Decomposable Submodular Maximization in Federated Setting

Akbar Rafiey [1]

Abstract

Submodular functions, as well as the sub-class of decomposable submodular functions, and their optimization appear in a wide range of applications in machine learning, recommendation systems, and welfare maximization. However, optimization of decomposable submodular functions with millions of component functions is computationally prohibitive. Furthermore, the component functions may be private (they might represent user preference function, for example) and cannot be widely shared. To address these issues, we propose a *federated optimization* setting for decomposable submodular optimization. In this setting, clients have their own preference functions, and a weighted sum of these preferences needs to be maximized. We implement the popular *continuous greedy* algorithm in this setting where clients take parallel small local steps towards the local solution and then the local changes are aggregated at a central server. To address the large number of clients, the aggregation is performed only on a subsampled set. Further, the aggregation is performed only intermittently between stretches of parallel local steps, which reduces communication cost significantly. We show that our federated algorithm is guaranteed to provide a good approximate solution, even in the presence of above cost-cutting measures. Finally, we show how the federated setting can be incorporated in solving fundamental discrete submodular optimization problems such as Maximum Coverage and Facility Location.

1. Introduction

Submodularity of a set function implies a natural diminishing returns property where the marginal benefit of any given element decreases as we select more and more elements. Formally, a set function $F\colon 2^E \to \mathbb{R}$ is *submodular* if for any $S \subseteq T \subseteq E$ and $e \in E \setminus T$ it holds that $F(S \cup \{e\}) - F(S) \geq F(T \cup \{e\}) - F(T)$. Decomposable submodular functions is an important subclass of submodular functions which can be written as sums of several component submodular functions: $F(S) = \sum_{i=1}^{N} f_i(S)$, for all $S \subseteq E$, where each $f_i : 2^E \to \mathbb{R}$ is a submodular function on the ground set E with $|E| = n$.

Decomposable submodular functions include some of the most fundamental and well-studied submodular functions such as max coverage, graph cuts, welfare maximization etc., and have found numerous applications in machine learning (Dueck and Frey, 2007; Gomes and Krause, 2010; Mirzasoleiman, Badanidiyuru, and Karbasi, 2016a; Mirzasoleiman, Karbasi, Sarkar, and Krause, 2016b; Mirzasoleiman, 2017), economics (Dobzinski & Schapira, 2006; Feige, 2006; Feige & Vondrák, 2006; Papadimitriou et al., 2008; Vondrák, 2008), and data summarization and recommender systems (Dueck & Frey, 2007; Gomes & Krause, 2010; Tschiatschek et al., 2014; Lin & Bilmes, 2011). The main approach in these cases is the *centralized* and *sequential* greedy.

The need for scalable and efficient optimization methods, which do not require collecting raw data in a central server and ensure secure information collection, is widespread in applications handling sensitive data such as medical data, web search queries, salary data, and social networks. In many such cases, individuals and companies are reluctant to share their data and collecting their data in a central server is a violation of their privacy. Moreover, collecting and storing all data on a single server or cluster is computationally expensive and infeasible for large-scale datasets, particularly when working with high-dimensional data or complex models. Thus, there is a widespread demand for scalable optimization algorithms that are both *decentralized* and prioritize *privacy*. Below are some examples that motivate the focus of this paper.

Example 1.1 (Welfare Maximization)**.** The welfare maximization problem aims to maximize the overall utility or welfare of a group of individuals or agents $a_1, \ldots, a_N$. In this problem, there is a set of items or goods E, and each individual a_i has a certain preference or utility expressed as a submodular function $f_i : 2^E \to \mathbb{R}_+$ that assigns a value

[1]Halıcıoğlu Data Science Institute, University of California, San Diego, USA. Correspondence to: Akbar Rafiey <arafiey@ucsd.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

to each combination of items. The goal is to partition E into disjoint subsets $S_1, \ldots, S_N$ in order to maximize the social welfare $\sum_{i=1}^{N} f_i(S_i)$. In an important special case, called Combinatorial Public Projects (Gupta et al., 2010; Papadimitriou et al., 2008), the goal is to find a subset $S \subseteq E$ of size at most k maximizing $F(S) = \sum_{i=1}^{N} f_i(S)$. This problem appears in different fields, such as resource allocation, public goods provision, market design, and has been intensively studied (Khot et al., 2008; Lehmann et al., 2006; Mirrokni et al., 2008). An optimal approximation algorithm is known in *value oracle* model in which it is required to have access to the value of $f_i(S)$ for each agent and any $S \subseteq E$ (Călinescu et al., 2011). However, in scenarios where agents are hesitant to disclose their data to a central server and storing all data on a single server is computationally infeasible, the demand for a decentralized and private submodular maximization algorithm becomes imperative.

Example 1.2 (Feature Selection)**.** Enabling privacy-protected data sharing among clinical centers is crucial for global collaborations. Consider geographically dispersed hospitals $H_1, \ldots, H_N$, with each hospital maintaining its own data that it is unwilling to share. The goal is to identify a small subset of features that effectively classifies the target variable across the entire dataset over all hospitals. A decentralized and efficient feature selection algorithm is crucial for uncovering hidden patterns while maintaining data ownership. By adopting a decentralized approach, hospitals can balance collaborative knowledge discovery and data privacy. One approach is to maximize a submodular function capturing the mutual information between features and the class labels (Krause & Guestrin, 2005).

Federated setting for learning and optimization. *Federated setting* (Konečnỳ et al., 2016; McMahan et al., 2017) is a novel and practical framework that addresses issues regarding privacy, data sharing, and centralized computation. On one hand, it is a distributed and collaborative approach that allows multiple parties, such as different organizations or devices, to train a shared model or collaboratively optimize an objective function while keeping their data locally. This approach helps to protect the privacy of data by ensuring that the raw data is never shared or moved outside of the individuals' systems. Instead, only the model updates are exchanged and aggregated to improve the shared model and improve the objective value.

On the other hand, the federated framework reduces the amount of data that needs to be transferred and processed at any one time, which can significantly reduce the computational complexity of the overall process. Additionally, federated setting can also take advantage of the computational resources available at each party, such as the processing power of mobile devices or edge devices, which can further reduce computational load on the server. This way, federated learning can train models more efficiently, even with large-scale datasets and complex models, and provide a scalable solution for distributed learning.

Problem definition and setting. In this paper, we introduce the problem of maximizing a submodular function in the federated setting. Let E be a ground set of size n and $c_1, \ldots, c_N$ be N clients each of whom has a *private* interest over E. Each client's interest is expressed as a submodular function. Let $f_i : 2^E \to \mathbb{R}_+$ be the associated submodular function of the i-th client. A central server wants to solve the following constrained distributed optimization model

$$\max_{S \in \mathcal{I}} \left\{ F(S) = \sum_{i=1}^{N} p_i f_i(S) \right\}, \tag{1}$$

where $\mathcal{I}$ is the independent sets of a matroid $\mathcal{M}$ with ground set E, and p_i are pre-defined weights such that $\sum_{i=1}^{N} p_i = 1$. For instance, they can be set to $1/N$, or the fraction of data owned by each client. The constraint implies sets of particular properties, e.g., subsets of size at most k. Note that, the unconstrained optimization is a special case of this.

In the optimization problem (1) the data can be massively distributed over the number of clients N, which can be huge. Moreover unlike the traditional distributed setting, in the federated setting the server does not have control over clients' devices nor on how data is distributed. For example, when a mobile phone is turned off or WiFi access is unavailable, the central server will lose connection to this device. Furthermore, client's objective can be very different depending on their local datasets. To minimize communication overhead and server computation load, the number of communication rounds need to be minimal.

Constraints. We formally discuss factors of efficiency and restrictions that should be considered.

1. Privacy: One of the main appeals of decentralized and federated setting is to preserve the privacy. There are several models of privacy and security that have been considered in the literature such as Differential Privacy (DP) and Secure Aggregator (SecAgg), and a mix of these two. Single-server SecAgg is a cryptographic secure multi-party computation (MPC) that enables clients to submit vector inputs, such that the server (an aggregator) can *only* decipher the combined update, not individual updates. This is usually achieved via additive masking over a finite group (Bell et al., 2020; Bonawitz et al., 2016). Note that secure aggregation alone does not provide any privacy guarantees. To achieve a DP-type guarantee, noise can be added locally, with the server aggregating the perturbed local information via SecAgg. This user-level DP framework has recently been adopted in private federated learning (Agarwal et al., 2018; 2021; Kairouz et al., 2021; Chen et al., 2022a; Wang et al., 2023).

In this paper we use Single-server SecAgg model of privacy, a dominant and well-established approach in the field.

We leave other notions of privacy, such as a mix of DP and SecAgg, for future works. There has been a recent and concurrent progress towards this direction in terms of cardinality constraints (Wang et al., 2023).

2. Communication and bit complexity: There are a few aspects to this, firstly the number of communication rounds should be as small as possible. Second, the information communicated between should require low bandwidth and they better require small bit complexity to encode.

3. Convergence and utility: While the above impose strong restrictions, a good decentralized submodular maximization algorithm should not scarify the convergence rate by too much and should yield to an accurate and acceptable result in comparison to the centeralized methods.

Our contributions. We present the first federated (constrained) submodular maximization algorithm converging close to optimum guarantees known in centralized settings.

• We propose a decentralized version of the popular Continuous Greedy algorithm `Federated Continuous Greedy` (`FedCG`) and prove its convergence whenever the client functions are nonnegative monotone submodular achieving the optimal multiplicative approximation factor $(1 - 1/e)$ with a small additive (Section 3).

• We incorporate important and practical scenarios that are relevant for federated setting such as partial client selection, low communication rounds and computation cost. We give rigorous theoretical guarantees under each scenarios matching the optimal multiplicative approximation of $(1 - 1/e)$ and small additive error (Section 4).

• We introduce a new algorithm that serves as a discrete federated optimization algorithm for submodular maximization. Its convergence and applications to discrete problems such as `Facility Location` and `Maximum Coverage` are explored (Section 5).

