Federated Functional Gradient Boosting

Zebang Shen
University of Pennsylvania

Hamed Hassani
University of Pennsylvania

Satyen Kale
Google Research

Amin Karbasi
Yale

Abstract

Motivated by the tremendous success of boosting methods in the standard centralized model of learning, we initiate the theory of boosting in the Federated Learning setting. The primary challenges in the Federated Learning setting are heterogeneity in client data and the requirement that no client data can be transmitted to the server. We develop *federated functional gradient boosting* (FFGB) an algorithm that is designed to handle these challenges. Under appropriate assumptions on the weak learning oracle, the FFGB algorithm is proved to efficiently converge to certain neighborhoods of the global optimum. The radii of these neighborhoods depend upon the level of heterogeneity measured via the total variation distance and the much tighter Wasserstein-1 distance, and diminish to zero as the setting becomes more homogeneous. In practice, as suggested by our theoretical findings, we propose using FFGB to warm-start existing Federated Learning solvers and observe significant performance boost in highly heterogeneous settings. The code can be found here.

1 Introduction

Federated learning (FL) is a machine learning paradigm in which multiple clients cooperate to learn a model under the orchestration of a central server [McMahan et al., 2017b]. In FL, clients tend to have very heterogeneous data, and this data is never sent to the central server due to privacy and communication constraints. Thus, it is necessary to offload as much computation as possible to the clients. The challenge in FL is to train a single unified model via decentralized computation even though the clients have heterogeneous data.

In this paper, we initiate a study of *boosting* in the FL setting. Boosting [Schapire and Freund, 2012] is a classical ensemble method for building additive models in a greedy, stagewise manner. Boosting can be viewed as solving a non-parametric *functional minimization* problem in an iterative manner. In each stage of boosting, a new model from a certain base class of models is added to the current ensemble to decrease the training loss. The new model is found via *functional gradient descent*, which involves finding an approximation to the functional gradient of the loss in the base class via a corresponding training procedure (see, e.g., [Mason et al., 2000, Friedman, 2001a]). This training procedure is commonly referred to as a "weak learning oracle".

The power of boosting in the centralized setting arises from the flexibility afforded by operating over arbitrary base function classes equipped with such a weak learning oracle. In other words, boosting operates generically over function classes which can be carefully designed to encode expert domain knowledge. This flexibility has manifested in tremendous practical success of specific boosting methods such as AdaBoost [Freund and Schapire, 1997] and Gradient Boosted Decision Trees [Friedman, 2001b]. Motivated by the success of centralized boosting methods, in this paper we aim to adapt the theory of boosting to the FL setting. The primary challenges in this adaptation are the heterogeneity of client data, and the requirement that no client data be transmitted to the server. Therefore, the server needs to orchestrate the aggregation of the models learned from clients (without having access to their data) while at the same time guide the learning on the clients via appropriate feedback. In this paper, we show how aggregation and learning can be combined seamlessly in order to construct a strong model by learning local weak models in a federated model of computation.

Contributions. This paper formulates the federated functional minimization problem and develops federated functional gradient boosting algorithms and convergence analyses for various settings of client heterogeneity. The main contributions are as follows:

1. We propose a *federated functional gradient boosting*

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

(FFGB) algorithm for solving functional minimization under the federated learning paradigm. The algorithm adapts to different data heterogeneity settings via appropriate choices of weak learning oracles. The algorithm relies on the clients running multiple *restricted functional gradient descent* (RFGD) steps locally, but augmented with *residual* variables that are crucial for proving global convergence of the algorithm.

2. We consider the most general setting without specifying the loss function or the relations between the marginal distributions of the feature vectors on the clients. We show that FFGB converges to a neighborhood centered around the global minimizer, whose radius depends on the average Total Variation distance between the local marginal distributions and the population marginal distribution. We also construct an example that shows that convergence to the exact minimizer is not possible without further assumptions.

3. We improve the above convergence analysis under the setting that the marginal distributions are identical and heterogeneity arises due to differing conditional distributions of labels. Under a weaker assumption on the weak learning oracle, we show that FFGB converges to the global minimizer in a sublinear rate. Interestingly, our analysis suggests that the number of local steps of FFGB should be $\Omega(\sqrt{T})$ in order to have the best convergence rate where T is the number of boosting stages. This is a rather surprising result that shows the benefit of taking multiple local steps in FL.

4. We tighten the bound on the convergence radius from the TV distance to *Wasserstein-1* distance for the special case of square loss. Such an improvement can be significant especially when there are mismatches between the support of the local marginal distributions.

5. From a practical perspective, we incorporate the knowledge distillation technique in the model aggregation step of FFGB. Inspired by our theoretical analyses, we propose to use the resulting method FFGB-DISTILL as a warm start for SOTA FL solvers. Our experiments show clear evidence for the benefits of such a scheme on practical datasets.

1.1 Related work

In the FL setting, FEDAVG [McMahan et al., 2017b] is the most popular algorithm for solving standard parametric optimization problems. In every round of FEDAVG, the server sends a global average parameter to the local machines, which then perform multiple local steps (usually stochastic gradient descent) to update the received parameters. These improved local parameters are aggregated by the server for the next round. It has been noted by several papers (see, e.g. [Karimireddy et al., 2020b] and the references therein) that FEDAVG deteriorates in the presence of client heterogeneity. Federated optimization algorithms such as SCAFFOLD [Karimireddy et al., 2020b], FEDPROX [Li et al., 2020b], MIME [Karimireddy et al., 2020a], FEDDYN [Acar et al., 2020] etc. were designed to tackle this issue.

In the centralized setting of gradient boosting, rigorous analysis with rates of convergence were given by Duffy and Helmbold [2002], Rätsch et al. [2001], Zhang and Yu [2005], Grubb and Bagnell [2011]. The algorithms in this paper and their analyses build upon the prior work of Grubb and Bagnell [2011].

In the FL setting, Li et al. [2020a] propose a stagewise algorithm named SIMFL that trains gradient Gradient Boosted Decision Trees. However, we note that a prerequisite for the proposed method is the pairwise similarity over all the data points. The concept of similarity is heavily task specific as it completely depends on the data labelling, and requires strong domain knowledge. In contrast, in our work we focus on methods that are agnostic to the data similarity in order to have a broader applicability. We therefore did not compare with SIMFL in our experiment. Hamer et al. [2020] have developed a method called FEDBOOST which, despite its name, is *not* a boosting method in the traditional sense, but rather a communication-efficient method to learn a linear classifier over features provided by pre-trained classifiers in the FL setting. While this linear classifier is an ensemble of the pre-trained models, FEDBOOST doesn't guide the training of new models. Hence we do not provide further comparison with this method.

2 Preliminaries

In this section, we define the necessary notation as well as the federated functional minimization problem that is solved via boosting algorithms.

2.1 Notation.

For a positive integer n, we define $[n] := \{1, 2, \dots, n\}$. For a vector $x \in \mathbb{R}^d$, we use x_i to denote its i^{th} entry. We use $\langle \cdot, \cdot \rangle$ to denote the standard Euclidean inner product in $\mathbb{R}^d$ and use $\| \cdot \|$ for the corresponding standard Euclidean norm.
Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the input space, and $\mathcal{M}_+^1(\mathcal{X})$ denote the set of probability measures on $\mathcal{X}$. For two measures $\alpha, \beta \in \mathcal{M}_+^1(\mathcal{X})$, let $TV(\alpha, \beta)$ denote their total variation distance and $\mathrm{W}_p(\alpha, \beta)$ denote their Wasserstein-p distance (definitions in appendix).
For a fixed distribution $\alpha \in \mathcal{M}_+^1(\mathcal{X})$, we define the corresponding weighted $\mathcal{L}^p$ space $(1 \le p < \infty)$ of functions from $\mathcal{X}$ to $\mathbb{R}^c$ as follows:

$\mathcal{L}^p(\alpha)\overset{\Delta}{=}\{f : \mathcal{X} \to \mathbb{R}^c \mid \left(\int_{\mathcal{X}} \|f(x)\|^p d\alpha(x)\right)^{\frac{1}{p}} < \infty\}$. The space $\mathcal{L}^2(\alpha)$ is endowed with natural inner product and norm: for two functions $f, g \in \mathcal{L}^2(\alpha)$, we have $\langle f, g\rangle_\alpha = \int_{\mathcal{X}} \langle f(x), g(x)\rangle d\alpha(x)$ and $\|f\|_\alpha = \sqrt{\langle f, f\rangle_\alpha}$. In the limiting case when $p \to \infty$, we define $\mathcal{L}^\infty(\alpha)\overset{\Delta}{=}\left\{f : \mathcal{X} \to \mathbb{R}^c \,\middle|\, \forall x \in \mathrm{supp}(\alpha), \|f(x)\| < \infty\right\}$, where $\mathrm{supp}(\alpha)$ denotes the support of α. We define the α-infinity norm of a function $f : \mathcal{X} \to \mathbb{R}^c$ by $\|f\|_{\alpha,\infty}\overset{\Delta}{=}\sup_{x\in\mathrm{supp}(\alpha)} \|f(x)\|$. We will also use $\mathcal{L}^\infty\overset{\Delta}{=}\left\{f : \mathcal{X} \to \mathbb{R}^c \,\middle|\, \forall x \in \mathcal{X}, \|f(x)\| < \infty\right\}$, and define $\|f\|_\infty\overset{\Delta}{=}\sup_{x\in\mathcal{X}} \|f(x)\|$. We denote the Lipschitz constant of a continuous function $f : \mathcal{X} \to \mathbb{R}^c$ as

$$\|f\|_{\mathrm{lip}}\overset{\Delta}{=}\inf\{L : \forall x, x' \in \mathcal{X}, \ \|f(x)-f(x')\| \le L\|x-x'\|\}.$$

2.2 Functional Minimization in $\mathcal{L}^2$ Space

For some output space $\mathcal{Y}$, let $\ell : \mathbb{R}^c \times \mathcal{Y} \to \mathbb{R}$ be a loss function that is *convex* in the first argument. An important example of ℓ is the cross entropy loss, where $\mathcal{Y} = [c]$, and

$$\ell(y', y) = -\log\left(\frac{\exp(y'_y)}{\sum_{i=1}^c \exp(y'_i)}\right). \tag{1}$$

Given a joint distribution P on $\mathcal{X} \times \mathcal{Y}$, we use α to denote the marginal distribution of P on $\mathcal{X}$, and for every $x \in \mathcal{X}$, we let $\beta^x \in \mathcal{M}_+^1(\mathcal{Y})$ be the distribution on $\mathcal{Y}$ under P conditioned on x. We then define the risk functional $\mathcal{R} : \mathcal{L}^2(\alpha) \to \mathbb{R}$ as $\mathcal{R}[f] \overset{\Delta}{=} \mathbb{E}_{(x,y)\sim P}[\ell(f(x), y)]$. We consider the following Tikhonov regularized functional minimization problem:

$$\min_{f\in\mathcal{L}^2(\alpha)} \mathcal{F}[f] \overset{\Delta}{=} \mathcal{R}[f] + \frac{\mu}{2}\|f\|_\alpha^2. \tag{2}$$

Here, $\mu \ge 0$ is some regularization parameter. To solve such a problem, we define the *subgradient* $\nabla\mathcal{F}[f]$ of $\mathcal{F}$ at f as the set of all $g \in \mathcal{L}^2(\alpha)$ such that $\mathcal{F}[g] \ge \mathcal{F}[f] + \langle g - f, \nabla\mathcal{F}[f]\rangle_\alpha$. Since $\mathcal{L}^2(\alpha)$ is a Hilbert space, according to the Riesz representation theorem, the subgradient can also be represented by a function in $\mathcal{L}^2(\alpha)$. Concretely, the subgradient can be explicitly computed as follows: $\nabla\mathcal{F}[f]$ is the collection of functions $h \in \mathcal{L}^2(\alpha)$, such that for any $x \in \mathrm{supp}(\alpha)$,

$$h(x) = \mathbb{E}_{y\sim\beta^x}[\nabla_1\ell(f(x), y)] + \mu f(x), \tag{3}$$

where $\nabla_1\ell(y', y) = \partial\ell(y', y)/\partial y'$. In particular, when only empirical measures of α and β^x are available, i.e. when P is the empirical distribution on the set $\{(x_1, y_1), \ldots, (x_M, y_M) \in \mathcal{X}\times\mathcal{Y}\}$, the empirical version of the above subgradient computation (3) is as follows: tor all $j \in [M]$, $h(x_j) = \nabla_1\ell(f(x_j), y_j) + \mu f(x_j)$.

Restricted Functional Gradient Descent (Boosting). A standard approch to solve the functional minimization problem (2) is the functional gradient descent method $f^{t+1} := f^t - \eta^t h^t$, $h^t \in \nabla\mathcal{F}[f^t]$. In the empirical case described above, one can construct $h^t \in \nabla\mathcal{F}[f^t]$ by interpolating $\{x_j, \nabla_1\ell(f^t(x_j), y_j) + \mu f^t(x_j)\}_{j=1}^M$. However, this choice of h^t has at least two drawbacks: 1. Evaluating h^t at a single point requires going through the whole dataset; 2. h^t is constructed explicitly on the data points (x_j, y_j) and thus cannot be transmitted to the server in the FL setting. One alternative to the explicit functional gradient descent method is restricted functional gradient descent, also known as Boosting [Mason et al., 2000]:

$$f^{t+1} := f^t - \eta^t h^t_{weak}, \text{ where } \ h^t_{weak} = \mathcal{Q}^2_\alpha(\nabla\mathcal{F}[f^t]). \tag{4}$$

Here, $\mathcal{Q}^2_\alpha$ is a *weak learning oracle* such that for any $\phi \in \mathcal{L}^2(\alpha)$, the output $h = \mathcal{Q}^2_\alpha(\phi)$ is a function in $\mathcal{L}^2(\alpha)$ satisfying the following *weak learning assumption*:

$$\|h - \phi\|_\alpha \le (1-\gamma)\|\phi\|_\alpha, \tag{5}$$

for some positive constant $0 < \gamma \le 1$. Replacing the interpolating h^t with h^t_{weak} alleviates the aforementioned two drawbacks and the restricted functional gradient descent can be proved [Grubb and Bagnell, 2011] to converge to the global optimum under standard regularity conditions.

Implementing the Weak Learning Oracle. Let ϕ be the input to the oracle and let h_θ be the candidate weak learner, e.g. a neural network with parameter θ. We can implement $\mathcal{Q}^2_\alpha$ by solving

$$\min_\theta \sum_{x\in\mathrm{supp}(\alpha)} \|\phi(x) - h_\theta(x)\|^2. \tag{6}$$

3 Federated Functional Minimization

In this paper, we consider the federated functional minimization problem. We assume that there are N client machines. Client machine i draws samples from distribution P_i over $\mathcal{X} \times \mathcal{Y}$. We denote the marginal distribution of P_i on $\mathcal{X}$ by α_i and the conditional distribution on $\mathcal{Y}$ given x by β_i^x. Due to the heterogeneous nature of the federated learning problems, the P_i's differ across different clients.

We define α to be the “arithmetic" mean of the local input probability measures[1]:

$$\alpha \overset{\Delta}{=} \frac{1}{N}\sum_{i=1}^N \alpha_i. \tag{7}$$

[1] More precisely, for any Borel subset of $\mathcal{X}$, its measure under α is the average of the measures under α_i's.

Algorithm 1 SERVER procedure

```
1: procedure SERVER($f^0$, $T$, $\mathcal{C}$)
2:     for $t \leftarrow 0$ to $T-1$ do
3:         Sample set $\mathcal{S}^t$ of clients.
4:         $f^{t+1} = \frac{1}{|\mathcal{S}^t|}\sum_{i\in\mathcal{S}^t}$ CLIENT($i, t, f^t$)
5:     return $f^T$.
```

It is easy to see that $\mathcal{L}^2(\alpha) \subseteq \mathcal{L}^2(\alpha_i)$ for all i. We define $\mathcal{R}_i : \mathcal{L}^2(\alpha) \to \mathbb{R}$ to be

$$\mathcal{R}_i[f] \triangleq \mathbb{E}_{(x,y)\sim P_i}[\ell(f(x), y)] \tag{8}$$

and denote $\mathcal{F}_i[f] = \mathcal{R}_i[f] + \frac{\mu}{2}\|f\|^2_{\alpha_i}$. The goal of federated learning is to minimize the average loss functional

$$\min_{f\in\mathcal{L}^2(\alpha)} \mathcal{F}[f] = \frac{1}{N}\sum_{i=1}^{N}\mathcal{F}_i[f]. \tag{9}$$

In the rest of the paper, we use f^* to denote the optimizer of (9). We emphasize that in federated optimization, due to the privacy requirement, the local inner product structure $\langle\cdot,\cdot\rangle_{\alpha_i}$ as well as the subgradient *cannot* be shared with the server during training phase, which constitutes the major challenge of the federated functional minimization problem (9). In the following, we first present the proposed *federated functional gradient boosting* (FFGB) method and show that it converges to certain neighborhoods of the global optimum whose radii depend upon the level of heterogeneity.

