Free-rider Attacks on Model Aggregation in Federated Learning

Yann Fraboni
Université Côte d'Azur, Inria Sophia Antipolis,
Epione Research Group, France
Accenture Labs, Sophia Antipolis, France
yann.fraboni@accenture.com

Richard Vidal
Accenture Labs, Sophia Antipolis, France
richard.vidal@accenture.com

Marco Lorenzi
Université Côte d'Azur, Inria Sophia Antipolis,
Epione Research Group, France
marco.lorenzi@inria.fr

Abstract

Free-rider attacks on federated learning consist in dissimulating participation to the federated learning process with the goal of obtaining the final aggregated model without actually contributing with any data. We introduce here the first theoretical and experimental analysis of free-rider attacks on federated learning schemes based on iterative parameters aggregation, such as FedAvg or FedProx, and provide formal guarantees for these attacks to converge to the aggregated models of the fair participants. We first show that a straightforward implementation of this attack can be simply achieved by not updating the local parameters during the iterative federated optimization. As this attack can be detected by adopting simple countermeasures at the server level, we subsequently study more complex disguising schemes based on stochastic updates of the free-rider parameters. We demonstrate the proposed strategies on a number of experimental scenarios, in both iid and non-iid settings. We conclude by providing recommendations to avoid free-rider attacks in real world applications of federated learning, especially in sensitive domains where security of data and models is critical.

1 Introduction

Federated learning is a training paradigm that has gained popularity in the last years as it enables different clients to jointly learn a global model without sharing their respective data. It is particularly suited for Machine Learning applications in domains where data security is critical, such as healthcare [1, 2]. The relevance of this approach is witnessed by current large scale federated learning initiatives on the way in the medical domain, for instance for learning predictive models of breast cancer[1], or for drug discovery and development[2].

In this setting, aggregation results entail critical information beyond data itself: a model trained on exclusive datasets may have very high commercial or intellectual value. This critical aspect can lead to the emergence of opportunistic behaviors in federated learning, where ill-intentioned clients may be participating with the unique scope of obtaining the federated model, without actually contributing with any data during the training process. In particular, the attacker, or free-rider, aims at disguising

[1] blogs.nvidia.com/blog/2020/04/15/federated-learning-mammogram-assessment/
[2] www.imi.europa.eu/projects-results/project-factsheets/melloddy

Preprint. Under review.

its participation to federated learning while ensuring that the iterative training process ultimately converges to the wished target: the aggregated model of the fair participants.

The study of security and safety of federated learning is an active research domain, and several kind of attacks have been studied. For example, an attacker may interfere during the iterative federated learning procedure to degrade/modify models performances [3, 4, 5, 6, 7], or retrieve information about other clients data [8, 9]. Yet, free-riding for federated learning has been so far under-investigated in the literature. To the best of our knowledge, the only investigation is in a preliminary work [10] focusing on attack strategies operated on federated learning based on gradient aggregation. However, no theoretical guarantees are provided for the effectiveness of this kind of proposed attacks. Furthermore this setup is unpractical in many real world applications, where federated training schemes based on model averaging are instead more common, due to the reduced data exchange across the network. FedAvg [11] is the most representative framework of this kind, as is based on the iterative averaging of the clients models' parameters, after updating each client model for a given number of training epochs at the local level. To improve the robustness of FedAvg in non-iid and heterogeneous learning scenarios, FedProx [12] extends FedAvg by including a regularization term penalizing local departures of clients' parameters from the global model.

The contribution of this work consists in the development of a theoretical framework for the development and study of free-rider attacks in federated learning schemes based on model averaging, such as in FedAvg and FedProx. The problem is here formalized via the reformulation of federated learning as a stochastic dynamical system describing the evolution of the aggregated parameters across iterations. A critical requirement for this kind of opportunistic attacks is to ensure the convergence of the training process to the wished target represented by the aggregated model of the fair clients. We show that the proposed framework allows to derive explicit conditions for guaranteeing the success of the attack. This is an important theoretical feature as it is of primary interest for the attacker not to interfere with the learning process.

We first derive in Section 2.3 a basic free-riding strategy to guarantee the convergence of federated learning to the model of the fair participants. This strategy simply consists in returning at each iteration the received global parameters. As this behavior can easily be detected by the server, we build more complex strategies to disguise the free-rider contribution to the optimization process, based on opportune stochastic perturbations of the parameters. We demonstrate in Section 2.4 that this strategy does not alter the global model convergence, and in Section 3 we experimentally demonstrate our theory on a number of learning scenarios in both iid and non-iid settings. We conclude by investigating advanced strategies for designing and identifying free-riders' attacks. All proofs and additional material are provided in the Appendix.

2 Methods

Before introducing in Section 2.2 the core idea of free-rider attacks, we first recapitulate in Section 2.1 the general context of parameter aggregation in federated learning.

2.1 Federated learning through model aggregation: FedAvg and FedProx

In federated learning, we consider a set I of participating clients respectively owning dataset $\mathcal{D}_i$ composed by M_i samples. During optimization, it is generally assumed that the D elements of the clients' parameters vector $\tilde{\boldsymbol{\theta}}_i(t) = (\tilde{\theta}_{i,0}(t), \tilde{\theta}_{i,1}(t), .., \tilde{\theta}_{i,D}(t))$, and the global parameters $\boldsymbol{\theta}(t) = (\theta_0(t), \theta_1(t), .., \theta_D(t))$ are aggregated independently at each iteration round t. Following this assumption, and for simplicity of notation, in what follows we restrict our analysis to a single parameter entry, that in will be generally denoted by $\tilde{\theta}_i(t)$ and $\theta(t)$ for clients and server respectively.

In this setting, to estimate a global model across clients, FedAvg [11] is an iterative training strategy based on the aggregation of local model parameters $\tilde{\theta}_i(t)$. At each iteration step t, the server sends the current global model parameters $\theta(t)$ to the clients. Each client updates the model by minimizing over E epochs the local cost function $\mathcal{L}(\tilde{\theta}_i(t), \mathcal{D}_i)$ initialized with $\theta(t)$, and subsequently returns the updated local parameters $\tilde{\theta}_i(t)$ to the server. The global model parameters $\theta(t+1)$ at the iteration

step $t+1$ are then estimated as a weighted average:

$$\theta(t+1) = \sum_{i \in I} \frac{M_i}{N} \tilde{\theta}_i(t), \tag{1}$$

where $N = \sum_{i \in I} M_i$ represents the total number of samples across distributed datasets. FedProx [12] builds on FedAvg by adding to the cost function a L2 regularization term penalizing the deviation of the local parameters $\tilde{\theta}_i(t)$ from the global parameters $\theta(t)$. The new cost function is $\mathcal{L}_{Prox}(\tilde{\theta}_i(t), \mathcal{D}_i) = \mathcal{L}(\tilde{\theta}_i(t), \mathcal{D}_i) + \frac{\mu}{2} \left\| \tilde{\theta}_i(t) - \theta(t) \right\|^2$ where μ is the hyperparameter monitoring the regularization by enforcing proximity between local updates and the global model.

2.2 Formalizing Free-rider attacks

The strategy of a free-rider consists in participating to federated learning by dissimulating local updating through the sharing of opportune counterfeited parameters with the aim of obtaining the aggregated model of the fair clients.

We denote by J the set of fair clients, i.e. clients following the federated learning strategy of Section (2.1) and by K the set of free-riders, i.e. malicious clients pretending to participate to the learning process, such that $I = J \cup K$ and $J \neq \emptyset$. We denote by M_K the number of samples declared by the free-riders, and we introduce the parameter $\beta = \frac{M_K}{N} - 1 \in [-1, 0]$ which quantifies the amount of free-riders' samples relative to the total number of training samples N.

We assume that, in absence of free-riders, clients' parameters observed during federated learning are realisations from time-varying processes $\tilde{\theta}_j(t) = \theta_j(t) + \rho_j \zeta_j(t)$, where $\theta_j(t)$ are smoothly convergent functions and $\zeta_j(t)$ is a noisy process, here assumed to be delta-correlated unit variance Gaussian white noise, modulated by the parameter ρ_j (this assumption will be relaxed in Section 2.4.2). The participation of the free-riders to federated learning implies that the processes of the fair clients are being perturbed by the attacks throughout training. In particular, perturbing the aggregation (1) at the server level implies, on each client's side, a perturbation of the initial condition for the local optimization problem. We therefore assume that each fair client's parameters evolution under free-rider attacks follows the perturbed trajectory $\tilde{\theta}_j(t) = \theta_j(t) + \rho_j \zeta_j(t) + \eta(1+\beta)\phi_j(t)$, with $\phi_j(t)$ a Gaussian white noise process. The perturbation is proportional to the number of samples declared by the free-riders, and its magnitude is controlled by the parameter η. This assumption accounts for the potential non-convexity of the problem at the client's level where, for a certain perturbation of the aggregated parameters $\theta(t)$, the parameters returned by the fair client would fall in a neighbourhood of the original process $\tilde{\theta}_j(t)$. As we shall see in the following Sections, the extent of the perturbation, i.e. the noise level η, determines the quality of the free-rider attack.

In the next Section we show that a basic free-rider strategy simply consists in returning at each iteration the received global parameters: $\theta_k(t) = \theta(t)$. We call this type of attack *plain free-riding*.

2.3 Plain free-riding

The plain free-rider returns the same model parameters as the received ones, i.e. $\forall k \in K,\ \theta_k(t) = \theta(t)$. In this setting, the server aggregation process (1) can be rewritten as:

$$\theta(t+1) = \sum_{j \in J} \frac{M_j}{N} \tilde{\theta}_j(t) + \frac{M_K}{N} \theta(t)\ . \tag{2}$$

Before proceeding with the development of equation (2), to have a better intuition of FedAvg under free-rider attacks, let us consider the simplified setting in which the fair clients parameters θ_j are assumed to be constant, i.e. $\forall j \in J,\ \tilde{\theta}_j(t) = \theta_j$. This is equivalent to consider that the optimization process is close enough to the optimum for all the clients, which return the same local parameters θ_j at each global update step. We rewrite the server aggregation process (2) as follows:

$$\theta(t+1) - \theta(t) = \sum_{j \in J} \frac{M_j}{N} \theta_j + \beta \theta(t). \tag{3}$$

