Label Private Deep Learning Training based on Secure Multiparty Computation and Differential Privacy

Sen Yuan, Milan Shen, Ilya Mironov, Anderson Nascimento
Facebook Inc.
yuansen@fb.com, milanshen@fb.com, imironov@fb.com, andclay@fb.com

Introduction Secure multiparty computation (MPC) [1, 2] is one of the basic building blocks of privacy preserving machine learning (PPML). MPC is particularly useful in a scenario where several parties need to compute a function of a dataset that cannot be directly accessed by the computing parties. This can happen, for example, when the private input data does not come directly from data holders but is the result of previous secure computations. In such scenarios, no party holds the private data in the clear: it is secret shared among the computing parties. The cost of training the a machine learning model using solely MPC can be several orders of magnitude more expensive than training in the clear. In this contribution, we show that it is possible to dramatically speed-up applications of privacy preserving machine learning on top of MPC by releasing differentially private information about the players inputs throughout the computation. Because of the guarantees of differential privacy, the information release does not affect the privacy of individual entries, while runtimes can be improved by orders of magnitude. We apply our ideas to the problem of training machine learning models when one of the parties is in possession of training features, and the corresponding labels for each input are secret, potentially coming from previous computations, and are secret shared between the computing parties. Note that since the labels are not directly accessible by any of the computing parties, a direct application of local differential privacy is not possible. This problem is relevant for a wealth of applications. For example, this scenario is relevant for computing models for predicting ads conversions where one party knows the training features, while the labels (conversion or no conversion) are usually not held neither of them and need to be computed using MPC with the parties' inputs (timestamps of clicks and/or impressions). A direct application of MPC to solve this problem would secret share all the inputs to the computation (features and labels) and use them as input to an MPC protocol. In the case of a deep learning model, that would imply the computation of several activation functions, gradients of loss functions, inner products - all expensive computations when carried over MPC. For a network with thousands of neurons, that would imply millions of secure computations per dataset entry per epoch. We will show that by releasing information during the MPC execution (in a differentially private way), we can drastically reduce the computational complexity of such protocols. We present two different protocols for solving this problem: one based on local differential privacy [3] and another one based on a modification of the well-known differentially private stochastic gradient descent (DP-SGD). [4].

Label Privacy and Privacy Preserving Machine Learning: Chase et. al. [5] proposed a protocol for training a neural network with differential privacy in a scenario where a dataset is distributed horizontally across several parties, so labels and features are known to these parties. MPC is used to aggregate and add differential privacy to batches locally trained in the clear by each one of these parties. This is clearly different from our scenario where labels are not known by any party. Thus, training batches using local computations in the clear is not possible. Label privacy was initially studied by [6] and was applied to linear regression in [7]. In a recent paper [8], Ghazi et. al. attacked the problem of using deep learning when only labels need to be kept private, which is similar to our problem. However, in their paper, they assume that the labels are accessible either by a trusted curator responsible for implementing differential privacy mechanisms or by the data owners themselves. It differs from our practical applications hence MPC computation is well suited to be leveraged.

Label Differential Privacy [8]- Let ϵ and δ be non-negative constants. A randomized training algorithm $\mathcal{A}$ taking as input a dataset is said to be (ϵ, δ) label differentially private (ϵ, δ)-LabelDP) if

Preprint. Under review.

for any two training datasets D and D' that differ in the label of a single example, and for any subset S of outputs of $\mathcal{A}$, we have that $Pr[\mathcal{A}(D) \in S] \leq e^{\epsilon} Pr[\mathcal{A}(D') \in S] + \delta$.

