
CodedPrivateML: A Fast and Privacy-Preserving Framework for Distributed Machine Learning

Jinhyun So^{*1} Bařak Güler^{*1} A. Salman Avestimehr¹ Payman Mohassel²

Abstract

How to train a machine learning model while keeping the data private and secure? We present CodedPrivateML, a fast and scalable approach to this critical problem. CodedPrivateML keeps both the data and the model information-theoretically private, while allowing efficient parallelization of training across distributed workers. We characterize CodedPrivateML’s privacy threshold and prove its convergence for logistic (and linear) regression. Furthermore, via experiments over Amazon EC2, we demonstrate that CodedPrivateML can provide an order of magnitude speedup (up to $\sim 34\times$) over the state-of-the-art cryptographic approaches.

1. Introduction

Modern machine learning models are breaking new ground by achieving unprecedented performance in various application domains. Training such models, however, is a daunting task. Due to the typically large volume of data and complexity of models, training is a compute and storage intensive task. Furthermore, training should often be done on sensitive data, such as healthcare records, browsing history, or financial transactions, which raises the issues of security and privacy of the dataset. This creates a challenging dilemma. On the one hand, due to its complexity, training is often desired to be outsourced to more capable computing platforms, such as the cloud. On the other hand, the training dataset is often sensitive and particular care should be taken to protect the privacy of the dataset against potential breaches in such platforms. This dilemma gives rise to the main problem that we study here: *How can we offload the training task to a distributed computing platform, while maintaining the privacy of the dataset?*

More specifically, we consider a scenario in which a data-

owner (e.g., a hospital) wishes to train a logistic regression model by offloading the large volume of data (e.g., healthcare records) and computationally-intensive training tasks (e.g., gradient computations) to N machines over a cloud platform, while ensuring that any collusions between T out of N workers do not leak information about the training dataset. We focus on the semi-honest adversary setup, where the corrupted parties follow the protocol but may leak information in an attempt to learn the training dataset.

We propose CodedPrivateML for this problem, which has three salient features:

1. provides strong information-theoretic privacy guarantees for both the training dataset and model parameters in the presence of colluding workers.
2. enables fast training by distributing the training computation load effectively across several workers.
3. leverages a new method for secret sharing the dataset and model parameters based on coding and information theory principles, which significantly reduces the communication overhead and the complexity for distributed training.

At a high level, CodedPrivateML can be described as follows. It secret shares the dataset and model parameters at each round of the training in two steps. First, it employs stochastic quantization to convert the dataset and the weight vector at each round into a finite domain. It then combines (or *encodes*) the quantized values with random matrices, using a novel coding technique named Lagrange coding (Yu et al., 2019), to guarantee privacy (in an information-theoretic sense) while simultaneously distributing the workload among multiple workers. The challenge is however that Lagrange coding can only work for computations that are in the form of polynomial evaluations. The gradient computation for logistic regression, on the other hand, includes non-linearities that cannot be expressed as polynomials. CodedPrivateML handles this challenge through polynomial approximations of the non-linear sigmoid function in the training phase.

Upon secret sharing of the encoded dataset and model parameters, each worker performs the gradient computations using the chosen polynomial approximation of the sigmoid

* Equal contribution. ¹ Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA, jinhyuns@usc.edu, bguler@usc.edu, avestimehr@ee.usc.edu. ² Visa Research, payman.mohassel@gmail.com

function, and sends the result back to the master. It is useful to note that the workers perform the computations over the quantized and encoded data *as if they were computing over the true dataset*. That is, the structure of the computations are the same for computing over the true dataset versus computing over the encoded dataset.

Finally, the master collects the results from a subset of fastest workers and decodes the gradient over the finite field. It then converts the decoded gradients to the real domain, updates the weight vector, and secret shares it with the worker nodes for the next round. We note that since the computations are performed in a finite domain while the weights are updated in the real domain, the update process may lead to undesired behaviour as weights may not converge. Our system guarantees convergence through the proposed stochastic quantization technique while converting between real and finite fields.

We theoretically prove that CodedPrivateML guarantees the convergence of the model parameters, while providing information-theoretic privacy for the training dataset. Our theoretical analysis also identifies a trade-off between privacy and parallelization. More specifically, each additional worker can be utilized either for more privacy, by protecting against a larger number of collusions T , or more parallelization, by reducing the computation load at each worker. We characterize this trade-off for CodedPrivateML.

Furthermore, we empirically demonstrate the impact of CodedPrivateML by comparing it with state-of-the-art cryptographic approaches based on secure multi-party computing (MPC) (Yao, 1982; Ben-Or et al., 1988), that can also be applied to enable privacy-preserving machine learning tasks (e.g. see (Nikolaenko et al., 2013; Gascón et al., 2017; Mohassel & Zhang, 2017; Lindell & Pinkas, 2000; Dahl et al., 2018; Chen et al., 2019)). In particular, we envision a master who secret shares its data and model parameters among multiple workers who collectively perform the gradient computation using a multi-round MPC protocol. Given our focus on information-theoretic privacy, the most relevant MPC-based scheme for empirical comparison is the BGW-style (Ben-Or et al., 1988) approach based on Shamir’s secret sharing (Shamir, 1979). While several more recent work design MPC-based private learning solutions with information-theoretic security (Wagh et al., 2018; Mohassel & Rindal, 2018), their constructions are limited to three or four parties.

We run extensive experiments over Amazon EC2 cloud to empirically demonstrate the performance of CodedPrivateML. We train a logistic regression model for image classification over the MNIST dataset (LeCun et al., 2010), while the computation workload is distributed to up to $N = 40$ machines over the cloud. We demonstrate that CodedPrivateML can provide substantial speedup in training

time (up to $\sim 34.1\times$), compared with MPC-based schemes, while guaranteeing the same level of accuracy. The primary disadvantage of the MPC-based scheme is its reliance on extensive communication and coordination between the workers for distributed private computing, and not benefiting from parallelization among the workers as the whole computation is repeated by all players who take part in MPC. They however guarantee a higher privacy threshold (i.e., larger T) compared with CodedPrivateML.

Other related works. Apart from MPC-based schemes to this problem, one can consider two other solutions. One is based on Homomorphic Encryption (HE) (Gentry & Boneh, 2009) which allows for computation to be performed over encrypted data, and has been used to enable privacy-preserving machine learning solutions (Gilad-Bachrach et al., 2016; Hesamifard et al., 2017; Graepel et al., 2012; Yuan & Yu, 2014; Li et al., 2017; Kim et al., 2018; Wang et al., 2018; Han et al., 2019). The privacy guarantees of HE are based on computational assumptions, whereas our system provides strong information-theoretic security. Moreover, HE requires computations to be performed over encrypted data which leads to many orders of magnitude slow down in training. For example, for image classification on the MNIST dataset, HE takes 2 hours to learn a logistic regression model with 96% accuracy (Han et al., 2019). In contrast, in CodedPrivateML there is no slow down in performing the coded computations which allows for a faster implementation. As a trade-off, HE allows collusion between a larger number of workers whereas in CodedPrivateML this number is determined by other system parameters such as number of workers and the computation load assigned to each worker.

Another possible solution is based on differential privacy (DP), which is a release mechanism that preserves the privacy of personally identifiable information, in that the removal of any single element from the dataset does not change the computation outcomes significantly (Dwork et al., 2006). In the context of machine learning, DP is mainly used for training when the model parameters are to be released for public use, to ensure that the individual data points from the dataset cannot be identified from the released model (Chaudhuri & Monteleoni, 2009; Shokri & Shmatikov, 2015; Abadi et al., 2016; Pathak et al., 2010; McMahan et al., 2018; Rajkumar & Agarwal, 2012; Jayaraman et al., 2018). The main difference between these approaches and our work is that we can guarantee strong information-theoretic privacy that leaks no information about the dataset, and preserve the accuracy of the model throughout the training. We note however that it is in principle possible to compose techniques of CodedPrivateML with differential privacy to obtain the best of both worlds if the intention is to publicly release the final model, but we leave this as future work.

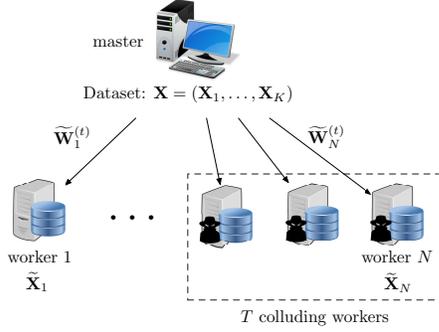


Figure 1. The distributed training setup consisting of a master and N worker nodes. The master shares with each worker a coded version of the dataset (denoted by $\tilde{\mathbf{X}}_i$'s) and the current estimate of the model parameters (denoted by $\tilde{\mathbf{W}}_i^{(t)}$'s) to guarantee the information-theoretic privacy of the dataset against any T colluding workers. Workers perform computations locally over the coded data and send the results back to the master.

2. Problem Setting

We study the problem of training a logistic regression model. The training dataset is represented by a matrix $\mathbf{X} \in \mathbb{R}^{m \times d}$ consisting of m data points with d features and a label vector $\mathbf{y} \in \{0, 1\}^m$. Row i of \mathbf{X} is denoted by \mathbf{x}_i .

The model parameters (weights) $\mathbf{w} \in \mathbb{R}^d$ are obtained by minimizing the cross entropy function,

$$C(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^m (-y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i)) \quad (1)$$

where $\hat{y}_i = g(\mathbf{x}_i \cdot \mathbf{w}) \in (0, 1)$ is the estimated probability of label i being equal to 1 and $g(\cdot)$ is the sigmoid function

$$g(z) = 1/(1 + e^{-z}). \quad (2)$$

The problem in (1) can be solved via gradient descent, through an iterative process that updates the model parameters in the opposite direction of the gradient. The gradient for (1) is given by $\nabla C(\mathbf{w}) = \frac{1}{m} \mathbf{X}^\top (g(\mathbf{X} \times \mathbf{w}) - \mathbf{y})$. Accordingly, model parameters are updated as,

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \frac{\eta}{m} \mathbf{X}^\top (g(\mathbf{X} \times \mathbf{w}^{(t)}) - \mathbf{y}) \quad (3)$$

where $\mathbf{w}^{(t)}$ holds the estimated parameters from iteration t , η is the learning rate, and function $g(\cdot)$ operates element-wise over the vector given by $\mathbf{X} \times \mathbf{w}^{(t)}$.

