratory for Information and Communication Systems (LICOS). He is currently a Professor, in the Department of Electrical Engineering, at the University of California, Los Angeles, where he directs the Information Theory and Systems laboratory.

His research interests include information theory and its applications to several areas including learning, security and privacy, data compression, wireless networks, cyber-physical systems, genomics and neuroscience; more information can be found at http://licos.ee.ucla.edu. He has received several recognitions for his research including 2013 IEEE Information Theory Society & Communications Society Joint Paper Award, the 2013 ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc) best paper award, the 2006 IEEE Donald Fink prize paper award and the 2019 Google Faculty Research Award. He served as a Distinguished Lecturer and also currently serves on board of governors for the IEEE Information theory society. He is a Fellow of the IEEE.

He has been an associate editor for IEEE Transactions on Information Theory, ACM/IEEE Transactions on Networking, IEEE Communication Letters, a guest editor for IEEE Selected Topics in Signal Processing and in the program committees of several IEEE conferences. He has also helped organize IEEE and ACM conferences including serving as the Technical Program Co-Chair for 2012 IEEE Information Theory Workshop (ITW), the Technical Program Co-Chair for the 2015 IEEE International Symposium on Information Theory (ISIT) and General co-chair for ACM Mobihoc 2018. He has 8 issued patents.