**SCIENCE CHINA**
Information Sciences

• **Supplementary File** •

# ACCEL: An Efficient and Privacy-Preserving Federated Logistic Regression Scheme over Vertically Partitioned Data

Jiaqi Zhao[1], Hui Zhu[1*], Fengwei Wang[1], Rongxing Lu[2], Hui Li[1], Zhongmin Zhou[3] & Haitao Wan[3]

[1]*State Key Laboratory of Integrated Networks Services, Xidian University, Xi'an 710071, China;*
[2]*Faculty of Computer Science, University of New Brunswick, Fredericton E3B 5A3, Canada;*
[3]*China Mobile (Suzhou) Software Technology Co., Ltd, Suzhou 215153, China*

## Appendix A Preliminaries

This section briefly reviews logistic regression, vertical federated learning, and the SHE technique, which serve as the fundamental building blocks of ACCEL. Moreover, the related notations of ACCEL are shown in Table A1.

**Table A1** Notations of ACCEL

| Notations | Definition |
|---|---|
| $k_0, k_1, k_2$ | Security parameters. |
| $PP, SK$ | Public parameter and secret key of SHE. |
| $\oplus, \odot$ | Homomorphic addition and multiplication. |
| $K$ | The number of participants. |
| $\theta^{(r)}$ | Global model of $r$th training round. |
| $D, N$ | The total number of data features and data samples. |
| $\alpha, \lambda$ | Learning rate and regularization parameter. |
| $\kappa$ | Accuracy parameter. |
| $\mathcal{D}^{(k)}$ | Local training data of $P_k$. |
| $X_n^{(k)}$ | The $n$th data sample of $P_k$. |
| $x_{n,d}^{(k)}$ | The $d$th feature value of $X_n^{(k)}$. |
| $y_n$ | The label of $n$th data sample. |
| $D_k$ | The number of $P_k$'s data features. |
| $\mathcal{V}^{(k)}, \mathcal{M}^{(k)}$ | Data aggregation subvector and submatrix of $P_k$. |
| $\mathcal{M}$ | Data aggregation matrix. |
| $\mathcal{M}^{(u,v)}$ | Data aggregation submatrix of $P_u$ and $P_v$. |
| $[\![x]\!]$ | Ciphertext of $x$. (x can be an interger, vector, or matrix.) |

### Appendix A.1 Logistic Regression

Logistic regression [1] is one of the most widely applied machine learning models in practice, which gains huge success in many fields such as spam classification, disease diagnosis, malicious traffic detection, and so on.

Logistic regression focus on describing the relationship between multiple continuous predictor variables and a classification result. Specifically, given the logistic regression model parameter $\theta = (\theta_0, \theta_1, \cdots, \theta_D)$ and a data sample $X = (x_1, x_2, \cdots, x_D)$, the output $y$ can be obtained by computing

* Corresponding author (email: zhuhui@xidian.edu.cn)

$$y = \frac{1}{1 + e^{-h(\theta, X)}},$$

where $h(\theta, X) = \theta_0 + \sum_{d=1}^{D} \theta_d x_d$, and $y$ is a classification probability in $[0, 1]$.

For training a logistic regression model, gradient descent [2] algorithm is used in our scheme. Specifically, given training data $\mathcal{D} = \{X_n, y_n\}_{n=1}^{N}$ and $X_n = (x_{n,1}, x_{n,2}, \cdots, x_{n,D})$, $\theta$ is updated iteratively to minimize the loss between the predicted value and the true value. At first, the loss function $L(\theta)$ can be computed as

$$L(\theta) = \frac{1}{N} \sum_{n=1}^{N} log(1 + e^{-y_n h(\theta, X^n)}) + \lambda \left\| \theta \right\|^2,$$

where $\lambda$ is the $l_2$-norm regularization term. For facilitating the construction of data aggregation matrix, $L(\theta)$ can be approximated by 2-order Taylor series expansion as

$$L(\theta) = \frac{1}{N} \sum_{n=1}^{N} log2 - \frac{1}{2} y_n h(\theta, X_n) + \frac{1}{8} h^2(\theta, X_n) + \lambda \left\| \theta \right\|^2.$$