As shown in Figure 1, we consider a master-worker distributed computing architecture, where the master offloads the computationally-intensive operations to N workers. These operations correspond to gradient computations in (3). In doing so, master wishes to protect the privacy of the dataset \mathbf{X} against any potential collusions between up to T workers, where T is the *privacy parameter* of the system.

At the beginning of the training, dataset \mathbf{X} is shared in a privacy-preserving manner among the workers. To do so, \mathbf{X} is first partitioned into K submatrices $\mathbf{X} = [\mathbf{X}_1^\top \cdots \mathbf{X}_K^\top]^\top$, for some $K \in \mathbb{N}$. Parameter K is related to the computation

load at each worker (i.e., what fraction of the dataset is processed at each worker), as well as the number of workers the master has to wait for, to reconstruct the gradient at each step. The master then creates N *encoded* submatrices, denoted by $\tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_N$, by combining the K parts of the dataset together with some random matrices to preserve privacy, and sends $\tilde{\mathbf{X}}_i$ to worker $i \in [N]$. This process should only be performed once for the dataset \mathbf{X} .

At each iteration t of the training, the master also needs to send to worker $i \in [N]$ the current estimate of the model parameters (i.e., $\mathbf{w}^{(t)}$ in (3)). However, it is recently shown that the intermediate model parameters can also leak substantial information about the dataset (Melis et al., 2019). The master also needs to prevent the leakage of these intermediate parameters. To that end, the master creates an encoded matrix $\tilde{\mathbf{W}}_i^{(t)}$ to secret share the current estimate of model parameters with worker $i \in [N]$. This coding strategy should also be private against any T colluding workers.

More specifically, the coding strategy that is used for secret sharing the dataset (i.e., creating $\tilde{\mathbf{X}}_i$'s) and model parameters (i.e., creating $\tilde{\mathbf{W}}_i^{(t)}$'s) should be such that any subset of T colluding workers can not learn any information, in the strong information-theoretic sense, about the training dataset \mathbf{X} . Formally, for every subset of workers $\mathcal{T} \subseteq [N]$ of size at most T , we should have,

$$I(\mathbf{X}; \tilde{\mathbf{X}}_{\mathcal{T}}, \{\tilde{\mathbf{W}}_{\mathcal{T}}^{(t)}\}_{t \in [J]}) = 0 \quad (4)$$

where I denotes the mutual information, J is the number of iterations, and $\tilde{\mathbf{X}}_{\mathcal{T}}, \{\tilde{\mathbf{W}}_{\mathcal{T}}^{(t)}\}_{t \in [J]}$ is the collection of the coded matrices and coded parameter estimations stored at workers in \mathcal{T} . We refer to a protocol that guarantees privacy against T colluding workers as a T -private protocol.

At each iteration, worker $i \in [N]$ performs its computation locally using $\tilde{\mathbf{X}}_i$ and $\tilde{\mathbf{W}}_i^{(t)}$ and sends the result back to the master. After receiving the results from a sufficient number of workers, the master recovers $\mathbf{X}^\top g(\mathbf{X} \times \mathbf{w}^{(t)}) = \sum_{k=1}^K \mathbf{X}_k^\top g(\mathbf{X}_k \times \mathbf{w}^{(t)})$, reconstructs the gradients, and updates the model parameters using (3). In doing so, the master needs to wait only for the fastest workers. We define the *recovery threshold* of the protocol as the minimum number of workers the master needs to wait for. The relations between the recovery threshold and parameters N , K , and T will be detailed in our theoretical analysis.

Remark 1. Although our presentation is based on logistic regression, CodedPrivateML can also be applied to linear regression with minor modifications.

3. The Proposed CodedPrivateML Strategy

CodedPrivateML strategy consists of four main phases that are first described at a high-level below, and then with details in the rest of this section.

Phase 1: Quantization. In order to guarantee information-theoretic privacy, one has to mask the dataset and the weight vector in a finite field \mathbb{F} using uniformly random matrices, so that the added randomness can make each data point appear equally likely. In contrast, the dataset and weight vectors for the training task are defined in the domain of real numbers. We address this by employing a stochastic quantization technique to convert the parameters from the real domain to the finite domain and vice versa. Accordingly, in the first phase of our system, master quantizes the dataset and weights from the real domain to the domain of integers, and then embeds them in a field \mathbb{F}_p of integers modulo a prime p . The quantized version of the dataset \mathbf{X} is given by $\bar{\mathbf{X}}$. The quantization of the weight vector $\mathbf{w}^{(t)}$, on the other hand, is represented by a matrix $\bar{\mathbf{W}}^{(t)}$, where each column holds an independent stochastic quantization of $\mathbf{w}^{(t)}$. This structure will be important in ensuring the convergence of the model. Parameter p is selected to be sufficiently large to avoid wrap-around in computations. Its value depends on the bitwidth of the machine as well as the number of additive and multiplicative operations. For example, in a 64-bit implementation, we select $p = 15485863$ (the largest prime with 24 bits) as detailed in our experiments.

Phase 2: Encoding and Secret Sharing. In the second phase, the master partitions the quantized dataset $\bar{\mathbf{X}}$ into K submatrices and encodes them using the recently proposed Lagrange coding technique (Yu et al., 2019), which we will describe in detail in Section 3.2. It then sends to worker $i \in [N]$ a coded submatrix $\tilde{\mathbf{X}}_i \in \mathbb{F}_p^{\frac{m}{K} \times d}$. As we will illustrate later, this encoding ensures that the coded matrices do not leak any information about the true dataset, even if T workers collude. In addition, the master has to ensure the weight estimations sent to the workers at each iteration do not leak information about the dataset. This is because the weights updated via (3) carry information about the whole training set, and sending them directly to the workers may breach privacy. In order to prevent this, at iteration t , master also quantizes the current weight vector $\mathbf{w}^{(t)}$ to the finite field and encodes it again using Lagrange coding.

Phase 3: Polynomial Approximation and Local Computations. In the third phase, each worker performs the computations using its local storage and sends the result back to the master. We note that the workers perform the computations over the encoded data as if they were computing over the true dataset. That is, the structure of the computations are the same for computing over the true dataset versus computing over the encoded dataset. A major challenge is that Lagrange coding is designed for distributed polynomial computations. However, the computations in the training phase are not polynomials due to the sigmoid function. We overcome this by approximating the sigmoid with a polynomial of a selected degree r . This allows us to represent the gradient computations in terms of polynomials that can be

computed locally by each worker.

Phase 4: Decoding and Model Update. The master collects the results from a subset of fastest workers and decodes the gradient over the finite field. Finally, master converts the decoded gradients to the real domain, updates the weight vector, and secret shares it with workers for the next round.

We next provide the details of each phase. The overall algorithm of CodedPrivateML, and each of its four phases, are also presented in Appendix A.1 of supplementary materials.

3.1. Quantization

We consider an element-wise lossy quantization scheme for the dataset and weights. For quantizing the dataset $\mathbf{X} \in \mathbb{R}^{m \times d}$, we use a simple deterministic rounding technique:

$$\text{Round}(x) = \begin{cases} \lfloor x \rfloor & \text{if } x - \lfloor x \rfloor < 0.5 \\ \lfloor x \rfloor + 1 & \text{otherwise} \end{cases} \quad (5)$$

where $\lfloor x \rfloor$ is the largest integer less than or equal to x . We define the quantized dataset as

$$\bar{\mathbf{X}} \triangleq \phi \left(\text{Round}(2^{l_x} \cdot \mathbf{X}) \right) \quad (6)$$

where the rounding function from (5) is applied element-wise to the elements of matrix \mathbf{X} and l_x is an integer parameter that controls the quantization loss. Function $\phi : \mathbb{Z} \rightarrow \mathbb{F}_p$ is a mapping defined to represent a negative integer in the finite field by using two's complement representation,

$$\phi(x) = \begin{cases} x & \text{if } x \geq 0 \\ p + x & \text{if } x < 0 \end{cases} \quad (7)$$

Note that the domain of (6) is $\left[-\frac{p-1}{2^{(l_x+1)}}, \frac{p-1}{2^{(l_x+1)}} \right]$. To avoid a wrap-around which may lead to an overflow error, prime p should be large enough, i.e., $p \geq 2^{l_x+1} \max\{|\mathbf{X}_{i,j}|\} + 1$.

At each iteration, master also quantizes the weight vector $\mathbf{w}^{(t)}$ from real domain to the finite field. This proves to be a challenging task as it should be performed in a way to ensure the convergence of the model. Our solution to this is a quantization technique inspired by (Zhang et al., 2017; 2016). Initially, we define a stochastic quantization function:

$$Q(x; l_w) \triangleq \phi \left(\text{Round}_{stoc}(2^{l_w} \cdot x) \right) \quad (8)$$

where l_w is an integer parameter to control the quantization loss. $\text{Round}_{stoc} : \mathbb{R} \rightarrow \mathbb{Z}$ is a stochastic rounding function:

$$\text{Round}_{stoc}(x) = \begin{cases} \lfloor x \rfloor & \text{with prob. } 1 - (x - \lfloor x \rfloor) \\ \lfloor x \rfloor + 1 & \text{with prob. } x - \lfloor x \rfloor \end{cases}$$

The probability of rounding x to $\lfloor x \rfloor$ is proportional to the proximity of x to $\lfloor x \rfloor$ so that stochastic rounding is unbiased (i.e., $\mathbb{E}[\text{Round}_{stoc}(x)] = x$).

