A Federated Algorithm for Privacy-Preserving Empirical Risk Minimization

Zonghao Huang
zonghao.huang@okstate.edu
School of Electrical and Computer Engineering
Oklahoma State University

Editor: Zonghao Huang

1. Problem Statement

We consider a group of data owners $[m] = 1, 2, \ldots, m$, who collaboratively learn a model over their private datasets. We use $D_j = \{(x_i^{(j)}, y_i^{(j)}), i = 1, 2, \ldots, n_j\}$ to denote the dataset possessed by data owner j, where $x_i^{(j)}$ is the data feature vector and $y_i^{(j)}$ is the corresponding data label.

The goal of our problem is to learn a model over the aggregated dataset $\{D_j\}_{j\in[m]}$, which is formulated as a regularized empirical risk minimization:

$$\min_{\theta} J(\theta) = \min_{\theta} \frac{1}{m} \sum_{j=1}^{m} J_j(\theta) = \min_{\theta} \frac{1}{m} \sum_{j=1}^{m} \frac{1}{n_j} \sum_{i=1}^{n_j} \ell(\theta, x_i^{(j)}, y_i^{(j)}) + \lambda N(\theta). \tag{1}$$

The data owners will solve the above minimization problem via the gradient-based method with calibrated noise under Secure Multi-Party Computation (MPC).

2. Our Algorithm

Following the previous work (Jayaraman et al., 2018), our algorithm is based on the gradient-based iterative learning with calibrated noise under MPC. Instead of just computing the gradient, our algorithm allows the data owners to update their local model with calibrated noise, and aggregates the local models in every K iterations, which is similar to the federated learning framework proposed by (McMahan et al., 2016). The details of our algorithm are shown in Algorithm 1.

In our algorithm, we define that $\theta^t = \frac{1}{m} \sum_{j=1}^{m} \theta_j^t$, which is inaccessible by each worker when $t \notin \{nK, n = 1, 2, \ldots\}$.

Theorem 1 (Privacy Guarantee) *Given a estimator θ^T obtained by Algorithm 1 involving a class of T gradient descent-based updates perturbed with noise $z \in N(0, \sigma^2 \mathrm{I})$. Assumed $\ell(\theta)$ is G-Lipschitz over $\theta \in C$. The estimator θ^T is (ϵ, δ)-differentially private if we set:*

$$\sigma^2 = \frac{8G^2 T \log(1/\delta)}{m^2 n_{(1)}^2 \epsilon^2}. \tag{2}$$

Proof We can prove the theorem by following the proof of Theorem 3.4 in (Jayaraman et al., 2018). ■

Algorithm 1 Differentially Private Federated Algorithm Under MPC

Initialize $\{\theta_i^0\}$;
for $t = 1, 2, \ldots, T$ **do**
  for $j = 1, 2, \ldots, m$ **do**
    $\theta_j^t = \theta_j^{t-1} - \eta_t(\nabla J_j(\theta_j^{t-1}) + z)$, $z \sim N(0, \sigma^2 \mathrm{I})$;
  end for
  if $t \in \{nK, n = 1, 2, \ldots\}$ **then**
    $\theta^t = \frac{1}{m}\sum_{j=1}^m \theta_j^t$;
    for $j = 1, 2, \ldots, m$ **do**
      $\theta_j^t = \theta^t$;
    end for
  end if
end for

Note: When K is set to be 1, our approach (Algorithm 1) is equivalent to the algorithm proposed in (Jayaraman et al., 2018). Compared with their work, our algorithm has less communication cost and less computation cost since workers only communicate and perform MPC when $t \in \{nK, n = 1, 2, \ldots\}$.

3. Theoretical Result

In this section, we will give the utility analysis of our algorithm by the excess empirical risk.

Lemma 2 *Assumed that $J_j(\theta)$ is λ-strongly convex, and L-smooth over $\theta \in C$. If $\eta_t \leq \frac{1}{4L}$ for all t, we have:*

$$\mathbb{E}[\|\theta^{t+1} - \theta^*\|^2] \leq (1 - \eta^t\lambda)\mathbb{E}[\|\theta^t - \theta^*\|^2] + 6L\eta_t^2\Gamma + 2\mathbb{E}[\frac{1}{m}\sum_{j=1}^m \|\theta^t - \theta_j^t\|^2] + \eta_t^2 d\sigma^2, \quad (3)$$

where $\Gamma = J^* - \frac{1}{m}\sum_{j=1}^m J_j^*$. J^* is the minimal value of $J(\theta)$ while J_j^* is the minimal value of $J_j(\theta)$. Γ quantifies the heterogeneity degree of the data distribution.

