Rizk *et al.*
EURASIP Journal on Advances in Signal Processing (2023) 2023:87
https://doi.org/10.1186/s13634-023-01049-4

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Privatized graph federated learning

Elsa Rizk[1*], Stefan Vlaski[2] and Ali H. Sayed[1]

*Correspondence:
elsa.rizk@epfl.ch

[1] School of Engineering, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[2] Department of Electrical and Electronic Engineering, Imperial College London, London, UK

Abstract

Federated learning is a semi-distributed algorithm, where a server communicates with multiple dispersed clients to learn a global model. The federated architecture is not robust and is sensitive to communication and computational overloads due to its one-master multi-client structure. It can also be subject to privacy attacks targeting personal information on the communication links. In this work, we introduce graph federated learning, which consists of multiple federated units connected by a graph. We then show how graph-homomorphic perturbations can be used to ensure the algorithm is differentially private on the server level. While on the client level, we show that improvement in the differentially private federated learning algorithm can be attained through the addition of random noise to the updates, as opposed to the models. We conduct both convergence and privacy theoretical analyses and illustrate performance by means of computer simulations.

Keywords: Federated learning, Distributed learning, Privatized learning, Differntial privacy

1 Introduction

Federated learning (FL) [1] is one particular distributed structure where users no longer need to send their data to a server for training. Instead, data remains local, and training happens in collaboration between different clients and the server. Compared to a fully decentralized solution, communication occurs between the server and the clients (or agents), instead of directly between the agents themselves. Such a solution is advantageous in the sense that users no longer need to worry about sharing their data with an unknown party, and the high cost of sending all their raw data is eliminated. In this way, the data stays locally safe on a user's device, and no extra communication cost is incurred for transferring the data remotely. However, such a distributed architecture is not robust to communication failures and computational overloads, nor it is immune to privacy attacks when agents are required to share their local updates. In standard FL, millions of users can be connected to *one* server at a time. This means one server will need to be responsible for the communication with all clients with significant computational burden, thus rendering the system susceptible to communication failures. Furthermore, whether clients send their gradient updates or their local models, information about their data can be inferred from the exchanges and leaked [2–5]. Consider for instance the logistic risk; the gradient of the loss function is a constant multiple of

the feature vector. Thus, even though the actual data samples are not sent to the server, information about them can still be inferred from the gradient updates or the models.

These considerations motivate us to propose an architecture for federated learning with privacy guarantees. In particular, we introduce the graph federated architecture, which consists of multiple servers, and we privatize the algorithm by ensuring the communication occuring between the servers and the clients is secure. Graph-homomorphic perturbations, which were initially introduced in [6], focus on the communication between servers. They are based on adding correlated noise to the messages sent between servers such that the noise cancels out if we were to take the average of all messages across all servers. As for the privatization between the clients and their servers, we share noisy updates as opposed to models. The two protocols make sure the effect of the added noise is reduced.

Other works have also contributed to addressing the same challenges we are considering in this work, albeit differently. For example, the work [7] introduces a hierarchical architecture, where it is assumed there are multiple servers connected in a tree structure. Such a solution still has one main server and thus faces the same robustness problem as FL. The graph federated learning architecture in this work (and which appeared in the earlier conference publication [8]) is a more general structure. The work [9] generalizes the standard distributed learning framework to include local updates, while [10] has a similar architecture to the GFL architecture proposed earlier in [8], it nevertheless does not deal with privacy and employs different objective functions and a different learning algorithm based on the alternating direction method of multipliers. Likewise, a plethora of solutions exist that relate to privacy issues. These methods may be split into two subgroups: those using random perturbations to ensure a certain level of differential privacy [11–20], or those that rely on cryptographic methods [21–25]. Both have their advantages and disadvantages. While differential privacy is easy to implement, it hinders the performance of the algorithm by reducing the model utility. As for cryptographic methods, they are generally harder to implement since they require more computational and communication power [26, 27]. Furthermore, they restrict the number of participating users. Moving forward, we go ahead with the study of differentially private methods.

The main contribution in this work is three-fold. We introduce a new generalized and more realistic architecture for the federated setting where we now consider multiple servers connected by some graph structure. Furthermore, many earlier works have proposed adding Laplacian noise sources to the shared information among agents in order to ensure some level of privacy. However, these works have largely ignored the fact that these noises degrade the mean-square error (MSE) performance of the network from $O(\mu)$ down to $O(\mu^{-1})$, where μ is the small learning parameter. To resolve this issue, we define a new noise generation scheme that mantains the MSE at $O(1)$ while ensuring privacy. Although the work [20] proposed a noisy-distributed consensus strategy, this reference lacks a useful construction method for the perturbations. In this work, we devise a construction scheme. Therefore, the main difference between our proposed method and previous works is that we devise a noise construction scheme that ensures the total sum of the added noise cancels out centrally. This results in the improved MSE bound of $O(1)$. Finally, we prove that clients sharing noisy updates as opposed to noisy models lead to improved performance relative to what is commonly done in the prior literature.

Moreover, we do not assume bounded gradients, as commonly assumed in previous works [12, 15, 16], since this condition does not actually hold in most situations in practice. Note, for instance, that even quadratic risks do not have bounded gradients. For this reason, we will not rely on this condition, and will instead be able to show that our noise construction is able to ensure differential privacy with high probability for most cases of interest. The main results shown in this work are as follows:

1. Privatized GFL under graph-homomorphic perturbations converges in the MSE sense to an $O(1)$ neighbourhood of the true model w^o as opposed to $O(\mu^{-1})$ when random perturbations are used instead.
2. Privatized FL under perturbed gradients converges in the MSE sense to an $O(\mu)$ neighbourhood of the true model w^o as opposed to $O(\mu^{-1})$ when perturbed models are shared instead.
3. GFL with graph-homomorphic perturbations and perturbed gradients is $\epsilon(i)$-differentially private with high probability.

2 Graph federated architecture

In the graph federated architecture, which we initially introduced in [8], we consider P federated units connected by a graph structure. Each federated unit consists of a server and a set of K agents. Thus, the overall architecture can be represented as a graph depicted in Fig. 1. We denote the combination matrix connecting the servers by $A \in \mathbb{R}^{P\times P}$, and we write a_{mp} to refer to the elements of A. We assume each agent of every server has its own dataset $\{x_{p,k,n}\}_{n=1}^{N_{p,k}}$ that is non-iid when compared to the other agents. The subscript p refers to the federated unit, k to the agent, and n to the data sample. We note the difference between our proposed architecture and a fully distributed setting. The graph federated architecture consists of a network of federated units while

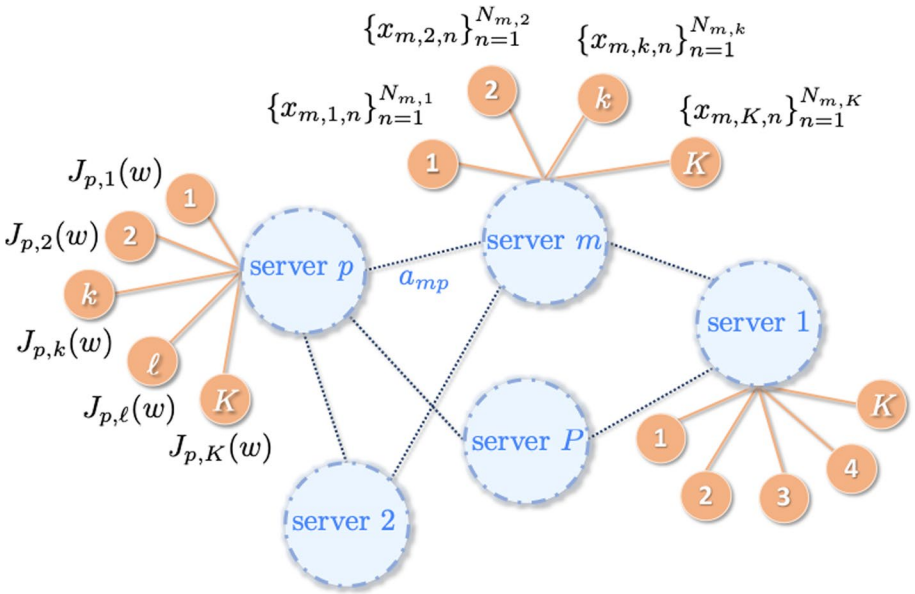


Fig. 1 The graph federated learning architecture

a fully distributed network removes the need for servers and assumes clients are connected to each other based on some graph structure. Such an architecture is an improvement on the original federated architecture and not necessarily on the fully distributed architecture. Instead of clients communicating with the same server, we split the load among multiple servers.

With this architecture, we associate a convex optimization problem that will take into account the cost function at each federated unit. Thus, the optimization goal is to find the optimal global model w^o that minimizes an average empirical risk:

$$w^o \triangleq \underset{w\in\mathbb{R}^M}{\operatorname{argmin}} \frac{1}{P}\sum_{p=1}^{P}\frac{1}{K}\sum_{k=1}^{K} J_{p,k}(w), \tag{1}$$

where each individual cost is an empirical risk defined over the local loss functions $Q_{p,k}(\cdot;\cdot)$:

$$J_{p,k}(w) \triangleq \frac{1}{N_{p,k}}\sum_{n=1}^{N_{p,k}} Q_{p,k}(w; x_{p,k,n}). \tag{2}$$

To solve problem (1) each federated unit p runs the standard federated averaging (FedAvg) algorithm [1]. An iteration i of the algorithm consists of the server p selecting a subset of L participating agents $\mathcal{L}_{p,i}$. Then, in parallel, each agent runs a series of stochastic gradient descent (SGD) steps. We call these local steps epochs, and denote an epoch by the letter e and the total number of epochs by $E_{p,k}$. The sampled data point at an agent k in the federated unit p during the e^{th} epoch of iteration i is denoted by b. Thus, during an iteration i, each participating agent $k \in \mathcal{L}_{p,i}$ updates the last model $\boldsymbol{w}_{p,i-1}$ and sends its new model $\boldsymbol{w}_{p,k,E_{p,k}}$ to the server after $E_{p,k}$ epochs. During a single epoch e, the agent updates its current local model $w_{p,k,e-1}$ by running a single SGD step. Thus, an agent repeats the following adaptation step for $e = 1, 2, \ldots, E_{p,k}$:

$$\boldsymbol{w}_{p,k,e} = \boldsymbol{w}_{p,k,e-1} - \frac{\mu}{E_{p,k}}\nabla_{w^\intercal} Q_{p,k}(\boldsymbol{w}_{p,k,e-1}; \boldsymbol{x}_{p,k,b}), \tag{3}$$

with $\boldsymbol{x}_{p,k,b}$ be the sampled data of agent k in federated unit p, and $\boldsymbol{w}_{p,k,0} = \boldsymbol{w}_{p,i-1}$. After all the participating agents $k \in \mathcal{L}_{p,i}$ run all their epochs, the server aggregates their final models $\boldsymbol{w}_{p,k,E_{p,k}}$, which we rename as $\boldsymbol{w}_{p,k,i}$ since it is the final local model at iteration i:

$$\boldsymbol{\psi}_{p,i} = \frac{1}{L}\sum_{k\in\mathcal{L}_{p,i}} \boldsymbol{w}_{p,k,i}. \tag{4}$$

Next, at the server level, these estimates are combined across neighbourhoods using a diffusion type strategy, where we first consider the previous steps (3) and (4) as the adaptation step and the following step as the combination step:

$$\boldsymbol{w}_{p,i} = \sum_{m\in\mathcal{N}_p} a_{pm}\boldsymbol{\psi}_{m,i}. \tag{5}$$

To introduce privacy, the models communicated at each round between the agents and the servers need to be encrypted in some way. We could either apply secure multiparty

computation (SMC) tools, like secret sharing, or use differential privacy. We focus on differential privacy or masking tools that can be represented by added noise. Thus, we let agent 1 in federated unit 2 add a noise component $\boldsymbol{g}_{2,1,i}$ to its final model $\boldsymbol{w}_{2,1,i}$ at iteration i, and then let serever 2 add $\boldsymbol{g}_{12,i}$ to the message $\boldsymbol{\psi}_{2,i}$ it sends to server 1. More generally, we denote by $\boldsymbol{g}_{pm,i}$ the noise added to the message sent by server m to server p at iteration i. Similarly, we denote by $\boldsymbol{g}_{p,k,i}$ the noise added to the model sent by agent k to server p during the ith iteration. We use unseparated subscripts pm for the inter-server noise components to point out their ability to be combined into a matrix structure. Contrarily, the agent-server noise components' subscripts are separated by a comma to highlight a hierarchical structure. Thus, the privatized algorithm can be written as a client update step (6), a server aggregation step (7), and a server combination step (8):

$$\boldsymbol{w}_{p,k,i} = \boldsymbol{w}_{p,i-1} - \frac{\mu}{E_{p,k}} \sum_{e=1}^{E_{p,k}} \nabla_{w^\intercal} Q_{p,k}(\boldsymbol{w}_{p,k,e-1}; \boldsymbol{x}_{p,k,b}), \tag{6}$$

$$\boldsymbol{\psi}_{p,i} = \frac{1}{L} \sum_{k \in \mathcal{L}_{p,i}} \boldsymbol{w}_{p,k,i} + \boldsymbol{g}_{p,k,i}, \tag{7}$$

$$\boldsymbol{w}_{p,i} = \sum_{m \in \mathcal{N}_p} a_{pm}(\boldsymbol{\psi}_{m,i} + \boldsymbol{g}_{pm,i}). \tag{8}$$

The client update step (6) follows from (3) by combining the multiple epochs for $e = 1, 2, \ldots, E_{p,k}$ into one update step, with $\boldsymbol{w}_{p,k,i} = \boldsymbol{w}_{p,k,E_{p,k}}$ and $\boldsymbol{w}_{p,k,0} = \boldsymbol{w}_{p,i-1}$, namely:

$$\begin{aligned} \boldsymbol{w}_{p,k,E_{p,k}} &= \boldsymbol{w}_{p,k,E_{p,k}-1} - \frac{\mu}{E_{p,k}} \nabla_{w^\intercal} Q_{p,k}(\boldsymbol{w}_{p,k,E_{p,k}-1}; \boldsymbol{x}_{p,k,b}) \\ &= \boldsymbol{w}_{p,k,E_{p,k},-2} - \frac{\mu}{E_{p,k}} \sum_{e=E_{p,k}-1}^{E_{p,k}} \nabla_{w^\intercal} Q_{p,k}(\boldsymbol{w}_{p,k,e-1}; \boldsymbol{x}_{p,k,b}) \\ &= \boldsymbol{w}_{p,k,0} - \frac{\mu}{E_{p,k}} \sum_{e=1}^{E_{p,k}} \nabla_{w^\intercal} Q_{p,k}(\boldsymbol{w}_{p,k,e-1}; \boldsymbol{x}_{p,k,b}). \end{aligned} \tag{9}$$

3 Performance analysis

In this section, we show a list of results on the performance of the algorithm. We study the convergence of the privatized algorithm (6)–(8), and examine the effect of privatization on performance.

