Maximizing Global Model Appeal in Federated Learning

Anonymous authors
Paper under double-blind review

Abstract

Federated learning (FL) aims to collaboratively train a global model using local data from a network of clients. To warrant collaborative training, each federated client may expect the resulting global model to satisfy some individual requirement, such as achieving a certain loss threshold on their local data. However, in real FL scenarios, the global model may not satisfy the requirements of all clients in the network due to the data heterogeneity across clients. In this work, we explore the problem of *global model appeal* in FL, which we define as *the total number of clients* that find that the global model satisfies their individual requirements. We discover that global models trained using traditional FL approaches can result in a significant number of clients unsatisfied with the model based on their local requirements. As a consequence, we show that global model appeal can directly impact how clients participate in training and how the model performs on new clients at inference time. Our work proposes MaxFL, which maximizes the number of clients that find the global model appealing. MaxFL achieves a 22-40% and 18-50% improvement in the test accuracy of training clients and (unseen) test clients respectively, compared to a wide range of FL approaches that tackle data heterogeneity, aim to incentivize clients, and learn personalized/fair models.

1 Introduction

Federated learning (FL) is a distributed learning framework that considers training a machine learning model using a network of clients (e.g., mobile phones, hospitals), without directly sharing client data with a central server McMahan et al. (2017). FL is typically performed by aggregating clients' updates over multiple communication rounds to produce a global model Kairouz et al. (2019). In turn, each client may have its own requirement that it expects to be met by the resulting global model under different settings such as at inference or with some fine-tuning. For example, clients such as hospitals or edge devices may expect that the global model performs *at least* better than a local model trained in isolation on the client's limited local data before contributing to FL training. Unfortunately, due to data heterogeneity across the clients, the global model may fail to meet the requirements of all clients Yu et al. (2020).

Previously, a plethora of works in FL on techniques such as variance reduction Karimireddy et al. (2019), personalization Fallah et al. (2020); Dinh et al. (2020), and fairness Li et al. (2019) have proposed to train a global model or several personalized models that can better cater to the needs of the clients or the server. However, prior work has not directly focused on *the total number of clients* that are satisfied with *the single global model* based on their individual requirements, and has not explored how this may affect the training of the global model from the server's perspective when clients have the autonomy to freely join or leave the federation. Recent closely related works focus on clients' incentives from a game-theoretic lens Han et al. (2022); Kang et al. (2019); Zhang et al. (2021) and establish useful insights for simple linear tasks, but it is difficult to extend these to practical non-convex machine learning problems. Other related works design strategies specifically to prevent client dropout Wang & Xu (2022); Gu et al. (2021), but these algorithms are stateful, i.e., they require saving the gradient information from previously participating clients for the current updates, making them impractical to implement in cross-device settings Kairouz et al. (2019). Moreover, these works only provide convergence guarantees to the global minimum with respect to the standard FedAvg

objective McMahan et al. (2017), lacking theoretical justification that their objectives can yield solutions that guarantee more participating clients compared to other classic FL objectives, even in simplified settings; we provide such guarantees for our proposed objective for mean estimation problems, which helps to shed light on our strong empirical performance (Section 2.2).

Proposing a new and formal metric to evaluate FL systems, we define that a global model is *appealing* to a client if it satisfies the client's specified requirement, such as incurring at most some max training loss. Subsequently, we define the number of clients which find the global model appealing as *global model appeal* (GM-Appeal; formalized in Definition 1). We show that having a high global model appeal is critical for the server to maintain a large pool of clients to select from for training, and for gathering additional willingly participating clients. This is especially true in the light of clients possibly opting out of FL due to the significant costs associated with training (e.g., computational overhead, privacy risks, logistical challenges), which is a practical concern not typically considered in prior work. With a larger pool of clients to select from, a server can not only improve privacy-utility tradeoffs McMahan et al. (2018), but can also improve test accuracy on participating clients, and produce a global model that generalizes better at inference to new unseen clients (see Fig. 1 and Table 1).

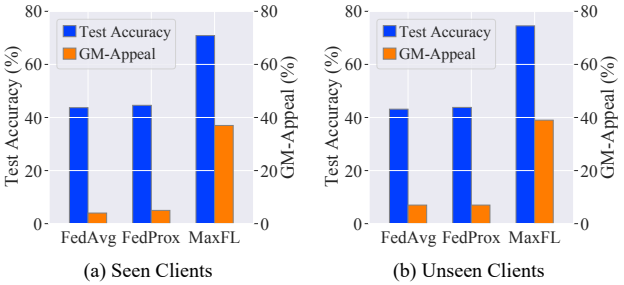


Figure 1: Test acc. and GM-Appeal of the global model for FMNIST. A higher GM-Appeal results in a higher test accuracy for both the seen clients that have participated during training and unseen clients that have not, due to the server having a larger pool of clients to select from. MaxFL, which aims to maximize GM-Appeal, results in the highest test accuracy compared to the other baselines that do not consider GM-Appeal.

In this work, we seek to understand: (1) What benefits exist when maximizing global model appeal relative to other common federated modeling approaches, and (2) What strategies exist to maximize global model appeal in federated settings. Our key contributions are summarized as follows:

- We introduce the notion of global model appeal (referred to as GM-Appeal), the fraction of clients that have their local requirements met by the global model. We then propose MaxFL, an objective that directly maximizes global model appeal, and show that having a high global model appeal can lead to better test accuracy on training clients, as well as better generalization to new unseen clients at inference time.
- We theoretically show for mean estimation that the MaxFL objective yields a solution that guarantees higher GM-Appeal than standard FL objectives and provide convergence guarantees for our MaxFL solver which allows partial client participation, is applicable to non-convex objectives, and is stateless (does not require clients to maintain local parameters during training).
- We empirically evaluate the performance of MaxFL to thoroughly understand the benefits of maximizing GM-Appeal with experiments where i) clients can flexibly opt-out of training, ii) there are new incoming (unseen) clients, iii) there are Byzantine clients participating in training, and iv) clients perform personalization via local fine-tuning with the global model.
- We show that MaxFL significantly improves the global model appeal in practice, leading to a 22-40% and 18-50% test accuracy improvement for the seen clients and unseen clients respectively, compared to a wide range of FL methods, including those that tackle data heterogeneity, aim for variance reduction or incentivizing clients, or provide personalization or fairness.

Overall, our goal in comparing MaxFL with a variety of other FL methods that have varying goals is not necessarily to compete against these methods, but rather to understand and demonstrate the potential benefits of our proposed notion of global model appeal relative to other objectives under different scenarios, such as with flexible client participation or new incoming clients. As global model appeal has not been studied

previously in FL, our work is the first to explore its possible implications, and then propose an objective to train a global model that can maximize the number of clients whose individual requirements are satisfied. We hope our proposed perspective of viewing and evaluating FL systems can inspire future works in different FL scenarios and applications where client participation is not necessarily taken for granted, and can potentially be used in conjunction with prior approaches in FL (e.g., see Section D.3). We provide a more detailed review of prior work and related areas of fairness, personalization and client incentives in Section 4.