Secure Multi Party Computation (MPC) Building Blocks: Our proposed solutions work under any secure general multiparty computational protocol. We provide implementations using Crypten[1] [9]. We refer to [9] for a detailed description of the implementation of our basic MPC building blocks. For ease of comprehension, we will state our protocols for n=2 (two computing parties) and using additive secret sharing. An additive secret sharing of a value $x \in \mathbb{Z}_q = \{0, 1, \cdots, q-1\}$ is a pair of random numbers x_a, x_b chosen uniformly from $\mathbb{Z}_q$ subject to the restriction that $x = x_a + x_b \mod q$. We denote the secret sharing of x by $[\![x]\!]_q = (x_a, x_b)$. Notice that two players can easily compute shares of the addition of two secrets by just locally adding their respective shares, i.e. $[\![x]\!]_q + [\![y]\!]_q = [\![x+y]\!]_q = (x_a + y_a \mod q, x_b + y_b \mod q)$. To compute the multiplication of two secrets we use the protocol proposed by Beaver in [10], which is based on random pre-computed triples and requires one round of communications between Alice and Bob and the exchange of two ring elements from $\mathbb{Z}_q$. Let $x_1, x_2, \cdots, x_l$ be the binary representation of x. It is convenient to transform $[\![x]\!]_q$ into $[\![x_1]\!]_2, [\![x_2]\!]_2, \cdots, [\![x_l]\!]_2$, that is to transform a secret x shared over $\mathbb{Z}_q$ into binary secret shares of the bit of the binary representation of x. Such a task is accomplished by bit decomposition protocols [11]. There are also simple procedures for doing the reverse task, computing $[\![x]\!]_q$ from $[\![x_1]\!]_2, [\![x_2]\!]_2, \cdots, [\![x_l]\!]_2$ [12]. We also use secure comparison protocols [13, 14, 15]. In a secure comparison protocol, two parties hold secret inputs x and y respectively, and would like to check if $x < y$ or not. The outcome of the protocol is a bit shared between the two players which indicates if the input x is smaller than the input y or not but without revealing any other information about the inputs. We slightly abuse our notation and also use $[\![x]\!]_q$ when x is a vector with coordinates belonging into $\mathbb{Z}_q$. We denote the secure component wise multiplication (Hadamard product) of two vectors x and y by $[\![x]\!]_q \odot [\![y]\!]_q$. Finally, we note that since all of our computations happen over $\mathbb{Z}_q$, we need to map finite precision real valued inputs into appropriate integers. Details on how this mapping happens are presented in [9].

Problem Description: We work in a two-party scenario. We are motivated by the use case when the model label is created from two data holders combining their data together and it is considered private such that we don't want to reveal it to neither party. Thus, in our scenario, one of the parties holds training features, while the label is secret shared between them. The parties have shared identifiers for each row of data, or training example, that can be used to align the dataset during the training process. We'll refer to the party holding the training features as Alice and the other party as Bob. We are interested in training a deep learning model in such a situation. The model will be made available to Alice after training. Bob will not receive any output in our protocols. Informally, we say a protocol is secure if, upon protocol completion, Bob learns nothing, and Alice learns solely differentially private information about the labels. We work in a scenario where there is no trusted curator to obtain global differential privacy. Moreover, since the labels, cannot be directly accessed by any party and are secretly shared between Alice and Bob, a direct use of local differential privacy is also out of question. The model will be trained using pairs $(x_i, y_i), 1 \leq i \leq n$. Alice is in possession of x_i, while y_i is a label secret shared between Alice and Bob and not known to any of them. The desired output is a vector of weights $W = (w_1, ..., w_d)$ that will be known only to Alice. Bob receives no output at the end of the protocol.

Protocol I - Randomized Response: Our first protocol is based on the randomized response mechanism. For the sake of simplicity, we restrict our analysis to the case where the labels are binary. An extension to multiple classes is presented in the appendix A.1. The main idea here is for Alice and Bob randomize the label y_i and reveal it in the clear. Since Alice and Bob cannot directly access the labels in the clear, by randomizing we mean that Alice and Bob will compute (using a secure two-party computation protocol) a bit $\hat{y}_i$ so that $\hat{y}_i = 1$ with probability p, for an appropriately chosen bias p. Alice and Bob then privately XOR $y_i$and $\hat{y}_i$ and release the result. A direct approach to generate $\hat{y}_i$ is for Alice and Bob to generate λ random bits each one of them and interpret these random bits as the binary expansion of a uniform number r in $[0, 1]$. Alice and Bob then use a private comparison protocol and check if $r < p$. Alice and Bob can privately evaluate the XOR of bits a and b by computing the formula: a XOR $b = a + b - 2ab$ over MPC. Once the $\hat{y}_i$ XOR y_i are available, we can use the training strategy proposed in Ghazi et. al. [8] to train a deep learning model with the randomized labels and the corresponding features x_i. Before presenting our protocol, we briefly remark that Alice and Bob can non-interactively produce shares of a random bit

[1] https://crypten.ai

$[\![r]\!]_2$ by locally picking up random bits r_a and r_b and defining $r = r_a + r_b \mod 2$. The privacy of the protocol follows from the security of the MPC building blocks and the definition of local label differential privacy.