The FFGB algorithm shares a similar structure as FEDAVG. The SERVER procedure (Algorithm 1) performs global variable aggregation in each round, and then sends the global consensus function f^t to the clients. While in line 4 we aggregate the functions received from the clients by averaging them, for a practical implementation we can also use *knowledge distillation*, as discussed in Section 5. The CLIENT procedure (Algorithm 2) performs K local steps of the RFGD update (4) tracked via the local variable $g_i^{k,t}$, for $k = 1, 2, \ldots, K$, with the initialization $g_i^{1,t} = f^t$. The crucial twist on the RFGD update is the use of an additional residual variable Δ_i, which is initialized to the constant zero function (Alg. 2, line 2). This residual variable accumulates the approximation error of the descent direction $h_i^{k,t}$ incurred by the weak oracle (Alg. 2, line 6). Such a residual is then used to compensate the next functional subgradient (Alg. 2, line 4) and is used in the query to the weak learning oracle. Since the gradient of the Tikhonov's regularization in $\mathcal{F}_i$ exactly known, there is no need to approximate that part with the weak learning oracle. The local variable function $g_i^{k,t}$ is then refined by the approximated functional gradient $\left(h_i^{k,t} + \mu g_i^{k,t}\right)$ (Alg. 2, line 5). Note that the residual Δ_i only tracks the error of estimating $\nabla\mathcal{R}_i$ since the rest of $\nabla\mathcal{F}_i$ is exactly available. After K such updates, the local function $g^{K+1,t}$ is communicated to the server which aggregates these functions across the clients.

Algorithm 2 CLIENT procedure for Federated Functional Gradient Boosting (FFGB)

```
1: procedure CLIENT($i$, $t$, $f$)
2:     $\Delta_i^{0,t} = 0$, $g_i^{1,t} = f^t$ ;
3:     for $k \leftarrow 1$ to $K$ do
4:         $h_i^{k,t} := \mathcal{Q}^*_{\alpha_i}(\Delta_i^{k-1,t} + \nabla\mathcal{R}_i[g_i^{k,t}])$
5:         $g_i^{k+1,t} := g_i^{k,t} - \eta^{k,t}\left(h_i^{k,t} + \mu g_i^{k,t}\right)$
6:         $\Delta_i^{k,t} := \Delta_i^{k-1,t} + \nabla\mathcal{R}_i[g_i^{k,t}] - h_i^{k,t}$
7:     return $g_i^{K+1,t}$.
```

Residual. The residual Δ_i used in FFGB (Algorithm 2) is crucial to our convergence analysis which smoothly interpolates different heterogeneous setting. When we have homogeneous local input distributions, i.e. $\forall i \in [N], \alpha_i = \alpha$, our goal is to show that FFGB converges to the *global minimizer* (note that in this case the conditional distributions may still be heterogeneous, i.e. β_i^x may not be equal to β_j^x for $i \neq j$). However, such a result is not possible without the residual term: in its absence, the error accumulated by the RFGD updates via the calls to the weak learning oracle may not vanish since in FL the local subgradient is *non-zero* even at the global optimal solution due to client heterogeneity. This is in sharp contrast to the single machine case where, at the global optimal solution, the subgradient vanishes and so does the approximation error incurred by the weak oracle, leading to convergence of RFGD. This technique has also been applied for functional minimization in the much simpler centralized setting [Grubb and Bagnell, 2011].

Weak Learning Oracle. Given the local input distribution α_i, the description of FFGB in Algorithm 2 does not specify the implementation of the weak learning oracle $\mathcal{Q}^*_{\alpha_i}$ in line 4. The $*$ in the superscript in $\mathcal{Q}^*_{\alpha_i}$ indicates that in our convergence analyses, three types of weak learning oracle are used for different heterogeneous settings: ℓ^2 oracle $\mathcal{Q}^2_{\alpha}$, ℓ^∞ oracle $\mathcal{Q}^\infty_{\alpha}$, and Lipschitz oracle $\mathcal{Q}^{\text{lip}}_{\alpha}$, as listed in Table 1.

4 Convergence Analysis

In this section, we present the convergence results of FFGB under different settings which are summarized in Table 1. Note that we present the convergence rate in the case when all N clients participate in each round. The analysis easily extends to the case when only a few clients are sampled in each round, incurring a penalty for the variance in the sampling. Qualitatively the convergence bound stays the same. Details can be found in the appendix.

4.1 General Setting

First, we consider the most general setting where we assume no specific form of the loss functional ℓ in the objective (8) and the local input distributions are potentially heterogeneous, i.e. $\exists\ i, j \in [N]$ such that $\alpha_i \neq \alpha_j$. To analyze the convergence of FFGB, we make the following standard regularity assumption, which is satisfied e.g. by the cross entropy loss (1).

Assumption 4.1. *The subgradients of $\mathcal{R}_i[f]$ are G-bounded under the $\mathcal{L}^\infty(\alpha_i)$ norm, i.e.*

$$\forall f \in \mathcal{L}^\infty(\alpha_i), \|\nabla \mathcal{R}_i[f]\|_{\alpha_i,\infty} \leq G.$$

Additionally, we make the following assumption on the weak learning oracle $\mathcal{Q}^\infty_{\alpha_i}$ which is slightly stronger than the standard $\mathcal{L}^2$ weak learning oracle (see (5)): for any $h \in \mathcal{L}^\infty(\alpha_i)$, we assume that

$$\|\mathcal{Q}^\infty_{\alpha_i}(h) - h\|_{\alpha_i,\infty} \leq (1-\gamma)\|h\|_{\alpha_i,\infty}, \tag{10}$$

for some positive constant $0 < \gamma \leq 1$ and that

$$\|\mathcal{Q}^\infty_{\alpha_i}(h)\|_\infty \leq \bar{G}_\gamma, \tag{11}$$

for $\|h\|_{\alpha_i,\infty} \leq \bar{G}_\gamma$, where $\bar{G}_\gamma$ is some constant that depends on both G and γ and will be determined in the following analysis.

The following theorem states that FFGB.C converges to a neighborhood of the global minimizer, with a radius proportional to the average TV distance between the local input distribution α_i and the arithmetic mean α:

Theorem 4.1. *Let f^0 be the initializer function. We define a proxy of the heterogeneity among the local input distributions α's in the* TV *distance as*

$$\omega_{\mathrm{TV}} \triangleq \frac{1}{N}\sum_{i=1}^{N} \mathrm{TV}(\alpha, \alpha_i). \tag{12}$$

Under Assumption 4.1, and supposing the weak learning oracle $\mathcal{Q}^\infty_\alpha$ satisfies (10) *and* (11) *with constant γ and $\bar{G}_\lambda = 2G/\lambda$, using the step sizes $\eta^{k,t} = 4/(\mu(tK+k+1))$, the output of* FFGB *satisfies*

$$\|f^T - f^*\|_\alpha^2 = O\Big(\frac{\|f^0 - f^*\|_\alpha^2}{KT} + \frac{KG^2 log(KT)}{T\mu^2\gamma^2} + \frac{(1-\gamma)^2 G^2}{K\mu^2\gamma^2} + \frac{G^2\omega_{\mathrm{TV}}}{\mu\gamma^2}\Big).$$

Table 1: Convergence results and oracle assumptions under different heterogeneous settings. The radii ω_{TV} and ω_{W_1} are defined in (12) and (16) respectively. These two quantities measure the degree of the data heterogeneity under the Total-Variation and the Wasserstein-1 norm respectively. The capability of a weak learning oracle is measured by the corresponding constant γ, which appears in the requirements (5), (10), and (15). For the three settings below, the quantities γ's of the listed oracles are obtained under the $\mathcal{L}^2$ norm, the $\mathcal{L}^\infty$ norm, and a Sobolev-type norm respectively.

Setting	Oracle	Result
$\alpha_i = \alpha$ (general loss)	$\mathcal{Q}^2_{\alpha_i}$, see (5)	$\|f^T - f^*\|_\alpha^2 \xrightarrow{T\to\infty} 0$
$\alpha_i \neq \alpha$ (general loss)	$\mathcal{Q}^\infty_{\alpha_i}$, see (10)	$\|f^T - f^*\|_\alpha^2 \xrightarrow{T\to\infty} O(\omega_{\mathrm{TV}})$
$\alpha_i \neq \alpha$ (square loss)	$\mathcal{Q}^{\mathrm{lip}}_{\alpha_i}$, see (15)	$\|f^T - f^*\|_\alpha^2 \xrightarrow{T\to\infty} O(\omega_{W_1})$

The key to prove the above theorem is to ensure that the local variable function $g_i^{k,t}$ remains bounded during the entire optimization procedure, which allows us to exploit the variational formulation of the total variation (TV) distance in order to relate the local inner product structure to the global one:

$$|\langle f, g\rangle_{\alpha_i} - \langle f, g\rangle_\alpha| \leq 2\|f\|_\infty\|g\|_\infty TV(\alpha, \alpha_i).$$

This is the reason we need the stronger weak learner oracle (10): while the standard $\mathcal{L}^2(\alpha)$ oracle (5) ensures the residual is reduced in average, it may still have large spiky values on $\mathrm{supp}(\alpha_i)$.

For sufficiently large K and T, the first three terms of Theorem 4.1 diminishes to zero. However, the last term that depends on ω_{TV}, the heterogeneity of the local input distributions, will not vanish, which prevents FFGB converging to the global minimizer f^*. The following example shows that it is impossible to reach the exact minimizer without further assumptions for a large class of deterministic algorithms.

Theorem 4.2. *Consider the federated functional minimization problem* (9)*. For any deterministic algorithm $\mathcal{A}$ such that its output f^t is the affine combination of outputs of the weak learning oracle $\mathcal{Q}^\infty_{\alpha_i}$. There exists an instance of the problem* (9) *with $G = O(1)$ and $\omega_{\mathrm{TV}} = 1$ and an adversarial weak learning oracle $\mathcal{Q}^\infty_{\alpha_i} = \mathcal{Q}^\infty_{\alpha_i}(\mathcal{A})$ constructed according to the update rule of $\mathcal{A}$ with $\gamma = 1$ such that $f^t(x) = \mathbf{0}$ for any t and $x \in \mathrm{supp}(\alpha)$.*

As a consequence of the above example, the output of algorithm $\mathcal{A}$ is independent of the conditional distributions β_i^x which is clearly not optimal.

4.2 Special Case: Homogeneous α_i's

A direct implication of Theorem 4.1 is that, when the input distributions are homogeneous, i.e. $\forall i \in [N]$, $\alpha_i = \alpha$, FFGB converges to the global minimizer of the federated functional minimization problem (9). Note that, in this case, the conditional distribution β_i^x can still vary among different clients and so does the joint distributions P_i. In the following, we show that, as long as the input distributions are homogeneous, we can obtain the global convergence of FFGB under weaker regularity and oracle assumptions: We relax Assumption 4.1 to the following one.

Assumption 4.2. *For all $i \in [N]$, the subgradients of $\mathcal{R}_i[f]$ are G-bounded under the $\mathcal{L}^2(\alpha_i)$ norm, i.e. for any $f \in \mathcal{L}^2(\alpha)$, we have $\|\nabla \mathcal{R}_i[f]\|_{\alpha_i} \leq G$.*

and we relax the oracle assumption (10) to the standard one (5), i.e. the one in the $\mathcal{L}^2$ sense.

Theorem 4.3 (Convergence result of FFGB)**.** *Let f^0 be the initializer function. Suppose that Assumption 4.2 holds, and suppose the weak learning oracle $\mathcal{Q}_\alpha^2$ satisfies* (5) *with constant γ. Using the step size $\eta^{k,t} = \frac{2}{\mu(tK+k+1)}$, the output of* FFGB *satisfies*

$$\|f^T - f^*\|_\alpha^2 = O\Bigg(\frac{\|f^0 - f^*\|_\alpha^2}{KT} + \frac{KG^2\log(KT)}{T\gamma^2\mu^2} + \frac{(1-\gamma)G^2}{K\mu^2\gamma^2} + \frac{(1-\gamma)G^2\log(KT)}{KT\mu\gamma^2}\Bigg).$$

In the limit case when $\gamma = 1$, the weak oracle exactly approximates its input and hence the above result degenerates to the one of FedAvg [Li et al., 2019]: $\|f^T - f^*\|_\alpha^2 = O\left(\frac{\|f^0 - f^*\|_\alpha^2}{KT} + \frac{KG^2\log(KT)}{T\mu^2}\right)$. Note that when $\gamma < 1$, in order to have the best convergence rate (up to log factors), we need to set the number of local steps $K = \Omega(\sqrt{T})$, leading to the following corollary.

Corollary 4.1. *Under the conditions of Theorem 4.3, when $\gamma < 1$, the best convergence rate (up to log factors) is $\|f^T - f^*\|_\alpha^2 = O(1/\sqrt{T})$ by choosing $K = \Omega(\sqrt{T})$.*

4.3 Special Case: Square Loss

We show that when the loss ℓ in the local objective functional (8) is the square loss $\ell(y', y) = \frac{1}{2}\|y' - y\|^2$ for $y,\ y' \in \mathcal{Y} \subseteq \mathbb{R}^c$, we can improve the bound on radius of convergence by replacing the TV distance with the Wasserstein-1 distance, if the optimal solution f^* is L-Lipschitz continuous. Note that the Wasserstein distance usually provides a better characterization for the distance between two distributions with mismatched supports. The purpose of this section is to show that at least for certain problems the TV norm type result can be further strengthened. The following improved convergence radius heavily relies on the special structure of the functional gradient $\nabla \mathcal{R}_i$ and the choice of the weak learning oracle. For simplicity, we assume that the domain $\mathcal{Y} \subseteq \mathbb{R}$, the extension to $\mathbb{R}^c$ is analogous.

Assumption on local dataset. In this special case, we assume that P_i is the empirical measure on a finite set of client data $\{(x_{i,j}, y_{i,j}) : \ j \in [M]\}$. We also assume that the labels are generated from the inputs via the L-Lipschitz optimal solution f^* with *no additive noise*, i.e., the data satisfies the following Lipschitzness property: for any $j, j' \in [M]$, we have $|y_{i,j} - y_{i,j'}| \leq L\|x_{i,j} - x_{i,j'}\|$. This assumption is crucial to our choice of subgradient in order to obtain a improved convergence radius.

Choice of subgradient. Note that the functional subgradient $\nabla \mathcal{R}_i[g]$ at g is any function h satisfying that for all $j \in [M]$, $h(x_j) = g(x_{i,j}) - y_{i,j}$. The above assumption allows us to construct an L-Lipschitz function $u_i : \mathbb{R}^d \to \mathbb{R}$ that interpolates the data $\{(x_{i,j}, y_{i,j}) : \ j \in [M]\}$: specifically, this Lipschitz extension is defined as[2]

$$u_i(x) \overset{\Delta}{=} \min_{j \in [M]} \left(y_{i,j} + L\|x - x_{i,j}\|\right). \tag{13}$$

Importantly, the function $h = g - u_i$ is one such subgradient $\nabla \mathcal{R}_i[g]$ and will be used when querying the weak learning oracle.

Choice of weak learning oracle. To obtain the desired convergence radius, the weak learning oracle needs to exploit the special structure of the functional gradient $\nabla \mathcal{R}_i[g] = g - u_i$. Specifically, let g be the variable function which is explicitly known. Upon the query $\phi + g$ with $\|\phi\|_{\mathrm{lip}} < \infty$, the weak learning oracle $\mathcal{Q}_{\alpha_i}^{\mathrm{lip}}$ outputs $h + g$ such that $\|h\|_{\mathrm{lip}} \leq \|\phi\|_{\mathrm{lip}}$ and the following two conditions hold simultaneously

$$\|h - \phi\|_{\alpha,\infty} \leq (1 - \gamma)\|\phi\|_{\alpha,\infty} \tag{14}$$

$$\|h - \phi\|_{\mathrm{lip}} \leq (1 - \gamma)\|\phi\|_{\mathrm{lip}}, \tag{15}$$

for some positive constant $0 < \gamma \leq 1$. Note that, in the notation, we use the superscript "lip" to emphasize that we measure the quality the oracle's output with additional consideration on its Lipschitz continuity. We can implement this oracle using the Sobolev training [Czarnecki et al., 2017]. This is further discussed in the appendix.

Before we present the convergence result of FFGB in this special case, we need the following boundedness assumptions:

Assumption 4.3. *All labels are B-bounded: i.e. $\forall i \in [N], j \in [M], -B \leq y_{i,j} \leq B$, for some $B > 0$. For every pair of measures α_i and $\alpha_{i'}$, $\forall x_{i,j} \in \mathrm{supp}(\alpha_i)$, $\exists x_{i',j'} \in \mathrm{supp}(\alpha_{i'})$ such that $\|x_{i,j} - x_{i',j'}\| \leq D$.*

Theorem 4.4. *Consider the special case of the federated functional minimization problem with square loss. Assume that the optimal solution f^* is L-Lipschitz continuous. Let f^0 be the initializer. We define a proxy of the heterogeneity among the local input distributions*

[2] If the output domain $\mathcal{Y}$ is high dimensional, i.e. $\mathcal{Y} \subseteq \mathbb{R}^c$, then the construction of u_i follows from Kirszbaum's Lipschitz extension theorem [Schwartz, 1969].