2 Problem Formulation

Setting. We consider a setup where M clients are connected to a central server to collaboratively train a global model. For each client $k \in [M]$, its true loss function is given by $f_k(\mathbf{w}) = \mathbb{E}_{\xi \sim \mathcal{D}_k}[\ell(\mathbf{w}, \xi)]$ where $\mathcal{D}_k$ is the true data distribution of client k, and $\ell(\mathbf{w}, \xi)$ is the composite loss function for the model $\mathbf{w} \in \mathbb{R}^d$ for data sample ξ. In practice, each client only has access to its local training dataset $\mathcal{B}_k$ with $|\mathcal{B}_k| = N_k$ data samples sampled from $\mathcal{D}_k$. Client k's empirical loss function is $F_k(\mathbf{w}) = \frac{1}{|\mathcal{B}_k|} \sum_{\xi \in \mathcal{B}_k} \ell(\mathbf{w}, \xi)$. While some of the take-aways of our work (e.g., improving performance on unseen clients) may be more specific to cross-device applications, our general setup and method are applicable to both cross-device and cross-silo FL.

Defining Global Model Appeal. Each client's natural aim is to find a model that minimizes its true local loss $f_k(\mathbf{w})$. Clients can have different thresholds of how small this loss should be, and we denote such self-defined threshold for each client as $\rho_k,\ k \in [M]$. For instance, each client can perform solo training on its local dataset $\mathcal{B}_k$ to obtain an approximate local model $\widehat{\mathbf{w}}_k$ and have its threshold to be the true loss from this local model, i.e., $\rho_k = f_k(\widehat{\mathbf{w}}_k)$[1]. Based on these client requirements, we provide the formal definition of global model appeal below:

Definition 1 (Global Model Appeal)**.** *A global model* $\mathbf{w}$ *is said to be appealing to client* $k \in [M]$ *if* $f_k(\mathbf{w}) < \rho_k$*, i.e., the global model* $\mathbf{w}$ *yields a smaller local true loss than the self-defined threshold of the client. Accordingly, we define the fraction of clients to which the global model is appealing as global model appeal (GM-Appeal) with* $\mathbb{I}$ *being the indicator function:*

$$\textit{GM-Appeal} = \frac{1}{M} \sum_{k=1}^{M} \mathbb{I}\{f_k(\mathbf{w}) < \rho_k\} \tag{1}$$

Our GM-Appeal metric measures the exact *fraction of clients* that find the global model appealing by focusing on whether the global model satisfies the clients' requirements or not instead of looking at the gap between $f_k(\mathbf{w})$ and ρ_k. Another variation of (1) could be to measure the margin $\sum_k \max\{\rho_k - f_k(\mathbf{w}), 0\}$, but this does not capture the motivation behind our work which is to understand how the number of clients that find the global model appealing affects the global model performance.

Why Explore GM-Appeal in FL? GM-Appeal measures how many clients have their requirements satisfied by the global model. Thus, it can gauge important characteristics of the server's global model such as i) how many clients are likely to dropout with the current global model or ii) how many new incoming clients that do not have the capacity for additional training will likely be satisfied with the current global model without any additional training to the model. Ultimately, a high global model appeal can lead to a larger pool of clients for the server to select from. The standard FL objective McMahan et al. (2017) or its popular variants Karimireddy et al. (2019); Li et al. (2020); Fallah et al. (2020) does not consider whether the global model satisfies the clients' requirements, and implicitly assumes that the server will have a large number of clients to select from. However, this may not necessarily be true if clients are allowed to dropout when they find the global model unappealing. We show that acquiring a larger pool of clients by improving global model appeal is useful for improving the global model for both the seen clients at training as well as the unseen clients at inference (see Fig. 1 and Table 1). In fact, we find that other baselines such as those that aim to tackle data heterogeneity, improve fairness, or provide personalization have low GM-Appeal, leading to a large number of clients opting out. Due to this, the global model is trained on just a few limited data points, resulting in poor performance. Our work explores a new notion of global model appeal, showing its significance in FL.

[1] The client can have held-out data used for calculating the true loss $f_k(\cdot)$ or use its training data as a proxy. We explain in more detail of defining ρ_k in Section 3.

2.1 Proposed MaxFL Objective

In this section, we first introduce MaxFL whose aim is to train a global model that maximizes GM-Appeal. A naïve approach can be to directly maximize GM-Appeal defined in (1) as follows:

$$\underset{\mathbf{w}}{\operatorname{argmax}}\, \text{GM-Appeal} = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{k=1}^{M} \operatorname{sign}(f_k(\mathbf{w}) - \rho_k). \tag{2}$$

where $\operatorname{sign}(x) = 1$ if $x \geq 0$ and 0 otherwise. There are two immediate difficulties in minimizing (2). First, clients may not know their true data distribution $\mathcal{D}_k$ to compute $f_k(\mathbf{w}) - \rho_k$. Second, the sign function makes the objective nondifferentiable and limits the use of common gradient-based methods. We resolve these issues by proposing a "proxy" for (2) with the following relaxations.

i) Replacing the Sign function with the Sigmoid function $\sigma(\cdot)$: Replacing the non-differentiable 0-1 loss with a smooth differentiable loss is a standard tool used in optimization Nguyen & Sanner (2013); Masnadi-shirazi & Vasconcelos (2008). Given the many candidates (e.g. hinge loss, ReLU, sigmoid), we find that using the sigmoid function is essential for our objective to faithfully approximate the true objective in (2). We further discuss the theoretical implications of using the sigmoid loss in Section 2.2.

ii) Replacing $f_k(\mathbf{w})$ with $F_k(\mathbf{w})$: As clients do not have access to their true distribution $\mathcal{D}_k$ to compute $f_k(\cdot)$ we propose to use an empirical estimate $\sigma(F_k(\mathbf{w}) - \rho_k)$. This is again similar to what is done in standard FL where we minimize $F_k(\mathbf{w})$ instead of $f_k(\mathbf{w})$ at client k. Note that the global model $\mathbf{w}$ is trained on the data of all clients, making it unlikely to overfit to the local data of any particular client, leading to $f_k(\mathbf{w}) \approx F_k(\mathbf{w})$, which we also show empirically in Appendix D.3.

With the two relaxations above, we present our proposed MaxFL objective:

$$\text{MaxFL Obj.:}\ \min_{\mathbf{w}} \widetilde{F}(\mathbf{w}) = \min_{\mathbf{w}} \frac{1}{M} \sum_{i=1}^{M} \widetilde{F}_i(\mathbf{w}),\ \text{ where } \widetilde{F}_i(\mathbf{w}) := \sigma(F_i(\mathbf{w}) - \rho_i). \tag{3}$$

Before presenting our proposed solver for the MaxFL objective, we first present a motivating toy example with mean estimation which shows that MaxFL's objective leads to a different solution that has higher GM-Appeal compared to the solution obtained from the classic FL objective.

