AN AGNOSTIC APPROACH TO FEDERATED LEARNING WITH CLASS IMBALANCE

Zebang Shen, Juan Cervino, Hamed Hassani, Alejandro Ribeiro
Department of Electrical and Systems Engineering
University of Pennsylvania
Philadelphia, PA 19104, USA
{zebang,jcervino,hassani,aribeiro}@seas.upenn.edu

ABSTRACT

Federated Learning (FL) has emerged as the tool of choice for training deep models over heterogeneous and decentralized datasets. As a reflection of the experiences from different clients, severe class imbalance issues are observed in real-world FL problems. Moreover, there exists a drastic mismatch between the imbalances from the local and global perspectives, i.e. a local majority class can be the minority of the population. Additionally, the privacy requirement of FL poses an extra challenge, as one should handle class imbalance without identifying the minority class. In this paper we propose a novel agnostic constrained learning formulation to tackle the class imbalance problem in FL without requiring further information beyond the standard FL objective. A meta algorithm, CLIMB, is designed to solve the target optimization problem, with its convergence property analyzed under certain oracle assumptions. Through an extensive empirical study over various data heterogeneity and class imbalance configurations, we showcase that CLIMB considerably improves the performance in the minority class without compromising the overall accuracy of the classifier, which significantly outperforms previous arts. In fact, we observe the greatest performance boost in the most difficult scenario where every client only holds data from one class. The code can be found here.

1 INTRODUCTION

Class imbalance is ubiquitous in real world supervised learning problems. Examples span from medical applications (Lee & Shin, 2020; Roy et al., 2019; Choudhury et al., 2019), fraud detection (Yang et al., 2019; Chan et al., 1999), to consumer based applications (Wang et al., 2021b; Wu et al., 2020; Long et al., 2020). In these scenarios, data belonging to a subset of classes constitute a great proportion of the population while data from the minority classes, generated by uncommon events, are scarce (He & Garcia, 2009). Having a non-uniform number of samples per class deteriorates the performance of the classifier in the minority class (Huang et al., 2016), resulting in low training and testing accuracy. More importantly, unintended consequences of mistreating the minority class can be catastrophic (Van Hulse et al., 2007) if the problem is not handled appropriately.

From the perspectives of *data heterogeneity* and *privacy*, the issue of class imbalance is even more significant in the setting of Federated Learning (FL). Due to the heterogeneity of the local data distributions, there can be a significant mismatch between the local and global imbalance, i.e. the class that is a minority locally can actually be a majority class globally. Moreover, for the purpose of privacy protection in FL, one should tackle the class-imbalance problem in an agnostic manner, i.e. the proposed algorithms *should not* require the minority class to be identified.

In the centralized training setting, techniques like balanced sampling, loss re-weighting, and gradient tuning have achieved many successes (Johnson & Khoshgoftaar, 2019). However, due to the extra difficulty of handling the class imbalance problem in FL, previous arts that rely on the explicit identification of the minority class do not directly apply. While research along this line is quite limited, a commonly used heuristic is to estimate the portion of a class using the norm of the gradient per class. When used as proxies, these quantities are then utilized to reweight the losses

corresponding to different classes (Wang et al., 2021a; Yang et al., 2020). A key drawback of (Wang et al., 2021a) is that a subset of the client's local dataset needs to be shared, which is not ideal for privacy preserving purposes. Moreover, the effectiveness of such approaches degrades when there is a notable mismatch between the local imbalance and the global imbalance (Wang et al., 2021a).

In this work, we target the most difficult yet possibly most interesting setting in FL, where there is a significant mismatch between the local and global imbalance. Concretely, on certain clients, a minority class from a global perspective is in the majority locally. Practitioners often encounters such a mismatch due to the highly heterogeneous data configuration in FL and the *principle of locality* (Wang et al., 2021a). For example, when the local data are produced by a minority user, the corresponding minority data could occupy a large portion of the local distribution. In this scenario, due to the scarceness of the minority data from a global perspective, the corresponding underrepresented client will usually experience poor service quality from the trained model, giving rise to potentially severe fairness issue.

To overcome the aforementioned challenges, we propose a constrained learning formulation to handle the class imbalance issue in FL while accounting for both *heterogeneity* and *privacy*. In brief, we impose constraints on the standard FL formulation so that the empirical loss on every client should not overly exceed the average empirical loss. Such constraints are shown to force the classifier to account for all classes equally and hence mitigate the detrimental effect of class imbalance, under a type of heterogeneous data configuration that captures the mismatch between the local and global imbalance. The advantages of our formulation are threefold: First, in contrast to previous arts which usually rely on some heuristics to reweight the loss functions for different classses, our approach is principled, yielding a simple optimization interpretation. Second, unlike existing methods, our formulation requires no additional information compared to the original FL formulation and it treats data from different classes agnostically, and hence is less likely to leak the private information of the clients. Third, from our extensive empirical study, our formulation can significantly improve the testing accuracy on the minority class, without compromising the overall performance.

Contribution. We summarize our contributions as follows.

1. **Problem Formulation:** To address the challenge of class imbalance in FL with data heterogeneity, we propose a novel constrained FL formulation with explicit enforcement on the similarity between the local empirical losses. Using only information from the standard FL objective, our approach is completely agnostic to class distribution of the client data and its proxies, as opposed to the existing literature (which mostly violate privacy requirements of FL).

2. **Meta-Algorithm:** We solve our constrained optimization problem via a primal-dual approach. For a fixed dual variable, the corresponding Lagrangian function enjoys a similar structure as the standard FL loss, but with non-uniform weights on the local objectives. Accordingly, we propose an efficient (meta-) algorithm called `CLIMB` that can use any FL optimization method as a subroutine, with nearly negligible communication overhead. Furthermore, we analyze the convergence properties of `CLIMB` under certain oracle assumptions.

3. **Accuracy Improvement in the minority class:** On benchmark datasets, `CLIMB` with Fed-Avg as base solver achieves significant enhancement of the accuracy in the minority class without compromising the overall accuracy under various data heterogeneity and class imbalance scenarios.

2 Related Works

Class Imbalance in Centralized Setting. In the centralized learning setting, the number of samples per class is known. We categorize existing solutions that exploit such information as follows.
Balanced sampling. In order to balance the data used for gradient calculation, the most common approaches are either to under sample from the majority classes (Liu et al., 2008), or to augment the minority class (Chawla et al., 2002; Guo & Viktor, 2004). Both methods seek to generate an artificial uniform distribution of data (Buda et al., 2018; Pouyanfar et al., 2018).
Loss reweighting. Methods in this category focus on re-weighting the loss functions for different classes, with extra emphases on the mistakes made in the minority class (Cui et al., 2019; Ling & Sheng, 2008; Sun et al., 2007; Wang et al., 2016). There also exist works like (Lin et al., 2017) that adjust the scale of the loss sample-wise, without exploiting the global data distribution.
Gradient tuning. In the context of neural networks, some other works focus on tuning the gradient per class (Anand et al., 1993) showing faster rates of convergence in the class imbalance setting.

To do this, the gradient is reshaped in order to have an equal magnitude when projected to all the gradients per class, while keeping the same norm as the original gradient.

Class Imbalance in Federated Learning. Combating with the issue of class imbalance is more challenging in FL since the data composition in such a setting is generally unknown and the minority classes are often difficult to identify Li et al. (2020); Wang et al. (2020b). While research in this direction are quite limited, to the best of our knowledge, we classify them into the following two classes based on whether a proxy of the data composition is explicitly built.

(i) A common strategy in existing methods is to explicitly build proxies of the data composition based on some heuristics. These proxies are dynamically adjusted during the learning procedure and allow the aforementioned three techniques, balanced sampling, loss re-weighting and gradient tuning, to be utilized in the FL setting: Wang et al. (2021a) suggests that there exists a proportional relation between the magnitude of the gradients the corresponds to the last layer of the neural network and the sample quantity. Based on such observation, the proposed the `Ratio-Loss`, a class-wise re-weighted version of the standard cross entropy loss. We note that to define the `Ratio-Loss`, one needs to collect a subset of client data as "auxiliary data", which may not be ideal in the setting of FL. On the same vein, Yang et al. (2020) argues that the gradient per class can be used as a proxy to infer the imbalances in the distributions of the clients, and thus clients should be selected according to how uniform the magnitude of gradient per class is. We also notice the work `Astraea` which introduces extra virtual components called mediators between the FL server and clients (Duan et al., 2019). The mediator is assumed to have access to the local data distributions of the clients so that client rescheduling and data augmentation can be carried out accordingly. Through the prepossessing on the mediator, the gradients communicated to the server are made balanced class-wise. However, when privacy is taken into consideration, methods like `Astraea` may not be appropriate as in essence it is directly built on the global data distribution.

(ii) Without explicitly constructing an estimation of the data composition, there are works that resort to techniques like active learning and reinforcement learning to implicitly learn the data composition as the optimization progresses. Works along this research line usually rely on client selection to mitigate the effects of class imbalances. In (Goetz et al., 2019), to address the class imbalance problem, there will a higher probability to sampling clients that posses the global minority class. Other works, leverage client selection as a multi-armed bandit problem, and design a client selection policy to balance the gradients (Xia et al., 2020). Other arts leverage Q-learning techniques to select clients in order to minimize the overall loss Wang et al. (2020a). However, we believe in the most practical setting of FL, clients that are available per communication round is incoercible and hence these strategies have limited applicability.

Comparison with Previous Works. Our method significantly differs from the above strategies in the following ways: (1) Our method requires no knowledge of the data composition, nor any proxy of such information. Consequently, our approach is agnostic to the minority class and better preserves clients' privacy. Such a property draws clear distinction between our work and the works listed in class (i) above. (2) As will be more clear in the following section, we only perform client-wise re-weighting as opposed to the class-wise re-weighting schemes used in previous works, which further emphasize the agnostic nature of our approach. (3) In contrast to the methods in class (ii) above, our approach does not require active client selection, enjoying a broader applicability in the most practical and interesting settings.

3 Federated Learning with Empirical Loss Constraints

We consider the problem of multi-class classification. Let $x \in \mathcal{X} \subseteq \mathbb{R}^d$ be the input and use $y \in \mathcal{Y} = \{1, \dots, C\}$ to denote the target label, where C is the total number of classes. In the context of FL, we assume to have N clients, each of which posses its own private data distribution $p_i(x, y)$. Here, $p_i(x, y)$ is a joint probability distribution of the input x and the output label y. One can decompose $p_i(x, y)$ as $p_i(x, y) = p(x|y)p_i(y)$, where $p(x|y)$ is the conditional distribution of the input x given class y and $p_i(y)$ is the marginal distribution of class y on client i. We assume that the conditional distribution $p(x|y)$ is identical on all devices, but the marginal distribution $p_i(y)$ can vary significantly due to the heterogeneity of the data configuration.

Let $\mathcal{H} = \{\phi : \mathcal{X} \times \Theta \to \mathbb{R}^C\}$ be a family of parameterized predictors with parameter $\theta \in \Theta \subseteq \mathbb{R}^Q$. Let $\ell : \mathbb{R}^C \times \mathcal{Y} \to \mathbb{R}_+$ be the loss function, then the local objective function of client i is defined as

$$f_i(\theta) = \mathbb{E}_{(x,y)\sim p_i}[\ell(\phi(x, \theta), y)]. \tag{1}$$

For multi-class classification, ℓ is often chosen to be the cross entropy loss. In the standard formulation of FL, the goal is to minimize the global average of local objectives

$$\min_{\theta \in \Theta} \bar{f}(\theta) := \frac{1}{N} \sum_{i=1}^{N} f_i(\theta). \tag{2}$$

While it is well known that, in the presence of class imbalance, the above vanilla formulation will produce models that perform poorly on the minority data, we consider the following configuration of heterogeneous local class distributions in order to make a quantitative analysis of such a phenomenon for a concrete class imbalance setting. More importantly, such a setting will also motivate our constrained FL formulation. We emphasize that the following setting is just used as a motivating example to showcase our claims, and our results apply generally to any FL setting.