Considering an infinitesimal increment of time, we obtain the first order differential equation:

$$\dot{\theta}(t) = \beta\theta(t) + \sum_{j\in J} \frac{M_j}{N}\theta_j, \tag{4}$$

for which the solution is $\theta(t) = e^{\beta t}\theta(0) + \sum_{j\in J} \frac{M_j}{N-M_K}\theta_j(1-e^{\beta t})$, where $\theta(0)$ is the training initialization. This expression shows that in this simple setting, the global model with plain free-riders converges to the aggregated model of the fair clients: since $\beta < 0$ we obtain $\theta(t) \xrightarrow{t\to+\infty} \sum_{j\in J} \frac{M_j}{N-M_K}\theta_j$. Also, the relative sample size declared by the free-riders $\frac{M_K}{N}$ influences the convergence with exponential speed $O(e^{\beta t})$. In practice, the smaller the ratio of data declared by the free-riders, the faster the trained model converges to its optimum, thus approaching the final model with fair clients. The limit case $M_K = N$, i.e. $\beta = 0$, which corresponds to only free-riders participating to federated learning ($J = \emptyset$), leads to the trivial result $\theta(t) = \theta(0)$. In this case there is no learning throughout the training process. We now generalize equation (3) to the time varying setting of equation (2):

$$\theta(t+1) - \theta(t) = \sum_{j\in J} \frac{M_j}{N}\tilde{\theta}_j(t) + \beta\theta(t), \tag{5}$$

leading to the first order stochastic differential equation with Wiener noise variables Wt_j^i:

$$\mathrm{d}\theta(t) = \left(\beta\theta(t) + \sum_{j\in J} \frac{M_j}{N}\theta_j(t)\right)\mathrm{d}t + \sum_{j\in J}\left(\frac{M_j}{N}(1+\beta)\eta\mathrm{d}Wt_j^1 + \frac{M_j}{N}\rho_j\mathrm{d}Wt_j^2\right). \tag{6}$$

Theorem 1 (Plain free-riding)**.** Assuming that $\forall j \in J\ \exists\alpha_j \in \mathbb{R}^+$ s.t. $|\dot{\theta}_j(t)| = O(t^{-\alpha_j})$, federated learning with plain free-riders (4) converges in expectation to the aggregated federated model of the fair clients:

$$\mathbb{E}\left[\theta(t)\right] \xrightarrow{t\to+\infty} \sum_{j\in J} \frac{M_j}{N-M_K}\theta_j(t), \tag{7}$$

$$\mathrm{Var}\left[\theta(t)\right] = V_{\eta^2,\{\rho_j\}} \xrightarrow{t\to+\infty} \frac{(N-\sum_j M_j)^2}{2N^2}\frac{1}{\sum_j \frac{M_j}{N}}\sum_j (\frac{M_j}{N})^2\eta^2 + \frac{1}{2\sum_j \frac{M_j}{N}}\sum_j (\frac{M_j}{N})^2\rho_j^2. \tag{8}$$

Since the output of plain free-riders is deterministic, this result arises from the perturbed process of the fair clients. On one hand, as soon as the perturbation η introduced by the attack is small enough, plain free-riders expect to converge to the fair clients' aggregation with convergence speed $O(e^{\beta t})$. On the other hand, Theorem 1 shows that the uncertainty of the process depends on the trade-off between free-riders and fair clients contributions. We note that in the limit case $\sum_j M_j = N$ (only fair participants) the uncertainty is uniquely due to the the fair clients' variability governed by ρ_i, as the first term of (8) is zero. With only free-riders ($\sum_j M_j = 0$) we obtain the trivial solution discussed for equation (4).

2.4 Disguised free-riding

Plain free-riders can be easily detected by the server, since for each iteration the condition $\theta_k(t) - \theta(t) = 0$ is true. We study here improved attack strategies based on the sharing of opportunely disguised parameters. Free-riders should ideally return disguised parameters $\theta_k(t) = f_k(\theta(t))$ such that in expectation we obtain $\mathbb{E}\left[\theta_k(t)\right] = \theta(t)$. Simulating the behavior of a fair client also requires the free-rider to simulate convergence: the magnitude of the update should be more important at the beginning of the training and approaching zero towards the end.

In what follows we investigate sufficient conditions on the perturbation models to obtain the desired convergence behaviour of free-rider attacks.

2.4.1 Attacks based on additive stochastic perturbations

A disguised free-rider with additive noise generalizes the plain one, and uploads parameters $\theta_k(t) = \theta(t) + \varphi(t)\epsilon(t)$. Here, the perturbation $\epsilon(t)$ is assumed to be Gaussian white noise, and $\varphi(t) > 0$ is a

suitable time-varying perturbation compatible with the free-rider attack. In this new setting, we can rewrite the FedAvg aggregation process (1) as follows:

$$\theta(t+1) = \sum_{j \in J} \frac{M_j}{N} \tilde{\theta}_j(t) + \frac{M_K}{N} \theta(t) + \frac{M_K}{N} \varphi(t) \epsilon(t). \tag{9}$$

Extending the plain case, the relationship leads to the following stochastic differential equation:

$$\begin{aligned} \mathrm{d}\theta(t) = & \left(\beta \theta(t) + \sum_{j \in J} \frac{M_j}{N} \theta_j(t) \right) \mathrm{d}t + \sum_{j \in J} \left(\frac{M_j}{N} (1+\beta) \eta \mathrm{d}Wt_j^1 + \frac{M_j}{N} \rho_j \mathrm{d}Wt_j^2 \right) \\ & + \frac{M_K}{N} \varphi(t) \mathrm{d}Wt. \end{aligned} \tag{10}$$

Theorem 2 (Single disguised free-rider)**.** Under the conditions of Theorem 1, let the perturbation model $\varphi(t)$ in equation (10) be such that $\varphi(t) = O(t^{-\gamma})$, with $\gamma > 0$. Then, federated learning (9) converges to the aggregated model of the fair clients. We have:

$$\mathbb{E}\left[\theta(t)\right] \xrightarrow{t \to +\infty} \sum_{j \in J} \frac{M_j}{N - M_K} \theta_j(t), \text{ and } \mathrm{Var}\left[\theta(t)\right] \xrightarrow{t \to +\infty} V_{\eta^2, \{\rho_j\}} \tag{11}$$

The extension to this result to the case of multiple free-riders requires to account in equation (10) for an attack of the form $\sum_{k \in K} \frac{M_K^{(k)}}{N} \varphi_k(t) \mathrm{d}Wt_k$, where $M_K^{(k)}$ is the sample size declared by free-rider k. Corollary 1 follows from the linearity of this functional.

Corollary 1 (Multiple disguised free-riders)**.** If the perturbation model $\varphi_k(t)$ of each free-rider is such that $\varphi_k(t) = O(t^{-\gamma_k})$, with $\gamma_k > 0$, the result of Theorem 2 still holds when combining multiple free-riders in equation (9).

2.4.2 Time-varying noise model of fair-clients evolution

Theorem 2 provides us with conditions on the decay of the noisy update an attacker should design to ensure convergence of the process. Interestingly, the general decaying shape identified for $\varphi(t)$ can be seamlessly translated to define sufficient conditions for the time-varying variability $\rho_j(t)$ of the fair clients, to ensure compatibility with the federated learning process.

Corollary 2. Let the fair clients evolve according to $\tilde{\theta}_j(t) = \theta_j(t) + \rho_j(t)\zeta_j(t)$, where $\theta_j(t)$ are smoothly convergent functions and $\zeta_j(t)$ is delta-correlated unit variance Gaussian white noise. If the functions $\rho_j(t)$ are such that $\rho_j(t) = O(t^{-a_j})$, with $a_j > 0$, then the aggregation process of federated learning in equation (1) is such that $\mathbb{E}\left[\theta(t)\right] \xrightarrow{t \to +\infty} \sum_{j \in J} \frac{M_j}{N - M_K} \theta_j(t)$, and $\mathrm{Var}\left[\theta(t)\right] \xrightarrow{t \to +\infty} 0$. Under the same conditions, the asymptotic variance of Theorems 1 and 2 reduces to $\mathrm{Var}\left[\theta(t)\right] \xrightarrow{t \to +\infty} \frac{(N - \sum_j M_j)^2}{2N^2} \frac{1}{\sum_j \frac{M_j}{N}} \sum_j (\frac{M_j}{N})^2 \eta^2$.

Corollary 2 enables the generalization of our theory to more realistic noise models for the fair clients. We can indeed relax the initial stationarity assumption on the variability of the parameters' evolution, to account for smoothly decaying perturbations when approaching the client optima during training. Furthermore, it is interesting to notice that in this case the variability of the global model under free-rider attacks is uniquely due to the perturbation η^2 induced by the free-riders.

3 Experiments

This experimental section focuses on a series of benchmarks for the proposed free-rider attacks. The methods being of general application, the focus here is to empirically demonstrate our theory on diverse experimental setups and model specifications. All code, data and experiments are available at `https://github.com/Accenture/Labs-Federated-Learning/tree/free-rider_attacks/`

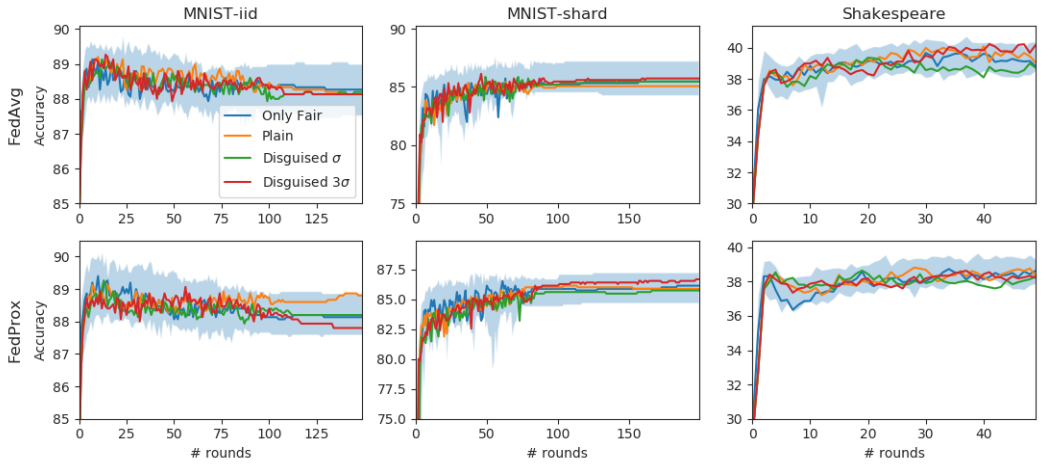

Figure 1: Accuracy performances for (a) FedAvg and (b) FedProx in the different experimental scenarios (20 local training epochs). The shaded blue region indicates the variability of federated learning model with fair clients only, estimated from 30 different training initialization.

3.1 Experimental Details

We consider 5 fair clients for each of the following scenarios:

MNIST (classification in iid and non-iid settings). We study a standard classification problem on MNIST [13] and create two benchmarks: an iid dataset (MNIST iid) where we give 600 training digits and 300 testing digits to each client, and a non-iid dataset (MNIST non-iid), where for each digit we create two shards with 150 training samples and 75 testing samples, and allocate 4 shards for each client. For each scenario, we use a logistic regression predictor.

Shakespeare (LSTM prediction). We study a LSTM model for next character prediction on the dataset of *The Complete Works of William Shakespeare* [11]. We randomly choose 5 clients with more than 3000 samples, and assign 70% of the dataset to training and 30% to testing. Each client has on average 6415.4 samples (± 1835.6) . We use a two-layer LSTM classifier containing 100 hidden units with an 8 dimensional embedding layer. The model takes as an input a sequence of 80 characters, embeds each of the characters into a learned 8-dimensional space and outputs one character per training sample after 2 LSTM layers and a fully connected one.