2.2 Toy Example: Maximizing GM-Appeal in Mean Estimation

We consider a toy setup with $M = 2$ clients where the true loss function at each client is given by $f_k(w) = (w - \theta_k)^2$. In practice, clients only have N_k samples drawn from the distribution given by $e_{k,j} \sim \mathcal{N}(\theta_k, \nu^2)$, $\forall j \in [N_k]$. We further assume that the empirical loss function at each client is given by $F_k(w) = (w - \widehat{\theta}_k)^2 + (\widehat{\theta}_k - \theta_k)^2$ where $\widehat{\theta}_k$ is the empirical mean, $\widehat{\theta}_k = \frac{1}{|\mathcal{B}_k|} \sum_{j=1}^{N_k} e_{k,j}$. It is easy to see that the minimizer of $F_k(w)$ is the empirical mean $\widehat{\theta}_k$. Thus, we set the solo-trained model at each client as $\widehat{w}_k = \widehat{\theta}_k$ and the loss threshold requirement at a client as $\rho_k = F_k(\widehat{w}_k) = (\widehat{\theta}_k - \theta_k)^2$.

GM-Appeal for Standard FL Model Decreases Exponentially with Heterogeneity. For simplicity let us assume $N_1 = N_2 = N$ where $\gamma^2 = \nu^2/N$ is the variance of the local empirical means and $\gamma_G^2 = ((\theta_1 - \theta_2)/2)^2 > 0$ is the measure of heterogeneity between the true means. The standard FL objective will always set the FL model to be the average of the local empirical means (i.e. $w = (\widehat{\theta}_1 + \widehat{\theta}_2)/2$) and does not take into account the heterogeneity among the clients. As a result, the GM-Appeal of the global model decreases *exponentially* as γ_G^2 increases.

Lemma 2.1. *The expected GM-Appeal of the standard FL model is upper bounded by* $2\exp\left(-\gamma_G^2/(5\gamma^2)\right)$, *where the expectation is taken over the randomness in the local datasets* $\mathcal{B}_1, \mathcal{B}_2$.

Maximizing GM-Appeal with Relaxed Objective. We now maximize the GM-Appeal for this setting by solving a relaxed version of the objective in (2) as proposed earlier. We replace the true loss $f_k(\cdot)$ with the empirical loss $F_k(\cdot)$ and replace the 0-1 (sign) loss with a differentiable approximation $h(\cdot)$. We first show that setting $h(\cdot)$ to be a standard convex surrogate for the 0-1 loss (e.g. log loss, exponential loss, ReLU) leads to our new objective behaving the same as the standard FL objective.

Lemma 2.2. *Let h be any function that is convex, twice differentiable, and strictly increasing in $[0, \infty)$. Then our relaxed objective is strictly convex and has a unique minimizer at $w^* = (\widehat{\theta}_1 + \widehat{\theta}_2)/2$.*

MaxFL Objective Leads to Increased GM-Appeal. Based on Lemma 2.2, we see that we need nonconvexity in $h(\cdot)$ for the objective to behave differently than standard FL. We set $h(x) = \sigma(x) = \exp(x)/1 + \exp(x)$, as proposed in our MaxFL objective in (3) and find that the MaxFL objective *adapts* to the empirical heterogeneity parameter $\widehat{\gamma}_G^2 = (\widehat{\theta}_2 - \widehat{\theta}_1/2)^2$. If $\widehat{\gamma}_G^2 < 1$ (small data heterogeneity), the objective encourages collaboration by setting the global model to be the average of the local models. Conversely, if $\widehat{\gamma}_G^2 > 2$ (large data heterogeneity), the objective encourages *separation* by setting the global model close to the local model of either client (see Fig. 2 below). Based on this observation, we have the following theorem.

Theorem 2.1. *Let w be a local minima of the MaxFL objective. The expected GM-Appeal using w is lower bounded by $\exp\left(-\gamma^{-2}\right)/16$ where the expectation is over the local datasets $\mathcal{B}_1, \mathcal{B}_2$.*

Observe that even with $\gamma_G^2 \gg 0$, MaxFL will keep satisfying the requirement of at least one client by adapting its objective accordingly. We show the behavior of MaxFL in a 3-client setup which further highlights the non-trivialness of our proposed MaxFL's formulation in Appendix A along with the simulation details for Fig. 2. Details of our proof in this section can be found in Appendix B.

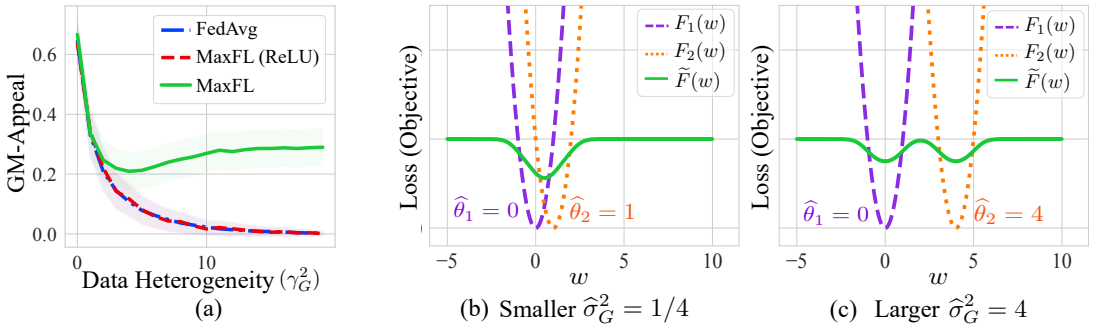


Figure 2: (a): GM-Appeal for FedAvg decays exponentially while GM-Appeal for MaxFL is lower bounded. Replacing the sigmoid approximation with ReLU in MaxFL leads to the same solution as FedAvg. (b-c): MaxFL adapts to the heterogeneity of the problem—for small heterogeneity it encourages collaboration by having a single global minima, for large heterogeneity it encourages separation by having far away local minimas.

3 Proposed MaxFL Solver

In this section, we present our MaxFL objective's solver. The MaxFL algorithm enjoys the following properties: i) uses the same local SGD procedure as in standard FedAvg, ii) allows partial client participation, and iii) is stateless. By stateless, we mean that clients do not carry varying local parameters throughout training rounds, preventing issues from stale parameters Wang et al. (2021). With the sigmoid approximation of sign loss and for differentiable $F_k(\mathbf{w})$, our objective $\widetilde{F}(\mathbf{w})$ in (3) is differentiable and can be minimized with gradient descent and its variants. Its gradient is given by:

$$\nabla \widetilde{F}(\mathbf{w}) = \frac{1}{M} \sum_{k=1}^{M} \underbrace{(1 - \widetilde{F}_k(\mathbf{w}))\widetilde{F}_k(\mathbf{w})}_{\text{aggregating weight}:=q_k(\mathbf{w})} \nabla F_k(\mathbf{w}). \tag{4}$$

Observe that $\nabla \widetilde{F}(\mathbf{w})$ is a **weighted aggregate** of the gradients of the clients' empirical losses, similar in spirit to the gradient $\nabla F(\mathbf{w})$ in standard FL. The key difference is that in MaxFL, the weights $q_k(\mathbf{w}) := (1 - \widetilde{F}_k(\mathbf{w}))\widetilde{F}_k(\mathbf{w})$ depend on how much the global model appeals to the clients and are dynamically updated based on the current model $\mathbf{w}$, as we discuss below.