Motivating Example. Let u be the uniform distribution over the classes, i.e. for $y \sim u$, $\Pr(y = c) = \frac{1}{C}, \forall c \in \mathcal{Y}$ and let δ_c be the Dirac distribution of class c. We assume for some small but fixed $\alpha \in [0, 1]$, the local class distribution of the client i is a mixture of the uniform distribution over all classes and the Dirac distribution of a fixed class c_i, i.e. $p_i = \alpha u + (1 - \alpha)\delta_{c_i}$. We use N_c to denote the number clients with $c_i = c$. Note that in the limit setting when $\alpha = 0$, the aforementioned configuration captures the most heterogeneous setting: clients only have data from a single class.
We consider an extreme case of class imbalance under the above configuration. Without loss of generality, we consider the binary classification problem, i.e. $C = 2$, and we assume class 1 to be the minority class with $N_1 = 1$. We define $g_i(\theta) := \mathbb{E}_{x \sim p(x|y=i)}[\ell(\phi(x, \theta), i)]$ as the loss of the predictor $\phi(\cdot, \theta)$ on the data with $y = i$. We can calculate that

$$\bar{f}(\theta) = \left(\frac{\alpha}{2} + \frac{1 - \alpha}{N}\right) g_1(\theta) + \left(\frac{\alpha}{2} + \frac{(1 - \alpha)(N - 1)}{N}\right) g_2(\theta). \tag{3}$$

Clearly, when α, the portion of data with uniform label distribution, is small, e.g. $\alpha = 0$, and N, the total number of clients, is large, the loss of the predictor on class 1 has negligible weight, which often leads to the poor performance on the minority class in the trained classifier as the majority classes dominate the gradient. Moreover, observe that when α is small, there is a significant mismatch between the local and global class imbalance: the global minority class 1 is in the majority on the corresponding client locally. Such phenomenon is pertinent to the FL setting due to the data heterogeneity and poses a great challenge for tackling the issue of class imbalance in FL.

3.1 Constrained FL Formulation

In our work, to address the class imbalance challenge, we propose to minimize the following constrained FL (CFL) formulation

$$P_\epsilon^* = \min_{\theta \in \Theta} \bar{f}(\theta) := \frac{1}{N} \sum_{i=1}^{N} f_i(\theta) \tag{CFL}$$
$$\text{s.t. } f_i(\theta) - \bar{f}(\theta) \leq \epsilon, \forall i \in [1, \ldots, N].$$

Here, ϵ is a tolerance constant that controls the enforced closeness in the training loss among clients, and we emphasize the dependence of the optimal value P_ϵ^* on ϵ by encoding it in the subscript. As a motivation, we show that the constraint in our formulation (CFL) can be translated to a constraint on the performance of the minority class, under the above class imbalance setting: Consider the setting where the tolerance constant is very small, i.e. ϵ is close to zero. WLOG, we assume that the first device is the one that has $c_i = 1$. Recall the definition of g_i above Eq.(3). One can compute that

$$f_1(\theta) - \bar{f}(\theta) = \frac{(1 - \alpha)(N - 1)}{N} (g_1(\theta) - g_2(\theta)) \leq \epsilon \iff g_1(\theta) - g_2(\theta) \leq \frac{N\epsilon}{(1 - \alpha)(N - 1)}.$$

Therefore, our formulation (CFL) enjoys a clear class-balancing interpretation in the highly heterogeneous setting when α is small and N is large. Note that this is achieved *without* the identification of the minority class and server collects *no* additional information compared to the vanilla FL.

While the above nice interpretation of the proposed constraint may not hold exactly for general class imbalance settings, in spirit, we want the resulting classifier from our algorithm to perform similarly on every class, or in other words to account for the minority class and majority class equally. In the following section, we discuss how to solve the proposed formulation by alternating the primal and dual updates of an equivalent Lagrangian formulation.

Algorithm 1 CLIMB: CLass IMBalance Federated Learning

1: **Input**: initial model θ^0, a subroutine *ClientUpdate*, dual step size η_D, maximum round T;
2: Initialize the dual variables $\boldsymbol{\lambda} = [0, \ldots, 0]$;
3: **for** $t = 1, 2, \ldots, T$ **do**
4: Compute weights: $\forall i \in [N], w_i = 1 + \lambda_i - \bar{\lambda}$, with $\bar{\lambda} = \frac{1}{N}\sum_{i=1}^N \lambda_i$;
5: **Primal Update**: $\theta^{t+1} \leftarrow$ *ClientUpdate*$(\{w_i\}_{i=1}^N, \theta^t)$;
6: **Dual Update**: $\forall i \in [N], \lambda_i \leftarrow [\lambda_i + \eta_D(f_i(\theta^{t+1}) - \bar{f}(\theta^{t+1}) - \epsilon)]_+$, with $\bar{f} = \frac{1}{N}\sum_{i=1}^N f_i$;
7: **end for**
8: **Output**: model θ^{T+1}

3.2 Algorithm Construction

In order to solve problem (CFL), we resort to the method of Lagrange multipliers. By introducing the dual variables $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_N] \in \mathbb{R}_+^N$, we define the Lagrangian function as,

$$\mathcal{L}(\theta, \boldsymbol{\lambda}) = \frac{1}{N}\sum_{i=1}^N f_i(\theta) + \lambda_i \left(f_i(\theta) - \frac{1}{N}\sum_{j=1}^N f_j(\theta) - \epsilon \right) \tag{4}$$

$$= \frac{1}{N}\sum_{i=1}^N (1 + \lambda_i - \bar{\lambda}) f_i(\theta) - \lambda_i \epsilon, \qquad \text{with } \bar{\lambda} = \frac{1}{N}\sum_{i=1}^N \lambda_i. \tag{5}$$

With this in hand, we can construct a lower bound of the above constrained optimization problem using the Lagrangian $\mathcal{L}(\theta, \boldsymbol{\lambda})$ as follows:

$$D_\epsilon^* = \max_{\boldsymbol{\lambda} \in \mathbb{R}_+^N} \min_{\theta \in \Theta} \mathcal{L}(\theta, \boldsymbol{\lambda}) \leq \min_{\theta \in \Theta} \max_{\boldsymbol{\lambda} \in \mathbb{R}_+^N} \mathcal{L}(\theta, \boldsymbol{\lambda}) = P_\epsilon^*,$$

where we often refer to D_ϵ^* as the dual problem. To solve the dual problem, we propose CLIMB, a method described in Algorithm 1, which is discussed in detail as follows.
CLIMB proceeds by alternatingly optimizing over the primal variable θ and the dual variable $\boldsymbol{\lambda}$.
Primal Update. For a fixed $\boldsymbol{\lambda}$, the minimization of $\mathcal{L}$ with respect to θ is equivalent to a re-weighted version of the standard unconstrained FL objective (2):

$$\min_{\theta \in \Theta} \mathcal{L}(\theta, \boldsymbol{\lambda}) \Longleftrightarrow \min_{\theta \in \Theta} \frac{1}{N}\sum_{i=1}^N (1 + \lambda_i - \bar{\lambda}) f_i(\theta). \tag{6}$$

Importantly, this simple and canonical form allows us to perform the update on θ using any FL solver as the base optimizer. For flexibility, we do not explicitly choose the base optimizer in our algorithm description, but refer to the update on θ as the subroutine *ClientUpdate*$(\{w_i\}_{i=1}^N, \theta^t) \to \theta^{t+1}$. Such a subroutine takes the non-uniform weights $\{w_i\}_{i=1}^N$ on the local objectives and the current consensus model θ^t as inputs, and returns an updated model θ^{t+1}. Note that every call to *ClientUpdate* may consists of multiple communication rounds.
Dual Update. Once we have obtained the new consensus model θ^{t+1}, we perform dual update on λ by taking a *single* dual ascent step of the following equivalent objective (given some fixed θ),

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}_+^N} \mathcal{L}(\theta, \boldsymbol{\lambda}) \Longleftrightarrow \max_{\boldsymbol{\lambda} \in \mathbb{R}_+^N} \frac{1}{N}\sum_{i=1}^N \lambda_i \left(f_i(\theta) - \frac{1}{N}\sum_{j=1}^N f_j(\theta) - \epsilon \right). \tag{7}$$

To evaluate the gradient of $\mathcal{L}$ with respect to $\boldsymbol{\lambda}$, we need to first broadcast the consensus model θ^{t+1} and then aggregate the function values $f(\theta^{t+1})$. Since the broadcast model can be used for the next round of primal update, the only overhead of CLIMB compared to the standard FL solver *ClientUpdate*, is to transmit the functional value, which is negligible.

Remark 3.1 *We emphasize that CLIMB can be implemented in a privacy-preserving manner: A client can carry out its update locally given the access to the global average of the dual variables $\bar{\lambda}$ and the global average of the loss functions $\bar{f}(\theta)$. These quantities can be computed via the standard FL technique of Homomorphic Encryption without revealing the exact value of the dual variable λ_i and local loss $f_i(\theta)$ to the server, as elaborated in Appendix D.*

3.3 THEORETICAL GUARANTEES

Based on the recent progress in constrained learning (Chamon et al., 2021), we show that under mild regularity conditions, the duality gap between the dual problem D_ϵ^* and the primal problem P_ϵ^* can be controlled by some quantity that describes the capability of the parametric function class $\mathcal{H}$.

Assumption 3.1 *The loss function ℓ in the definition of the local objective (1) is L-Lipschitz, i.e.$\|\ell(x,\cdot)-\ell(z,\cdot)\| \leq L\|x-z\|$, and bounded by B.*

Assumption 3.2 *The conditional distribution $p(x|y)$ is non-atomic for all $y \in \mathbb{R}^C$.*

While usually the local data distribution is discrete, we can always augment it by randomly perturbing the data points with white noise, this is often used as data augmentation in vision tasks.

Assumption 3.3 *There exists a convex hypothesis class $\hat{\mathcal{H}}$ such that $\mathcal{H} \subseteq \hat{\mathcal{H}}$, and there exists a constant $\xi > 0$ such that $\forall \hat{\phi} \in \hat{\mathcal{H}}$, there exists $\theta \in \Theta$ such that $\sup_{x\in\mathcal{X}} \|\hat{\phi}(x) - \phi(x,\theta)\| \leq \xi$.*

A simple strategy to construct $\hat{\mathcal{H}}$ is to take the convex hull of $\mathcal{H}$. When $\mathcal{H}$ is sufficiently rich, ξ can be expected to be small. Notice that this bound can be decreased by increasing the richness of the function class $\mathcal{H}$. All the proofs can be found in the Appendix.

Theorem 3.1 (Near Zero Duality Gap) *Under Assumptions 3.1, 3.2,3.3, and the Contrained Federated Learning problem is feasible in $\hat{\mathcal{H}}$ with constraint $\epsilon - 2L\xi$, the Constrained Federated Learning problem has near zero-duality gap,*

$$P_\epsilon^* - D_\epsilon^* \leq (2|\boldsymbol{\lambda}^*_{\epsilon-2L\xi}|_1 + 1)L\xi, \tag{8}$$

where $\boldsymbol{\lambda}^*_{\epsilon-2L\xi}$ is the optimal dual variable associated with the Constrained Federated Learning problem (CFL) *with constraints $\epsilon - 2L\xi$ over the space of functions $\hat{\mathcal{H}}$.*

Theorem 3.1 establishes an upper bound on the duality gap of the Constrained Federated Learning Problem CFL. The gap depends on the Lipschitz constant of the loss function L, the richness of the function class ξ, and the optimal dual variable of a more restrictive problem. Note that we required the Constrained Federated Learning problem to be feasible for constraint $\epsilon - 2L\xi$. In the case of the cross-entropy loss, as long as $\epsilon - 2L\xi > 0$, this can be satisfied by a classifier that assigns a uniform label for every sample, as each individual loss will be equal to each other.
Since the minimization of the Lagrangian function $\mathcal{L}$ is non-convex, to show the convergence of `CLIMB`, we need an additional oracle assumption as follows.