For quantizing the weight vector $\mathbf{w}^{(t)}$, the master creates r independent quantized vectors:

$$\bar{\mathbf{w}}^{(t),j} \triangleq Q_j(\mathbf{w}^{(t)}; l_w) \in \mathbb{F}_p^{d \times 1} \quad \text{for } j \in [r] \quad (9)$$

where the quantization function (8) is applied element-wise to the vector $\mathbf{w}^{(t)}$ and each $Q_j(\cdot; \cdot)$ denotes an independent realization of (8). The number of quantized vectors r is equal to the degree of the polynomial approximation for the sigmoid function, which we will describe later in Section 3.3. The intuition behind creating r independent quantizations is to ensure that the gradient computations performed using the quantized weights are unbiased estimators of the true gradients. As detailed in Section 4, this property is fundamental for the convergence analysis of our model. The specific values of parameters l_x and l_w provide a trade-off between the rounding error and overflow error. In particular, a larger value reduces the rounding error while increasing the chance of an overflow. We denote the quantization of the weight vector $\mathbf{w}^{(t)}$ as

$$\bar{\mathbf{W}}^{(t)} = [\bar{\mathbf{w}}^{(t),1} \quad \dots \quad \bar{\mathbf{w}}^{(t),r}]. \quad (10)$$

by arranging the quantized vectors from (9) in matrix form.

3.2. Encoding and Secret Sharing

The master first partitions the quantized dataset $\bar{\mathbf{X}}$ into K submatrices $\bar{\mathbf{X}} = [\bar{\mathbf{X}}_1^\top \dots \bar{\mathbf{X}}_K^\top]^\top$, where $\bar{\mathbf{X}}_i \in \mathbb{F}_p^{\frac{m}{K} \times d}$ for $i \in [K]$. It also selects $K + T$ distinct elements $\beta_1, \dots, \beta_{K+T}$ from \mathbb{F}_p . It then employs Lagrange coding (Yu et al., 2019) to encode the dataset. More specifically, it finds a polynomial $u : \mathbb{F}_p \rightarrow \mathbb{F}_p^{\frac{m}{K} \times d}$ of degree at most $K + T - 1$ such that $u(\beta_i) = \bar{\mathbf{X}}_i$ for $i \in [K]$, and $u(\beta_i) = \mathbf{Z}_i$ for $i \in \{K + 1, \dots, K + T\}$, where \mathbf{Z}_i 's are chosen uniformly at random from $\mathbb{F}_p^{\frac{m}{K} \times d}$ (the role of \mathbf{Z}_i 's is to mask the dataset and provide privacy against up to T colluding workers). This is accomplished by letting u be the respective *Lagrange interpolation polynomial*

$$u(z) \triangleq \sum_{j \in [K]} \bar{\mathbf{X}}_j \cdot \prod_{k \in [K+T] \setminus \{j\}} \frac{z - \beta_k}{\beta_j - \beta_k} + \sum_{j=K+1}^{K+T} \mathbf{Z}_j \cdot \prod_{k \in [K+T] \setminus \{j\}} \frac{z - \beta_k}{\beta_j - \beta_k}. \quad (11)$$

Master then selects N distinct elements $\{\alpha_i\}_{i \in [N]}$ from \mathbb{F}_p such that $\{\alpha_i\}_{i \in [N]} \cap \{\beta_j\}_{j \in [K]} = \emptyset$, and encodes the dataset by letting $\tilde{\mathbf{X}}_i = u(\alpha_i)$ for $i \in [N]$. By defining an encoding matrix $\mathbf{U} = [\mathbf{u}_1 \dots \mathbf{u}_N] \in \mathbb{F}_p^{(K+T) \times N}$ whose $(i, j)^{th}$ element is given by $u_{ij} = \prod_{\ell \in [K+T] \setminus \{i\}} \frac{\alpha_j - \beta_\ell}{\beta_i - \beta_\ell}$, one can also represent the encoding of the dataset as,

$$\tilde{\mathbf{X}}_i = u(\alpha_i) = (\bar{\mathbf{X}}_1, \dots, \bar{\mathbf{X}}_K, \mathbf{Z}_{K+1}, \dots, \mathbf{Z}_{K+T}) \cdot \mathbf{u}_i. \quad (12)$$

At iteration t , the quantized weights $\bar{\mathbf{W}}^{(t)}$ are also encoded using a Lagrange interpolation polynomial,

$$v(z) \triangleq \sum_{j \in [K]} \bar{\mathbf{W}}^{(t),j} \cdot \prod_{k \in [K+T] \setminus \{j\}} \frac{z - \beta_k}{\beta_j - \beta_k} + \sum_{j=K+1}^{K+T} \mathbf{V}_j \cdot \prod_{k \in [K+T] \setminus \{j\}} \frac{z - \beta_k}{\beta_j - \beta_k}. \quad (13)$$

where \mathbf{V}_j for $j \in [K + 1, K + T]$ are chosen uniformly at random from $\mathbb{F}_p^{d \times r}$. The coefficients $\beta_1, \dots, \beta_{K+T}$ are the same as the ones in (11). We note that the polynomial in (13) has the property $v(\beta_i) = \bar{\mathbf{W}}^{(t)}$ for $i \in [K]$.

The master then encodes the quantized weight vector by using the same evaluation points $\{\alpha_i\}_{i \in [N]}$. Accordingly, the weight vector is encoded as

$$\tilde{\mathbf{W}}_i^{(t)} = v(\alpha_i) = (\bar{\mathbf{W}}^{(t),1}, \dots, \bar{\mathbf{W}}^{(t),r}, \mathbf{V}_{K+1}, \dots, \mathbf{V}_{K+T}) \cdot \mathbf{u}_i \quad (14)$$

for $i \in [N]$, using the encoding matrix \mathbf{U} from (12). The degree of the polynomials $u(z)$ and $v(z)$ are both $K + T - 1$.

3.3. Polynomial Approximation and Local Computation

Upon receiving the encoded (and quantized) dataset and weights, workers should proceed with gradient computations. However, a major challenge is that Lagrange coding is originally designed for polynomial computations, while the gradient computations that the workers need to do are not polynomials due to the sigmoid function. Our solution is to use a polynomial approximation of the sigmoid function,

$$\hat{g}(z) = \sum_{i=0}^r c_i z^i, \quad (15)$$

where r and c_i denotes the degree and coefficients of the polynomial, respectively. The coefficients are obtained by fitting the sigmoid function via least squares estimation.

Using this polynomial approximation we can rewrite (3) as,

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \frac{\eta}{m} \bar{\mathbf{X}}^\top (\hat{g}(\bar{\mathbf{X}} \times \mathbf{w}^{(t)}) - \mathbf{y}). \quad (16)$$

where $\bar{\mathbf{X}}$ is the quantized version of \mathbf{X} , and $\hat{g}(\cdot)$ operates element-wise over the vector $\bar{\mathbf{X}} \times \mathbf{w}^{(t)}$.

Another challenge is to ensure the convergence of weights. As we detail in Section 4, this necessitates the gradient estimations to be unbiased using the polynomial approximation with quantized weights. We solve this by utilizing the computation technique from Lemma 4.1 in (Zhang et al., 2016) using the quantized weights formed in Section 3.1. Specifically, given a degree r polynomial from (15) and r independent quantizations from (10), we define a function,

$$\bar{g}(\bar{\mathbf{X}}, \bar{\mathbf{W}}^{(t)}) \triangleq \sum_{i=0}^r c_i \prod_{j \leq i} (\bar{\mathbf{X}} \times \bar{\mathbf{w}}^{(t),j}) \quad (17)$$

where the product $\prod_{j \leq i}$ operates element-wise over the vectors $(\bar{\mathbf{X}} \times \bar{\mathbf{w}}^{(t),j})$ for $j \leq i$. Lastly, we note that (17) is an unbiased estimator of $\hat{g}(\bar{\mathbf{X}} \times \mathbf{w}^{(t)})$,

$$E[\bar{g}(\bar{\mathbf{X}}, \bar{\mathbf{W}}^{(t)})] = \hat{g}(\bar{\mathbf{X}} \times \mathbf{w}^{(t)}), \quad (18)$$

where $\hat{g}(\cdot)$ acts element-wise over the vector $\bar{\mathbf{X}} \times \mathbf{w}^{(t)}$, and the result follows from the independence of quantizations.

Using (17), we rewrite the update equations from (16) in terms of the quantized weights,

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \frac{\eta}{m} \bar{\mathbf{X}}^\top (\bar{g}(\bar{\mathbf{X}}, \bar{\mathbf{W}}^{(t)}) - \mathbf{y}). \quad (19)$$

The computations are then performed at each worker locally. In particular, at each iteration, worker $i \in [N]$ locally computes $f : \mathbb{F}_p^{\frac{m}{K} \times d} \times \mathbb{F}_p^{d \times r} \rightarrow \mathbb{F}_p^d$:

$$f(\tilde{\mathbf{X}}_i, \tilde{\mathbf{W}}_i^{(t)}) = \tilde{\mathbf{X}}_i^\top \bar{g}(\tilde{\mathbf{X}}_i, \tilde{\mathbf{W}}_i^{(t)}) \quad (20)$$

using $\tilde{\mathbf{X}}_i$ and $\tilde{\mathbf{W}}_i^{(t)}$ and sends the result back to the master. This computation is a polynomial function evaluation in finite field arithmetic and the degree of f is $\deg(f) = 2r + 1$.

3.4. Decoding and Model Update

After receiving the evaluation results in (20) from a sufficient number of workers, master decodes $\left\{ f(\tilde{\mathbf{X}}_k, \tilde{\mathbf{W}}_k^{(t)}) \right\}_{k \in [K]}$ over the finite field. The minimum number of workers the master needs to wait for is termed the recovery threshold of the system and is equal to $(2r + 1)(K + T - 1) + 1$ as we demonstrate in Section 4.