1.1. Related Work

FedAvg, its convergence, and assumptions. The concept of Federated Learning (FL) (McMahan et al., 2017) has found application in various domains such as natural language processing, computer vision, and healthcare. The popular FL algorithm, Federated Averaging (`FedAvg`), is an extension of Local SGD that aims to reduce communication costs in distributed settings (Gorbunov et al., 2021; Stich, 2019; Wang & Joshi, 2021; Yu et al., 2019). However, despite its practical benefits in addressing privacy, data heterogeneity, and computational constraints, it may not converge to a "good enough" solution in general (Pathak & Wainwright, 2020; Zhang et al., 2020). Analyzing the convergence of `FedAvg` and providing theoretical guarantees is challenging and necessitates making certain assumptions. Assumptions related to bounded gradients, convexity, Lipschitzness, statistical heterogeneity, and bounded variance of stochastic gradients for each client have been explored in recent works (Karimireddy et al., 2020; Li et al., 2020; Woodworth et al., 2020; Wang et al., 2019; Yu et al., 2019).

Decentralized / distributed submodular maximization. The main approach to submodular maximization is the greedy approach which in fact, in the centralized setting yields the tight approximation guarantee in various scenarios and constraints e.g., see (Nemhauser et al., 1978; Vondrák, 2008; Călinescu et al., 2011). Centeralized submodular maximization under privacy constraints is an active research area (Mitrovic et al., 2017; Rafiey & Yoshida, 2020; Chaturvedi et al., 2021). However the sequential nature of the greedy approach makes it challenging to scale it to massive datasets. This issue is partially addressed by the means of Map-Reduce style algorithms (Kumar et al., 2015) as well as several elegant algorithms in the distributed setting (Mirzasoleiman et al., 2016b; Barbosa et al., 2015). Recent work of Mokhtari et al. (2018b) ventures towards decentralized submodular maximization for continuous submodular functions. In general continuous submodular functions are not convex nor concave and there has been a line of work to optimize continuous submodular functions using SGD methods (Hassani et al., 2017). Mokhtari et al. under several assumptions, such as assuming clients' local objective functions are monotone, DR-submodular, Lipschitz continuous, and have bounded gradient norms, prove that *Decenteralized Continuous Greedy* algorithm yields a feasible solution with quality $O(1 - 1/\mathrm{e})$ times the optimal solution. The setting in (Mokhtari et al., 2018b) is fundamentally different from the federated setting in a sense that they require sharing gradient information of the clients with the server or with the neighboring nodes in an underlying graph. Perhaps the most closely related method to our work is due to Dadras et al. (2022); Zhang et al. (2022). Dadras et al. (2022) consider Frank-Wolfe Algorithm (Frank & Wolfe, 1956) in the federated setting and propose `Federated Frank-Wolfe` (`FedFW`) algorithm and analyze its convergence for both convex and non-convex functions and under the L-Lipschitzness and bounded gradients assumptions.

2. Preliminaries

Let E denote the ground set, $|E| = n$. For a vector $\mathbf{x} \in \mathbb{R}^{|E|}$ and a set $S \subseteq E$, $\mathbf{x}(S)$ denotes $\sum_{e \in S} \mathbf{x}(e)$. A submodular function $f : 2^E \to \mathbb{R}$ is monotone if $f(S) \leq f(T)$ for every $S \subseteq T \subseteq E$. Throughout this paper we assume $f(\emptyset) = 0$.

Multilinear extension. The multilinear extension $\widehat{f} : [0,1]^{|E|} \to \mathbb{R}$ of a set function $f : \{0,1\}^{|E|} \to \mathbb{R}$ is

$$\widehat{f}(\mathbf{x}) = \sum_{S \subseteq E} f(S) \prod_{e \in S} \mathbf{x}(e) \prod_{e \notin S} (1 - \mathbf{x}(e)) = \mathbb{E}_{R \sim \mathbf{x}}[f(R)]$$

where $R \subseteq E$ is a random set that contains each element $e \in E$ independently with probability $\mathbf{x}(e)$ and excludes it with probability $1 - \mathbf{x}(e)$. We write $R \sim \mathbf{x}$ to denote that $R \subseteq E$ is a random set sampled according to $\mathbf{x}$.

Observe that for all $S \subseteq E$ we have $\widehat{f}(\mathbf{1}_S) = f(S)$. For monotone non-decreasing submodular function f, $\widehat{f}$ has the following properties (Călinescu et al., 2011) that are crucial in analyses of our algorithms:

1. $\widehat{f}$ is monotone, meaning $\frac{\partial \widehat{f}}{\partial \mathbf{x}(e)} \geq 0$. Hence, $\nabla \widehat{f}(\mathbf{x}) = (\frac{\partial \widehat{f}}{\partial \mathbf{x}(1)}, \ldots, \frac{\partial \widehat{f}}{\partial \mathbf{x}(n)})$ is a nonnegative vector.

2. $\widehat{f}$ is concave along any direction $\mathbf{d} \geq \mathbf{0}$.

Note that $\frac{\partial \widehat{f}}{\partial \mathbf{x}(e)} = \mathbb{E}_{R \sim \mathbf{x}}[f(R \cup \{e\}) - f(R \setminus \{e\})]$. That is the expected marginal contribution for e where the expectation is taken over $R \subseteq E \setminus \{e\}$ sampled according to $\mathbf{x}$. By submodularity, for any $\mathbf{x} \in [0,1]^n$,

$$|\nabla \widehat{f}(\mathbf{x})|_\infty \leq \max_{e \in E} f(\{e\}) := m_f. \tag{2}$$

Matroids and matroid polytopes. A pair $\mathcal{M} = (E, \mathcal{I})$ of a set E and $\mathcal{I} \subseteq 2^E$ is called a *matroid* if 1) $\emptyset \in \mathcal{I}$, 2) $A \in \mathcal{I}$ for any $A \subseteq B \in \mathcal{I}$, and 3) for any $A, B \in \mathcal{I}$ with $|A| < |B|$, there exists $e \in B \setminus A$ such that $A \cup \{e\} \in \mathcal{I}$. We call a set in $\mathcal{I}$ an *independent set*. We sometimes abuse notation and use $S \in \mathcal{M}$. The *rank function* $r_{\mathcal{M}} \colon 2^E \to \mathbb{Z}_+$ of $\mathcal{M}$ is $r_{\mathcal{M}}(S) = \max\{|I| : I \subseteq S, I \in \mathcal{I}\}$. An independent set $S \in \mathcal{I}$ is called a *base* if $r_{\mathcal{M}}(S) = r_{\mathcal{M}}(E)$. We denote the rank of $\mathcal{M}$ by $r(\mathcal{M})$. The *matroid polytope* $\mathcal{P}(\mathcal{M}) \subseteq \mathbb{R}^E$ of $\mathcal{M}$ is $\mathcal{P}(\mathcal{M}) = \operatorname{conv}\{\mathbf{1}_I : I \in \mathcal{I}\}$ where conv denotes the convex hull. Or equivalently (Edmonds, 2001), $\mathcal{P}(\mathcal{M}) = \{\mathbf{x} \geq \mathbf{0} : \mathbf{x}(S) \leq r_{\mathcal{M}}(S);\ \forall S \subseteq E\}$.

The Continuous Greedy Algorithm. Our algorithms for maximizing a submodular function in federated settings are based on the `Continuous Greedy` (`CG`) algorithm. We briefly explain this algorithm. The results mentioned are from (Călinescu et al., 2011; Vondrák, 2008). Let $\mathcal{M} = (E, \mathcal{I})$ be a matroid and $\mathcal{P}(\mathcal{M})$ be its matroid polytope of rank r, let f be a nonnegative and monotone submodular function and $\widehat{f}$ be its multilinear extension. `CG` starts with $\mathbf{x}^{(0)} = \mathbf{0}$. For every $t \in \{0, 1, 2, \ldots, T-1\}$ it computes $\mathbf{x}^{(t+1)}$ using the following update step $\mathbf{x}^{(t+1)} \leftarrow \mathbf{x}^{(t)} + \eta \mathbf{v}^{(t)}$, where $\mathbf{v}^{(t)} = \operatorname{argmax}_{\mathbf{w} \in \mathcal{P}} \langle \mathbf{w}, \nabla \widehat{f}(\mathbf{x}^{(t)}) \rangle$. For $\texttt{OPT} = \max_{\mathbf{x} \in \mathcal{P}} \widehat{f}(\mathbf{x})$ we have $(1 - (1-\eta)^T)\texttt{OPT} \leq \widehat{f}(\mathbf{x}^{(T)}) + C\eta^2/2$. Here, the constant C depends on the Lipschitz of the function, and $\mathbf{x}^{(T)} \in \mathcal{P}$ as it is a convex combination of vectors from the polytope. For $\eta = 1/T$ and large enough T we get $(1 - 1/\mathrm{e})\texttt{OPT} \leq \widehat{f}(\mathbf{x}^{(T)}) + \epsilon$. Given $\mathbf{x}^{(T)} \in \mathcal{P}$, there are rounding procedures to obtain $S \in \mathcal{I}$ with $\widehat{f}(\mathbf{x}^{(T)}) \leq f(S)$. The approximation factor $1 - 1/\mathrm{e}$ is the best possible assuming $\mathbf{P} \neq \mathbf{NP}$ (Feige, 1998).

3. Federated Continuous Greedy

In this section, we propose our `Federated Continuous Greedy` (`FEDCG`) method. We start with a simplistic scenario of federated model with full participation which already shows some of the challenges that we have to overcome before delving into the partial participation model which is more computationally feasible.

Consider optimization problem (1) where each f_i is a nonnegative monotone submodular function, $\sum_{i=1}^N p_i = 1$, and $\mathcal{I}$ is the independent sets of matroid $\mathcal{M} = (E, \mathcal{I})$ of rank r.