Assumption 3.4 (Approximate Solution) *For every dual variable $\boldsymbol{\lambda} \in \mathbb{R}_+^N$, and precision $\delta > 0$ there exists an oracle approximate solution $\theta_{\boldsymbol{\lambda}}$ such that $\mathcal{L}(\theta_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) \leq \min_{\theta\in\mathbb{R}^Q} \mathcal{L}(\theta, \boldsymbol{\lambda}) + \delta$.*

Theorem 3.2 (Convergence) *Define the dual function $d(\boldsymbol{\lambda}) = \min_{\theta\in\Theta} \mathcal{L}(\theta, \boldsymbol{\lambda})$. Under Assumptions 3.1 to 3.4, for a fixed tolerance $r > 0$, the iterates generated by Algorithm 1 converge to a neighborhood of the dual problem D_ϵ^* in at most $T_r = \mathcal{O}(1/r)$ steps, i.e.,*

$$d(\boldsymbol{\lambda}^{T_r}) \geq D_\epsilon^* - \delta - \frac{\eta_D}{2}B^2 - r, \tag{9}$$

4 EXPERIMENTS

In this section, we evaluate our formulation (CFL) against the competitors on various FL scenarios at the presence of class imbalance. Our results highlight the benefits of our approach especially when the local data distributions are severely heterogeneous and there is a significant mismatch between local and global imbalance. To ensure a fair comparison, we use Fed-Avg as the base optimizer in the current experiment for *all formulations*: the standard FL objective in Eq.(2), the proposed formulation in Eq.(CFL), `Ratio-Loss` (Wang et al., 2021a), and `Focal-Loss` (Lin et al., 2017). We emphasize that the novelty of our work lies in the new constrained FL formulation (CFL) and is orthogonal to how the formulation is solved. We now describe the datasets and models used in our experiments with more details provided in Appendix A.
Datasets Three benchmark datasets are used in our experiments with the default train/test splits,

Imbalance ratio	Dataset	Level of heterogeneity	Baseline (Eq.(2))	CLIMB **(this work)**	Ratio-Loss	Focal-Loss
$\rho = \mathbf{20}$	CIFAR10		1 minority class out of 10 total classes			
		$\alpha = 0.1$	0.0532 (0.6754)	**0.2080 (0.6829)**	0.1140 (0.6727)	0.0450 (0.6445)
		$\alpha = 0.2$	0.137 (0.7121)	**0.3230 (0.7121)**	0.1790 (0.7037)	0.0430 (0.6914)
			3 minority classes out of 10 total classes			
		$\alpha = 0.1$	0 (0.5669)	**0.2810 (0.6031)**	0 (0.5746)	0 (0.6565)
		$\alpha = 0.2$	0.1279 (0.7098)	**0.3240 (0.7167)**	0.1790 (0.7054)	0.0552 (0.6905)
	MNIST		1 minority class out of 10 total classes			
		$\alpha = 0.1$	0.6889 (0.9375)	**0.8552 (0.9556)**	0.8472 (0.9544)	0.6186 (0.9278)
		$\alpha = 0.2$	0.7925 (0.9540)	**0.8748 (0.9616)**	0.8052 (0.9555)	0.7784 (0.9479)
			3 minority classes out of 10 total classes			
		$\alpha = 0.1$	0.3425 (0.8260)	**0.6987 (0.9158)**	0.4134 (0.8484)	0.1944 (0.7938)
		$\alpha = 0.2$	0.4720 (0.8596)	**0.7290 (0.9153)**	0.6717 (0.9063)	0.4602 (0.8654)
$\rho =$**10**	CIFAR10		1 minority class out of 10 total classes			
		$\alpha = 0.1$	0.2058 (0.6841)	**0.3629 (0.7041)**	0.2164 (0.6839)	0.0414 (0.6543)
		$\alpha = 0.2$	0.1813 (0.7113)	**0.3743 (0.7312)**	0.2657 (0.7083)	0.1347 (0.6911)
			3 minority classes out of 10 total classes			
		$\alpha = 0.1$	0.0492 (0.5933)	**0.2280 (0.6358)**	0.0315 (0.5825)	0 (0.5548)
		$\alpha = 0.2$	0.1064 (0.6380)	**0.2734 (0.6696)**	0.0993 (0.6211)	0.0298 (0.5982)
	MNIST		1 minority class out of 10 total classes			
		$\alpha = 0.1$	0.8473 (0.9534)	**0.9305 (0.9584)**	0.8851 (0.9558)	0.8469 (0.9470)
		$\alpha = 0.2$	0.8962 (0.9634)	**0.9239 (0.9657)**	0.8798 (0.9615)	0.8953 (0.9586)
			3 minority classes out of 10 total classes			
		$\alpha = 0.1$	0.6045 (0.8906)	**0.8084 (0.9356)**	0.6981 (0.9119)	0.4690 (0.8670)
		$\alpha = 0.2$	0.7272 (0.9190)	**0.8195 (0.9392)**	0.7663 (0.9320)	0.6961 (0.9099)

Table 1: The minority class testing accuracy and the overall testing accuracy (the quantity in the parentheses) after 5000 communication rounds. If there are multiple minority classes, we report the worst of them. Here N, the number of devices, is 500. The base FL solver is Fed-Avg with *partial-participation*: 100 devices participate in every communication round.

which are MNIST (LeCun et al., 1998), CIFAR10 (Krizhevsky et al., 2009) and Fashion-MNIST (Xiao et al., 2017). The results on the last dataset are deferred to the appendix due to space limitation.
Heterogeneity. We generate heterogeneity in the local data distributions according to the strategy from (Karimireddy et al., 2020; Hsu et al., 2019): Let $\alpha \in [0, 1]$ be some constant that determines the level of heterogeneity. For a fixed α, we divide the dataset among $N = 100$ (moderate) or $N = 500$ (massive) clients as follows: for we allocate to each client a portion of α i.i.d. data and the remaining portion of $(1 - \alpha)$ by sorting according to label. In our appendix, we also consider the Dirichlet type heterogeneous data allocation scheme which is wildly used in the literature of Federated Learning, for example (Hsu et al., 2019; Acar et al., 2020).
Data Imbalance. We simulate the phenomenon of class imbalance by removing data belong to the minority classes: Observe that the datasets included in our experiments both have 10 perfectly balanced classes. For the minority class(es), we retain only $1/\rho$ portion of the corresponding data. Here, $\rho \geq 1$ the ratio between the numbers of data in the majority class and in the minority class

and is termed the *imbalance ratio*. For example, when there are 3 minority classes with $\rho = 10$, 90% of the data belong to classes 0, 1, 2 (without loss of generality) are manually removed. In our experiments, we consider the setting of 1 or 3 minority classes and we take $\rho = 5, 10, 20$.
Models We follow the choice of model architectures in (Acar et al., 2020; McMahan et al., 2017). Specifically, we use a 2 hidden layer fully-connected neural network for MNIST, where the numbers of neurons are (128, 128). For CIFAR10, we use a CNN model consisting of 2 convolutional layers with 64 5 × 5 filters followed by 2 fully connected layers with 394 and 192 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.

4.1 Results Summary

To evaluate the effectiveness of an approach against the challenge of class imbalance, one needs to take into consideration both the performance on the minority class(es) and the average performance on all the classes. We report both quantities after sufficient communication rounds (5000 rounds for CIFAR10 and 1000 rounds for MNIST) under various experiment settings in Tables 1 and 2. In every cell of these tables, the quantity above denote the minority class testing accuracy and the quantity in the parentheses is the average performance on all the classes. We also considered the case where there are multiple minority classes and we report the worst accuracy among the minority classes. Our approach outperforms previous arts in all cases, often by a large margin.
Imbalance Ratio and Number of Minority Classes. The imbalance ratio ρ and the number of minority classes are two important quantities to measure the difficulty of a class imbalance problem. Under all the included choices, `CLIMB` consistently beats previous arts in both minority class testing accuracy and average performance on all the classes. Therefore, we conclude that `CLIMB` is able to boost the performance on the minority class *without* compromising the performance on the other classes. This is a rare merit when addressing the class imbalance problem. In fact, methods like `Ratio-Loss`is able to improve the testing accuracy on the minority class, but it also sacrifices the performance on the rest classes, leading to an inferior average testing accuracy.
Level of Heterogeneity. We test the performance of `CLIMB` under different levels of heterogeneity and observe that `CLIMB` outperforms the included methods considerably in all settings. It has the biggest advantage over the competitors in the most heterogeneous setting, $\alpha = 0$. There are situations that existing methods completely fail in the minority class with less than 10% accuracy, but `CLIMB` is still able to correctly classify most of the minority data, e.g. see Table 2$\rightarrow \rho = 5 \rightarrow$ MNIST$\rightarrow$ 3 minority classes$\rightarrow \alpha = 0$.
Moderate vs. Large Number of Devices An important goal of FL is to exploit the computational resources of the IoT devices. Hence, the scalability to a large number of devices is a critical property of an FL method. We hence test `CLIMB` on both moderate ($N = 100$, see Table 2) and massive devices ($N = 500$, see Table 1) settings. To make things more practical, we instantiate the subroutine *ClientUpdate* using Fed-Avg with the partial-participation scheme in the massive device setting. Specifically, 100 devices participate model training after every Fed-Avg global communication round. We clearly observe the advantage of `CLIMB` in all of these setups.

Conclusion

In this paper we proposed a novel agnostic constrained learning formulation to tackle the problem of class imbalance in the Federated Learning setting. By introducing constraints in the learning procedure we enforced the performance to be similar in all clients, thus accounting for the class imbalances. In terms of privacy protection, our formulation requires no further information than the standard FL objective to be collected in the server and it never estimates the data composition as opposed to all previous approaches. Moreover, compared with previous arts which are usually heuristic based, our approach is principled as it is purely optimization based and can be efficiently solved via the proposed meta-algorithm `CLIMB`, yielding major practical benefits. Our extensive empirical study showcases the superiority of proposed constrained formulation over previous arts.

Acknowledgments

This research is supported by AFOSR Award 19RT0726, NSF HDR TRIPODS award 1934876, NSF award CPS-1837253, NSF award CIF-1910056, NSF CAREER award CIF-1943064, and NSF award CCF-2112665.

Imbalance ratio	Dataset	Level of heterogeneity	Baseline (Eq.(2))	CLIMB **(this work)**	Ratio-Loss	Focal-Loss
$\rho = \mathbf{10}$	CIFAR10		1 minority class out of 10 total classes			
		$\alpha = 0.0$	0.0229 (0.5734)	**0.5575** **(0.6076)**	0 (0.4836)	0 (0.4205)
		$\alpha = 0.1$	0.2753 (0.7143)	**0.5054** **(0.7246)**	0.2929 (0.6951)	0.2284 (0.6860)
		$\alpha = 0.2$	0.2988 0.7348	**0.4689** **(0.7511)**	0.3825 (0.7329)	0.2618 0.7249
			3 minority classes out of 10 total classes			
		$\alpha = 0.0$	0.0402 (0.5534)	**0.2756** **(0.5598)**	0 (0.4678)	0 (0.4527)
		$\alpha = 0.1$	0.1316 (0.6189)	**0.3399** **(0.6637)**	0.0690 (0.615)	0.0408 (0.5976)
		$\alpha = 0.2$	0.2566 (0.6659)	**0.3292** **(0.6983)**	0.1916 (0.6504)	0.1213 (0.6346)
	MNIST		1 minority class out of 10 total classes			
		$\alpha = 0.0$	0.3092 (0.8630)	**0.9175** **(0.9341)**	0.4650 (0.8630)	0.3078 (0.8556)
		$\alpha = 0.1$	0.8597 (0.9586)	**0.9428** **(0.9675)**	0.9189 (0.9648)	0.8348 (0.9529)
		$\alpha = 0.2$	0.8750 (0.9640)	**0.9377** **(0.9715)**	0.9283 (0.9706)	0.8882 (0.9631)
			3 minority classes out of 10 total classes			
		$\alpha = 0.0$	0.0153 (0.7133)	**0.8029** **(0.9115)**	0.0631 0.7222	0.0717 0.7401
		$\alpha = 0.1$	0.7263 (0.9189)	**0.8828** **(0.9522)**	0.7870 (0.9341)	0.6674 (0.9069)
		$\alpha = 0.2$	0.7950 (0.9352)	**0.8004** **(0.9416)**	0.7801 (0.9364)	0.7785 (0.9274)
$\rho =$**5**	CIFAR10		1 minority class out of 10 total classes			
		$\alpha = 0.0$	0.2892 (0.6382)	**0.5987** **(0.6468)**	0.0942 (0.5506)	0.1491 (0.5631)
		$\alpha = 0.1$	0.4101 (0.7186)	**0.6075** **(0.7351)**	0.4011 (0.7008)	0.3558 (0.6994)
		$\alpha = 0.2$	0.5335 (0.7536)	**0.6063** **(0.7556)**	0.5054 (0.7427)	0.4359 (0.7380)
			3 minority classes out of 10 total classes			
		$\alpha = 0.0$	0.0313 (0.4742)	**0.3813** **(0.5639)**	0.0512 (0.5108)	0 (0.4514)
		$\alpha = 0.1$	0.5135 (0.6636)	**0.6420** **(0.6977)**	0.4600 (0.6632)	0.4213 (0.6384)
		$\alpha = 0.2$	0.5251 (0.6960)	**0.6328** **(0.7202)**	0.5751 (0.6883)	0.5311 (0.6749)
	MNIST		1 minority class out of 10 total classes			
		$\alpha = 0.0$	0.7245 (0.8953)	**0.9154** **(0.9224)**	0.8020 (0.9016)	0.7429 (0.8992)
		$\alpha = 0.1$	0.9378 (0.9666)	**0.9582** **(0.9693)**	0.9286 (0.9651)	0.9408 (0.9624)
		$\alpha = 0.2$	0.9448 (0.9711)	**0.9670** **(0.9734)**	0.9561 (0.9733)	0.9481 (0.9686)
			3 minority classes out of 10 total classes			
		$\alpha = 0.0$	0.0160 (0.7722)	**0.8744** **(0.9137)**	0 (0.7370)	0.0544 (0.7826)
		$\alpha = 0.1$	0.8294 (0.9422)	**0.9217** **(0.9606)**	0.8520 (0.9501)	0.7987 (0.9361)
		$\alpha = 0.2$	0.8557 (0.9536)	**0.8818** **(0.9595)**	0.8730 (0.9536)	0.8321 (0.9442)