Algorithm 1: Randomized Response based Solution

```
input  : Alice inputs $x_i$, Alice and Bob input shares $[\![y]\!]_i$, $\{1 \leq i \leq n\}$, $p$ is a public parameter
         in $[0, 1]$
output : Trained Model
1  Alice and Bob locally generate lambda shares of random bits $[\![r_j]\!]_2, 1 \leq j \leq \lambda$, and define
     $r_1, ..., r_\lambda$ as the binary extension of a real number $r$ in $[0, 1]$;
2. Alice and Bob run a secure comparison protocol and obtain $[\![r < p]\!]_q$. The output of this
     protocol is a bit that is equal to one if $r < p$ or zero if $r \geq p$. Denote the output by $o_i$ and note it
     is secret shared between Alice and Bob. None of these parties know its value;
3  Alice and Bob compute $[\![\hat{y}_i]\!]_q = [\![y_i]\!]_q$ XOR $[\![o_i]\!]_q = [\![y_i]\!]_q + [\![o_i]\!]_q - 2[\![y_i]\!]_q[\![o_i]\!]_q$. Bob announces
     his shares of this computation in the clear to Alice. Alice recovers $\hat{y}_i$ in the clear;
4  Alice and Bob repeat steps 1,2 and 3 $n$ times for $y_1, ..., y_n$;
5  Alice uses $(x_i, \hat{y}_i)$ to compute the model in the clear using the training strategy proposed in [8];
```

Second Protocol - Label DP-SGD: While our randomized response algorithm is efficient and provides good accuracy, the utility of randomized response data decreases substantially for the case when the label space cardinality is high. Additionally, it does not work in the case one is interested in regression, rather than classification. In order to cope with these cases, we propose another protocol - an adaptation of DP-SGD to a label privacy scenario. Interestingly, we show that when only label privacy is required, no gradient clippings is necessary, contrary to what happens in the original DP-SGD [4]. Assume our dataset is of the form (x_i, y_i), where x_i represents the features in possession of Alice for the i-th data set entry, and y_i represents the corresponding attributed labels that are secret shared between Alice and Bob. The goal is to train a model that depends on d parameters here represented by W. Our start idea is to run DP-SGD [4] on top of MPC. Since Alice has all the features x_i, forward passes can be performed in the clear. MPC will be used for computing gradients, aggregating them and adding noise. The result is then released in the clear and Alice can do a back propagation, updating the weights in the clear. Note that label information is only revealed in an aggregated format and after noise is added, so no violation of individual label privacy happens. Moreover, Alice's information is never sent to the Bob. However, we now show that since in our problem features are known by Alice and only the labels are unknown, we can improve the utility of differentially private stochastic gradient descent. In the usual DP-SGD algorithm, gradients need to be kept differentially private. Thus, they need to be clipped, aggregated and noise should be added to them. We now point out that when only labels need to be kept private, clipping is not needed. We work with cross-entropy loss function and with a binary classification problem (for the sake of simplicity). However, our results naturally extend to other loss functions as well as to multi class problems - these extensions are presented in the appendix A.2 . In the following, $\mathcal{L}$ denotes the loss function, p_i denotes the output probability of the output neuron for input x_i, and W represents the set of parameters. The design hinges on a key observation from the chain rule: $\frac{d\mathcal{L}(y_i,p_i)}{dW} = \frac{d\mathcal{L}(y_i,p_i)}{dZ_i} \cdot \frac{dZ_i}{dW} = (p_i - y_i) \cdot \frac{dZ_i}{dW}$, where Z_i is the output layer neuron value before the sigmoid transformation. Assume that the size of the mini batch is N.

Thus, the gradient depends on a scalar component $(p_i - y_i)$ times a vector dZ_i/dW (which depends on the label y_i). The scalar quantity is bounded $[-1, 1]$ and since only the labels need to be kept differentially private, we compute the sensitivity of $d\mathcal{L}(y_i, p_i)/dW$ based on variations on y_i. The overall sensitivity of $d\mathcal{L}(y_i, p_i)/dW$ is bounded by the maximum L_2 norm among all the vector dZ_i/dW within a mini batch. Denote such value by g and note it is known by Alice. We then should compute the aggregated noisy gradients $1/N(\sum_{i=1}^{N} d\mathcal{L}(y_i, p_i)/dW + \mathcal{N}(0, (g\sigma)^2 I_d))$, where $\mathcal{N}(0, (g\sigma)^2 I_d)$ denotes a Gaussian vector of dimension d (the total number of parameters of the model) and variance $(g\sigma)^2$ for an appropriately chosen constant $\sigma > 0$. The aggregated noisy gradient is then revealed and Alice can use it to update the d parameters of her model. Note that when computing $d\mathcal{L}(y_i, p_i)/dW$ only the quantity $(p_i - y_i)$ needs to be computed jointly by Alice and Bob. dZ_i/dW can be computed in the clear by Alice.