α's in the Wasserstein-1 distance as

$$\omega_{W_1} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{W}_1(\alpha, \alpha_i) \tag{16}$$

Define $G^2 = \frac{2L^2}{N^2}\sum_{i,s=1}^{N} \mathrm{W}_2^2(\alpha_s, \alpha_i) + 2B^2$. Under Assumption 4.3, and supposing the weak learning oracle $\mathcal{Q}_\alpha^{\mathrm{lip}}$ satisfies (14) *and* (15) *with constant γ, using the step sizes $\eta^{k,t} = 4/(\mu(tK + k + 1))$, the output of* FFGB *satisfies*

$$\|f^T - f^*\|_\alpha^2 = O\bigg(\frac{K\left(L(LD+B)\omega_{W_1} + G^2\right) log(KT)}{T\mu^2\gamma^2} + \frac{\|f^0 - f^*\|^2}{KT} + \frac{(1-\gamma)^2B^2}{\mu^2\gamma^2K} + \frac{L(LD+B)\omega_{W_1}}{\gamma^2\mu}\bigg).$$

The key component in the above analysis is to ensure that the variable function remains Lipschitz continuous along the entire optimization trajectory. However, maintaining such a property of the variable is subtle and heavily relies on the structure of the chosen subgradient. We elaborate this in the appendix.

5 Practical Usage of FFGB

Reducing memory and communication costs via knowledge distillation. While the FFGB algorithm has theoretical convergence guarantees, it does have two drawbacks for a direct practical implementation. The first is that aggregating the client functions via the direct averaging scheme in line (4) of Algorithm 1 causes the ensemble model to grow in size. A solution to this issue in settings where an *unlabeled* dataset that is very similar to the client data is publicly available, is the *knowledge distillation* technique [Hinton et al., 2015, Bucila et al., 2006]. Suppose ρ denotes the empirical distribution that describes the distillation dataset and let h_θ be the candidate model, e.g. a neural network with parameter θ. An alternative implementation of the client aggregation is to solve:

$$\min_\theta \sum_{x\in\mathrm{supp}(\rho)} \|h_\theta(x) - \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} g_i(x)\|^2, \tag{17}$$

and return h_θ as the output of $\textsc{Aggregate}(\{g_i\}_{i\in\mathcal{S}})$. We note that Lin et al. [2020] has also explored the idea of exploiting knowledge distillation in FL, though with a distillation loss function different from (17). The other drawback is that the function returned by the clients is also an ensemble of size K. This increases communication complexity, and hence, in our experiments, we simply use $K = 1$.

FFGB as Warm Start. The example in Theorem 4.2 shows that fundamental barrier in the federated functional minimization problem under the fully heterogeneous setting, which precludes the exact convergence to the global minimizer. On the other hand, the convergence result in Theorem 4.1 suggests that FFGB is able to quickly converge to a neighborhood of the global minimizer. These theoretical findings inspire us to use FFGB to warm start the existing FL solvers: few rounds of FFGB can be used to give a very good initial model to optimize further with other standard FL methods like FedAvg.

6 Experiments

In this section, we show how FFGB can be used with knowledge distillation to improve the performance of the baseline algorithms on the computer vision task of multiclass classification. The implementation of knowledge distillation requires the access to a unlabelled public dataset, which is usually available for vision tasks. We use FFGB-distill to denote the method that distills the global ensemble into a single model after every FFGB round. Consequently, the per-round communication cost of FFGB-distill is reduced to K which is significantly smaller than $O(TNK)$ in FFGB. Moreover, to have the best performance under communication budgets, we fix $K = 1$ in our current experiments. Note that while the boosting method produces ensemble models with high testing accuracy in the centralized setting, it is excluded from comparison since such a setting is never available in FL. Our code is available here `https://github.com/shenzebang/Federated-Learning-Pytorch`.

Datasets. Three datasets are used in our experiment, CIFAR10, CIFAR100 [Krizhevsky et al., 2009], and EMNIST [Cohen et al., 2017]. We use CIFAR100 as the distillation dataset for CIFAR10 and vice versa. For EMNIST, we use the digits as the distillation dataset for the letters and vice versa. These two subsets of EMNIST, letters and digits, are denoted by EMNIST-L and EMNIST-D respectively. In all experiments, the labels of the distillation dataset are never used.

Heterogeneous client distributions The heterogeneity across local datasets is controlled by dividing the dataset among N clients in the following manner: we choose $s \in [0, 1]$ as the degree of homogeneity. We then randomly select a portion $s \times 100\%$ of the data from the dataset and allocate them equally to all clients; for the remaining $(1 - s) \times 100\%$ portion of the data, we sort the data points by their labels and assign them to the clients sequentially. This is the same scheme as employed in [Karimireddy et al., 2020b, Hsu et al., 2020] to induce heterogeneity. In our experiment, we are interested in the heterogeneous setting and s takes

Table 2: Number of transmitted models for different methods to reach a *moderate* target testing accuracy, a performance metric used in the previous work [Acar et al., 2020]. The s numbers indicate degree of homogeneity in the client datasets. In the first column, the datasets in bold are the training sets and the ones in parenthesis are the distillation datasets.

Dataset	Level of heterogeneity	Target Accuracy	FFGB-distill **(this work)**	FedAvg-distill	FedDyn (FedPD)	FedAvg
CIFAR10 (CIFAR100)	s = 0.1	0.50	**1**	4	20	25
		0.55	**3**	19	30	39
	s = 0.2	0.54	**1**	3	16	18
		0.60	**3**	25	24	32
	s = 0.3	0.55	**1**	3	15	16
		0.60	**3**	11	20	24
CIFAR100 (CIFAR10)	s = 0.1	0.2	**4**	> 100	70	> 100
		0.25	**20**	> 100	> 100	> 100
	s = 0.2	0.2	**1**	81	23	74
		0.25	**17**	> 100	37	> 100
	s = 0.3	0.2	**1**	66	19	25
		0.25	**7**	95	29	40
EMNIST-L (EMNIST-D)	s = 0.1	0.89	**8**	>100	93	>100
	s = 0.2	0.89	**5**	>100	68	>100
	s = 0.3	0.89	**3**	>100	43	>100
EMNIST-D (EMNIST-L)	s = 0.1	0.99	**4**	>100	>100	>100
	s = 0.2	0.99	**4**	>100	>100	>100
	s = 0.3	0.99	**4**	>100	>100	>100

value from $\{0.1, 0.2, 0.3\}$. We set $N = 100$ in all cases.

Architecture of the Neural Network. We follow the choice of model architectures in [Acar et al., 2020, McMahan et al., 2017a]. For CIFAR10 and CIFAR100, we use a CNN model consisting of 2 convolution layers with 64 5×5 filters followed by 2 fully connected layers with 394 and 192 neurons. For EMNIST, we use the same convolution layers and reduce the fully connected layers to (120, 84) neurons. We note that higher testing accuracy on the included datasets can be obtain by using models with high capacity, but is orthogonal to our research.

Choice of hyperparameters. In our implementation of FFGB-distill, we choose $\mathcal{Q}^2_{\alpha_i}$ as the weak learning oracle. To solve the corresponding optimization problem (6) on the clients, we run Adam for 100 epochs on the local data, with learning rate 0.001, batch size 64, without weight decay. The learning rate of FFGB-distill is fixed as 10 which according to our observation gives the best results. We minimize the distillation loss (17) using Adam with the same hyperparameters adopted when computing the weak learning oracle. For FedAvg-distill, the hyperparameters are chosen as suggested by the original paper. For a fair comparison, the number of local epochs is set to 100, same as FFGB-distill. For FedDyn (or equivalently FedPD) and FedAvg, after a global communication round, we run SGD for 10 local epochs with batch size 64. The learning rate is grid searched from $\{0, 01, 0.05, 0.1\}$ (FedAvg can diverge with a large learning rate). We use gradient clipping for all experiment setup with maximum gradient norm 5.

Data augmentation. The data augmentation technique has a huge impact on the output accuracy. For the experiments on CIFAR10 and CIFAR100, we use random horizontal flip and random crop with size 32 and padding 4 to transform the minibatch used in each SGD/Adam step. For the experiments on EMNIST, we conduct no data augmentation.

6.1 Result Summary

We run FFGB-distill and its competitors on CIFAR10 and CIFAR100 with different levels of data heterogeneity. The results are reported in Table 2 and Figure 1 and are summarized as follows.

- **FFGB-distill quickly converges to a moderate accuracy.** In Table 2, we report the number of models transmitted so that the included methods reach a *moderate* target accuracy. We emphasize that the reported target accuracy are obtained after only a few communication rounds, e.g. FFGB-distill obtains a 50% testing accuracy after a single iteration, and are presented to motivate that FFGB-distill provides an effective warm-start. This should not be con-

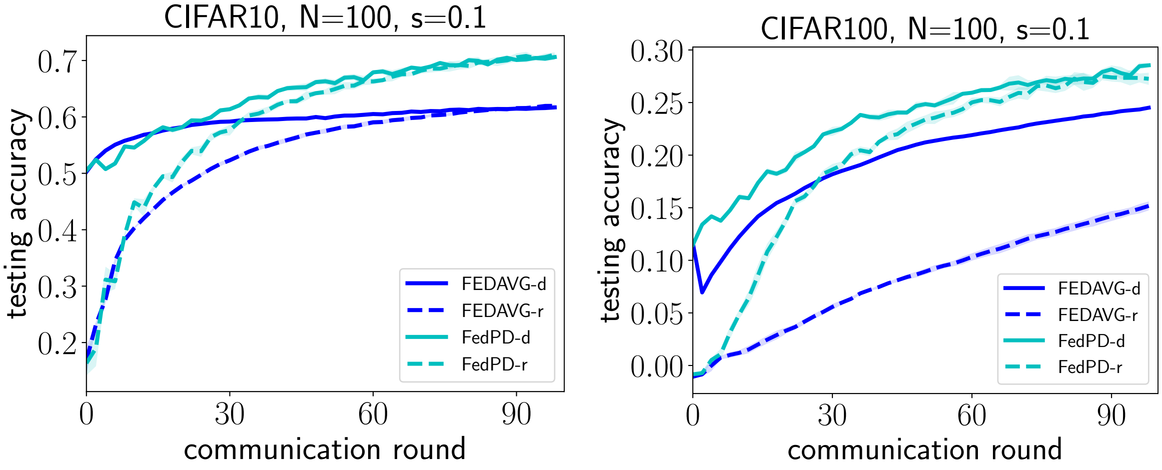

Figure 1: Performance of baseline methods when initialized with a single FFGB-distill step ("<alg>-d") and randomly initialized ("<alg>-r"). Here FedPD is a synonym for FedDyn as these two methods are proved to be equivalent.

fused with the high accuracy obtained after hundreds of communication rounds. From our results, we see that FFGB-distill has a clear advantage over the baselines, including FedDyn which to the best of our knowledge is the SOTA FL solver. Note that the comparisons between FFGB-distill and FedAvg-distill can also be regarded as an ablation study: Both methods use knowledge distillation to fuse the ensemble on the server, but each with a different client update scheme. The clear advantage of FFGB-distill over FedAvg-distill shows that our functional minimization scheme also has a significant contribution to the observed improvements. Moreover, while FedAvg-distill improves the performance of FedAvg on CIFAR10, on the more difficult CIFAR100 in with higher data heterogeneity, FedAvg-distill has worse performance than FedAvg. This phenomenon is possibly due to that FedAvg fails to generate reasonable local models with limited communication rounds, a necessity for the success of knowledge distillation.

• **FFGB-distill boosts SOTA FL solver as a warm start.** The empirical results in Table 2 show that FFGB-distill achieve a moderate accuracy within the first few rounds. This observation motivates us to utilize FFGB-distill as a warm start method for existing FL solves. To this end, we conduct the ablation study where we run FedAvg and FedDyn using either the output of *a single* FFGB-distill *step* as initialization or random initialization. Figure 1 shows that one step of FFGB-distill significantly boosts the performance of all included baselines.

Conclusion

In this paper, we initiate the theory of boosting in the Federated Learning setting. We develop *federated functional gradient boosting* (FFGB) an algorithm that is designed to handle the challenge of data heterogeneity. Under appropriate assumptions on the weak learning oracle, the FFGB algorithm is proved to efficiently converge to certain neighborhoods of the global optimum. The radii of these neighborhoods depend upon the level of heterogeneity measured via the total variation distance and the much tighter Wasserstein-1 distance, and diminish to zero as the setting becomes more homogeneous. While our work serves as the first step towards extending the theory of boosting to the Federated Learning setting, more work is needed to ensure that the memory and communication costs are reduced to have an impact on FL in practice. One of the possible applications, as hinted by our theoretical findings, is to use FFGB with knowledge distillation to warm-start existing Federated Learning solvers. In our experiments, we observe this strategy boosts performance in highly heterogeneous settings.

Acknowledgement

Amin Karbasi's research is partially supported by NSF CAREER Award (IIS-1845032) and ONR YIA (N00014-19-1-2406). The research of Zebang Shen and Hamed Hassani is supported by AFOSR Award 19RT0726, NSF HDR TRIPODS award 1934876, NSF award CPS-1837253, NSF award CIF-1910056, and NSF CAREER award CIF-1943064.

References

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2020.

Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *SIGKDD*, pages 535–541. ACM, 2006.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pages 2921–2926. IEEE, 2017.

Wojciech M Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. Sobolev training for neural networks. In *Advances in Neural Information Processing Systems*, pages 4278–4287, 2017.

Nigel Duffy and David P. Helmbold. Boosting methods for regression. *Mach. Learn.*, 47(2-3):153–200, 2002.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55 (1):119–139, 1997.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), October 2001a.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001b.

Alexander Grubb and J Andrew Bagnell. Generalized boosting algorithms for convex optimization. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1209–1216, 2011.

Jenny Hamer, Mehryar Mohri, and Ananda Theertha Suresh. Fedboost: A communication-efficient algorithm for federated learning. In *International Conference on Machine Learning*, pages 3973–3983. PMLR, 2020.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. URL `http://arxiv.org/abs/1503.02531`.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Federated visual classification with real-world data distribution. *arXiv preprint arXiv:2003.08082*, 2020.

Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv preprint arXiv:2008.03606*, 2020a.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020b.

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

Qinbin Li, Zeyi Wen, and Bingsheng He. Practical federated gradient boosting decision trees. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4642–4649, 2020a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *MLSys*. mlsys.org, 2020b.

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*, 2019.

Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, 33:2351–2363, 2020.

Llew Mason, Jonathan Baxter, Peter L Bartlett, and Marcus R Frean. Boosting algorithms as gradient descent. In *Advances in neural information processing systems*, pages 512–518, 2000.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017a.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017b.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11 (5-6):355–607, 2019.

Gunnar Rätsch, Sebastian Mika, and Manfred K. Warmuth. On the convergence of leveraging. In *NeurIPS*, pages 487–494, 2001.

Robert E. Schapire and Yoav Freund. *Boosting: Foundations and Algorithms*. MIT Press, 2012.

Jacob T Schwartz. *Nonlinear functional analysis*, volume 4. CRC Press, 1969.

Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *Annals of Statistics*, 33(4):1538–1579, 2005.

A Total Variation Distance and Wasserstein-1 Distance between Probability Measures

Given two probability distributions $\alpha, \beta \in \mathcal{M}_+^1(\mathcal{X})$, the total variation distance between α and β is

$$\mathrm{TV}(\alpha, \beta) = \sup_{A \in \mathcal{F}} |\alpha(A) - \beta(A)|, \tag{18}$$

where $\mathcal{F}$ is the Borel sigma algebra over $\mathcal{X}$.

The p-Wasserstein metric between α and β is defined as

$$W_p(\alpha, \beta) \overset{\Delta}{=} \min_{\pi \in \Pi} \left(\int_{\mathcal{X}^2} \|x - y\|^p d\pi(x, y) \right)^{1/p}, \tag{19}$$

where $\Pi(\alpha, \beta) \overset{\Delta}{=} \{\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{X}) | \sharp_1 \pi = \alpha, \sharp_2 \pi = \beta\}$ is the set of joint distributions with given marginal distributions α and β. Here $\sharp_i$ denotes the marginalization.

B Experiments: Implementing the Weak Learning Oracles

We now discuss the implementations of the weak learning oracles. In our experiments, we only use the weak learning oracle $\mathcal{Q}_\alpha^2$. We discuss the implementation of the oracles $\mathcal{Q}_\alpha^\infty$ and $\mathcal{Q}_\alpha^{\mathrm{lip}}$ for completeness, but the suggested schemes may not necessarily be very efficient in practice.

Implementing $\mathcal{Q}_\alpha^\infty$. Let ϕ be the input to the oracle and let h_θ be the candidate weak learner to be trained. Here h_θ is a neural network with parameter θ.