We train federated models following in turn FedAvg and FedProx aggregation processes. In FedProx, the hyperparameter μ monitoring the regularization is chosen according to the best performing scenario reported in [12]: $\mu = 1$ for MNIST (iid and non-iid), and $\mu = 0.001$ for Shakespeare. For the free-rider we declare a number of samples equal to the average sample size across fair clients. We test federated learning with 5 and 20 local epochs using SGD optimization with learning rate $\lambda = 0.0005$ for MNIST (iid and non-iid), and $\lambda = 0.5$ for Shakespeare, and batch size of 100.

3.2 Free-rider attacks: convergence and performances

Designing a free-rider attack requires to specify a perturbation function $\varphi(t) \in O(t^{-\gamma})$, for a general parameter $\gamma > 0$. While the design of optimally disguised free-rider attacks is out of the scope of this study, here we propose heuristics to tune the perturbations parameters according to practical hypothesis on the parameters evolution. These hypothesis are discussed and refined in Section 3.3.

In the following experiments, we investigate free-rider attacks taking the simple form $\varphi(t) = \sigma t^{-\gamma}$. The parameter γ is chosen among a panel of testing parameters $\gamma \in \{0.5, 1, 2\}$. After random initialization at the initial federated learning step, the parameter σ is instead opportunely estimated to mimic the extent of the distribution of the update $\Delta\theta = \theta(1) - \theta(0)$ observed between consecutive rounds of federated learning. We can simply model these increments as a zero-centered univariate Gaussian distribution, and assign the parameter σ to the value of the fitted standard deviation.

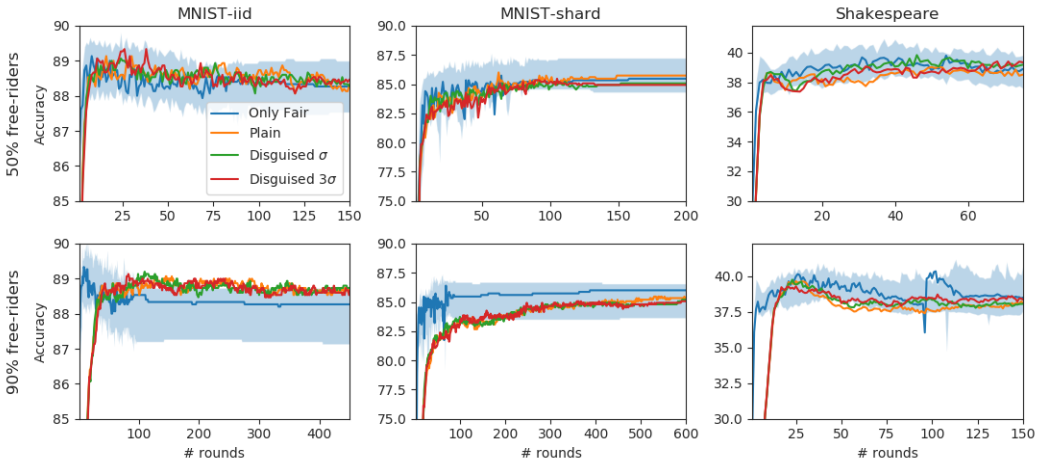

Figure 2: Plots for FedAvg and $E = 20$. Accuracy performances for FedAvg according to the number of free-riders participating in the learning process: 50% (top), and 90% (bottom) of the total amount of clients.

According to this strategy, the free-rider would return parameters $\theta_k(t)$ with perturbations distributed as the ones observed between two consecutive optimization rounds.

Figure 1 shows the evolution of the model obtained with FedAvg (20 local training epochs) with respect to different scenarios: 1) fair clients only, 2) plain free-rider, 3) disguised free-rider with decay parameter $\gamma = 1$, and estimated noise level σ, and 4) disguised free-rider with noise level increased to 3σ. For each scenario, we compare the federated model obtained under free-rider attacks with respect to the equivalent model obtained with the participation of the fair clients only. For this latter setting, to assess the model training variability, we repeated the training 30 times with different parameter initializations. The results show that, independently from the chosen free-riding strategy, the resulting models attain comparable performances with respect to the one of the model obtained with fair clients only. Similar results are obtained for the setup with 5 local training epochs and different values of γ (Appendix C.1). We also quantified the equivalence of the models parameter-wise, via the average L2 distance, and in terms of the overall parameter distribution, through the Kolmogorov-Smirnov (KS) test (Appendix C.2), confirming that for each scenario the free-riders converge to the fair client's model, whereas the 3σ scenario seems to lead to larger dissimilarities. This result is in agreement with Theorem 2 and also suggests that the perturbation η induced by the attacks is generally small.

We investigate the same training setup under the influence of multiple free-riders. In particular, we test the scenarios where the free-riders declare respectively 50% and 90% of the total training sample size. In practice, we maintain the same experimental setting composed by 5 fair clients, and we increase the number of free-riders to respectively 5 and 45, while declaring for each free-rider a sample size equal to the average number of samples of the fair clients.

Figure 2 shows that, independently from the magnitude of the perturbation function, the number of free-riders does not seem to affect the performance of the final aggregated model. On the contrary, the convergence speed is importantly impacted, as it is sensibly slower in the 90% free-riders scenario. This result is confirmed when inspecting the dissimilarity with respect to the fair clients' model in terms of L2 and KS measures (Appendix C.2), and is similar for the setup with 5 local training epochs, and with FedProx (Appendix C.1).

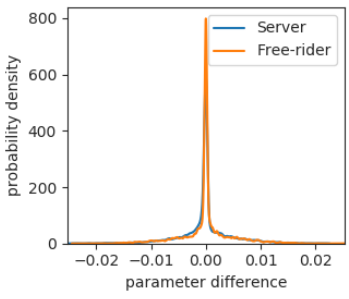

Figure 3: Distribution of the parameter update $\Delta\theta$ for FedAvg with MNIST non-iid and $E = 20$. The distribution is optimally modelled by a mixture of 3 Gaussians.

3.3 Advanced Free-rider attacks

This section illustrates practical directions for improving the disguising scheme by leveraging on the general result of Theorem 2. The key observation is that the form of a model update

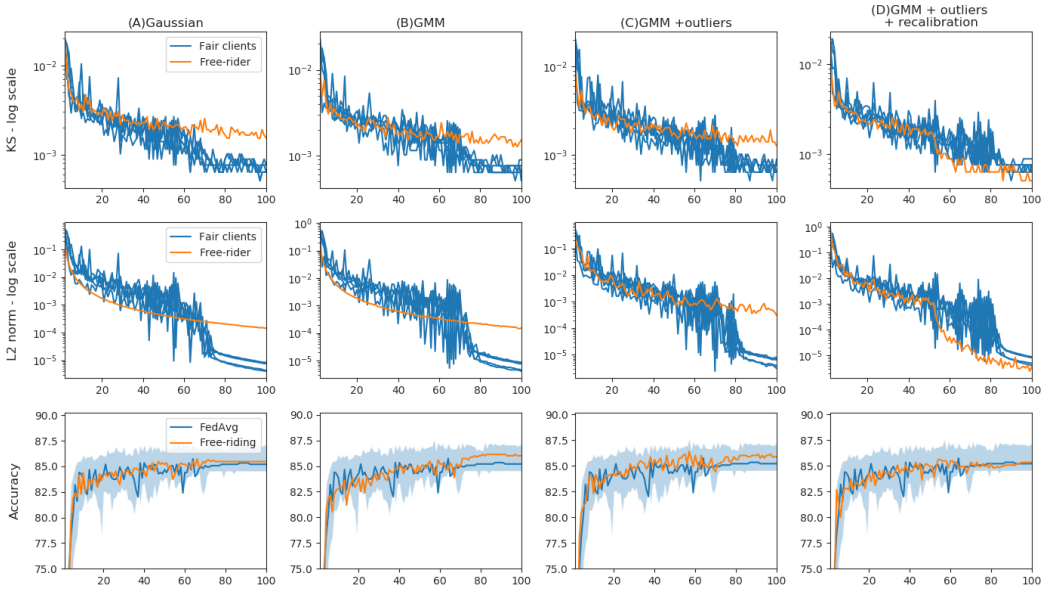

Figure 4: Comparison of free-rider and fair client's parameters with respect to the global model (FedAvg, MNIST non-iid, 20 local training epochs). Evolution of Kolmogorov-Smirnov (KS) statistic (top), L2 distance (middle), and model accuracy (bottom) across optimization rounds.

during training is generally non Gaussian (Figure 3). In most cases, a general parameter update $\Delta\theta(t) = \theta(t) - \theta(t-1)$ is zero-centered and heavily skewed, with only some parameters affected by large changes between optimization rounds. For this reason, the creation of a synthetic update based on a Gaussian model may still be easily identified at the server level, for example by simple comparison of the distribution of the free-rider parameters with respect to the global model's one (Figure 4, column A). To improve the realism of the attack, we investigate disguising schemes based on the fitting of multimodal distribution forms for the update. In particular, we model the initial global update, $\Delta\theta = \theta(1) - \theta(0)$, through Gaussian Mixture Modeling (GMM), where the optimal number of 3 Gaussian components was established according to the associated Bayesian Information Criterion (BIC).

Each parameter of the model is assigned to one of the clusters, depending on the extent of the relative update. Similarly as in Section 3.2, each parameter is associated with a perturbation σ equal to the standard deviation of the respective Gaussian component, thus obtaining 3 different evolution profiles characterized by increasing magnitude. While this strategy aims at obtaining a more realistic representation of the variability of the features' update across optimization rounds, the GMM may still lead to overly smooth simulated parameters, since it is based on the modeling of the average quantity $\Delta\theta$ (Figure 4, column B). This issue can be overcome by random generation of 'skews', for example by assigning a subset of parameters to outlier values, to mimic specificity of the training on the local dataset. The subset is here chosen as 10 parameters belonging to the Gaussian component with highest variance, to which we assign a perturbation value σ equal to 25 times their variance (Figure 4, column C). Finally, the profile of the perturbation $\varphi(t)$ may not faithfully follow the model evolution over long time horizons. Figure 4, column D, shows that re-calibration of the perturbation parameters σ after a fixed number of rounds (here 50) can improve the realism of the update.

4 Conclusion and discussion

In this work, we introduced a theoretical framework for the study of free-riding attacks on model aggregation in Federated Learning. Based on the proposed methodology, we proved that simple strategies based on returning the global model at each iteration already lead to successful free-rider attacks (plain free-riding), and we investigated more sophisticated disguising techniques relying on stochastic perturbations of the parameters (disguised free-riding). The convergence of each attack was demonstrated through theoretical proofs and experimental results.