We now show that we can reduce the amount of noise necessary to make our solution differentially private. We do so by exploiting the fact we are interested in label privacy and by using the iterative nature of the back propagation algorithm. Let L denote the total number of layers in our network, the L-th layer being the output layer. Rather than adding noise to the entire d-dimensional vector $1/N(\sum_{i=1}^{N} d\mathcal{L}(y_i, p_i)/dW)$ at once, we add noise to the aggregated gradient corresponding to the $L-1$ layer (representing the weights connected to the output layer of the neural network). Denote such gradient by $1/N(\sum_{i=1}^{N} d\mathcal{L}(y_i, p_i)/dW^{L-1})$. We compute $1/N(\sum_{i=1}^{N} d\mathcal{L}(y_i, p_i)/dW^{L-1} + \mathcal{N}(0, (g_t\sigma)^2 I_t))$, where g_t is the sensitivity of $1/N(\sum_{i=1}^{N} d\mathcal{L}(y_i, p_i)/dW^{L-1})$ wrt variations on y_i. Alice uses the noisy aggregated t-dimensional gradient $1/N(\sum_{i=1}^{N} d\mathcal{L}(y_i, p_i)/dW^{L-1} + \mathcal{N}(0, (g_t\sigma)^2 I_t))$ to update the t parameters corresponding to the weights connected to the output layer of the network. These t dimensional noisy gradients are then back propagated to the parameters in remaining layers of the network. Differential privacy is ensured by its post-processing property. We remark that the variance of the noise added is independent of the total number of parameters of the model d, depending only on the number of weights connected to the last layer t. Our solution is presented in Algorithm 2. The security analysis will be presented in an extended version.

Algorithm 2: Label DP-SGD

```
input  : Alice inputs $x_i$, Alice and Bob input shares $[\![y_i]\!]_q$ $\{1 \leq i \leq n\}$, mini-batch size $N$, the
         number of weights connected to the output layer $t$, and $\sigma > 0$.
output : Trained model for Alice. No output for Bob
for each mini batch do
    Alice forward prop and outputs $p_i$ (output probability of the output neuron) using input $x_i$;
    Alice and Bob Compute $[\![p_i - y_i]\!]_q$ ;
    Alice computes $dZ_i/dW^{L-1}$ and secret shares it with Bob;
    Alice and Bob compute per sample gradient
      $[\![d\mathcal{L}(y_i, p_i)/dW]\!]_q = [\![(p_i - y_i)]\!]_q [\![dZ_i/dW^{L-1}]\!]_q$;
    Alice and Bob compute aggregated gradient $[\![\sum_1^N (p_i - y_i) dZ_i/dW^{L-1}]\!]_q$;
    Alice computes $g_t$ (maximum $L_2$ norm of $dZ_i/dW^{L-1}$ across all the mini batch);
    Bob generates in the clear $\mathcal{N}(0, I_t)$, a $t$-dimensional Gaussian noise vector with mean zero
      and variance one;
    Bob computes the secret share of the Gaussian noise vector $[\![\mathcal{N}(0, I_t)]\!]_q$ with Alice;
    Alice generates in the clear the square root of the variance needed for DP noise, $(g_t\sigma)^2$;
    Alice computes the secret shares the square root of the variance $[\![(g_t\sigma)^2]\!]_q$ with Bob;
    Alice and Bob multiply $[\![\mathcal{N}(0, I_t)]\!]_q$ times $[\![\sqrt{(g_t\sigma)^2}]\!]_q$. This result is added to the
      aggregated gradients for the mini batch and divided by $N$ resulting in
      $[\![1/N\{\sum_1^N (p_i - y_i) dZ_i/dW^{L-1} + \mathcal{N}(0, (g_t\sigma)^2 I_t)\}]\!]$;
    Bob sends his shares to Alice. Alice opens the noisy average gradients and updates the
      weights connected to the output layer of her model. Alice then back propagates these noisy
      weights to the remaining layers/weights;
end
```

Implementation Results We trained a neural network with one convolutional layer (kernel size =5), three fully connected layers (with 256, 256, and 128 neurons respectively) and one output layer (1 neuron) using the traditional approach (secret share all the inputs and train the model using MPC - no release of DP information). The time required for MPC operations in the traditional approach was 38.4s per 100 data set entries. Using Label DP-SGD for (2 classes and $t = 128$) the runtime for the required MPC operations was 8ms per 100 data set entries. To observe what is the effect of Label DP-SGD on accuracy, we have trained, we trained a model with the CIFAR-10[2] image dataset. We applied ResNet18 [16] to the model training. The baseline accuracy (best accuracy for the model without any noise) is 99.2% under our experiment setup. Using Label DP-SGD, the accuracy drop is less than 4.2% when epsilon is greater than or equal to 4. This is a dramatic improvement over DP-SGD. See appendix B for a detailed description of our results.