We can implement the oracle by solving

$$\min_\theta \max_{x \in \mathrm{supp}(\alpha)} \|\phi(x) - h_\theta(x)\|^2. \tag{20}$$

Implementing $\mathcal{Q}_\alpha^{\mathrm{lip}}$. Recall that we assume the input to $\mathcal{Q}_\alpha^{\mathrm{lip}}$ to be of the form $-\phi + g$, where g is some function that is explicitly available (usually, the variable function) and recall that $\mathcal{Q}_\alpha^{\mathrm{lip}}$ outputs $-h_\theta + g$ where h_θ is a neural network with parameter θ. In other words, $\mathcal{Q}_\alpha^{\mathrm{lip}}$ only approximates ϕ in the input with h_θ and leaves the known part, g, untouched.

We can implement the oracle $\mathcal{Q}_\alpha^{\mathrm{lip}}$ by solving

$$\min_\theta \left(\max_{x \in \mathrm{supp}(\alpha)} \|\phi(x) - h_\theta(x)\|^2 \right) + \int_{\mathcal{X}} \|\nabla_x \phi(x) - \nabla_x h_\theta(x)\|^2 dx. \tag{21}$$

Note that the gradient of the input ϕ is available as we have the explicit expression of $\phi = u_i + \Delta_i^{k,t}$ (u_i is defined in (13)). The above scheme is similar to the Sobolev training scheme (1) in Czarnecki et al. [2017].

C Proof of Theorem 4.2

Suppose that we have two clients each with Dirac local distributions, i.e., $\alpha_1 = \delta_{x_1}$ and $\alpha_2 = \delta_{x_2}$ with $x_1 \neq x_2$. Consider the cross entropy loss ℓ defined in Eq.(1) with ℓ_2 regularization. Hence the local objective functional is $\mathcal{F}_i[f] = \ell(f(x_i), y_i) + ||f(x_i)||^2 * \mu/2$ (there is only one term since α_i is a Dirac). We can compute that the functional gradient is $\nabla \mathcal{F}_i[f] = \nabla_1 \ell(f(x_i), y_i)\delta_{x_i} + \mu f(x_i)\delta_{x_i}$. For simplicity we assume every client takes a single local step, i.e. $K = 1$ and that the algorithm $\mathcal{A}$ updates the variable by $f^{t+1} = f^t - \eta^t(g_1^t + g_2^t)$ given the outputs of the weak learner g_1^t and g_2^t. The following example can be easily generalized to other update rules where f^t is the affine combination of g_i^t's and there are more than one local steps, i.e. $K > 1$.

We construct an adversarial weak learning oracle $\mathcal{Q}(\mathcal{A})$ according to the update rule of $\mathcal{A}$ such that $f^t(x_i) = [0, 0]$ for any t and i under the initialization $f^0(x_i) = [0, 0]$, in the following manner: The adversarial oracle $\mathcal{Q}(\mathcal{A})$ returns g_1^t and g_2^t such that

$$g_1^t(x_1) = \nabla_1 \ell(f(x_1), y_1), \quad g_1^t(x_2) = -\nabla_1 \ell(f(x_2), y_2);$$

$$g_2^t(x_1) = -\nabla_1 \ell(f(x_1), y_1), \quad g_2^t(x_2) = \nabla_1 \ell(f(x_2), y_2).$$

Here we note that the regularization term vanishes since f^t is a zero function. We can check that in this case, the adversarial oracle $\mathcal{O}(\mathcal{A})$ satisfies (10) with $\gamma = 1$ since on the support of α_i, we have $g_i^t(x_i) = \nabla \mathcal{F}_i[f^t](x_i)$. And we can simply set $\bar{G}_\gamma$ to be G in (11).

However, we always have $f^{t+1} = f^t - \eta^t(g_1^t + g_2^2) \equiv 0$ given that $f^t \equiv 0$. Consequently, the output ensemble model of algorithm $\mathcal{A}$ is independent of the conditional distributions $p(y_i|x_i)$ which is clearly not optimal. In fact, we can compute in this special case that

$$\forall t, \ \|f^t - f^*\|_\alpha^2 = \|f^*\|_\alpha^2. \tag{22}$$

This lower bound agrees with the result in Theorem 4.1 for $T \to \infty$ if we pick $\|f^*\| = G/\mu$ (see (31)).

D Proof of Theorem 4.1

We now present the proof of FFGB in the most general setting where we make no assumption on the homogeneity of α's, the distributions over the input space, and the loss function ℓ in the objective (8) can be any function that is convex w.r.t. its first input. Unfortunately the statement of Theorem 4.1 in the main body contains a typo which is fixed in the following restatement and is marked in red.

Theorem D.1 (Theorem 4.1 restated)**.** *Let f^0 be the initializer function. We define a proxy of the heterogeneity among the local input distributions α's in the* TV *distance as*

$$\omega_{\mathrm{TV}} \triangleq \frac{1}{N}\sum_{i=1}^{N} \mathrm{TV}(\alpha, \alpha_i). \tag{23}$$

Under Assumption 4.1, and supposing the weak learning oracle $\mathcal{Q}_\alpha^\infty$ satisfies (10) *and* (11) *with constant γ and $\bar{G}_\lambda = 2G/\lambda$, using the step sizes $\eta^{k,t} = 4/(\mu(tK + k + 1))$, the output of* FFGB *satisfies*

$$\|f^T - f^*\|_\alpha^2 = O\Big(\frac{\|f^0 - f^*\|_\alpha^2}{KT} + \frac{KG^2 log(KT)}{T\mu^2\gamma^2} + \frac{(1-\gamma)^2 G^2}{K\mu^2\gamma^2} + \frac{G^2\omega_{\mathrm{TV}}}{\mu^2\gamma^2}\Big).$$

Proof. Similar to the proof of Theorem 4.3, for simplicity, in this proof, we define $\hat{h}_i^{k,t} := h_i^{k,t} + \mu g_i^{k,t}$.

We first show that $\Delta_i^{k,t}$, the residual variable, and $h_i^{k,t}$, the output of the weak learing oracle $\mathcal{Q}_{\alpha_i}^\infty$, are bounded under the $\mathcal{L}^\infty(\alpha_i)$ norm: Note that $\Delta_i^{0,t} \equiv 0$. Besides, for $x \in \mathrm{supp}(\alpha_i)$, in each iteration $|\Delta_i^{k,t}(x)|$ is first increased at most by G after adding $\nabla \mathcal{R}_i[g_i^{k,t}]$ and is then reduced by at least $1 - \gamma$ after subtracting the weak learner $h_i^{k,t}$. Consequently, we have

$$\|\Delta_i^{k,t}\|_{\alpha_i,\infty} \le (1-\gamma)\left(\|\Delta_i^{k-1,t}\|_{\alpha_i,\infty} + G\right) \text{ and } \|\Delta_i^{0,t}\|_\infty = 0 \Rightarrow \forall k, \|\Delta_i^{k,t}\|_{\alpha_i,\infty} \le \frac{1-\gamma}{\gamma} G = G_\gamma^1. \tag{24}$$

Further, using the triangle inequality we have $\|\Delta_i^{k-1,t} + \nabla \mathcal{R}_i[g_i^{k,t}]\|_{\alpha_i,\infty} \le G + \frac{1-\gamma}{\gamma} G = \frac{G}{\gamma}$ and hence

$$\|h_i^{k,t}\|_{\alpha_i,\infty} \le \|\Delta_i^{k-1,t} + \nabla \mathcal{R}_i[g_i^{k,t}] - h_i^{k,t}\|_{\alpha_i,\infty} + \|\Delta_i^{k-1,t} + \nabla \mathcal{R}_i[g_i^{k,t}]\|_{\alpha_i,\infty} \le \frac{2-\gamma}{\gamma} G = G_\gamma^2. \tag{25}$$

Based on the above results, the following lemmas characterize the boundedness of $h_i^{k,t}$, $g_i^{k,t}$, and $\hat{h}_i^{k,t}$.

Lemma D.1. *Assume that the initial function satisfies $\|f^0\|_\infty \le \frac{2G}{\gamma\mu}$. Then for all $t \ge 0$, $\|f^t\| \le \frac{2G}{\mu\gamma}$ and $\|\bar{g}^{k,t}\|_\infty \le \frac{2G}{\gamma\mu}$ for all $1 \le k \le K$.*

Proof. For $t = 0$, $\bar{g}^{1,0} = f^0$ and hence $\|\bar{g}^{1,0}\|_\infty \le \frac{2G}{\gamma\mu}$ due to the initialization. Now assume that for $t = \tau$ the statement holds. Therefore $\|f^\tau\|_\infty \le \frac{2G}{\gamma\mu}$. So for $t = \tau + 1$, $\|\bar{g}^{1,t}\|_\infty \le \frac{2G}{\gamma\mu}$. From the update rule in line (5) of Algorithm 2, we have

$$\bar{g}^{k+1,t} = (1 - \mu\eta^{k,t})\bar{g}^{k,t} + \frac{1}{N}\sum_{i=1}^{N} \eta^{k,t} h_i^{k,t} \tag{26}$$

Since we have $\|h_i^{k,t}\|_\infty \leq \frac{2G}{\gamma\mu}$ from (25) and the assumption (11) on the weak learning oracle $\mathcal{Q}_{\alpha_i}^\infty$. Recursively, we have

$$\|\bar{g}^{k+1,t}\|_\infty - \frac{2G}{\gamma\mu} \leq (1-\mu\eta^{k,t})\left(\|\bar{g}^{k,t}\|_\infty - \frac{2G}{\gamma\mu}\right) \leq 0,$$

which leads to the conclusion. □

Combing the above results, we have the boundedness of $\hat{h}_i^{k,t}$ and $\nabla F_i[g_i^{k,t}]$: $\|\hat{h}_i^{k,t}\|_\infty \leq \frac{4G}{\gamma}$ and $\|\nabla\mathcal{F}_i[g_i^{k,t}]\|_\infty \leq \frac{4G}{\gamma}$.

For a fixed communication round t, we define two hypothetical sequences $\bar{g}^{k,t} = \frac{1}{N}\sum_{i=1}^N g_i^{k,t}$ and $\bar{h}^{k,t} = \frac{1}{N}\sum_{i=1}^N \hat{h}_i^{k,t}$. Note that $\bar{g}^{1,t} = f^t$. From the update rule in line 5 of Algorithm 2, we write

$$\|\bar{g}^{k+1,t} - f^*\|_\alpha^2 = \|\bar{g}^{k,t} - f^*\|_\alpha^2 + (\eta^{k,t})^2\|\bar{h}^{k,t}\|_\alpha^2 - 2\eta^{k,t}\langle\bar{g}^{k,t} - f^*, \bar{h}^{k,t}\rangle_\alpha. \tag{27}$$

For the second term, we have $\|\bar{h}^{k,t}\|_\alpha^2 \leq \|\bar{h}^{k,t}\|_\infty^2 = O(\frac{G^2}{\gamma^2})$.

The last term of (27) can be split as

$$\begin{aligned}
&-2\langle\bar{g}^{k,t} - f^*, \bar{h}^{k,t}\rangle_\alpha & (28)\\
=& -\frac{2}{N}\sum_{i=1}^N\langle\bar{g}^{k,t} - f^*, \hat{h}_i^{k,t}\rangle_{\alpha_i} + \left(\langle\bar{g}^{k,t} - f^*, \hat{h}_i^{k,t}\rangle_\alpha - \langle\bar{g}^{k,t} - f^*, \hat{h}_i^{k,t}\rangle_{\alpha_i}\right) &\\
=& \frac{2}{N}\sum_{i=1}^N\langle g_i^{k,t} - \bar{g}^{k,t}, \hat{h}^{k,t}\rangle_{\alpha_i} + \langle f^* - g_i^{k,t}, \nabla\mathcal{F}_i[g_i^{k,t}]\rangle_{\alpha_i} + \langle f^* - g_i^{k,t}, \hat{h}^{k,t} - \nabla\mathcal{F}_i[g_i^{k,t}]\rangle_{\alpha_i} &\\
&\quad + \left(\langle\bar{g}^{k,t} - f^*, \hat{h}^{k,t}\rangle_{\alpha_i} - \langle\bar{g}^{k,t} - f^*, \hat{h}^{k,t}\rangle_\alpha\right). & (29)
\end{aligned}$$

Using the variational formulation of the TV norm, with the boundedness of $\bar{g}^{k,t}$ and $\hat{h}_i^{k,t}$ one has

$$|\langle\bar{g}^{k,t} - f^*, \hat{h}_i^{k,t}\rangle_\alpha - \langle\bar{g}^{k,t} - f^*, \hat{h}_i^{k,t}\rangle_{\alpha_i}| \leq O\left(G^2/\mu\gamma^2 \cdot \mathrm{TV}(\alpha, \alpha_i)\right). \tag{30}$$

where we know from the first-order optimalily solution of f^*:

$$\nabla\mathcal{F}[f^*] = \mu f^* + \frac{1}{N}\sum_{i\in[N]}\nabla\mathcal{R}_i[f^*] \equiv 0 \Rightarrow \|f^*\|_{\alpha,\infty} \leq \frac{1}{\mu}\max_i\|\nabla\mathcal{R}_i[f^*]\|_{\alpha_i,\infty} \leq \frac{G}{\mu}. \tag{31}$$

Denote $\omega_{TV} \triangleq \frac{1}{N}\sum_{i=1}^N \mathrm{TV}(\alpha, \alpha_i)$. We hence have

$$\frac{2}{N}\sum_{i=1}^N\left(\langle\bar{g}^{k,t} - f^*, h_i^{k,t}\rangle_{\alpha_i} - \langle\bar{g}^{k,t} - f^*, h_i^{k,t}\rangle_\alpha\right) \leq \delta \triangleq O(G^2\omega_{TV}/\mu\gamma^2). \tag{32}$$

The first term of (29) can be bounded by

$$\begin{aligned}
&\frac{2}{N}\sum_{i=1}^N\langle g_i^{k,t} - \bar{g}^{k,t}, h_i^{k,t}\rangle_{\alpha_i}\\
\leq& \frac{1}{N}\sum_{i=1}^N\eta^{k,t}\|h_i^{k,t}\|_{\alpha_i}^2 + \frac{1}{\eta^{k,t}}\|g_i^k - \bar{g}^k\|_{\alpha_i}^2\\
\leq& \eta^{k,t}4G^2/\gamma^2 + \frac{1}{N}\sum_{i=1}^N\frac{1}{\eta^{k,t}}\|g_i^k - \bar{g}^k\|_{\alpha_i}^2 = O(\frac{\eta^{k,t}G^2}{\gamma^2}) + \frac{1}{\eta^{k,t}}\cdot\frac{1}{N}\sum_{i=1}^N\|g_i^k - \bar{g}^k\|_{\alpha_i}^2.
\end{aligned}$$

The second term of (29) can be bounded by using the μ-strong convexity of $\mathcal{F}_i$:

$$\frac{2}{N}\sum_{i=1}^N\langle f^* - g_i^{k,t}, \nabla\mathcal{F}_i[g_i^{k,t}]\rangle_{\alpha_i} \leq -\frac{2}{N}\sum_{i=1}^N\left(\mathcal{F}_i[g_i^{k,t}] - \mathcal{F}_i[f^*]\right) - \frac{2}{N}\sum_{i=1}^N\frac{\mu}{2}\|f^* - g_i^{k,t}\|_{\alpha_i}^2.$$

For the first term above, using the optimality of f^*, we have

$$\begin{aligned}
& -\frac{2}{N}\sum_{i=1}^{N}\Big(\mathcal{F}_i[g_i^{k,t}] - \mathcal{F}_i[f^*]\Big) \\
= & -\frac{2}{N}\sum_{i=1}^{N}\Big(\mathcal{F}_i[g_i^{k,t}] - \mathcal{F}_i[\bar{g}^{k,t}] + \mathcal{F}_i[\bar{g}^{k,t}] - \mathcal{F}_i[f^*]\Big) \leq -\frac{2}{N}\sum_{i=1}^{N}\Big(\mathcal{F}_i[g_i^{k,t}] - \mathcal{F}_i[\bar{g}^{k,t}]\Big).
\end{aligned}$$

For the second term, we have

$$\|f^* - g_i^{k,t}\|_{\alpha_i}^2 \leq 2\|f^* - \bar{g}^{k,t}\|_{\alpha_i}^2 + 2\|\bar{g}^{k,t} - g_i^{k,t}\|_{\alpha_i}^2.$$