Proof The proof follows the proof of Lemma 1 in (Li et al., 2019). ■

Lemma 3 (Lemma 3 in (Li et al., 2019)) *Assumed that*

$$\mathbb{E}[\|\nabla J_j(\theta)\|^2] \leq H^2. \quad (4)$$

If η_t is non-increasing and $\eta_t \leq 2\eta_{t+K}$ for all t, we have:

$$\mathbb{E}[\frac{1}{m}\sum_{j=1}^m \|\theta^t - \theta_j^t\|^2] \leq 4\eta_t^2(K-1)^2H^2. \quad (5)$$

Theorem 4 (Utility Analysis) *Given the estimator θ^T obtained by Algorithm 1 where $J_j(\theta)$ is G-Lipschitz, λ-strongly convex, and L-smooth over $\theta \in C$. If the learning rate η_t is non-increasing, $\eta_t \leq \frac{1}{4L}$, and $\eta_t \leq 2\eta_{t+K}$ for all t, and the gradients are perturbed with noise $z \in N(0, \sigma^2 \mathrm{I})$ with σ^2 defined by Eq. (2), we have the following excess empirical risk:*

$$\mathbb{E}[J(\theta^T)] - J(\theta^*) \leq \frac{2L}{\lambda(\gamma + T)}(\frac{B}{\lambda} + 2L\|\theta^0 - \theta^*\|^2), \tag{6}$$

where $B = 6L\Gamma + 8(K-1)^2G^2 + d\frac{8G^2T\log(1/\delta)}{m^2n_{(1)}^2\epsilon^2}$.

Proof We follow (Li et al., 2019) to give the proof. Since $J_j(\theta)$ is $G - Lipschitz$, thus $\mathbb{E}[\|\nabla J_j(\theta)\|^2] \leq G^2$. Based on Lemma 2 and Lemma 3, we have:

$$\mathbb{E}[\|\theta^{t+1} - \theta^*\|^2] \leq (1 - \eta^t\lambda)\mathbb{E}[\|\theta^t - \theta^*\|^2] + \eta_t^2 B, \tag{7}$$

where $B = 6L\Gamma + 8(K-1)^2G^2 + d\frac{8G^2T\log(1/\delta)}{m^2n_{(1)}^2\epsilon^2}$.

By setting $\eta_t = \frac{2}{\lambda(t+\gamma)}$ where $\gamma = \max\{\frac{8L}{\lambda} - 1, K\}$ to enforce that $\eta_t \leq \frac{1}{4L}$ and $\eta_t \leq 2\eta_{t+E}$, we can prove that $\mathbb{E}[\|\theta^t - \theta^*\|^2] \leq \frac{v}{\lambda + t}$ where $v = \max\{\frac{4B}{\lambda^2}, (\gamma+1)\|\theta^0 - \theta^*\|^2\}$. By the smoothness of $J(\theta)$, we have:

$$\mathbb{E}[J(\theta^T)] - J(\theta^*) \leq \frac{L}{2}\|\theta^T - \theta^*\|^2 \leq \frac{Lv}{2(\lambda + T)} \leq \frac{2L}{\lambda(\gamma + T)}(\frac{B}{\lambda} + 2L\|\theta^0 - \theta^*\|^2). \tag{8}$$

■

4. Discussion

As shown in Theorem 4, the choice of large K will introduce the error by $\frac{16L(K-1)^2G^2}{\lambda^2(\gamma+T)}$, but reduce the communication cost by a factor of $\frac{1}{K}$. What is the optimal choice of K to balance the utility and the communication cost?

It is still a problem whether the federated algorithm with differential privacy could achieve the state-of-art excess empirical risk bound.

References

Bargav Jayaraman, Lingxiao Wang, David Evans, and Quanquan Gu. Distributed learning without distress: Privacy-preserving empirical risk minimization. In *Advances in Neural Information Processing Systems*, pages 6343–6354, 2018.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629*, 2016.