Behavior of the Aggregation Weights. For a given $\mathbf{w}$, the aggregation weights $q_k(\mathbf{w})$ depend on the *GM-Appeal Gap*, $F_k(\mathbf{w}) - \rho_k$ (see Fig. 3). When $F_k(\mathbf{w}) \ll \rho_k$, the global model $\mathbf{w}$ sufficiently meets the

client's requirement. Therefore, MaxFL sets $q_k(\mathbf{w}) \approx 0$ to focus on the updates of other clients. Similarly, if $F_k(\mathbf{w}) \gg \rho_k$, MaxFL sets $q_k(\mathbf{w}) \approx 0$. This is because $F_k(\mathbf{w}) \gg \rho_k$ implies that the current model $\mathbf{w}$ is incompatible with the requirement of client k and hence it is better to avoid optimizing for this client at the risk of sacrificing other clients' requirements. MaxFL gives the highest weight to clients for which the global model performs similarly to the clients' requirements since this allows it to increase the GM-Appeal without sabotaging other clients' requirements.

A Practical MaxFL Solver. Directly minimizing the MaxFL objective using gradient descent can be slow to converge and impractical, as it requires all clients to be available for training. Instead, we propose a practical MaxFL algorithm, which uses multiple local updates at each client to speed up convergence as done in standard FL McMahan et al. (2017) and allow partial client availability. We use the superscript (t, r) to denote the communication round t and the local iteration index r. In each round t, the server selects a new set of clients $\mathcal{S}^{(t,0)}$ uniformly at random and sends the most recent global model $\mathbf{w}^{(t,0)}$ to the clients in $\mathcal{S}^{(t,0)}$. Clients in $\mathcal{S}^{(t,0)}$ perform τ local iterations with a learning rate η_l to calculate their updates as $\mathbf{w}_k^{(t,r+1)} = \mathbf{w}_k^{(t,r)} - \eta_l \mathbf{g}(\mathbf{w}_k^{(t,r)}, \xi_k^{(t,r)}), \forall\ r \in \{0, ..., \tau - 1\}$ where $\mathbf{g}(\mathbf{w}_k^{(t,r)}, \xi_k^{(t,r)}) = \frac{1}{b} \sum_{\xi \in \xi_k^{(t,r)}} \nabla f(\mathbf{w}_k^{(t,r)}, \xi)$ is the stochastic gradient computed using a mini-batch $\xi_k^{(t,r)}$ of size b that is randomly sampled from client k's local dataset $\mathcal{B}_k$. The weight $q_k(\mathbf{w}_k^{(t,0)})$ can be computed at each client by calculating the loss over its training data with $\mathbf{w}_k^{(t,0)}$, which is a simple inference step. Clients in $\mathcal{S}^{(t,0)}$ then send their local updates $\Delta \mathbf{w}_k^{(t,0)} := \mathbf{w}_k^{(t,\tau)} - \mathbf{w}_k^{(t,0)}$ and weights $q_k(\mathbf{w}_k^{(t,0)})$ back to the server, which updates the global model as $\mathbf{w}^{(t+1,0)} = \mathbf{w}^{(t,0)} - \eta_g^{(t,0)} \sum_{k \in \mathcal{S}^{(t,0)}} q_k(\mathbf{w}^{(t,0)}) \Delta \mathbf{w}_k^{(t,0)}$ where $\eta_g^{(t,0)} = \frac{\eta_g}{\sum_{k \in \mathcal{S}^{(t,0)}} q_k(\mathbf{w}^{(t,0)}) + \epsilon}$ is the adaptive server learning rate with global learning rate η_g and $\epsilon > 0$. We discuss the reasoning for such a learning rate below. The pseudo-code for our MaxFL solver can be found in Algorithm 1 in Appendix D.1.

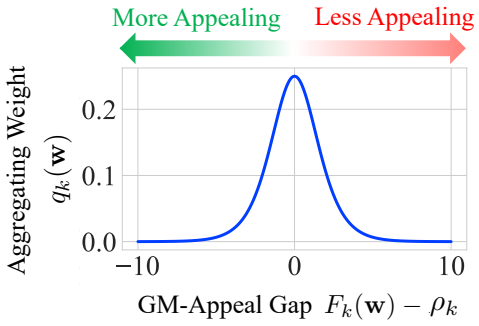


Figure 3: Aggregating weight $q_k(\mathbf{w})$ versus the GM-Appeal gap defined as $F_k(\mathbf{w}) - \rho_k$ for any client $k \in [M]$.

Adaptive Server Learning Rate for MaxFL. With L_c continuous and L_s smooth $F_k(\mathbf{w})$, $\forall k \in [M]$ (see Assumption 3.1), the objective $\widetilde{F}(\mathbf{w})$ is $\widetilde{L}_s$ smooth where $\widetilde{L}_s = \frac{L_s}{M} \sum_{k=1}^{M} q_k(\mathbf{w}) + \frac{L_c}{4}$ (see Appendix C). Hence, the optimal learning rate $\tilde{\eta}$ for the MaxFL is given by, $\widetilde{\eta} = 1/\widetilde{L}_s = M\eta / \left(\sum_{k=1}^{M} q_k(\mathbf{w}) + \epsilon\right)$, where $\eta = \frac{1}{L_s}$ is the optimal learning rate for standard FL and $\epsilon = \frac{ML_c}{4L_s} > 0$ is a constant. The denominator of the optimal $\widetilde{\eta}$ is proportional to the sum of the aggregation weights $q_k(\mathbf{w})$ and acts as a dynamic normalizing factor. Therefore, we propose using an adaptive global learning rate $\eta_g^{(t,0)} = \eta_g / (\sum_{k \in \mathcal{S}^{(t,0)}} q_k(\mathbf{w}^{(t,0)}) + \epsilon)$ with hyperparameters η_g, ϵ.

Setting ρ_k as $F_k(\widehat{\mathbf{w}}_k)$ for MaxFL. One intuitive way to set ρ_k for each client is to set it as the training loss value $F_k(\widehat{\mathbf{w}}_k)$ where $\widehat{\mathbf{w}}_k$ is a client local model that is solo-trained with a few warm-up local SGD steps on its local data. The loss value only needs to be computed once and saved as a constant beforehand at each client. How to train the local model $\widehat{\mathbf{w}}_k$, such as deciding the number of steps of training or whether to add regularization, is dependent on the personal resources and requirements of the clients. Therefore, the quality of the local model (such as whether it has overfitted or underfitted) is not explicitly controlled by MaxFL. Nevertheless, we do provide an ablation study on the number of local SGD steps we take for training the local trained model and provide the results in Appendix D.3 which shows that MaxFL performance is in fact robust to the number of local SGD steps we take and the performance does not vary much on this number. In the essence, ρ_k is a client-dependent parameter that the client can choose, and our proposed MaxFL is a more general framework that can be used for any client-defined ρ_k. One might raise concerns that adversarial clients can send arbitrarily small ρ_k in the hope to get high weights, but as Fig. 3 clearly shows this will in fact make the client get smaller weights. Further, we add experiments with byzantine clients in Section 5 and show that MaxFL is robust to these scenarios.