Bit complexity and accuracy trade-off. For every client i and every $\mathbf{x} \in \mathcal{P}$, the vector $\mathbf{v}_i = \operatorname{argmax}_{\mathbf{w} \in \mathcal{P}} \langle \mathbf{w}, \nabla \widehat{f}_i(\mathbf{x}) \rangle$ is determined by maximizing a linear function $\langle \mathbf{w}, \nabla \widehat{f}_i(\mathbf{x}) \rangle$ over matroid polytope $\mathcal{P}$. This problem can be solved very efficiently. We can assume that $\mathbf{v}_i$ is a vertex of $\mathcal{P}$ and furthermore, since $\nabla \widehat{f}_i$ is a nonnegative vector, that this vertex corresponds to a base of matroid $\mathcal{M}$. Hence, without loss of generality $\mathbf{v}_i$ is the indicator vector of a base with r ones and $n - r$ zeros. Thus it can be encoded using $O(r \log(n))$ bits, sublinear in the size of the ground set. On the other hand, the vector $\nabla \widehat{f}_i(\mathbf{x})$ itself requires $\tilde{O}(n)$ bits for encoding. In what follows we will see how restricting the bit complexity effects the accuracy, and the amount of computation that server should do.

`FEDCG` with full participation. First consider the case where clients can send their gradients. We proceeds in rounds. Initially $\mathbf{x}^{(0)} = 0$. On the t-th round, first, the central server broadcasts the latest model $\mathbf{x}^{(t)}$ to all clients. Each client i after receiving the update sets $\mathbf{x}_i^{(t)} = \mathbf{x}^{(t)}$ and computes $\nabla \widehat{f}_i(\mathbf{x}^{(t)})$. The server then aggregates local information via SecAgg, and computes $\nabla \widehat{F}(\mathbf{x}^{(t)}) = \sum_{i=1}^N p_i \nabla \widehat{f}_i(\mathbf{x}^{(t)})$. After receiving $\nabla \widehat{F}(\mathbf{x}^{(t)})$, the server computes $\mathbf{v}^{(t)} = \operatorname{argmax}_{\mathbf{w} \in \mathcal{P}} \langle \mathbf{w}, \nabla \widehat{F}(\mathbf{x}^{(t)}) \rangle$ by maximizing a linear function subject to the matroid constraint and produces the new global model with learning rate η: $\mathbf{x}^{(t+1)} \leftarrow \mathbf{x}^{(t)} + \eta \mathbf{v}^{(t)}$. It is clear that, similar to the centeralized `CG`, large enough T yields $F(\mathbf{x}^{(T)}) \geq (1 - 1/\mathrm{e})\texttt{OPT}$. This simple framework has an advantage over centralized `CG`, it is taking advantage of the computational resources available at each client.

Second consider a more challenging case where clients can send at most $\tilde{O}(r)$ bits information. We see how this restriction effects the accuracy. Our algorithm, `FEDCG`, proceeds in rounds. Initially $\mathbf{x}^{(0)} = 0$. On the t-th round, first, the central server broadcasts the latest model $\mathbf{x}^{(t)}$ to all clients. Each client i after receiving the update sets $\mathbf{x}_i^{(t)} = \mathbf{x}^{(t)}$ and performs one step of continuous greedy approach to find a direction that best aligns with her local gradient:

$$\mathbf{v}_i^{(t)} \leftarrow \operatorname*{argmax}_{\mathbf{v} \in \mathcal{P}} \langle \mathbf{v}, \nabla \widehat{f}_i(\mathbf{x}_i^{(t)}) \rangle \tag{3}$$

Lastly, clients send their update directions $\mathbf{v}_1^{(t)}, \dots, \mathbf{v}_N^{(t)}$ to the secure aggregator to compute $\Delta^{(t)} = \sum_{i=1}^N p_i \mathbf{v}_i^{(t)}$. After computing $\Delta^{(t)}$, the server produces the new global model with learning rate η:

$$\mathbf{x}^{(t+1)} \leftarrow \mathbf{x}^{(t)} + \eta \Delta^{(t)} \tag{4}$$

Even in this unrealistic setting where all clients participate in each round the convergence analysis requires new insights. In order to provide an approximation guarantee for our algorithm, we shall obtain a lower bound on the function value improvement by taking the direction $\Delta^{(t)}$, that is providing a lower bound for $\langle \Delta^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)}) \rangle$. However, each $\mathbf{v}_i^{(t)}$ is the projection of the local gradient into the matroid polytope and it does not carry information about the magnitude of the expected marginal contributions i.e. $\|\nabla \widehat{f}_i(\mathbf{x}^{(t)})\|$. Without assuming an assumption on the *heterogeneity* of local functions one can construct examples where a single element and corresponding client's marginal contribution are significantly more dominant than others and hence taking direction $\Delta^{(t)}$ results in a very bad approximation guarantee.

We therefore need an assumption that acts as a tool in constraining the level of heterogeneity which poses a significant obstacle in federated optimization. A common way to handle heterogeneity is to impose a bound on the magnitude of gradients over each clients' local function i.e $\|\nabla \widehat{f}_i(\mathbf{x})\| \leq \gamma$. This types of assumption is not only common in the literature regarding submodular function maximization in decentralized settings (Mokhtari et al., 2018b; Zhang et al., 2020), but also in studies of convex and non-convex optimization in federated learning (Chen et al., 2022b; Dadras et al., 2022; Karimireddy et al., 2020; Yu et al., 2019; Li et al., 2020).

In the case of submodular functions, each coordinate of the gradient of the multilinear extension corresponds to the marginal gain of adding one single element. In this paper we impose the following assumption which is much more relaxed than assuming a bound on the magnitude of the gradients from each client.

Assumption 3.1. For all $i = 1, \dots, N$ and $t = 1, \dots, T$ we have $|\nabla \widehat{f}_i(\mathbf{x}^{(t)}) - \nabla \widehat{F}(\mathbf{x}^{(t)})|_\infty \leq \gamma_t$. Note that, by submodularity and monotonicity we have

$$\max_{t \in [T]} \gamma_t \leq 2 \max_i \max_{e \in E} f_i(\{e\}) = 2 \max_i m_{f_i}. \tag{5}$$

Monotonicity of f_i implies that for every $\mathbf{x} \leq \mathbf{y}$ coordinate-wise, it holds that $\widehat{f}_i(\mathbf{x}) \leq \widehat{f}_i(\mathbf{x})$. Additionally, gradients are antitone i.e., for every $\mathbf{x} \leq \mathbf{y}$ coordinate-wise, it holds that $\nabla \widehat{f}_i(\mathbf{x}) \geq \nabla \widehat{f}_i(\mathbf{y})$. Thus as the algorithm advances and t grows, γ_t's change but never exceed the upper bound in (5). Additionally, in numerous instances, $\max_{t \in [T]} \gamma_t$ is relatively small. For instance, in `Max Coverage` problem each $f_i(S)$ is either 0 or 1 , depending if a client is covered by S or not, thus for this problem $\gamma = 1$.

Algorithm 1 Federated Continuous Greedy (FEDCG)

Input: Matroid polytope $\mathcal{P}$, number of communication rounds T, learning rate η, and K.
$\mathbf{x}^{(0)} = \mathbf{0}$
for $t = 0$ to $T - 1$ **do**
  Server selects a subset of K *active* clients $A^{(t)}$ according to Client Sampling Scheme, and sends $\mathbf{x}^{(t)}$ to them.
  for Client i in $A^{(t)}$ in parallel **do**
    $\mathbf{v}_i^{(t)} \leftarrow \operatorname{argmax}_{\mathbf{v} \in \mathcal{P}} \langle \mathbf{v}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle$
    Send $\mathbf{v}_i^{(t)}$ back to the secure aggregator.
  end for
  SecAgg: $\Delta^{(t)} = \frac{1}{|A^{(t)}|} \sum_{i \in A^{(t)}} \mathbf{v}_i^{(t)}$
  Server updates: $\mathbf{x}^{(t+1)} \leftarrow \mathbf{x}^{(t)} + \eta \Delta^{(t)}$
end for
Apply a proper rounding scheme on $\mathbf{x}^{(T)}$ to obtain a solution for (1)

We now have enough ingredients to prove the following convergence theorem for the case where all clients participate in every communication round. Let $D = \sum_{t=1}^T \gamma_t$

Theorem 3.2 (Full participation)**.** *Let $\mathcal{M}$ be a matroid of rank r and $\mathcal{P}$ be its matroid polytope. Under the full participation assumption and Assumption 3.1, for every $\eta > 0$, Algorithm 1 returns a $\mathbf{x}^{(T)} \in \mathcal{P}$ such that*

$$\left(1 - (1-\eta)^T\right) OPT \leq \widehat{F}(\mathbf{x}^{(T)}) + \eta r \sum_{t=1}^T \gamma_t + \frac{T \eta^2 r^2 m_F}{2}$$

In particular, for large enough T, setting $\eta = 1/T$, Algorithm 1 requires at most $\tilde{O}(r)$ bits of communication per user per round ($\tilde{O}(NTr)$ in total) and obtains

$$(1 - 1/\mathrm{e}) OPT \leq \widehat{F}(\mathbf{x}^{(T)}) + \left(\frac{rD}{T} + \frac{r^2 m_F}{2T} \right).$$

3.1. Partial Participation and Client Selection

Client sampling in the FL optimization framework is imperative for various practical reasons, including the following:

• Large scale and dynamic nature. In real-world applications, a server usually serves several billions of devices/clients who can join or leave the federated optimization system due to several reasons like intermittent connectivity, technical issue, or simply based on their availability or preferences. Hence, it is computationally inefficient and often impossible to get updates from all clients.

• Communication and bandwidth. On one hand, waiting for the slowest client to finish can increase the expected round

duration as the number of participating clients per training increases, a phenomenon known as "straggler's effect". On the other hand, communication can be a primary bottleneck for federated settings because of clients bandwidth limitation and the possibility of server throttling.

• Small models and redundancy. It is often the case that FL models are small because of clients limited computational power or memory, it therefore is unnecessary to train an FL model on billions of clients. Note that for optimization problem (1) in many practical scenarios models have smaller size in comparison to the number of clients, this is because of the matroid constraint or simply because the size of the ground set is much smaller than the number of clients.

Here we discuss our sampling scheme which crucially does not violate clients' privacy; see Algorithm 1.

Unbiased client sampling scheme. At each communication round the server chooses an active client from $i \in [N]$ with probability p_i, and repeats this process K times to obtain a multiset $A^{(t)}$ of size K which may contain a client more than once. Then the aggregation step is $\Delta^{(t)} = \frac{1}{K}\sum_{i\in A^{(t)}} \mathbf{v}_i^{(t)}$ where $\mathbf{v}_i^{(t)}$ defined in (3).