Combine the above inequality to yield

$$\begin{aligned}
& \frac{2}{N}\sum_{i=1}^{N}\langle f^* - g_i^{k,t}, \nabla\mathcal{F}_i[g_i^{k,t}]\rangle_{\alpha_i} \\
\leq & -\frac{2}{N}\sum_{i=1}^{N}\Big(\mathcal{F}_i[g_i^{k,t}] - \mathcal{F}_i[\bar{g}^{k,t}]\Big) - \frac{\mu}{4}\|f^* - \bar{g}^{k,t}\|_{\alpha}^2 \\
& + \frac{\mu}{2N}\sum_{i=1}^{N}\|g_i^{k,t} - \bar{g}^{k,t}\|_{\alpha_i}^2 - \frac{\mu}{2N}\sum_{i=1}^{N}\|f^* - g_i^{k,t}\|_{\alpha_i}^2 \\
\leq & -\frac{2}{N}\sum_{i=1}^{N}\langle\nabla\mathcal{F}_i[\bar{g}^{k,t}], g_i^{k,t} - \bar{g}^{k,t}\rangle_{\alpha_i} - \frac{\mu}{4}\|f^* - \bar{g}^{k,t}\|_{\alpha}^2 \\
& + \frac{\mu}{2N}\sum_{i=1}^{N}\|g_i^{k,t} - \bar{g}^{k,t}\|_{\alpha_i}^2 - \frac{\mu}{2N}\sum_{i=1}^{N}\|f^* - g_i^{k,t}\|_{\alpha_i}^2 \\
\leq & \frac{1}{N}\sum_{i=1}^{N}\eta^{k,t}\|\nabla\mathcal{F}_i[\bar{g}^{k,t}]\|_{\alpha_i}^2 + \frac{1}{\eta^{k,t}}\|g_i^{k,t} - \bar{g}^{k,t}\|_{\alpha_i}^2 \\
& - \frac{\mu}{4}\|f^* - \bar{g}^{k,t}\|_{\alpha}^2 + \frac{\mu}{2N}\sum_{i=1}^{N}\|g_i^{k,t} - \bar{g}^{k,t}\|_{\alpha_i}^2 - \frac{\mu}{2N}\sum_{i=1}^{N}\|f^* - g_i^{k,t}\|_{\alpha_i}^2 \\
\leq & \eta^{k,t}\frac{16G^2}{\gamma^2} + (\frac{\mu}{2} + \frac{1}{\eta^{k,t}})\frac{1}{N}\sum_{i=1}^{N}\|g_i^{k,t} - \bar{g}^{k,t}\|_{\alpha_i}^2 - \frac{\mu}{4}\|f^* - \bar{g}^{k,t}\|_{\alpha}^2 - \frac{\mu}{2N}\sum_{i=1}^{N}\|f^* - g_i^{k,t}\|_{\alpha_i}^2 \qquad (33)
\end{aligned}$$

Note that $\|\bar{g}^{k,t} - f^t\|_{\alpha_i}^2 = \|\sum_{\kappa=1}^{k-1}\eta^{\kappa,t}\bar{h}^{\kappa,t}\|_{\alpha_i}^2$ and $\eta^{t,\kappa} \leq 2\eta^{t,k}$ for $\kappa \leq k$. Therefore, $\frac{1}{N}\sum_{i=1}^{N}\|g_i^{k,t} - \bar{g}^{k,t}\|_{\alpha_i}^2$ can be bounded by (we use $\|\cdot\|_{\alpha_i} \leq \|\cdot\|_{\infty}$ in the following)

$$\begin{aligned}
& \frac{1}{N}\sum_{i=1}^{N}\|g_i^{k,t} - \bar{g}^{k,t}\|_{\alpha_i}^2 \\
= & \frac{1}{N}\sum_{i=1}^{N}\|g_i^{k,t} - f^t + f^t - \bar{g}^{k,t}\|_{\alpha_i}^2 \\
\leq & \frac{1}{N}\sum_{i=1}^{N}2\|g_i^{k,t} - f^t\|_{\alpha_i}^2 + 2\|f^t - \bar{g}^{k,t}\|_{\alpha_i}^2 \leq 36\sum_{\kappa=1}^{k-1}(\eta^{\kappa,t})^2G^2/\gamma^2 \leq 144(\eta^{k,t})^2K^2G^2/\gamma^2 \\
= & O(\frac{(\eta^{k,t})^2K^2G^2}{\gamma^2}).
\end{aligned}$$

Plug in the above results into (27) to yield (note that $\hat{h}_i^{k,t} - \nabla\mathcal{F}_i[g_i^{k,t}] = h_i^{k,t} - \nabla\mathcal{R}_i[g_i^{k,t}]$)

$$\begin{aligned}\|\bar{g}^{k+1,t} - f^*\|_\alpha^2 \leq & (1 - \frac{\mu\eta^{k,t}}{4})\|\bar{g}^{k,t} - f^*\|_\alpha^2 + O(\frac{(\eta^{k,t})^2K^2G^2}{\gamma^2}) \\ & + \frac{2\eta^{k,t}}{N}\sum_{i=1}^N \langle f^* - g_i^{k,t}, h_i^{k,t} - \nabla\mathcal{R}_i[g_i^{k,t}]\rangle_{\alpha_i} + \eta^{k,t}\delta - \frac{\mu}{2N}\sum_{i=1}^N \|f^* - g_i^{k,t}\|_{\alpha_i}^2\end{aligned}$$

Recall that $\eta^{k,t} = \frac{4}{\mu(tK+k+1)}$ and multiply both sides by $(Kt+k+1)$

$$\begin{aligned}& (Kt+k+1)\|\bar{g}^{k+1,t} - f^*\|^2 \\ \leq\ & (Kt+k)\|\bar{g}^{k,t} - f^*\|^2 + O(\frac{K^2G^2}{\mu^2\gamma^2(Kt+k+1)}) \\ & + \frac{4}{\mu N}\sum_{i=1}^N \langle f^* - g_i^{k,t}, h_i^{k,t} - \nabla\mathcal{R}_i[g_i^{k,t}]\rangle_{\alpha_i} + \frac{2\delta}{\mu} - \frac{\mu}{2N}\sum_{i=1}^N \|f^* - g_i^{k,t}\|_i^2\end{aligned}$$

Sum from $k = 1$ to K

$$\begin{aligned}& (Kt+K+1)\|\bar{g}^{k+1,t} - f^*\|^2 \\ \leq\ & (Kt+1)\|\bar{g}^{1,t} - f^*\|^2 + O\left(\frac{K^2G^2}{\mu^2\gamma^2}(\log(Kt+K+1) - \log(Kt+1))\right) \\ & + \frac{4}{\mu N}\sum_{i=1}^N\sum_{k=1}^K \langle f^* - g_i^{k,t}, h_i^{k,t} - \nabla\mathcal{R}_i[g_i^{k,t}]\rangle_{\alpha_i} + \frac{2\delta K}{\mu} - \frac{2}{N}\sum_{k=1}^K\sum_{i=1}^N \|f^* - g_i^{k,t}\|_{\alpha_i}^2\end{aligned}$$

We analyze the first term in the second line above as follows.

$$\begin{aligned}& \sum_{k=1}^K \langle f^* - g_i^{k,t}, \hat{h}_i^{k,t} - \nabla\mathcal{F}_i[g_i^{k,t}]\rangle_\alpha = \sum_{k=1}^K \langle f^* - g_i^{k,t}, h_i^{k,t} - \nabla\mathcal{R}_i[g_i^{k,t}]\rangle_\alpha \\ = & \sum_{k=1}^K \langle f^* - g_i^{k,t}, h_i^{k,t} - (\nabla\mathcal{R}_i[g_i^{k,t}] + \Delta_i^{k-1})\rangle_\alpha + \sum_{k=1}^K \langle f^* - g_i^{k,t}, \Delta_i^{k-1}\rangle_\alpha \\ = & \sum_{k=1}^K \langle f^* - g_i^{k,t}, -\Delta_i^k\rangle_\alpha + \sum_{k=2}^K \langle f^* - g_i^{k,t}, \Delta_i^{k-1}\rangle_\alpha + \langle f^* - g_i^1, \Delta_i^0\rangle_\alpha \qquad \&\Delta_i^0 = 0 \\ = & \sum_{k=1}^K \langle f^* - g_i^{k,t}, -\Delta_i^k\rangle_\alpha + \sum_{k=1}^{K-1} \langle f^* - g_i^{k+1}, \Delta_i^k\rangle_\alpha \\ = & \sum_{k=1}^K \langle f^* - g_i^{k,t}, -\Delta_i^k\rangle_\alpha + \sum_{k=1}^{K-1} \langle f^* - g_i^{k,t}, \Delta_i^k\rangle_\alpha + \sum_{k=1}^{K-1} \langle \eta_t^k h_i^{k,t}, \Delta_i^k\rangle_\alpha \\ = & \langle f^* - g_i^{K,t}, -\Delta_i^K\rangle_\alpha + \sum_{k=1}^{K-1} \langle \eta_t^k h_i^{k,t}, \Delta_i^k\rangle_\alpha \\ \leq & \frac{\mu}{2}\|f^* - g_i^{K,t}\|^2 + \frac{1}{2\mu}(\frac{1-\gamma}{\gamma})^2G^2 + \frac{(1-\gamma)(2-\gamma)}{\gamma^2}G^2\sum_{k=1}^{K-1}\eta^{k,t} \\ \leq & \frac{\mu}{2}\|f^* - g_i^{K,t}\|^2 + \frac{1}{2\mu}(\frac{1-\gamma}{\gamma})^2G^2 + \frac{(1-\gamma)(2-\gamma)}{\gamma^2}G^2(\log(tK+K) - \log(tK+2)). \qquad (34)\end{aligned}$$

Using this result, we obtain

$$\begin{aligned}& (K(t+1)+1)\|f^{t+1} - f^*\|_\alpha^2 \\ \leq\ & (Kt+1)\|f^t - f^*\|_\alpha^2 + O\left(\frac{K^2G^2}{\mu^2\gamma^2}(\log(K(t+1)+1) - \log(Kt+1))\right) \\ & + O(G^2\frac{(1-\gamma)^2}{\mu^2\gamma^2}) + O\left(G^2\frac{1-\gamma}{\mu\gamma^2}(\log(K(t+1)) - \log(Kt))\right) + \frac{2\delta K}{\mu}\end{aligned}$$

Sum from $t = 0$ to $T - 1$ and use the non-expensiveness of the clip operation to yield

$$\begin{aligned}
&(KT+1)\|f^T - f^*\|_\alpha^2 \\
\leq\ &(k_0+1)\|f^0 - f^*\|_\alpha^2 + O\left(\frac{K^2G^2\log(KT+1)}{\mu^2\gamma^2}\right) + O(\frac{(1-\gamma)G^2\log(TK)}{\mu\gamma^2}) \\
&+ O(TG^2\frac{(1-\gamma)^2}{\mu^2\gamma^2}) + O(\frac{G^2TK\omega_{TV}}{\mu^2\gamma^2}),
\end{aligned}$$

and hence

$$\|f^T - f^*\|_\alpha^2 = O(\frac{\|f^0 - f^*\|_\alpha^2}{KT} + \frac{KG^2log(KT)}{T\mu^2\gamma^2} + \frac{(1-\gamma)^2G^2}{K\mu^2\gamma^2} + \frac{G^2\omega_{TV}}{\mu^2\gamma^2}). \tag{35}$$

□

E Proof of Theorem 4.3

In this section, we present the convergence analysis of the setting where the distributions over the input space are i.i.d.. Under this setting, we show that FFGB converges to the global minimizer in a sublinear rate.

Theorem E.1 (Theorem 4.3 restated.)**.** *Let f^0 be the initializer function. Suppose that Assumption 4.2 holds, and suppose the weak learning oracle $\mathcal{Q}_\alpha^2$ satisfies* (5) *with constant γ. Using the step size $\eta^{k,t} = \frac{2}{\mu(tK+k+1)}$, the output of* FFGB *satisfies*

$$\|f^T - f^*\|_\alpha^2 = O\left(\frac{\|f^0 - f^*\|_\alpha^2}{KT} + \frac{KG^2\log(KT)}{T\gamma^2\mu^2} + \frac{(1-\gamma)G^2}{K\mu^2\gamma^2} + \frac{(1-\gamma)G^2\log(KT)}{KT\mu\gamma^2}\right).$$

Proof. Since we are considering the setting where $\alpha = \alpha_i$, we ignore the subscript i and simply write $\|\cdot\|_\alpha$ and $\langle\cdot,\cdot\rangle_\alpha$ for the norm and the inner product in $\mathcal{L}^2(\alpha)$.

For simplicity, we denote $\hat{h}_i^{k,t} = h_i^{k,t} + \mu g_i^{k,t}$.

We define two hypothetical global average sequences $\bar{g}^{k,t} = \frac{1}{N}\sum_{i=1}^N g_i^{k,t}$ and $\bar{h}^{k,t} = \frac{1}{N}\sum_{i=1}^N \hat{h}_i^{k,t}$. In particular, we have $\bar{g}^{1,t} = f^t$. From the update rule in line (5) of Algorithm 2, we write

$$\|\bar{g}^{k+1,t} - f^*\|_\alpha^2 = \|\bar{g}^{k,t} - f^*\|_\alpha^2 + (\eta^{k,t})^2\|\bar{h}^{k,t}\|_\alpha^2 - 2\eta^{k,t}\langle\bar{g}^{k,t} - f^*, \bar{h}^{k,t}\rangle_\alpha \tag{36}$$

The last term of (36) can be split as

$$-2\langle\bar{g}^{k,t} - f^*, \bar{h}^{k,t}\rangle_\alpha = -\frac{2}{N}\sum_{i=1}^N\langle\bar{g}^{k,t} - f^*, \hat{h}_i^{k,t}\rangle_\alpha \tag{37}$$

$$\begin{aligned}
&= -\frac{2}{N}\sum_{i=1}^N\langle\bar{g}^{k,t} - g_i^{k,t}, \hat{h}_i^{k,t}\rangle_\alpha + \langle g_i^{k,t} - f^*, \hat{h}_i^{k,t}\rangle_\alpha \\
&= \frac{2}{N}\sum_{i=1}^N\langle g_i^{k,t} - \bar{g}^{k,t}, \hat{h}_i^{k,t}\rangle_\alpha + \langle f^* - g_i^{k,t}, \nabla\mathcal{F}_i[g_i^{k,t}]\rangle_\alpha + \langle f^* - g_i^{k,t}, \hat{h}_i^{k,t} - \nabla\mathcal{F}_i[g_i^{k,t}]\rangle_\alpha.
\end{aligned} \tag{38}$$

The second term of (38) can be bounded using the μ-strong convexity of $\mathcal{F}_i$

$$\frac{2}{N}\sum_{i=1}^N\langle f^* - g_i^{k,t}, \nabla\mathcal{F}_i[g_i^{k,t}]\rangle_\alpha \leq\ -\frac{2}{N}\sum_{i=1}^N\mathcal{F}_i[g_i^{k,t}] - \mathcal{F}_i[f^*] - \frac{2}{N}\sum_{i=1}^N\frac{\mu}{2}\|f^* - g_i^{k,t}\|_\alpha^2.$$

Note that using the optimality of f^* we have

$$\sum_{i=1}^N\mathcal{F}_i[g_i^{k,t}] - \mathcal{F}_i[f^*] = \sum_{i=1}^N\mathcal{F}_i[g_i^{k,t}] - \mathcal{F}_i[\bar{g}^{k,t}] + \mathcal{F}_i[\bar{g}^{k,t}] - \mathcal{F}_i[f^*] \leq \sum_{i=1}^N\mathcal{F}_i[g_i^{k,t}] - \mathcal{F}_i[\bar{g}^{k,t}]$$

and that by recalling that $\bar{g}^{k,t} = \frac{1}{N}\sum_{i=1}^N g_i^{k,t}$ and using the Cauchy–Schwarz inequality we have

$$\frac{1}{N}\sum_{i=1}^N \|f^* - g_i^{k,t}\|_\alpha^2 \geq \|f^* - \bar{g}^{k,t}\|_\alpha^2.$$

Therefore, we can bound

$$\begin{aligned}
&\frac{2}{N}\sum_{i=1}^N \langle f^* - g_i^{k,t}, \nabla\mathcal{F}_i[g_i^{k,t}]\rangle_\alpha \\
\leq\ & -\frac{2}{N}\sum_{i=1}^N \mathcal{F}_i[g_i^{k,t}] - \mathcal{F}_i[\bar{g}^{k,t}] - \frac{\mu}{2}\|f^* - \bar{g}^{k,t}\|_\alpha^2 - \frac{1}{N}\sum_{i=1}^N \frac{\mu}{2}\|f^* - g_i^{k,t}\|_\alpha^2 \\
\leq\ & -\frac{2}{N}\sum_{i=1}^N \langle \nabla\mathcal{F}_i[\bar{g}^{k,t}], g_i^{k,t} - \bar{g}^{k,t}\rangle_\alpha - \frac{\mu}{2}\|f^* - \bar{g}^{k,t}\|_\alpha^2 - \frac{1}{N}\sum_{i=1}^N \frac{\mu}{2}\|f^* - g_i^{k,t}\|_\alpha^2 \\
\leq\ & \frac{1}{N}\sum_{i=1}^N \eta^{k,t}\|\nabla\mathcal{F}_i[\bar{g}^{k,t}]\|_\alpha^2 + \frac{1}{\eta^{k,t}}\|g_i^{k,t} - \bar{g}^{k,t}\|_\alpha^2 - \frac{\mu}{2}\|f^* - \bar{g}^{k,t}\|_\alpha^2 - \frac{1}{N}\sum_{i=1}^N \frac{\mu}{2}\|f^* - g_i^{k,t}\|_\alpha^2,
\end{aligned}$$

where we use Young's inequality in the last inequality.

