Privacy preserving / Genaralization bound

Instructed by Dr.Yassaee

AmirHossein Naghdi Ali Sadeghian MohammadParsa Dini

Sharif University of Technology

July 7, 2024

Contents

Federated Learning

1. Background
2. Definition
3. System Components
4. Privacy Mechanisms

Background and Definition

Background

1. Data Limitations in Size and Approximating Real Distributions
2. Data Sharing Constraints within the Organization Due to Policies and Regulations

Definition

In a federated learning system, multiple parties collaboratively train machine learning models without directly exchanging their raw data. Instead, each party trains its model locally using its own data. The output of the system is a customized machine learning model for each participating party

System Components

Parties

In FLSs, the parties are the data owners and the beneficiaries of FL. They can be organizations or mobile devices.

Manager

the manager is usually a powerful central server. It conducts the training of the global machine learning model and manages the communication between the parties and the server.

System Components

Communication-Computation Framework

In FLSs, the computation happens on the parties and the manager, while the communication happens between the parties and the manager. Usually, the aim of the computation is for the model training and the aim of the communication is for exchanging the model parameters.

1. Federated Averaging (FedAvg)
2. SimFL (Decentralized Framework)

System Components

Federated Averaging (FedAvg)

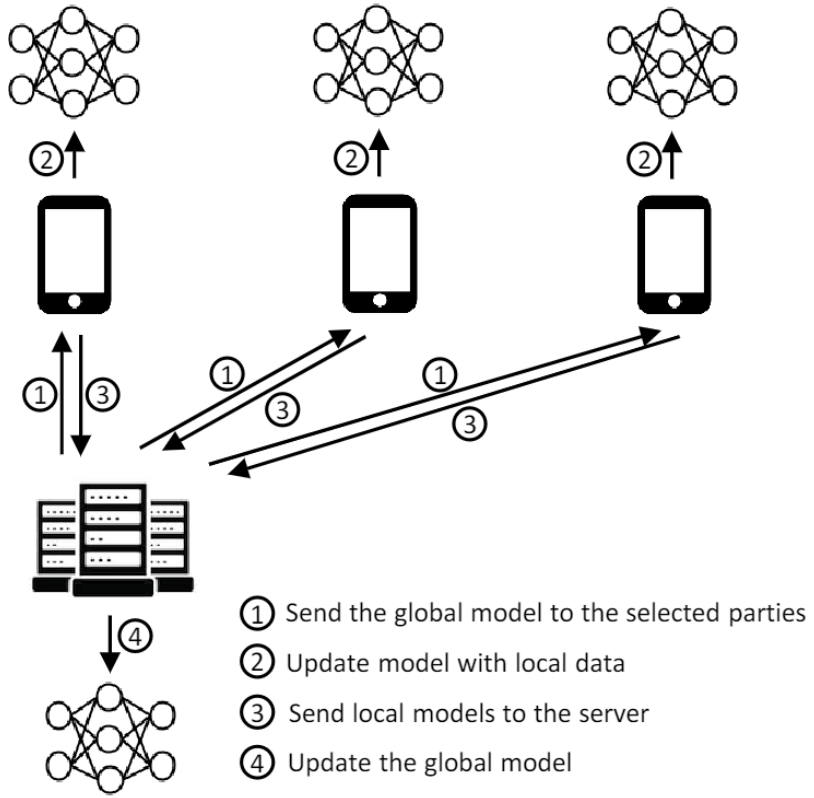
- Proposed in 2016, FedAvg is a widely used framework for federated learning.
- In each iteration:
  - The server sends the current global model to selected parties.
  - Selected parties update the global model with their local data.
  - Updated models are sent back to the server.
  - The server averages received local models to create a new global model.
- FedAvg repeats this process until reaching the specified number of iterations, with the server's global model as the final output.