Table 2: The minority class testing accuracy and the overall testing accuracy (the quantity in the parentheses) after sufficiently many communication rounds. If there are multiple minority classes, we report the worst of them. Here N, the number of devices, is 100. The base FL solver is Fed-Avg with *full-participation*: all devices participate in every communication round.

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.

Rangachari Anand, Kishan G Mehrotra, Chilukuri K Mohan, and Sanjay Ranka. An improved algorithm for neural network classification of imbalanced training sets. *IEEE Transactions on Neural Networks*, 4(6):962–969, 1993.

Dimitri Bertsekas. *Convex optimization theory*. Athena Scientific, 2009.

Dimitri Bertsekas. *Convex optimization algorithms*. Athena Scientific, 2015.

Stephen Boyd and Almir Mutapcic. Subgradient methods.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

Luiz Chamon and Alejandro Ribeiro. Probably approximately correct constrained learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Luiz FO Chamon, Santiago Paternain, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained learning with non-convex losses. *arXiv preprint arXiv:2103.05134*, 2021.

Philip K Chan, Wei Fan, Andreas L Prodromidis, and Salvatore J Stolfo. Distributed data mining in credit card fraud detection. *IEEE Intelligent Systems and Their Applications*, 14(6):67–74, 1999.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Olivia Choudhury, Yoonyoung Park, Theodoros Salonidis, Aris Gkoulalas-Divanis, Issa Sylla, et al. Predicting adverse drug reactions on distributed health data using federated learning. In *AMIA Annual symposium proceedings*, volume 2019, pp. 313. American Medical Informatics Association, 2019.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

Moming Duan, Duo Liu, Xianzhang Chen, Yujuan Tan, Jinting Ren, Lei Qiao, and Liang Liang. Astraea: Self-balancing federated learning for improving classification accuracy of mobile deep learning applications. In *2019 IEEE 37th international conference on computer design (ICCD)*, pp. 246–254. IEEE, 2019.

Jack Goetz, Kshitiz Malik, Duc Bui, Seungwhan Moon, Honglei Liu, and Anuj Kumar. Active federated learning. *arXiv preprint arXiv:1909.12641*, 2019.

Hongyu Guo and Herna L Viktor. Learning from imbalanced data sets with boosting and data generation: the databoost-im approach. *ACM Sigkdd Explorations Newsletter*, 6(1):30–39, 2004.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5375–5384, 2016.

Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

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*, pp. 5132–5143. PMLR, 2020.

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

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Geun Hyeong Lee and Soo-Yong Shin. Federated learning on clinical benchmark data: Performance assessment. *J Med Internet Res*, 22(10):e20891, Oct 2020. ISSN 1438-8871. doi: 10.2196/20891. URL http://www.jmir.org/2020/10/e20891/.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data, 2020.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Charles X Ling and Victor S Sheng. Cost-sensitive learning and the class imbalance problem. *Encyclopedia of machine learning*, 2011:231–235, 2008.

Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2008.

Guodong Long, Yue Tan, Jing Jiang, and Chengqi Zhang. *Federated Learning for Open Banking*, pp. 240–254. Springer International Publishing, Cham, 2020. ISBN 978-3-030-63076-8. doi: 10.1007/978-3-030-63076-8_17. URL https://doi.org/10.1007/978-3-030-63076-8_17.

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 (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Samira Pouyanfar, Yudong Tao, Anup Mohan, Haiman Tian, Ahmed S Kaseb, Kent Gauen, Ryan Dailey, Sarah Aghajanzadeh, Yung-Hsiang Lu, Shu-Ching Chen, et al. Dynamic sampling in convolutional neural networks for imbalanced data classification. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, pp. 112–117. IEEE, 2018.

Alejandro Ribeiro. Optimal resource allocation in wireless communication and networking. *EURASIP Journal on Wireless Communications and Networking*, 2012(1):1–19, 2012.

Abhijit Guha Roy, Shayan Siddiqui, Sebastian Pölsterl, Nassir Navab, and Christian Wachinger. Braintorrent: A peer-to-peer environment for decentralized federated learning. *arXiv preprint arXiv:1905.06731*, 2019.

Yanmin Sun, Mohamed S Kamel, Andrew KC Wong, and Yang Wang. Cost-sensitive boosting for classification of imbalanced data. *Pattern recognition*, 40(12):3358–3378, 2007.

Fabio Tardella. A new proof of the lyapunov convexity theorem. *SIAM journal on control and optimization*, 28(2):478–481, 1990.

Jason Van Hulse, Taghi M Khoshgoftaar, and Amri Napolitano. Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th international conference on Machine learning*, pp. 935–942, 2007.

Hao Wang, Zakhary Kaplan, Di Niu, and Baochun Li. Optimizing federated learning on non-iid data with reinforcement learning. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pp. 1698–1707. IEEE, 2020a.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization, 2020b.

Lixu Wang, Shichao Xu, Xiao Wang, and Qi Zhu. Addressing class imbalance in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10165–10173, 2021a.

Shoujin Wang, Wei Liu, Jia Wu, Longbing Cao, Qinxue Meng, and Paul J Kennedy. Training deep neural networks on imbalanced data sets. In *2016 international joint conference on neural networks (IJCNN)*, pp. 4368–4374. IEEE, 2016.

Yi Wang, Imane Lahmam Bennani, Xiufeng Liu, Mingyang Sun, and Yao Zhou. Electricity consumer characteristics identification: A federated learning approach. *IEEE Transactions on Smart Grid*, 12(4):3637–3647, 2021b. doi: 10.1109/TSG.2021.3066577.

Qiong Wu, Xu Chen, Zhi Zhou, and Junshan Zhang. Fedhome: Cloud-edge based personalized federated learning for in-home health monitoring. *IEEE Transactions on Mobile Computing*, pp. 1–1, 2020. doi: 10.1109/TMC.2020.3045266.

Wenchao Xia, Tony QS Quek, Kun Guo, Wanli Wen, Howard H Yang, and Hongbo Zhu. Multi-armed bandit-based client scheduling for federated learning. *IEEE Transactions on Wireless Communications*, 19(11):7108–7123, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Miao Yang, Akitanoshou Wong, Hongbin Zhu, Haifeng Wang, and Hua Qian. Federated learning with class imbalance reduction, 2020.

Wensi Yang, Yuhang Zhang, Kejiang Ye, Li Li, and Cheng-Zhong Xu. Ffd: A federated learning based method for credit card fraud detection. In Keke Chen, Sangeetha Seshadri, and Liang-Jie Zhang (eds.), *Big Data – BigData 2019*, pp. 18–32, Cham, 2019. Springer International Publishing. ISBN 978-3-030-23551-2.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning*, pp. 7252–7261. PMLR, 2019.

Xinwei Zhang, Mingyi Hong, Sairaj Dhople, Wotao Yin, and Yang Liu. Fedpd: A federated learning framework with optimal rates and adaptivity to non-iid data. *arXiv preprint arXiv:2005.11418*, 2020.

A Additional Details of the Experiments

A.1 Choice of Hyperparameters

Implementation of *ClientUpdate*. As discussed in Section 3.2 and Algorithm 1, the subroutine *ClientUpdate* needs to be instantiated in order to carry out the actual computation of `CLIMB`. While in principle one can choose *ClientUpdate* to be any existing FL solver, in our paper we use Fed-Avg. Specifically, in every call to *ClientUpdate*, we run Fed-Avg with 5 communication rounds. In the moderate device setting, i.e. $N = 100$, all devices participate the training procedure in every round, while in the massive device setting, i.e. $N = 500$, we randomly (without replacement) select 100 devices to participate in training.

After receiving the most recent consensus model, each device takes 25 local SGD steps, each of which uses 20% of the local data as a minibatch. The step size is set to 0.05 uniformly in all experiments, which is selected by grid searching in the set $\{0.1, 0.05, 0.02\}$ to ensure the fastest convergence rate but without diverging.

To ensure a fair comparison, in all the included competing formulations, we adopt the same setting for the base FL solver. The only difference is that we set the step size of Fed-Avg in `Focal-Loss` and `Ratio-Loss` to be 0.1, which gives better performance than the choice of 0.05 for these two formulations.

Choice of the dual learning rate η_D and the tolerance parameter ϵ. We choose η_D to be 0.1 for all experiments on CIFAR10 and 2 for all experiments on MNIST. Just like the optimization of any other min-max problems, the dual step size is critical to the stability of the training. We made these choices by grid searching in the set $\{4, 2, 1, 0.5, 0.1, 0.05\}$ to ensure the fastest convergence rate but without diverging.
We choose the tolerance parameter ϵ to be 0.1 for all experiments on CIFAR10 and 0.01 for all experiments on MNIST. We made these choices by grid searching in the set $\{1, 0.1, 0.01, 0.001\}$ to ensure the fastest convergence rate but without diverging.

A.2 Deep Learning Training Techniques

Training the deep neural network is known to be tricky. We list all the techniques we used in our experiments to ensure the reproducibility of our results.

- **Data augmentation.** For all the experiments on CIFAR10, we use random horizontal flip and random crop with size 32 and padding 4 to transform the minibatch used in each SGD step. For MNIST, no data augmentation is used.
- **Weight decay.** For all experiments, we use weight decay with parameter 0.001.
- **Gradient clipping.** We use gradient clipping for all experiment setup with maximum gradient norm set to 5.

B Additional Experiments

B.1 Dirichlet Heterogeneity

To showcase that `CLIMB` enjoys a superior performance over previous arts beyond the "sort and allocate" type heterogeneity setting (the one used in our experiment section), we conduct experiments where the heterogeneity of the client data distributions are generated according to a Dirichlet distribution (Yurochkin et al., 2019). Such a choice of heterogeneity generating process is widely used in the literature of Federated Learning, for example (Hsu et al., 2019; Acar et al., 2020). Our implementation of the Dirichlet distribution exactly follows the implementation of (Acar et al., 2020), which is available in the Github repository `https://github.com/alpemreacar/FedDyn`. In Figure 3, we plot the clients' data composition under the Dirichlet type heterogeneity, which is very different from the "sort and allocate" type heterogeneity setting used in our experiments. We conduct experiments under this drastically different type of heterogeneity to showcase that our approach has clear advantages under a broad settings.