[2] https://www.cs.toronto.edu/ kriz/cifar.html

References

[1] David Chaum, Claude Crépeau, and Ivan Damgard. Multiparty unconditionally secure protocols. In *Proceedings of the twentieth annual ACM symposium on Theory of computing*, pages 11–19, 1988.

[2] Oded Goldreich, Silvio Micali, and Avi Wigderson. How to play any mental game, or a completeness theorem for protocols with honest majority. In *Providing Sound Foundations for Cryptography: On the Work of Shafi Goldwasser and Silvio Micali*, pages 307–328. 2019.

[3] Graham Cormode, Somesh Jha, Tejas Kulkarni, Ninghui Li, Divesh Srivastava, and Tianhao Wang. Privacy at scale: Local differential privacy in practice. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1655–1658, 2018.

[4] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[5] Melissa Chase, Ran Gilad-Bachrach, Kim Laine, Kristin E Lauter, and Peter Rindal. Private collaborative neural network learning. *IACR Cryptol. ePrint Arch.*, 2017:762, 2017.

[6] Kamalika Chaudhuri and Daniel Hsu. Sample complexity bounds for differentially private learning. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 155–186. JMLR Workshop and Conference Proceedings, 2011.

[7] Di Wang and Jinhui Xu. On sparse linear regression in the local differential privacy model. In *International Conference on Machine Learning*, pages 6628–6637. PMLR, 2019.

[8] Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. On deep learning with label differential privacy. *arXiv preprint arXiv:2102.06062*, 2021.

[9] Brian Knott, Shobha Venkataraman, Awni Hannun, Shubho Sengupta, Mark Ibrahim, and Laurens van der Maaten. Crypten: Secure multi-party computation meets machine learning. In *Proceedings of the NeurIPS Workshop on Privacy-Preserving Machine Learning*, 2020.

[10] Donald Beaver. One-time tables for two-party computation. In *International Computing and Combinatorics Conference*, pages 361–370. Springer, 1998.

[11] Tord Reistad and Tomas Toft. Linear, constant-rounds bit-decomposition. In *International Conference on Information Security and Cryptology*, pages 245–257. Springer, 2009.

[12] Martine De Cock, Rafael Dowsley, Caleb Horst, Raj Katti, Anderson CA Nascimento, Wing-Sea Poon, and Stacey Truex. Efficient and private scoring of decision trees, support vector machines and logistic regression models based on pre-computation. *IEEE Transactions on Dependable and Secure Computing*, 16(2):217–230, 2017.

[13] Thijs Veugen, Frank Blom, Sebastiaan JA de Hoogh, and Zekeriya Erkin. Secure comparison protocols in the semi-honest model. *IEEE Journal of Selected Topics in Signal Processing*, 9(7):1217–1228, 2015.

[14] Juan Garay, Berry Schoenmakers, and José Villegas. Practical and secure solutions for integer comparison. In *International Workshop on Public Key Cryptography*, pages 330–342. Springer, 2007.

[15] Florian Kerschbaum, Debmalya Biswas, and Sebastiaan de Hoogh. Performance comparison of secure comparison protocols. In *2009 20th International Workshop on Database and Expert Systems Application*, pages 133–136. IEEE, 2009.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

A Extension to Multi Class Cases

A.1 Randomized Response based Protocol

We now extend our randomized response-based protocol to the multi class case. Wlog we assume that $y_i \in \mathbb{Z}_c = \{0, 1, \cdots, c-1\}$, for some integer $c > 2$. Denote by y_i the label for the $i-$th input and by $\hat{y}_i$ its randomized version. The probability that y_i is not flipped is $Pr[y_i = \hat{y}_i] = \frac{e^{\epsilon}}{e^{\epsilon}+c-1}$. Otherwise, the label y_i will be flipped (with probability $\frac{1}{e^{\epsilon}+c-1}$) into one of the $c-1$ values $\mathbb{Z}_c \setminus y_i$ according to a uniform probability distribution.