This work opens the way to the investigation of optimal disguising and defense strategies for free-rider attacks, beyond the proposed heuristics. Our experiments show that inspection of the client's distribution should be established as a routine practice for the detection of free-rider attacks in federated learning. This result motivates the study of more effective free-riding strategies, based on different noise model distributions and perturbation schemes. We will also work on the improvement of detection at the server level, through better modeling of the heterogeneity of the incoming clients' parameters. Finally, this study relies on a number of hypothesis concerning the evolution of the clients' parameters during federated learning. This choice provides us with a convenient theoretical setup for the formalization of the proposed theory which may be modified in the future, for example, for investigating more complex forms variability and parameters aggregation.

Broader Impact

The problem of free-rider attacks in federated learning may have significant economical and societal impact. Since it allows clients' participation without sharing the data, federated learning is becoming the de-facto training setup in current large-scale machine learning projects in several critical applications, such as healthcare, banking, and telecommunication. The resulting models derived from sensitive and protected data may have high commercial and intellectual value as well, due to their exclusive nature. Our research proves that if precautions are not taken, malicious clients can disguise their participation to federated learning to appropriate a federated model without providing any contribution. Our research therefore stimulates the investigation of novel verification techniques for the implementation of secured federated learning projects, to avoid intellectual property or commercial losses.

Acknowledgments and Disclosure of Funding

This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002, and by the ANR JCJC project Fed-BioMed 19-CE45-0006-01. The project was also supported by Accenture and the Inria Sophia Antipolis - Méditerranée, "NEF" computation cluster.

References

[1] Theodora Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Paschalidis, and Wei Shi. Federated learning of predictive models from federated electronic health records. International Journal of Medical Informatics, 112, 01 2018.

[2] Santiago Silva, Boris A Gutman, Eduardo Romero, Paul M Thompson, Andre Altmann, and Marco Lorenzi. Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data. In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pages 270–274. IEEE, 2019.

[3] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Analyzing federated learning through an adversarial lens. 36th International Conference on Machine Learning, ICML 2019, 2019-June:1012–1021, 2019.

[4] Bo Li, Yining Wang, Aarti Singh, and Yevgeniy Vorobeychik. Data poisoning attacks on factorization-based collaborative filtering. Advances in Neural Information Processing Systems, (Nips):1893–1901, 2016.

[5] Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. 35th International Conference on Machine Learning, ICML 2018, 13:8947–8956, 2018.

[6] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In International Conference on Learning Representations, 2019.

[7] Shiqi Shen, Shruti Tople, and Prateek Saxena. AUROR: Defending against poisoning attacks in collaborative deep learning systems. In ACM International Conference Proceeding Series, volume 5-9-Decemb, pages 508–519, 2016.

[8] Zhibo Wang, Mengkai Song, Zhifei Zhang, Yang Song, Qian Wang, and Hairong Qi. Beyond Inferring Class Representatives: User-Level Privacy Leakage from Federated Learning. Proceedings - IEEE INFOCOM, 2019-April:2512–2520, 2019.

[9] Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz. Deep Models under the GAN: Information leakage from collaborative deep learning. Proceedings of the ACM Conference on Computer and Communications Security, pages 603–618, 2017.

[10] Jierui Lin, Min Du, and Jian Liu. Free-riders in Federated Learning: Attacks and Defenses. http://arxiv.org/abs/1911.12560, 2019.

[11] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 54, 2017.

[12] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated Optimization in Heterogeneous Networks. Proceedings of the 1 st Adaptive & Multitask Learning Workshop, Long Beach, California, 2019, pages 1–28, 2018.

[13] Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Ha. LeNet. Proceedings of the IEEE, (November):1–46, 1998.

A Complete Proofs

A.1 Proof of Theorem 1

Proof. The aggregation at the server level follows:

$$\theta(t+1) = \sum_{j \in J} \frac{M_j}{N} \tilde{\theta}_j(t) + \frac{M_K}{N} \theta(t). \tag{12}$$

By considering an infinitesimal increment of time we obtain the following first order stochastic differential equation:

$$\dot{\theta}(t) = \sum_{j \in J} \frac{M_j}{N} \tilde{\theta}_j(t) + \beta \theta(t), \tag{13}$$

where we define $\beta = \frac{M_K}{N} - 1$. According to the noise process hypothesis of Section (2.2), we thus obtain the stochastic differential equation:

$$\mathrm{d}\theta(t) = \left(\beta \theta(t) + \sum_{j \in J} \frac{M_j}{N} \theta_j(t) \right) \mathrm{d}t + \sum_{j \in J} \left(\frac{M_j}{N} (1+\beta) \eta \mathrm{d}W_j^1(t) + \frac{M_j}{N} \rho_j \mathrm{d}W_j^2(t) \right), \tag{14}$$

with associated solution:

$$\begin{aligned} \theta(t) = e^{\beta t} [c_0 + \sum_{j \in J} \frac{M_j}{N} \underbrace{\int_{t_0}^{t} e^{-\beta x} \, \theta_j(x) dx}_{\mathbf{A}} + \sum_{j \in J} \frac{M_j}{N} (1+\beta) \eta \underbrace{\int_{t_0}^{t} e^{-\beta x} dW_{j\,x}^1}_{\mathbf{B1}} \\ + \sum_{j \in J} \frac{M_j}{N} \rho_j \underbrace{\int_{t_0}^{t} e^{-\beta x} dW_{j\,x}^2}_{\mathbf{B2}}], \end{aligned} \tag{15}$$

where we denote $c_0 = e^{-\beta t_0} \theta(t_0)$. In the following part of the proof we study the asymptotic properties of solution 15 separately.

- **Asymptotic convergence of A**. We study the asymptotic properties of (A) by first integrating by parts the integrals $\int_{t_0}^{t} e^{-\beta x} \, \theta_j(x) \, dx$:

$$\begin{aligned} \int_{t_0}^{t} e^{-\beta x} \theta_j(x) dx &= [-\frac{1}{\beta} e^{-\beta x} \theta_j(x)]_{t_0}^{t} + \int_{t_0}^{t} \frac{1}{\beta} e^{-\beta x} \dot{\theta}_j(x) dx \\ &= -\frac{1}{\beta} e^{-\beta t} \theta_j(t) + \frac{1}{\beta} \int_{t_0}^{t} e^{-\beta x} \dot{\theta}_j(x) dx + C. \end{aligned} \tag{16}$$

  To investigate the asymptotic properties of the combination of the integrals (16) in equation (15), we first note that $\frac{M_j}{N} \frac{1}{\beta} = \frac{M_j}{M_K - N}$. We can therefore study the limit $t \to \infty$ of the quantity:

$$g(t) = \sum_{j \in J} \frac{M_j}{M_K - N} e^{\beta t} \int_{t_0}^{t} e^{-\beta x} \, \dot{\theta}_j(x) \, dx. \tag{17}$$

  We assume smooth convergence for $\theta_j(t)$ under the condition $|\dot{\theta}_j| = O(\frac{1}{t^{\alpha_j}})$, i.e. $\exists \, t_j, R_i \in \mathbb{R}^+$ such that $\forall t \geq t_j, |\theta_j(t)| \leq \frac{R_i}{t^{\alpha_j}}$. By considering $t_0 = \max_{j \in J}(t_j)$, $\alpha = \min_{j \in J}(\alpha_j)$ and $t > t_0$, the triangular inequality gives:

$$|g(t)| \leq \sum_{j \in J} \frac{M_j}{M_K - N} e^{\beta t} | \int_{t_0}^{t} e^{-\beta x} \dot{\theta}_j(x) \, dx |, \tag{18}$$

and for the assumption of piece-wise continuity of $\dot{\theta}_j$ in $[t_0, t]$ we get:

$$|g(t)| \leq \sum_{j\in J} \frac{M_j}{N - M_K} e^{\beta t} \int_{t_0}^{t} e^{-\beta x} |\dot{\theta}_j(x)| dx$$
$$\leq \sum_{j\in J} \frac{M_j}{N - M_K} R_i e^{\beta t} \int_{t_0}^{t} \frac{e^{-\beta x}}{x^{\alpha}} dx. \tag{19}$$

Since that $\alpha > 0$, $t_0 > 0$ and $-\beta > 0$, lemma in Proposition B.1 shows that

$$|g(t)| \leq e^{\beta t} \int_{t_0}^{t} \frac{e^{-\beta x}}{x^{\alpha}}\, dx \xrightarrow{t\to+\infty} 0. \tag{20}$$

$$\tag{21}$$

- **Asymptotic convergence of B1 and B2**. The asymptotic properties of the stochastic integrals B1 and B2 follow from the general properties of Ito's integrals. For any constant L, we have:

$$\mathbb{E}\left[L e^{\beta t} \int_{t_0}^{t} e^{-\beta x} dW_x \right] = 0, \tag{22}$$
$$\mathrm{Var}\left[L e^{\beta t} \int_{t_0}^{t} e^{-\beta x} dW_x \right] = L^2 e^{2\beta t} \int_{t_0}^{t} e^{-2\beta x} dW_x \tag{23}$$
$$= -\frac{L^2 e^{2\beta t}}{2\beta} (e^{-2\beta t} - e^{-2\beta t_0}) \xrightarrow{t\to+\infty} -\frac{L^2}{2\beta}, \tag{24}$$

  where the first equality in line (23) is due to Ito's isometry.

By substituting in equation (15) the convergence results for the quantities (16), (22) and (23), we finally conclude that:

$$\mathbb{E}\left[\theta(t)\right] \xrightarrow{t\to+\infty} \sum_{j\in J} \frac{M_j}{N - M_K} \theta_j(t), \tag{25}$$
$$\mathrm{Var}\left[\theta(t)\right] \xrightarrow{t\to+\infty} -\frac{1}{2\beta} \sum_{j\in J} (\frac{M_j}{N})^2 \left((1+\beta)^2 \eta^2 + \rho_j^2 \right) \tag{26}$$
$$= -\frac{(1+\beta)^2 \eta^2}{2\beta} \sum_{j\in J} (\frac{M_j}{N})^2 - \frac{1}{2\beta} \sum_{j\in J} (\frac{M_j}{N})^2 \rho_j^2 \tag{27}$$
$$= \frac{(N - \sum_j M_j)^2 \eta^2}{2N^2} \frac{N}{\sum_j M_j} \sum_j (\frac{M_j}{N})^2 + \frac{N}{2\sum_j M_j} \sum_j (\frac{M_j}{N})^2 \rho_j^2 \tag{28}$$
$$= \frac{(N - \sum_j M_j)^2 \eta^2}{2N^2} \frac{1}{\sum_j \frac{M_j}{N}} \sum_j (\frac{M_j}{N})^2 + \frac{1}{2\sum_j \frac{M_j}{N}} \sum_j (\frac{M_j}{N})^2 \rho_j^2. \tag{29}$$

Note: in the special case $\beta = 0$, then equation (12) can be expressed as $\theta(t+1) = \theta(t)$ thus $\theta(t) = \theta_0(t)$.