Besides, recall that $\|\nabla\mathcal{R}_i[g]\|_\alpha \leq G$ from Assumption 4.2. Together with the property of the oracle, we have

$$\|\Delta_i^{k,t}\|_\alpha \leq (1-\gamma)\left(\|\Delta_i^{k-1,t}\|_\alpha + G\right) \text{ and } \|\Delta_i^{0,t}\|_\alpha = 0 \Rightarrow \forall k, \|\Delta_i^{k,t}\|_\alpha \leq \frac{1-\gamma}{\gamma}G. \tag{39}$$

Consequently, we also have

$$\|\bar{h}^k\|_\alpha \leq \frac{1}{N}\sum_{i=1}^N \|h_i^{k,t}\|_\alpha = \frac{1}{N}\sum_{i=1}^N \|\nabla\mathcal{R}_i[g_i^{k,t}] + \Delta_i^{k-1}\|_\alpha \leq \frac{2-\gamma}{\gamma}G.$$

From line 5 of Algorithm 2, we have $g_i^{k+1,t} = (1-\mu\eta^{k,t})g_i^{k,t} + \eta^{k,t}h_i^{k,t}$ and therefore

$$\|g_i^{k+1,t}\|_\alpha - \frac{2G}{\gamma\mu} \leq (1-\mu\eta^{k,t})\left(\|g_i^{k+1,t}\|_\alpha - \frac{2G}{\gamma\mu}\right), \tag{40}$$

where we use $\|h_i^{k,t}\|_\alpha \leq 2G/\gamma$. Therefore, if we have initially $\|f^t\|_\alpha \leq \frac{2G}{\gamma\mu}$, we always have $\|g_i^{k,t}\|_\alpha \leq \frac{2G}{\gamma\mu}$ (hence so is f^{t+1} as it is the global average $\bar{g}^{K+1,t}$). Further, together with $\|h_i^{k,t}\|_\alpha \leq 2G/\gamma$, we have $\|\hat{h}_i^{k,t}\|_\alpha \leq 4G/\gamma$.

Additionally, $\frac{1}{N}\sum_{i=1}^N \|g_i^{k,t} - \bar{g}^{k,t}\|_\alpha^2$ can be bounded by (we use $E[(X-E[X])^2] \leq E[X^2]$)

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^N \|g_i^{k,t} - \bar{g}^{k,t}\|_\alpha^2 &= \frac{1}{N}\sum_{i=1}^N \|g_i^{k,t} - g_i^{1,t} + g_i^{1,t} - \bar{g}^{k,t}\|_\alpha^2 \\
&\leq \frac{1}{N}\sum_{i=1}^N \|g_i^{k,t} - g_i^{1,t}\|_\alpha^2 \leq \sum_{\kappa=1}^k 16(\eta^{\kappa,t})^2G^2/\gamma^2 \leq 64(\eta^{k,t})^2K^2G^2/\gamma^2,
\end{aligned}$$

where we use $\eta^{\kappa,t} \leq 2\eta^{k,t}$ for any $t \geq 0$ and $1 \leq \kappa \leq k$. Therefore we can bound the first term of (38) by

$$\frac{2}{N}\sum_{i=1}^N \langle g_i^{k,t} - \bar{g}^{k,t}, \hat{h}_i^{k,t}\rangle_\alpha \leq \frac{1}{N}\sum_{i=1}^N \frac{1}{\eta^{k,t}}\|g_i^{k,t} - \bar{g}^{k,t}\|_\alpha^2 + \frac{\eta^{k,t}}{N}\sum_{i=1}^N \|\hat{h}_i^{k,t}\|_\alpha^2 \leq (64K^2+16)\eta^{k,t}G^2/\gamma^2$$

Plug in the above results into (36) to yield

$$\begin{aligned}
\|\bar{g}^{k+1,t} - f^*\|_\alpha^2 \leq (1 - \frac{\mu\eta^{k,t}}{2})\|\bar{g}^{k,t} - f^*\|_\alpha^2 + O\left((\eta^{k,t})^2K^2G^2/\gamma^2\right) \\
+ \frac{2\eta^{k,t}}{N}\sum_{i=1}^N \langle f^* - g_i^{k,t}, \hat{h}_i^{k,t} - \nabla\mathcal{F}_i[g_i^{k,t}]\rangle_\alpha - \frac{\eta^{k,t}}{N}\sum_{i=1}^N \frac{\mu}{2}\|f^* - g_i^{k,t}\|^2.
\end{aligned}$$

Recall that $\eta^{k,t} = \frac{2}{\mu(tK+k+1)}$ and multiply both sides by $(tK+k+1)$

$$\begin{aligned}(tK+k+1)\|\bar{g}^{k+1,t}-f^*\|_\alpha^2 &\le (tK+k)\|\bar{g}^{k,t}-f^*\|_\alpha^2 + O(\frac{K^2G^2}{\gamma^2\mu^2(tK+k+1)})\\ &\quad + \frac{4}{\mu N}\sum_{i=1}^N\langle f^*-g_i^{k,t}, \hat{h}_i^{k,t}-\nabla\mathcal{F}_i[g_i^{k,t}]\rangle_\alpha - \frac{2}{\mu N}\sum_{i=1}^N\frac{\mu}{2}\|f^*-g_i^{k,t}\|_\alpha^2.\end{aligned}$$

Sum from $k=1$ to K

$$\begin{aligned}&(tK+K+1)\|\bar{g}^{K+1,t}-f^*\|_\alpha^2\\ &\le (tK+1)\|\bar{g}^{1,t}-f^*\|_\alpha^2 + O(\frac{K^2G^2}{\gamma^2\mu^2})\left(\log(tK+K)-\log(tK+1)\right)\\ &\quad + \frac{4}{\mu N}\sum_{i=1}^N\sum_{k=1}^K\langle f^*-g_i^{k,t}, \hat{h}_i^{k,t}-\nabla\mathcal{F}_i[g_i^{k,t}]\rangle_\alpha - \frac{2}{\mu N}\sum_{i=1}^N\sum_{k=1}^K\frac{\mu}{2}\|f^*-g_i^{k,t}\|_\alpha^2.\end{aligned} \tag{41}$$

For the first term of the second line above, the following equality holds for the same reason as (34)

$$\begin{aligned}&\sum_{k=1}^K\langle f^*-g_i^{k,t}, \hat{h}_i^{k,t}-\nabla\mathcal{F}_i[g_i^{k,t}]\rangle_\alpha = \langle f^*-g_i^{K,t}, -\Delta_i^K\rangle_\alpha + \sum_{k=1}^{K-1}\langle\eta_t^k h_i^{k,t}, \Delta_i^k\rangle_\alpha\\ &\le \frac{\mu}{2}\|f^*-g_i^{K,t}\|^2 + \frac{1}{2\mu}(\frac{1-\gamma}{\gamma})^2G^2 + \frac{(1-\gamma)(2-\gamma)}{\gamma^2}G^2\sum_{k=1}^{K-1}\eta^{k,t}\\ &\le \frac{\mu}{2}\|f^*-g_i^{K,t}\|^2 + \frac{1}{2\mu}(\frac{1-\gamma}{\gamma})^2G^2 + \frac{(1-\gamma)(2-\gamma)}{\gamma^2}G^2(\log(tK+K)-\log(tK+2)).\end{aligned} \tag{42}$$

Using this result, we obtain (we cancel $\frac{\mu}{2}\|f^*-g_i^{K,t}\|^2$ with the last term of (41))

$$\begin{aligned}&(K(t+1)+1)\,\|f^{t+1}-f^*\|_\alpha^2\\ &\le (Kt+1)\|f^t-f^*\|_\alpha^2 + O(\frac{K^2G^2}{\gamma^2\mu^2})(\log(K(t+1)+1)-\log(Kt+1))\\ &\quad + \frac{4}{\mu}(\frac{1}{2\mu}(\frac{1-\gamma}{\gamma})^2G^2 + \frac{(1-\gamma)(2-\gamma)}{\gamma^2}G^2(\log(K(t+1)-1)-\log(Kt+1)))\end{aligned}$$

Sum from $t=0$ to $T-1$ to yield

$$\begin{aligned}&(KT+1)\|f^T-f^*\|_\alpha^2\\ &\le \|f^0-f^*\|_\alpha^2 + O(\frac{K^2G^2\log(KT)}{\gamma^2\mu^2}) + O(\frac{(1-\gamma)TG^2}{\mu^2\gamma^2}) + O(\frac{(1-\gamma)G^2\log(KT)}{\mu\gamma^2}),\end{aligned}$$

and hence

$$\begin{aligned}&\|f^T-f^*\|_\alpha^2\\ &\le O(\frac{\|f^0-f^*\|_\alpha^2}{KT}) + O(\frac{KG^2\log(KT)}{T\gamma^2\mu^2}) + O(\frac{(1-\gamma)G^2}{K\mu^2\gamma^2}) + O(\frac{(1-\gamma)G^2\log(KT)}{KT\mu\gamma^2}).\end{aligned}$$

□

F Proof of Theorem 4.4

To show that FFGB converges to a neighborhood of f^*, the global minimizer of the federated functional minimization problem (9), in the regression loss special case, we need to following lemma that characterizes the difference between the inner products in $\mathcal{L}^2$ spaces with different weights.

Lemma F.1. *For two functions $f, g \in \mathcal{L}^\infty(\alpha)$ with $\|f\|_{\text{lip}} < \infty$ and $\|g\|_{\text{lip}} < \infty$, denote $\xi := \|f\|_{\text{lip}}\|g\|_{\alpha,\infty} + \|g\|_{\text{lip}}\|f\|_{\alpha,\infty}$. Then*

$$|\langle f, g\rangle_{\alpha_i} - \langle f, g\rangle_\alpha| \leq \xi \mathrm{W}_1(\alpha, \alpha_i). \tag{43}$$

Proof. Recall (19) where the Wasserstein-1 distance between two discrete distribution μ and ν can be written as

$$W_1(\mu, \nu) = \min_{\Pi \geq 0} \int_{\mathcal{X}^2} \|x - y\| d\Pi(x, y), \quad s.t. \sharp_1 \Pi = \mu, \sharp_2 \Pi = \nu.$$

Note that the constraint of the above problem implies that $\mathrm{supp}(\Pi) \subseteq \mathrm{supp}(\mu) \times \mathrm{supp}(\nu)$, otherwise Π must be infeasible. The above minimization problem is equivalent to

$$W_1(\mu, \nu) = \min_{\Pi \geq 0} \max_{\phi, \psi} \int_{\mathcal{X}^2} \|x - y\| d\Pi(x, y) + \int_{\mathcal{X}} \phi(x) d\left(\mu - \sharp_1 \Pi\right)(x) + \int_{\mathcal{X}} \psi(y) d\left(\nu - \sharp_2 \Pi\right)(y).$$

Change the order of min-max to max-min (due to convexity) and rearrange terms:

$$W_1(\mu, \nu) = \max_{\phi, \psi} \left\{ \min_{\Pi \geq 0} \int_{\mathcal{X}^2} \|x - y\| - \phi(x) - \psi(y) d\Pi(x, y) \right\} + \int_{\mathcal{X}} \phi(x) d\mu(x) + \int_{\mathcal{X}} \psi(y) d\nu(y).$$

Therefore, we must have that for $(x, y) \in \mathrm{supp}(\Pi) \subseteq \mathrm{supp}(\mu) \times \mathrm{supp}(\nu)$, $\phi(x) + \psi(y) \leq \|x - y\|$ which leads to

$$\begin{aligned} W_1(\mu, \nu) = \max_{\phi, \psi} & \int_{\mathcal{X}} \phi(x) d\mu(x) + \int_{\mathcal{X}} \psi(y) d\nu(y), \\ s.t. \ & \phi(x) + \psi(y) \leq \|x - y\|, \forall (x, y) \in \mathrm{supp}(\mu) \times \mathrm{supp}(\nu). \end{aligned}$$

Now, recall that every local distribution α_i is described by a set of data feature points: $\alpha_i = \frac{1}{M}\sum_{j=1}^M \delta_{x_{i,j}}$, where δ_x is the Dirac distribution; and the global distribution α is described by the union of all these points: $\alpha = \frac{1}{MN}\sum_{i,j=1}^{N,M} \delta_{x_{i,j}}$. Clearly we have $\mathrm{supp}(\alpha_i) \subseteq \mathrm{supp}(\alpha)$. Using Proposition 6.1. of [Peyré et al., 2019] with $\mathcal{X} = \mathrm{supp}(\alpha)$, the above bi-variable problem is equivalent to the single-variable problem

$$\begin{aligned} W_1(\mu, \nu) = \max_{\phi} & \int_{\mathcal{X}} \phi(x) d\mu(x) - \int_{\mathcal{X}} \phi(y) d\nu(y), \\ s.t. \ & |\phi(x) - \phi(y)| \leq \|x - y\|, \forall (x, y) \in \mathrm{supp}(\mu) \times \mathrm{supp}(\nu). \end{aligned}$$

Recall that in (43), $\phi(x) = f(x)g(x)$ and $\xi = \|g\|_{\text{lip}}\|f\|_\infty + \|f\|_{\text{lip}}\|g\|_\infty$. One can check that

$$\begin{aligned} &\|\phi(x)/\xi - \phi(y)/\xi\| = \|f(x)\left(g(x) - g(y)\right) + g(y)\left(f(x) - f(y)\right)\|/\xi \\ \leq \ &\left(\|g\|_{\text{lip}}\|f\|_\infty + \|f\|_{\text{lip}}\|g\|_\infty\right)/\xi \|x - y\| = \|x - y\|. \end{aligned}$$

Therefore

$$W_1(\mu, \nu) \geq |\int_{\mathcal{X}} f(x)g(x)/\xi d\alpha(x) - \int_{\mathcal{X}} f(x)g(x)/\xi d\alpha_i(y)|,$$

which is equivalent to (43). □

Recall that we assume the input to $\mathcal{Q}_\alpha^{\text{lip}}$ to be of the form $-\phi + g$, where g is some function that is explicitly available (usually, the variable function) and recall that $\mathcal{Q}_\alpha^{\text{lip}}$ outputs $-h_\theta + g$ where h_θ is a neural network with parameter θ. In other words, $\mathcal{Q}_\alpha^{\text{lip}}$ only approximates ϕ in the input with h_θ and leaves the known part, g, untouched. Therefore, the client procedure of FFGB in Algorithm 2 can be equivalently written as Algorithm 3. In the following, we will analyze the convergence of Algorithm 3.

Algorithm 3 CLIENT procedure of Federated Functional Gradient Boosting for regression loss

```
procedure CLIENT($i$, $t$, $f$)
    $\Delta_i^0 = 0$, $g_i^{1,t} = f^t$ ;
    for $k \leftarrow 1$ to $K$ do
        $h_i^k := \mathcal{Q}_{\alpha_i}^{\text{lip}}(\Delta_i^{k-1} - u_i)$
        $g_i^{k+1,t} := g_i^{k,t} - \eta^{k,t}(g_i^{k,t} - h_i^k)$
        $\Delta_i^k := \Delta_i^{k-1} - u_i + h_i^k$
    return $g_i^{K,t}$.
```

Theorem F.1 (Theorem 4.4 restated)**.** *Consider the special case of the federated functional minimization problem with square loss. Assume that the optimal solution f^* is L-Lipschitz continuous. Let f^0 be the initializer. We define a proxy of the heterogeneity among the local input distributions α's in the Wasserstein-1 distance as*

$$\omega_{W_1} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{W}_1(\alpha, \alpha_i) \tag{44}$$

Suppose that Assumption 4.3 holds and define $G^2 = \frac{2L^2}{N^2}\sum_{i,s=1}^{N} \mathrm{W}_2^2(\alpha_s, \alpha_i) + 2B^2$. Moreover, suppose the weak learning oracle $\mathcal{Q}_\alpha^{\mathrm{lip}}$ satisfies (14) *and* (15) *with constant γ. Using the step sizes $\eta^{k,t} = 4/(\mu(tK+k+1))$, the output of* FFGB *satisfies*

$$\|f^T - f^*\|_\alpha^2 = O\left(\frac{K\left(L(LD+B)\omega_{W_1} + G^2\right)log(KT)}{T\mu^2\gamma^2} + \frac{\|f^0 - f^*\|^2}{KT} + \frac{(1-\gamma)^2B^2}{\mu^2\gamma^2K} + \frac{L(LD+B)\omega_{W_1}}{\gamma^2\mu}\right).$$

Proof. For a fixed communication round t, we define a hypothetical sequence $\bar{g}^{k,t} = \frac{1}{N}\sum_{i=1}^{N} g_i^{k,t}$. We also define $\bar{h}^k = \frac{1}{N}\sum_{i=1}^{N} h_i^k$. Note that $\bar{g}^{1,t} = f^t$.