The next lemma shows that this sampling scheme is unbiased and in expectation the average update from chosen clients $A^{(t)}$ is equal to the average update from all clients.

Lemma 3.3 (Unbiased sampling scheme)**.** *For Client Sampling Scheme, we have* $\mathbb{E}_{A^{(t)}}\left[\langle \Delta^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle\right] = \sum_{i=1}^N p_i \langle \mathbf{v}_i^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle$.

The following lemma allows us to bound the variance which in turn helps to provide our convergence guarantees. To show the following result, it is required to upper bound the difference between the improvement on the function value by taking the direction suggested by a selected client versus taking the direction obtained by averaging all the directions from the clients. That is bounding $|\langle \mathbf{v}_s, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle - \langle \sum_{i=1}^N \mathbf{v}_i^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle|$, for a selected client s in $A^{(t)}$. This in turn needs providing an upper bound for $\langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle - \langle \mathbf{v}_s, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle$, for all $i \neq s$. At the heart, our proof relies on the properties of multilinear extensions of local submodular functions, the fact that each $\mathbf{v}_i^{(t)}$ corresponds to a base of the matroid, and Assumption 3.1.

Lemma 3.4 (Bounded variance)**.** *Using Client Sampling Scheme we have* $\mathrm{Var}(\langle \Delta^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle) \leq 36 r^2 \gamma_t^2 / K$.

Armed with the above lemmas and concentration inequalities e.g., Chebyshev's inequality, we can prove the convergence of Algorithm 1. This essentially is done by bounding the error introduced by the decentralized setting and carefully carrying the error through the analysis.

Theorem 3.5. *Let $\mathcal{M}$ be a matroid of rank r and $\mathcal{P}$ be its matroid polytope. Using Client Sampling Scheme, for every $\eta, \delta > 0$, Algorithm 1 returns a $\mathbf{x}^{(T)} \in \mathcal{P}$ so that with probability at least $1-\delta$*

$$\left(1-(1-\eta)^T\right) \mathtt{OPT} \leq \widehat{F}(\mathbf{x}^{(T)}) + \eta\left(r\sum_{t=1}^T \gamma_t + \frac{6r\sum_{t=1}^T \gamma_t}{\sqrt{K\delta/T}}\right) + \frac{T\eta^2 r^2 m_F}{2}$$

In particular, by setting $\eta = 1/T$, Algorithm 1 requires at most $\tilde{O}(r)$ bits of communication per user per round ($\tilde{O}(KTr)$ in total) and yields

$$(1-1/\mathrm{e})\,\mathtt{OPT} \leq \widehat{F}(\mathbf{x}^{(T)}) + \left(\frac{rD}{T} + \frac{6rD}{\sqrt{TK\delta}} + \frac{r^2 m_F}{2T}\right)$$

4. Practical Federated Continuous Greedy

One of the main considerations in federated optimization is the number of communication rounds. In this section, we show how Algorithm 1 can be further improved to reduce communication rounds while simultaneously incorporating partial participation.

Algorithm description. Initially $\mathbf{x}^{(0)} = 0$. On the t-th round of the `Practical Federated Continuous Greedy (FedCG+)`, the central server first broadcasts the latest model $\mathbf{x}^{(t)}$ to a subset of active clients of size K denoted by $A^{(t)}$. Next, each client $i \in A^{(t)}$ sets $\mathbf{x}_i^{(t,0)} = \mathbf{x}^{(t)}$ and performs τ steps of continuous greedy approach locally. More precisely, let $j \in \{0, 1, \ldots, \tau-1\}$ and $\mathbf{x}_i^{(t,j)}$ denote the i-th client's local model at communication round t and local update step j, then the local updates are

$$\widetilde{\mathbf{v}}_i^{(t,j)} \leftarrow \operatorname*{argmax}_{\mathbf{v}\in\mathcal{P}} \langle \mathbf{v}, \nabla \widetilde{f}_i(\mathbf{x}_i^{(t,j)}, \zeta_i^{(t,j)})\rangle; \tag{6}$$

$$\mathbf{x}_i^{(t,j+1)} = \mathbf{x}_i^{(t,j)} + \widetilde{\mathbf{v}}_i^{(t,j)}/\tau \tag{7}$$

Here $\zeta_i^{(t,j)}$ is a set of subsets from the ground set E sampled according to $\mathbf{x}_i^{(t,j)}$, and $\nabla \widetilde{f}_i(\mathbf{x}_i^{(t,j)}, \zeta_i^{(t,j)}) \in \mathbb{R}^n_{\geq 0}$ is an estimation of the gradient $\nabla \widehat{f}_i(\mathbf{x}_i^{(t,j)})$ (more on this later).

After τ steps of local update, the i-th client from $A^{(t)}$ send her update $\widetilde{\Delta}_i^{(t+\tau)} = \mathbf{x}_i^{(t,\tau)} - \mathbf{x}_i^{(t,0)}$ to the secure aggregator to compute $\widetilde{\Delta}^{(t+\tau)} = \frac{1}{|A^{(t)}|}\sum_{i\in A^{(t)}} \widetilde{\Delta}_i^{(t+\tau)}$. Note that each $\widetilde{\Delta}_i^{(t+\tau)}$ belongs to $\mathcal{P}$ since it is a convex combination of vectors $\widetilde{\mathbf{v}}_i^{(t,j)} \in \mathcal{P}$. However, $\widetilde{\Delta}_i^{(t+\tau)}$ may not be an integral vector and in the worst case $O(n)$ bits are required to encode it. After receiving $\widetilde{\Delta}^{(t+\tau)}$, the server produces the new global model with learning rate η:

$$\mathbf{x}^{(t+\tau)} \leftarrow \mathbf{x}^{(t)} + \eta \widetilde{\Delta}^{(t+\tau)} \tag{8}$$

Gradient estimation. Evaluating the multilinear extension involves summing over all subsets S of E, there are $2^{|E|}$ such subsets. However, recall $\partial \widehat{f}/\partial \mathbf{x}(e) = \mathbb{E}[f(R\cup\{e\})] -$

$\mathbb{E}[f(R \setminus \{e\})]$ where $R \subseteq E$ is a random subset sampled according to $\mathbf{x}$. Hence a simple application of Chernoff's bound tells us by sampling sufficiently many subsets we can obtain a good estimation of $\nabla\widehat{f}(\mathbf{x})$ (Călinescu et al., 2011; Vondrák, 2008) (more details in the Appendix). For large enough m, let $\zeta_i^{(t,j)} = \{R_i^{(t,j,1)}, \dots, R_i^{(t,j,m)}\}$ be subsets of E that are sampled independently according to i-th client's local model $\mathbf{x}_i^{(t,j)}$. Let $\nabla\widetilde{f}_i(\mathbf{x}_i^{(t,j)}, \zeta_i^{(t,j)})$ denote the *stochastic* approximation of $\nabla\widehat{f}_i(\mathbf{x}_i^{(t,j)})$. Then with probability $1-\delta$ we have $\|\nabla\widehat{f}_i(\mathbf{x}_i^{(t,j)}) - \nabla\widetilde{f}_i(\mathbf{x}_i^{(t,j)}, \zeta_i^{(t,j)})\| \le \sigma$.

In convergence analyses of our algorithm there are two sources of randomness, one in the sampling schemes for client selection and the other in data sampling at each clients local data to estimate the gradients. In Algorithm 2 there are $\frac{T}{\tau}$ communications rounds between the server and clients. Define $\mathcal{I}_\tau = \{\tau i \mid i = 1, 2, 3, \dots\}$ to denote the set of communication rounds with the server. Similar to Lemma 3.3:

Lemma 4.1 (Unbiased sampling scheme)**.** *For Client Sampling Scheme, at every communication round $t+\tau \in \mathcal{I}_\tau$, we have $\mathbb{E}_{A^{(t)}}\left[\langle\widetilde{\Delta}^{(t+\tau)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle\right] = \langle\sum_{i=1}^N p_i\widetilde{\Delta}_i^{(t+\tau)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle$.*

Bounding the variance of $\langle\widetilde{\Delta}^{(t+\tau)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle$ is much more delicate than in Lemma 3.4. The are two main reasons, one is the deviation caused by local steps, second is the deviation from the true $\nabla\widehat{F}(\mathbf{x}^{(t)})$ caused by gradient estimations. To handle the deviation cause by local steps we assume all $\widehat{f}_i$ have L-Lipschitz continuous gradients i.e., $\|\nabla\widehat{f}_i(\mathbf{x}) - \nabla\widehat{f}_i(\mathbf{y})\| \le L\|\mathbf{x}-\mathbf{y}\|$, $\forall \mathbf{x}, \mathbf{y} \in \mathcal{P}$. In fact as shown in Lemma 3 of (Mokhtari et al., 2018a), for each $\widehat{f}_i$ and local model $\mathbf{x}_i^{(t,j+1)} = \mathbf{x}_i^{(t,j)} + \widetilde{\mathbf{v}}_i^{(t,j)}/\tau$, it holds that

$$\left\|\nabla\widehat{f}_i(\mathbf{x}_i^{(t,j+1)}) - \nabla\widehat{f}_i(\mathbf{x}_i^{(t,j)})\right\| \le \frac{m_{f_i}\sqrt{r}}{\tau}\left\|\widetilde{\mathbf{v}}_i^{(t,j)}\right\| \le \frac{m_{f_i} r}{\tau}$$

The factor $-m_{f_i}$ is in fact a lower bound on the entries of the Hessian matrix. That is $\frac{\partial\widehat{f}_i}{\partial\mathbf{x}(i)\partial\mathbf{x}(j)} \ge -m_{f_i}$ (Hassani et al., 2017).

Let $Q = \sum_{t\in I_\tau} L_t$ where L_t is such that for all $i \in [N]$ and $j \in [\tau]$ it holds

$$\left\|\nabla\widehat{f}_i(\mathbf{x}_i^{(t,j)}) - \nabla\widehat{f}_i(\mathbf{x}^{(t)})\right\| \le L_t\|\mathbf{x}_i^{(t,j)} - \mathbf{x}^{(t)}\|$$

Observe that L_t is upper bounded by $\max_i m_{f_i}\sqrt{r}$.