3.1 Modeling conditions

To go forward with our analysis, we require certain reasonable assumptions on the graph structure and cost functions.

Assumption 1 (Combination matrix) The combination matrix A describing the graph is symmetric and doubly-stochastic, i.e.:

$$a_{pm} = a_{mp}, \quad \sum_{m=1}^{P} a_{mp} = 1. \tag{10}$$

Furthermore, the graph is strongly-connected and A satisfies:

$$\iota_2 \triangleq \rho\left(A - \frac{1}{P}\mathbb{1}\mathbb{1}^{\mathsf{T}}\right) < 1. \tag{11}$$

□

Assumption 2 (*Convexity and smoothness*) The empirical risks $J_{p,k}(\cdot)$ are $\nu-$ *strongly* convex, and the loss functions $Q_{p,k}(\cdot;\cdot)$ are convex, namely for $\nu > 0$,:

$$J_{p,k}(w_2) \geq J_{p,k}(w_1) + \nabla_{w^{\mathsf{T}}} J_{p,k}(w_1)(w_2 - w_1) + \frac{\nu}{2}\|w_2 - w_1\|^2, \tag{12}$$

$$Q_{p,k}(w_2;\cdot) \geq Q_{p,k}(w_1;\cdot) + \nabla_{w^{\mathsf{T}}} Q_{p,k}(w_1;\cdot)(w_2 - w_1). \tag{13}$$

Furthermore, the loss functions have δ-Lipschitz continuous gradients, meaning there exists $\delta > 0$ such that for any data point $x_{p,n}$:

$$\|\nabla_{w^{\mathsf{T}}} Q_{p,k}(w_2; x_{p,k,n}) - \nabla_{w^{\mathsf{T}}} Q_{p,k}(w_1; x_{p,k,n})\| \leq \delta\|w_2 - w_1\|. \tag{14}$$

□

We also require a bound on the difference between the global optimal model w^o and the local optimal models $w^o_{p,k}$ that optimize $J_{p,k}(\cdot)$. This assumption is used to bound the gradient noise and the incremental noise defined further ahead. It is not a restrictive assumption, and it imposes a condition on when collaboration is sensical among different agents. In other words, since the agents have non-iid data, sometimes their optimal models are too different and collaboration would hurt their individual performance. For example, when considering recommender systems, people in the same country are more likely to get the same movie recommended as opposed to across different countries. This means, people of the same country might have different models but relatively close contrary to different countries.

Assumption 3 (*Model drifts*) The distance of each local model $w^o_{p,k}$ to the global model w^o is uniformly bounded, i.e., there exists $\xi \geq 0$ such that $\|w^o - w^o_p\| \leq \xi$.

3.2 Network centroid convergence

We study the convergence of the algorithm from the network centroid's $\boldsymbol{w}_{c,i}$ perspective:

$$\boldsymbol{w}_{c,i} \triangleq \frac{1}{P}\sum_{p=1}^{P} \boldsymbol{w}_{p,i}. \tag{15}$$

We write the central recursion as:

$$\begin{aligned}\boldsymbol{w}_{c,i} = \boldsymbol{w}_{c,i-1} - \mu \frac{1}{PL} \sum_{p=1}^{P} \sum_{k \in \mathcal{L}_{p,i}} \frac{1}{E_{p,k}} \sum_{e=1}^{E_{p,k}} \nabla_{w^\mathsf{T}} Q_{p,k}(\boldsymbol{w}_{p,k,e-1}; \boldsymbol{x}_{p,k,b}) \\ + \frac{1}{PL} \sum_{p=1}^{P} \sum_{k \in \mathcal{L}_{p,i}} \boldsymbol{g}_{p,k,i} + \frac{1}{P} \sum_{p,m=1}^{P} a_{pm} \boldsymbol{g}_{pm,i}.\end{aligned} \tag{16}$$

Next, we define the model error as $\widetilde{\boldsymbol{w}}_{c,i} \triangleq w^o - \boldsymbol{w}_{c,i}$ and the average gradient noise:

$$\boldsymbol{s}_i \triangleq \frac{1}{P} \sum_{p=1}^{P} \boldsymbol{s}_{p,i}, \tag{17}$$

with the per-unit gradient noise $\boldsymbol{s}_{p,i}$:

$$\boldsymbol{s}_{p,i} \triangleq \widehat{\nabla_{w^\mathsf{T}} J_p}(\boldsymbol{w}_{p,i-1}) - \nabla_{w^\mathsf{T}} J_p(\boldsymbol{w}_{p,i-1}), \tag{18}$$

and

$$\widehat{\nabla_{w^\mathsf{T}} J_p}(\cdot) \triangleq \frac{1}{L} \sum_{k \in \mathcal{L}_{p,i}} \frac{1}{E_{p,k}} \sum_{e=1}^{E_{p,k}} \nabla_{w^\mathsf{T}} Q_{p,k}(\cdot; \boldsymbol{x}_{p,k,b}). \tag{19}$$

We introduce the average incremental noise $\boldsymbol{q}_i$ and the local incremental noise $\boldsymbol{q}_{p,i}$, which capture the error introduced by the multiple local update steps:

$$\boldsymbol{q}_i \triangleq \frac{1}{P} \sum_{p=1}^{P} \boldsymbol{q}_{p,i}, \tag{20}$$

$$\boldsymbol{q}_{p,i} \triangleq \frac{1}{L} \sum_{k \in \mathcal{L}_{p,i}} \frac{1}{E_{p,k}} \sum_{e=1}^{E_k} \left(\nabla_{w^\mathsf{T}} Q_{p,k}(\boldsymbol{w}_{p,k,e-1}; \boldsymbol{x}_{p,k,b}) - \nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{p,i-1}; \boldsymbol{x}_{p,k,b}) \right) \tag{21}$$

We then arrive at the following error recursion:

$$\widetilde{\boldsymbol{w}}_{c,i} = \widetilde{\boldsymbol{w}}_{c,i-1} + \mu \frac{1}{P} \sum_{p=1}^{P} \nabla_{w^\mathsf{T}} J_p(\boldsymbol{w}_{p,i-1}) + \mu \boldsymbol{s}_i + \mu \boldsymbol{q}_i - \boldsymbol{g}_i, \tag{22}$$

where $\boldsymbol{g}_i$ is the total added noise at iteration i:

$$\boldsymbol{g}_i \triangleq \frac{1}{PL} \sum_{p=1}^{P} \sum_{k \in \mathcal{L}_{p,i}} \boldsymbol{g}_{p,k,i} + \frac{1}{P} \sum_{p,m=1}^{P} a_{pm} \boldsymbol{g}_{pm,i} \tag{23}$$

We estimate the first and second-order moments of the gradient noise in the following lemma. To do so, we use the fact, shown in previous work (Lemma 1 in [28]), that the individual gradient noise is zero-mean with a bounded second order moment:

$$\mathbb{E}\left\{ \|\boldsymbol{s}_{p,i}\|^2 | \mathcal{F}_{i-1} \right\} \leq \beta_{s,p}^2 \|\widetilde{\boldsymbol{w}}_{p,i-1}\|^2 + \sigma_{s,p}^2, \tag{24}$$

where the constants are defined as:

$$\beta_{s,p}^2 \triangleq \frac{6\delta^2}{L}\left(1+\frac{1}{K}\sum_{k=1}^{K}\frac{1}{E_{p,k}}\right), \tag{25}$$

$$\sigma_{s,p}^2 \triangleq \frac{1}{LK}\sum_{k=1}^{K}\left(\frac{12}{E_{p,k}}+3\right)\frac{1}{N_{p,k}}\sum_{n=1}^{N_{p,k}}\|\nabla_{w^\intercal}Q_{p,k}(w^o;x_{p,k,n})\|^2, \tag{26}$$

and $\mathcal{F}_{i-1}$ is the filtration defined over the randomness introduced by all the past subsampling of the data for the calculation of the stochastic gradient. Using Assumption 3, we can guarantee that $\sigma_{s,p}^2$ is bounded by bounding:

$$\|\nabla_{w^\intercal}Q_{p,k}(w^o;x_{p,k,n})\|^2 \leq 2\|\nabla_{w^\intercal}Q_{p,k}(w_{p,k}^o;x_{p,k,n})\|^2 + 2\delta^2\xi^2. \tag{27}$$

Lemma 1 (*Estimation of first and second-order moments of the gradient noise*) *The gradient noise defined in* (17) *is zero-mean and has a bounded second-order moment*:

$$\mathbb{E}\left\{\|\boldsymbol{s}_i\|^2|\mathcal{F}_{i-1}\right\} \leq \beta_s^2\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + \sigma_s^2 + \frac{2}{P}\sum_{p=1}^{P}\beta_{s,p}^2\|\boldsymbol{w}_{p,i-1}-\boldsymbol{w}_{c,i-1}\|^2 \tag{28}$$

where the constants β_s^2 *and* σ_s^2 *are given by*:

$$\beta_s^2 \triangleq \frac{2}{P}\sum_{p=1}^{P}\beta_{s,p}^2, \quad \sigma_s^2 \triangleq \frac{1}{P}\sum_{p=1}^{P}\sigma_{s,p}^2. \tag{29}$$

Proof The above result follows from applying the Jensen's inequality and the bounds on the per-unit gradient noise $\boldsymbol{s}_{p,i}$. □

The new term found in the bound of the gradient term is what we call the network disagreement:

$$\frac{1}{P}\sum_{p=1}^{P}\|\boldsymbol{w}_{p,i}-\boldsymbol{w}_{c,i}\|^2. \tag{30}$$

It captures the difference in the path taken by the individual models versus the network centroid. We bound this difference in Lemma 3. However, before doing so, we show that the second order moment of the incremental noise is on the order of $O(\mu)$. From Lemma 5 in [28], we can bound the individual incremental noise:

$$\mathbb{E}\|\boldsymbol{q}_{p,i}\|^2 \leq a\mu^2\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,i-1}\|^2 + a\mu^2\xi^2 + \frac{1}{K}\sum_{k=1}^{K}(b_k\mu^4+c_k\mu^2)\sigma_{q,p,k}^2, \tag{31}$$

where the constants are given by:

$$a \triangleq \frac{4\delta^2}{K}\sum_{k=1}^{K}\frac{(E_{p,k}+1)(1-\lambda)-1+\lambda^{E_{p,k}+1}}{E_{p,k}^2(1-\lambda)^2}, \tag{32}$$

$$b_k \triangleq \frac{2E_{p,k}(E_{p,k}+1)(1-\lambda)^2 - 4E_{p,k}(1-\lambda)+4\lambda}{E_{p,k}^2(1-\lambda)^3} - \frac{2\lambda^{E_{p,k}+1}}{E_{p,k}^2(1-\lambda)^3}, \tag{33}$$

$$c_k \triangleq \frac{E_{p,k}-1}{3E_{p,k}}, \tag{34}$$

$$\lambda \triangleq 1 - 2\nu\mu + 4\delta^2\mu^2, \tag{35}$$

$$\sigma_{q,p,k}^2 \triangleq 3\sum_{n=1}^{N_{p,k}} \|\nabla_{w^\mathsf{T}} Q_{p,k}(w_{p,k}^o; x_{p,k,n})\|^2. \tag{36}$$

The following result follows.

Lemma 2 (*Estimation of second-order moment of the incremental noise*) *The incremental noise defined in* (20) *has a bounded second-order moment*:

$$\begin{aligned}\mathbb{E}\|\boldsymbol{q}_i\|^2 \le{}& O(\mu)\mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + O(\mu)\xi^2 + O(\mu^2)\sigma_q^2 \\ &+ \frac{O(\mu)}{P}\sum_{p=1}^{P}\mathbb{E}\|\boldsymbol{w}_{p,i-1} - \boldsymbol{w}_{c,i-1}\|^2,\end{aligned} \tag{37}$$

where the constant σ_q^2 *is the average of* $\sigma_{q,p,k}^2$:

$$\sigma_q^2 \triangleq \frac{1}{PK}\sum_{p=1}^{P}\sum_{k=1}^{K}(b_k\mu^4 + c_k\mu^2)\sigma_{q,p,k}^2. \tag{38}$$

Proof The above result follows from applying the Jensen inequality and the bounds on the per-unit incremental noise $\boldsymbol{q}_{p,i}$. Furthermore, $a = O(\mu^{-1})$, $b_k = O(\mu^{-1})$, and $c_k = O(1)$ reduce the expression to (37). □

We now bound the network disagreement. To do so, we first introduce the eigendecomposition of $A = QHQ^\mathsf{T}$:

$$Q \triangleq \begin{bmatrix}\frac{1}{\sqrt{P}}\mathbb{1} & Q_\theta\end{bmatrix}, \quad H \triangleq \begin{bmatrix}1 & 0\\ 0 & H_\theta\end{bmatrix}, \tag{39}$$

where H_θ is a diagonal matrix that includes the last $(P-1)$ eigenvalues of A and Q_θ their corresponding eigenvectors.