From the Lipschitz extension construction, we have $\|u_i\|_{\mathrm{lip}} \le L$. Additionally, using the assumptions (15) on the oracle, the residual is inductively proved to be $\frac{(1-\gamma)}{\gamma}L$-Lipschitz continuous as follows. For the base case, note that $\|\Delta_i^0\|_{\mathrm{lip}} \equiv 0$. Now, assume that for some $k \ge 1$, we have $\|\Delta_i^{k-1}\|_{\mathrm{lip}} \le \frac{(1-\gamma)}{\gamma}L$. Then

$$\|\Delta_i^k\|_{\mathrm{lip}} = \|h_i^k - (u_i - \Delta_i^{k-1})\|_{\mathrm{lip}} \le (1-\gamma)\|u_i - \Delta_i^{k-1}\|_{\mathrm{lip}} \le (1-\gamma)(\|\Delta_i^{k-1}\|_{\mathrm{lip}} + L) \le \tfrac{(1-\gamma)}{\gamma}L.$$

Therefore, the query to the weak learning oracle is also Lipschitz continuous: $\|\Delta_i^{k-1} - u_i\|_{\mathrm{lip}} \le L/\gamma$, and so is the output, $\|h_i^k\|_{\mathrm{lip}} \le L/\gamma$. Now, the update rule of $g_i^{k,t}$ (line 5), and the boundedness of $\|h_i^k\|_{\mathrm{lip}}$ imply the boundedness of $\|g_i^{k,t}\|_{\mathrm{lip}}$ for sufficiently small $\eta^{k,t}$:

$$\begin{aligned}\|g_i^{k+1,t}\|_{\mathrm{lip}} = \|(1-\eta^{k,t})g_i^{k,t} + \eta^{k,t}h_i^k\|_{\mathrm{lip}} \le (1-\eta^{k,t})\|g_i^{k,t}\|_{\mathrm{lip}} + L/\gamma\cdot\eta^{k,t}\\ \Rightarrow \|g_i^{k,t}\|_{\mathrm{lip}} \le L/\gamma \text{ (via induction using} \|g_i^{1,t}\|_{\mathrm{lip}} \le L/\gamma).\end{aligned}$$

Lemma F.2. *The residual Δ_i^k and the output h_i^k of the oracle $\mathcal{Q}_{\alpha_i}^{\mathrm{lip}}$ are bounded under the $\mathcal{L}^\infty(\alpha_i)$ norm:*

$$\|\Delta_i^k\|_{\alpha_i,\infty} \le \frac{(1-\gamma)B}{\gamma} \text{ and } \|h_i^k\|_{\alpha_i,\infty} \le B/\gamma.$$

Proof. From property (15) of the weak leaner oracle $\mathcal{Q}_{\alpha_i}^{\mathrm{lip}}$, we have

$$\|\Delta_i^k\|_{\alpha_i,\infty} = \|\Delta_i^{k-1} - u_i + h_i^k\|_{\alpha_i,\infty} \le (1-\gamma)\|\Delta_i^{k-1} - u_i\|_{\alpha_i,\infty} \le (1-\gamma)\|\Delta_i^{k-1}\|_{\alpha_i,\infty} + (1-\gamma)B, \tag{45}$$

where the second inequality uses the boundedness of $y_{i,j} = f_i^*(x_{i,j})$ in Assumption 4.3. We hence have

$$\|\Delta_i^k\|_{\alpha_i,\infty} - \frac{(1-\gamma)B}{\gamma} \le (1-\gamma)\left(\|\Delta_i^{k-1}\|_{\alpha_i,,\infty} - \frac{(1-\gamma)B}{\gamma}\right) \Rightarrow \|\Delta_i^k\|_{\alpha_i,\infty} \le \frac{(1-\gamma)B}{\gamma}. \tag{46}$$

The boundedness of $\|h_i^k\|_{\alpha_i,\infty}$ can be obtained from the above inequality: $\|h_i^k\|_{\alpha_i,\infty} \le \|u_i\|_{\alpha_i,\infty} + \|\Delta_i^k\|_{\alpha_i,\infty} \le B/\gamma$. □

Lemma F.3. *The local variable function $g_i^{k,t}$ is bounded under the $\mathcal{L}^\infty(\alpha_i)$ norm: $\|g_i^{k,t}\|_{\alpha_i,\infty} \le B/\gamma$.*

Proof. Using the update rule in line 5 of Algorithm 3, we have

$$\begin{aligned}&\|g_i^{k+1,t}\|_{\alpha_i,\infty} = \|(1-\eta^{k,t})g_i^{k,t} + \eta^{k,t}h_i^k\|_{\alpha_i,\infty}\\ &\le (1-\eta^{k,t})\|g_i^{k,t}\|_{\alpha_i,\infty} + \eta^{k,t}\|h_i^k\|_{\alpha_i,\infty} \le (1-\eta^{k,t})\|g_i^{k,t}\|_{\alpha_i,\infty} + \eta^{k,t}B/\gamma.\end{aligned}$$

Inductively, we have the boundedness of $\|g_i^{k+1,t}\|_{\alpha_i,\infty}$

$$\|g_i^{k+1,t}\|_{\alpha_i,\infty} - B/\gamma \le (1-\eta^{k,t})\left(\|g_i^{k,t}\|_{\alpha_i,\infty} - B/\gamma\right) \Rightarrow \|g_i^{k,t}\|_{\alpha_i,\infty} \le B/\gamma. \tag{47}$$

□

Lemma F.4. *The local variable function $g_i^{k,t}$, the global average function $\bar{g}^{k,t}$, and the output of the oracle $\mathcal{Q}_{\alpha_i}^{\mathrm{lip}}$ are $(LD+B)/\gamma$-bounded under the $\mathcal{L}^\infty(\alpha)$ norm.*

Proof. From Lemmas F.3 and F.2, $g_i^{k,t}$, $\bar{g}^{k,t}$ and $h_i^{k,t}$ are B/γ on the support of α_i. Using Assumption 4.3 together with the L/γ-Lipschitz continuity of $g_i^{k,t}$, $\bar{g}^{k,t}$ and $h_i^{k,t}$, we have the results. □

While the above lemma implies the boundedness of $\bar{g}^{k,t}$ under the $\mathcal{L}^2(\alpha)$ norm, we can tighten the analysis with the following lemma. Important, the following result does not depend on the constant D in Assumption 4.3.

Lemma F.5. *The hypothetical global sequences $\bar{g}^{k,t}$ and $\bar{h}^{k,t}$ are bounded under the local norm $\|\cdot\|_{\alpha_s}$: Denote $G_s^2 = \frac{2L^2}{N}\sum_{i=1}^N \mathrm{W}_2^2(\alpha_s,\alpha_i) + 2B^2$. We have that $\|\bar{g}^{k,t}\|_{\alpha_s} \le G_s^2/\gamma^2$ and $\|\bar{h}^{k,t}\|_{\alpha_s} \le G_s^2/\gamma^2$, where $\mathrm{W}_2(\alpha_s,\alpha_i)$ is the Wasserstein-2 distance between measures α_i and α_s. Consequently, we have $\bar{g}^{k,t}$ and $\bar{h}^{k,t}$ are G-bounded under the $\mathcal{L}^2(\alpha)$ norm, where we further denote $G^2 = \frac{1}{N}\sum_{s=1}^N G_s^2$.*

Proof. Let $\Pi^{s,i} \in \mathbb{R}_+^{M\times M}$ be the Wasserstein-2 optimal transport plan (matrix) between α_s and α_i. The entry $\Pi_{j_1,j_2}^{s,i}$ denotes the portion of mass that should be transported from $x_{s,j_1} \in \mathrm{supp}(\alpha_s)$ to $x_{i,j_2} \in \mathrm{supp}(\alpha_i)$. Note that in α_i and α_s, the entries x_{s,j_1} and x_{i,j_2} have uniform weight $1/M$. As a transport plan, any row or column of $\Pi^{s,i}$ sums up to $1/M$. We now show that $\|\bar{g}^{k,t}\|_{\alpha_s}$ is bounded using the Lipschitz continuity of $g_i^{k,t}$.

$$\|\bar{g}^{k,t}\|_{\alpha_s}^2 = \frac{1}{M}\sum_{j=1}^M \|\frac{1}{N}\sum_{i=1}^N g_i^{k,t}(x_{s,j})\|^2 \le \frac{1}{M}\sum_{j=1}^M \frac{1}{N}\sum_{i=1}^N \|g_i^{k,t}(x_{s,j})\|^2 = \frac{1}{N}\sum_{i=1}^N \frac{1}{M}\sum_{j=1}^M \|g_i^{k,t}(x_{s,j})\|^2. \tag{48}$$

We analyze the summand as follows.

$$\begin{aligned}
\frac{1}{M}\sum_{j=1}^M \|g_i^{k,t}(x_{s,j})\|^2 =& \sum_{j_1=1}^M\sum_{j_2=1}^M \Pi_{j_1,j_2}^{s,i}\|g_i^{k,t}(x_{s,j_1}) - g_i^{k,t}(x_{i,j_2}) + g_i^{k,t}(x_{i,j_2})\|^2 \\
\le& \sum_{j_1=1}^M\sum_{j_2=1}^M \Pi_{j_1,j_2}^{s,i}\left(2\|g_i^{k,t}(x_{s,j_1}) - g_i^{k,t}(x_{i,j_2})\|^2 + 2\|g_i^{k,t}(x_{i,j_2})\|^2\right) \\
\le& \sum_{j_1=1}^M\sum_{j_2=1}^M \Pi_{j_1,j_2}^{s,i}\left(2L^2/\gamma^2\cdot\|x_{s,j_1} - x_{i,j_2}\|^2 + 2\|g_i^{k,t}(x_{i,j_2})\|^2\right) \\
=& \, 2L^2/\gamma^2\cdot \mathrm{W}_2^2(\alpha_s,\alpha_i) + \frac{2}{M}\sum_{j_2=1}^M \|g_i^{k,t}(x_{i,j_2})\|^2 \\
=& \, 2L^2/\gamma^2\cdot \mathrm{W}_2^2(\alpha_s,\alpha_i) + 2\|g_i^{k,t}\|_{\alpha_i}^2,
\end{aligned}$$

where we used the definition of the Wasserstein-2 distance. Therefore, (48) can be bounded by

$$\|\bar{g}^{k,t}\|_{\alpha_s}^2 \le \frac{1}{N}\sum_{i=1}^N 2L^2/\gamma^2\cdot \mathrm{W}_2^2(\alpha_s,\alpha_i) + 2\|g_i^{k,t}\|_{\alpha_i}^2 \le \frac{2L^2}{N\gamma^2}\sum_{i=1}^N \mathrm{W}_2^2(\alpha_s,\alpha_i) + 2B^2/\gamma^2. \tag{49}$$

Following the similar proof above, we have the same bound for $\|\bar{h}^{k,t}\|_{\alpha_s}$ as $h_i^{k,t}$ is also B/γ- bounded and L/γ-Lipschitz continuous:

$$\|\bar{h}^{k,t}\|_{\alpha_s}^2 \le \frac{2L^2}{N\gamma^2}\sum_{i=1}^N \mathrm{W}_2^2(\alpha_s,\alpha_i) + 2B^2/\gamma^2. \tag{50}$$

□

We now present the convergence analysis of Algorithm 3. From the update rule in line 5 of Algorithm 3, we write

$$\|\bar{g}^{k+1,t} - f^*\|_\alpha^2 = \|\bar{g}^{k,t} - f^*\|_\alpha^2 + (\eta^{k,t})^2\|\bar{g}^{k,t} - \bar{h}^k\|_\alpha^2 - 2\eta^{k,t}\langle \bar{g}^{k,t} - f^*, \bar{g}^{k,t} - \bar{h}^k\rangle_\alpha. \tag{51}$$

To bound the second term, note that

$$\|\bar{g}^{k,t} - \bar{h}^k\|_\alpha^2 \le \frac{1}{N}\sum_{i=1}^N \|g_i^{k,t} - h_i^k\|_\alpha^2. \tag{52}$$

For each individual term on the R.H.S. of the above inequality, we have

$$\begin{aligned}\|g_i^{k,t} - h_i^k\|_\alpha^2 &= \left(\|g_i^{k,t} - h_i^k\|_\alpha^2 - \|g_i^{k,t} - h_i^k\|_{\alpha_i}^2\right) + \|g_i^{k,t} - h_i^k\|_{\alpha_i}^2 \\ \le O(L(LD+B)/\gamma^2 &\cdot \mathrm{W}_1(\alpha,\alpha_i)) + O(B^2/\gamma^2),\end{aligned}$$

where we use the variational formulation (43) of the Wasserstein-1 distance as well as the Lipschitz continuity and boundedness of $g_i^{k,t}$ and $h_i^{k,t}$ under the $\mathcal{L}^2(\alpha)$ norm. Therefore the second term is bounded by

$$\|\bar{g}^{k+1,t} - f^*\|_\alpha^2 \le O(L(LD+B)/\gamma^2 \cdot \omega) + O(B^2/\gamma^2), \quad \omega = \frac{1}{N}\sum_{i=1}^N \mathrm{W}_1(\alpha,\alpha_i). \tag{53}$$

The third term of (51) can be split as

$$\begin{aligned}& -2\langle \bar{g}^{k,t} - f^*, \bar{g}^{k,t} - \bar{h}^k\rangle_\alpha \qquad (54)\\ =& -\frac{2}{N}\sum_{i=1}^N \langle \bar{g}^{k,t} - f^*, g_i^{k,t} - h_i^k\rangle_{\alpha_i} + \left(\langle \bar{g}^{k,t} - f^*, g_i^{k,t} - h_i^k\rangle_\alpha - \langle \bar{g}^{k,t} - f^*, g_i^{k,t} - h_i^k\rangle_{\alpha_i}\right)\\ =& \frac{2}{N}\sum_{i=1}^N \langle g_i^{k,t} - \bar{g}^{k,t}, g_i^{k,t} - h_i^k\rangle_{\alpha_i} + \langle f^* - g_i^{k,t}, g_i^{k,t} - u_i\rangle_{\alpha_i} + \langle f^* - g_i^{k,t}, u_i - h_i^{k,t}\rangle_{\alpha_i}\\ &\quad + \left(\langle \bar{g}^{k,t} - f^*, g_i^{k,t} - h_i^k\rangle_\alpha - \langle \bar{g}^{k,t} - f^*, g_i^{k,t} - h_i^k\rangle_{\alpha_i}\right). \qquad (55)\end{aligned}$$

The last term of R.H.S. of the above equality can be bounded using the Lipschitz continuity and the boundedness of $(\bar{g}^{k,t} - f^*)$ and $(g_i^{k,t} - h_i^k)$ and the variational formulation of W_1 (see (43)):

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^N \left(\langle \bar{g}^{k,t} - f^*, g_i^{k,t} - h_i^k\rangle_\alpha - \langle \bar{g}^{k,t} - f^*, g_i^{k,t} - h_i^k\rangle_{\alpha_i}\right)\\ &= O(L(LD+B)/\gamma^2 \cdot \omega), \quad \omega = \frac{1}{N}\sum_{i=1}^N \mathrm{W}_1(\alpha,\alpha_i).\end{aligned}$$

The first term of of the R.H.S. of (55) can be bounded by

$$\begin{aligned}&\frac{2}{N}\sum_{i=1}^N \langle g_i^{k,t} - \bar{g}^{k,t}, g_i^{k,t} - h_i^k\rangle_{\alpha_i} \le \frac{1}{N}\sum_{i=1}^N \eta^{k,t}\|g_i^{k,t} - h_i^k\|_{\alpha_i}^2 + \frac{1}{\eta^{k,t}}\|g_i^{k,t} - \bar{g}^{k,t}\|_{\alpha_i}^2\\ \le& O(\eta^{k,t} L(LD+B)/\gamma^2 \cdot \mathrm{W}_1(\alpha,\alpha_i)) + O(\eta^{k,t} B^2/\gamma^2) + \frac{1}{\eta^{k,t}} \cdot \frac{1}{N}\sum_{i=1}^N \|g_i^{k,t} - \bar{g}^k\|_{\alpha_i}^2.\end{aligned}$$

The second term of (55) can be bounded by using the μ-strong convexity of $\mathcal{R}_i$ (note that $\mu = 1$ and we use $\nabla\mathcal{R}_i[g_i^{k,t}]$ to denote $(g_i^{k,t} - u_i)$ as they are identical on the support of α_i). The following inequality holds for the same reason as (33).

$$\begin{aligned}&\frac{2}{N}\sum_{i=1}^N \langle f^* - g_i^{k,t}, \nabla\mathcal{R}_i[g_i^{k,t}]\rangle_{\alpha_i}\\ \le& O\left(\eta^{k,t} G^2/\gamma^2\right) + \left(\frac{\mu}{2} + \frac{1}{\eta^{k,t}}\right)\frac{1}{N}\sum_{i=1}^N \|g_i^{k,t} - \bar{g}^{k,t}\|_{\alpha_i}^2 - \frac{\mu}{4}\|f^* - \bar{g}^{k,t}\|^2 - \frac{\mu}{2N}\sum_{i=1}^N \|f^* - g_i^{k,t}\|_{\alpha_i}^2\end{aligned}$$