Therefore, $\theta$ is updated as

$$\begin{cases} \theta_0^{(r+1)} = (1 - 2\lambda\alpha)\theta_0^{(r)} - \dfrac{\alpha}{N} \displaystyle\sum_{n=1}^{N} (\frac{1}{4} h(\theta^{(r)}, X_n) - \frac{1}{2} y_n) \\ \theta_d^{(r+1)} = (1 - 2\lambda\alpha)\theta_d^{(r)} - \dfrac{\alpha}{N} \displaystyle\sum_{n=1}^{N} (\frac{1}{4} h(\theta^{(r)}, X_n) - \frac{1}{2} y_n) x_{n,d} \end{cases},$$

where $d = 1, \cdots, D$, $\alpha$ is the learning rate, and $r$ is current training round.

Finally, the above training process terminates until $\theta$ converges or the maximum iteration number is reached.

## Appendix A.2 The SHE Technique

The SHE technique [3] can compute homomorphic addition and multiplication efficiently, which mainly contains three functions named KeyGen, Enc, and Dec, which are described detailedly in the following.

- KeyGen$(k_0, k_1, k_2) \rightarrow PP, SK$: Given the security parameters $(k_0, k_1, k_2)$ satisfying $k_1 \ll k_2 < k_0$, select two large prime numbers $p, q$ with $|p| = |q| = k_0$ and calculate $\mathcal{N} = pq$. Then, select a random number $\mathcal{L}$ with $|\mathcal{L}| = k_2$. Finally, the secret key $SK$ is $(p, q, \mathcal{L})$ and the public parameter $PP$ is $(k_0, k_1, k_2, \mathcal{N})$.
- Enc$(m) \rightarrow c$: Given a message $m \in (-2^{k_1}, 2^{k_1})$, the ciphertext $c$ is calculated with $SK$ as

$$c = \text{Enc}(m) = (r\mathcal{L} + m)(1 + r'p) \bmod \mathcal{N},$$

where $r \in (0, 2^{k_2})$ and $r' \in (0, 2^{k_0})$ are two random numbers.
- Dec$(c) \rightarrow m$: Given a ciphertext $c$, the corresponding plaintext $m$ can be retrieved with $SK$ through computing

$$\begin{aligned} m' &= (c \bmod p) \bmod \mathcal{L}, \\ m &= \text{Dec}(c) = \begin{cases} m', (m' < \dfrac{\mathcal{L}}{2}) \\ m' - \mathcal{L}, (else) \end{cases}. \end{aligned}$$

Sepecifically, given two ciphertexts $c_1 = \text{Enc}(m_1)$, $c_2 = \text{Enc}(m_2)$, and a plaintext $m_3$, the SHE technique has the following four homomorphic properties:

- Ciphertext homomorphic addition: $c_1 \oplus c_2 = c_1 + c_2 \bmod \mathcal{N} = \text{Enc}(m_1 + m_2)$,
- Plaintext-ciphertext homomorphic addition: $c_1 \oplus m_3 = c_1 + m_3 \bmod \mathcal{N} = \text{Enc}(m_1 + m_3)$,
- Ciphertext homomorphic multiplication: $c_1 \odot c_2 = c_1 \cdot c_2 \bmod \mathcal{N} = \text{Enc}(m_1 \cdot m_2)$,
- Plaintext-ciphertext homomorphic multiplication: $c_1 \odot m_3 = c_1 \cdot m_3 \bmod \mathcal{N} = \text{Enc}(m_1 \cdot m_3), (m_3 > 0)$,

where $\oplus$ and $\odot$ represent the homomorphic addition and multiplication.

**References**


1 Hosmer Jr D W, Lemeshow S, Sturdivant R X. Applied logistic regression. John Wiley & Sons, 2013
2 Yang Q, Liu Y, Cheng Y, et al. Federated learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2019, 13(3): 1-207
3 Mahdikhani H, Lu R X, Zheng Y D, et al. Achieving o(log3n) communication-efficient privacy-preserving range query in fog-based iot. IEEE Internet Things J., 2020, 7(6): 5220-5232