Lemma 3 (*Network disagreement*) *The average deviation from the centroid is bounded during each iteration i*:

$$\frac{1}{P}\sum_{p=1}^{P}\mathbb{E}\|\boldsymbol{w}_{p,i}-\boldsymbol{w}_{c,i}\|^2 \le \frac{\iota_2^i}{P}\mathbb{E}\|(Q_\epsilon\otimes I)\boldsymbol{\mathcal{W}}_0\|^2+\frac{\iota_2^2}{P}\sum_{j'=0}^{i-1}\iota_2^{j'}\sum_{p=1}^{P}\Bigg\{\mu^2\Bigg(\frac{2\delta^2}{\iota_2(1-\iota_2)} + \beta_{s,p}^2+O(\mu)\Bigg)\Bigg(\lambda_p^{j'}A^{j'}[p]\,\mathrm{col}\left\{\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,0}\|^2\right\}_{p=1}^{P}+\sum_{j=0}^{j'-1}\lambda_p^j \times A^j[p]\,\mathrm{col}\left\{\mu^2\sigma_{s,p}^2+O(\mu^2)\xi^2+O(\mu^3)\sigma_{q,p}^2+\sigma_{g,p}^2\right\}_{p=1}^{P}\Bigg) + \mu^2\frac{2\|\nabla_{w^\intercal}J_p(w^o)\|^2}{\iota_2(1-\iota_2)}+\mu^2\sigma_{s,p}^2+O(\mu^3)\xi^2+O(\mu^4)\sigma_{q,p}^2 + \frac{1}{\iota_2^2}\sigma_{g,p}^2\Bigg\}, \tag{40}$$

where $\boldsymbol{\mathcal{W}}_0 \triangleq \mathrm{col}\left\{\boldsymbol{w}_{p,0}\right\}_{p=1}^{P}$ *and* $\lambda_p \triangleq \sqrt{1-2\nu\mu+\delta^2\mu^2}+\beta_{s,p}^2\mu^2+O(\mu^2)\in(0,1)$. *Then, in the limit*:

$$\limsup_{i\to\infty}\frac{1}{P}\sum_{p=1}^{P}\mathbb{E}\|\boldsymbol{w}_{p,i}-\boldsymbol{w}_{c,i}\|^2 \le \frac{\iota_2^2}{P(1-\iota_2)}\sum_{p=1}^{P}\mu^2\sigma_{s,p}^2+\frac{1}{\iota_2^2}\sigma_{g,p}^{+}O(\mu)\sigma_{g,p}^2+O(\mu^3). \tag{41}$$

Proof See "Appendix 2". □

Thus, from the above lemma, we see that the individual models gravitate to the centroid model with an error introduced due to the added privatization. The effect of the added noise overpowers that of the gradient and incremental noise, since the later is on the order of the step-size.

Then, using the above result, we can establish the convergence of the centroid model to a neighbourhood of the true optimal model w^o in the mean-square-error (MSE) sense.

Theorem 1 (*Centroid MSE convergence*) *Under Assumptions 1, 2 and 3, the network centroid converges to the optimal point* w^o *exponentially fast for a sufficiently small step-size* μ:

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i}\|^2 \le \lambda_c\mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2+\mu^2\sigma_s^2+O(\mu^2)\xi^2+O(\mu^3)\sigma_q^2+\mathbb{E}\|\boldsymbol{g}_i\|^2 + \frac{O(\mu)}{P}\sum_{p=1}^{P}\mathbb{E}\|\boldsymbol{w}_{p,i-1}-\boldsymbol{w}_{c,i-1}\|^2, \tag{42}$$

where $\lambda_c=\sqrt{1-2\nu\mu+\delta^2\mu^2}+\beta_s^2\mu^2+O(\mu^2)\in(0,1)$. *Then, letting i tend to infinity, we get*:

$$\limsup_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i}\|^2 \le \frac{\mu^2\sigma_s^2+O(\mu^2)\xi^2+O(\mu^3)\sigma_q^2+\mathbb{E}\|\boldsymbol{g}\|^2}{1-\lambda_c}+\sum_{p=1}^{P}O(1)\sigma_{g,p}^2+O(\mu). \tag{43}$$

Proof See "Appendix 3". □

The main term in the above bound is the variance of the added noise with a dominating factor of μ^{-1}, since:

$$1-\lambda_c = 1-\sqrt{1-O(\mu)+O(\mu^2)}-O(\mu^2) = O(\mu)-O(\mu^2) = O(\mu) \tag{44}$$

which allows us to rewrite the bound as follows:

$$\begin{aligned}\limsup_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i}\|^2 \leq\ & O(\mu)\sigma_s^2 + O(\mu)\xi^2 + O(\mu^2)\sigma_q^2 + O(\mu^{-1})\mathbb{E}\|\boldsymbol{g}\|^2 \\ &+ \sum_{p=1}^{P} O(1)\sigma_{g,p}^2 + O(\mu),\end{aligned} \tag{45}$$

with $\mathbb{E}\|\boldsymbol{g}\|^2$ representing the variance of the total added noise, independent of time. While in general decreasing the step-size improves performance, the above result shows that this need not be the case with privatization. Thus, since the added noise impacts the model utility negatively, it is important to choose a privatization scheme that reduces the effect. In what follows, we look closely at such a scheme.

3.3 Graph-homomorphic perturbations

We consider a specific privatization scheme and specialize the above results. The goal of the scheme is to remove the $O(\mu^{-1})$ term from the MSE bounds. Thus, focusing on the centroid model expression (16), we wish to cancel out the total added noise amongst servers, i.e.,

$$\sum_{p,m=1}^{P} a_{pm}\boldsymbol{g}_{pm,i} = 0. \tag{46}$$

To achieve this, we introduce graph-homomorphic perturbations defined as follows [6]. We assume each server p draws a sample $\boldsymbol{g}_{p,i}$ independently from the Laplace distribution $Lap(0,\sigma_g/\sqrt{2})$ with variance σ_g^2. Server p then sets the noise $\boldsymbol{g}_{mp,i}$ added to the message sent to its neighbour m as:

$$\boldsymbol{g}_{mp,i} = \begin{cases} \boldsymbol{g}_{p,i} & m \neq p \\ -\frac{1-a_{pp}}{a_{pp}}\boldsymbol{g}_{p,i}. \end{cases} \tag{47}$$

With such a construction, condition (46) is satisfied:

$$\begin{aligned}\sum_{p,m=1}^{P} a_{pm}\boldsymbol{g}_{pm,i} &= \sum_{p\neq m} a_{pm}\boldsymbol{g}_{p,i} - \sum_{p=1}^{P} a_{pp}\left(\frac{1-a_{pp}}{a_{pp}}\right)\boldsymbol{g}_{p,i} \\ &= \sum_{p=1}^{P}(1-a_{pp})\boldsymbol{g}_{p,i} - (1-a_{pp})\boldsymbol{g}_{p,i} = 0.\end{aligned} \tag{48}$$

Thus, with such a scheme, the noise components proportional to $O(\mu^{-1})$ resulting from the noise added between the servers cancel out in the error recursions, however since

gradients are evaluated at the local models $\boldsymbol{w}_{p,i}$ and not at the centroid $\boldsymbol{w}_{c,i}$, thus the effect of the noise is still evident. Yet, this remaining error introduced by the noise is controlled by the step-size. Thus, its effect can be mitigated by using a smaller step-size. In the next corollary, we show that if no noise is added amongst the clients and graph-homomorphic perturbations are used amongst servers, then the error converges to $O(1)\sigma_g^2$.

Corollary 1 (*Centroid MSE convergence under graph-homomorphic perturbations*) *Under Assumptions 1, 2 and 3, the network centroid with graph-homomorphic perturbations converges to the optimal point w^o exponentially fast for a sufficiently small step-size μ*:

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i}\|^2 \leq \lambda_c \mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + \mu^2\sigma_s^2 + O(\mu^2)\xi^2 + O(\mu^3)\sigma_q^2 + \frac{O(\mu)}{P}\sum_{p=1}^{P}\mathbb{E}\|\boldsymbol{w}_{p,i-1} - \boldsymbol{w}_{c,i-1}\|^2. \tag{49}$$

Then, letting i tend to infinity, we get:

$$\limsup_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i}\|^2 \leq \frac{\mu^2\sigma_s^2 + O(\mu^2)\xi^2 + O(\mu^3)\sigma_q^2}{1-\lambda_c} + \sum_{p=1}^{P} O(1)\sigma_{g,p}^2 + O(\mu). \tag{50}$$

Proof Starting from (43), and replacing $\mathbb{E}\|\boldsymbol{g}\|^2 = 0$ because $\boldsymbol{g}_i = 0$, we get the final result. □

3.4 Sharing gradients as opposed to weight estimates

We next show that sharing gradients versus models is better for the performance under added noise. In the remainder of this section and for the sake of simplicity, we illustrate this conclusion by considering one federated unit, say for $p = 1$. Thus, if we were to introduce differential privacy to federated learning, then a random Laplacian noise should be added to each model by the client before aggregation by the server, and the new privatized aggregation step will become:

$$\boldsymbol{w}_{1,i} = \frac{1}{L}\sum_{k\in\mathcal{L}_{1,i}}\left(\boldsymbol{w}_{1,k,i} + \boldsymbol{g}_{1,k,i}\right). \tag{51}$$

However, if we were to study the MSE convergence of this privatized algorithm, we would notice a new $O(\mu^{-1})\sigma_g^2$ term in the bound (Theorem 1). To address this degradation, we now describe an alternative implementation that shares *gradients* as opposed to weight estimates. Note first that the FL algorithm can be expressed in a single step taken from the server's perspective:

$$\boldsymbol{w}_{1,i} = \boldsymbol{w}_{1,i-1} - \mu \frac{1}{L} \sum_{k \in \mathcal{L}_{1,i}} \frac{1}{E_{1,k}} \sum_{e=1}^{E_{1,k}} \widehat{\nabla_{w^\intercal} J_{1,k}}(\boldsymbol{w}_{1,k,e-1}). \tag{52}$$

This suggests that instead of every agent sharing its final model $\boldsymbol{w}_{1,k,i}$, they could share the total update:

$$\frac{1}{E_{1,k}} \sum_{e=1}^{E_{1,k}} \widehat{\nabla_{w^\intercal} J_{1,k}}(\boldsymbol{w}_{1,k,e-1}). \tag{53}$$

The server then aggregates the updates from all participating agents and updates the previous model $\boldsymbol{w}_{1,i-1}$. In this case, if we were to privatize this new version of the algorithm, we would add random noise to the updates which are then scaled by the step-size:

$$\boldsymbol{\psi}_{1,k,i-1} = \frac{1}{E_{1,k}} \sum_{e=1}^{E_{1,k}} \widehat{\nabla_{w^\intercal} J_{1,k}}(\boldsymbol{w}_{1,k,e-1}), \tag{54}$$

$$\boldsymbol{w}_{1,i} = \boldsymbol{w}_{1,i-1} - \mu \frac{1}{L} \sum_{k \in \mathcal{L}_{1,i}} \left(\boldsymbol{\psi}_{1,k,i-1} + \boldsymbol{g}_{1,k,i}\right). \tag{55}$$

We show in the following theorem the effect of the added noise to the new FL algorithm. It turns out, the noise introduces an $O(\mu)$ error instead of $O(\mu^{-1})$.

Theorem 2 (*MSE convergence of privatized FL*) *Under Assumptions 2 and 3, the privatized FL algorithm (54)–(55) converges exponentially fast for a small enough step-size to a neighbourhood of the optimal model*:

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_{1,i}\|^2 \leq \lambda \mathbb{E}\|\widetilde{\boldsymbol{w}}_{1,i-1}\|^2 + O(\mu^2)\sigma_{s,1}^2 + O(\mu^2)\xi^2 + \frac{\mu^2}{L}\sigma_{g,1}^2 + O(\mu^3). \tag{56}$$

where $\lambda = \sqrt{1 - 2\nu\mu + (\beta_{s,1}^2 + \delta^2)\mu^2} + O(\mu^2) \in (0,1)$. *Then, in the limit*:

$$\limsup_{i \to \infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{1,i}\|^2 \leq O(\mu)(\sigma_{s,1}^2 + \xi^2 + \sigma_{g,1}^2) + O(\mu^2). \tag{57}$$

Proof See "Appendix 6". □

Thus, sharing the updates instead of the models is advantageous since the effect of the added noise on the performance is reduced. The $O(\mu)$ factor allows us to increase the value of the noise variance while ensuring the model utility does not deteriorate significantly. Therefore, to guarantee an $\epsilon(i)$-DP algorithm, we let the added noise be a zero-mean Laplacian random variable with σ_g^2 variance.

4 Privacy analysis

We study the privacy of the algorithm (6)–(8) in terms of differential privacy. We focus on graph-homomorphic perturbations and show that the adopted scheme is differentially private. To do so, we first define what it means for an algorithm to be ϵ-differentially private. Therefore, without loss of generality, assume agent 1 in federated unit 1 decides to not

participate, and its data samples $x_{1,1}$ are replaced by a new set $x'_{1,1}$ with a different distribution. Then, with the new data, the algorithm takes a different path. We denote the new models by $\boldsymbol{w}'_{p,k,i}$. The idea behind differential privacy is that an outside observant should not be able to distinguish between the two trajectories $\boldsymbol{w}_{p,k,i}$ and $\boldsymbol{w}'_{p,k,i}$ and conclude whether agent one participated in the training. More formally, differential privacy is defined bellow.

Definition 1 ($\epsilon(i)$-Differential Privacy) We say that the algorithm given in (6)–(8) is $\epsilon(i)$-differentially private for server p at time i if the following condition holds on the joint distribution $f(\cdot)$:

$$\frac{f\left(\left\{\left\{\boldsymbol{\psi}_{p,j}+\boldsymbol{g}_{pm,j}\right\}_{m\in\mathcal{N}_p\setminus\{p\}}\right\}_{j=0}^{i}\right)}{f\left(\left\{\left\{\boldsymbol{\psi}'_{p,j}+\boldsymbol{g}_{pm,j}\right\}_{m\in\mathcal{N}_p\setminus\{p\}}\right\}_{j=0}^{i}\right)}\leq e^{\epsilon(i)}. \tag{58}$$

□

Thus, the above definition states that minimaly varried trajectories have comparable probabilities. In addition, the smaller the value of ϵ is, the higher the privacy guarantee will be. Thus, the goal will be to decrease ϵ as long as the model utility is not strongly affected.