Note that $\|f^t - \bar{g}^{k,t}\|^2_{\alpha_i} = \| \sum_{\kappa=1}^k \eta^{\kappa,t} \left(\bar{g}^{\kappa,t} - \bar{h}^{\kappa}\right) \|^2_{\alpha_i}$ and $\eta^{t,\kappa} \leq 2\eta^{t,k}$ for $\kappa \leq k$. Therefore, $\frac{1}{N}\sum_{i=1}^N \|g_i^{k,t} - \bar{g}^{k,t}\|^2_{\alpha_i}$ can be bounded by

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N \|g_i^{k,t} - \bar{g}^{k,t}\|^2_{\alpha_i} \\
=& \frac{1}{N}\sum_{i=1}^N \|g_i^{k,t} - f^t + f^t - \bar{g}^{k,t}\|^2_{\alpha_i} \\
\leq& \frac{1}{N}\sum_{i=1}^N 2\|g_i^{k,t} - f^t\|^2_{\alpha_i} + 2\|f^t - \bar{g}^{k,t}\|^2_{\alpha_i} = O(\sum_{\kappa=1}^k (\eta^{\kappa,t})^2 G^2/\gamma^2) = O(\eta^{k,t})^2 K^2 G^2/\gamma^2 \\
=& O(\frac{(\eta^{k,t})^2 K^2 G^2}{\gamma^2}).
\end{aligned}$$

Plug in the above results into (51) to yield

$$\begin{aligned}
\|\bar{g}^{k+1,t} - f^*\|^2_\alpha \leq (1 - \frac{\mu\eta^{k,t}}{2})\|\bar{g}^{k,t} - f^*\|^2_\alpha &+ O\left(\frac{(\eta^{k,t})^2 K^2}{\gamma^2}\left(L(LD+B)\omega + G^2 + B^2\right)\right) \\
&+ \frac{2\eta^{k,t}}{N}\sum_{i=1}^N \langle f^* - g_i^{k,t}, u_i - h_i^k\rangle_{\alpha_i} \\
&+ O(\eta^{k,t} L(LD+B)/\gamma^2 \cdot \omega) - \frac{\mu\eta^{k,t}}{2N}\sum_{i=1}^N \|f^* - g_i^{k,t}\|^2_{\alpha_i}
\end{aligned}$$

Set $\eta^{k,t} = \frac{4}{\mu(Kt+k+1)}$ and multiply both sides by $(Kt+k+1)$

$$\begin{aligned}
&(Kt+k+1)\|\bar{g}^{k+1,t} - f^*\|^2 \\
\leq& (Kt+k)\|\bar{g}^{k,t} - f^*\|^2 + O(\frac{K^2\left(L(LD+B)\omega + G^2 + B^2\right)}{\mu^2\gamma^2(Kt+k+1)}) \\
&+ \frac{4}{\mu N}\sum_{i=1}^N \langle f^* - g_i^{k,t}, u_i - h_i^k\rangle_{\alpha_i} + O\left((L(LD+B)\cdot\omega/(\gamma^2\mu)\right) - \frac{1}{N}\sum_{i=1}^N \|f^* - g_i^{k,t}\|^2_i
\end{aligned}$$

Sum from $k = 1$ to K

$$\begin{aligned}
&(Kt+K+1)\|\bar{g}^{k+1,t} - f^*\|^2 \\
\leq& (Kt+1)\|\bar{g}^{1,t} - f^*\|^2 + O\left(\frac{K^2\left(L(LD+B)\omega + G^2 + B^2\right)}{\mu^2\gamma^2}(\log(Kt+K+1) - \log(Kt+1))\right) \\
&+ \frac{4}{\mu N}\sum_{i=1}^N\sum_{k=1}^K \langle f^* - g_i^{k,t}, u_i - h_i^k\rangle_{\alpha_i} + O\left(KL(LD+B)\cdot\omega/(\gamma^2\mu)\right) - \frac{1}{N}\sum_{k=1}^K\sum_{i=1}^N \|f^* - g_i^{k,t}\|^2_{\alpha_i}
\end{aligned}$$

We now focus on the last term

$$
\begin{aligned}
&\sum_{k=1}^{K}\langle f^* - g_i^{k,t}, u_i - h_i^k\rangle_{\alpha_i} \\
=& \sum_{k=1}^{K}\langle f^* - g_i^{k,t}, u_i - (u_i - \Delta_i^{k-1} + \Delta_i^k)\rangle_{\alpha_i} = \sum_{k=1}^{K}\langle f^* - g_i^{k,t}, \Delta_i^{k-1} - \Delta_i^k\rangle_{\alpha_i} \\
=& \sum_{k=1}^{K}\langle f^* - g_i^{k,t}, -\Delta_i^k\rangle_{\alpha_i} + \sum_{k=2}^{K}\langle f^* - g_i^{k,t}, \Delta_i^{k-1}\rangle_{\alpha_i} + \langle f^* - g_i^1, \Delta_i^0\rangle_{\alpha_i} \qquad \&\Delta_i^0 = 0 \\
=& \sum_{k=1}^{K}\langle f^* - g_i^{k,t}, -\Delta_i^k\rangle_{\alpha_i} + \sum_{k=1}^{K-1}\langle f^* - g_i^{k+1}, \Delta_i^k\rangle_{\alpha_i} \\
=& \sum_{k=1}^{K}\langle f^* - g_i^{k,t}, -\Delta_i^k\rangle_{\alpha_i} + \sum_{k=1}^{K-1}\langle f^* - g_i^{k,t}, \Delta_i^k\rangle_{\alpha_i} + \sum_{k=1}^{K-1}\langle \eta_t^k\left(g_i^{k,t} - h_i^k\right), \Delta_i^k\rangle_{\alpha_i} \\
=& \langle f^* - g_i^{K,t}, -\Delta_i^K\rangle_{\alpha_i} + \sum_{k=1}^{K-1}\langle \eta_t^k\left(g_i^{k,t} - h_i^k\right), \Delta_i^k\rangle_{\alpha_i} \\
\leq& \mu\|f^* - g_i^{k,t}\|_{\alpha_i}^2 + O((1-\gamma)^2B^2/(\gamma^2\mu)) + O(G^2\frac{1-\gamma}{\gamma^2})\sum_{k=1}^{K-1}\eta^{k,t} \\
=& \mu\|f^* - g_i^{k,t}\|_{\alpha_i}^2 + O((1-\gamma)^2B^2/(\gamma^2\mu)) + O\left(G^2\frac{1-\gamma}{\gamma^2}(\log(KT+K) - \log(Kt))\right).
\end{aligned}
$$

Using this result, we obtain

$$
\begin{aligned}
&(K(t+1)+1)\|f^{t+1} - f^*\|^2 \\
\leq& (Kt+1)\|f^t - f^*\|^2 + O\left(\frac{K^2\left(L(LD+B)\omega + G^2 + B^2\right)}{\mu^2\gamma^2}(\log(K(t+1)+1) - \log(Kt+1))\right) \\
&+ \frac{2}{\mu}(O((1-\gamma)^2B^2/(\gamma^2\mu)) + O\left(G^2\frac{1-\gamma}{\gamma^2}(\log(K(t+1)) - \log(Kt))\right)) \\
&+ O\left(KL(LD+B)\cdot\omega/(\gamma^2\mu)\right)
\end{aligned}
$$

Sum from $t = 0$ to $T-1$ and use the non-expensiveness of the projection operation (note that $\mathcal{C}$ is a convex set) to yield

$$
\begin{aligned}
(KT+1)\|f^T - f^*\|^2 \leq (k_0+1)\|f^0 - f^*\|^2 &+ O\left(\frac{K^2\left(L(LD+B)\omega + G^2 + B^2\right)\log(KT+1)}{\mu^2\gamma^2}\right) \\
&+ O(\frac{(1-\gamma)^2TB^2}{\mu^2\gamma^2} + \frac{(1-\gamma)G^2\log(TK)}{\mu\gamma^2}) + O\left(TKL^2\cdot\omega/(\gamma^2\mu)\right),
\end{aligned}
$$

and hence ($B = O(G)$)

$$
\|f^T - f^*\|^2 = O(\frac{\|f^0 - f^*\|^2}{KT} + \frac{K\left(L(LD+B)\omega + G^2\right)log(KT)}{T\mu^2\gamma^2} + \frac{(1-\gamma)^2B^2}{\mu^2\gamma^2K} + \frac{L(LD+B)\omega}{\gamma^2\mu}).
$$

□

G Partial Device Participation

In this section, we consider the setting of partial device participation. In the following discussion, we take the setting of homogeneous input distributions for example. Similar arguments hold for the other two settings.

In round t of Algorithm 2, we randomly sample without replacement a subset $\mathcal{S}_t \subseteq [N]$ of clients and only compute the average of their returns to update the global variable function. We assume that all $\mathcal{S}_t$ has the same cardinality m. Conceptually, we can imagine all the clients are participating in the update, but we only utilize the results in the set $\mathcal{S}_t$.

Similar to the proof for the setting of full device participation, we define the hypothetical global average function $\bar{g}^{k,t} = \frac{1}{N}\sum_{i=1}^{N} g_i^{k,t}$. In particular, we have $\bar{g}^{1,t} = f^t$. From the derivation therein (see Section E), we have

$$\begin{aligned}
&(tK+K+1)\|\bar{g}^{K+1,t} - f^*\|_\alpha^2 \\
\leq\, &(tK+1)\|\bar{g}^{1,t} - f^*\|_\alpha^2 + O(\frac{K^2G^2}{\gamma^2\mu^2})(\log(tK+K+1) - \log(tK++1)) \\
&+ \frac{4}{\mu}(\frac{1}{2\mu}(\frac{1-\gamma}{\gamma})^2G^2 + \frac{(1-\gamma)(2-\gamma)}{\gamma^2}G^2(\log(tK+K) - \log(tK)))
\end{aligned}$$

However, unlike the setting of full device participation, we do not have $f^{t+1} = \bar{g}^{K+1,t}$. Instead, $f^{t+1} = \frac{1}{m}\sum_{i\in\mathcal{S}^t} g_i^{K+1}$. We have the following simple but useful lemma. The proof of this lemma is similar to scheme II of Lemma 5 in [Li et al., 2019].

Lemma G.1. $\mathbb{E}_{\mathcal{S}_t}\|f^{t+1} - \bar{g}^{K+1,t}\|_\alpha^2 = O\left(\frac{N-m}{N-1}\frac{(\eta^{K,t})^2K^2G^2}{m\gamma^2}\right)$.

Moreover, f^{t+1} is an unbiased estimator of $\bar{g}^{K+1,t}$. Therefore $\mathbb{E}_{\mathcal{S}^t}\langle f^{t+1} - \bar{g}^{K+1,t}, \bar{g}^{K+1,t} - f^*\rangle_\alpha = 0$ and

$$\mathbb{E}_{\mathcal{S}^t}\|f^{t+1} - f^*\|_\alpha^2 = \mathbb{E}_{\mathcal{S}^t}\|f^{t+1} - \bar{g}^{K+1,t}\|^2 + \|\bar{g}^{K+1,t} - f^*\|_\alpha^2. \tag{56}$$

Recall that $\eta^{k,t} = \frac{2}{\mu(tK+k+1)}$. Combining the above results, we have

$$\begin{aligned}
&((t+1)K+1)\,\mathbb{E}_{\mathcal{S}_t}\|f^{t+1} - f^*\|_\alpha^2 \\
\leq\, &(tK+1)\|f^t - f^*\|_\alpha^2 + O(\frac{K^2G^2}{\gamma^2\mu^2})(\log(tK+K+1) - \log(tK+1)) \\
&+ \frac{4}{\mu}(\frac{1}{2\mu}(\frac{1-\gamma}{\gamma})^2G^2 + \frac{(1-\gamma)(2-\gamma)}{\gamma^2}G^2(\log(tK+K) - \log(tK))) \\
&+ O\left(\frac{N-m}{N-1}\frac{\eta^{K,t}K^2G^2}{\mu m\gamma^2}\right).
\end{aligned}$$

Sum the above results from $t=0$ to T, we have

$$\begin{aligned}
&((T+1)K+1)\,\mathbb{E}\|f^{T+1} - f^*\|_\alpha^2 \\
\leq\, &\|f^0 - f^*\|_\alpha^2 + O(\frac{K^2G^2\log(KT)}{\gamma^2\mu^2}) + O(\frac{TG^2}{\mu^2}(\frac{1-\gamma}{\gamma})^2) \\
&+ O\left(\frac{(1-\gamma)}{\mu\gamma^2}G^2\log(TK+K)\right) + O(\frac{N-m}{N-1}\frac{KG^2\log T}{\mu^2 m\gamma^2}) \\
\Rightarrow\, &\mathbb{E}\|f^{T+1} - f^*\|_\alpha^2 = O\bigg(\frac{\|f^0 - f^*\|_\alpha^2}{KT} + \frac{KG^2\log(KT)}{T\gamma^2\mu^2} + \frac{G^2}{K\mu^2}(\frac{1-\gamma}{\gamma})^2 \\
&\qquad + \frac{(1-\gamma)G^2\log(TK)}{KT\mu\gamma^2} + \frac{N-m}{N-1}\frac{G^2\log T}{T\mu^2 m\gamma^2}\bigg).
\end{aligned}$$

Theorem G.1. *Let f^0 be the initializer function. Suppose that Assumption 4.2 holds, and suppose the weak learning oracle $\mathcal{Q}_\alpha^2$ satisfies (5) with constant γ. We pick the step size $\eta^{k,t} = \frac{2}{\mu(tK+k+1)}$ and in each round the server randomly selects a subset $\mathcal{S}_t \subseteq [N]$ without replacement with $|\mathcal{S}_t| = m$. The output of* FFGB *(Algorithm 2) satisfies*

$$\|f^T - f^*\|_\alpha^2 = O(\frac{\|f^0 - f^*\|_\alpha^2}{KT} + \frac{KG^2\log(KT)}{T\gamma^2\mu^2} + \frac{(1-\gamma)G^2}{K\mu^2\gamma^2} + \frac{(1-\gamma)G^2\log(KT)}{KT\mu\gamma^2} + \frac{N-m}{N-1}\frac{G^2\log T}{T\mu^2 m\gamma^2}).$$

Theorem G.2. *Let f^0 be the initializer function. We define a proxy of the heterogeneity among the local input distributions α's in the* TV *distance as*

$$\omega_{\mathrm{TV}} \triangleq \frac{1}{N}\sum_{i=1}^{N}\mathrm{TV}(\alpha, \alpha_i). \tag{57}$$

Suppose Assumption 4.1 holds, and suppose that the weak learning oracle $\mathcal{Q}_\alpha^\infty$ *satisfies* (10) *and* (11) *with constant* γ *and* $\bar{G}_\lambda = 2G/\lambda$. *We use the step sizes* $\eta^{k,t} = 4/(\mu(tK+k+1))$ *and in each round the server randomly selects a subset* $\mathcal{S}_t \subseteq [N]$ *without replacement with* $|\mathcal{S}_t| = m$. *The output of* FFGB *satisfies*

$$\|f^T - f^*\|_\alpha^2 = O\bigg(\frac{\|f^0 - f^*\|_\alpha^2}{KT} + \frac{KG^2 log(KT)}{T\mu^2\gamma^2} + \frac{(1-\gamma)^2G^2}{K\mu^2\gamma^2} + \frac{G^2\omega_{\mathrm{TV}}}{\mu^2\gamma^2} + \frac{N-m}{N-1}\frac{G^2\log T}{T\mu^2 m\gamma^2}\bigg).$$

Theorem G.3. *Consider the special case of the federated functional minimization problem with square loss. Assume that the optimal solution* f^* *is L-Lipschitz continuous. Let* f^0 *be the initializer. We define a proxy of the heterogeneity among the local input distributions* α*'s in the Wasserstein-1 distance as*

$$\omega_{W_1} = \frac{1}{N}\sum_{i=1}^N \mathrm{W}_1(\alpha, \alpha_i) \tag{58}$$

Suppose that Assumption 4.3 holds and define $G^2 = \frac{2L^2}{N^2}\sum_{i,s=1}^N \mathrm{W}_2^2(\alpha_s, \alpha_i) + 2B^2$. *Moreover, suppose the weak learning oracle* $\mathcal{Q}_\alpha^{\mathrm{lip}}$ *satisfies* (14) *and* (15) *with constant* γ. *We pick the step sizes* $\eta^{k,t} = 4/(\mu(tK+k+1))$ *and in each round the server randomly selects a subset* $\mathcal{S}_t \subseteq [N]$ *without replacement with* $|\mathcal{S}_t| = m$. *The output of* FFGB *satisfies*

$$\|f^T - f^*\|_\alpha^2 = O\bigg(\frac{K\left(L(LD+B)\omega_{W_1} + G^2\right) log(KT)}{T\mu^2\gamma^2} + \frac{\|f^0 - f^*\|^2}{KT} + \frac{(1-\gamma)^2B^2}{\mu^2\gamma^2K} + \frac{L(LD+B)\omega_{W_1}}{\gamma^2\mu} + \frac{N-m}{N-1}\frac{G^2\log T}{T\mu^2 m\gamma^2}\bigg).$$