We now proceed to the details of decoding. By construction of the Lagrange polynomials in (11) and (13), one can define a univariate polynomial $h(z) = f(u(z), v(z))$ such that

$$\begin{aligned} h(\beta_i) &= f(u(\beta_i), v(\beta_i)) \\ &= f(\bar{\mathbf{X}}_i, \bar{\mathbf{W}}^{(t)}) = \bar{\mathbf{X}}_i^\top \bar{g}(\bar{\mathbf{X}}_i, \bar{\mathbf{W}}^{(t)}) \end{aligned} \quad (21)$$

for $i \in [K]$. On the other hand, from (20), the computation result from worker i equals to

$$\begin{aligned} h(\alpha_i) &= f(u(\alpha_i), v(\alpha_i)) \\ &= f(\tilde{\mathbf{X}}_i, \tilde{\mathbf{W}}_i^{(t)}) = \tilde{\mathbf{X}}_i^\top \bar{g}(\tilde{\mathbf{X}}_i, \tilde{\mathbf{W}}_i^{(t)}) \end{aligned} \quad (22)$$

The main intuition behind the decoding process is to use the computations from (22) as evaluation points $h(\alpha_i)$ to interpolate the polynomial $h(z)$. Specifically, the master can obtain all coefficients of $h(z)$ from $(2r + 1)(K + T - 1) + 1$ evaluation results as long as $\deg(h(z)) \leq (2r + 1)(K + T - 1)$. After $h(z)$ is recovered, the master can recover (21) by computing $h(\beta_i)$ for $i \in [K]$ and evaluating

$$\sum_{k=1}^K f(\bar{\mathbf{X}}_k, \bar{\mathbf{W}}^{(t)}) = \sum_{k=1}^K \bar{\mathbf{X}}_k^\top \bar{g}(\bar{\mathbf{X}}_k, \bar{\mathbf{W}}^{(t)}) = \bar{\mathbf{X}}^\top \bar{g}(\bar{\mathbf{X}}, \bar{\mathbf{W}}^{(t)}) \quad (23)$$

Lastly, master converts (23) from the finite field to the real domain and updates the weights according to (19). This conversion is attained by the function,

$$Q_p^{-1}(\bar{x}; l) = 2^{-l} \cdot \phi^{-1}(\bar{x}). \quad (24)$$

where we let $l = l_x + r(l_x + l_w)$, and $\phi^{-1} : \mathbb{F}_p \rightarrow \mathbb{R}$ is defined as follows,

$$\phi^{-1}(\bar{x}) = \begin{cases} \bar{x} & \text{if } 0 \leq \bar{x} < \frac{p-1}{2} \\ \bar{x} - p & \text{if } \frac{p-1}{2} \leq \bar{x} < p \end{cases} \quad (25)$$

4. Convergence and Privacy Guarantees

Consider the cost function (1) that we aim to minimize in logistic regression when dataset \mathbf{X} is replaced with the quantized dataset $\bar{\mathbf{X}}$ using (6). Also denote \mathbf{w}^* as the optimal weight vector that minimizes (1) when $\hat{y}_i = g(\bar{\mathbf{x}}_i \cdot \mathbf{w})$, where $\bar{\mathbf{x}}_i$ is row i of $\bar{\mathbf{X}}$. In this section we prove that CodedPrivateML would guarantee convergence to the optimal model parameters (i.e., \mathbf{w}^*) while maintaining the privacy of the dataset against colluding workers.

Recall that the model update at the master node in CodedPrivateML follows (19), which is

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \frac{\eta}{m} \bar{\mathbf{X}}^\top (\bar{g}(\bar{\mathbf{X}}, \bar{\mathbf{W}}^{(t)}) - \mathbf{y}). \quad (26)$$

We first state a lemma, which is proved in Appendix A.2 in supplementary materials.

Lemma 1. Let $\mathbf{p}^{(t)} \triangleq \frac{1}{m} \bar{\mathbf{X}}^\top (\bar{g}(\bar{\mathbf{X}}, \bar{\mathbf{W}}^{(t)}) - \mathbf{y})$ denote the gradient computation using the quantized weights $\bar{\mathbf{W}}^{(t)}$ in CodedPrivateML. Then we have

- (Unbiasedness) Vector $\mathbf{p}^{(t)}$ is an asymptotically unbiased estimator of the true gradient. $\mathbb{E}[\mathbf{p}^{(t)}] = \nabla C(\mathbf{w}^{(t)}) + \epsilon(r)$, and $\epsilon(r) \rightarrow \mathbf{0}$ as $r \rightarrow \infty$ where r is the degree of polynomial in (15) and expectation is taken with respect to the quantization errors,
- (Variance bound) $\mathbb{E}[\|\mathbf{p}^{(t)} - \mathbb{E}[\mathbf{p}^{(t)}]\|_2^2] \leq \frac{1}{2^{-2l_w m^2}} \|\bar{\mathbf{X}}\|_F^2 \triangleq \sigma^2$ where $\|\cdot\|_2$ and $\|\cdot\|_F$ denote the l_2 -norm and Frobenius norm, respectively.

We also need the following basic lemma, which is proved in Appendix A.3 of supplementary materials.

Lemma 2. The gradient of the cost function (1) with quantized dataset $\bar{\mathbf{X}}$ (as defined in (6)) is L -Lipschitz with $L \triangleq \frac{1}{4} \|\bar{\mathbf{X}}\|_2^2$, i.e., for all $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$ we have

$$\|\nabla C(\mathbf{w}) - \nabla C(\mathbf{w}')\| \leq L \|\mathbf{w} - \mathbf{w}'\|, \quad (27)$$

We now state our main theorem for CodedPrivateML.

Theorem 1. Consider the training of a logistic regression model in a distributed system with N workers using CodedPrivateML with dataset $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_K)$, initial weight vector $\mathbf{w}^{(0)}$, and constant step size $\eta = 1/L$ (where L is defined in Lemma 2). Then, CodedPrivateML guarantees,

- (Convergence) $\mathbb{E} \left[C \left(\frac{1}{J} \sum_{t=0}^J \mathbf{w}^{(t)} \right) \right] - C(\mathbf{w}^*) \leq \frac{\|\mathbf{w}^{(0)} - \mathbf{w}^*\|^2}{2\eta^J} + \eta\sigma^2$ in J iterations, where σ^2 is given in Lemma 1,
- (Privacy) \mathbf{X} remains information-theoretically private against any T colluding workers, i.e., $I(\mathbf{X}; \tilde{\mathbf{X}}_{\mathcal{T}}, \{\tilde{\mathbf{W}}_{\mathcal{T}}^{(t)}\}_{t \in [J]}) = 0, \forall \mathcal{T} \subset [N], |\mathcal{T}| \leq T$, as long as we have $N \geq (2r + 1)(K + T - 1) + 1$, where r is the degree of the polynomial approximation in (15).

Remark 2. *Theorem 1 reveals an important trade-off between privacy and parallelization in CodedPrivateML. The parameter K reflects the amount of parallelization in CodedPrivateML, since the computation load at each worker node is proportional to $1/K$ -th of the dataset. The parameter T also reflects the privacy threshold in CodedPrivateML. Theorem 1 shows that, in a cluster with N workers, we can achieve any K and T as long as $N \geq (2r + 1)(K + T - 1) + 1$. This condition further implies that, as the number of workers N increases, the parallelization (K) and privacy threshold (T) of CodedPrivateML can also increase linearly, leading to a scalable solution.*

Remark 3. *Theorem 1 also applies to the simpler linear regression problem. The proof follows the same steps.*

Proof. (Convergence) First, we show that the master can decode $\bar{\mathbf{X}}^\top \bar{g}(\bar{\mathbf{X}}, \bar{\mathbf{W}}^{(t)})$ over the finite field as long as $N \geq (2r + 1)(K + T - 1) + 1$. As described in Sections 3.3 and 3.4, given the polynomial to approximation of the sigmoid function in (15), the degree of $h(z)$ in (21) is a most $(2r + 1)(K + T - 1)$. The decoding process uses the computations from workers as evaluation points $h(\alpha_i)$ to interpolate the polynomial $h(z)$. The master can obtain all coefficients of $h(z)$ as long as the master collects at least $\deg(h(z)) + 1 \leq (2r + 1)(K + T - 1) + 1$ evaluation results of $h(\alpha_i)$. After $h(z)$ is recovered, the master can decode the sub-gradient $\bar{\mathbf{X}}_i^\top \bar{g}(\bar{\mathbf{X}}_i, \bar{\mathbf{W}}^{(t)})$ by computing $h(\beta_i)$ for $i \in [K]$. Hence, the recovery threshold is given by $(2r + 1)(K + T - 1) + 1$ to decode $\bar{\mathbf{X}}^\top \bar{g}(\bar{\mathbf{X}}, \bar{\mathbf{W}}^{(t)})$.

Next, we consider the update equation in CodedPrivateML (see (26)) and prove its convergence to \mathbf{w}^* . From the L -Lipschitz continuity of $\nabla C(\mathbf{w})$ stated in Lemma 2, we have

$$\begin{aligned} C(\mathbf{w}^{(t+1)}) &\leq C(\mathbf{w}^{(t)}) + \langle \nabla C(\mathbf{w}^{(t)}), \mathbf{w}^{(t+1)} - \mathbf{w}^{(t)} \rangle \\ &\quad + \frac{L}{2} \|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\|^2 \\ &= C(\mathbf{w}^{(t)}) - \eta \langle \nabla C(\mathbf{w}^{(t)}), \mathbf{p}^{(t)} \rangle + \frac{L}{2} \|\mathbf{p}^{(t)}\|^2, \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ is the inner product. By taking the expectation with respect to the quantization noise on both sides,

$$\begin{aligned} &\mathbb{E}[C(\mathbf{w}^{(t+1)})] \\ &\leq C(\mathbf{w}^{(t)}) - \eta \|\nabla C(\mathbf{w}^{(t)})\|^2 + \frac{L\eta^2}{2} (\|\nabla C(\mathbf{w}^{(t)})\|^2 + \sigma^2) \\ &\leq C(\mathbf{w}^{(t)}) - \eta(1 - L\eta/2) \|\nabla C(\mathbf{w}^{(t)})\|^2 + L\eta^2 \sigma^2 / 2 \\ &\leq C(\mathbf{w}^{(t)}) - \eta/2 \|\nabla C(\mathbf{w}^{(t)})\|^2 + \eta\sigma^2 / 2 \end{aligned} \quad (28)$$

$$\begin{aligned} &\leq C(\mathbf{w}^*) + \langle \nabla C(\mathbf{w}^{(t)}), \mathbf{w}^{(t)} - \mathbf{w}^* \rangle \\ &\quad - \frac{\eta}{2} \|\nabla C(\mathbf{w}^{(t)})\|^2 + \eta\sigma^2 / 2 \end{aligned} \quad (29)$$