Imbalance ratio	Dataset	Level of heterogeneity	Baseline (Eq.(2))	CLIMB **(this work)**	Ratio-Loss	Focal-Loss
$\rho = \mathbf{10}$	CIFAR10		1 minority class out of 10 total classes			
		Dirichlet (0.3)	0.1967 (0.7512)	**0.3745 (0.7739)**	0.1446 (0.7471)	0.0237 (0.7278)
		Dirichlet (10.0)	0.1938 (0.7690)	**0.2773 (0.7739)**	0.2071 (0.7715)	0.1250 0.7606
			3 minority classes out of 10 total classes			
		Dirichlet (0.3)	0.1729 (0.6795)	**0.3359 (0.7045)**	0.1991 (0.6621)	0.0 (0.6287)
		Dirichlet (10.0)	0.1154 (0.6611)	**0.2886 (0.7119)**	0.1621 (0.6695)	0.0980 0.6522
	Fashion-MNIST		1 minority class out of 10 total classes			
		Dirichlet (0.3)	0.2921 (0.8710)	**0.4960 (0.8799)**	03977. (0.8614)	0.2112 (0.8499)
		Dirichlet (10.0)	0.2890 (0.8666)	**0.4662 (0.8842)**	0.3968 (0.8760)	0.2067 (0.8562)
			3 minority classes out of 10 total classes			
		Dirichlet (0.3)	0.2303 (0.7625)	**0.5657 (0.8437)**	0.098 (0.7294)	0.0 (0.6899)
		Dirichlet (10.0)	0.2441 (0.7579)	**0.5920 (0.8521)**	0.2464 (0.7994)	0.1241 (0.7280)

Table 3: Performance of CLIMB under the Dirichlet type heterogeneity. The minority class testing accuracy and the overall testing accuracy (the quantity in the parentheses) after sufficiently many communication rounds. If there are multiple minority classes, we report the worst of them. Here N, the number of devices, is 100. The base FL solver is Fed-Avg with *full-participation*: all devices participate in every communication round.

In Table 3, we report the results in the highly heterogeneous, i.e. Dirichlet distribution with parameter 0.3 (see Figure) and moderate heterogeneous (Dirichlet(10.0)) settings, on the CIFAR10 dataset. We can observe that CLIMB has clear advantages in terms of both the minority class accuracy and the overall accuracy.

Moreover, as suggested by the reviewer, we include a new dataset Fashion-MNIST (Xiao et al., 2017) in Table 3. Similar to MNIST, Fashion-MNIST consists of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28x28 grayscale image, associated with a label from 10 classes. For Fashion-MNIST, we did not perform data augmentation and we use the same NN training technique as described in section A. In terms of hyperparameter setting, we retain the choice for CIFAR10, with the only difference that the dual learning rate for CLIMBis increased from 0.1 to 0.3.

B.2 Performance of CLIMB under Various Heterogeneous Levels

As suggested by the reviewer, we perform the ablation study on α, the parameter that controls the heterogeneity level. We focus on the results of the CIFAR10 dataset with the imbalance ratio ρ set to 10. We can observe CLIMB has the biggest advantage in the most heterogeneous setting when α is close to 0. The performance difference between CLIMB gradually decays as α increases and when $\alpha = 0.5$, the Ratio-Loss proposed by (Wang et al., 2021a) outperforms CLIMB. Consequently, CLIMB is suitable in the highly heterogeneous setting which is usually encountered in practical settings.

B.3 Convergence Curve

We plot the convergence curves for the minority testing accuracy and overall testing accuracy under different levels of heterogeneity settings on the CIFAR10 dataset in Figure 1. We can observe that CLIMB has clear advantages on both results in the included settings

Imbalance ratio	Dataset	Level of heterogeneity	Baseline (Eq.(2))	CLIMB **(this work)**	Ratio-Loss	Focal-Loss
$\rho = \mathbf{10}$	CIFAR10		1 minority class out of 10 total classes			
		$\alpha = 0.0$	0.0229 (0.5734)	**0.5575 (0.6076)**	0 (0.4836)	0 (0.4205)
		$\alpha = 0.1$	0.2753 (0.7143)	**0.5054 (0.7246)**	0.2929 (0.6951)	0.2284 (0.6860)
		$\alpha = 0.2$	0.2988 0.7348	**0.4689 (0.7511)**	0.3825 (0.7329)	0.2618 (0.7249)
		$\alpha = 0.3$	0.3357 (0.7511)	**0.4425 (0.7544)**	0.4312 (0.7534)	0.3304 (0.7455)
		$\alpha = 0.4$	0.3652 (0.7610)	**0.4046 (0.7634)**	0.3948 (0.7627)	0.3239 0.(7526)
		$\alpha = 0.5$	0.4127 (0.7714)	0.4369 **(0.7774)**	**0.5015** (0.7734)	0.3289 (0.7639)

Table 4: Ablation study on α. The minority class testing accuracy and the overall testing accuracy (the quantity in the parentheses) after sufficiently many communication rounds. If there are multiple minority classes, we report the worst of them. Here N, the number of devices, is 100. The base FL solver is Fed-Avg with *full-participation*: all devices participate in every communication round.

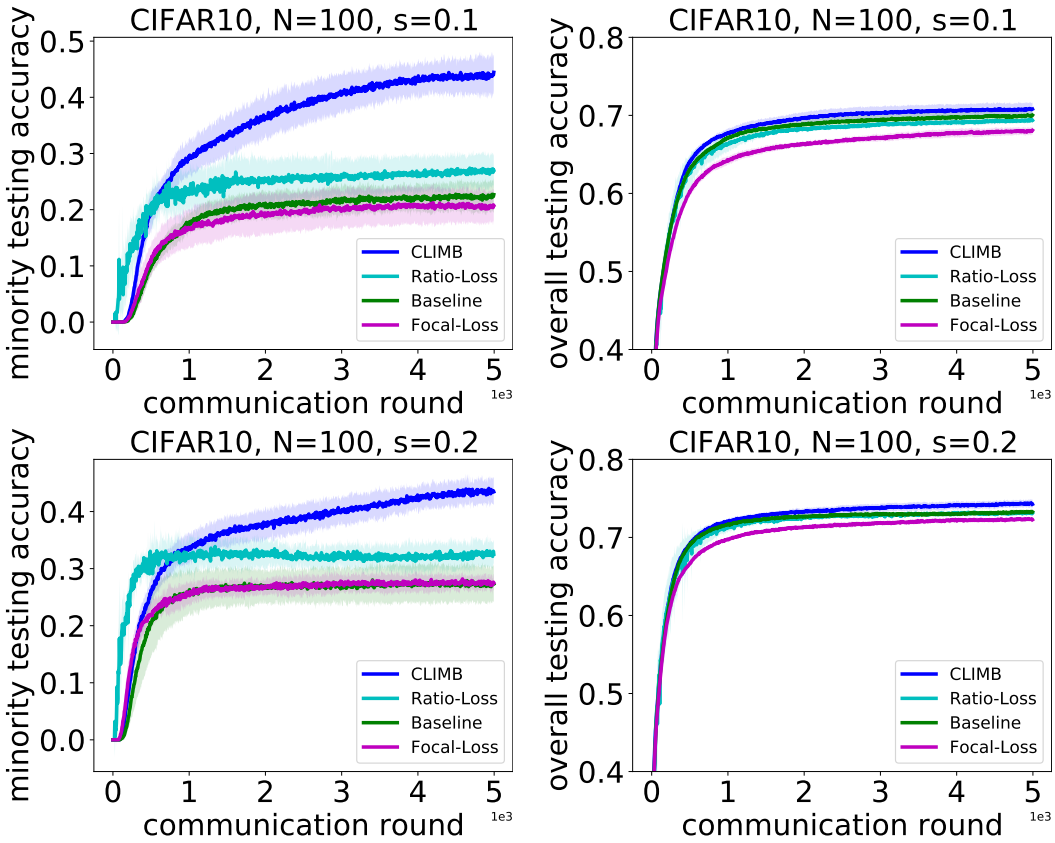


Figure 1: Convergence curve. The experiment setup follows the one in Table 2.

B.4 Performance of FedPD

We report the results of using FedPD (Zhang et al., 2020) which is equivalent to FedDyn (Acar et al., 2020) as base FL solver to solve the standard FL objective (2). We observe that while FedPD is more efficient than FedAvg at the early stage of the training, but is less stable than FedAvg and we hence pick FedAvg to be the base FL solver for CLIMB. Our implementation of FedPD follows exactly

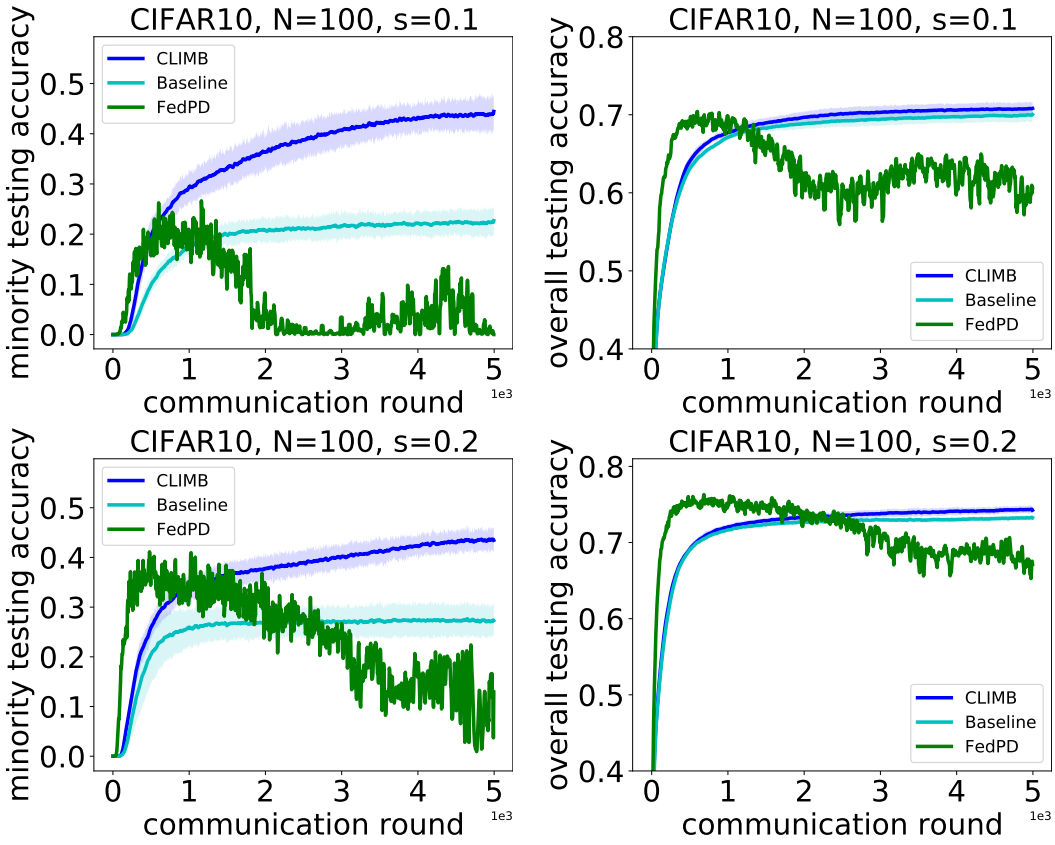


Figure 2: Convergence curve. The experiment setup follows the one in Table 2.

the code of the original implementation by Acar et al. (2020) and the hyperparameter η of FedPD is set to 10 which gives the best performance in our experience.