System Components

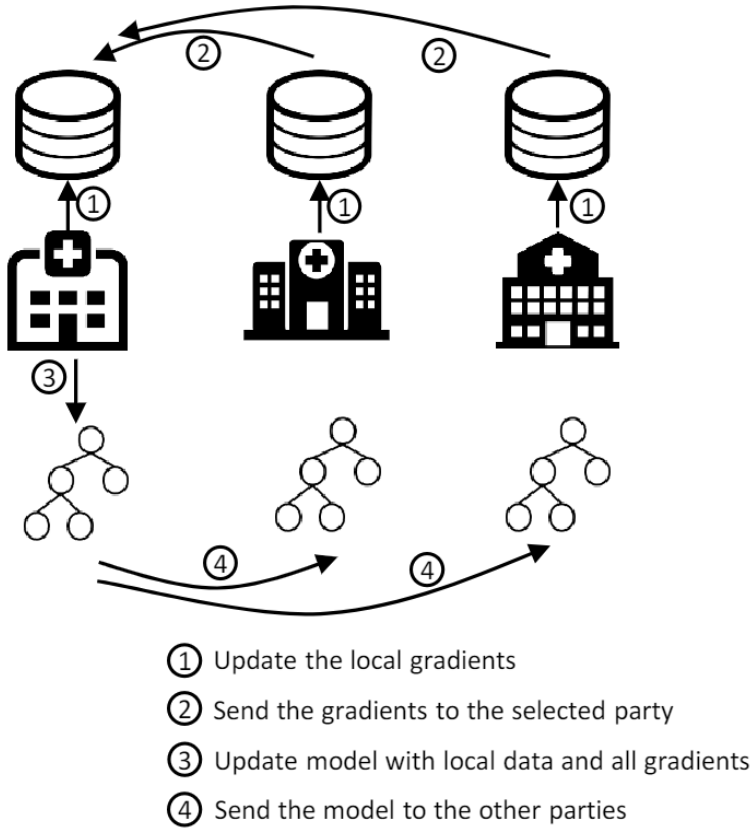
SimFL (Decentralized Framework)

- SimFL, proposed by Li et al., represents a decentralized federated learning framework.
- No trusted server is needed.
- In each iteration:
  - Parties update gradients using their local data.
  - Gradients are sent to a selected party.
  - The selected party uses its local data and gradients to update the model.
  - The updated model is sent to all other parties.
- To ensure fairness, every party updates the model for a similar number of rounds.
- SimFL repeats a specified number of iterations and produces the final model.

System Components



(a) FedAvg



(b) SimFL

Privacy Mechanismss

Cryptographic Techniques

Homomorphic Encryption:

- **Definition:** Homomorphic encryption allows computation on encrypted data without decrypting it first.
- **How It Works:**
  - Parties encrypt their messages before sending them.
  - Computation is performed on the encrypted messages.
  - The final result is obtained by decrypting the encrypted output.
- **Privacy Protection:**
  - User privacy is well protected.
  - Secure aggregation of transferred gradients.
- **Limitations:**
  - No privacy guarantees for the final model.
  - Vulnerable to inference attacks and model inversion.
  - High computation overhead due to encryption and decryption.

Privacy Mechanismss

Differential Privacy

- **Definition:**
  - Differential privacy quantifies and limits the privacy loss of an individual when their private data enters a dataset.
  - It adds random noise to data or model parameters.
- **Privacy Guarantees:**
  - **Individual Records:**
    - Differential privacy ensures that one record's contribution doesn't significantly impact the model output.
    - Protects against inference attacks.
  - **Aggregated Models:**
    - Noise prevents exact reconstruction of individual records.
    - Statistical privacy guarantees for the entire model.
- **Trade-Off:**
  - **Accuracy vs. Privacy:**
    - Noise affects model accuracy.
    - Striking a balance is crucial.

The setting and aim of the paper

Here the aim is to extract Differential Privacy parameters (ϵ, δ) in a new setting. In our federated learning model, *n* distributed users send their updates of a shared model to a trusted aggregator. At each iteration, *m* number of users are chosen uniformly without replacement.
We also make two assumptions:

- users communicate over encrypted channels with a trusted aggregator.
- the aggregator releases the model parameters only after a certain number of iterations and hide all intermediate updates.

However, approximate DP was obtained in order to avoid the Exact DP, since we would have to deal with Renyi divergence which is not convex jointly on (ν, μ).