$$\leq C(\mathbf{w}^*) + \langle \mathbb{E}[\mathbf{p}^{(t)}], \mathbf{w}^{(t)} - \mathbf{w}^* \rangle - \frac{\eta}{2} \mathbb{E}[\|\mathbf{p}^{(t)}\|^2] + \eta\sigma^2 \quad (30)$$

$$\begin{aligned} &= C(\mathbf{w}^*) + \eta\sigma^2 + \mathbb{E}[\langle \mathbf{p}^{(t)}, \mathbf{w}^{(t)} - \mathbf{w}^* \rangle - \frac{\eta}{2} \|\mathbf{p}^{(t)}\|^2] \\ &= C(\mathbf{w}^*) + \eta\sigma^2 + \frac{1}{2\eta} (\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 - \|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|^2) \end{aligned}$$

where (28) follows from $L\eta \leq 1$, (29) from the convexity of C , and (30) holds since $\mathbb{E}[\mathbf{p}^{(t)}] = \nabla C(\mathbf{w}^{(t)})$ and $\mathbb{E}[\|\mathbf{p}^{(t)}\|^2] - \|\nabla C(\mathbf{w}^{(t)})\|^2 \leq \sigma^2$ from Lemma 1 with assuming the arbitrarily large r . Summing the above equations for $t = 0, \dots, J - 1$, we have

$$\begin{aligned} &\sum_{t=0}^{J-1} (\mathbb{E}[C(\mathbf{w}^{(t+1)})] - C(\mathbf{w}^*)) \\ &\leq \frac{1}{2\eta} (\|\mathbf{w}^{(0)} - \mathbf{w}^*\|^2 - \|\mathbf{w}^{(J)} - \mathbf{w}^*\|^2) + J\eta\sigma^2 \\ &\leq \frac{\|\mathbf{w}^{(0)} - \mathbf{w}^*\|^2}{2\eta} + J\eta\sigma^2. \end{aligned}$$

Finally, since C is convex, we observe that

$$\begin{aligned} \mathbb{E}\left[C\left(\frac{1}{J} \sum_{t=0}^{J-1} \mathbf{w}^{(t)}\right)\right] &\leq \sum_{t=0}^{J-1} (\mathbb{E}[C(\mathbf{w}^{(t+1)})] - C(\mathbf{w}^*)) \\ &\leq \frac{\|\mathbf{w}^{(0)} - \mathbf{w}^*\|^2}{2\eta J} + \eta\sigma^2, \end{aligned}$$

which completes the proof of convergence.

(Privacy) Proof of T -privacy is deferred to Appendix A.4 in the supplementary materials.

5. Experiments

We now experimentally demonstrate the impact of CodedPrivateML, and make comparisons with existing cryptographic approaches to the problem. Our focus is on training a logistic regression model for image classification, while the computation load is distributed to multiple machines on the Amazon EC2 Cloud Platform.

Setup. We train the logistic regression model from (1) for binary image classification on the MNIST dataset (LeCun et al., 2010) to experimentally examine two things: the accuracy of CodedPrivateML and the performance gain in terms of training time. The size of dataset is $(m, d) = (12396, 1568)$ ¹. Experiments with additional dataset sizes are provided in Appendix A.6 of supplementary material.

We implement CodedPrivateML using the MPI4Py (Dalcín et al., 2005) message passing interface on Python. Computations are performed in a distributed manner on Amazon EC2 clusters using `m3.xlarge` machine instances.

We then compare CodedPrivateML with the MPC-based approach when applied to our problem. In particular, we implement a BGW-style construction (Ben-Or et al., 1988) based on Shamir's secret sharing scheme (Shamir, 1979) where we secret share the dataset among N workers who proceed with a multiround protocol to compute the gradient. We further incorporate the quantization and approximation techniques introduced here as BGW-style protocols are also bound to arithmetic operations over a finite field. See Appendix A.5 of supplementary materials for additional detail.

¹To have a larger dataset we duplicate the MNIST dataset.

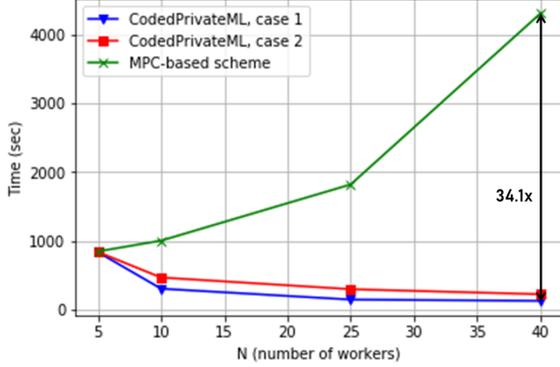


Figure 2. Performance gain of CodedPrivateML over the MPC-based scheme. The plot shows the total training time for accuracy 95.04% (25 iterations) for different number of workers N in Amazon EC2 Cloud Platform.

Table 1. Breakdown of the total run time with $N = 40$ workers.

Protocol	Encode time (s)	Comm. time (s)	Comp. time (s)	Total run time (s)
MPC approach	845.55	49.51	3457.99	4304.60
CodedPrivateML (Case 1)	50.97	3.01	66.95	126.20
CodedPrivateML (Case 2)	90.65	6.45	110.97	222.50

CodedPrivateML parameters. There are several system parameters in CodedPrivateML that should be set. Given that we have a 64-bit implementation, we select the field size to be $p = 15485863$, which is the largest prime with 24 bits to avoid the overflow on intermediate multiplication. We then optimize the quantization parameters, l_x in (6) and l_w in (9), by taking into account the trade-off between the rounding and overflow error. In particular, we choose $l_x = 2$ and $l_w = 4$. We also need to set the parameter r , the degree of the polynomial for approximating the sigmoid function. We consider both $r = 1$ and $r = 2$ and as we show later empirically observe that a degree one approximation provides very good accuracy. We finally need to select T (privacy threshold) and K (amount of parallelization) in CodedPrivateML. As stated in Theorem 1, these parameters should satisfy $N \geq (2r + 1)(K + T - 1) + 1$. Given our choice of $r = 1$, we consider two cases:

- **Case 1 (maximum parallelization).** All resources to parallelization by setting $K = \lfloor \frac{N-1}{3} \rfloor$ and $T = 1$,
- **Case 2 (equal parallelization and privacy).** The resources are split equally by setting $K = T = \lfloor \frac{N+2}{6} \rfloor$,

Training time. In the first set of experiments, we measure the training time while increasing the number of workers N gradually. The results are demonstrated in Figure 2. We make the following observations.²

- CodedPrivateML provides substantial speedup over the MPC approach, in particular, up to $34.1\times$ and $19.4\times$

²For $N = 5$, all schemes have almost same performance because they use same system parameters, $K = T = 1$.

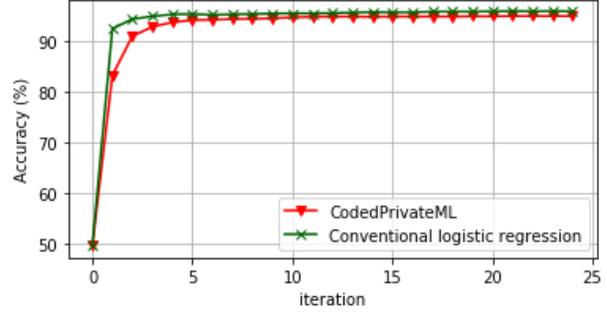


Figure 3. Comparison of the accuracy of CodedPrivateML (demonstrated for Case 2 and $N = 40$ workers) vs conventional logistic regression that uses the sigmoid function without quantization. Accuracy is measured with MNIST dataset restructured for binary classification problem between 3 and 7 (using 12396 samples for the training set and 2038 samples for the test set).

speedup in Cases 1 and 2, respectively. The breakdown of the total run time for one scenario is shown in Table 1. One can note that CodedPrivateML provides significant improvement in all three categories of dataset encoding and secret sharing; communication time between the workers and the master; and computation time. One reason for this is that, in MPC-based schemes, size of the secret shared dataset at each worker is the same as the original dataset, while in CodedPrivateML it is $1/K$ -th of the dataset. This provides a large parallelization gain for CodedPrivateML. The other reason is the communication complexity of MPC-based schemes. We provide the results for more scenarios in Appendix A.6 of supplementary material.

- We note that the total run time of CodedPrivateML decreases as the number of workers increases. This is again due to the parallelization gain of CodedPrivateML (i.e., increasing K while N increases). This parallelization gain is not achievable in MPC-based scheme, since the whole computation has to be repeated by all players who take part in MPC. We should however point out that MPC-based scheme could attain a higher privacy threshold ($T = N/2 - 1$), while CodedPrivateML can achieve $T = \lfloor \frac{N+2}{6} \rfloor$ (Case 2).

Accuracy. We also examine the accuracy and convergence of CodedPrivateML in the experiments. Figure 3 illustrates the test accuracy of the binary classification problem between digits 3 and 7. With 25 iterations, the accuracy of CodedPrivateML with degree one polynomial approximation and conventional logistic regression are 95.04% and 95.98%, respectively. This result shows that CodedPrivateML guarantees almost the same level of accuracy, while being privacy preserving. Our experiments also show that CodedPrivateML achieves convergence with comparable rate to conventional logistic regression. Those results are provided in Appendix A.6 in the supplementary materials.