C PROOFS

C.1 PROOF OF THEOREM 3.1

Lemma C.1 *The non-parametric problem has zero duality gap, i.e.*

$$\hat{D}_\epsilon^* = \max_{\boldsymbol{\lambda}\in\mathbb{R}_+^N} \min_{\hat{\phi}\in\hat{\mathcal{H}}} \mathcal{L}(\phi,\lambda) = \min_{\hat{\phi}\in\hat{\mathcal{H}}} \max_{\boldsymbol{\lambda}\in\mathbb{R}_+^N} \mathcal{L}(\phi,\lambda) = \hat{P}_\epsilon^*. \tag{10}$$

Proof C.1 *This proof follows the lines of (Ribeiro, 2012; Chamon et al., 2021), which exploits the fact that the perturbation set of the primal problem is convex by the Lyapunov convexity theorem (Tardella, 1990), and connects the Supporting Hyperplane Theorem (Bertsekas, 2009) with the dual problem.*

To begin with, recall that the dual problem $\hat{D}_\epsilon^*$ is a relaxation of the primal problem $\hat{P}_\epsilon^*$, and thus it is always a lower bound, i.e. $\hat{D}_\epsilon^* \leq \hat{P}_\epsilon^*$ (Boyd & Vandenberghe, 2004; Bertsekas, 2009). In order to show that zero duality holds, it then suffices to show that $\hat{P}_\epsilon^* \leq \hat{D}_\epsilon^*$, in which case $\hat{P}_\epsilon^* = \hat{D}_\epsilon^*$ by the anti-symmetry property of the inequality.

To express problem CFL a in a more succinct way, we express the objective function as $\bar{f}(\hat{\phi}) = \frac{1}{N}\sum_{k=1}^N f_k(\hat{\phi})$, and define the constraint function $g_i(\hat{\phi}) = f_i(\hat{\phi}) - \frac{1}{N}\sum_{k=1}^N f_k(\hat{\phi})$. We define the perturbation set $\mathcal{P}$ of problem (P_ϕ) as follows,

$$\mathcal{P} = \{(p,\mathbf{c}) \in \mathbb{R}^{N+1},\ \textit{with}\ \mathbf{c} = [c_1,\dots,c_N] \big| \exists \hat{\phi} \in \hat{\mathcal{H}}, \bar{f}(\phi) \leq p, g_i(\phi) \leq c_i\} \tag{11}$$

Given Assumption 3.2, the perturbation set $\mathcal{P}$ is a convex set by (Chamon et al., 2021, Lemma 1). Now note that the point $[\hat{P}_\epsilon^*, \boldsymbol{\epsilon}]$ does not belong to the interior of the perturbation set $\mathcal{P}$. This is true as, by definition P^* is the minimum p_0 that satisfies the constraints $\boldsymbol{\epsilon}$ in $\mathcal{P}$. We can now leverage the Supporting Hyperplane Theorem (Bertsekas, 2009, Proposition 1.5.1), to claim that there exists at least one vector $[v_0, \mathbf{v}] \in \mathbb{R}^{N+1}$ such that,

$$v_0 p + \mathbf{v}^T\mathbf{c} \geq v_0 \hat{P}_\epsilon^* + \mathbf{v}^T\boldsymbol{\epsilon}, \forall (p,\mathbf{c}) \in \mathcal{P}. \tag{12}$$

First, note that for (12) to hold it is necessary that the vector $[v_0, \mathbf{v}] \in \mathbb{R}^{N+1}$ is non-negative in all its components. If the supporting vector $[v_0, \mathbf{v}] \in \mathbb{R}^{N+1}$ could take negative values, then there is a value for which the inequality is reversed, thus $[v_0, \mathbf{v}] \in \mathbb{R}^{N+1}$ are non-negative. Second, note that the first element of the perturbation vector must be positive $v_0 > 0$. As if $v_0 = 0$, would imply that $[\hat{P}_\epsilon^*, \boldsymbol{\epsilon}]$ is an interior point of $\mathcal{P}$, and thus there exists a feasible point $\phi^{'}$ with $[P^{'}, \boldsymbol{\epsilon}]$ such that $P^{'} < \hat{P}_\epsilon^*$ contradicting the definition of $\hat{P}_\epsilon^*$. By defining $\bar{\mathbf{v}} = [v_1/v_0, \dots, v_N/v_0]$, we can rewrite (12) as,

$$p + \bar{\mathbf{v}}^T(\mathbf{c} - \boldsymbol{\epsilon}) \geq \hat{P}_\epsilon^*, \forall (p,\mathbf{c}) \in \mathcal{P}. \tag{13}$$

It can now be seen that (13) is nothing but the non-parametric version of the Lagrangian of the Federated Learning problem as stated in (5),

$$\bar{f}(\phi) + \sum_{i=1}^N \bar{v}_i(g_i(\phi) - \boldsymbol{\epsilon}) \geq \hat{P}_\epsilon^*, \forall \hat{\phi} \in \mathcal{P}. \tag{14}$$

We can thus apply the minimum operator on the right hand for every $\hat{v}_i$ as follows

$$\underset{\phi\in\mathcal{P}}{\textit{minimum}}\ \bar{f}(\phi) + \sum_{i=1}^N \bar{v}_i(g_i(\phi) - \boldsymbol{\epsilon}) \geq \hat{P}_\epsilon^*. \tag{15}$$

Thus showing that strong duality holds as we showed that $\hat{D}_\epsilon^* \geq \hat{P}_\epsilon^*$. ■

Proof C.2 (Theorem 3.1) *Given that $\mathcal{H} \subseteq \hat{\mathcal{H}}$, and that the functional problem has zero duality gap (i.e. $\hat{P}^*_\epsilon = \hat{D}^*_\epsilon$) by Lemma C.1, we can write the following inequality,*

$$\hat{P}^*_\epsilon = \hat{D}^*_\epsilon = \max_{\boldsymbol{\lambda} \in \mathbb{R}^N_+} \min_{\hat{\phi} \in \hat{\mathcal{H}}} \mathcal{L}(\hat{\phi}, \boldsymbol{\lambda}) \leq \max_{\boldsymbol{\lambda} \in \mathbb{R}^N_+} \min_{\theta \in \Theta} \mathcal{L}(\theta, \boldsymbol{\lambda}) \leq D^*_\epsilon \tag{16}$$

Now we need to obtain a bound between the parametric problem P^*_ϵ, and the functional version $\hat{P}^*_\epsilon$ of the problem. To this end we define the set of feasible solutions of the problem CFL with respect to the perturbation as follows,

$$\mathcal{S}_\epsilon = \{h : f_i(h) - \bar{f}(h) \leq \epsilon \quad \forall i \in [1, \dots, N], h(\mathcal{X}) \to \mathbb{R}^C\} \tag{17}$$

We also define the set of optimal solutions of the non-parametric problem as follows,

$$\begin{aligned} \{\hat{\phi}^*_\epsilon\} = \underset{\hat{\phi} \in \hat{\mathcal{H}}}{argmin}\ & \bar{f}(\hat{\phi}) \\ \textit{s.t. } & f_i(\hat{\phi}) - \bar{f}(\hat{\phi}) \leq \epsilon, \forall i \in [1, \dots, N]. \end{aligned} \tag{18}$$

Note that as $\{\hat{\phi}^*_\epsilon\}$ are solutions to the problem CFL, they are feasible for constraint ϵ and thus $\{\hat{\phi}^*\} \in \mathcal{S}_\epsilon \cap \hat{\mathcal{H}}$. Given that the problem is feasible with constraint $\epsilon - 2L\xi$, we can now pick one element of the set of optimal solutions, which we will denote $\hat{\phi}^*_{\epsilon-2L\xi}$, and obtain one parametric function from $\mathcal{H}$ which we will denote $\phi(\tilde{\theta}_{\epsilon-2L\xi}, \cdot) \in \mathcal{H}, \tilde{\theta}_{\epsilon-2L\xi} \in \Theta$ that satisfies

$$\sup_{x \in \mathcal{X}} \|\hat{\phi}^*_{\epsilon-2L\xi}(x) - \phi(x, \tilde{\theta}_{\epsilon-2L\xi})\| \leq \xi \tag{19}$$

which exists by assumption 3.3. We know that as the loss function is Lipschitz continuous, and $\phi(\tilde{\theta}, \cdot)$ is close to $\hat{\phi}$ by construction (cf. (19)*), thus the following holds*

$$\|f_i(\hat{\phi}^*_{\epsilon-2L\xi}) - f_i(\tilde{\theta}_{\epsilon-2L\xi})\| \leq L\xi \tag{20}$$

Using (20)*, we can show that $\tilde{\theta}_{\epsilon-2L\xi}$ is feasible for the Constrained Federated Learning problem. Given that $\phi^*_{\epsilon-2L\xi}$ is feasible for the constrained federated learning problem* (CFL) *with constraint $\epsilon - 2L\xi$, then the constraint difference for each constraint i is bounded by,*

$$\begin{aligned} f_i(\hat{\phi}^*_{\epsilon-2L\xi}) - \bar{f}(\hat{\phi}^*_{\epsilon-2L\xi}) &\leq \epsilon - 2L\xi \\ f_i(\tilde{\theta}_{\epsilon-2L\xi}) - \bar{f}(\tilde{\theta}_{\epsilon-2L\xi}) + \Big(f_i(\hat{\phi}^*_{\epsilon-2L\xi}) - \bar{f}(\hat{\phi}^*_{\epsilon-2L\xi})) - (f_i(\tilde{\theta}_{\epsilon-2L\xi}) - \bar{f}(\tilde{\theta}_{\epsilon-2L\xi}))\Big) &\leq \epsilon - 2L\xi \\ f_i(\tilde{\theta}_{\epsilon-2L\xi}) - \bar{f}(\tilde{\theta}_{\epsilon-2L\xi}) - \|f_i(\hat{\phi}^*_{\epsilon-2L\xi}) - \bar{f}(\hat{\phi}^*_{\epsilon-2L\xi})) - (f_i(\tilde{\theta}_{\epsilon-2L\xi}) - \bar{f}(\tilde{\theta}_{\epsilon-2L\xi}))\| &\leq \epsilon - 2L\xi \\ f_i(\tilde{\theta}_{\epsilon-2L\xi}) - \bar{f}(\tilde{\theta}_{\epsilon-2L\xi}) - 2L\xi &\leq \epsilon - 2L\xi \end{aligned} \tag{21}$$

where the last inequality holds as the loss function is L-Lipschitz by assumption (3.1)*, and $\phi(\tilde{\theta}, \cdot)$ satisfies* (19)*. Hence, we can say that $\phi(\tilde{\theta}_{\epsilon-2L\xi}, \cdot) \in \mathcal{S}_\epsilon \cap \mathcal{H}$, and thus $\phi(\tilde{\theta}_{\epsilon-2L\xi}, \cdot)$ is a feasible solution of the constrained federated learning problem with constraint ϵ. Now we can express the value of the duality gap as follows,*

$$P^*_\epsilon \geq D^*_\epsilon \geq \min_{\theta \in \Theta} \mathcal{L}(\theta, \boldsymbol{\lambda}) \geq \min_{\phi \in \hat{\mathcal{H}}} \mathcal{L}(\phi, \boldsymbol{\lambda}) \quad \forall \boldsymbol{\lambda} \in \mathbb{R}^N_+. \tag{22}$$

Where the previous inequality holds as $\mathcal{H} \in \hat{\mathcal{H}}$ by assumption 3.3. Now, given that there $\phi^*_{\epsilon-2L\xi}$ is the solution to the constrained federated learning problem with constraint $\epsilon - 2L\xi$, which attains the value $\hat{P}^*_{\epsilon-2L\xi}$ and that the functional version of the problem has zero-duality gap by Lemma C.1, defining the optimal dual variables as $\boldsymbol{\lambda}^*_{\epsilon-2L\xi}$, it yields

$$P^*_\epsilon \geq D^*_\epsilon \geq \min_{\phi \in \hat{\mathcal{H}}} \mathcal{L}(\phi, \boldsymbol{\lambda}^*_{\epsilon-2L\xi}) + 2L\xi|\boldsymbol{\lambda}^*_{\epsilon-2L\xi}|_1 - 2L\xi|\boldsymbol{\lambda}^*_{\epsilon-2L\xi}|_1 = \hat{P}^*_{\epsilon-2L\xi} - 2L\xi|\boldsymbol{\lambda}^*_{\epsilon-2L\xi}|_1. \tag{23}$$