Lemma 4.2 (Bounded variance)**.** *For $t+\tau \in \mathcal{I}_\tau$ we have $\mathrm{Var}(\langle\widetilde{\Delta}^{(t+\tau)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle) \le \frac{1}{K}\left(6r\gamma_t + 2(\sigma r + L_t r^{1.5})\right)^2$.*

While the variance can be made arbitrary small by sampling more clients, the additive error caused by local steps cannot be controlled by sampling more clients. This shows up in the next theorem.

Algorithm 2 Practical FedCG (FEDCG+)

Input: Matroid polytope $\mathcal{P}$, number of communication rounds T/τ, server's learning rate η, $\sigma, \delta > 0$, and K.
$\mathbf{x}^{(0)} = \mathbf{0}$, $m = O(\log(TK/\delta)/\sigma^2)$
for $t = 0, \tau, 2\tau, \dots, (T-1)/\tau$ **do**
  Server selects a subset of K *active* clients $A^{(t)}$ according to Client Sampling Scheme, and sends $\mathbf{x}^{(t)}$ to them.
  for Client i in $A^{(t)}$ in parallel **do**
    $\mathbf{x}_i^{(t,0)} \leftarrow \mathbf{x}^{(t)}$
    for $j = 0, \dots, \tau-1$ **do**
      Randomly sample m sets $\zeta_i^{(t,j)} = \{R_i^{(t,j,1)}, \dots, R_i^{(t,j,m)}\}$ according to $\mathbf{x}_i^{(t,j)}$
      Let $\nabla\widetilde{f}_i(\mathbf{x}_i^{(t,j)}, \zeta_i^{(t,j)})$ be the estimate of $\nabla\widehat{f}_i(\mathbf{x}_i^{(t,j)})$
      $\widetilde{\mathbf{v}}_i^{(t,j)} \leftarrow \mathrm{argmax}_{\mathbf{v}\in\mathcal{P}}\langle\mathbf{v}, \nabla\widetilde{f}_i(\mathbf{x}_i^{(t,j)}, \zeta_i^{(t,j)})\rangle$
      $\mathbf{x}_i^{(t,j+1)} \leftarrow \mathbf{x}_i^{(t,j)} + \widetilde{\mathbf{v}}_i^{(t,j)}/\tau$
    end for
    $\widetilde{\Delta}_i^{(t+\tau)} \leftarrow \mathbf{x}_i^{(t,\tau)} - \mathbf{x}_i^{(t,0)}$ {Local model change}
    Send $\widetilde{\Delta}_i^{(t+\tau)}$ back to the secure aggregator.
  end for
  SecAgg: $\widetilde{\Delta}^{(t+\tau)} = \frac{1}{|A^{(t)}|}\sum_{i\in A^{(t)}}\widetilde{\Delta}_i^{(t+\tau)}$.
  Server updates: $\mathbf{x}^{(t+\tau)} \leftarrow \mathbf{x}^{(t)} + \eta\widetilde{\Delta}^{(t+\tau)}$
end for
Apply a proper rounding scheme on $\mathbf{x}^{(T)}$ to obtain a solution for (1)

Theorem 4.3. *Let $\mathcal{M}$ be a matroid of rank r and $\mathcal{P}$ be its matroid polytope. Using Client Sampling Scheme, for every $\eta, \delta > 0$, Algorithm 2 returns a $\mathbf{x}^{(T/\tau)} \in \mathcal{P}$ such that with probability at least $1-\delta$ it holds*

$$\begin{aligned}(1-(1-\eta)^{T/\tau})OPT &\le \widehat{F}(\mathbf{x}^{(T/\tau)}) + \frac{T\eta^2 r^2 m_F}{2\tau} \\ &+ \overbrace{\frac{T}{\tau}}^{\text{heterogeneity}} \overbrace{(\eta r\sum_{t=1}^{T/\tau}\gamma_t + 2\sigma r + 2\eta r^{1.5}\sum_{t=1}^{T/\tau}L_t)}^{\text{local steps}} \\ &+ \overbrace{\sqrt{T}(6\eta r\sum_{t=1}^{T/\tau}\gamma_t + 2\sigma r + 2\eta r^{1.5}\sum_{t=1}^{T/\tau}L_t))/\sqrt{K\tau\delta}}^{\text{client sampling}}\end{aligned}$$

In particular, for $\eta = \tau/T$, Algorithm 2 has at most $\tilde{O}(n)$ bits of communication per user per round ($\tilde{O}(KTn/\tau)$ in

Algorithm 3 Federated Discrete Greedy

Input: Matroid $\mathcal{M}$ of rank r, importance factors $\{w_i\}_{i=1}^N$, $\varepsilon \in (0,1)$.
$S \leftarrow \emptyset$, $\kappa \leftarrow \tilde{O}(rn/\varepsilon^2)$
for $t = 0$ to $r-1$ **do**
  Server sends S to all clients.
  for each client i in parallel **do**
    $\kappa_i \leftarrow \min\{\kappa \cdot w_i, 1\}$
    for all e such that $S \cup \{e\} \in \mathcal{M}$ **do**
      $\Delta_i[e] \leftarrow (f_i(S \cup \{e\}) - f_i(S))/\kappa_i$
    end for
    {/* Randomized Response */}
    With probability κ_i sends Δ_i to the secure aggregator
    With probability $1-\kappa_i$ does nothing
  end for
  SecAgg: $\Delta^{(t)}$ (the sum of Δ_i received in this round.)
  Server updates: $S \leftarrow S \cup \{\operatorname{argmax}_{e:S\cup\{e\}\in\mathcal{M}} \Delta^{(t)}[e]\}$
end for
Output: S

total) and yields

$$(1-1/\mathrm{e})\,OPT \le \widehat{F}(\mathbf{x}^{(T)}) + \frac{\tau r^2 m_F}{2T} + \frac{rD\tau}{T} + 2\sigma r(1+\frac{\sqrt{\tau}}{\sqrt{KT\delta}}) + \frac{2r^{1.5}Q\tau}{T} + \frac{\sqrt{\tau}(6Dr+2r^{1.5}Q)}{\sqrt{TK\delta}}$$

Remark 4.4. From a fractional solution $\mathbf{x}^{(T)} \in \mathcal{P}$ returned by Algorithms 1 and 2 one can obtain a solution for (1) using appropriate rounding schemes. Rounding schemes such as pipage rounding (Călinescu et al., 2011), swap rounding (Chekuri et al., 2010), or greedy rounding are oblivious; they do not require access to the objective function. Therefore, the server can utilize these rounding schemes without needing to know the decomposable function itself.

5. Discrete Algorithm in Federated Setting

While parallel SGD and continuous methods such as ours in this paper are commonly used as the main tool in federated optimization, we introduce a rather discrete approach to the field and believe it will find further applications. Our approach is inspired by recent works of Rafiey & Yoshida (2022); Kenneth & Krauthgamer (2023) on a seemingly unrelated topic. Rafiey & Yoshida (2022) introduced a method to sparsify a sum of submodular functions in a centralized setting which was improved by (Kenneth & Krauthgamer, 2023). We tailor their approach to the federated setting and discuss its effectiveness for discrete problems such as `Facility Location` and `Maximum Coverage` problems. At the heart of this approach is for clients to know their "importance" without sharing sensitive information. In the monotone case, the *importance factor* for client i is defined as:

$$w_i = \max_{e\in E} \frac{f_i(\{e\})}{F(\{e\})}.$$

In several cases such as `Max Facility Location` and `Maximum Coverage` computing the importance factor can be done efficiently and with constant number of communication rounds. For now, let us continue by assuming each client knows its own importance factor.

Algorithm 3 description. Let $\varepsilon \in (0,1)$ and set $S = \emptyset$. The server gradually adds elements to S for only r rounds. At round t the central server broadcasts the current set S to all clients (or a subset of active clients). Then each client i computes the marginal contribution of each element from $E \setminus S$ to its local function f_i

$$\Delta_i[e] = f_i(S\cup\{e\}) - f_i(S); \quad \forall e : S\cup\{e\} \in \mathcal{I}$$

Let $\kappa = \tilde{O}(rn/\varepsilon^2)$. Then each client sends its update in a *randomized response* manner, that is, the i-th client with probability $\kappa_i = \min\{1, \kappa \cdot w_i\}$ sends the scaled vector $(\frac{1}{\kappa_i})\Delta_i$ to the secure aggregator, and with the complement probability does not send anything. The secure aggregator computes $\Delta^{(t)}$; the sum of the update vectors it has received. The server updates $S \leftarrow S \cup \{\operatorname{argmax}_{e:S\cup\{e\}\in\mathcal{M}} \Delta^{(t)}[e]\}$. The following theorem follows from the sparsification results in (Rafiey & Yoshida, 2022; Kenneth & Krauthgamer, 2023) and approximation guarantees of the greedy algorithm (Nemhauser et al., 1978).

Theorem 5.1. *For every $\varepsilon \in (0,1)$, Algorithm 3, with probability at least $1-1/n$, returns a subset $S \in \mathcal{I}$ such that $(1/2-\varepsilon)OPT \le F(S)$.*

Under a cardinality constraint (uniform matroid), Algorithm 3, with probability at least $1-1/n$, returns a subset $S \in \mathcal{I}$ such that $(1-1/\mathrm{e}-\varepsilon)OPT \le F(S)$.

Moreover, in expectation, at each round $\tilde{O}(rn^2/\varepsilon^2)$ clients participate where r is the rank of the matroid.

Facility Location Problem. Let $\mathcal{C}$ be a set of N clients and E be a set of facilities with $|E| = n$. For $c : \mathcal{C} \times E \to \mathbb{R}$ let the i-th client's score function over a subset of facilities be $f_i(A) = \max_{j\in A} c(i,j)$. The objective for `Max Facility Location` is $\max_{S\subseteq E, |S|\le k} \sum_{i=1}^N \max_{j\in S} c(i,j)$. For each client i the importance factor is $w_i = \max_{j\in E} \frac{c(i,j)}{F(\{j\})}$.

In several situations computing the importance factors is straightforward. For instance, in movie recommendation systems where a facility location objective is used (see Appendix D), the average rating i.e., $F(\{j\})$, for all movies

are publicly available. Hence, each client knows its own importance factor. Having w_i, we can apply Algorithm 3. Doing so, note that at each round the expected number of clients participating is $\tilde{O}(kn^2/\varepsilon^2)$, **independent of** N.