References

- Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016.
- Ben-Or, M., Goldwasser, S., and Wigderson, A. Completeness theorems for non-cryptographic fault-tolerant distributed computation. In *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing*, pp. 1–10. ACM, 1988.
- Brinkhuis, J. and Tikhomirov, V. *Optimization: Insights and Applications*. Princeton Series in Applied Mathematics. Princeton University Press, 2011.
- Chaudhuri, K. and Monteleoni, C. Privacy-preserving logistic regression. In *Advances in Neural Information Processing Systems*, pp. 289–296, 2009.
- Chen, V., Pastro, V., and Raykova, M. Secure computation for machine learning with SPDZ. *arXiv:1901.00329*, 2019.
- Cover, T. M. and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, 2012.
- Dahl, M., Mancuso, J., Dupis, Y., Decoste, B., Giraud, M., Livingstone, I., Patriquin, J., and Uhma, G. Private machine learning in TensorFlow using secure computation. *arXiv:1810.08130*, 2018.
- Dalcín, L., Paz, R., and Storti, M. MPI for Python. *Journal of Parallel and Distributed Computing*, 65(9):1108–1115, 2005.
- Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pp. 265–284. Springer, 2006.
- Gascón, A., Schoppmann, P., Balle, B., Raykova, M., Dohner, J., Zahur, S., and Evans, D. Privacy-preserving distributed linear regression on high-dimensional data. *Proceedings on Privacy Enhancing Technologies*, 2017 (4):345–364, 2017.
- Gentry, C. and Boneh, D. *A fully homomorphic encryption scheme*, volume 20. Stanford University, Stanford, 2009.
- Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M., and Wernsing, J. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In *International Conference on Machine Learning*, pp. 201–210, 2016.
- Graepel, T., Lauter, K., and Naehrig, M. ML confidential: Machine learning on encrypted data. In *International Conference on Information Security and Cryptology*, pp. 1–21. Springer, 2012.
- Han, K., Hong, S., Cheon, J. H., and Park, D. Logistic regression on homomorphic encrypted data at scale. *Thirty-First Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-19)*, Available online: <https://daejunpark.github.io/iaai19.pdf>, 2019.
- Hesamifard, E., Takabi, H., and Ghasemi, M. CryptoDL: Deep neural networks over encrypted data. *arXiv:1711.05189*, 2017.
- Jayaraman, B., Wang, L., Evans, D., and Gu, Q. Distributed learning without distress: Privacy-preserving empirical risk minimization. In *Advances in Neural Information Processing Systems*, pp. 6346–6357, 2018.
- Kim, A., Song, Y., Kim, M., Lee, K., and Cheon, J. H. Logistic regression model training based on the approximate homomorphic encryption. *BMC Medical Genomics*, 11 (4):23–55, Oct 2018.
- LeCun, Y., Cortes, C., and Burges, C. MNIST handwritten digit database. [Online]. Available: <http://yann.lecun.com/exdb/mnist>, 2, 2010.
- Li, P., Li, J., Huang, Z., Gao, C.-Z., Chen, W.-B., and Chen, K. Privacy-preserving outsourced classification in cloud computing. *Cluster Computing*, pp. 1–10, 2017.
- Lindell, Y. and Pinkas, B. Privacy preserving data mining. In *Annual International Cryptology Conference*, pp. 36–54. Springer, 2000.
- McMahan, H. B., Ramage, D., Talwar, K., and Zhang, L. Learning differentially private recurrent language models. In *International Conference on Learning Representations*, 2018.
- Melis, L., Song, C., Cristofaro, E. D., and Shmatikov, V. Exploiting unintended feature leakage in collaborative learning. *arXiv:1805.04049*, 2019.
- Mohassel, P. and Rindal, P. ABY 3: A mixed protocol framework for machine learning. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 35–52, 2018.
- Mohassel, P. and Zhang, Y. SecureML: A system for scalable privacy-preserving machine learning. In *38th IEEE Symposium on Security and Privacy*, pp. 19–38. IEEE, 2017.
- Nikolaenko, V., Weinsberg, U., Ioannidis, S., Joye, M., Boneh, D., and Taft, N. Privacy-preserving ridge regression on hundreds of millions of records. In *IEEE*

- Symposium on Security and Privacy*, pp. 334–348. IEEE, 2013.
- Pathak, M., Rane, S., and Raj, B. Multiparty differential privacy via aggregation of locally trained classifiers. In *Advances in Neural Information Processing Systems*, pp. 1876–1884, 2010.
- Rajkumar, A. and Agarwal, S. A differentially private stochastic gradient descent algorithm for multiparty classification. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS'12)*, volume 22 of *Proceedings of Machine Learning Research*, pp. 933–941, La Palma, Canary Islands, Apr 2012.
- Shamir, A. How to share a secret. *Communications of the ACM*, 22(11):612–613, 1979.
- Shokri, R. and Shmatikov, V. Privacy-preserving deep learning. In *Proceedings of the 2015 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1310–1321, 2015.
- Wagh, S., Gupta, D., and Chandran, N. Securenn: Efficient and private neural network training. Cryptology ePrint Archive, Report 2018/442, 2018. <https://eprint.iacr.org/2018/442>.
- Wang, Q., Du, M., Chen, X., Chen, Y., Zhou, P., Chen, X., and Huang, X. Privacy-preserving collaborative model learning: The case of word vector training. *IEEE Transactions on Knowledge and Data Engineering*, 30(12): 2381–2393, Dec 2018.
- Yao, A. C. Protocols for secure computations. In *IEEE Annual Symposium on Foundations of Computer Science*, pp. 160–164, 1982.
- Yu, Q., Raviv, N., Kalan, S. M. M., Soltanolkotabi, M., and Avestimehr, A. S. Lagrange coded computing: Optimal design for resiliency, security and privacy. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.
- Yuan, J. and Yu, S. Privacy preserving back-propagation neural network learning made practical with cloud computing. *IEEE Transactions on Parallel and Distributed Systems*, 25(1):212–221, 2014.
- Zhang, H., Li, J., Kara, K., Alistarh, D., Liu, J., and Zhang, C. The ZipML framework for training models with end-to-end low precision: The cans, the cannots, and a little bit of deep learning. *arXiv:1611.05402*, 2016.
- Zhang, H., Li, J., Kara, K., Alistarh, D., Liu, J., and Zhang, C. ZipML: Training linear models with end-to-end low precision, and a little bit of deep learning. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 4035–4043, Sydney, Australia, Aug 2017.

A. Supplementary Materials

A.1. Algorithms

The overall procedure of the CodedPrivateML protocol is given in Algorithm 1. Procedures for individual phases are shown in Algorithms 2-5 for Sections 3.1-3.4, respectively.

Algorithm 1 CodedPrivateML

input Dataset \mathbf{X}, \mathbf{y}
output Model parameters (weights) $\mathbf{w}^{(J)}$

- 1: (Master) Compute the quantized dataset $\bar{\mathbf{X}}$ using (6).
- 2: (Master) Form the encoded matrices $\{\tilde{\mathbf{X}}_i\}_{i \in [N]}$ in (12).
- 3: (Master) Send $\tilde{\mathbf{X}}_i$ to worker $i \in [N]$.
- 4: (Master) Initialize the weights $\mathbf{w}^{(0)} \in \mathbb{R}^{d \times 1}$.
- 5: **for** iteration $t = 0, \dots, J - 1$ **do**
- 6: (Master) Find the quantized weights $\bar{\mathbf{W}}^{(t)}$ from (10).
- 7: (Master) Encode $\bar{\mathbf{W}}^{(t)}$ into $\{\tilde{\mathbf{W}}_i^{(t)}\}_{i \in [N]}$ using (14).
- 8: (Master) Send $\tilde{\mathbf{W}}_i^{(t)}$ to worker $i \in [N]$.
- 9: (Worker $i = 1, \dots, N$) Compute $f(\tilde{\mathbf{X}}_i, \tilde{\mathbf{W}}_i^{(t)})$ from (20) and send the result back to the master.
- 10: **if** Master received results from $(2r+1)(K+T-1)+1$ workers **then**
- 11: (Master) Decode $\{f(\bar{\mathbf{X}}_k, \bar{\mathbf{W}}_k^{(t)})\}_{k \in [K]}$ via polynomial interpolation from the received results.
- 12: **end if**
- 13: (Master) Compute $\sum_{k=1}^K f(\bar{\mathbf{X}}_k, \bar{\mathbf{W}}_k^{(t)})$ in (23) and convert it from finite field to real domain using (24).
- 14: (Master) Update the weight vector via (19).
- 15: **end for**
- 16: **return** $\mathbf{w}^{(J)}$

Algorithm 2 Quantization

input Dataset \mathbf{X} and weights $\mathbf{w}^{(t)}$
output Quantized dataset $\bar{\mathbf{X}}$ and weights $\bar{\mathbf{W}}^{(t)}$

- 1: (Master) Compute the quantized dataset from (6),

$$\bar{\mathbf{X}} = \phi(\text{Round}(2^{l_x} \cdot \mathbf{X}))$$
 using function $\text{Round}(\cdot)$ from (5) and $\phi(\cdot)$ from (7).
- 2: (Master) Compute r independent stochastic quantizations of vector $\mathbf{w}^{(t)}$ given in (9),

$$\bar{\mathbf{w}}^{(t),j} \triangleq Q_j(\mathbf{w}^{(t)}; l_w), \quad j = 1, \dots, r,$$
 by applying the quantization function (8) element-wise over the vector $\mathbf{w}^{(t)}$.
- 3: (Master) Construct the quantized weight matrix in (10),

$$\bar{\mathbf{W}}^{(t)} = [\bar{\mathbf{w}}^{(t),1} \quad \dots \quad \bar{\mathbf{w}}^{(t),r}]$$
 using the quantized vectors $\bar{\mathbf{w}}^{(t),j}$ for $j = 1, \dots, r$.
- 4: **return** $\bar{\mathbf{X}}$ and $\bar{\mathbf{W}}^{(t)}$

Algorithm 3 Encoding and Secret Sharing

input Quantized dataset $\bar{\mathbf{X}}$ and weights $\bar{\mathbf{W}}^{(t)}$
output Encoded dataset $\tilde{\mathbf{X}}_i$ and weights $\tilde{\mathbf{W}}_i^{(t)}$ for $i \in [N]$

- 1: (Master) Partition the quantized dataset $\bar{\mathbf{X}}$ into K submatrices $\bar{\mathbf{X}} = [\bar{\mathbf{X}}_1^\top \dots \bar{\mathbf{X}}_K^\top]^\top$.
- 2: (Master) Construct the encoded matrices $\tilde{\mathbf{X}}_i$ for $i \in [N]$ as in (12) using the Lagrange polynomial from (11).
- 3: (Master) Construct the encoded weights $\tilde{\mathbf{W}}_i^{(t)}$ for $i \in [N]$ as in (14) using the Lagrange polynomial from (13).
- 4: (Master) Send $\tilde{\mathbf{X}}_i$ and $\tilde{\mathbf{W}}_i^{(t)}$ to worker i , where $i \in [N]$.