Next, in order to show that the algorithm is differentially private, we require the sensitivity of the algorithm to be bounded. The sensitivity at time i is defined as:

$$\Delta(i)=\|\boldsymbol{\mathcal{W}}_i-\boldsymbol{\mathcal{W}}'_i\|. \tag{59}$$

It measures the distance between the original and perturbed weight vectors. It is shown in "Appendix 4" that $\Delta(i)$ can be bounded as follows:

$$\Delta(i)\leq B+B'+\sqrt{P}\|w^o-w'^o\|, \tag{60}$$

for constants B and B' chosen by the designer. Moreover, the above bound holds with high probability given by:

$$\begin{aligned}\mathbb{P}(\Delta(i)\leq B+B'+\sqrt{P}\|w^o-w'^o\|)\geq&\left(1-\frac{\lambda^i_{\max}\mathbb{E}\|\boldsymbol{\mathcal{W}}_0\|^2+O(\mu)+O(\mu^{-1})}{B^2}\right)\\&\times\left(1-\frac{\lambda'^i_{\max}\mathbb{E}\|\boldsymbol{\mathcal{W}}'_0\|^2+O(\mu)+O(\mu^{-1})}{B'^2}\right).\end{aligned} \tag{61}$$

This result shows that the sensitivity can be bounded with high probability, which in turn is dependent on the values chosen for B and B'. Larger values for these constants increase the probability, but nevertheless lead to a looser bound for privacy (as shown in Theorem 3). Therefore, the choice of B and B' needs to be balanced judiciously to ensure the desired level of privacy.

Using the bound on the sensitivity and from the definition of differential privacy, we can finally show that the algorithm is differentially private with high probability.

Theorem 3 (*Privacy of GFL algorithm*) *If the algorithm* (6)–(8) *adopts graph-homomorphic perturbations, then it is* $\epsilon(i)$*-differentially private with high probability, at time* i *for a standard deviation of* $\sigma_g = \sqrt{2}(B + B'\sqrt{P}\|w^o - w'^o\|)(i+1)/\epsilon(i)$.

Proof See "Appendix 5". □

Thus, the above theorem suggests, if we wish the algorithm to be $\epsilon(i)$-differentially private, then we need to choose the noise variance accordingly. The larger the variance is, the more private the algorithm will be. However, the longer the algorithm is run, we will require a larger noise variance to keep the same level of privacy guarantee. Said differently, if we fix the added noise, then as time passes, the algorithm becomes less private, and more information is leaked. However, with graph-homomorphic perturbations, we can afford to increase the variance since its effect is constant on the MSE, and thus decreases the leakage.

Moreover, we study the effect of the model drift on the privacy of the algorithm. Thus, if we examine closely the probability that the sensitivity is bounded, the model drift ξ appears in the $O(\mu)$ term. The smaller the model drift is, we note that the higher the probability that the sensitivity is bounded. This in turn implies that the algorithm is differentially private with higher probability. Furthermore, if we study the average $\epsilon(i)$, we see that:

$$\begin{aligned}
\mathbb{E}\,\epsilon(i) &= \frac{\sqrt{2}}{\sigma_g}\sum_{j=1}^{i}\mathbb{E}\Delta(j)\\
&\le \frac{\sqrt{2}}{\sigma_g}\sum_{j=1}^{i}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,j}\| + \mathbb{E}\|\widetilde{\boldsymbol{w}}'_{p,j}\| + \|w^o - w'^o\|\\
&\le \frac{\sqrt{2}}{\sigma_g}\sum_{j=1}^{i}\lambda^{j/2}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,0}\| + \frac{1}{\sqrt{1-\lambda}}\Big(O(\mu)(\sigma_{s,p}^2 + \xi^2 + \sigma_g^2) + O(\mu^{3/2})\Big)\\
&\quad + \lambda'^{j/2}\mathbb{E}\|\widetilde{\boldsymbol{w}}'_{p,0}\| + \frac{1}{\sqrt{1-\lambda'}}\Big(O(\mu)(\sigma_{s,p}'^2 + \xi'^2 + \sigma_g^2) + O(\mu^{3/2})\Big)\\
&\quad + \|w^o - w'^o\|\\
&\le \frac{1-\lambda^{i/2}}{1-\lambda^{1/2}}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,0}\| + \frac{1-\lambda'^{i/2}}{1-\lambda'^{1/2}}\mathbb{E}\|\widetilde{\boldsymbol{w}}'_{p,0}\| + i\|w^o - w'^o\|\\
&\quad + \frac{i}{\sqrt{1-\lambda}}\Big(O(\mu)(\sigma_{s,p}^2 + \xi^2 + \sigma_g^2) + O(\mu^{3/2})\Big)\\
&\quad + \frac{i}{\sqrt{1-\lambda'}}\Big(O(\mu)(\sigma_{s,p}'^2 + \xi'^2 + \sigma_g^2) + O(\mu^{3/2})\Big),
\end{aligned} \tag{62}$$

as the model drift decreases, so does $\epsilon(i)$ on average. Therefore, with smaller model drift we can achieve higher privacy with more certainty.

5 Experimental analysis

We conduct a series of experiments to study the influence of privatization on the GFL algorithm. The aim of the experiments is to show the superior performance of graph-homomorphic perturbations to random perturbations and perturbations to gradients versus models, and to study the effect of different parameters on the performance of the algorithm.

5.1 Regression

We first start by studying a regression problem on simulated data. We do so for the tractability of the problem. We consider the quadratic loss that has a closed form solution, i.e., a formal expression for the true model w^o is known, which makes the calculation of the mean square error feasible and more accurate.

Therefore, consider a streaming feature vector $\boldsymbol{u}_{p,k,n} \in \mathbb{R}^M$ with output variable $\boldsymbol{d}_{p,k}(n) \in \mathbb{R}$ given by:

$$\boldsymbol{d}_{p,k}(n) = \boldsymbol{u}_{p,k,n}^{\mathsf{T}} w^{\star} + \boldsymbol{v}_{p,k}(n), \tag{63}$$

where $w^{\star} \in \mathbb{R}^M$ is some generating model, and $\boldsymbol{v}_{p,k}(n)$ is some zero-mean Guassian random variable with $\sigma^2_{v_{p,k}}$ variance and independent of $\boldsymbol{u}_{p,k,n}$. Then, the optimal model that solves the following problem:

$$\min_{w} \frac{1}{P}\sum_{p=1}^{P}\frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_{p,k}}\sum_{n=1}^{N_{p,k}} \|\boldsymbol{d}_{p,k}(n) - \boldsymbol{u}_{p,k,n}^{\mathsf{T}} w\|^2 + \rho\|w\|^2 \tag{64}$$

is found to be:

$$w^o = (\widehat{R}_u + \rho I)^{-1}(\widehat{R}_u w^{\star} + \widehat{r}_{uv}), \tag{65}$$

where $\widehat{R}_u$ and $\widehat{r}_{uv}$ are defined as:

$$\widehat{R}_u \triangleq \frac{1}{P}\sum_{p=1}^{P}\frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_{p,k}}\sum_{n=1}^{N_k} \boldsymbol{u}_{p,k,n}\boldsymbol{u}_{p,k,n}^{\mathsf{T}}, \tag{66}$$

$$\widehat{r}_{uv} \triangleq \frac{1}{P}\sum_{p=1}^{P}\frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_{p,k}}\sum_{n=1}^{N_k} \boldsymbol{v}_{p,k}(n)\boldsymbol{u}_{p,k,n}. \tag{67}$$

We consider $P = 10$ units, each with $K = 100$ total agents. We assume, $N_{p,k} = 100$ for each agent. We randomly generate two-dimensional feature vectors $\boldsymbol{u}_{p,k}(n)$ from a Gaussian random vector with zero-mean and a randomly generated covariance matrix $R_{u_{p,k}}$. We then calculate the corresponding outputs according to (63). To make the data non-iid across agents, we assume the covariance matrix $R_{u_{p,k}}$ is different for each agent, as well as the variance $\sigma^2_{v_{p,k}}$ of the added noise. When running the algorithm, we assume

each unit samples at random $L = 11$ agents, and each agent runs $E_{p,k} \in [1, 10]$ epochs and uses a mini-batch of $B_{p,k} \in [5, 10]$ samples.

We compare three algorithms: the standard GFL algorithm, the privatized GFL algorithm with random perturbations, and the privatized GFL with homomorphic perturbations. We do not add noise between the clients and their server to focus on the effect of the perturbations between the servers. In the first set of simulations, we fix the step-size $\mu = 0.7$ and the regularization parameter $\rho = 0.1$. We fix the variance of the added noise for privatization in both schemes to $\sigma_g^2 = 0.1$. We then plot the mean-square-deviation (MSD) at each time step for the centroid model:

$$\text{MSD}_i \stackrel{\Delta}{=} \|\boldsymbol{w}_{c,i} - w^o\|^2, \tag{68}$$

as seen in Fig. 2. We observe that the privatized GFL with random perturbations has lower performance compared to the other two algorithms. While, using homomorphic perturbations does not result in such a decay in performance. Thus, our suggested scheme does a good job at tracking the performance of the original GFL algorithm, while not compromising with the privacy level.

We next study the extent of the effect of the noise on the model utility. Thus, we run a series of experiments with varying added noise $\sigma_g^2 = \{0.001, 0.01, 0.1, 1, 2, 10\}$ for the two privatized GFL algorithms. We plot the resulting MSD curves in Fig. 3a. We observe for a fixed step-size, as we increase the variance, the MSD of the algorithm with random perturbations increases significantly as opposed to the algorithm with homomorphic perturbations. Thus, we conclude that the algorithm with random perturbations is more sensitive to the variance of the added noise. In fact, at some point, while using random perturbations, for some variance, the algorithm breaks down. While using graph-homomorphic perturbations, delays that effect for much larger variance. In addition, as long as the step-size is small enough, we can always control the effect of the graph-homomorphic perturbations.

However, if we were to look at the individual MSD for one federated unit, we would discover that the performance of the algorithm decays as the noise variance is increased. Nonetheless, it is not to the extent of random perturbations. We plot in Fig. 3b the average individual MSD for the varying noise variance:

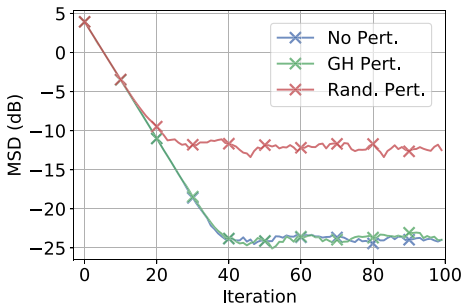
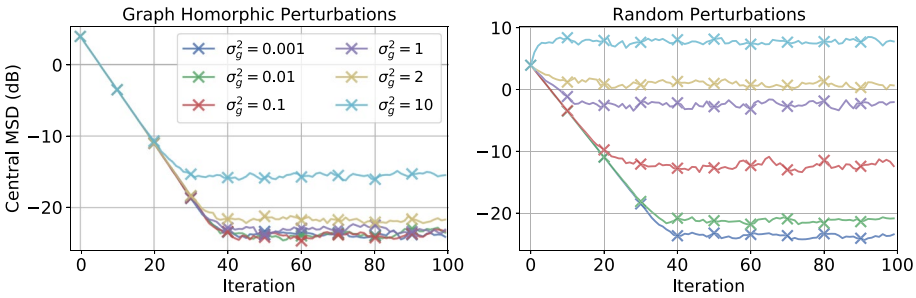
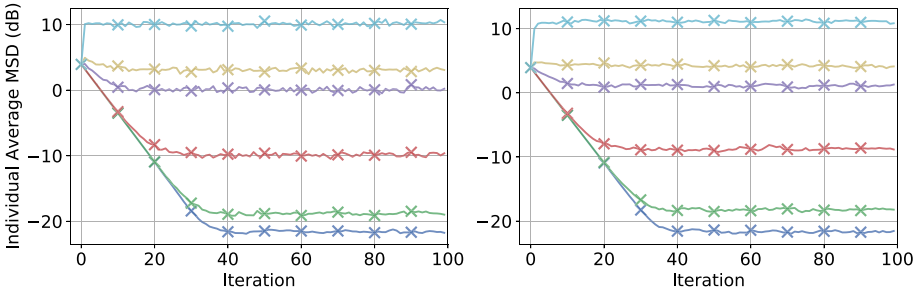


Fig. 2 Performance of GFL with no perturbations (blue), with graph-homomorphic perturbations (green), and random perturbations (red)



(a) centroid model



(b) individual models

Fig. 3 Performance curves of privatized GFL with varying noise variance

$$\text{MSD}_{\text{avg},i} \triangleq \frac{1}{P}\sum_{p=1}^{P} \|\boldsymbol{w}_{p,i} - w^o\|^2. \tag{69}$$

We observe that for a fixed noise variance, homomorphic perturbations results in a better performance. Furthermore, as we increase the noise variance, the network disagreement increases for both schemes. This comes as no surprise and is in accordance with Lemma 3. Furthermore, as previously mentioned, graph-homomorphic perturbations have the added value of not being negatively affected by the decrease in the step-size. In addition, even though the improvement does not seem significant, the source of the error of the two schemes is different. Furthermore, the information of the true model is distributed in the network and can be retrieved by running at the end of the learning algorithm a consensus-type step. At that point, the local models no longer contain information about the local data, and thus agents can safely share their models. However, when random perturbations are used, reconstruction is not possible since the information has been lost in the network due to the added perturbations.