□

A.2 Proof of Theorem 2

Proof. The differential equation governing the disguised attack is

$$\mathrm{d}\theta(t) = \left(\beta\theta(t) + \sum_{j\in J} \frac{M_j}{N} \theta_j(t) \right) \mathrm{d}t + \sum_{j\in J} \left(\frac{M_j}{N} (1+\beta) \eta \mathrm{d}Wt_j^1 + \frac{M_j}{N} \rho_j \mathrm{d}Wt_j^2 \right)$$
$$+ \frac{M_K}{N} \varphi(t) \mathrm{d}Wt. \tag{30}$$

We note that this equation differs from the one in proof (A.1) for the last term only. We derive conditions for the perturbation $\varphi(t) > 0$ for ensuring convergence, by studying the integral

$$\mathcal{L}(t) = e^{\beta t} \int_{t_0}^{t} \varphi(x) e^{-\beta x} dW_x. \tag{31}$$

We prove the convergence of $\mathcal{L}(t)$ when $\varphi(t) = O(t^{-\gamma})$, $\gamma > 0$. In particular, let $\sigma > 0$ and $t_0 > 0$ such that $\varphi(t) \leq \sigma t^{-\gamma}$, for $t > t_0$. We thus have $\mathcal{L}(t) \leq \mathcal{L}'(t) = \sigma e^{\beta t} \int_{t_0}^{t} \frac{e^{-\beta x}}{x^{\gamma}} dW_x$, and

$$\mathbb{E}\left[\mathcal{L}'(t)\right] \xrightarrow{t \to +\infty} 0, \tag{32}$$

$$\mathrm{Var}\left[\mathcal{L}'(t)\right] = \sigma^2 e^{2\beta t} \int_{t_0}^{t} \frac{e^{-2\beta x}}{x^{2\gamma}} dW_x \xrightarrow{t \to +\infty} 0, \tag{33}$$

where the limit of the variance is a consequence of Proposition B.1.

□

A.3 Proof of Corollary 1

Proof. The differential equation governing the disguised attack for many free-riders:

$$\begin{aligned} \mathrm{d}\theta(t) = &\left(\beta\theta(t) + \sum_{j \in J} \frac{M_j}{N} \theta_j(t) \right) \mathrm{d}t + \sum_{j \in J} \left(\frac{M_j}{N} (1+\beta) \eta \mathrm{d}Wt_j^1 + \frac{M_j}{N} \rho_j \mathrm{d}Wt_j^2 \right) \\ &+ \sum_{k \in K} \frac{M_k}{N} \varphi_k(t) \mathrm{d}Wt_k. \end{aligned} \tag{34}$$

We note that this equation differs from the one in proof (A.1) for the last term only. Since proof (A.2) already proves the convergence for a single term $\mathcal{L}_k(t) = e^{\beta t} \int_{t_0}^{t} \varphi_k(x) e^{-\beta x} dW_x$. when $\varphi_k(t) = O(t^{-\gamma_k})$, $\gamma_k > 0$. The corollary thus follows from the linearity of the last term:

$$\mathbb{E}\left[\theta(t)\right] \xrightarrow{t \to +\infty} \sum_{j \in J} \frac{M_j}{N - M_K} \theta_j(t) \text{ and } \mathrm{Var}\left[\theta(t)\right] \xrightarrow{t \to +\infty} V_{\eta^2, \{\rho_j\}}. \tag{35}$$

□

A.4 Proof of Corollary 2

Proof. We consider $\rho_j(t) = O(t^{-a_j})$ with $\alpha_j > 0$ for $t > t_0 > 0$. We show in proof (A.2) that for $\mathcal{L}_{\rho_j}(t) = \frac{M_j}{N} e^{\beta t} \int_{t_0}^{t} \rho_j(t) e^{-\beta x} dW^2{}_x^j$, we have:

$$\mathbb{E}\left[\mathcal{L}_{\rho_j}(t)\right] \xrightarrow{t \to +\infty} 0 \text{ and } \mathrm{Var}\left[\mathcal{L}_{\rho_j}(t)\right] \xrightarrow{t \to +\infty} 0 \tag{36}$$

Thus for Theorem 1 and 2 and for Corollary 1, the asymptotic convergence of $\theta(t)$ follows:

$$\mathbb{E}\left[\theta(t)\right] \xrightarrow{t \to +\infty} \sum_{j \in J} \frac{M_j}{N - M_K} \theta_j(t), \tag{37}$$

$$\mathrm{Var}\left[\theta(t)\right] \xrightarrow{t \to +\infty} -\frac{1}{2\beta} \sum_{j \in J} (\frac{M_j}{N})^2 (1+\beta)^2 \eta^2 \tag{38}$$

$$= -\frac{(1+\beta)^2 \eta^2}{2\beta} \sum_{j \in J} (\frac{M_j}{N})^2 \tag{39}$$

$$= \frac{(N - \sum_j M_j)^2 \eta^2}{2N^2} \frac{N}{\sum_j M_j} \sum_j (\frac{M_j}{N})^2 \tag{40}$$

$$= \frac{(N - \sum_j M_j)^2 \eta^2}{2N^2} \frac{1}{\sum_j \frac{M_j}{N}} \sum_j (\frac{M_j}{N})^2. \tag{41}$$

□

B Calculus

Proposition B.1. $\forall\, t_0, k, \alpha \in \mathbb{R}^+$, the following limit holds:

$$e^{-kt} \int_{t_0}^{t} \frac{e^{kx}}{x^\alpha}\, dx \xrightarrow{t\to+\infty} 0.$$

Proof. Let us consider $t, t_0, k \in \mathbb{R}^+$, $\alpha \in (0,1)$, and define h such that $h(t, t_0, k, \alpha) = e^{-kt}\int_{t_0}^{t} \frac{e^{kx}}{x^\alpha}\, dx$.
Writing the exponential as a power series, we get:

$$h(t, t_0, k, \alpha) = e^{-kt} \int_{t_0}^{t} \sum_{n=0}^{+\infty} \frac{(kx)^n}{n!} \frac{1}{x^\alpha}\, dx. \tag{42}$$

Considering that $\forall (x, n),\ \frac{(kx)^n}{n!}\frac{1}{x^\alpha} \geq 0$, we can use the Fubini/Tonelli theorem and permute the sum and the integral which gives:

$$\begin{aligned}
h(t, t_0, k, \alpha) &= e^{-kt} \sum_{n=0}^{+\infty} \int_{t_0}^{t} \frac{k^n}{n!} x^{n-\alpha}\, dx && (43)\\
&= e^{-kt} \sum_{n=0}^{+\infty} \frac{k^n}{n!} [\frac{x^{n-\alpha+1}}{n-\alpha+1}]_{t_0}^{t} && (44)\\
&= e^{-kt} [\sum_{n=0}^{+\infty} \frac{k^n}{n!} \frac{t^{n-\alpha+1}}{n-\alpha+1} - \sum_{n=0}^{+\infty} \frac{k^n}{n!} \frac{t_0^{n-\alpha+1}}{n-\alpha+1}] && (45)\\
&= e^{-kt} [t^{1-\alpha} l(t, k, \alpha) - t_0^{1-\alpha} l(t_0, k, \alpha)], && (46)
\end{aligned}$$

with $l(x, k, \alpha) = \sum_{n=0}^{+\infty} \frac{k^n}{n!} \frac{x^n}{n-\alpha+1}$.

The theorem of Cauchy-Hadamard tells us that the radius of convergence of $l(x, \alpha)$ is $\mathbb{R}$

$$\frac{\frac{k^{n+1}}{(n+1)!}\frac{1}{(n+1)-\alpha+1}}{\frac{k^n}{n!}\frac{1}{n-\alpha+1}} = \frac{k}{n+1}\frac{n-\alpha+1}{n-\alpha+2} \xrightarrow{n\to+\infty} 0. \tag{47}$$

We find a convenient upper bound $l(x)$ without a power series:

$$\begin{aligned}
l(x, k, \alpha) &\leq \frac{1}{1-\alpha} + \sum_{n=1}^{+\infty} \frac{k^n}{n!} \frac{2}{n+1} t^n && (48)\\
&= \frac{1}{1-\alpha} + \frac{2}{kt} \sum_{n=2}^{+\infty} \frac{k^n}{n!} t^n && (49)\\
&= \frac{1}{1-\alpha} + \frac{2}{kt} [e^{kt} - kt - 1]. && (50)
\end{aligned}$$

which gives the following upper bound for $l(t)$

$$h(t, t_0, k, \alpha) \leq \frac{e^{-kt} t^{1-\alpha}}{1-\alpha} + \frac{2}{kt^\alpha} - 2e^{-kt} t^{1-\alpha} - \frac{2}{t^\alpha} e^{-kt} - t_0^{1-\alpha} l(t_0) e^{-kt}. \tag{51}$$

Given that k and α are positive and that $h(t, t_0, k, \alpha) \geq 0$ we finally get:

$$h(t, t_0, k, \alpha) = e^{-kt} \int_{t_0}^{t} \frac{e^{kx}}{x^\alpha}\, dx \xrightarrow{t\to+\infty} 0. \tag{52}$$

In the case where $\alpha \geq 1$, $\exists \gamma \in (0,1)$ s.t. $\frac{1}{\alpha} \leq \frac{1}{\gamma}$, and we obtain:

$$h(t, t_0, k, \alpha) \leq h(t, t_0, k, \gamma) \xrightarrow{t\to+\infty} 0. \tag{53}$$

Finally:

$$\forall (t_0, k, \alpha) \in {\mathbb{R}^+}^3,\ h(t, t_0, k, \alpha) = e^{-kt} \int_{t_0}^{t} \frac{e^{kx}}{x^\alpha}\, dx \xrightarrow{t\to+\infty} 0. \tag{54}$$

□

C Additional experimental results

C.1 Accuracy Performances

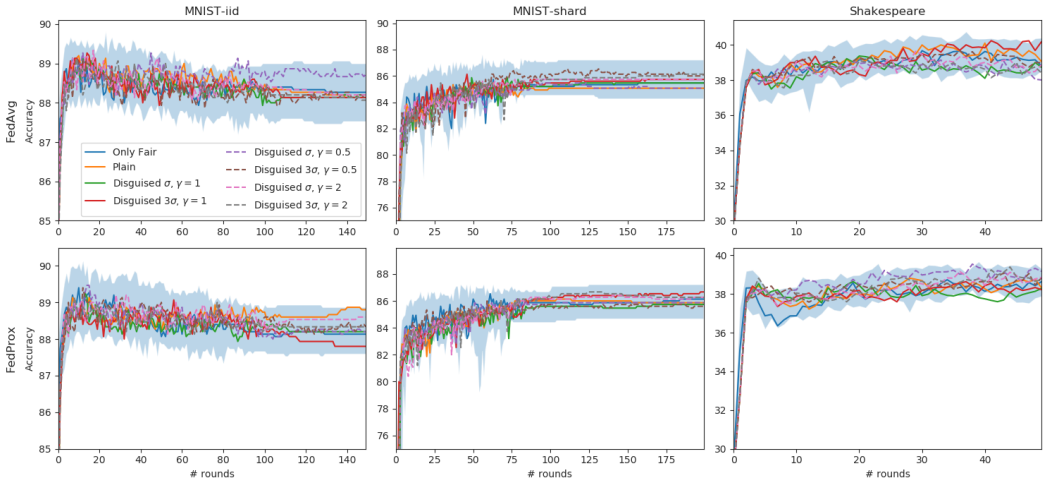

Figure 5: Accuracy performances for (a) FedAvg and (b) FedProx in the different experimental scenarios (20 local training epochs). The shaded blue region indicates the variability of federated learning model with fair clients only, estimated from 30 different training initialization.