Preliminaries

Remark

Given a convex function $f : [0; 1) \to R$ with $f(1) = 0$, the f-divergence between two probability measures ν and μ is defined as:

$$D_f(\mu||\nu) := \mathbf{E}_\nu[f(\frac{d\mu}{d\nu})]$$

Preliminaries

Remark

E_ϵ-divergence is the f-divergence associated with $f(t) = (t - e^\epsilon)_+ \max\{0, t - e^\epsilon\}$:

$$\mathbf{E}_\epsilon(\mu||\nu) = \sup_{A \subseteq Y} [\mu(A) - e^\epsilon \nu(A)] = \frac{1}{2} \int |\mu(A) - e^\epsilon \nu(A)| - \frac{1}{2}(e^{d(A)} - 1) =$$

$$\mu(\log \frac{d\mu}{d\nu} > \epsilon) - e^\epsilon \nu(\log \frac{d\mu}{d\nu} > \epsilon)$$

Preliminaries

Lemma 1

For $m_1, m_2 \in R^d$ and $\sigma > 0$, let $\mathcal{N}_1$ and $\mathcal{N}_2$ denote $\mathcal{N}(m_1, \sigma^2 I)$ and $\mathcal{N}(m_2, \sigma^2 I)$, respectively. Then we have:

$$E_\epsilon(\mathcal{N}_1 || \mathcal{N}_2) = Q(\frac{\epsilon}{k} - \frac{k}{2}) - e^\epsilon Q(\frac{\epsilon}{k} + \frac{k}{2})$$

where $k = \frac{||m_2 - m_1||_2}{\sigma}$ and $Q(t) = \int_t^\infty e^{\frac{-u^2}{2}} du$.

we will be needing this quite a lot later!

Remark

Furthermore, let us define $\theta_\epsilon(r)$ as:

$$\theta_\epsilon(r) := E_\epsilon(\mathcal{N}(r, \sigma^2 I) || \mathcal{N}(0, \sigma^2 I)) = Q(\frac{\epsilon}{r} - \frac{r}{2}) - e^\epsilon Q(\frac{\epsilon}{r} + \frac{r}{2})$$

Preliminaries

remark

Properties of E_ϵ:

- $0 \leq E_\epsilon(\mu||\nu) \leq TV(\mu,\nu)$ for any $\epsilon > 0$. The upper bound is equality if and only if $\epsilon = 0$,
- $\epsilon \to E_\epsilon(\mu||\nu)$ is continuous and strictly decreasing on $[0, TV(\mu,\nu)]$,
- $E_\epsilon(\mu||\nu)$ decreases by post-processing (data processing inequality),
- $(\mu,\nu) \to E_\epsilon(\mu||\nu)$ is convex

Preliminaries

remark

The contraction coefficient: $1 \geq \eta_f(K) := \sup_{\mu,\nu: D_f(\mu||\nu)\neq 0} \frac{D_f(\mu K||\nu K)}{D_f(\mu||\nu)}$

Dobrushin in his paper resulted that:

remark

The contraction coefficient: $\eta_{TV}(K) = sup_{y_1,y_2\in Y} TV(K(y_1), K(y_2))$.

Preliminaries

Since TV can be generalized to E_ϵ, we can deduce that:

remark

$1 \geq \eta_f(K) := \sup_{\mu,\nu: E_\epsilon(\mu||\nu)\neq 0} \frac{E_\epsilon(\mu K||\nu K)}{e_\epsilon(\mu||\nu)}$

The following theorem establishes a two-point characterization for η_ϵ similar to the Dobrushin's characterization in.

remark

The generalized contraction coefficient:
$\eta_\epsilon(K) = sup_{y_1,y_2\in Y} E_\epsilon(K(y_1), K(y_2))$.

Preliminaries

Lemma 2

Let $Y \subset R^d$ be a bounded set. For the Markov kernel specified by $K(y) = \mathcal{N}(y, \sigma^2 I)$ for $y \in Y$ and $\sigma > 0$, we have $\eta_\epsilon(K) = \theta_\epsilon(\frac{||Y||}{\sigma})$ where: where $||Y|| := \max_{y_1, y_2 \in Y} ||y_1 - y_2||_2$.