In other situations, clients can compute their importance factors in a federated setting, using a secure aggregator and without sharing their data with other clients. Further details on this and discussions about the `Max Coverage` problem are presented in Appendix D.

6. Conclusions and Future Work

We present FEDCG, the first algorithm for decomposable submodular maximization in federated setting under matroid constraints. FEDCG is based on the continuous greedy algorithm and achieves the best possible approximation factor i.e., $1 - 1/\mathrm{e}$ under mild assumptions even faced with client selection and low communication rounds. Additionally, we introduce a new federated framework for discrete problems. Our work leads to many interesting directions for future work, such as providing stronger privacy guarantee using differentially private methods (Mitrovic et al., 2017; Rafiey & Yoshida, 2020; Chaturvedi et al., 2021), finding a lower bound on the bit complexity, and improving the additive errors of our algorithms.

Acknowledgements

I would like to thank Arya Mazumdar and Barna Saha for their many useful discussions and for reading several drafts of this paper. I would also like to thank Yuichi Yoshida for his valuable discussions and inspiring comments on the paper. Additionally, I am grateful to Nazanin Mehrasa, Wei-Ning Chen, Heng Zhu, and Quanquan C. Liu for insightful comments on this work.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agarwal, N., Suresh, A. T., Yu, F. X., Kumar, S., and McMahan, B. cpsgd: Communication-efficient and differentially-private distributed SGD. In *NeurIPS 2018*, pp. 7575–7586, 2018.

Agarwal, N., Kairouz, P., and Liu, Z. The skellam mechanism for differentially private federated learning. In *NeurIPS 2021*, pp. 5052–5064, 2021.

Barbosa, R. D. P., Ene, A., Nguyen, H. L., and Ward, J. The power of randomization: Distributed submodular maximization on massive datasets. In *ICML*, volume 37, pp. 1236–1244. JMLR.org, 2015.

Bell, J. H., Bonawitz, K. A., Gascón, A., Lepoint, T., and Raykova, M. Secure single-server aggregation with (poly)logarithmic overhead. In *CCS*, pp. 1253–1269. ACM, 2020.

Bonawitz, K. A., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., and Seth, K. Practical secure aggregation for federated learning on user-held data. *CoRR*, abs/1611.04482, 2016.

Călinescu, G., Chekuri, C., Pál, M., and Vondrák, J. Maximizing a monotone submodular function subject to a matroid constraint. *SIAM J. Comput.*, 40(6):1740–1766, 2011.

Chaturvedi, A., Nguyen, H. L., and Zakynthinou, L. Differentially private decomposable submodular maximization. In *AAAI*, pp. 6984–6992, 2021.

Chekuri, C., Vondrák, J., and Zenklusen, R. Dependent randomized rounding via exchange properties of combinatorial structures. In *FOCS*, pp. 575–584. IEEE Computer Society, 2010.

Chen, W., Choquette-Choo, C. A., Kairouz, P., and Suresh, A. T. The fundamental price of secure aggregation in differentially private federated learning. In *ICML*, volume 162, pp. 3056–3089. PMLR, 2022a.

Chen, W., Horváth, S., and Richtárik, P. Optimal client sampling for federated learning. *Transactions on Machine Learning Research*, 2022b. ISSN 2835-8856.

Dadras, A., Prakhya, K., and Yurtsever, A. Federated frank-wolfe algorithm. In *Workshop on Federated Learning: Recent Advances and New Challenges (in Conjunction with NeurIPS)*, 2022.

Dobzinski, S. and Schapira, M. An improved approximation algorithm for combinatorial auctions with submodular bidders. In *SODA*, pp. 1064–1073, 2006.

Dueck, D. and Frey, B. J. Non-metric affinity propagation for unsupervised image categorization. In *ICCV*, pp. 1–8, 2007.

Edmonds, J. Submodular functions, matroids, and certain polyhedra. In *Combinatorial Optimization - Eureka, You Shrink!*, pp. 11–26, 2001.

Feige, U. A threshold of ln *n* for approximating set cover. *J. ACM*, 45(4):634–652, 1998.

Feige, U. On maximizing welfare when utility functions are subadditive. In *STOC*, pp. 41–50, 2006.

Feige, U. and Vondrák, J. Approximation algorithms for allocation problems: Improving the factor of 1 - 1/e. In *FOCS*, pp. 667–676, 2006.

Frank, M. and Wolfe, P. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2): 95–110, 1956.

Gomes, R. and Krause, A. Budgeted nonparametric learning from data streams. In *ICML*, pp. 391–398, 2010.

Gorbunov, E., Hanzely, F., and Richtárik, P. Local SGD: unified theory and new efficient methods. In *AISTATS*, volume 130, pp. 3556–3564. PMLR, 2021.

Gupta, A., Ligett, K., McSherry, F., Roth, A., and Talwar, K. Differentially private combinatorial optimization. In *SODA*, pp. 1106–1125. SIAM, 2010.

Hassani, S. H., Soltanolkotabi, M., and Karbasi, A. Gradient methods for submodular maximization. In *NeurIPS*, pp. 5841–5851, 2017.

Kairouz, P., Liu, Z., and Steinke, T. The distributed discrete gaussian mechanism for federated learning with secure aggregation. In *ICML*, volume 139, pp. 5201–5212. PMLR, 2021.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. SCAFFOLD: stochastic controlled averaging for federated learning. In *ICML*, volume 119, pp. 5132–5143. PMLR, 2020.

Kenneth, Y. and Krauthgamer, R. Cut sparsification and succinct representation of submodular hypergraphs. *CoRR*, abs/2307.09110, 2023.

Khot, S., Lipton, R. J., Markakis, E., and Mehta, A. Inapproximability results for combinatorial auctions with submodular utility functions. *Algorithmica*, 52(1):3–18, 2008.

Konečnỳ, J., McMahan, H. B., Ramage, D., and Richtárik, P. Federated optimization: distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

Krause, A. and Guestrin, C. Near-optimal nonmyopic value of information in graphical models. In *UAI*, pp. 324–331, 2005.

Kumar, R., Moseley, B., Vassilvitskii, S., and Vattani, A. Fast greedy algorithms in mapreduce and streaming. *ACM Trans. Parallel Comput.*, 2(3):14:1–14:22, 2015.

Lehmann, B., Lehmann, D., and Nisan, N. Combinatorial auctions with decreasing marginal utilities. *Games Econ. Behav.*, 55(2):270–296, 2006.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of fedavg on non-iid data. In *ICLR*. OpenReview.net, 2020.

Lin, H. and Bilmes, J. A. A class of submodular functions for document summarization. In *HLT*, pp. 510–520, 2011.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, volume 54, pp. 1273–1282. PMLR, 2017.

Mirrokni, V. S., Schapira, M., and Vondrák, J. Tight information-theoretic lower bounds for welfare maximization in combinatorial auctions. In *EC*, pp. 70–77. ACM, 2008.

Mirzasoleiman, B. *Big Data Summarization Using Submodular Functions*. PhD thesis, ETH Zurich, Zürich, Switzerland, 2017.

Mirzasoleiman, B., Badanidiyuru, A., and Karbasi, A. Fast constrained submodular maximization: Personalized data summarization. In *ICML*, volume 48, pp. 1358–1367, 2016a.

Mirzasoleiman, B., Karbasi, A., Sarkar, R., and Krause, A. Distributed submodular maximization. *J. Mach. Learn. Res.*, 17:238:1–238:44, 2016b.

Mitrovic, M., Bun, M., Krause, A., and Karbasi, A. Differentially private submodular maximization: Data summarization in disguise. In *ICML*, pp. 2478–2487, 2017.

Mokhtari, A., Hassani, H., and Karbasi, A. Conditional gradient method for stochastic submodular maximization: Closing the gap. In *AISTATS*, volume 84, pp. 1886–1895. PMLR, 2018a.

Mokhtari, A., Hassani, H., and Karbasi, A. Decentralized submodular maximization: Bridging discrete and continuous settings. In *ICML*, volume 80, pp. 3613–3622. PMLR, 2018b.

Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. An analysis of approximations for maximizing submodular set functions - I. *Math. Program.*, 14(1):265–294, 1978.

Papadimitriou, C. H., Schapira, M., and Singer, Y. On the hardness of being truthful. In *FOCS*, pp. 250–259. IEEE Computer Society, 2008.

Pathak, R. and Wainwright, M. J. Fedsplit: an algorithmic framework for fast federated optimization. In *NeurIPS*, volume 33, pp. 7057–7066, 2020.

Rafiey, A. and Yoshida, Y. Fast and private submodular and k-submodular functions maximization with matroid constraints. In *ICML*, pp. 7887–7897, 2020.

Rafiey, A. and Yoshida, Y. Sparsification of decomposable submodular functions. In *AAAI*, pp. 10336–10344. AAAI Press, 2022.

Stan, S., Zadimoghaddam, M., Krause, A., and Karbasi, A. Probabilistic submodular maximization in sub-linear time. In *ICML*, volume 70, pp. 3241–3250. PMLR, 2017.

Stich, S. U. Local SGD converges fast and communicates little. In *ICLR*. OpenReview.net, 2019.

Tschiatschek, S., Iyer, R. K., Wei, H., and Bilmes, J. A. Learning mixtures of submodular functions for image collection summarization. In *NeurIPS*, pp. 1413–1421, 2014.

Vondrák, J. Optimal approximation for the submodular welfare problem in the value oracle model. In *STOC*, pp. 67–74, 2008.

Wang, J. and Joshi, G. Cooperative SGD: A unified framework for the design and analysis of local-update SGD algorithms. *J. Mach. Learn. Res.*, 22:213:1–213:50, 2021.

Wang, S., Tuor, T., Salonidis, T., Leung, K. K., Makaya, C., He, T., and Chan, K. Adaptive federated learning in resource constrained edge computing systems. *IEEE J. Sel. Areas Commun.*, 37(6):1205–1221, 2019.

Wang, Y., Zhou, T., Chen, C., and Wang, Y. Federated submodular maximization with differential privacy. *IEEE Internet of Things Journal*, 2023.