Algorithm 3: Randomized Response based Solution - Multi Class Extension

```
input  : Alice inputs $x_i$, Alice and Bob input the secret shares shares $\llbracket y_i \rrbracket, \{1 \leq i \leq n\}$, $p$ is a
         public parameter in $[0, 1]$
output : Trained Model
2  Alice and Bob locally generate lambda shares of random bits $\llbracket r_j \rrbracket_2, 1 \leq j \leq \lambda$, and define
     $r_1, ..., r_\lambda$ as the binary extension of a real number $r$ in $[0, 1]$;
3  Alice and Bob privately compute $\llbracket r - p \rrbracket_q$ and run a private comparison protocol to check if
     $r - p < 0$. The output of this protocol is a bit $o_i$ secret shared between Alice and Bob and it is
     equal to one if $r - p < 0$ or zero if $r - p \geq 0$;
4  Alice and Bob secret share the publicly known value $\llbracket c - 1 \rrbracket_q$;
5  Alice picks up a random integer $a_i \in \mathbb{Z}_{c-1}$ according to a uniform distribution;
6  Alice picks up a random integer $b_i \in \mathbb{Z}_{c-1}$ according to a uniform distribution;
7  Define $f_i = a_i + b_i \mod c - 1$;
8  Alice and Bob convert $\llbracket f_i \rrbracket_{c-1}$ into $\llbracket f_i \rrbracket_q$;
9  Alice and Bob run a private equality test protocol between $\llbracket y_i \rrbracket_q$ and $\llbracket f_i \rrbracket_q$. Denote the outcome
     of the comparison by $\llbracket h_i \rrbracket_q$. So, $h_i = 1$ iff $y_i = f_i$;
10 Alice and Bob compute $\llbracket \hat{y}_i \rrbracket_q = \llbracket o_i \rrbracket_q \llbracket y_i \rrbracket_q + (1 - \llbracket o_i \rrbracket_q)(\llbracket h_i \rrbracket_q \llbracket c - 1 \rrbracket + (1 - \llbracket h_i \rrbracket_q)\llbracket f_i \rrbracket_q)$;
11 Bob announces his shares of this computation in the clear to Alice. Alice recovers $\hat{y}_i$ in the clear;
12 Alice and Bob repeat steps $1, 2, \cdots, 12$ $n$ times for $y_1, ..., y_n$;
13 Alice uses $(x_i, \hat{y}_i)$ to compute the model in the clear using the training strategy proposed in [8];
```

the arguments for proving correctness and security are the same as in the binary case.

A.2 Local DP-SGD

In order to generalize protocol 2 to the multi class case, we need to basically change the sensitivity analysis and the amount of noise that needs to be added to gradients to obtain an (ϵ, δ) label private solution. Let's start by the sensitivity analysis.

Assume the multi class label is one hot encoded and c is the position of the ground truth label. For example, if the label is $(0, 0, 1, 0)$, then c is 2 assuming our index starting from 0. Then the cross entropy loss can be defined as

$$\mathcal{L}(c, \mathbf{z}) = -\log \frac{e^{z_c}}{\sum_j e^{z_j}} = -\log p_c \tag{1}$$

where $\mathbf{z} = (z_1, ..., z_K)$ is output represented by logits and p_j is the output probability for output neuron j. Let W_j be the weight vector connecting to the jth logit in the output layer. And let $W = (W_1, ..., W_K)^T$ be the vector representing all weights connecting to output layer. By chain rule,

$$\frac{d\mathcal{L}}{dW} = \sum_{j=1}^{K} [p_j - I(j = c)] \frac{dz_j}{dW} \tag{2}$$

As we change the ground truth label c to c', the sensitivity is

$$||\frac{d\mathcal{L}}{dW}(c) - \frac{d\mathcal{L}}{dW}(c')|| = ||\sum_{j=1}^{K}[I(j=c) - I(j=c')]\frac{dz_j}{dW}|| \leq 2\max_{j}||\frac{dz_j}{dW}|| \tag{3}$$

If we consider sample index i, then equation (2) becomes

$$\frac{d\mathcal{L}}{dW} = \frac{1}{N}\sum_{ij}^{N}[p_{ji} - I(j=c_i)]\frac{dz_{ji}}{dW} \tag{4}$$

Then for any c_i and c_i' the sensitivity analysis becomes:

$$||\frac{d\mathcal{L}}{dW}(c_i) - \frac{d\mathcal{L}}{dW}(c_i')|| = ||\frac{1}{N}\sum_{ij}[I(j=c_i) - I(j=c_i')]\frac{dz_{ji}}{dW}|| \leq \frac{2}{N}\max_{ij}||\frac{dz_{ji}}{dW}|| \tag{5}$$

The nominator is 2 is because there $|I(j=c_i) - I(j=c_i')|$ is 1 only if j is c_i or c_i', otherwise 0. Therefore, the noise added to gradient (average across samples) $\frac{d\mathcal{L}}{dW}$ should be $N(0, \frac{4\sigma^2}{N^2}\max_{ij}||\frac{dz_{ji}}{dW}||^2)$.