Setting of the problem

In our federated learning model, n distributed users send their updates of a shared model to a trusted aggregator. At each iteration, m number of users are chosen uniformly without replacement.
we assume $m = qn$ and since the subsampling is performed without replacement, the total number of iteration is $T = \frac{n}{m} = \frac{1}{q}$.

Batch size of size-1

- Let $\pi \in S_n$ be a random permutation map and S_n is the symmetric group on $[n]$. The federated learning algorithm iterates as follows:
- The aggregator samples the initial parameter $W_0 \sim \mu_0$ in $ball(\rho)$, the ball of radius ρ in R^d , according to a distribution μ_0 and sends it to user $\pi(1)$.
- User $\pi(1)$ uses W_0 and her local data $x_{\pi(1)}$ to compute the update $W_1 := \eta\nabla(W_0, x_{\pi(1)}) + \eta\sigma Z_1$, where $Z_1 \sim \mathcal{N}(0, I)$. This update is then sent back to the aggregator.
- Upon receipt of W_1, the aggregator computes $W_1 = proj_\rho(W_0 - W_1)$, where $proj_\rho()$ denotes the projection operator onto $ball(\rho)$. Then W_1 is sent to user $\pi(2)$.
- Continue the above procedure until all n users send the aggregator their updates (i.e., $T = n$ is the number of iterations). The aggregator releases W_T .

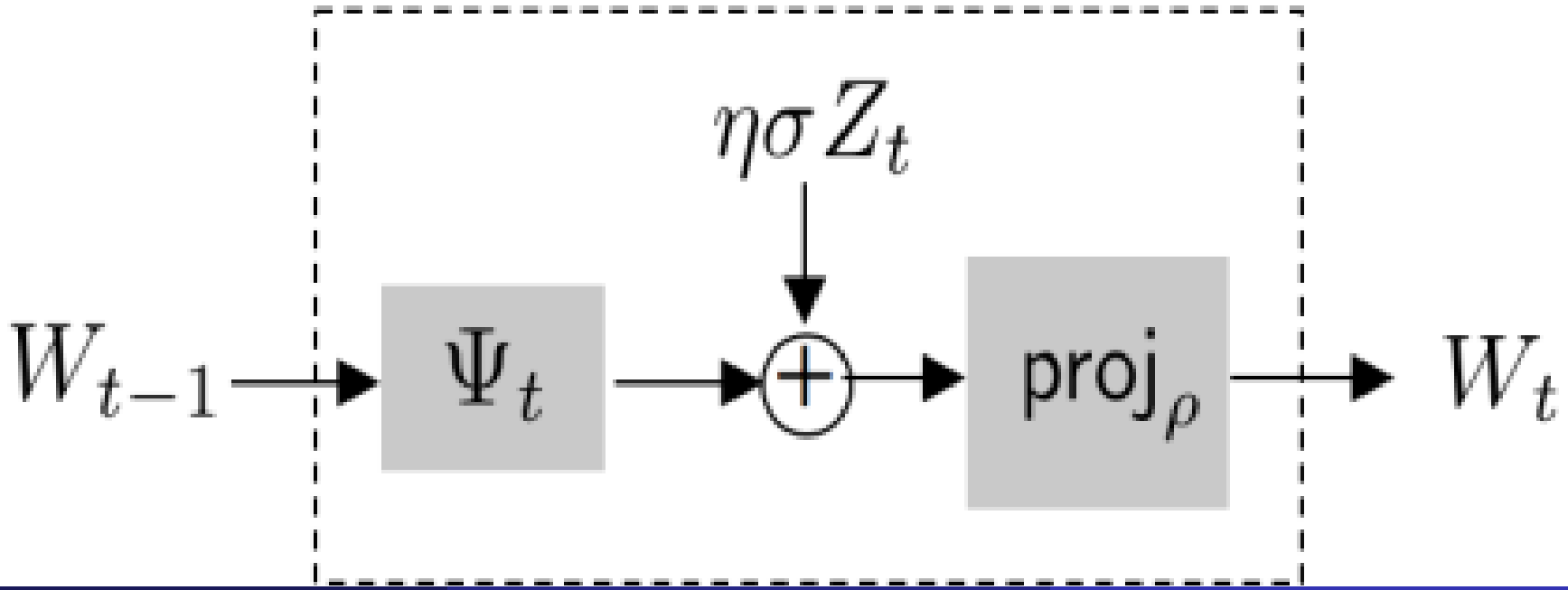
Batch size of size-1

- the update function:

$$\Psi_t(w) = w - \eta \nabla l(w, x_{\pi(t)})$$

- the projected Gaussian Markov kernel:

$$K_t(w) = proj_\rho(\mathcal{N}(\Psi_t(w), \eta^2 \sigma^2 I)).$$



Now consider a pair of neighboring datasets x and x' that differ in the i-th entry. We have $x_i \neq x'_i$ and $x_j = x'_j$ for $j \in [n]\{i\}$) and let μ_t and μ'_t be the distributions of the W_t when algorithm runs on x and x'.
Now using the DPI inequality it implies that:
$E_\epsilon(\mu_T||\mu'_T) \leq E_\epsilon(\mu_{T-1}||\mu'_{T-1})\eta_\epsilon(K_T) \leq E_\epsilon(\mu_{T-2}||\mu'_{T-2})\eta_\epsilon(K_T)\eta_\epsilon(K_{T-1})$
Applying this inequality T times will lead to:

$$E_\epsilon(\mu_T||\mu'_T) \leq E_\epsilon(\mu_t||\mu'_t)\Pi_{j=t+1}^{T}\eta_\epsilon(K_j)$$

Batch size of size-m

Batch size of size-m

- the update function:

$$\Psi_t(w) = w - \frac{\eta}{m}\nabla l(w, x_{\pi(t)})$$

- the projected Gaussian Markov kernel:

$$K_t(w) = proj_\rho(\mathcal{N}(\Psi_t(w), \eta^2\sigma^2 I)).$$

The Theorem

Theorem 1

Let the loss function $w \to (w, x)$ be convex, L-Lipschitz and *betta*-smooth for all $x \in X$ and also $\eta \leq 2/$. Then Algorithm 1 is (ϵ, δ)-DP for $\epsilon \geq 0$ and

$$\delta = \frac{m}{n}\theta_\epsilon\left(\frac{2L}{\sqrt{m}\sigma}\right)\frac{\left(1 - \theta_\epsilon\left(\frac{2\rho\sqrt{m}}{\eta\sigma}\right)\right)^{\frac{m}{n}}}{1 - \theta_\epsilon\left(\frac{2\rho\sqrt{m}}{\eta\sigma}\right)}$$

Proof

The proof begins using DPI and using Kernels

$$E_\epsilon(\mu_T||\mu'_T) \leq \sum_{t=1}^{T} Pr(\pi(i)=t) E_\epsilon(\mu_t||\mu'_t) Pi_{j=t+1}^{T} \eta_\epsilon(K_j) =$$

q $\sum_{t=1}^{T} E_\epsilon(\mu_t||\mu'_t) \Pi_{j=t+1}^{T} \eta_\epsilon(K_j)$

Proof

It implies that:

$$\eta_\epsilon(K_j) =$$

$\sup_{w_1,w_2 \in \text{ball}(\rho)} E_\epsilon(K_j(w_1)||K_j(w_2)) \leq$

$\sup_{w_1,w_2 \in \text{ball}(\rho)} E_\epsilon(\mathcal{N}(\Psi_j(w_1), \hat{\sigma}^2 I)||\mathcal{N}(\Psi_j(w_2), \hat{\sigma}^2 I)) =$

$\theta_\epsilon(\frac{\Psi_j(ball())}{\hat{\sigma}}) \leq$

$\theta_\epsilon(\frac{2\rho}{\sigma}) = \theta_\epsilon(\frac{2\rho}{\sqrt{m}\eta\sigma})$

Proof

Lets review our update function for these datasets:

- $\Psi_t(w) := w - \frac{\eta}{m}\nabla l(w, x_i) + \sum_{j \in B_t - \{i\}} \nabla l(w, x_j)$
- $\Psi'_t(w) := w - \frac{\eta}{m}\nabla l(w, x'_i) + \sum_{j \in B_t - \{i\}} \nabla l(w, x_j)$