We next fix the noise variance $\sigma_g^2 = 0.1$ and varying the step-size $\mu = \{0.1, 0.5, 1, 5\}$. According to Theorem 4, the MSD resulting from random perturbations includes an $O(\mu^{-1})$ term, which is not the case when using graph-homomorphic perturbations. Thus, we expect a decrease in the step-size will not significantly affect the privatized algorithm with graph-homomorphic perturbations as opposed to random perturbations. Indeed, as seen in Fig. 4, as μ is increased, the final MSD increases; this is probably due to the $O(\mu)\sigma_s^2$ term in the bound. In contrast, for significantly small or large μ, the performance of the privatized algorithm with random perturbations decreases.

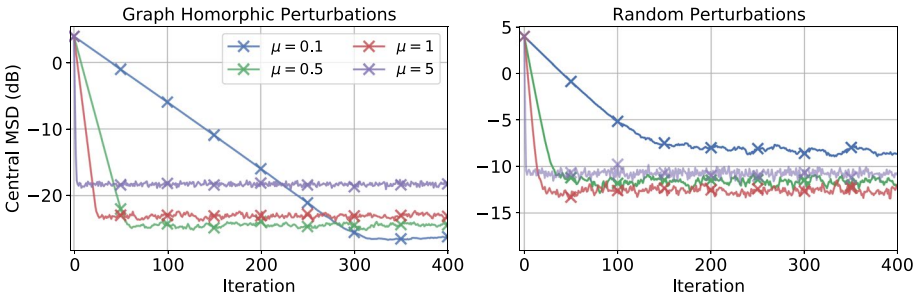


Fig. 4 Performance curves of privatized GFL with varying step-size

In addition, what we observe for both privacy schemes, is that the rate of convergence slows down as we decrease the step-size. Thus, there exists an optimal step-size that achieves a good compromise between a fast convergence and a low MSD.

5.2 Privatized federated learning

We focus on the single server FL setting (i.e., $P = 1$), where we assume we have $K = 1000$ agents of which we choose $L = 30$ at a time. We generate non-iid datasets of varying size for each agent as in the previous section. We allow each agent to run varying epochs $E_k \in [1, 10]$ during an iteration of the algorithm. We set the step-size $\mu = 0.2$, $\rho = 0.007$ and $\sigma_g^2 = 0.02$. We compare three algorithms: the standard FL algorithm, the privatized FL algorithm with sharing of models, and the privatized FL algorithm with sharing of updates. We plot the average MSD curves after repeating the experiment 100 times. As expected, the effect of the added noise is worse when models are shared (yellow curve Fig. 5) than when updates are shared (red curve Fig. 5).

We next study the effect of the step-size on the MSD of the privatized FL algorithm. We expect that as μ is increased the MSD increases for the FL algorithm when updates are shared. While, when models are shared, since the gradient noise variance is tuned by μ and the added noise variance by μ^{-1}, we expect to observe a trade-off. On one hand, as μ is increased the effect of the gradient noise is increased while that of the added noise is diminished. On the other hand, as μ is decreased, the effect of the added noise overpowers that of the gradient noise. Indeed, we observe this phenomenon in (a) and (b) of Fig. 6.

Finally, we study the effect of the variance of the added noise. We fix the step-size at $\mu = 0.2$ and vary the noise variance $\sigma_g^2 = \{0.01, 0.05, 0.1, 0.5\}$. In the two cases, as we increase σ_g^2 the performance diminishes ((c), (d) of Fig. 6). However, the larger values of

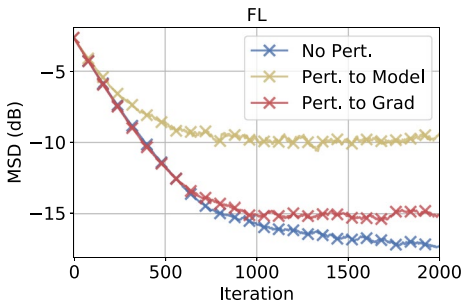
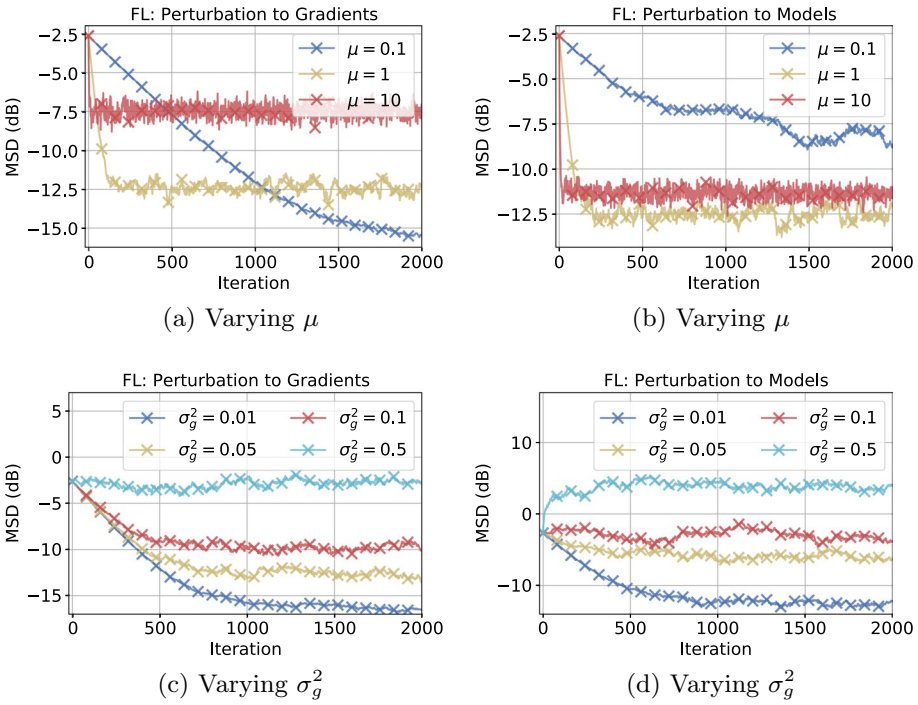


Fig. 5 MSD plots of FL



(a) Varying μ (b) Varying μ

(c) Varying σ_g^2 (d) Varying σ_g^2

Fig. 6 MSD plots of privatized FL with varying step-size and variance of added noise

the added noise variance affect the perturbed models more than the perturbed gradients. The algorithm diverges for lower values of σ_g^2 in the case when models are shared as opposed to when gradients are shared. Thus, sharing updates can handle larger values of σ_g^2 before the algorithm diverges. In addition, since the variance is tuned by the step-size, we can always find a suitable μ to decrease its effect.

5.3 Classification

We now focus on a classification problem applied to a dataset on click rate prediction of ads. We consider the Avazu click through dataset [29]. We split the 5101 data unequally among a total of 50 agents. We assume there are $P = 5$ units each with $K = 10$ agents. We add non-idd noise to the data at each agent to change their distributions. We again compare three algorithms: standard GFL, privatized GFL with homomorphic perturbations, and privatized GFL with random perturbations. We use a regularized logistic risk with regularization parameter $\rho = 0.03$. We set the step-size $\mu = 0.5$. We repeat the algorithms for multiple levels of privacy. We then settle on a noise variance $\sigma_g^2 = 0.6$ for which the privatized algorithm with random perturbations still converges. We plot in Fig. 7 the testing error on a set of 256 clean samples that were not perturbed with noise to change their distributions. We use the centriod model learned during each iteration to calculate the corresponding testing error. We observe that the graph-homomorphic perturbations do not hinder the performance of the privatized model. As for random perturbations, they significantly reduce the utility of the learnt model.

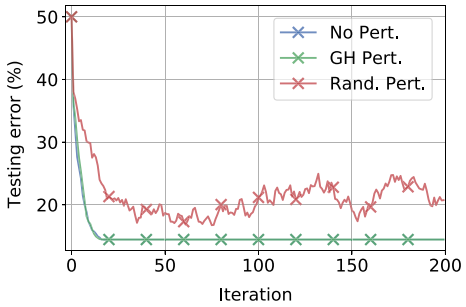


Fig. 7 Testing error of GFL with no perturbations (blue), with graph-homomorphic perturbations (green), and random perturbations (red)

6 Conclusion

In this work, we introduced graph federated learning and implemented an algorithm that guarantees privacy of the data in a differential privacy sense. We showed general privatization based on adding random perturbations to updates in federated learning have a negative effect on the performance of the algorithm. Random perturbations drive the algorithm farther away from the true optimal model. However, we showed by adding graph-homomorphic perturbations, which exploit the graph structure, performance can be recovered with guaranteed privacy. We also showed that using dependent perturbations does not result in the same trade-off between privacy and efficiency. In federated learning, we proved that sharing perturbed gradients versus perturbed models significantly reduces the effect of the added noise on the model utility. Thus, we no longer have to choose what to prioritize, and instead, we can have both a highly privatized algorithm with a good model utility.

Appendix 1: Secondary result on individual MSE performance

We first introduce the following theorem, which will be used to bound the network disagreement. We loosely bound the individual MSE for each federated unit. A tighter bound can be found, however, it is not needed.

Theorem 4 (*Individual MSE convergence*) *Under Assumptions 1, 2 and 3, the individual models converge to the optimal model w^o exponentially fast for a sufficiently small step-size*:

$$\begin{aligned}&\operatorname{col}\{\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,i}\|^2\}_{p=1}^P\\&\preceq \Lambda^i \operatorname{col}\{\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,0}\|^2\}_{p=1}^P + \sum_{j=0}^{i}\Lambda^j \operatorname{col}\{\mu^2\sigma_{s,p}^2 + O(\mu^2)\xi^2 + O(\mu^3)\sigma_{q,p}^2 + \sigma_{g,p}^2\}_{p=1}^P,\end{aligned} \tag{70}$$

where $\preceq$ is the elementwise comparison, Λ is a diagonal matrix with the p^{th} entry given by $\lambda_p = \sqrt{1-2\nu\mu+\delta^2\mu^2} + \beta_{s,p}^2\mu^2 + O(\mu^2) \in (0,1)$, $\sigma_{q,p}^2$ the average of $\sigma_{q,p,k}^2$, and $\sigma_{g,p}^2$ is the total variance introduced by the noise added at server p. Then, taking the limit of i to infinity:

$$\begin{aligned}&\limsup_{i\to\infty}\ \operatorname{col}\left\{\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,i}\|^2\right\}_{p=1}^{P}\\&\quad\preceq (I-\Lambda)^{-1}\ \operatorname{col}\left\{\mu^2\sigma_{s,p}^2+O(\mu^2)\xi^2+O(\mu^3)\sigma_{q,p}^2+\sigma_{g,p}^2\right\}_{p=1}^{P}.\end{aligned}\tag{71}$$

Proof Focusing on the error of a single server p, we can verify that:

$$\begin{aligned}&\mathbb{E}\{\|\widetilde{\boldsymbol{w}}_{p,i}\|^2|\mathcal{F}_{i-1}\}\\&\overset{(a)}{=}\mathbb{E}\left\{\Big\|\sum_{m\in\mathcal{N}_p}a_{pm}\big(\widetilde{\boldsymbol{w}}_{m,i-1}+\mu\nabla_w{}^{\mathsf{T}}J_m(\boldsymbol{w}_{m,i-1})+\mu\boldsymbol{q}_{m,i}\big)\Big\|^2\Big|\mathcal{F}_{i-1}\right\}\\&\quad+\mu^2\mathbb{E}\left\{\Big\|\sum_{m\in\mathcal{N}_p}a_{pm}\boldsymbol{s}_{m,i}\Big\|^2\Big|\mathcal{F}_{i-1}\right\}+\mathbb{E}\left\{\Big\|\sum_{m\in\mathcal{N}_p}\frac{a_{pm}}{L}\sum_{k\in\mathcal{L}_{m,i}}\boldsymbol{g}_{m,k,i}\Big\|^2\Big|\mathcal{F}_{i-1}\right\}\\&\quad+\mathbb{E}\left\{\Big\|\sum_{m\in\mathcal{N}_p}a_{pm}\boldsymbol{g}_{pm,i}\Big\|^2\Big|\mathcal{F}_{i-1}\right\},\\&\overset{(b)}{\leq}\sum_{m\in\mathcal{N}_p}a_{pm}\bigg(\frac{1}{\alpha}\|\widetilde{\boldsymbol{w}}_{m,i-1}+\mu\nabla_w{}^{\mathsf{T}}J_m(\boldsymbol{w}_{m,i-1})\|^2+\frac{\mu^2}{1-\alpha}\big(O(\mu)\|\widetilde{\boldsymbol{w}}_{m,i-1}\|^2\\&\quad+O(\mu)\xi^2+O(\mu^2)\sigma_{q,m}^2\big)+\mu^2\Big(\sigma_{s,m}^2+\beta_{s,m}^2\|\widetilde{\boldsymbol{w}}_{m,i-1}\|^2\Big)+\frac{1}{LK}\sum_{k=1}^{K}\mathbb{E}\|\boldsymbol{g}_{m,k,i}\|^2\\&\quad+\mathbb{E}\|\boldsymbol{g}_{pm,i}\|^2\bigg),\\&\overset{(c)}{\leq}\sum_{m\in\mathcal{N}_p}a_{pm}\bigg(\bigg(\frac{1-2\nu\mu+\delta^2\mu^2}{\alpha}+\beta_{s,m}^2\mu^2+\frac{O(\mu^3)}{1-\alpha}\bigg)\|\widetilde{\boldsymbol{w}}_{m,i-1}\|^2+\mu^2\sigma_{s,m}^2\\&\quad+\frac{O(\mu^3)\xi^2+O(\mu^4)\sigma_{q,m}^2}{1-\alpha}+\frac{1}{LK}\sum_{k=1}^{K}\mathbb{E}\|\boldsymbol{g}_{m,k,i}\|^2+\mathbb{E}\|\boldsymbol{g}_{pm,i}\|^2\bigg),\end{aligned}\tag{72}$$

where we define $\sigma_{q,m}^2$ to be the average of $\sigma_{q,m,k}^2$. Step (a) follows from independence of random variables and the zero-mean of the gradient noise and the added noise, (b) from Jensens' inequality and the bound on the gradient noise (24) and the incremental noise (37), (c) from ν-strong convexity and δ-Lipschtz continuity. Then, choosing $\alpha=\sqrt{1-2\nu\mu+\delta^2\mu^2}=1-O(\mu)$ and taking the expectation over the filtration, we get:

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,i}\|^2\leq\sum_{m\in\mathcal{N}_p}a_{pm}\Big(\lambda_m\mathbb{E}\|\widetilde{\boldsymbol{w}}_{m,i-1}\|^2+\mu^2\sigma_{s,m}^2+O(\mu^2)\xi^2+O(\mu^3)\sigma_{q,m}^2+\sigma_{g,m}^2\Big),\tag{73}$$

where we introduce the constants λ_m and $\sigma_{g,m}^2$:

$$\lambda_m \triangleq \sqrt{1-2\nu\mu+\delta^2\mu^2}+\beta_{s,m}^2\mu^2+O(\mu^2). \tag{74}$$

Next, taking the column vector of every local mean-square-error, we get the following bound in which we drop the indexing from the column vectors:

$$\begin{aligned}
&\operatorname{col}\left\{\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,i}\|^2\right\}\\
&\preceq \Lambda A\ \operatorname{col}\left\{\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,i-1}\|^2\right\}+A\ \operatorname{col}\left\{\mu^2\sigma_{s,p}^2+\sigma_{g,p}^2+O(\mu^2)\xi^2\right\}+A\ \operatorname{col}\left\{O(\mu^3)\sigma_{q,p}^2\right\},\\
&\preceq \Lambda^i A^i \operatorname{col}\left\{\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,0}\|^2\right\}+\sum_{j=0}^{i}\Lambda^j A^j \operatorname{col}\left\{\mu^2\sigma_{s,p}^2+\sigma_{g,p}^2\right\}\\
&\quad+\Lambda^j A^j \operatorname{col}\left\{O(\mu^2)\xi^2+O(\mu^3)\sigma_{q,p}^2\right\},\\
&\preceq \Lambda^i \operatorname{col}\left\{\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,0}\|^2\right\}+\sum_{j=0}^{i}\Lambda^j \operatorname{col}\left\{\mu^2\sigma_{s,p}^2+\sigma_{g,p}^2+O(\mu^2)\xi^2+O(\mu^3)\sigma_{q,p}^2\right\},
\end{aligned} \tag{75}$$

where we define the diagonal matrix Λ with λ_p as entries on the diagonal. Then choosing μ small enough such that $\lambda_p < 1$ for every p, we know the limit of Λ^i as i goes to infinity is zero. Furthermore, if the eigenvalues of Λ are less than 1, which they are, then the geometric series converges to $(I-\Lambda)^{-1}$. Thus, we get the desired result. □

Appendix 2: Proof of Lemma 3

Consider the aggregate model vector, i.e., $\boldsymbol{\mathcal{W}}_i \triangleq \operatorname{col}\left\{\boldsymbol{w}_{p,i}\right\}_{p=1}^{P}$, for which we write the model recursion as:

$$\begin{aligned}
\boldsymbol{\mathcal{W}}_i = (A\otimes I)^{\mathsf{T}}\Bigg(&\boldsymbol{\mathcal{W}}_{i-1}-\mu\operatorname{col}\left\{\nabla_{w^{\mathsf{T}}}J_p(\boldsymbol{w}_{p,i-1})+\boldsymbol{s}_{p,i}+\boldsymbol{q}_{p,i}\right\}\\
&+\operatorname{col}\left\{\frac{1}{L}\sum_{k\in\mathcal{L}_{p,i}}\boldsymbol{g}_{p,k,i}\right\}\Bigg)+\operatorname{diag}\big((A\otimes I)^{\mathsf{T}}\boldsymbol{\mathcal{G}}_i\big),
\end{aligned} \tag{76}$$

where $\boldsymbol{\mathcal{G}}_i$ is a matrix whose entries are the noise $\boldsymbol{g}_{pm,i}$, and the diag$(\cdot)$ function extracts the diagonal entries of a matrix and transforms them into a column vector.

Since A is doubly-stochastic, then it admits an eigendecomposition of the form $A = QHQ^{\mathsf{T}}$, with the first eigenvalue equal to 1 and its corresponding eigenvector equal to $\mathbb{1}/\sqrt{P}$.

Next, we define the extended centroid model $\boldsymbol{\mathcal{W}}_{c,i} \triangleq \left(\frac{1}{P}\mathbb{1}\mathbb{1}^{\mathsf{T}}\otimes I\right)\boldsymbol{\mathcal{W}}_i$, and write:

$$\begin{aligned}\boldsymbol{\mathcal{W}}_i - \boldsymbol{\mathcal{W}}_{c,i} &= \left(I - \frac{1}{P}\mathbb{1}\mathbb{1}^{\mathsf{T}} \otimes I\right)\boldsymbol{\mathcal{W}}_i \\ &= \left((Q^{\mathsf{T}} \otimes I)(Q \otimes I) - \frac{1}{P}\mathbb{1}\mathbb{1}^{\mathsf{T}} \otimes I\right)\boldsymbol{\mathcal{W}}_i \\ &= (Q_\epsilon^{\mathsf{T}} \otimes I)(Q_\epsilon \otimes I)\boldsymbol{\mathcal{W}}_i \\ &= (Q_\epsilon^{\mathsf{T}} \otimes I)H_\epsilon(Q_\epsilon \otimes I)\Bigg(\boldsymbol{\mathcal{W}}_{i-1} - \mu\ \mathrm{col}\left\{\nabla_{w^{\mathsf{T}}} J_p(\boldsymbol{w}_{p,i-1}) + \boldsymbol{s}_{p,i} + \boldsymbol{q}_{p,i}\right\} \\ &\quad + (Q_\epsilon^{\mathsf{T}} \otimes I)(Q_\epsilon \otimes I)\mathrm{diag}\Big((A \otimes I)^{\mathsf{T}}\boldsymbol{\mathcal{G}}_i\Big) + \ \mathrm{col}\left\{\frac{1}{L}\sum_{k\in\mathcal{L}_{p,i}} \boldsymbol{g}_{p,k,i}\right\}\Bigg).\end{aligned} \tag{77}$$

Then, taking the conditional expectation given the past models of $\|(Q_\epsilon \otimes I)\boldsymbol{\mathcal{W}}_i\|^2$, we can split the gradient noise and the added privacy noise from the model and the true gradient. Taking again the expectation over the past data, and then using the sub-multiplicity property of the norm followed by Jensen's inequality, we have:

$$\begin{aligned}&\mathbb{E}\|(Q_\epsilon \otimes I)\boldsymbol{\mathcal{W}}_i\|^2 \\ &\le \|H_\epsilon\|^2\Bigg(\mathbb{E}\Big\|(Q_\epsilon \otimes I)\boldsymbol{\mathcal{W}}_{i-1} - (Q_\epsilon \otimes I)\mu\ \mathrm{col}\left\{\nabla_{w^{\mathsf{T}}} J_p(\boldsymbol{w}_{p,i-1}) + \boldsymbol{q}_{p,i}\right\}\Big\|^2 \\ &\quad + \mu^2\|Q_\epsilon \otimes I\|^2\sum_{p=1}^{P}\mathbb{E}\|\boldsymbol{s}_{p,i}\|^2 + \|Q_\epsilon \otimes I\|^2\sum_{p=1}^{P}\mathbb{E}\left\|\frac{1}{L}\sum_{k\in\mathcal{L}_{p,i}}\boldsymbol{g}_{p,k,i}\right\|^2\Bigg) \\ &\quad + \|Q_\epsilon \otimes I\|^2\mathbb{E}\|\mathrm{diag}\Big((A \otimes I)^{\mathsf{T}}\boldsymbol{\mathcal{G}}_i\Big)\|^2 \\ &\le \|H_\epsilon\|^2\Bigg(\frac{1}{\|H_\epsilon\|}\mathbb{E}\|(Q_\epsilon \otimes I)\boldsymbol{\mathcal{W}}_{i-1}\|^2 + \frac{\mu^2\|Q_\epsilon \otimes I\|^2}{1 - \|H_\epsilon\|}\sum_{p=1}^{P}\mathbb{E}\|\nabla_{w^{\mathsf{T}}} J_p(\boldsymbol{w}_{p,i-1}) + \boldsymbol{q}_{p,i}\|^2 \\ &\quad + \mu^2\|Q_\epsilon \otimes I\|^2\sum_{p=1}^{P}\mathbb{E}\|\boldsymbol{s}_{p,i}\|^2 + \|Q_\epsilon \otimes I\|^2\sum_{p=1}^{P}\mathbb{E}\left\|\frac{1}{L}\sum_{k\in\mathcal{L}_{p,i}}\boldsymbol{g}_{k,p,i}\right\|^2\Bigg) \\ &\quad + \|Q_\epsilon \otimes I\|^2\mathbb{E}\|\mathrm{diag}\Big((A \otimes I)^{\mathsf{T}}\boldsymbol{\mathcal{G}}_i\Big)\|^2.\end{aligned} \tag{78}$$

Next, we focus on each individual term. Using Jensen for some constant α and then the Lipschitz condition and the bound on the incremental noise, we can bound the below norm as follows:

$$\begin{aligned}\mathbb{E}\|\nabla_{w^{\mathsf{T}}} J_p(\boldsymbol{w}_{p,i-1}) + \boldsymbol{q}_{p,i}\|^2 &\le \frac{2}{\alpha}\Big(\delta^2\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,i-1}\|^2 + \|\nabla_{w^{\mathsf{T}}} J_p(w^o)\|^2\Big) \\ &\quad + \frac{1}{1-\alpha}\Big(O(\mu)\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,i-1}\|^2 + O(\mu)\xi^2 + O(\mu^2)\sigma_{q,p}^2\Big).\end{aligned} \tag{79}$$

Using the bound on the gradient noise (24), we get another $\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,i-1}\|^2$ term, which can be bounded by the result in Theorem 4. Thus, we write:

$$
\begin{aligned}
&\frac{1}{1-\|H_\epsilon\|}\mathbb{E}\|\nabla_{w^\intercal}J_p(\boldsymbol{w}_{p,i-1})+\boldsymbol{q}_{p,i}\|^2+\mathbb{E}\|\boldsymbol{s}_{p,i}\|^2\\
&\quad\le\left(\frac{2\delta^2}{\alpha(1-\|H_\epsilon\|)}+\beta_{s,p}^2+\frac{O(\mu)}{(1-\alpha)(1-\|H_\epsilon\|)}\right)\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,i-1}\|^2+\frac{2\|\nabla_{w^\intercal}J_p(w^o)\|^2}{\alpha(1-\|H_\epsilon\|)}\\
&\qquad+\sigma_{s,p}^2+\frac{O(\mu)\xi^2+O(\mu^2)\sigma_{q,p}^2}{(1-\alpha)(1-\|H_\epsilon\|)}\\
&\quad\le\left(\frac{2\delta^2}{\alpha(1-\|H_\epsilon\|)}+\beta_{s,p}^2+\frac{O(\mu)}{(1-\alpha)(1-\|H_\epsilon\|)}\right)\left(\lambda_p^i A^i[p]\,\mathrm{col}\left\{\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,0}\|^2\right\}\right.\\
&\qquad\left.+\sum_{j=0}^{i-1}\lambda_p^j A^j[p]\,\mathrm{col}\left\{\mu^2\sigma_{s,p}^2+O(\mu^2)\xi^2+O(\mu^3)\sigma_{q,p}^2+\sigma_{g,p}^2\right\}\right)\\
&\qquad+\frac{2\|\nabla_{w^\intercal}J_p(w^o)\|^2}{\alpha(1-\|H_\epsilon\|)}+\sigma_{s,p}^2+\frac{O(\mu)\xi^2+O(\mu^2)\sigma_{q,p}^2}{(1-\alpha)(1-\|H_\epsilon\|)}.
\end{aligned}
\tag{80}
$$

The noise term can be witten in a more compact way, $\|Q_\epsilon\otimes I\|^2\sum_{p=1}^{P}\sigma_{g,p}^2$. Thus, putting everything together, we get:

$$
\begin{aligned}
&\mathbb{E}\|(Q_\epsilon\otimes I)\boldsymbol{\mathcal{W}}_i\|^2\\
&\quad\le\|H_\epsilon\|\mathbb{E}\|(Q_\epsilon\otimes I)\boldsymbol{\mathcal{W}}_{i-1}\|^2+\mu^2\|Q_\epsilon\otimes I\|^2\|H_\epsilon\|^2\sum_{p=1}^{P}\left(\left(\frac{2\delta^2}{\alpha(1-\|H_\epsilon\|)}\right.\right.\\
&\qquad\left.+\beta_{s,p}^2+\frac{O(\mu)}{(1-\alpha)(1-\|H_\epsilon\|)}\right)\left(\lambda_p^i A^i[p]\,\mathrm{col}\left\{\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,0}\|^2\right\}+\sum_{j=0}^{i-1}\lambda_p^j A^j[p]\right.\\
&\qquad\left.\times\ \mathrm{col}\left\{\mu^2\sigma_{s,p}^2+O(\mu^2)\xi^2+O(\mu^3)\sigma_{q,p}^2+\sigma_{g,p}^2\right\}\right)+\frac{2\|\nabla_{w^\intercal}J_p(w^o)\|^2}{\alpha(1-\|H_\epsilon\|)}\\
&\qquad\left.+\sigma_{s,p}^2+\frac{O(\mu)\xi^2+O(\mu^2)\sigma_{q,p}^2}{(1-\alpha)(1-\|H_\epsilon\|)}\right)+\|Q_\epsilon\otimes I\|^2\sum_{p=1}^{P}\sigma_{g,p}^2\\
&\quad\le\|H_\epsilon\|^i\mathbb{E}\|(Q_\epsilon\otimes I)\boldsymbol{\mathcal{W}}_0\|^2+\sum_{j'=0}^{i-1}\|H_\epsilon\|^{j'+2}\|Q_\epsilon\otimes I\|^2\left\{\mu^2\sum_{p=1}^{P}\left(\left(\frac{2\delta^2}{\alpha(1-\|H_\epsilon\|)}\right.\right.\right.\\
&\qquad\left.+\beta_{s,p}^2+\frac{O(\mu)}{(1-\alpha)(1-\|H_\epsilon\|)}\right)\left(\lambda_p^{j'}A^{j'}[p]\,\mathrm{col}\left\{\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,0}\|^2\right\}+\sum_{j=0}^{j'-1}\lambda_p^j A^j[p]\right.\\
&\qquad\left.\times\ \mathrm{col}\left\{\mu^2\sigma_{s,p}^2+O(\mu^2)\xi^2+O(\mu^3)\sigma_{q,p}^2+\sigma_{g,p}^2\right\}\right)+\frac{2\|\nabla_{w^\intercal}J_p(w^o)\|^2}{\alpha(1-\|H_\epsilon\|)}\\
&\qquad\left.\left.+\sigma_{s,p}^2+\frac{O(\mu)\xi^2+O(\mu^2)\sigma_{q,p}^2}{(1-\alpha)(1-\|H_\epsilon\|)}\right)+\frac{1}{\|H_\epsilon\|^2}\sum_{p=1}^{P}\sigma_{g,p}^2\right\}.
\end{aligned}
\tag{81}
$$