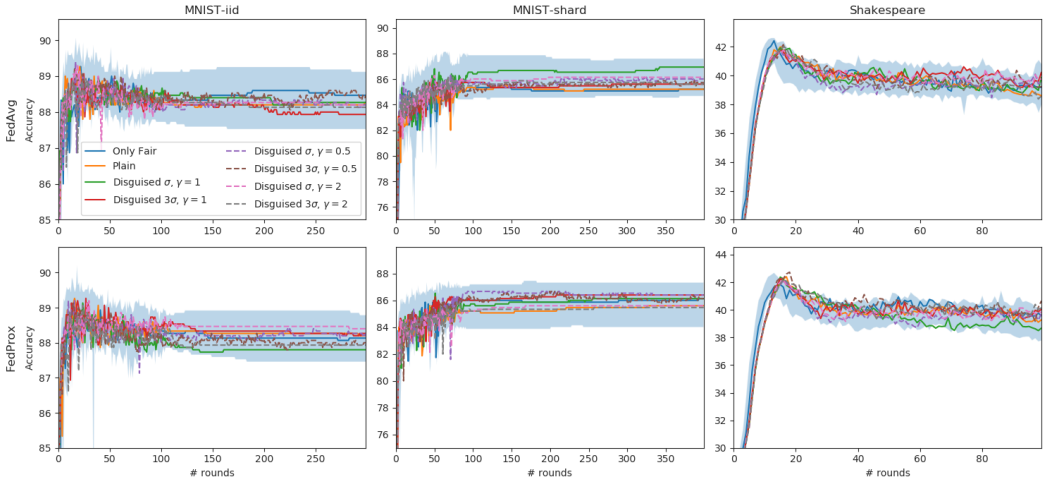

Figure 6: Accuracy performances for (a) FedAvg and (b) FedProx in the different experimental scenarios (5 local training epochs). The shaded blue region indicates the variability of federated learning model with fair clients only, estimated from 30 different training initialization.

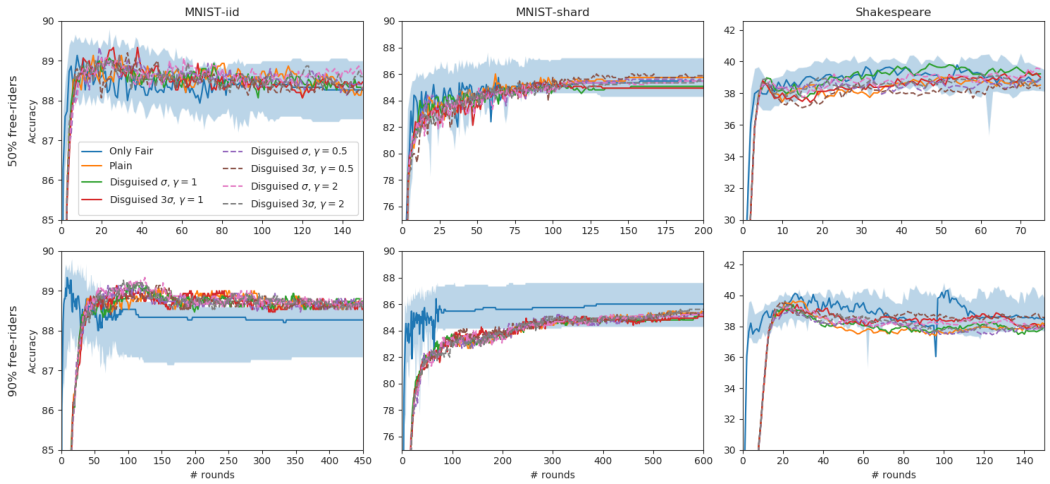

Figure 7: Plots for FedAvg and $E = 20$. Accuracy performances for FedAvg according to the number of free-riders participating in the learning process: 50% (top), and 90% (bottom) of the total amount of clients. The shaded blue region indicates the variability of federated learning model with fair clients only, estimated from 30 different training initialization.

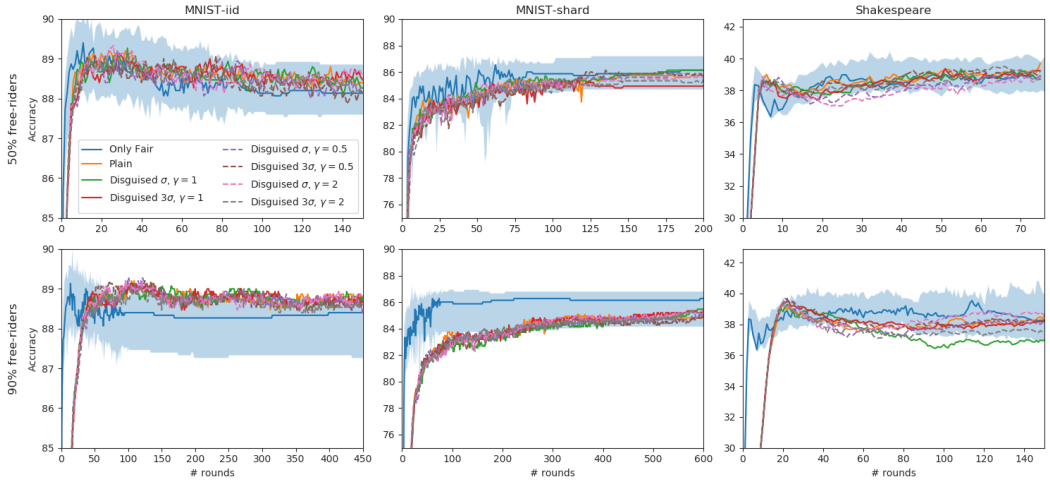

Figure 8: Plots for FedProx and $E = 20$. Accuracy performances for FedProx according to the number of free-riders participating in the learning process: 50% (top), and 90% (bottom) of the total amount of clients.

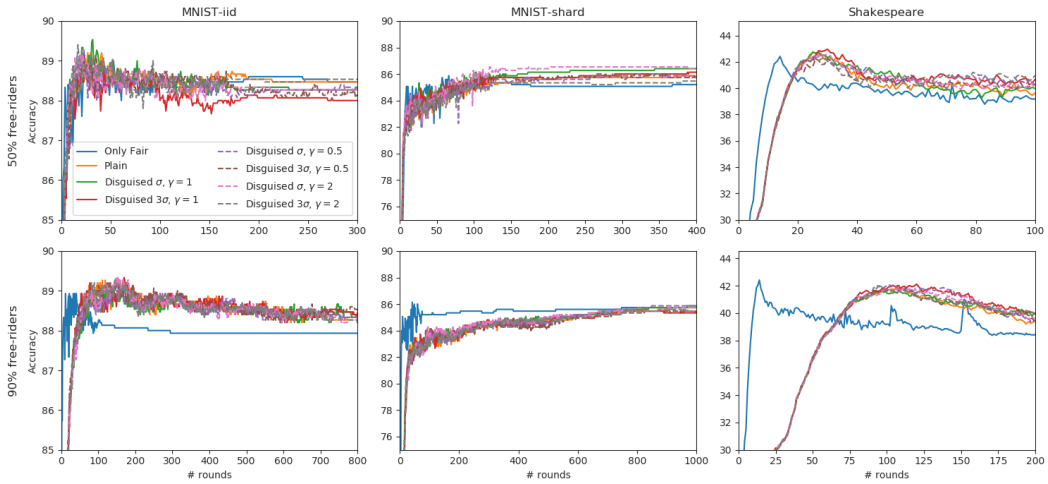

Figure 9: Plots for FedAvg and $E = 5$. Accuracy performances for FedAvg according to the number of free-riders participating in the learning process: 50% (top), and 90% (bottom) of the total amount of clients.

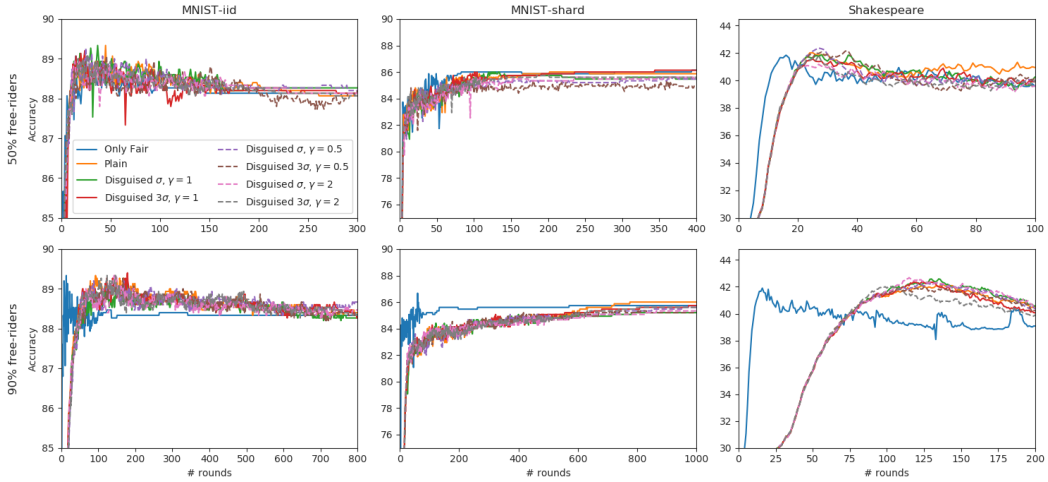

Figure 10: Plots for FedProx and $E = 5$. Accuracy performances for FedProx according to the number of free-riders participating in the learning process: 50% (top), and 90% (bottom) of the total amount of clients.