Algorithm 4 Polynomial Approximation and Local Computations

input Encoded dataset $\tilde{\mathbf{X}}_i$ and weights $\tilde{\mathbf{W}}_i^{(t)}$ for $i \in [N]$
output Computation results $f(\tilde{\mathbf{X}}_i, \tilde{\mathbf{W}}_i^{(t)})$ for $i \in [N]$

- 1: (Master) Find the polynomial approximation coefficients $\{c_i\}_{i=0}^r$ from (15), by fitting the sigmoid function to a degree r polynomial via least squares.
- 2: (Master) Send the coefficients $\{c_i\}_{i=0}^r$ to all workers.
- 3: (Worker $i = 1, \dots, N$) Locally compute the function,

$$f(\tilde{\mathbf{X}}_i, \tilde{\mathbf{W}}_i^{(t)}) = \tilde{\mathbf{X}}_i^\top \bar{g}(\tilde{\mathbf{X}}_i, \tilde{\mathbf{W}}_i^{(t)})$$

using $\tilde{\mathbf{X}}_i$ and $\tilde{\mathbf{W}}_i^{(t)}$ as given in (20), and send the result back to the master.

Algorithm 5 Decoding and Model Update

input Computation results $f(\tilde{\mathbf{X}}_i, \tilde{\mathbf{W}}_i^{(t)})$ from fastest $(2r+1)(K+T-1)+1$ workers
output Updated weights $\mathbf{w}^{(t+1)}$

- 1: (Master) Collect the results from $(2r+1)(K+T-1)+1$ fastest workers.
- 2: (Master) Decode $\left\{f(\bar{\mathbf{X}}_k, \bar{\mathbf{W}}_k^{(t)})\right\}_{k \in [K]}$ from (21) through polynomial interpolation using the computations $f(\tilde{\mathbf{X}}_i, \tilde{\mathbf{W}}_i^{(t)})$ from (22) that are received from $(2r+1)(K+T-1)+1$ workers.
- 3: (Master) Compute $\sum_{k=1}^K f(\bar{\mathbf{X}}_k, \bar{\mathbf{W}}_k^{(t)}) = \bar{\mathbf{X}}^\top \bar{g}(\bar{\mathbf{X}}, \bar{\mathbf{W}}^{(t)})$ and convert the result from finite field to the real domain using (24).
- 4: (Master) Update the weight vector according to (19),

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \frac{\eta}{m} \bar{\mathbf{X}}^\top (\bar{g}(\bar{\mathbf{X}}, \bar{\mathbf{W}}^{(t)}) - \mathbf{y}).$$

5: **return** $\mathbf{w}^{(t+1)}$

A.2. Proof of Lemma 1

(Unbiasedness) Given $\bar{\mathbf{X}}$, we have

$$\begin{aligned}\mathbb{E}[\mathbf{p}^{(t)}] &= \frac{1}{m} \bar{\mathbf{X}}^\top \left(\mathbb{E}[\bar{g}(\bar{\mathbf{X}}, \bar{\mathbf{W}}^{(t)})] - \mathbf{y} \right) \\ &= \frac{1}{m} \bar{\mathbf{X}}^\top \left(\hat{g}(\bar{\mathbf{X}} \times \mathbf{w}^{(t)}) - \mathbf{y} \right)\end{aligned}\quad (31)$$

where (31) follows (18). Then, we obtain

$$\begin{aligned}\mathbb{E}[\mathbf{p}^{(t)}] - \nabla C(\mathbf{w}^{(t)}) &= \frac{1}{m} \bar{\mathbf{X}}^\top \left(\hat{g}(\bar{\mathbf{X}} \times \mathbf{w}^{(t)}) - g(\bar{\mathbf{X}} \times \mathbf{w}^{(t)}) \right).\end{aligned}\quad (32)$$

Assume $\mathbf{w}^{(t)}$ is constrained such that $\|\mathbf{w}^{(t)}\| \leq R$ for some real value $R \in \mathbb{R}$ (Zhang et al., 2016, Lemma 4.2). Then, from the Weierstrass approximation theorem (Brinkhuis & Tikhomirov, 2011), for every $\epsilon > 0$, there exists a polynomial that approximates the sigmoid arbitrarily well, i.e., $|\hat{g}(x) - g(x)| \leq \epsilon$ for all x in the constrained interval. Therefore, given $\bar{\mathbf{X}}$, there exists a polynomial making the norm of (32) arbitrarily small.

(Variance bound) The variance of $\mathbf{p}^{(t)}$ satisfies,

$$\begin{aligned}\mathbb{E}[\|\mathbf{p}^{(t)} - \mathbb{E}[\mathbf{p}^{(t)}]\|_2^2] &= \frac{1}{m^2} \mathbb{E}[\|\bar{\mathbf{X}}^\top (\bar{g}(\bar{\mathbf{X}}, \bar{\mathbf{W}}^{(t)}) - \hat{g}(\bar{\mathbf{X}} \times \mathbf{w}^{(t)}))\|_2^2] \\ &= \frac{1}{m^2} \mathbb{E}[\text{Tr}(\bar{\mathbf{X}}^\top \mathbf{q}^{(t)} \mathbf{q}^{(t)\top} \bar{\mathbf{X}})] \\ &= \frac{1}{m^2} \text{Tr}(\bar{\mathbf{X}}^\top \mathbb{E}[\mathbf{q}^{(t)} \mathbf{q}^{(t)\top}] \bar{\mathbf{X}})\end{aligned}\quad (33)$$

where $\text{Tr}(\cdot)$ denotes the trace of a matrix, and we let $\mathbf{q}^{(t)} \triangleq \bar{g}(\bar{\mathbf{X}}, \bar{\mathbf{W}}^{(t)}) - \hat{g}(\bar{\mathbf{X}} \times \mathbf{w}^{(t)})$.

From Lemma 4 of (Zhang et al., 2016), we have that

$$\mathbb{E}[q_i^{(t)} q_j^{(t)}] \begin{cases} \leq 2^{-2l_w} \left(\sum_{k=0}^r c_k (\bar{\mathbf{x}}_i \cdot \mathbf{w}^{(t)})^k \right)^2 & \text{if } i=j, \\ = 0 & \text{otherwise} \end{cases}\quad (34)$$

where $q_i^{(t)}$ denotes the i^{th} element of $\mathbf{q}^{(t)}$.

Combining equations (33) and (34) with the fact that $\left(\sum_{k=0}^r c_k (\bar{\mathbf{x}}_i \cdot \mathbf{w}^{(t)})^k \right)^2 \approx (g(\bar{\mathbf{x}}_i \cdot \mathbf{w}^{(t)}))^2 \leq 1$ for all $i \in [m]$, we obtain

$$\begin{aligned}\mathbb{E}[\|\mathbf{p}^{(t)} - \mathbb{E}[\mathbf{p}^{(t)}]\|_2^2] &\leq \frac{1}{2^{-2l_w} m^2} \text{Tr}(\bar{\mathbf{X}}^\top \bar{\mathbf{X}}) \\ &= \frac{1}{2^{-2l_w} m^2} \|\bar{\mathbf{X}}\|_F^2\end{aligned}$$

A.3. Proof of Lemma 2

For the logistic regression cost function $C(\mathbf{w})$, the Lipschitz constant L is less or equal than the largest eigenvalue of the Hessian $\nabla^2 C(\mathbf{w})$ for all \mathbf{w} and is given by

$$L = \frac{1}{4} \max \left\{ \text{eig}(\bar{\mathbf{X}}^\top \bar{\mathbf{X}}) \right\}.\quad (35)$$

A.4. Privacy Proof of Theorem 1

Let $\mathbf{U}^{\text{top}} \in \mathbb{F}_p^{K \times N}$ and $\mathbf{U}^{\text{bottom}} \in \mathbb{F}_p^{T \times N}$ are the top and bottom submatrix of the encoding matrix \mathbf{U} constructed in Section 3.2, respectively. From Lemma 2 of (Yu et al., 2019), $\mathbf{U}^{\text{bottom}}$ is an MDS matrix. Therefore, every $T \times T$ submatrix of $\mathbf{U}^{\text{bottom}}$ is invertible.

For a colluding set of workers $\mathcal{T} \subset [N]$ of size T , their received dataset satisfies

$$\tilde{\mathbf{X}}_{\mathcal{T}} = \bar{\mathbf{X}} \times \mathbf{U}_{\mathcal{T}}^{\text{top}} + \mathbf{Z} \times \mathbf{U}_{\mathcal{T}}^{\text{bottom}}\quad (36)$$

where $\mathbf{Z} = (\mathbf{Z}_K, \dots, \mathbf{Z}_{K+T})$, and $\mathbf{U}_{\mathcal{T}}^{\text{top}} \in \mathbb{F}_p^{K \times T}$ and $\mathbf{U}_{\mathcal{T}}^{\text{bottom}} \in \mathbb{F}_p^{T \times T}$ are the top and bottom submatrices which correspond to the columns in \mathbf{U} that are indexed by \mathcal{T} . Since $\mathbf{U}_{\mathcal{T}}^{\text{bottom}}$ is invertible, $\tilde{\mathbf{X}}_{\mathcal{T}}$ is completely masked by the random matrix \mathbf{Z} . Similarly, $\tilde{\mathbf{W}}_{\mathcal{T}}^{(t)}$ is completely masked by the random matrix $\mathbf{V} = (\mathbf{V}_{K+1}, \dots, \mathbf{V}_{K+T})$ for all $t \in [J]$, where J is the total number of iterations.