Going back to the network disagreement, it is bounded by the above bound multiplied by $\|Q_\epsilon^{\mathsf{T}}\otimes I\|^2/P$. If we were to drive i to infinity, since $\|H_\epsilon\|=\iota_2<1$, with ι_2 being the second eigenvalue of A, and choosing $\alpha=\iota_2$ we would have:

$$
\begin{aligned}
&\limsup_{i\to\infty}\frac{1}{P}\sum_{p=1}^{P}\mathbb{E}\|\boldsymbol{w}_{p,i}-\boldsymbol{w}_{c,i}\|^2\\
&\le\frac{\|Q_\epsilon\otimes I\|^4\iota_2^2}{P}\Bigg\{\mu^2\sum_{p=1}^{P}\Bigg(\Bigg(\frac{2\delta^2}{\iota_2(1-\iota_2)}+\beta_{s,p}^2+\frac{O(\mu)}{(1-\iota_2)^2}\Bigg)\sum_{j'=0}^{\infty}\iota_2^{j'}\sum_{j=0}^{j'-1}\lambda_p^j A^j[p]\\
&\quad\times\ \mathrm{col}\left\{\mu^2\sigma_{s,p}^2+O(\mu^2)\xi^2+O(\mu^3)\sigma_{q,p}^2+\sigma_{g,p}^2\right\}+\frac{2\|\nabla_{w^\intercal}J_p(w^o)\|^2}{\iota_2(1-\iota_2)^2}\\
&\quad+\frac{\sigma_{s,p}^2}{1-\iota_2}+\frac{O(\mu)\xi^2+O(\mu^2)\sigma_{q,p}^2}{(1-\iota_2)^3}\Bigg)+\frac{1}{(1-\iota_2)\iota_2^2}\sum_{p=1}^{P}\sigma_{g,p}^2\Bigg\}\\
&\le\frac{\iota_2^2}{P}\Bigg\{\mu^2\sum_{p=1}^{P}\Bigg(\Bigg(\frac{2\delta^2}{\iota_2(1-\iota_2)}+\beta_{s,p}^2+\frac{O(\mu)}{(1-\iota_2)^2}\Bigg)\\
&\quad\times\sum_{m\in\mathcal{N}_p}\frac{\iota_2(\mu^2\sigma_{s,m}^2+O(\mu^2)\xi^2+O(\mu^3)\sigma_{q,m}^2+\sigma_{g,m}^2)}{1-\iota_2\lambda_p a_{pm}}+\frac{2\|\nabla_{w^\intercal}J_p(w^o)\|^2}{\iota_2(1-\iota_2)^2}\\
&\quad+\frac{\sigma_{s,p}^2}{1-\iota_2}+\frac{O(\mu)\xi^2+O(\mu^2)\sigma_{q,p}^2}{(1-\iota_2)^3}\Bigg)+\frac{1}{(1-\iota_2)\iota_2^2}\sum_{p=1}^{P}\sigma_{g,p}^2\Bigg\}\\
&=\frac{\iota_2^2}{P(1-\iota_2)}\sum_{p=1}^{P}\mu^2\sigma_{s,p}^2+\frac{1}{\iota_2^2}\sigma_{g,p}^2+O(\mu)\sigma_{g,p}^2+O(\mu^3).
\end{aligned}
\tag{82}
$$

Appendix 3: Proof of Theorem 1

First taking the conditional mean of the ℓ_2-norm of the centroid error given the past models, splits the mean into three independent terms: the centralized recursion, the gradient noise and the added noise. Then, taking the expectation again, we get:

$$
\begin{aligned}
&\mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i}\|^2\\
&\quad=\mathbb{E}\left\|\widetilde{\boldsymbol{w}}_{c,i-1}+\mu\frac{1}{P}\sum_{p=1}^{P}\nabla_{w^\intercal}J_p(\boldsymbol{w}_{p,i-1})+\mu\boldsymbol{q}_i\right\|^2+\mu^2\mathbb{E}\|\boldsymbol{s}_i\|^2+\mathbb{E}\|\boldsymbol{g}_{c,i}\|^2\\
&\quad\overset{(a)}{\le}\frac{1}{\alpha^2}\mathbb{E}\left\|\widetilde{\boldsymbol{w}}_{c,i-1}+\mu\frac{1}{P}\sum_{p=1}^{P}\nabla_{w^\intercal}J_p(\boldsymbol{w}_{c,i-1})\right\|^2+\frac{\mu^2}{1-\alpha}\mathbb{E}\|\boldsymbol{q}_i\|^2\\
&\qquad+\frac{\delta^2\mu^2}{\alpha(1-\alpha)P}\sum_{p=1}^{P}\mathbb{E}\|\boldsymbol{w}_{p,i-1}-\boldsymbol{w}_{c,i-1}\|^2+\mu^2\mathbb{E}\|\boldsymbol{s}_i\|^2+\mathbb{E}\|\boldsymbol{g}_{c,i}\|^2\\
&\quad\overset{(b)}{\le}\left(\frac{1}{\alpha^2}(1-2\nu\mu+\delta^2\mu^2)+\beta_s^2\mu^2+\frac{O(\mu^3)}{1-\alpha}\right)\mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2+\mu^2\sigma_s^2+\mathbb{E}\|\boldsymbol{g}_{c,i}\|^2\\
&\qquad+\left(\frac{\delta^2}{\alpha(1-\alpha)}+\frac{O(\mu^3)}{1-\alpha}+\beta_{s,max}^2\right)\frac{\mu^2}{P}\sum_{p=1}^{P}\mathbb{E}\|\boldsymbol{w}_{p,i-1}-\boldsymbol{w}_{c,i-1}\|^2\\
&\qquad+\frac{O(\mu^3)\xi^2+O(\mu^4)\sigma_q^2}{1-\alpha},
\end{aligned}
\tag{83}
$$

where inequality (a) follows from Jensen with constant $\alpha \in (0,1)$ and Lipshcitz, and (b) from applying Lemma 1. Then, choosing $\alpha = \sqrt[4]{1-2\nu\mu+\delta^2\mu^2} = 1 - O(\mu)$, the bound becomes:

$$\begin{aligned}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i}\|^2 \leq \lambda_c \mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + \mu^2\sigma_s^2 + \mathbb{E}\|\boldsymbol{g}_{c,i}\|^2 + O(\mu^2)\xi^2 + O(\mu^3)\sigma_q^2 \\ + \frac{O(\mu)}{P}\sum_{p=1}^{P}\mathbb{E}\|\boldsymbol{w}_{p,i-1} - \boldsymbol{w}_{c,i-1}\|^2.\end{aligned} \tag{84}$$

Finally, using the result on the network disagreement, recusrively bounding the error, and taking the limit of i, we get the final result:

$$\limsup_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{c,i}\|^2 \leq \frac{\mu^2\sigma_s^2 + \mathbb{E}\|\boldsymbol{g}_c\|^2 + O(\mu^2)\xi^2 + O(\mu^3)\sigma_q^2}{1-\lambda_c} + \sum_{p=1}^{P} O(1)\sigma_{g,p}^2 + O(\mu). \tag{85}$$

Appendix 4: Secondary result involving a bound on $\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i\|^2$

To show the sensitivity of the algorithm is bounded with high probability, we require a bound on $\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i\|^2$ and $\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i'\|^2$. From Theorem 4 we can bound the individual errors by:

$$\begin{aligned}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,i}\|^2 &\leq \lambda_p\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,i-1}\|^2 + \mu^2\sigma_{s,p}^2 + O(\mu^2)\xi^2 + O(\mu^3)\sigma_{q,p}^2 + \sigma_{g,p}^2 \\ &\leq \lambda_{\max}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,i-1}\|^2 + \mu^2\sigma_{s,p}^2 + O(\mu^2)\xi^2 + O(\mu^3)\sigma_{q,p}^2 + \sigma_{g,p}^2 \\ &\leq \lambda_{\max}^i\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,0}\|^2 + \frac{1-\lambda_{\max}^i}{1-\lambda_{\max}}\left(\mu^2\sigma_{s,p}^2 + O(\mu^2)\xi^2 + O(\mu^3)\sigma_{q,p}^2 + \sigma_{g,p}^2\right) \\ &\leq \lambda_{\max}^i\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,0}\|^2 + O(\mu) + O(\mu^{-1}),\end{aligned} \tag{86}$$

where $\lambda_{\max} = \max_p \lambda_p$. Then, $\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i\|^2$ can be bounded as follows:

$$\begin{aligned}\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i\|^2 &= \sum_{p=1}^{P}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,i}\|^2 \\ &\leq \sum_{p=1}^{P}\lambda_{\max}^i\mathbb{E}\|\widetilde{\boldsymbol{w}}_{p,0}\|^2 + O(\mu) + O(\mu^{-1}) \\ &= \lambda_{\max}^i\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_0\|^2 + O(\mu) + O(\mu^{-1}).\end{aligned} \tag{87}$$

It follows that for some constants B and B', the probability that $\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i\|$ and $\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i'\|$ are unbounded can be bounded using Markov's inequality by:

$$\begin{aligned}\mathbb{P}(\|\widetilde{\boldsymbol{\mathcal{W}}}_i\| \geq B) &\leq \frac{\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i\|^2}{B^2} \\ &\leq \frac{\lambda_{\max}^i\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_0\|^2 + O(\mu) + O(\mu^{-1})}{B^2},\end{aligned} \tag{88}$$

and similarly for $\mathbb{P}(\|\widetilde{\boldsymbol{\mathcal{W}}}_i'\| \geq B')$.

Appendix 5: Proof of Theorem 3

To evaluate the probability distribution in Definition 1, we note that the randomness of the models $\boldsymbol{\psi}_{p,j}$ arises from the subsampling of the data for the calculation of the stochastic gradient at each iteration. Thus, given the subsampled dataset, the models are now deterministic and since the added noises $\boldsymbol{g}_{pm,j}$ are Laplacian random variables, the distribution of the added noise over the neighbourhood of agent p and over the iterations is given by:

$$\begin{aligned} f\left(\left\{\left\{\boldsymbol{\psi}_{p,j}+\boldsymbol{g}_{pm,j}\right\}_{m\in\mathcal{N}_p\setminus\{p\}}\right\}_{j=0}^{i}\right) &= f(\boldsymbol{y}_0)f(\boldsymbol{y}_1|\boldsymbol{y}_0)\cdots f(\boldsymbol{y}_i|\boldsymbol{y_0},\cdots,\boldsymbol{y}_{i-1}) \\ &= \prod_{j=0}^{i}\frac{1}{\sqrt{2}\sigma_g}\exp\left(-\frac{\sqrt{2}}{\sigma_g}\|\boldsymbol{\psi}_{p,j}+\boldsymbol{g}_{p,j}\|\right) \\ &= \frac{1}{\sqrt{2}\sigma_g}\exp\left(-\frac{\sqrt{2}}{\sigma_g}\sum_{j=0}^{i}\|\boldsymbol{\psi}_{p,j}+\boldsymbol{g}_{p,j}\|\right), \end{aligned} \tag{89}$$

where $\boldsymbol{y}_j=\left\{\boldsymbol{\psi}_{p,j}+\boldsymbol{g}_{pm,j}\right\}_{m\in\mathcal{N}_p\setminus\{p\}}$ and the ratio in Definition 1 is bounded with high probability:

$$\begin{aligned} &\exp\left(-\frac{\sqrt{2}}{\sigma_g}\sum_{j=0}^{i}\|\boldsymbol{\psi}_{p,j}+\boldsymbol{g}_{p,j}\|-\|\boldsymbol{\psi}'_{p,j}+\boldsymbol{g}_{p,j}\|\right) \\ &\le \exp\left(\frac{\sqrt{2}}{\sigma_g}\sum_{j=0}^{i}\|\boldsymbol{\psi}_{p,j}-\boldsymbol{\psi}'_{p,j}\|\right) \\ &\le \exp\left(\frac{\sqrt{2}}{\sigma_g}\sum_{j=0}^{i}\Delta(j)\right) \\ &\le \exp\left(\frac{\sqrt{2}}{\sigma_g}\sum_{j=0}^{i}(B+B'+\|w^o-w'^o\|)\right) \\ &= \exp\left(\frac{\sqrt{2}}{\sigma_g}(B+B'+\|w^o-w'^o\|)(i+1)\right), \end{aligned} \tag{90}$$

where the inequalities follow from the triangle inequality and the bound on the sensitivity of the algorithm.