We can now further compare the value of $\hat{P}^*_{\epsilon-2L\xi}$, with $\bar{f}(\theta_{\epsilon-2L\xi})$ as follows,

$$\hat{P}^*_{\epsilon-2L\xi} = \hat{P}^*_{\epsilon-2L\xi} + \bar{f}(\tilde{\theta}_{\epsilon-2L\xi}) - \bar{f}(\tilde{\theta}_{\epsilon-2L\xi}) \tag{24}$$

$$\geq \bar{f}(\tilde{\theta}_{\epsilon-2L\xi}) - \|\hat{P}^*_{\epsilon-2L\xi} - \bar{f}(\theta_{\epsilon-2L\xi})\| \tag{25}$$

$$\geq P^*_\epsilon - L\xi. \tag{26}$$

Note that the last inequality holds by optimality, i.e. $P^*_\epsilon \leq \bar{f}(\tilde{\theta}_{\epsilon-2L\xi})$, given that $\tilde{\theta}_{\epsilon-2L\xi}$ is a feasible point of the Constrained Learning Problem by (21)*. Combining* (26) *with* (23) *completes the proof.*

C.2 Comments on Theorem 3.1

The bound in Theorem 3.1 quantifies the distance between the Constrained Federated Learning problem (CFL) *and the dual problem $\hat{D}^*_\epsilon$. This bound is controlled by two terms the approximation bound and the perturbation bound. On the one hand, we have the approximation bound term given by $L\xi$. This term is related to the distance between functions in the parametric and non-parametric setting i.e. assumption* (3.3)*. The approximation bound is linked to the fact that the Constrained Federated Learning problem CFL has zero duality gap in the non-parametric domain $\phi \in \hat{\mathcal{H}}$ by Lemma* (C.1)*. Therefore, the value of the Constrained Federated Learning problem in the parametric case $\hat{P}^*_\epsilon$ should be at most $L\xi$ away, given that the loss function ℓ is L-lipschitz by assumption 3.1. On the other hand, we have the perturbation bound, which is $(2L\xi)\|\boldsymbol{\lambda}^*_{\epsilon-2L\xi}\|$. The important part of this bound is the inclusion of the norm of the optimal dual variable of the perturbed problem with constraints $\epsilon - 2L\xi$. This constant shows up given that we need to find a feasible solution of the non-parametric problem. In order to relate the non-parametric and the parametric feasibility sets, we perturb the non-parametric problem until we assure that the parametric function close to it is feasible in the Constrained Federated Learning Problem CFL. Therefore, this constant $\|\boldsymbol{\lambda}^*_{\epsilon-2L\xi}\|$ can be seen as the Lipschitz constant of the non-parametric Constrained Federated Learning problem with respect to the perturbation, i.e.,*

$$\hat{P}^*_{\epsilon-2L\xi} - \hat{P}^*_\epsilon \leq \|\boldsymbol{\lambda}^*_{\epsilon-2L\xi}\|(\epsilon - (\epsilon - 2L\xi)) \tag{27}$$

$$\leq \|\boldsymbol{\lambda}^*_{\epsilon-2L\xi}\|2L\xi. \tag{28}$$

Note that as constraint $\epsilon - 2L\xi$ is smaller than ϵ, then $\hat{P}^*_\epsilon$ will be smaller or equal than $\hat{P}^*_{\epsilon-2L\xi}$.

C.3 Proof of Theorem 3.2

In order to prove Theorem 3.2, we will first provide the definition of supergradient, and then show that evaluating the constraint slack for each minimizer $\theta_{\boldsymbol{\lambda}}$ is a supergradient. Recall that the dual function is a concave function given that it is a pointwise infimum of a family of affine functions of $\boldsymbol{\lambda}$ (Boyd & Vandenberghe, 2004). The proof that follows is based on the δ-subgradient proof that can be found in (Bertsekas, 2015).

Definition C.1 (δ-Supergradient) *(Bertsekas, 2015) Given the concave dual function $d(\boldsymbol{\lambda})$, and a scalar $\delta > 0$, we say that the vector $g(\boldsymbol{\lambda})$ is a δ-supergradient of $d(\boldsymbol{\lambda})$ at a point $\boldsymbol{\lambda} \in \mathbb{R}^N_+$ if*

$$d(\boldsymbol{\mu}) \leq d(\boldsymbol{\lambda}) + g(\boldsymbol{\lambda})^T(\boldsymbol{\mu} - \boldsymbol{\lambda}) + \delta, \quad \forall \boldsymbol{\mu} \in \mathbb{R}^N. \tag{29}$$

Lemma C.2 (Constraint Slack is a δ-Supergradient) *For every dual variable $\boldsymbol{\lambda}$, given the approximate minimizer $\theta_{\boldsymbol{\lambda}}$ as in Assumption 3.4, then the constraint slacks,*

$$[g(\boldsymbol{\lambda})]_i = f_i(\theta_\lambda) - \bar{f}(\theta_\lambda) - \epsilon \tag{30}$$

are a δ-supergradient of the dual function at point λ, i.e.,

$$d(\boldsymbol{\mu}) \leq d(\boldsymbol{\lambda}) + g(\boldsymbol{\lambda})^T(\boldsymbol{\mu} - \boldsymbol{\lambda}) + \delta, \quad \forall \boldsymbol{\mu} \in \mathbb{R}^N \tag{31}$$

Proof C.3 *From Assumption 3.4 we can express the dual function at any given $\boldsymbol{\lambda}$ as follows,*

$$\mathcal{L}(\theta_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) \leq d(\boldsymbol{\lambda}) + \delta \tag{32}$$

We can then add and subtract the dual function at a given point μ as follows,

$$d(\boldsymbol{\lambda}) \geq \mathcal{L}(\theta_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) + d(\boldsymbol{\mu}) - d(\boldsymbol{\mu}) - \delta \tag{33}$$

Given that $d(\boldsymbol{\mu})$ is the minimum, we can upper bound the dual function by the value at $\mathcal{L}(\theta_{\boldsymbol{\lambda}}, \boldsymbol{\mu})$ to obtain,

$$d(\boldsymbol{\lambda}) \geq d(\boldsymbol{\mu}) + \mathcal{L}(\theta_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) - \mathcal{L}(\theta_{\boldsymbol{\lambda}}, \boldsymbol{\mu}) - \delta \tag{34}$$

$$\geq d(\boldsymbol{\mu}) + \frac{1}{N}\sum_{i=1}^{N}(f_i(\theta_{\boldsymbol{\lambda}}) - \bar{f}(\theta_{\boldsymbol{\lambda}}) - \epsilon)(\lambda_i - \mu_i) - \delta, \tag{35}$$

for all $\boldsymbol{\mu}$.

Proof C.4 (Theorem 3.2) *The following proof is along the lines of other subgradient/supergradient method proofs (Boyd & Mutapcic; Chamon & Ribeiro, 2020; Bertsekas, 2015). Let V^{k+1} be the distance between $\boldsymbol{\lambda}^{k+1}$ and $\boldsymbol{\lambda}^*$,*

$$V^{k+1} = \frac{1}{N}\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^*\|^2 \tag{36}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\|\lambda_i^{k+1} - \lambda_i^*\|^2 \tag{37}$$

We begin by expressing V^{k+1} in terms of the previous iteration as follows,

$$V^{k+1} = \frac{1}{N}\sum_{i=1}^{N}\|[\lambda_i^k + \eta_D(f_i(\theta_\lambda^k) - \bar{f}(\theta_\lambda^k) - \epsilon]_+ - \lambda_i^*\|^2 \tag{38}$$

$$\leq \frac{1}{N}\sum_{i=1}^{N}\|\lambda_i^k + \eta_D(f_i(\theta_\lambda^k) - \bar{f}(\theta_\lambda^k) - \epsilon - \lambda_i^*)\|^2 \tag{39}$$

$$\leq V^k + \frac{1}{N}\sum_{i=1}^{N}\eta_D^2\|f_i(\theta_\lambda^k) - \bar{f}(\theta_\lambda^k) - \epsilon\|^2 + 2\eta_D(f_i(\theta_\lambda^k) - \bar{f}(\theta_\lambda^k) - \epsilon)(\lambda_i^k - \lambda_i^*), \tag{40}$$

Given that ℓ is bounded by B by Assumption 3.1, we can bound the first term of the previous expression by NB^2, and as $\lambda^k \in \mathcal{F}_k$, then

$$V^{k+1} \leq V^k + \eta_D^2 B^2 + 2\sum_{i=1}^{N}\eta_D(f_i(\theta_\lambda^k) - \bar{f}(\theta_\lambda^k) - \epsilon)(\lambda_i^k - \lambda_i^*) \tag{41}$$

We can now bound the previous inequality using the value of D_ϵ^* as follows,

$$V^{k+1} \leq V^k + 2\eta_D(d(\boldsymbol{\lambda}^k) - D_\epsilon^* + \delta + \frac{\eta_D}{2}B^2) \tag{42}$$

Hence, by taking K steps the recursion can be expressed as,

$$V^K \leq V^0 + \sum_{k=0}^{K-1} 2\eta_D(d(\boldsymbol{\lambda}^k) - D_\epsilon^* + \delta + \frac{\eta_D}{2}B^2) \tag{43}$$

Given a precision $r > 0$, we can set the stopping time T_r such that,

$$T_r = \min\{k \quad | \quad d(\boldsymbol{\lambda}^k) - D_\epsilon^* + \delta + \frac{\eta_D}{2}B^2 > -r\} \tag{44}$$

Then, we can conclude that at stopping time T_r,

$$d(\boldsymbol{\lambda}^{T_r}) \geq D_\epsilon^* - \delta - \frac{\eta_D}{2}B^2 - r \tag{45}$$

Notice that $d(\boldsymbol{\lambda}^{T_r}) \leq D^*$ for all dual variables $\boldsymbol{\lambda}^k$. To conclude, notice that as $V^k \geq 0$, and for all $k \leq T_\alpha$ we can expresses (41) as,

$$0 \leq V^0 - T_r 2\eta_D r \tag{46}$$

From where we can conclude that the stopping time is finite $T_r \leq \frac{V_0}{2\eta_D r}$ completing the proof.

■

C.4 COMMENTS ON THEOREM 3.2

In Theorem 3.2 we show the convergence of the dual function $d(\boldsymbol{\lambda}^t)$ to the optimal solution of the dual function, i.e. the dual problem D_ϵ^* up to some error that depends on the capability of the parametric function class ξ, the dual step size η_D, and the error r introduced by the optimization oracle in Assumption 3.4. We can only obtain such a result since the dual function is potentially non-smooth with respect to the dual variable, and hence we can only use the analysis in the literature of sub-gradient optimization. Moreover, since we do not know the actual optimal value of the dual problem, we cannot use the Polyak step size which leads to the dependence on the dual step size η_D in our convergence result in (9). Given the non-convexity of the problem, we would require further assumptions to strengthen our theoretical results to guarantee convergence in the primal variable.

C.5 ERGODIC CONVERGENCE

In this section we will present an alternative proof of convergence. The proof here presented goes along the lines of Chamon et al. (2021). We will prove an ergodic convergence of our algorithm that is near optimal and near feasible. We will still rely in assumptions 3.1, 3.2, 3.3, and 3.4 but in this case the result will not be a last iterate proof as in Theorem 3.2.

Theorem C.1 *Under Assumptions 3.1, 3.2, 3.3, and 3.4 if there exists a feasible point of the Constrained Federated Learning problem CFL with constraint $\epsilon - \beta, \beta > 0$, then the average of T steps of Algorithm 1 verifies*

$$\frac{1}{T}\sum_{t=0}^{T-1} \bar{f}(\theta_t) \leq P_\epsilon^* + \delta + \mathcal{O}(\eta_D) \tag{47}$$

$$\frac{1}{TN}\sum_{t=0}^{T-1} f_i(\theta_t) - \bar{f}(\theta_t) - \epsilon \leq \mathcal{O}\left(\frac{1}{T\eta_D}\right) \quad \forall i \in [1, \dots, N]. \tag{48}$$

In Theorem C.1, we show sub-optimality and near feasibility in an ergodic average. Apart from Assumptions 3.1, 3.2, 3.3, and 3.4, we require a strictly feasible point to exist. In the particular case of the cross-entropy loss, this requirement can be satisfied by a classifier that outputs the uniform distribution for all samples. This way, all losses will be equal. In (47), we show that the objective function is sub-optimal, and this sub-optimality is controlled by the oracle constant δ from Assumption 3.4, and the dual step size η_D. Note that the right hand side of (47) does not depend upon the number of iterations T, this is due to the fact that we assume oracle queries. The near feasibility part (48), is related to the fact that throughout the trajectory, the norm of the dual variables $\boldsymbol{\lambda}$ is bounded. Therefore, at any given time $\boldsymbol{\lambda}_t$ is obtained by the summation of all the sub-gradient steps taken up to time t. This corresponds to the summation of all constrained slackness (i.e. left hand side of (48)).