C.2 L2 norm and KS test

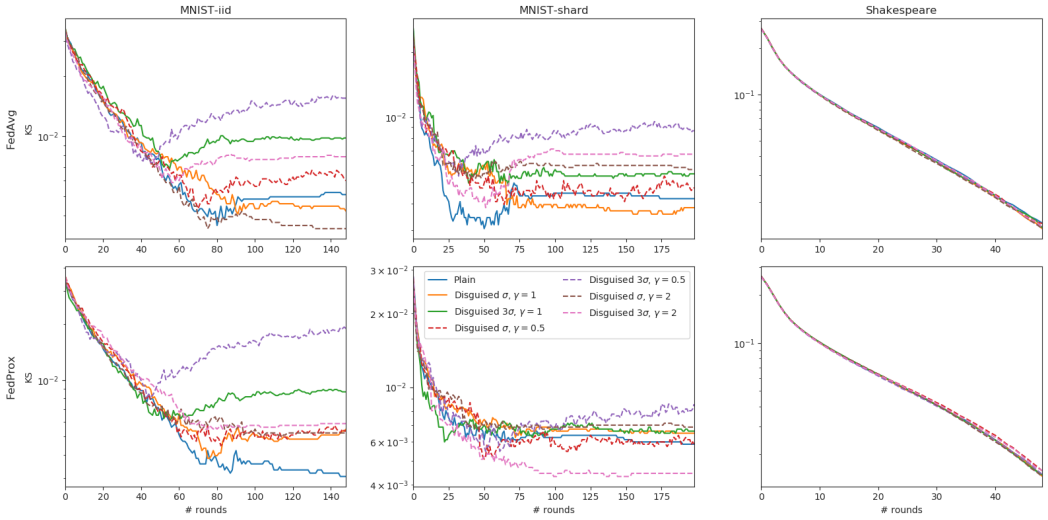

Figure 11: KS test for (a) FedAvg and (b) FedProx in the different experimental scenarios (20 local training epochs).

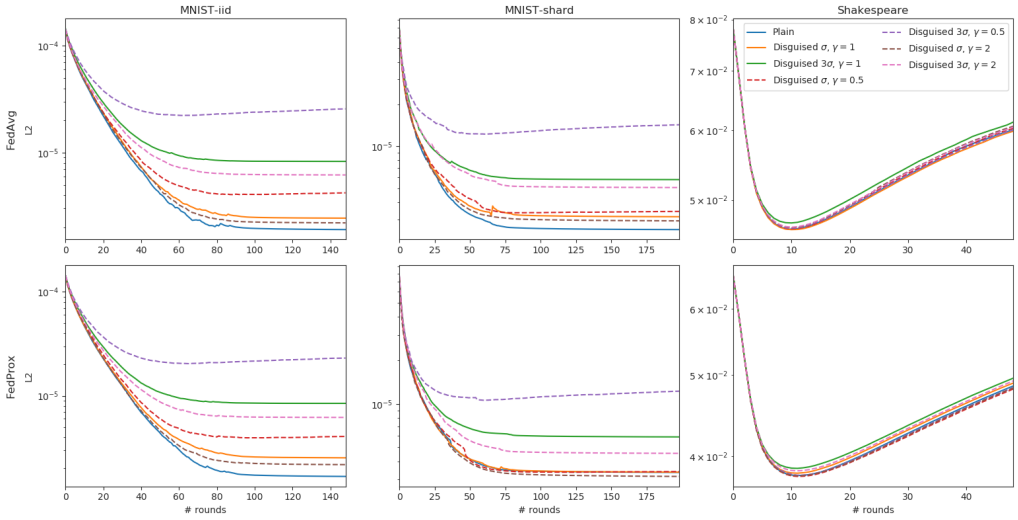

Figure 12: L2 norm for (a) FedAvg and (b) FedProx in the different experimental scenarios (20 local training epochs).

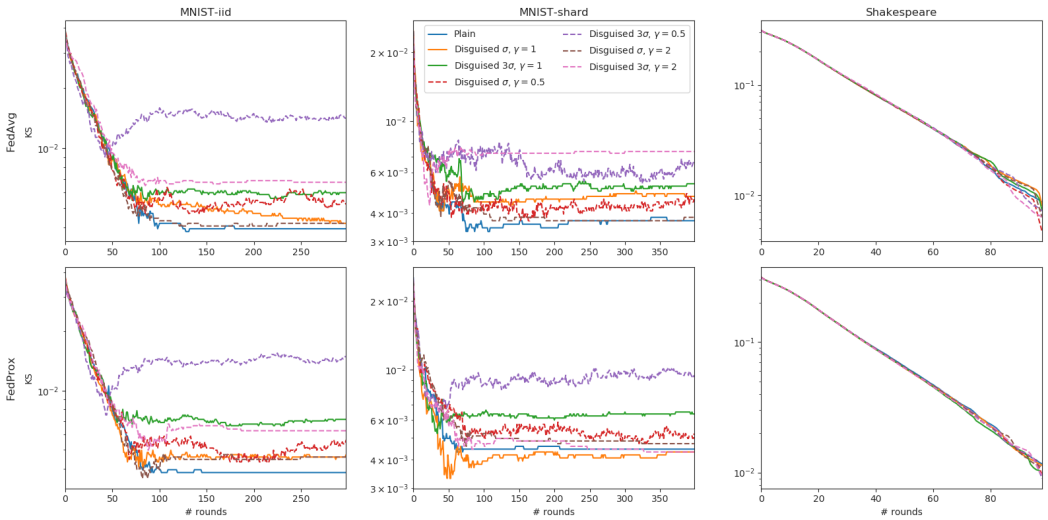

Figure 13: KS test for (a) FedAvg and (b) FedProx in the different experimental scenarios (5 local training epochs).

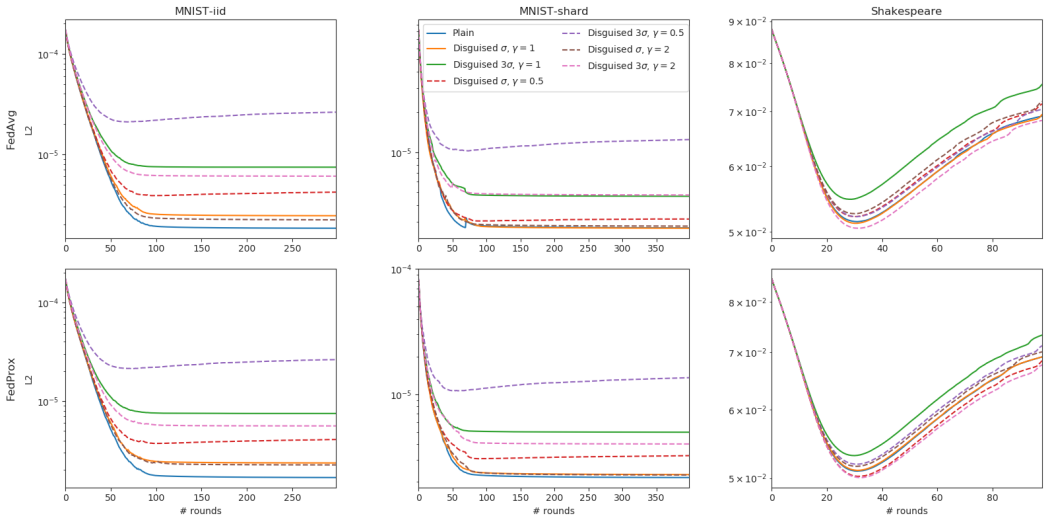

Figure 14: L2 norm for (a) FedAvg and (b) FedProx in the different experimental scenarios (5 local training epochs).

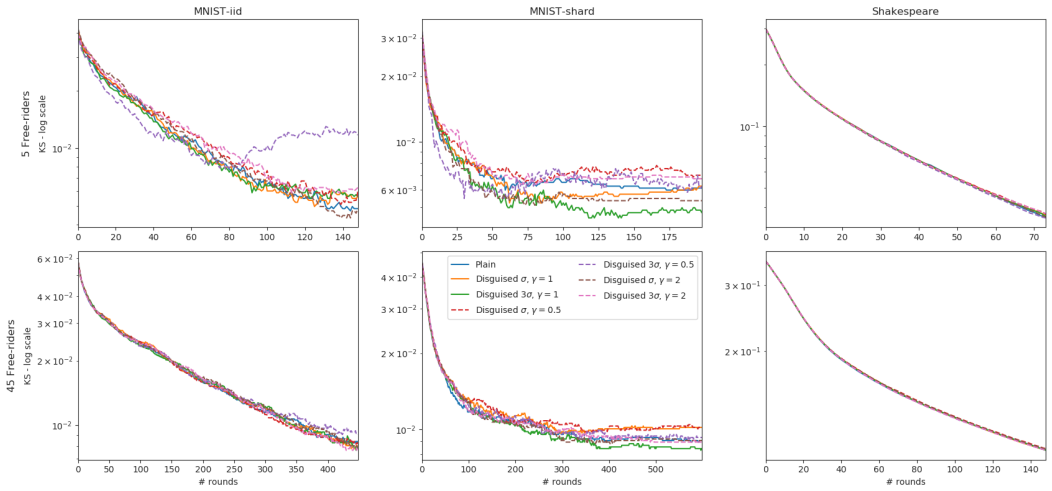

Figure 15: Plots for FedAvg and $E = 20$. KS Test for FedAvg according to the number of free-riders participating in the learning process: 50% (top), and 90% (bottom) of the total amount of clients.

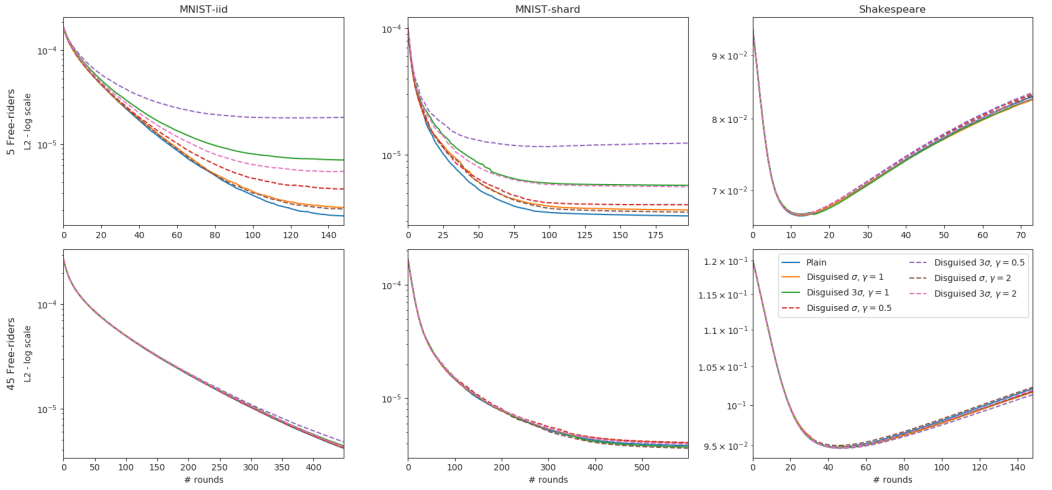

Figure 16: Plots for FedAvg and $E = 20$. L2 norm for FedAvg according to the number of free-riders participating in the learning process: 50% (top), and 90% (bottom) of the total amount of clients.