Appeal-based Flexible Client Participation. It may appear that our MaxFL solver requires clients to always participate in FL if selected even when the global model does not appeal to them. However, our algorithm is easily modified to allow clients to participate flexibly during training depending on whether they find the global model appealing or not. For such appeal-based flexible client participation, we assume that clients are available for training if selected only during a few initial training rounds. After these rounds, clients are included in the pool of clients where the server can select the clients from only if they find the global model appealing. We demonstrate this extension of MaxFL with appeal-based flexible client participation in Table 1 and Table 2. These experiments show that with flexible client participation, retaining a high global model is *even more imperative* for the server to achieve good test accuracy and generalization performance. We also show that even after we allow clients to participate flexibly, MaxFL retains a significantly higher number of clients that find the global model appealing compared to the other baselines.

3.1 Convergence Properties of MaxFL

In this section, we show the convergence guarantees of MaxFL in Algorithm 1. Our convergence analysis shows that the gradient norm of our global model goes to zero, and therefore we converge to a stationary point of our objective $\widetilde{F}(\mathbf{w})$. First, we introduce the assumptions and definitions below.

Assumption 3.1 (Continuity & Smoothness of $F_k(\mathbf{w})$, $\forall\, k$)**.** *The local objective functions $F_1(\mathbf{w})$, ..., $F_M(\mathbf{w})$, are L_c-continuous and L_s-smooth for any $\mathbf{w}$.*

Assumption 3.2 (Unbiased Stochastic Gradient with Bounded Variance for $F_k(\mathbf{w})$, $\forall\, k$)**.** *For mini-batch ξ_k uniformly sampled at random from $\mathcal{B}_k$, the resulting stochastic gradient is unbiased, i.e., $\mathbb{E}[\mathbf{g}_k(\mathbf{w}_k, \xi_k)] = \nabla F_k(\mathbf{w}_k)$, and its variance is bounded: $\mathbb{E}[\|\mathbf{g}_k(\mathbf{w}_k, \xi_k) - \nabla F_k(\mathbf{w}_k)\|^2] \leq \sigma_g^2$.*

Assumption 3.3 (Bounded Dissimilarity of $F(\mathbf{w})$)**.** *There exists $\beta^2 \geq 1$, $\kappa^2 \geq 0$ such that $\frac{1}{M}\sum_{i=1}^{M}\|\nabla F_i(\mathbf{w})\|^2 \leq \beta^2 \|\frac{1}{M}\sum_{i=1}^{M}\nabla F_i(\mathbf{w})\|^2 + \kappa^2$ for any $\mathbf{w}$.*

Assumption 3.1-3.3 are standard assumptions used in the optimization literature Stich (2019); Karimireddy et al. (2019); Bistritz et al. (2020); Wang et al. (2020), including the L_c-continuity assumption Shalev-Shwartz et al. (2009); Riis et al. (2021). Note that we do not assume anything for our proposed objective function $\widetilde{F}(\mathbf{w})$ and only have assumptions over the standard objective function $F(\mathbf{w})$ to prove the convergence of MaxFL over $\widetilde{F}(\mathbf{w})$ in Theorem 3.1.

Theorem 3.1 (Convergence to the MaxFL Objective $\widetilde{F}(\mathbf{w})$)**.** *Under Assumption 3.1-3.3, suppose the server uniformly selects m out of M clients without replacement in each round of Algorithm 1. With $\eta_l = \frac{1}{\sqrt{T}\tau}$, $\eta_g = \sqrt{\tau m}$, for a sufficiently large T we have:*

$$\min_{t\in[T]} \mathbb{E}\left[\left\|\nabla \widetilde{F}(\mathbf{w}^{(t,0)})\right\|^2\right] \leq \mathcal{O}\left(\frac{\sigma_g^2}{\sqrt{m\tau T}}\right) + \mathcal{O}\left(\frac{\sigma_g^2}{T\tau}\right) + \mathcal{O}\left(\frac{\sqrt{\tau}}{\sqrt{Tm}}\right) + \mathcal{O}\left(\frac{\kappa^2+\beta^2}{T}\right) \tag{5}$$

where $\mathcal{O}$ subsumes all constants (including L_s and L_c).

Theorem 3.1 shows that with a sufficiently large number of communication rounds T we reach a stationary point of our objective function $\widetilde{F}(\mathbf{w})$. The proof is deferred to Appendix C where we also show a version of this theorem that contains the learning rates η_g and η_l with the constants.

4 Related Work

To the best of our knowledge, the notion of GM-Appeal and the proposal to maximize it while considering flexible client participation have not appeared before in the previous literature. Previous works have focused on the notion of satisfying clients' personal requirements from a game-theoretic lens or designing strategies specifically to prevent client dropout, including the use of personalization, which have their limitations, as we discuss below.

4.1 Incentivizing Clients and Preventing drop-out

A recent line of work in game theory models FL as a coalition of self-interested agents and studies how clients can optimally satisfy their individual incentives defined differently from our goal. Instead of training a

single global model, Donahue & Kleinberg (2021a;b) consider the problem where each client tries to find the best possible coalition of clients to federate with to minimize its own error. Blum et al. (2021) consider an orthogonal setting where each client aims to satisfy its constraint of low expected error while simultaneously trying to minimize the number of samples it contributes to FL. While these works establish useful insights for simple linear tasks, it is difficult to extend these to practical non-convex machine learning tasks. In contrast to these works, in MaxFL we aim to directly maximize the *number* of satisfied clients using a global model. This perspective alleviates some of the analysis complexities occurring in game-theoretic formulations and allows us to consider general non-convex objective functions.

A separate line of work looks at how to prevent and deal with client drop-out in FL. Wang & Xu (2022) introduce a notion of 'friendship' among clients and proposes to use friends' local update as a substitute for the update of dropped-out clients. Gu et al. (2021) propose to use previous updates of dropped-out clients as a substitute for their current updates. Both algorithms are stateful. Another line of work Han et al. (2022); Kang et al. (2019); Zhang et al. (2021) aims to incentivize clients to contribute resources for FL and promote long-term participation by providing monetary compensation for their contributions, determined using game-theoretic tools. These techniques are orthogonal to MaxFL's formulation and can be combined if needed to further incentivize clients.

4.2 Personalized and Fair Federated Learning

Personalized federated learning (PFL) methods aim to increase performance by training multiple related models across the network (e.g., Smith et al., 2017). In contrast to PFL, MaxFL focuses on the challenging goal of training a *single* global model that can maximize the number of clients for which the global model outperforms their local model. Unlike PFL which may require additional training on new clients for personalization, MaxFL's global model can be used by new clients without additional training (see Table 3). Also, MaxFL is stateless, in that clients do not carry varying local parameters throughout training rounds as in many popular personalized FL methods Smith et al. (2017); Dinh et al. (2020); Fallah et al. (2020); Li et al. (2021), preventing parameter staleness problems which can be exacerbated by partial client participation Wang et al. (2021). Furthermore, MaxFL is orthogonal to and can be combined with PFL methods. We demonstrate this in Table 4, where we show results for MaxFL jointly used with personalization via fine-tuning Jiang et al. (2019). We compare MaxFL +Fine-tuning with another well known PFL method PerFedAvg Fallah et al. (2020) and show that MaxFL appeals to a significantly higher number of clients than the baseline.