Now the convexity of E_ϵ implies that:

$$E_\epsilon(\mu_t || \mu'_t) \leq \int E_\epsilon(K_t(y) || K'_t(y)) \mu_{t-1}(dy) \leq$$

$\int E_\epsilon(\mathcal{N}(\Psi_t(y), \hat{\sigma}^2 I) || \mathcal{N}(\Psi_t(y), \hat{\sigma}^2 I)) \mu_{t-1}(dy) =$
$\int \theta_\epsilon \left(\frac{||\Psi_t(y) - \Psi'_t(y)||_2}{\hat{\sigma}} \right) \mu_{t-1}(dy) \leq$
$\theta_\epsilon(\frac{2L\eta}{m\hat{\sigma}}) \leq \theta_\epsilon(\frac{2L}{\sqrt{m}\sigma})$

Proof

Now from L-liptschitzness we get:

$$||\Psi_t(y) - \Psi'_t(y)||_2 = ||\frac{\eta}{m}(\nabla l(y, x_i) - \nabla l(y, x'_i))||_2 \leq \frac{2L\eta}{m}$$

Now we can easily see that:

$$E_\epsilon(\mu_T||\mu'_T) \leq q\theta_\epsilon\left(\frac{2L\sqrt{m}}{\sigma}\right)\sum_{t=1}^{T}\theta_\epsilon\left(\frac{2\rho\sqrt{m}}{\eta\sigma}\right) =$$

$$\theta_\epsilon\left(\frac{2L}{\sqrt{m}\sigma}\right)\frac{(1-\theta_\epsilon(\frac{2\rho\sqrt{m}}{\eta\sigma}))^{\frac{m}{n}}}{1-\theta_\epsilon(\frac{2\rho\sqrt{m}}{\eta\sigma})}\frac{m}{n}$$

Generalization Error

Generalization Error

Let $S^{(i)} = (Z_1, \ldots, Z'_i, \ldots, Z_n)$ be a version of S with Z_i replaced by an i.i.d. copy Z'_i. Denote $S' = (Z'_1, \ldots, Z'_n)$. Then

$$\mathbb{E}_{S\sim\pi^n}[\Delta_A(S)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{S,S'}[\ell(A(S), Z'_i) - \ell(A(S^{(i)}), Z'_i)].$$

Assumption1

Assumption1

The loss function $\ell(\tilde{W}, \tilde{Z})$ satisfies

$$\log \mathbb{E}\left[\exp\left(\lambda\left(\ell(\tilde{W}, \tilde{Z}) - \mathbb{E}[\ell(\tilde{W}, \tilde{Z})]\right)\right)\right] \leq \psi(-\lambda)$$

for $\psi(\lambda) = \frac{R^2\lambda^2}{2}, \lambda \in (b, 0], \psi(0) = \psi'(0) = 0$, where $\tilde{W}, \tilde{Z}$ are taken independently from the marginals for W, Z, respectively.

Assumption2

Assumption2

The loss function $\ell(\tilde{W}, \tilde{Z})$ is sub-Gaussian with parameter R^2 in the sense that

$$\log \mathbb{E}\left[\exp\left(\lambda\left(\ell(\tilde{W}, \tilde{Z}) - \mathbb{E}[\ell(\tilde{W}, \tilde{Z})]\right)\right)\right] \leq \frac{R^2\lambda^2}{2}.$$

Local Bound

Local Bound

Suppose that $\ell(\cdot, z)$ is a convex function of $w \in \mathbb{R}^d$ for each z and that A_k represents the empirical risk minimization algorithm on local dataset S_k in the sense that

$$W_k = A_k(S_k) = \operatorname{argmin}_w \sum_{i=1}^{n} \ell(w, Z_{i,k}).$$

Then

$$\Delta_A(s) \leq \frac{1}{K} \sum_{k=1}^{K} \Delta_{A_k}(s_k).$$

Mutual Information Bound

Mutual Information Bound

$$\mathbb{E}_{S\sim\pi^n}[\Delta_A(S)] \leq \frac{1}{n}\sum_{i=1}^{n}\psi^{*-1}(I(W;Z_i))$$

where

$$\psi^{*-1}(y) = \inf_{\lambda\in[0,b)}\left(\frac{y+\psi(\lambda)}{\lambda}\right).$$