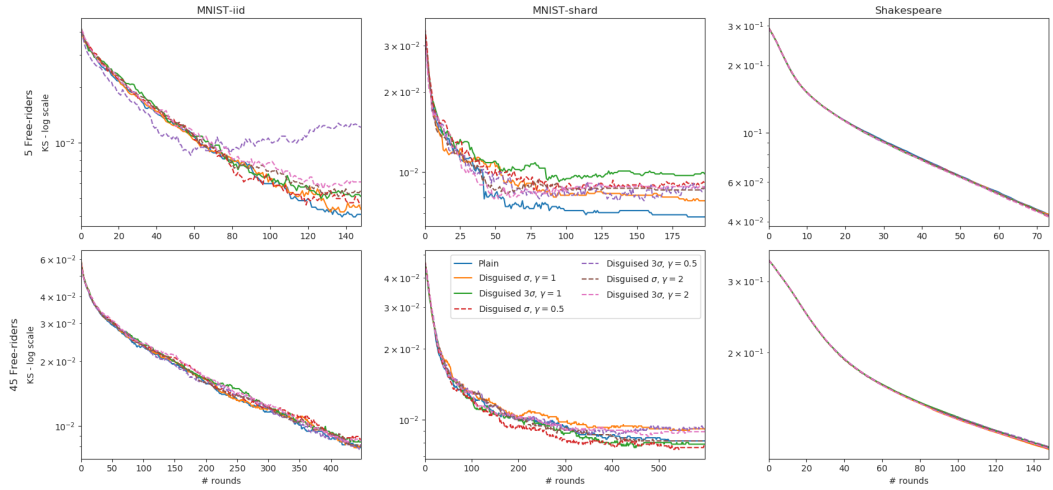

Figure 17: Plots for FedProx and $E = 20$. KS test for FedProx according to the number of free-riders participating in the learning process: 50% (top), and 90% (bottom) of the total amount of clients.

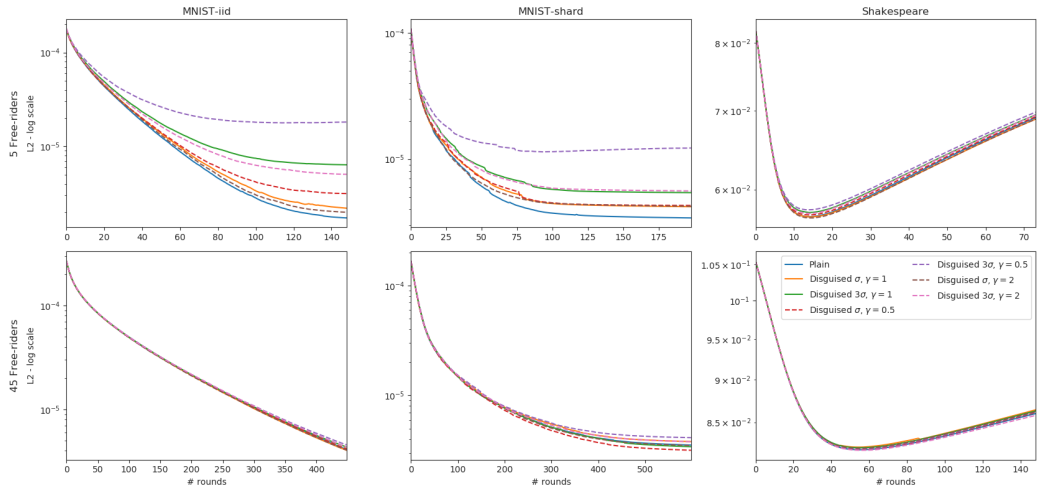

Figure 18: Plots for FedProx and $E = 20$. L2 norm for FedProx according to the number of free-riders participating in the learning process: 50% (top), and 90% (bottom) of the total amount of clients.

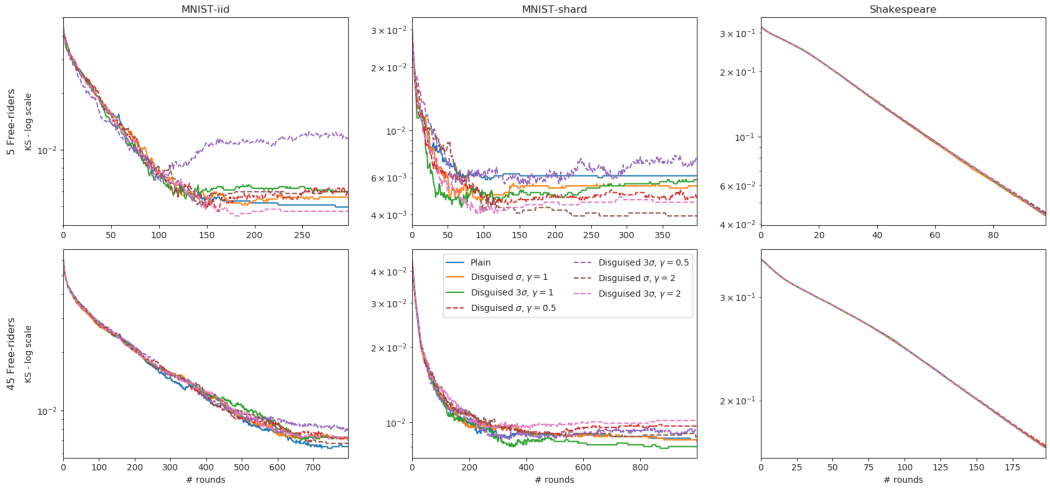

Figure 19: Plots for FedAvg and $E = 5$. KS Test for FedAvg according to the number of free-riders participating in the learning process: 50% (top), and 90% (bottom) of the total amount of clients.

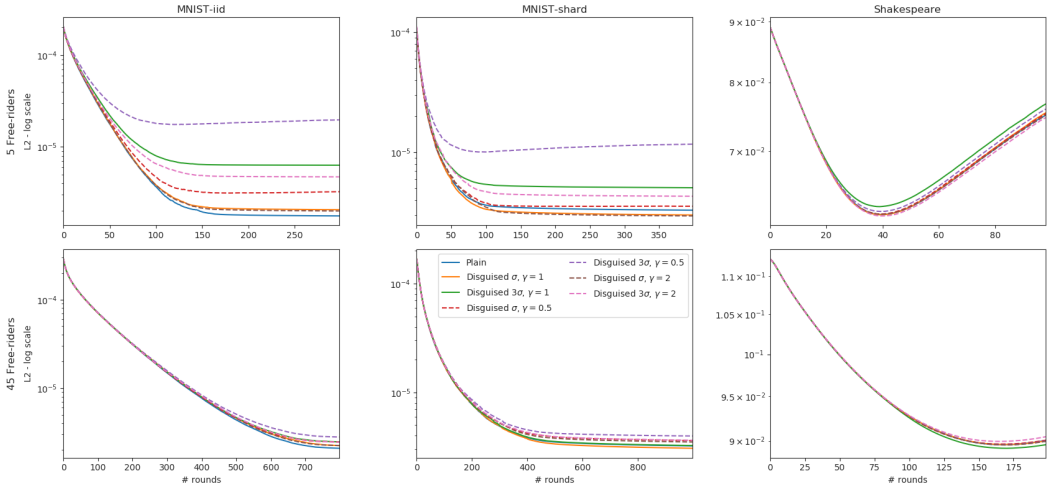

Figure 20: Plots for FedAvg and $E = 5$. L2 norm for FedAvg according to the number of free-riders participating in the learning process: 50% (top), and 90% (bottom) of the total amount of clients.

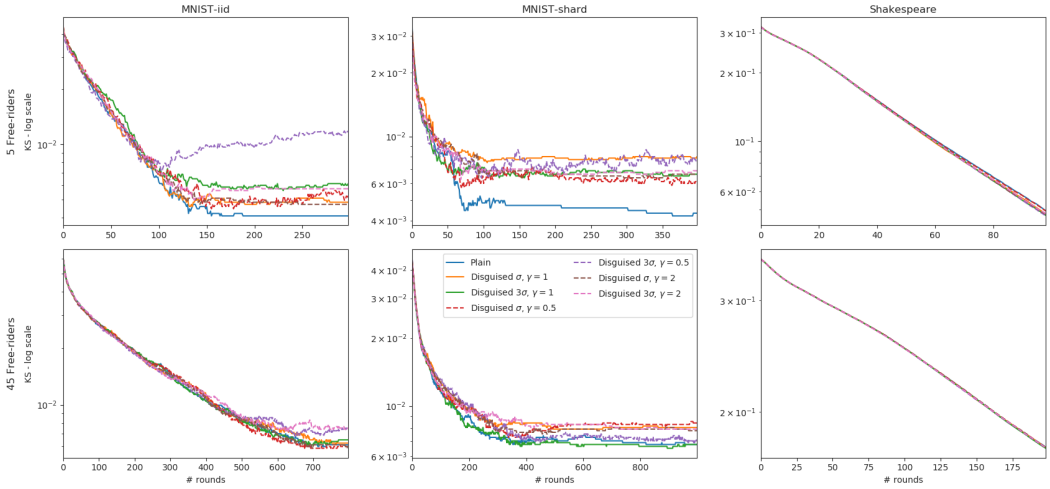

Figure 21: Plots for FedProx and $E = 5$. KS test for FedProx according to the number of free-riders participating in the learning process: 50% (top), and 90% (bottom) of the total amount of clients.

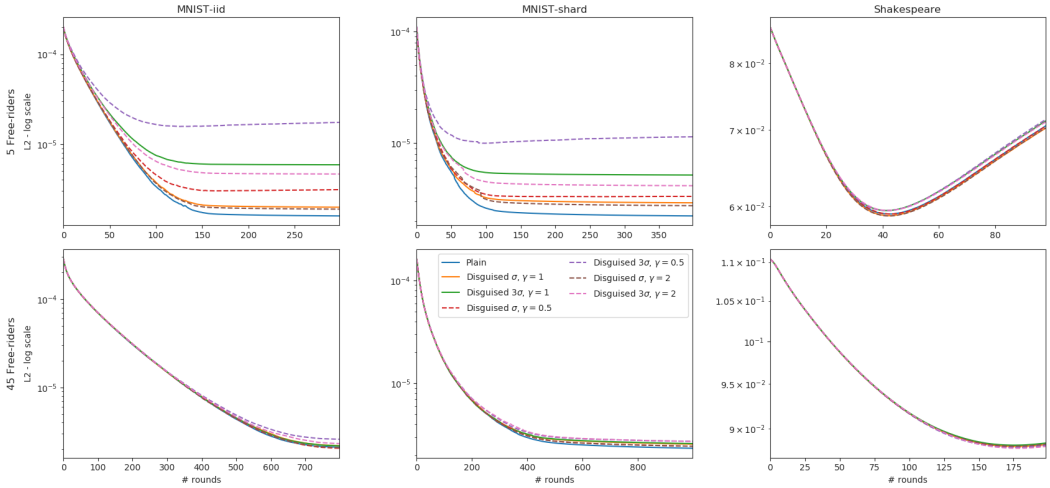

Figure 22: Plots for FedProx and $E = 5$. L2 Norm for FedProx according to the number of free-riders participating in the learning process: 50% (top), and 90% (bottom) of the total amount of clients.