Finally, another related area is fair FL, where a common goal is to train a global model whose accuracy has less variance across the client population than standard FedAvg Li et al. (2019); Mohri et al. (2019). A side benefit of these methods is that they can improve global model appeal for the worst performing clients. However, the downside is that the performance of the global model may be degraded for the best performing clients, thus making it unappealing for them to participate. We show in Appendix D.3 that fair FL methods are indeed not effective in increasing GM-Appeal.

5 Experiments

In this section we evaluate MaxFL for a number of different datasets while comparing with a wide range of baselines to show that maximizing GM-Appeal, i.e., training a global model that can appeal to a *larger number of clients*, provide many benefits for FL including: i) the server gaining more participating clients to select clients from for training a better global model for the seen clients, ii) the global model having a higher chance to have a good performance on unseen clients, iii) being robust against byzantine clients, and iv) clients gaining better performance with the global model when they combine MaxFL with local fine-tuning.

Datasets and Model. We evaluate MaxFL in three different settings: image classification for non-iid partitioned (i) FMNIST Xiao et al. (2017), (ii) EMNIST with 62 labels Cohen et al. (2017), and (iii) sentiment analysis for (iv) Sent140 Go et al. (2009)with a MLP. For FMNIST, EMNIST, and Sent140 dataset, we consider 100, 500, and 308 clients in total that are used for training where we select 5 and 10 clients uniformly at random per round for FMNIST and EMNIST, Sent140 respectively. These clients are active at some point in training the global model and we call them **'seen clients'**. We also sample the **'unseen clients'** from the same distribution from which we generate the seen clients, with 619 clients for Sent140, 100 clients for

	Seen Clients				Unseen Clients			
	FMNIST		EMNIST		FMNIST		EMNIST	
	Test Acc.	GM-Appeal	Test Acc.	GM-Appeal	Test Acc.	GM-Appeal	Test Acc.	GM-Appeal
FedAvg	43.70(±0.02)	0.04(±0.0)	35.15(±0.51)	0.02(±0.01)	43.14(±0.23)	0.07(±0.01)	37.14(±0.10)	0.06(±0.0)
FedProx	44.59(±1.94)	0.05(±0.01)	34.06(±1.21)	0.004(±0.0)	43.80(±1.67)	0.07(±0.01)	36.82(±0.22)	0.008(±0.0)
Scaffold	39.90(±0.59)	0.0(±0.0)	34.78(±2.05)	0.0(±0.0)	39.24(±0.68)	0.01(±0.0)	34.19(±1.25)	0.004(±0.0)
PerFedAvg	46.62(±1.0)	0.05(±0.0)	34.78(±1.05)	0.003(±0.0)	46.00(±0.87)	0.07(±0.0)	36.92(±0.51)	0.008(±0.0)
pFedme	31.06 (±2.06)	0.0 (±0.0)	9.78 (±2.13)	0.0 (±0.0)	20.11 (±3.4)	0.0 (±0.0)	7.05 (±1.03)	0.0 (±0.0)
qFFL	29.92(±3.13)	0.0(±0.0)	15.95(±3.02)	0.0(±0.0)	19.63(±2.17)	0.0(±0.0)	5.41(±0.52)	0.0(±0.0)
MW-Fed	44.41(±2.38)	0.04(±0.0)	30.44(±3.07)	0.01(±0.0)	43.46(±2.15)	0.06(±0.0)	36.54(±0.40)	0.01(±0.0)
MaxFL	**70.86**(±2.18)	**0.37**(±0.05)	**57.34**(±1.41)	**0.25**(±0.03)	**74.53**(±0.50)	**0.39**(±0.07)	**55.62**(±0.86)	**0.31**(±0.03)

Table 1: Avg. test accuracy and GM-Appeal where we train for 200 communication rounds. At the 10th communication round, we let clients flexibly opt-out or opt-in depending on whether the global model has met their requirements. We report the final avg. test accuracy and GM-Appeal at the 200th communication round.

FMNIST, and 500 for EMNIST. These unseen clients represent new incoming clients that have not been seen before during the training rounds of FL to evaluate the generalization performance at inference. Further details of the experimental settings are deferred to Appendix D.2.

Baselines. We compare MaxFL with numerous well-known FL algorithms such as standard FedAvg McMahan et al. (2017); FedProx Sahu et al. (2020) which aims to tackle data heterogeneity; SCAFFOLD Karimireddy et al. (2019) which aims for variance-reduction; PerFedAvg, pFedme Dinh et al. (2020) Fallah et al. (2020) which facilitates personalization; MW-Fed Blum et al. (2021) which incentivizes client participation; and qFFL which facilitates fairness Li et al. (2019). For all algorithms, we set ρ_k to be the same, i.e., $\rho_k = F_k(\widehat{\mathbf{w}}_k)$, where $\widehat{\mathbf{w}}_k$ is obtained by running a few warm-up local SGD steps on client k's data as outlined in Section 3 to ensure a fair comparison across baselines. We perform grid search for hyperparameter tuning for all baselines and choose the best performing ones.

Evaluation Metrics: GM-Appeal, Average Test Accuracy, and Preferred-model Test Accuracy. We evaluate MaxFL and other methods with three key metrics: 1) GM-Appeal, defined in (1), 2) average test accuracy (avg. test acc.) across clients, and a new metric that we propose called 3) preferred-model test accuracy. Preferred-model test accuracy is the average of the clients' test accuracies computed on either the global model $\mathbf{w}$ or their solo-trained local model $\widehat{\mathbf{w}}_k$, whichever one satisfies the client's requirement. We belive that average test accuracy is a more server-oriented metric as it assumes that clients will use the global model by default. On the other hand, preferred-model test accuracy is a more client-centric metric that allows clients to select the model which works best, thereby better reflecting their actual satisfaction.

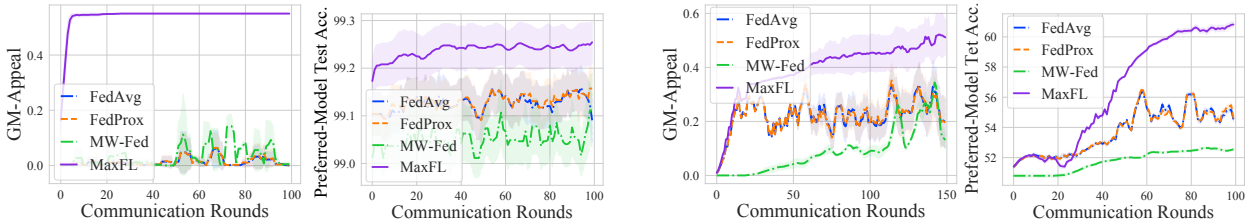
Note that the preferred-model test accuracy is a complementary novel metric we propose to gauge the satisfaction of the clients when they can choose between the global and local models. For instance, if we have a high preferred-model test accuracy but a low GM-Appeal where all clients are not interested in the global model, then the primary objective of MaxFL has not been achieved. However, in our results we show that in fact MaxFL has the complementary effect of not only being able to maximize the GM-Appeal, but also have higher preferred-model test accuracy. This is because MaxFL is able to find the global model that can perform better than a prefixed threshold for as many clients as possible, and therefore compared to the other global models trained from different baselines, the test accuracy for those clients who chose the global model is higher.