Linear or Location Models with Bregman Loss

Suppose that Assumption 1 holds for each node. Consider the following two cases:

(i) $\ell(w,(x,y)) = D_F(\langle x, w\rangle, y)$, then

$$\mathbb{E}_{S\sim\pi^{nK}}[\Delta_A(S)] \leq \frac{1}{nK^2}\sum_{i,k}\psi^{*-1}(I(W_k;Z_{i,k}))$$

$$\leq \frac{1}{K^2}\sum_{k=1}^{K}\psi^{*-1}\left(\frac{I(W_k;S_k)}{n}\right).$$

(ii) $\ell(w,z) = D_F(w,z)$, then

$$\mathbb{E}_{S\sim\pi^{nK}}[\Delta_A(S)] = \frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}_{S_k\sim\pi^{n}}[\Delta_{A_k}(S_k)]$$

Linear or Location Models with Bregman Loss

then we have:

$$\mathrm{E}_{S\sim\pi^{nK}}[\Delta_A(S)] \leq \frac{1}{nK^2}\sum_{i,k}\psi^{*-1}(I(W_k;Z_{i,k})) \leq \frac{1}{K^2}\sum_{k=1}^{K}\psi^{*-1}\left(\frac{I(W_k;S_k)}{n}\right).$$

Order reduction

$Z \sim \pi = \mathcal{N}(\mu, \sigma^2 I_d)$, $\ell(w, z) = \|w - z\|_2^2$

An obvious algorithm: $w_k = A_k(s_k) = \frac{1}{n}\sum_{i=1}^{n} z_{i,k}$.

For this algorithm, it can be shown that $I(\hat{W}; Z_{i,k}) = \frac{d}{2}\log\frac{nK}{(nK-1)}$ and

$$\psi^{*-1}(y) = 2\sqrt{d\left(1+\frac{1}{nK}\right)^2\sigma^4 y}$$

$$\mathbb{E}_{S\sim\pi^{nK}}[\Delta A(S)] \leq \sigma^2 d\sqrt{2(1+\tfrac{1}{nK})^2\log\tfrac{nK}{nK-1}} \to \mathcal{O}(\tfrac{1}{\sqrt{nK}}).$$

$$\mathbb{E}_{S\sim\pi^{nK}}[\Delta A(S)] \leq \tfrac{\sigma^2 d}{K}\sqrt{2(1+\tfrac{1}{n})^2\log\tfrac{n}{n-1}} \to \mathcal{O}(\tfrac{1}{K\sqrt{n}}).$$

Privacy Constraints

Privacy Constraints

Suppose each node's algorithm A_k is an ϵ-local differentially private mechanism in the sense that

$$\frac{p(w_k|s_k)}{p(w_k|s'_k)} \leq e^{\epsilon}$$

for each w_k, s_k, s'_k. Then for losses ℓ of the form in Theorem 4, and under Assumption 2,

$$\mathbb{E}_{S\sim\pi^{nK}}[\Delta_A(S)] \leq \frac{1}{K}\sqrt{\frac{2R^2\min\{\epsilon,(e-1)\epsilon^2\}}{n}}.$$

Communication Constraints

Communication Constraints

Suppose each node can only transit B bits of information to the model aggregator, meaning that each W_k can only take 2^B distinct possible values. Then for losses ℓ of the form in Theorem 4, and under Assumption 2:

$$E_{S\sim\pi^{nK}}[\Delta_A(S)] \leq \frac{1}{K}\sqrt{\frac{2(\log 2)R^2B}{n}}$$

References

- A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection, Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, Bingsheng He
- Differentially Private Federated Learning: An Information-Theoretic Perspective, Shahab Asoodeh, Flavio P. Calmon
- Improved Information Theoretic Generalization Bounds for Distributed and Federated Learning ,L. P. Barnes, Alex Dytso, H. V. Poor, Princeton University, Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Department of Electrical and Computer Engineering