Woodworth, B. E., Patel, K. K., Stich, S. U., Dai, Z., Bullins, B., McMahan, H. B., Shamir, O., and Srebro, N. Is local SGD better than minibatch sgd? In *ICML*, volume 119, pp. 10334–10343. PMLR, 2020.

Yu, H., Yang, S., and Zhu, S. Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *AAAI*, pp. 5693–5700. AAAI Press, 2019.

Zhang, M., Zhou, Y., Ge, Q., Zheng, R., and Wu, Q. Decentralized randomized block-coordinate frank-wolfe algorithms for submodular maximization over networks. *IEEE Trans. Syst. Man Cybern. Syst.*, 52(8):5081–5091, 2022.

Zhang, X., Hong, M., Dhople, S. V., Yin, W., and Liu, Y. Fedpd: A federated learning framework with optimal rates and adaptivity to non-iid data. *CoRR*, abs/2005.11418, 2020.

A. Preliminary Results and Proof of Convergence for Full Participation

Quick overview. At the heart of the analyses for the approximation ratio of the `Centralized Continuous Greedy` is to show that at each iteration the algorithm reduces the gap to the optimal solution by a significant amount (Călinescu et al., 2011). We follow the same general idea although there are several subtleties that we should address. Mainly, it is required to compare the improvement we obtain by taking direction $\mathbf{v}$ versus the improvement obtained by taking direction $\sum_{i=1}^{N} p_i \mathbf{v}_i$ where $\mathbf{v} = \operatorname{argmax}_{\mathbf{y}\in\mathcal{P}} \langle \mathbf{y}, \nabla \widehat{F}(\mathbf{x}) \rangle$ and $\mathbf{v}_i = \operatorname{argmax}_{\mathbf{y}\in\mathcal{P}} \langle \mathbf{y}, \nabla \widehat{f}_i(\mathbf{x}) \rangle$. To establish a comparison between the two, we need several intermediate lemmas.

We start off by focusing on providing an upper bound on $\texttt{OPT} - \widehat{F}(\mathbf{x})$.

Lemma A.1. *Let $F = \sum_{i=1}^{N} p_i f_i$ be a function where each $f_i : 2^E \to \mathbb{R}_+$ is a monotone submodular function. Suppose $\mathcal{P} \subseteq [0,1]^n$ is a polytope and define $\texttt{OPT} = \max_{\mathbf{x}\in\mathcal{P}} \widehat{F}(\mathbf{x})$. Then for any $\mathbf{x} \in [0,1]^n$ and $\mathbf{v}_i = \operatorname{argmax}_{\mathbf{y}\in\mathcal{P}} \langle \mathbf{y}, \nabla \widehat{f}_i(\mathbf{x}) \rangle$ we have*

$$\texttt{OPT} - \widehat{F}(\mathbf{x}) \leq \sum_{i=1}^{N} p_i \langle \mathbf{v}_i, \nabla \widehat{f}_i(\mathbf{x}) \rangle.$$

Proof of Lemma A.1. First, let us derive some observations on the decomposable submodular function F and its multilinear extension $\widehat{F}$. By definition $\widehat{F}(\mathbf{x}) = \mathbb{E}_{X\sim\mathbf{x}}[F(X)]$. Then

$$\widehat{F}(\mathbf{x}) = \mathbb{E}_{X\sim\mathbf{x}}[F(X)] = \mathbb{E}_{X\sim\mathbf{x}}\left[\sum_{i=1}^{N} p_i f_i\right] = \sum_{i=1}^{N} p_i (\mathbb{E}_{X\sim\mathbf{x}}[f_i(X)]) = \sum_{i=1}^{N} p_i \widehat{f}_i(\mathbf{x})$$

Then the gradient of $\widehat{F}$ at any point is equal to the average of gradients from local functions. That is $\nabla \widehat{F}(\mathbf{x}) = \nabla \left(\sum_{i=1}^{N} p_i \widehat{f}_i(\mathbf{x}) \right) = \sum_{i=1}^{N} p_i \nabla \widehat{f}_i(\mathbf{x})$.

Now let $\mathbf{w} \in \mathcal{P}$ be such that $\widehat{F}(\mathbf{w}) = \texttt{OPT}$.

$$\begin{aligned}
\langle \mathbf{w}, \nabla \widehat{F}(\mathbf{x}) \rangle &\leq \max_{\mathbf{v}\in\mathcal{P}} \langle \mathbf{v}, \nabla \widehat{F}(\mathbf{x}) \rangle = \max_{\mathbf{v}\in\mathcal{P}} \langle \mathbf{v}, \sum_{i=1}^{N} p_i \nabla \widehat{f}_i(\mathbf{x}) \rangle = \max_{\mathbf{v}\in\mathcal{P}} \sum_{i=1}^{N} p_i \langle \mathbf{v}, \nabla \widehat{f}_i(\mathbf{x}) \rangle \\
&\leq \sum_{i=1}^{N} p_i \max_{\mathbf{y}\in\mathcal{P}} \langle \mathbf{y}, \nabla \widehat{f}_i(\mathbf{x}) \rangle = \sum_{i=1}^{N} p_i \langle \mathbf{v}_i, \nabla \widehat{f}_i(\mathbf{x}) \rangle
\end{aligned}$$

In what follows, we prove $\texttt{OPT} - \widehat{F}(\mathbf{x}) \leq \langle \mathbf{w}, \nabla \widehat{F}(\mathbf{x}) \rangle$. Define $\mathbf{d} = (\mathbf{x} \vee \mathbf{w}) - \mathbf{x} = (\mathbf{w} - \mathbf{x}) \vee \mathbf{0}$. By the monotonicity, we have

$$\texttt{OPT} = \widehat{F}(\mathbf{w}) \leq \widehat{F}(\mathbf{x} \vee \mathbf{w}).$$

Note that $\mathbf{d} > \mathbf{0}$, and hence by concavity of $\widehat{F}$ along any positive direction we get

$$\widehat{F}(\mathbf{x} \vee \mathbf{w}) = \widehat{F}(\mathbf{x} + \mathbf{d}) \leq \widehat{F}(\mathbf{x}) + \langle \mathbf{d}, \nabla \widehat{F}(\mathbf{x}) \rangle.$$

Combining the above inequalities we obtain

$$\texttt{OPT} - \widehat{F}(\mathbf{x}) \leq \widehat{F}(\mathbf{x} \vee \mathbf{w}) - \widehat{F}(\mathbf{x}) \leq \langle \mathbf{d}, \nabla \widehat{F}(\mathbf{x}) \rangle$$

Now, since $\nabla \widehat{F}(\mathbf{x})$ is nonnegative and $\mathbf{d} \leq \mathbf{w}$ then we get

$$\texttt{OPT} - \widehat{F}(\mathbf{x}) \leq \langle \mathbf{w}, \nabla \widehat{F}(\mathbf{x}) \rangle$$

This completes the proof. □

The following is an immediate consequence of Lemma A.1.

Corollary A.2. *For any polytope $\mathcal{P} \subseteq [0,1]^n$, $\mathbf{x}^{(t)} \in [0,1]^n$, and $\mathbf{v}_i^{(t)} = \operatorname{argmax}_{\mathbf{y}\in\mathcal{P}} \langle \mathbf{y}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle$, $OPT - \widehat{F}(\mathbf{x}^{(t)}) \leq \sum_{i=1}^N p_i \langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle$.*

Next we use Assumption 3.1 to bound the difference between $\sum_{i=1}^N p_i \langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle$ and $\sum_{i=1}^N p_i \langle \mathbf{v}_i^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle$ at every iteration t.

Lemma A.3 (Bounded heterogeneity)**.** *Let $\mathcal{M}$ be a matroid of rank r and $\mathcal{P}$ be its corresponding matroid polytope. For any $\mathbf{x}^{(t)} \in [0,1]^n$ let $\mathbf{v}_i^{(t)} = \operatorname{argmax}_{\mathbf{y}\in\mathcal{P}} \langle \mathbf{y}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle$ and $\bar{\mathbf{v}}^{(t)} = \sum_{i=1}^N p_i \mathbf{v}_i^{(t)}$. Then, under the bounded gradient dissimilarity assumption 3.1, we have*

$$\langle \bar{\mathbf{v}}^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle \geq \sum_{i=1}^N p_i \langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle - r\gamma_t$$

Proof of Lemma A.3. The proof is straightforward.

$$\begin{aligned}
&\sum_{i=1}^N p_i \langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle - \langle \bar{\mathbf{v}}^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle = \sum_{i=1}^N p_i \langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle - \sum_{i=1}^N p_i \langle \mathbf{v}_i^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle \\
&= \sum_{i=1}^N p_i \langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) - \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle \leq \sum_{i=1}^N p_i r\gamma_t = r\gamma_t
\end{aligned}$$

where in the last inequality we used the fact that each $\mathbf{v}_i^{(t)}$ corresponds to a base in the matroid and we have $\mathbf{v}_i^{(t)} \in \{0,1\}^n$ and $|\mathbf{v}_i^{(t)}|_1 = r$. Given this and Assumption 3.1 we have $|\langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) - \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle| \leq r\gamma_t$. □

A.1. Putting Everything Together: Proof of Theorem 3.2

We now are ready to prove Theorem 3.2.