The new protocol will be exactly like protocol 2, but with an updated noise variance. In the following $\llbracket y_i \rrbracket_q$ denotes secret shares of an output vector y_i (a one-hot encoding of the label for input i). $\llbracket p_i \rrbracket_q$ represents secret shares of a vector where each component is the output of one neuron in the output layer. Accordingly, dZ_i/dW^{L-1} is a tensor where each coordinate is the corresponding gradient for one of the output neurons.

Algorithm 4: Label Private DP-SGD - Multi Class

```
input  : Alice inputs $x_i$, Alice and Bob input shares $\llbracket y_i \rrbracket_q$ $\{1 \leq i \leq n\}$, mini-batch size $N$, the
         number of weights connected to the output layer $t$, and $\sigma > 0$.
output : Trained model for Alice. No output for Bob
for each mini batch do
    Alice forward prop and outputs $p_i$ (vector with output probability of each output neuron)
      using input $x_i$;
    Alice and Bob Compute $\llbracket p_i - y_i \rrbracket_q$ ;
    Alice computes $dZ_i/dW^{L-1}$ and secret shares it with Bob;
    Alice and Bob compute per sample gradient
      $\llbracket d\mathcal{L}(y_i, p_i)/dW \rrbracket_q = \llbracket (p_i - y_i) \rrbracket_q \odot \llbracket dZ_i/dW^{L-1} \rrbracket_q$, where $\odot$ represents the
      coordinate-wise product;
    Alice and Bob compute aggregated gradient $\llbracket \sum_1^N (p_i - y_i) dZ_i/dW^{L-1} \rrbracket_q$;
    Alice computes $g_t = 2\max_{ij}||\frac{dz_{ji}}{dW}||$;
    Bob generates in the clear $\mathcal{N}(0, I_t)$, a $t$-dimensional Gaussian noise vector with mean zero
      and variance one;
    Bob computes the secret share of the Gaussian noise vector $\llbracket \mathcal{N}(0, I_t) \rrbracket_q$ with Alice;
    Alice generates in the clear the square root of the variance needed for DP noise, $(g_t\sigma)^2$;
    Alice computes the secret shares the square root of the variance $\llbracket (g_t\sigma)^2 \rrbracket_q$ with Bob;
    Alice and Bob multiply $\llbracket \mathcal{N}(0, I_t) \rrbracket_q$ times $\llbracket \sqrt{(g_t\sigma)^2} \rrbracket_q$. This result is added to the
      aggregated gradients for the mini batch and divided by $N$ resulting in
      $\llbracket 1/N\{\sum_1^N (p_i - y_i) dZ_i/dW^{L-1} + \mathcal{N}(0, (g_t\sigma)^2 I_t)\} \rrbracket$;
    Bob sends his shares to Alice. Alice opens the noisy average gradients and updates the
      weights connected to the output layer of her model. Alice then back propagates these noisy
      weights to the remaining layers/weights;
end
```

B Implementation Results

B.1 Complexity Analysis and Runtime

Complexity Analysis We measure the complexity of our protocols by the number of secure multiplications and the round complexity. Our protocol based on randomized response for the binary classification case performs one secure comparison and one secure multiplication per data set entry. The private comparison protocol needs 3λ multiplications. We use $\lambda = 16$ in our experiments. The round complexity of the comparison protocol is equal to λ [9]. The randomized response for c classes needs two secure comparison protocols plus four multiplications per data set entry. The cost of each one of the comparison protocols is 3λ and $\log_2 c$ multiplications, respectively. The round complexity of the protocol is $\max\{3\lambda, \log_2 c\}$. Our protocol based on Label Private DP-SGD requires $2tc$ secure multiplications per data set entry, where t is the number of neurons of the second last layer, and c is the number of classes. The round complexity of the protocol is 3 rounds per mini batch.

Experimental Setup and Runtimes We now describe runtimes for our protocols. We present runtime only for the operations that happen over MPC, computations that happen in the clear are not included. We start by presenting our results for the randomized response inspired protocols. We have chosen $q = 64$ bits, $\lambda = 16$ bits, and analyze binary and a ten classes classification problems. The results presented are averaged over 10 runs. We performed our simulations on a virtual machine - Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz with 251G of RAM.

	Time for 50k entries	Time for 100k entries
RR 2 Classes	1.77s	2.2s
RR 10 Classes	2.48s	4.2s

Table 1: Runtime for Randomized Response based Protocols

We now present results for our label private DP-SGD protocols. We have chosen $q = 64$ bits and batch size = 128. Results are averaged over 5 runs.