5.1 Experiment Results

Average Test Accuracy of Seen Clients & Unseen Clients. We first show that we improve the GM-Appeal and thus the average test accuracy performance for the 'seen clients' used during the training of the global model. In Table 1, we show the average test accuracy across clients where we let clients flexibly join or drop-out depending on whether the global model is appealing after 5% of communication rounds of mandatory participation. We show that MaxFL achieves the highest GM-Appeal than other baselines for

	FMNIST				EMNIST			
	Byz=0.1		Byz=0.05		Byz=0.1		Byz=0.05	
	Test Acc.	GM-Appeal	Test Acc.	GM-Appeal	Test Acc.	GM-Appeal	Test Acc.	GM-Appeal
MW-Fed	17.24 (±2.35)	0.01 (±0.0)	21.28 (±1.79)	0.02 (±0.0)	15.83 (±1.52)	0.004 (±0.0)	22.22 (±0.63)	0.008 (±0.001)
MaxFL	**69.42** (±2.87)	**0.35** (±0.05)	**70.60** (±2.76)	**0.42** (±0.03)	**52.74** (±0.44)	**0.20** (±0.01)	**56.10** (±0.77)	**0.23** (±0.01)

Table 2: Byzantine clients are included in the total clients where they artificially report large losses to the server and add noise to their gradients. The ratio of the Byzantine clients is denoted as 'Byz'. We report the final avg. test accuracy and GM-Appeal across the seen clients where we train for 200 communication rounds. At the 10th round, clients flexibly opt-out or opt-in depending on whether the global model has met their requirements.



(a) FMNIST (b) Sent140

Figure 4: GM-Appeal and preferred-model test accuracy for the seen clients are significantly higher for MaxFL. Therefore, clients can also benefit from choosing either the local or global model for best performance, while the server also gains a large number of clients to select from.

both FMNIST and EMNIST by 0.32-0.39 and 0.23-0.31 improvement, respectively. Since MaxFL is able to retain a larger pool of clients due to having a higher GM-Appeal, it therefore trains from selecting from a more larger client pool, leading to the highest average test accuracy compared to the baselines by 22-40% and 18-50% improvement respectively for the seen and unseen clients. Since the other baselines do not consider the notion of GM-Appeal entirely, it fails in preventing client dropouts leading to poor performance. Note that we do not use any of the 'unseen clients' during training and only calculate the GM-Appeal and test accuracy via inference with the global model trained with the 'seen clients'.

Robustness of MaxFL Against Byzantine Clients. One may think that MaxFL may be perceptible to attacks from Byzantine clients that intentionally send a greater GM-Appeal gap to the server to gain a higher aggregation weight. To show MaxFL's robustness against such attacks we show in Table 2 the performance of MaxFL with Byzantine clients attacks which send higher losses to gain higher weights and then send Gaussian noise mixed gradients to the server. We compare with MW-Fed Blum et al. (2021) which aims for incentivizing client participation by clients sending higher weights to the server and performing more local updates. In Table 2 we see that for both high and low byzantine client ratios, MaxFL achieves only 1-5% lower test accuracy for seen and unseen clients compared to the case where there are no Byzantine clients in Table 1. This is due to our objective giving lower weight to those clients that give a too high GM-Appeal gap (see Fig. 3). Hence MaxFL disregards these clients that send artificially high GM-Appeal gaps.

	GM-Appeal		Preferred-Model Test Acc.	
	FMNIST	Sent140	FMNIST	Sent140
FedAvg	0.08(±0.01)	0.37(±0.07)	98.53(±0.13)	57.05(±1.44)
FedProx	0.07(±0.01)	0.37(±0.07)	98.43(±0.21)	57.07(±1.42)
Scaffold	0.02(±0.01)	0.03(±0.05)	98.26(±0.20)	51.59(±0.11)
MW-Fed	0.05(±0.04)	0.17(±0.03)	98.32(±0.13)	55.57(±1.28)
MaxFL	**0.55**(±0.0)	**0.43**(±0.05)	**98.83**(±0.06)	**57.16**(±1.35)

Table 3: GM-Appeal and preferred-model test accuracy of the final global models for the unseen clients. MaxFL improves the GM-Appeal by at least 47% for FMNIST, and 6% for Sent140 and achieves the same or higher preferred-model test accuracy.

Preferred-model Test Accuracy: Clients' Perspective. Recall that a high preferred-model test accuracy implies that the client has a higher chance of satisfying its requirement by choosing between the

global or solo-trained local model, whichever performs better. In Fig. 4 we show that as the GM-Appeal increases across the communication round, preferred-model test accuracy also increases. MaxFL achieves the highest final GM-Appeal and preferred-model test accuracy indicating that it provides a win-win case for both the server and clients, since the clients have the highest accuracy by choosing the better model between the global model $\mathbf{w}$ and the local model $\widehat{\mathbf{w}}_k$, and the server has the highest fraction of participating clients. Similarly, in Table 3, MaxFL achieves the highest GM-Appeal and preferred-model test accuracy. Although the preferred-model test accuracy improvement compared to the other baselines may appear small, showing that MaxFL is able to maintain a high preferred-model test accuracy while also achieving a high GM-Appeal implies that it does not sabotage the benefit of clients while also bringing the server more clients to select from.

Local Tuning for Personalization. Personalized FL methods can be used to fine-tune the global model at each client before comparing it with the client's locally trained model. MaxFL can be combined with these methods by simply allowing clients to perform some fine-tuning iterations before computing the aggregation weights in Step 7 of Algorithm 1. Both for clients that are active during training and unseen test clients, we show in Table 4 that MaxFL increases the GM-Appeal by at least 10% compared to all baselines. For FMNIST and Sent140, the improvement in GM-Appeal over other methods is up to 27%, 28% respectively for active clients and 17%, 4% respectively for unseen clients.

	Seen Clients		Unseen Clients	
	FMNIST	Sent140	FMNIST	Sent140
FedAvg	0.38(±0.06)	0.25(±0.09)	0.39(±0.06)	0.42(±0.06)
FedProx	0.40(±0.07)	0.26(±0.09)	0.41(±0.07)	0.43(±0.12)
Scaffold	0.02(±0.02)	0.16(±0.22)	0.03(±0.02)	0.07(±0.01)
PerFedAvg	0.45(±0.05)	0.24(±0.10)	0.46(±0.06)	0.47(±0.06)
MW-Fed	0.28(±0.07)	0.08(±0.01)	0.39(±0.04)	0.20(±0.01)
MaxFL	**0.55**(±0.01)	**0.36**(±0.05)	**0.56**(±0.01)	**0.55**(±0.01)

Table 4: GM-Appeal of locally-tuned models with 5 local steps from the final global models for seen clients and unseen clients. Both for clients that are active during training and unseen test clients, MaxFL increases the fraction of clients that find the global model appealing by at least 10% as compared to all baselines.

6 Limitations and Concluding Remarks

In this work, we explore the notion of global model appeal by proposing to train a global model that maximizes the number of clients whose requirements are satisfied. We show that when participating clients drop out or clients do not join due to small global model appeal, the test accuracy for the current training clients and generalization performance to the new unseen clients can suffer significantly. With theoretical insights and guarantees, we show through extensive experiments that MaxFL retains many clients for training and thus achieves a high average test accuracy across the participating clients and also across the new incoming clients.