Since $\tilde{\mathbf{X}}_{\mathcal{T}}$ and $\tilde{\mathbf{W}}_{\mathcal{T}}^{(t)}$ are completely masked by the random padding matrices, T colluding workers get no information about $\bar{\mathbf{X}}$, i.e., $I(\bar{\mathbf{X}}; \tilde{\mathbf{X}}_{\mathcal{T}}, \{\tilde{\mathbf{W}}_{\mathcal{T}}^{(t)}\}_{t \in [J]}) = 0$. Then, from the data-processing inequality (Cover & Thomas, 2012),

$$I(\bar{\mathbf{X}}; \tilde{\mathbf{X}}_{\mathcal{T}}, \{\tilde{\mathbf{W}}_{\mathcal{T}}^{(t)}\}_{t \in [J]}) \geq I(\mathbf{X}; \tilde{\mathbf{X}}_{\mathcal{T}}, \{\tilde{\mathbf{W}}_{\mathcal{T}}^{(t)}\}_{t \in [J]}) \geq 0.\quad (37)$$

Therefore, $I(\mathbf{X}; \tilde{\mathbf{X}}_{\mathcal{T}}, \{\tilde{\mathbf{W}}_{\mathcal{T}}^{(t)}\}_{t \in [J]}) = 0$ and the original dataset remains information-theoretically private against T colluding workers.

A.5. Details of the Implemented MPC-based Scheme

We implement an MPC-based system with a similar privacy structure, that is, any collusions between T out of N workers do not reveal information (in an information-theoretic sense) about the dataset. To do so, we utilize the well-known BGW protocol, a secure MPC protocol that can compute polynomial evaluations privately, by ensuring collusions between up to T out of N workers do not leak information about the input variables (Ben-Or et al., 1988). Due to the polynomial nature of the computations supported by the protocol, we again use polynomial approximation for the sigmoid function. The protocol utilizes Shamir's secret sharing scheme for secret sharing the input variables (Shamir, 1979), which also requires the input variables to be represented in the finite field. Therefore, we again use our quantization technique to convert the dataset and weights from the real to finite domain. The system parameters used for quantization and polynomial approximation are selected to be the same as the ones used for CodedPrivateML.

In order to implement the MPC-based scheme in our problem, we encode the quantized dataset and weights using Shamir's secret sharing. For the (quantized) dataset

$\bar{\mathbf{X}} = [\bar{\mathbf{X}}_1^\top \dots \bar{\mathbf{X}}_K^\top]^\top$, where $\bar{\mathbf{X}}_i \in \mathbb{F}_p^{\frac{m}{K} \times d}$, this is achieved by creating a random polynomial

$$\mathbf{P}_i(z) = \bar{\mathbf{X}}_i + z\mathbf{R}_{i1} + \dots + z^T\mathbf{R}_{iT} \quad (38)$$

for each $i \in [K]$, where \mathbf{R}_{ij} for $j \in [T]$ are i.i.d. uniformly distributed random matrices. Then, each worker is assigned a secret share of the dataset using the polynomial from (38). We note that in this setup each worker receives a share for every $i \in [K]$. Therefore, the total amount of data stored at each worker is equal to the size of the whole dataset $\bar{\mathbf{X}}$. A similar polynomial is created for secret sharing the quantized weights $\bar{\mathbf{W}}^{(t)}$.

The workers then perform addition and multiplication operations on the secret shared data. For performing an addition operation, each worker locally adds its own shares. At the end of this phase, each worker will hold a secret share corresponding to the addition of the original variables. For performing a multiplication operation, workers first multiply their shares locally. After this phase, the protocol requires a communication step to take place between the workers, in which workers create new shares. We implement this communication phase also using the `MP4PY` message passing interface. One can reduce the number of communication rounds by using a vectorized form for operations involving vector products, and implementing a communication step between workers after each vectorized product. In our experiments, we implement this faster vectorized form. The protocol guarantees privacy against $\lfloor \frac{N-1}{2} \rfloor$ colluding workers (Ben-Or et al., 1988). In our experiments, the time spent during the communication phase between workers is included in the reported computation time.

A.6. Additional Experiments

A.6.1. BREAKDOWN OF THE TOTAL RUN TIME FOR ADDITIONAL SCENARIOS

We present the breakdown of the run time when training is done by different number of workers, using the dataset from Section 5. Tables 2 and 3 demonstrate the corresponding results for $N = 10$ and $N = 25$ workers, respectively. One can note that, in all scenarios, CodedPrivateML provides significant improvement in all three categories of dataset encoding and secret sharing; communication time between the workers and the master; and computation time.

Table 2. Breakdown of the total run time with $N = 10$ workers.

Protocol	Encode time (s)	Comm. time (s)	Comp. time (s)	Total run time (s)
MPC-based scheme	53.87	11.71	957.12	1001.53
CodedPrivateML (Case 1)	21.86	3.31	259.54	303.13
CodedPrivateML (Case 2)	32.20	5.55	390.98	465.52

Table 3. Breakdown of the total run time with $N = 25$ workers.

Protocol	Encode time (s)	Comm. time (s)	Comp. time (s)	Total run time (s)
MPC-based scheme	328.19	30.61	1492.44	1818.63
CodedPrivateML (Case 1)	33.27	3.06	97.46	144.77
CodedPrivateML (Case 2)	78.69	7.12	194.09	295.68

A.6.2. CONVERGENCE OF CODEDPRIVATEML

We also experimentally analyze the convergence behaviour of CodedPrivateML. Figure 4 presents the cross entropy loss for CodedPrivateML versus the conventional logistic regression model, over the dataset from Section 5. The latter setup uses the sigmoid function and no polynomial approximation, in addition, no quantization is applied to the dataset or the weight vectors. We observe that CodedPrivateML achieves convergence with comparable rate to conventional logistic regression. This result shows that CodedPrivateML guarantees almost the same convergence rate, while being privacy preserving.

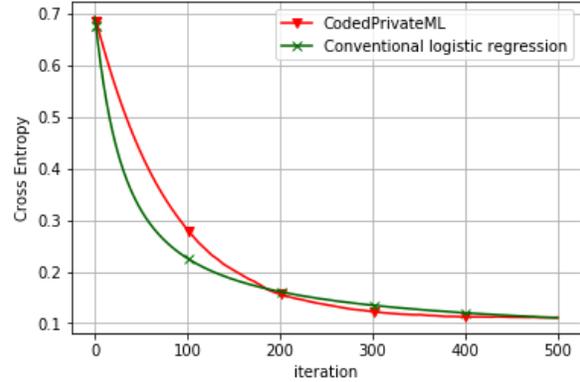


Figure 4. Convergence of CodedPrivateML (demonstrated for Case 2 and $N = 40$ workers) vs conventional logistic regression (using the sigmoid without polynomial approximation or quantization).

A.6.3. EXPERIMENTS FOR A SMALLER DATASET

In this section, we demonstrate the performance of CodedPrivateML on a smaller dataset, by considering $(m, d) = (12396, 784)$. Figure 5 illustrates the training time while increasing the number of workers N gradually. Tables 4-6 provide the breakdown of the run time for the training phase, with 10, 25, and 40 workers, respectively.

Upon inspecting the performance gains from Tables 1-3 versus Tables 6-4, we conclude that CodedPrivateML achieves a higher performance gain (from $26.2\times$ to $34.1\times$ when $N = 40$) as the dimension of dataset gets larger. Our interpretation of this behaviour is based on the following observation. Increasing the number of workers in the system has two major impacts on the training time of CodedPrivateML. The first one is reducing the computation load per worker, as each new worker can be used to increase the parameter K (i.e., the parallelization gain). This in turn reduces the

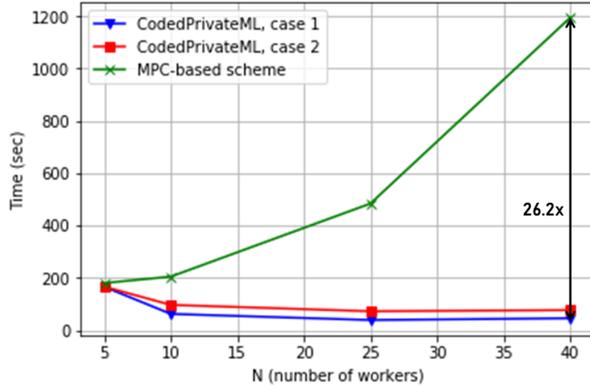


Figure 5. Performance gain of CodedPrivateML over the MPC-based scheme with the smaller dataset. The plot shows the total training time for accuracy 95.04% (25 iterations) for different number of workers N in Amazon EC2 Cloud Platform.

computation load per worker as the amount of work done by each worker is scaled with respect to $1/K$. The second one is that increasing the number of workers increases the encoding time. Therefore, for small datasets, i.e., when the computation load at each worker is small, the gain from increasing the number of workers beyond a certain point may be minimal and the system may saturate. A similar behaviour is observed in Figure 5 when the number of workers is increased from $N = 25$ to $N = 40$.

Therefore, in order to achieve the best performance gain, we find that CodedPrivateML is well suited for data-intensive distributed training environments for processing large datasets. Furthermore, it can be tuned to meet the specific performance guarantees required by different applications, i.e., a faster implementation versus more privacy.

Table 4. Breakdown of the run time with $(m, d) = (12396, 784)$ for $N = 10$ workers.

Protocol	Encode time (s)	Comm. time (s)	Comp. time (s)	Total run time (s)
MPC-based scheme	26.70	5.41	177.44	204.86
CodedPrivateML (Case 1)	8.15	1.26	50.97	62.23
CodedPrivateML (Case 2)	15.97	2.33	76.46	96.70

Table 5. Breakdown of the run time with $(m, d) = (12396, 784)$ for $N = 25$ workers.

Protocol	Encode time (s)	Comm. time (s)	Comp. time (s)	Total run time (s)
MPC-based scheme	166.04	14.87	316.55	484.09
CodedPrivateML (Case 1)	16.40	1.29	18.26	38.87
CodedPrivateML (Case 2)	28.62	3.12	38.26	72.39

Table 6. Breakdown of the run time with $(m, d) = (12396, 784)$ for $N = 40$ workers.

Protocol	Encode time (s)	Comm. time (s)	Comp. time (s)	Total run time (s)
MPC-based scheme	418.07	24.33	774.20	1194.12
CodedPrivateML (Case 1)	26.64	1.11	11.18	45.58
CodedPrivateML (Case 2)	46.52	2.79	21.07	76.81