Appendix 6: Proof of Theorem 2

We start by writing the error recursion:

$$\widetilde{\boldsymbol{w}}_{1,i}=\widetilde{\boldsymbol{w}}_{1,i-1}+\frac{\mu}{K}\sum_{k=1}^{K}\nabla_{w^{\mathsf{T}}}J_{1,k}(\boldsymbol{w}_{1,i-1})+\mu\boldsymbol{s}_{1,i}+\mu\boldsymbol{q}_{1,i}+\frac{\mu}{L}\sum_{k\in\mathcal{L}_{1,i}}\boldsymbol{g}_{1,k,i}, \tag{91}$$

where we introduce the gradient noise $\boldsymbol{s}_{1,i}$ and the incremental noise $\boldsymbol{q}_{1,i}$:

$$\boldsymbol{s}_{1,i} = \frac{1}{L}\sum_{k\in\mathcal{L}_{1,i}} \widehat{\nabla_{w^{\mathsf{T}}} J_{1,i}}(\boldsymbol{w}_{1,i-1}) - \frac{1}{K}\sum_{k=1}^{K} \widehat{\nabla_{w^{\mathsf{T}}} J_{1,k}}(\boldsymbol{w}_{1,i-1}), \tag{92}$$

$$\boldsymbol{q}_{1,i} = \frac{1}{L}\sum_{k\in\mathcal{L}_{1,i}} \frac{1}{E_{1,k}}\sum_{e=1}^{E_{1,k}} \widehat{\nabla_{w^{\mathsf{T}}} J_{1,k}}(\boldsymbol{w}_{1,k,e-1}) - \widehat{\nabla_{w^{\mathsf{T}}} J_{1,k}}(\boldsymbol{w}_{1,i-1}). \tag{93}$$

We have already shown in previous work that the gradient noise is zero-mean and has bounded second order-moment [28, Lemma 1], while the incremental noise has bounded second order-moment [28, Lemma 5]:

$$\mathbb{E}\{\|\boldsymbol{s}_{1,i}\|^2|\mathcal{F}_{i-1}\} \le \beta_{s,1}^2\|\widetilde{\boldsymbol{w}}_{1,i-1}\|^2 + \sigma_{s,1}^2, \tag{94}$$

$$\mathbb{E}\|\boldsymbol{q}_{1,i}\|^2 \le O(\mu)\mathbb{E}\|\widetilde{\boldsymbol{w}}_{1,i-1}\|^2 + O(\mu)\xi^2 + O(\mu^2)\sigma_{q,1}^2, \tag{95}$$

where the constants $\beta_{s,1}^2, \sigma_{s,1}^2, \sigma_{q,1}^2$ are given by:

$$\beta_{s,1}^2 = \frac{6\delta^2}{L}\left(1 + \frac{1}{K}\sum_{k=1}^{K}\frac{1}{E_{1,k}}\right), \tag{96}$$

$$\sigma_{s,1}^2 = \frac{1}{LK}\sum_{k=1}^{K}\left(\frac{12}{E_{1,k}} + 3\right)\frac{1}{N_{1,k}}\sum_{n=1}^{N_{1,k}}\|\nabla_{w^{\mathsf{T}}} Q_{1,k}(w^o; x_{1,k,n})\|^2, \tag{97}$$

$$\sigma_{q,1}^2 = \frac{3}{K}\sum_{k=1}^{K}\sum_{n=1}^{N_{1,k}}\|\nabla_{w^{\mathsf{T}}} Q_{1,k}(w_{1,k}^o; x_{1,k,n})\|^2. \tag{98}$$

Taking the conditional mean of the ℓ_2-norm of the error, we can split the noise term from the rest and then apply Jensen's inequality with some constant $\alpha \in (0,1)$:

$$\begin{aligned}
\mathbb{E}\{\|\widetilde{\boldsymbol{w}}_{1,i}\|^2|\mathcal{F}_{i-1},\mathcal{L}_{1,i}\} &= \mathbb{E}\Bigg\{\Bigg\|\widetilde{\boldsymbol{w}}_{1,i-1} + \frac{\mu}{K}\sum_{k=1}^{K}\nabla_{w^{\mathsf{T}}} J_{1,k}(\boldsymbol{w}_{1,i-1}) + \mu\boldsymbol{s}_{1,i} \\
&\quad + \mu\boldsymbol{q}_{1,i}\Bigg\|^2\Bigg|\mathcal{F}_{i-1},\mathcal{L}_{1,i}\Bigg\} + \frac{\mu^2}{L^2}\sum_{k\in\mathcal{L}_{1,i}}\mathbb{E}\|\boldsymbol{g}_{1,k,i}\|^2 \\
&\le \frac{1}{\alpha}\left\|\widetilde{\boldsymbol{w}}_{i-1} + \frac{\mu}{K}\sum_{k=1}^{K}\nabla_{w^{\mathsf{T}}} J_{1,k}(\boldsymbol{w}_{1,i-1})\right\|^2 \\
&\quad + \frac{\mu^2}{\alpha}\mathbb{E}\{\|\boldsymbol{s}_{1,i}\|^2|\mathcal{F}_{i-1},\mathcal{L}_{1,i}\} + \frac{\mu^2}{L}\sigma_{g,1}^2 \\
&\quad + \frac{\mu^2}{1-\alpha}\mathbb{E}\{\|\boldsymbol{q}_{1,i}\|^2|\mathcal{F}_{i-1},\mathcal{L}_{1,i}\}.
\end{aligned} \tag{99}$$

Using strong convexity and Lipschitz continuity of the functions we can bound the first term as:

$$\left\| \widetilde{\boldsymbol{w}}_{1,i-1} + \frac{\mu}{K} \sum_{k=1}^{K} \nabla_{w^\mathsf{T}} J_{1,k}(w_{1,i-1}) \right\|^2 \leq (1 - 2\nu\mu + \delta^2\mu^2) \|\widetilde{\boldsymbol{w}}_{1,i-1}\|^2. \tag{100}$$

Then, taking the expectations again over the past models and the selected agents, and using the bound on the gradient noise and incremental noise:

$$\begin{aligned} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{1,i}\|^2 \leq & \left(\frac{1 - 2\nu\mu + (\beta_{s,1}^2 + \delta^2)\mu^2}{\alpha} + \frac{O(\mu^3)}{1-\alpha} \right) \mathbb{E}\|\widetilde{\boldsymbol{w}}_{1,i-1}\|^2 + \frac{\mu^2}{\alpha}\sigma_{s,1}^2 \\ & + \frac{O(\mu^3)\xi^2 + O(\mu^4)\sigma_{q,1}^2}{1-\alpha} + \frac{\mu^2}{L}\sigma_{g,1}^2. \end{aligned} \tag{101}$$

Then, recursively bounding the error with $\alpha = \sqrt{1 - 2\nu\mu + (\beta_{s,1}^2 + \delta^2)\mu^2}$:

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_{1,i}\|^2 \leq \lambda^i \mathbb{E}\|\widetilde{\boldsymbol{w}}_{1,0}\|^2 + \frac{1-\lambda^i}{1-\lambda} \left(O(\mu^2)\sigma_{s,1}^2 + O(\mu^2)\xi^2 + \frac{\mu^2}{L}\sigma_{g,1}^2 + O(\mu^3)\sigma_{q,1}^2 \right), \tag{102}$$

and taking the limit of i:

$$\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{1,i}\|^2 \leq O(\mu)(\sigma_{s,1}^2 + \xi^2 + \sigma_{g,1}^2) + O(\mu^2)\sigma_{q,1}^2. \tag{103}$$

Abbreviations

GFL	Graph Federated Learning
FL	Federated Learning
FedAvg	Federated Averaging
SGD	Stochastic Gradient Descent
SMC	Secure Multiparty Computation
MSE	Mean-square-error
MSD	Mean-square-deviation

Acknowledgements
Not applicable.

Author contributions
ER is the first author and the corresponding author of this paper. Her main contributions include the mathematical analysis, derivation, simulations and writing. SV is the second author, and his contribution is conceptualization and reviewing. AHS is the third author, and his contribution is conceptualization, reviewing and supervision. All authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
The generated data is not available since it is randomly generated everytime the experiment is ran. The Avazu dataset used as a real world example can be found at http://www.csie.ntu.edu.tw/-cj1in/libsvmtools/.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 16 January 2023 Accepted: 15 August 2023
Published online: 25 August 2023

References

1. H.B. McMahan, E. Moore, D. Ramage, S. Hampson, Communication-efficient learning of deep networks from decentralized data, in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, vol. 54 (2017), pp. 1273–1282
2. B. Hitaj, G. Ateniese, F. Perez-Cruz, Deep models under the GAN: information leakage from collaborative deep learning, in *Proceedings of ACM SIGSAC Conference on Computer and Communications Security, New York, NY, USA* (2017), pp. 603–618
3. L. Melis, C. Song, E. De Cristofaro, V. Shmatikov, Exploiting unintended feature leakage in collaborative learning, in *IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA* (2019), pp. 691–706
4. M. Nasr, R. Shokri, A. Houmansadr, Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning, in *IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA* (2019), pp. 739–753
5. L. Zhu, S. Han, Deep leakage from gradients, in *Advances in Neural Information Processing Systems, Vancouver, Canada* (2019), pp. 17–31
6. S.Vlaski, A.H. Sayed, Graph-homomorphic perturbations for private decentralized learning, in *Proceedings of the ICASSP, Toronto, Canada* (2021), pp. 1–5
7. L. Liu, J. Zhang, S.H. Song, K.B. Letaief, Client-edge-cloud hierarchical federated learning, in *IEEE International Conference on Communications (ICC)* (2020), pp. 1–6
8. E. Rizk, A.H. Sayed, A graph federated architecture with privacy preserving learning, in *IEEE International Workshop on Signal Processing Advances in Wireless Communications, Lucca, Italy* (2021), pp. 1–5. arxiv:2104.13215
9. W. Liu, L. Chen, W. Zhang, Decentralized federated learning: balancing communication and computing costs. IEEE Trans. Signal Inf. Process. Over Netw. **8**, 131–143 (2022)
10. B. Wang, J. Fang, H. Li, X. Yuan, Q. Ling, Confederated learning: federated learning with decentralized edge servers. arXiv:2205.14905 (2022)
11. R.C. Geyer, T. Klein, M. Nabi, Differentially private federated learning: a client level perspective. arXiv:1712.07557 (2017)
12. R. Hu, Y. Guo, H. Li, Q. Pei, Y. Gong, Personalized federated learning with differential privacy. IEEE Internet Things J. **7**(10), 9530–9539 (2020)
13. A. Triastcyn, B. Faltings, Federated learning with Bayesian differential privacy, in *IEEE International Conference on Big Data, Los Angeles, California, USA* (2019), pp. 2587–2596
14. S. Truex, L. Liu, K.-H. Chow, M.E. Gursoy, W. Wei, LDP-FED: federated learning with local differential privacy, in *Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking* (2020), pp. 61–66
15. K. Wei, J. Li, M. Ding, C. Ma, H.H. Yang, F. Farokhi, S. Jin, T.Q. Quek, H.V. Poor, Federated learning with differential privacy: algorithms and performance analysis. IEEE Trans. Inf. Forensics Secur. **15**, 3454–3469 (2020)
16. B. Jayaraman, L. Wang, D. Evans, Q. Gu, Distributed learning without distress: privacy-preserving empirical risk minimization, in *Advances in Neural Information Processing Systems*, vol. 31. Montreal, Canada (2018)
17. C. Li, P. Zhou, L. Xiong, Q. Wang, T. Wang, Differentially private distributed online learning. IEEE Trans. Knowl. Data Eng. **30**(8), 1440–1453 (2018)
18. J. Zhu, C. Xu, J. Guan, D.O. Wu, Differentially private distributed online algorithms over time-varying directed networks. IEEE Trans. Signal Inf. Process. Over Netw. **4**(1), 4–17 (2018)
19. M.A. Pathak, S. Rane, B. Raj, Multiparty differential privacy via aggregation of locally trained classifiers, in *Advances in Neural Information Processing Systems, Vancouver, Canada* (2010), pp. 1876–1884
20. S. Gade, N.H. Vaidya, Private learning on networks. arXiv:1612.05236 (2016)
21. K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H.B. McMahan, S. Patel, D. Ramage, A. Segal, K. Seth, Practical secure aggregation for privacy-preserving machine learning, in *Proceedings of ACM SIGSAC Conference on Computer and Communications Security, New York, USA* (2017), pp. 1175–1191
22. A. Gascón, P. Schoppmann, B. Balle, M. Raykova, J. Doerner, S. Zahur, D. Evans, Privacy-preserving distributed linear regression on high-dimensional data. Proc. Priv. Enhanc. Technol. **2017**(4), 345–364 (2017)
23. P. Mohassel, Y. Zhang, SecureML: a system for scalable privacy-preserving machine learning, in *IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA* (2017), pp. 19–38
24. V. Nikolaenko, U. Weinsberg, S. Ioannidis, M. Joye, D. Boneh, N. Taft, Privacy-preserving ridge regression on hundreds of millions of records, in *IEEE Symposium on Security and Privacy, Berkeley, CA, USA* (2013), pp. 334–348
25. W. Zheng, R.A. Popa, J.E. Gonzalez, I. Stoica, Helen: Maliciously secure coopetitive learning for linear models, in *IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA* (2019), pp. 724–738
26. Y. Ishai, E. Kushilevitz, R. Ostrovsky, A. Sahai, Cryptography with constant computational overhead, in *Proceedings Annual ACM Symposium on Theory of Computing, Victoria British Columbia Canada* (2008), pp. 433–442
27. I. Damgård, Y. Ishai, M. Krøigaard, Perfectly secure multiparty computation and the computational overhead of cryptography, in *Annual International Conference on the Theory and Applications of Cryptographic Techniques, France* (2010), pp. 445–465
28. E. Rizk, S. Vlaski, A.H. Sayed, Federated learning under importance sampling (2020). arXiv:2012.07383
29. K. Avazu, Avazu's Click-Through Rate Prediction (2014). http://www.csie.ntu.edu.tw/-cj1in/libsvmtools/

Publisher's Note