Proof C.5 (Theorem C.1) *We will start by proving sub-optimality of the average over the iterates. To do so, we can begin with the summation of the objective function and relate it to the Lagrangian as follows,*

$$\frac{1}{T}\sum_{t=1}^{T} \bar{f}(\theta_t) = \frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}(\theta_t, \boldsymbol{\lambda}_t) - \frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N} \lambda_{it} \frac{(f_i(\theta) - \bar{f}(\theta) + \epsilon)}{N} \tag{49}$$

$$\leq \min_{\theta\in\Theta} \frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}(\theta_t, \boldsymbol{\lambda}_t) + \rho - \frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N} \lambda_{it} \frac{(f_i(\theta) - \bar{f}(\theta) + \epsilon)}{N} \tag{50}$$

$$\leq D_\epsilon^* + \rho - \frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N} \lambda_{it} \frac{(f_i(\theta) - \bar{f}(\theta) + \epsilon)}{N} \tag{51}$$

$$\leq P_\epsilon^* + \rho - \frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N} \lambda_{it} \frac{(f_i(\theta) - \bar{f}(\theta) + \epsilon)}{N} \tag{52}$$

Where (50) *holds by Assumption 3.4, inequality* (51) *holds since the dual problem is the maximum over λ, and* (52) *holds since the dual problem is always a lower bound of the primal problem Boyd*

& Vandenberghe (2004). In order to complete the near optimality proof, we need to bound the last term in the previous (52)*. To do so we express the squared norm of the iterates of* $\boldsymbol{\lambda}_t$ *as follows,*

$$\|\boldsymbol{\lambda}_{T+1} - \boldsymbol{\lambda}_0\|^2 = \|[\boldsymbol{\lambda}_t + \eta_D \frac{1}{N}[f_1(\theta_t) - \bar{f}(\theta_t) - \epsilon, \dots, f_N(\theta_t) - \bar{f}(\theta_t) - \epsilon]]_+ - \boldsymbol{\lambda}_0\|^2 \tag{53}$$

$$\leq \|\boldsymbol{\lambda}_t + \eta_D \frac{1}{N}[f_1(\theta_t) - \bar{f}(\theta_t) - \epsilon, \dots, f_N(\theta_t) - \bar{f}(\theta_t) - \epsilon] - \boldsymbol{\lambda}_0\|^2 \tag{54}$$

$$\leq \|\boldsymbol{\lambda}_t\|^2 + 2\eta_D \sum_{i=1}^{N} \lambda_{it} \frac{(f_i(\theta) - \bar{f}(\theta) + \epsilon)}{N} + \frac{\eta_D^2}{N^2} \sum_{i=1}^{N} \|f_i(\theta_t) - \bar{f}_i(\theta_t) - \epsilon\|^2 \tag{55}$$

$$\leq \|\boldsymbol{\lambda}_t\|^2 + 2\eta_D \sum_{i=1}^{N} \lambda_{it} \frac{(f_i(\theta) - \bar{f}(\theta) + \epsilon)}{N} + \frac{\eta_D^2}{N^2}(N2B^2) \tag{56}$$

$$\leq \|\boldsymbol{\lambda}_0\|^2 + 2\eta_D \sum_{t=0}^{T-1} \sum_{i=1}^{N} \lambda_{it-1} \frac{(f_i(\theta) - \bar{f}(\theta) + \epsilon)}{N} + \frac{\eta_D^2 2B^2 T}{N} \tag{57}$$

$$\leq 2\eta_D \sum_{t=0}^{T-1} \sum_{i=1}^{N} \lambda_{it} \frac{(f_i(\theta) - \bar{f}(\theta) + \epsilon)}{N} + \frac{\eta_D^2 2B^2 T}{N} \tag{58}$$

Where (54) *holds since the norm of the difference between two points is larger before projection them to a convex set. Note that the initialization of the dual variable is done with the zero vector i.e.,* $\boldsymbol{\lambda}_0 = [0, \dots, 0] \in \mathbb{R}^N$*. Given that the norm is always positive, we can express the previous inequality as*

$$-\frac{1}{T} \sum_{t=0}^{T-1} \sum_{i=1}^{N} \lambda_{it} \frac{(f_i(\theta) - \bar{f}(\theta) + \epsilon)}{N} \leq \frac{\eta_D B^2}{N} \tag{59}$$

Substituting (59) *into* (52) *completes the near optimality part of the proof.*

To complete the near feasiblity part of the proof, note that as we require the problem to be strictly feasible, then the norm of any of the optimal dual variables is finite i.e. $\|\boldsymbol{\lambda}^*\| < \infty, \forall \boldsymbol{\lambda}^* \in \Lambda^*$*, where* $\Lambda^* = argmax_{\boldsymbol{\lambda}}\, d(\boldsymbol{\lambda})$*. We denote* Λ^* *the set of optimal dual variables of the dual problem. By applying the iterates of the Algorithm 1, at each coordinate* i *we obtain,*

$$\lambda_{iT} \geq \lambda_{iT-1} + \eta_D \frac{1}{N}(f_i(\theta_{T-1}) - \bar{f}(\theta_{T-1}) - \epsilon) \tag{60}$$

$$\geq \lambda_{i0} + \eta_D \frac{1}{N} \sum_{t=0}^{T-1} (f_i(\theta_t) - \bar{f}(\theta_t) - \epsilon) \tag{61}$$

$$\geq \eta_D \frac{1}{N} \sum_{t=0}^{T-1} (f_i(\theta_t) - \bar{f}(\theta_t) - \epsilon) \tag{62}$$

Therefore, by bounding the norm of the dual variable we obtain

$$\frac{1}{TN} \sum_{t=0}^{T-1} (f_i(\theta_t) - \bar{f}(\theta_t) - \epsilon) \leq \frac{\lambda_{it+1}}{\eta_D T} \leq \frac{|\lambda_{it+1} - \lambda_i^*| + \lambda_i^*}{\eta_D T} \leq (\frac{B - D_\epsilon^*}{\beta} + \|\boldsymbol{\lambda}^*\|) \frac{1}{\eta_D T} \tag{63}$$

where the last inequality holds by (Chamon et al., 2021, Lemma 6).

D PRIVACY-PRESERVING IMPLEMENTATION OF CLIMB

As pointed out by one of the reviewers, we acknowledge that a client corresponding to a large dual variable is more likely to possess minority data. However, we emphasize that under the standard federated learning paradigm, CLIMB can be implemented in a privacy-preserving manner: one **cannot** recover the dual variables on the server side.

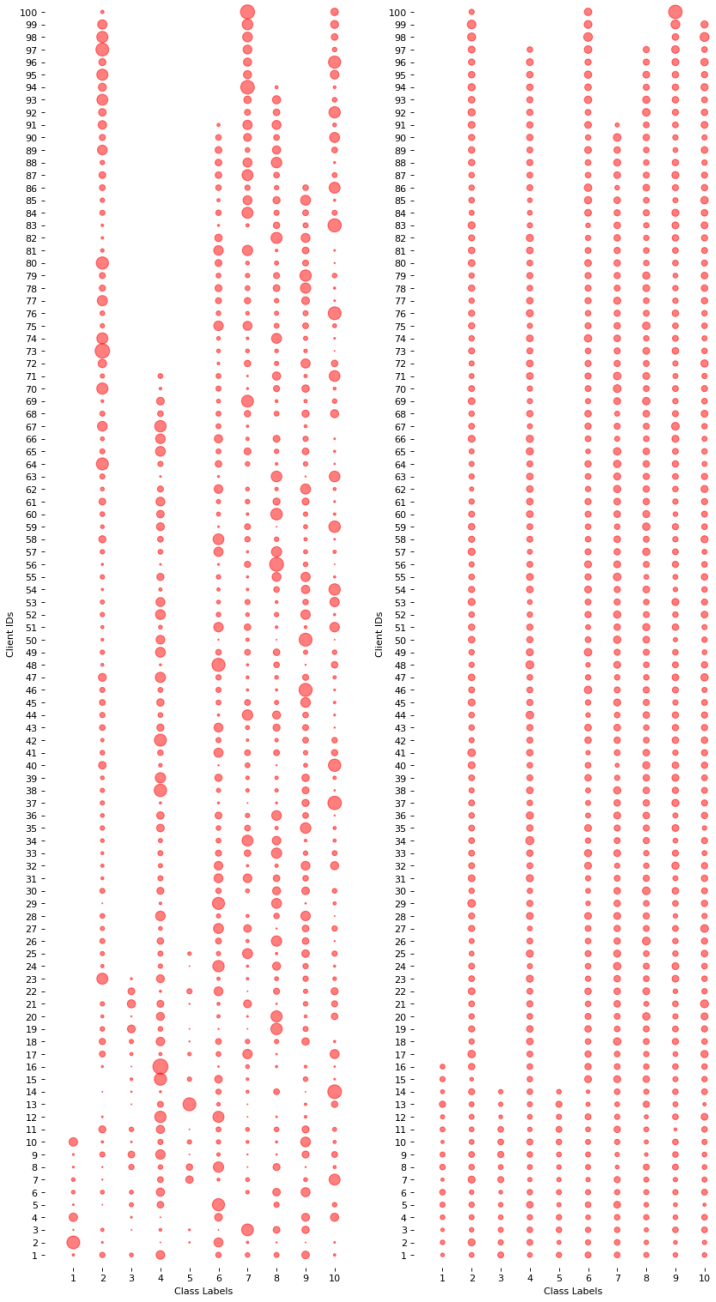
To achieve this goal, first observe that the weight computing step and dual update step of CLIMB (line 4 and line 6 of Algorithm 1 respectively) can be carried out locally as long as the client has access to the global average dual variable $\bar{\lambda}$ and the global average loss $\bar{f}(\theta^{t+1})$ since other terms only involve local information. In order to compute these two terms on the server in a privacy-preserving manner, one needs to address the following simplified problem:

> Assuming that each client privately holds a quantity a_i, how can we compute the global average $\bar{a} = \frac{1}{n}\sum_{i\in[n]} a_i$ without revealing the quatities a_i's to the server?

Assuming that there is a secret key that is available to the clients but not to the server, then with the Homomorphic Encryption technique the clients can

1. encrypt the dual variable λ_i the and local loss $f_i(\theta^{t+1})$,
2. communicate them with the server to compute the average homomorphically (the server cannot decrypt the individual contribution since it does not have the key),
3. receive the encrypted averages from the server and decrypt the received averages locally using the secret key.

In this way, both the average dual variable and the global average loss can be computed without being revealed to the server. The aforementioned scheme is often known as secure multi-party computation and the secret key is usually implemented through a trusted 3rd party (for example see the discussion on page 42 of the monograph Kairouz et al. (2019)).



Dirichlet(0.3) Dirichlet(10.0)

Figure 3: Client data composition under the Dirichlet type heterogeneity with three minority classes (Class Labels $\{1, 3, 5\}$) and Imbalance Ratio $\rho = 10$. The size of the dot represents the frequency of the corresponding class in a simple client. The left figure is generated with hyperparameter 0.3, corresponding to a highly heterogeneous setting, while the right one is generated with hyperparameter 10, corresponding to a moderate setting. We follow the implementation from (Acar et al., 2020), in which a Dirichlet prior is sampled for each client. One client at a time, we sample without replacement according to each client's prior. Once a class runs out of samples, the subsequent clients do not own samples of that class.