We note that our proposed metric GM-Appeal and MaxFL objective have some limitations. For instance, it is possible train a global model that sacrifices the performance of a few clients to maximize GM-Appeal, potentially reducing fairness. However, we expect that MaxFL could potentially be altered by modulating the ρ_k value to cover such limitations such as setting ρ_k to a constant that retains fairness. Another limitation in MaxFL is that it does not currently consider specific incentive mechanisms for various settings such as what cost we can set for the new incoming clients that wants to use the global model at inference without participating in training. Without setting this cost, one may raise the concern of free-rider problems. Nevertheless, our work presents a first step towards maximizing the set of clients that are interested in the global model; how to design the cost of using a global model would be an interesting, orthogonal direction of future work. As a similar notion of global model appeal has not been thoroughly examined previously, we hope our work can open up new research directions in understanding the role played by the server to prevent client dropout and recruit new clients by finding a global model that serves as many clients as possible.

Broader Impact Statement

Our work proposes a new objective to maximize global model appeal in federated learning, which can incentivize more clients to participate and prevent clients from dropping out, and improve generalization

performance on unseen clients. Despite these benefits, it is worth acknowledging the potential negative effects of the proposed objective and algorithm in terms of other metrics, such as unfairness (e.g., increased performance gap) between different subpopulations or vulnerability to Byzantine clients. Here, we note that (1) Our MaxFL framework is general in the sense that it could be adjusted to trade off fairness/utility by setting the local requirement parameters ρ_k appropriately for specific subpopulations, arriving at a fairer model that reduces the number of clients suffering from inferior performance; and (2) Our results demonstrate that MaxFL is in fact more robust to Byzantine clients than competitors due to its dynamic reweighting scheme (see Table 2). In general, however, it remains critical to carefully consider trade-offs between issues such as model appeal, fairness, and robustness for the application at hand. The goal of this work is to explore the implications of MaxFL both theoretically and empirically so that we can understand various benefits and limitations of GM-Appeal in different scenarios. We hope practitioners and researchers can thus appropriately adjust the framework and/or combine it with other learning schemes depending on the application of interest, weighing potential benefits of the approach relative to existing methods in terms of achieving higher accuracy, broader participation, and increased fairness or robustness.

References

I. Bistritz, A. J. Mann, and N. Bambos. Distributed distillation for on-device learning. In *Advances in Neural Information Processing Systems*, 2020.

Avrim Blum, Nika Haghtalab, Richard Lanas Phillips, and Han Shao. One for one, or all for all: Equilibria and optimality of collaboration in federated learning. In *International Conference on Machine Learning*, 2021.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: an extension of MNIST to handwritten letters. *arXiv preprint arXiv:1702.05373*, 2017.

Canh T. Dinh, Nguten H. Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes. In *Advances in Neural Information Processing Systems*, 2020.

Kate Donahue and Jon Kleinberg. Model-sharing games: Analyzing federated learning under voluntary participation. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, 2021a.

Kate Donahue and Jon Kleinberg. Optimality and stability in federated learning: A game-theoretic approach. In *Advances in Neural Information Processing Systems*, 2021b.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *Advances in Neural Information Processing Systems*, 2020.

Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 2009. URL https://www-cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf.

Xinran Gu, Kaixuan Huang, Jingzhao Zhang, and Longbo Huang. Fast federated learning in the presence of arbitrary device unavailability. *Advances in Neural Information Processing Systems*, 34:12052–12064, 2021.

Jingoo Han, Ahmad Faraz Khan, Syed Zawad, Ali Anwar, Nathalie Baracaldo Angel, Yi Zhou, Feng Yan, and Ali R. Butt. Tokenized incentive for federated learning. In *Proceedings of the Federated Learning Workshop at the Association for the Advancement of Artificial Intelligence (AAAI) Conference*, 2022.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. In *International Workshop on Federated Learning for User Privacy and Data Confidentiality in Conjunction with NeurIPS 2019 (FL-NeurIPS'19)*, December 2019.

Divyansh Jhunjhunwala, Pranay Sharma, Aushim Nagarkatti, and Gauri Joshi. Fedvarp: Tackling the variance due to partial client participation in federated learning. *arXiv*, 2022.

Yihan Jiang, Jakub Konecny, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv 1909.12488*, 2019.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurelien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adria Gascon, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecny, Aleksandra Korolova, Farinaz Koushanfar,

Sanmi Koyejo, Tancrede Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Ozgur, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramer, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Jiawen Kang, Zehui Xiong, Dusit Niyato, Han Yu, Ying-Chang Liang, and Dong In Kim. Incentive design for efficient federated learning in mobile networks: A contract theory approach. In *2019 IEEE VTS Asia Pacific Wireless Communications Symposium (APWCS)*, pp. 1–5, 2019. doi: 10.1109/VTS-APWCS.2019.8851649.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for on-device federated learning. *arXiv preprint arXiv:1910.06378*, 2019.

Yûsaku Komatu. Elementary inequalities for mills' ratio. *Reports of Statistical Application Research (Union of Japanese Scientific Engineers)*, 4:69–70, 1955.

Tian Li, Maziar Sanjabi, and Virginia Smith. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *Proceedings of the 38th International Conference on Machine Learning*, 2021. URL https://arxiv.org/abs/2012.04221.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations (ICLR)*, July 2020. URL https://arxiv.org/abs/1907.02189.

Hamed Masnadi-shirazi and Nuno Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In *Advances in Neural Information Processing Systems(NIPS)*, 2008.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agøura y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. *International Conference on Artificial Intelligenece and Statistics (AISTATS)*, April 2017. URL https://arxiv.org/abs/1602.05629.

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. *International Conference on Learning Representations*, 2018.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4615–4625, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Tan T. Nguyen and Scott Sanner. Algorithms for direct 0–1 loss optimization in binary classification. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, 2013.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Erlend S. Riis, Matthias J. Ehrhardt, G. R. W. Quispel, and Carola-Bibiane Schönlieb. A geometric integration approach to nonsmooth, nonconvex optimisation. *Foundations of Computational Mathematics*, 2021.

Anit Kumar Sahu, Tian Li, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith. Federated optimization for heterogeneous networks. In *Proceedings of the 3rd MLSys Conference*, January 2020.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, 2009.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, pp. 4424–4434. 2017. URL http://papers.nips.cc/paper/7029-federated-multi-task-learning.pdf.

Sebastian U Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations (ICLR)*, 2019.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

Heqiang Wang and Jie Xu. Friends to help: Saving federated learning from client dropout. *arXiv preprint arXiv:2205.13222*, 2022.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization. *preprint*, May 2020. URL https://arxiv.org/abs/2007.07481.

Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H Brendan McMahan, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *https://arxiv.org/abs/1708.07747*, aug 2017.

Tao Yu, Eugene Bagdasaryan, and Vitaly Shmatikov. Salvaging federated learning by local adaptation. *arXiv preprint arXiv:2002.04758*, 2020.

Meng Zhang, Ermin Wei, and Randall Berry. Faithful edge federated learning: Scalability and privacy. *IEEE Journal on Selected Areas in Communications*, 39(12):3790–3804, 2021. doi: 10.1109/JSAC.2021.3118423.