	Time for 50k entries and $t = 128$	Time for 100k entries and $t = 128$	Time for 50k entries and $t = 512$	Time for 100k entries and $t = 512$
Label DP-SGD 2 Classes	4.02s	7.72s	55.5s	1min 45s
Label DP-SGD 10 Classes	1min 32s	2min 40s	4min 8 s	8 min 25s

Table 2: Runtime for Label DP-SGD Protocols

We can see that the runtime performance of the randomized response-based mechanisms is better than that of Label DP-SGD. This result is expected since, for the parameters used in this experiments, we have a much higher number of private multiplications in Label DP-SGD. For ten classes, Label DP-SGD computes about $1.3 * 10^6$ secure multiplications per mini batch (128 data set entries), while the randomized response correspondent runs about $3k$ private multiplications per 128 data set entries. Additionally, the randomized response protocols are run only once, independent of the number of epochs used in the subsequent deep neural network training. For the sake of comparison, we trained a neural network with one convolutional layer (kernel size =5), three fully connected layers (with 256, 256, and 128 neurons respectively) and one output layer (1 neuron) using the traditional approach (secret share all the inputs and train the model using MPC - no release of DP information). The time required for MPC operations in the traditional approach was 38.4s per 100 data set entries. Using Label DP-SGD for (2 classes and $t = 128$) the runtime for the required MPC operations was 8ms per 100 data set entries.

Accuracy Estimation Experimental Setup The goal of accuracy analysis is to understand the accuracy deterioration due to DP noise for Protocol 2 - the label DP-SGD. The corresponding analysis for the randomized response-based protocol is exactly the same as presented in [8]. We use CIFAR-10[3] image dataset for model training and accuracy analysis. We apply ResNet18 [16] to the model

[3] https://www.cs.toronto.edu/ kriz/cifar.html

training. The ResNet18 is known to be one of the state of the art models on the CIFAR-10 dataset. The fully connected layer connecting to the output layer has 500 neurons. We train the model with 100 epochs, fixed learning rate 0.005, momentum 0.9 and batch size 128. The epsilon for Protocol 2 is calculated by the Moment Accountant [4], which is a function of noise multiplier, delta and number of epochs. Particularly, the epsilon grows as we increase the number of epochs. We set delta as 1e-5. Notice that the accuracy is also a function of the number of epochs. Instead of reporting the accuracy result for epoch=100, we choose the optimal epoch number to balance the epsilon and accuracy.

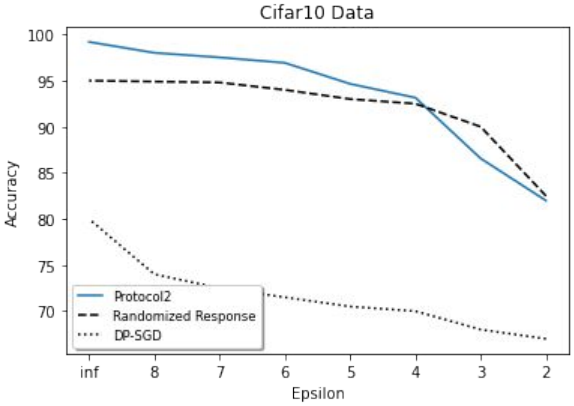

Figure 1: Accuracy results for the Randomized Response (RR) Protocol, Label DP-SGD (Protocol 2) and DP-SGD for the CIFAR10 data set (10 Classes). Results for the RR protocol are from [8]. Results for DP-SGD are from [4]

Observations for Accuracy Estimation The blue curve in Figure 1 depicts how accuracy changes as the epsilon decreases. The baseline accuracy (best accuracy for the model without any noise) is 99.2% under our experiment setup. The accuracy drop is less than 4.2% when epsilon is greater than or equal to 4. The accuracy falls relatively sharply when epsilon slides from 4 to 2. To better understand how our label privacy mechanism performs relative to existing methods, we add randomized response and DP-SGD lines in Figure 1. The numbers are from [8]. It is worth noting that the baselines (when epsilon is infinity) are different due to model training differences and care should be taken when comparing the mechanisms. However, the learnings are still insightful. Specifically, we observe that our Protocol 2is much better suited for label privacy compared to DP-SGD. The DP-SGD shows a sharper drop when epsilon goes from infinity to 8. Moreover, the accuracy degradation from randomized response looks flatter than our Protocol 2 when epsilon is above 4. When epsilon is less than 4, the decline slopes are similar.