Advanced Topics in Machine Learning 2024

Nirupam Gupta

Assignment 7: Privacy and Robustness in FL

Deadline: -

The assignments must be submitted individually – each student must write and submit a personal solution but we do not prevent you from discussing high-level ideas in small groups. If you use any LLM tool such as ChatGPT, please specify the purpose and manner in which you have utilized it.

We are interested in how *you solved the problems, and not just in the final answers. Please explain your solutions and ideas as clearly as you can.*

Late Penalty and multiple Submissions *Late submissions will incur a penalty of 10% of the total marks for every hour of delay (rounded up) with a maximum allowed delay of 5 hours after which the submission server will close. If you submit multiple submissions, only the last submission will be considered relevant both for grading answers as well as late penalty.*

Submission format: *Please upload your answers in a single* `.pdf` *file. If you have created code please include at least the key parts in the PDF as well as a link to the full code (on Github, Colab, or similar).*

Learning points: *The goal of this assignment is to provide a better understanding of the concepts covered in the class. It will help you grasp additional details that were skipped during the lecture.*

1. **Privacy Amplification in the Local Model.** In the class, we looked at privacy amplification under Poisson sampling of the input data. Specifically, consider an input $Z = (z_1, \dots, z_m)$. Let $\mathcal{S}$ denote the sampling operator, i.e., $\mathcal{S}([m]) = U$ is a random subset of $[m] \triangleq \{1, \dots, m\}$ such that, $\forall i \in [m]$, $i \in U$ with probability $q \in (0, 1)$. Suppose that we query Z by applying an (ε, δ)-DP mechanism $\mathcal{M}$ on $Z^U = (z_i \ ; \ i \in U)$. Then, the overall mechanism $\mathcal{M} \circ \mathcal{S}$ is (ε', δ')-DP w.r.t. Z, where

$$\varepsilon' = \log\left(1 + q(e^{\varepsilon} - 1)\right) \ , \text{ and } \ \delta' = q\delta. \tag{1}$$

Question 1 (Tightness of the above privacy amplification result). We demonstrate the tightness of the above result by considering a specific binary case where for all $i \in [m]$, $z_i \in \{0, 1\}$. Let $\mathcal{M}$ be an (ε, δ)-DP mechanism such that for any non-empty subset $U \subseteq [m]$, $\mathcal{M}\left(Z^U\right) = \mathbf{1}\left[\sum_{i \in U} z_i = 0\right]$ with probability p, otherwise $\mathcal{M}\left(Z^U\right) = \neg\mathbf{1}\left[\sum_{i \in U} z_i = 0\right]$. Here, $\mathbf{1}[\cdot]$ denotes the indicator function. If $U = \emptyset$ then $\mathcal{M}\left(Z^U\right) = \perp$. Essentially, $\mathcal{M}$ tells us whether the sum of its inputs is 0 or not, while preserving $(\varepsilon, 0)$-DP. Without using the privacy amplification result in (1), show that $\mathbf{M} \circ \mathcal{S}$ is $(\log\left(1 + q(e^{\varepsilon} - 1)\right), 0)$-DP.
(**Hint:** Start by determining the value of p for which $\mathcal{M}$ satisfies $(\varepsilon, 0)$-DP.)

Question 2 (Bridging the gap between the global and local models of privacy in FL). Keeping in mind the privacy amplification result in (1), modify the distributed gradient descent (DGD) method that we studied in the server-based FL setting with n clients in order to reduce the gap between the utilities of (ε, δ)-DP in the *global model* and *local model* cases.
(**Hint:** Instead of computing the full gradients over its local dataset comprising m data points, each client computes *stochastic* gradients using *Poisson sampled* data points.)

2. **Resilience and Redundancy.** Recall the problem of FL in the presence of adversarial (a.k.a. *Byzantine*-faulty) clients. Specifically, in the server-based setting comprising n clients we assume that up to f

clients can be adversarial and need not follow the prescribed instructions by the server correctly. Such adversarial clients may send arbitrary information regarding their local gradients or local datasets to the server during a learning algorithm. Recall that each client i holds m data points represented by set D_i. Assume that $n > 2f$. Now, consider the following properties of *f-Resilience* and *$2f$-Redundancy.*

(**f-Resilience.**) An FL algorithm Π is said to be *f-resilient* if and only if it enables the server to output $\widehat{h} \in \mathbb{H}$, where $\mathbb{H}$ denotes the hypothesis class, such that for all $S \subseteq [n]$ with $|S| = n - f$,

$$\widehat{h} \in \arg\min_{h \in \mathbb{H}} \sum_{i \in S} \mathcal{L}_i(h),$$

where $\mathcal{L}_i(h) \triangleq \frac{1}{m} \sum_{z \in D_i} \ell(h, z)$.

(**$2f$-Redundancy.**) The FL problem, represented by the tuple $(\mathbb{H}, \ell, D_1, \ldots, D_n)$, satisfies the property of *$2f$-redundancy* if and only if for all $S \subseteq [n]$, with $|S| = n - f$, and all $S' \subseteq S$, with $|S'| = n - 2f$,

$$\arg\min_{h \in \mathbb{H}} \sum_{i \in S'} \mathcal{L}_i(h) = \arg\min_{h \in \mathbb{H}} \sum_{i \in S} \mathcal{L}_i(h).$$

Question 3 (Redundancy implies resilience). Suppose that the server can ask for the entire dataset of each client. If client i is non-adversarial, it sends to the server its local dataset D_i correctly. However, if client i is adversarial, it can send to the server an arbitrary dataset, let's day $\widetilde{D}_i$, which comprises m data points that are different from those in D_i.
In this particular case, show that if the learning problem $(\mathbb{H}, \ell, D_1, \ldots, D_n)$ satisfies *$2f$-redundancy* then there exists an *f-resilient* algorithm Π.