Proof of Theorem 3.2. Recall that $\mathbf{x}^{(t+1)} \leftarrow \mathbf{x}^{(t)} + \eta\Delta^{(t)}$ where in the full participation case $\Delta^{(t)} = \sum_{i=1}^N p_i \mathbf{v}_i^{(t)}$. According to the Taylor's expansion of the function $\widehat{F}$ near the point $\mathbf{x}^{(t)}$ we can write

$$\widehat{F}(\mathbf{x}^{(t+1)}) = \widehat{F}(\mathbf{x}^{(t)}) + \langle \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle + \frac{1}{2}\langle \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}, \mathbf{H}_{\widehat{F}}(\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)})\rangle \tag{9}$$

$$= \widehat{F}(\mathbf{x}^{(t)}) + \eta\langle \Delta^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle + \frac{\eta^2}{2}\langle \Delta^{(t)}, \mathbf{H}_{\widehat{F}}\Delta^{(t)}\rangle \tag{10}$$

where $\mathbf{H}_{\widehat{F}}$ is the Hessian matrix i.e., the second derivative matrix. We provide a lower bound on each entry of $\mathbf{H}_{\widehat{F}}$. By the result of (Călinescu et al., 2011) and definition of the multilinear extension:

$$\frac{\partial \widehat{F}}{\partial \mathbf{x}(i)\partial \mathbf{x}(j)} = \mathbb{E}_{R\sim\mathbf{x}}[F(R\cup\{i,j\}) - F(R\cup\{i\}\setminus\{j\})] - \mathbb{E}_{R\sim\mathbf{x}}[F(R\cup\{j\}\setminus\{i\}) - F(R\setminus\{i,j\})] \tag{11}$$

$$\geq -\max\{F(\{i\}), F(\{j\})\} \geq -\max_{e\in E} F(\{e\}) = -m_F \tag{12}$$

where the second last inequality is a direct consequence of the submodularity of F. This means every entry of the Hessian is at least $-m_F$. Thus, we arrive at the following lower bound

$$\langle \Delta^{(t)}, \mathbf{H}_{\widehat{F}}\Delta^{(t)}\rangle \geq \sum_{i=1}^n \sum_{j=1}^n \Delta^{(t)}(i)\Delta^{(t)}(j)\mathbf{H}_{\widehat{F}}(i,j) \geq -m_F \sum_{i=1}^n \sum_{j=1}^n \Delta^{(t)}(i)\Delta^{(t)}(j) = -m_F \left(\sum_{i=1}^n \Delta^{(t)}(i)\right)^2 \geq -m_F r^2 \tag{13}$$

where in the last inequality we used the fact that $\Delta^{(t)} \in \mathcal{P}$ as it is a convex combination of vectors from $\mathcal{P}$, and hence $\sum_{i=1}^n \Delta^{(t)}(i) \leq r$.

Thus from (10) and (13) it follows that

$$\widehat{F}(\mathbf{x}^{(t+1)}) \geq \widehat{F}(\mathbf{x}^{(t)}) + \eta\langle\Delta^{(t)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle - \frac{\eta^2 \cdot m_F r^2}{2} \tag{14}$$

It is now required to provide a lower bound for $\langle\Delta^{(t)}, \nabla\widehat{F}(\mathbf{x}^{(t)}\rangle$ in terms of $\texttt{OPT}$. We prove the following lemma.

Lemma A.4. *Let γ_t be as in Assumption 3.1. Then we have $\texttt{OPT} - \widehat{F}(\mathbf{x}^{(t)}) \leq \langle\Delta^{(t)}, \nabla\widehat{F}(\mathbf{x}^{(t)}\rangle + r\gamma_t$.*

Proof of Lemma A.4. By Corollary A.2 we have that

$$\texttt{OPT} - \widehat{F}(\mathbf{x}^{(t)}) \leq \sum_{i=1}^{N} p_i \langle \mathbf{v}_i^{(t)}, \nabla\widehat{f}_i(\mathbf{x}^{(t)})\rangle$$

Furthermore, by Lemma A.3, we have

$$\sum_{i=1}^{N} p_i \langle \mathbf{v}_i^{(t)}, \nabla\widehat{f}_i(\mathbf{x}^{(t)})\rangle - \langle\Delta^{(t)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle \leq r\gamma_t$$

These two give the desired result. Moreover, it is worth noting

$$\sum_{i=1}^{N} p_i \langle \mathbf{v}_i^{(t)}, \nabla\widehat{f}_i(\mathbf{x}^{(t)})\rangle - \langle\Delta^{(t)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle \geq 0$$

This is because

$$\begin{aligned}
&\sum_{i=1}^{N} p_i \langle \mathbf{v}_i^{(t)}, \nabla\widehat{f}_i(\mathbf{x}^{(t)})\rangle - \langle\Delta^{(t)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle = \sum_{i=1}^{N} p_i \langle \mathbf{v}_i^{(t)}, \nabla\widehat{f}_i(\mathbf{x}^{(t)})\rangle - \sum_{i=1}^{N} p_i \langle\Delta^{(t)}, \nabla\widehat{f}_i(\mathbf{x}^{(t)})\rangle \\
&= \sum_{i=1}^{N} p_i \left(\langle \mathbf{v}_i^{(t)}, \nabla\widehat{f}_i(\mathbf{x}^{(t)})\rangle - \langle\Delta^{(t)}, \nabla\widehat{f}_i(\mathbf{x}^{(t)})\rangle\right) \geq 0
\end{aligned}$$

where in the last inequality by definition of $\mathbf{v}_i^{(t)}$ we have $\langle \mathbf{v}_i^{(t)}, \nabla\widehat{f}_i(\mathbf{x}^{(t)})\rangle - \langle\Delta^{(t)}, \nabla\widehat{f}_i(\mathbf{x}^{(t)})\rangle \geq 0$. □

Followed by Lemma A.4 and equation 14, we obtain

$$\widehat{F}(\mathbf{x}^{(t+1)}) - \widehat{F}(\mathbf{x}^{(t)}) \geq \eta\left(\texttt{OPT} - \widehat{F}(\mathbf{x}^{(t)}) - r\gamma_t\right) - \frac{\eta^2 \cdot m_F r^2}{2} \tag{15}$$

Now, by changing signs and adding $\texttt{OPT}$ to both sides, we get

$$\texttt{OPT} - \widehat{F}(\mathbf{x}^{(t+1)}) \leq (1-\eta)\left(\texttt{OPT} - \widehat{F}(\mathbf{x}^{(t)})\right) + \eta r\gamma_t + \frac{\eta^2 \cdot m_F r^2}{2} \tag{16}$$

Applying the same inequality inductively gives

$$\texttt{OPT} - \widehat{F}(\mathbf{x}^{(t+1)}) \leq (1-\eta)^{t+1}\left(\texttt{OPT} - \widehat{F}(\mathbf{0})\right) + \eta r\left(\sum_{t=0}^{T-1}(1-\eta)^{T-t-1}\gamma_t\right) + \frac{(\eta^2 \cdot m_F r^2)(\sum_{t=0}^{T-1}(1-\eta)^{T-t-1})}{2} \tag{17}$$

$$\leq (1-\eta)^{t+1}\left(\texttt{OPT} - \widehat{F}(\mathbf{0})\right) + \eta r\left(\sum_{t=0}^{T-1}\gamma_t\right) + \frac{T\eta^2 \cdot m_F r^2}{2} \qquad ((1-\eta) < 1)$$

Hence for $\mathbf{x}^{(T)}$ we have

$$\left(1-(1-\eta)^T\right)\texttt{OPT} \leq \widehat{F}(\mathbf{x}^{(T)}) + \eta r\left(\sum_{t=0}^{T-1}\gamma_t\right) + \frac{T\eta^2\cdot m_F r^2}{2} \tag{18}$$

Finally, setting $\eta = 1/T$ yields

$$\left(1-\frac{1}{\mathrm{e}}\right)\texttt{OPT} \leq \widehat{F}(\mathbf{x}^{(T)}) + \frac{r\left(\sum_{t=0}^{T-1}\gamma_t\right)}{T} + \frac{m_F r^2}{2T} \tag{19}$$

$$= \widehat{F}(\mathbf{x}^{(T)}) + \frac{rD}{T} + \frac{m_F r^2}{2T} \tag{20}$$

for $\sum_{t=0}^{T-1}\gamma_t = D$. Note that for any $\varepsilon \geq 0$ by setting $T = \max\{\frac{2rD}{\varepsilon}, \frac{m_F r^2}{\varepsilon}\}$ we obtain

$$\left(1-\frac{1}{\mathrm{e}}\right)\texttt{OPT} \leq \widehat{F}(\mathbf{x}^{(T)}) + \varepsilon \tag{21}$$

□

A.2. Upper Bound for γ_t in Assumption 3.1

Inequality (5) in Assumption 3.1 can be derived using monotonicity and submodularity as follows. For simplicity we drop t.

$$\begin{aligned}
|\nabla\widehat{f}_i(\mathbf{x}) - \nabla\widehat{F}(\mathbf{x})|_\infty &= \max_{e\in E}|[\nabla\widehat{f}_i(\mathbf{x})]_e - [\nabla\widehat{F}(\mathbf{x})]_e| \\
&\leq \max_{e\in E}[\nabla\widehat{f}_i(\mathbf{x})]_e + [\nabla\widehat{F}(\mathbf{x})]_e \quad \text{(by monotonicity and submodularity: } [\nabla\widehat{F}(\mathbf{x})]_e, [\nabla\widehat{f}_i(\mathbf{x})]_e \geq 0) \\
&\leq \max_{e\in E}\max_i [\nabla\widehat{f}_i(\mathbf{x})]_e + [\nabla\widehat{f}_i(\mathbf{x})]_e \qquad \text{(using } F = \textstyle\sum_i p_i f_i \text{ and } \sum_i p_i = 1) \\
&= 2\max_{e\in E}\max_i [\nabla\widehat{f}_i(\mathbf{x})]_e \\
&= 2\max_i \max_{e\in E} [\nabla\widehat{f}_i(\mathbf{x})]_e \\
&= 2\max_i m_{f_i}
\end{aligned}$$

B. Proof of Convergence for FedCG (Theorem 3.5)

In order to prove the approximation guarantee of Theorem 3.5, several steps are taken. First we show that our client sampling scheme is unbiased and furthermore we provide an upper bound for the variance.

B.1. Bounding the Variance

Unbiased client selection. Here we prove that our client selection is unbiased (Lemma 3.3).

Proof of Lemma 3.3. Recall that $\langle\Delta^{(t)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle = \langle\frac{1}{|A^{(t)}|}\sum_{i\in A^{(t)}}\mathbf{v}_i^{(t)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle$ where $A^{(t)}$ is a set of size K.

$$\mathbb{E}_{A^{(t)}}\left[\langle\Delta^{(t)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle\right] = \mathbb{E}_{A^{(t)}}\left[\langle\frac{1}{|A^{(t)}|}\sum_{i\in A^{(t)}}\mathbf{v}_i^{(t)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle\right] \tag{22}$$

$$= \frac{1}{K}\mathbb{E}_{A^{(t)}}\left[\langle\sum_{i\in A^{(t)}}\mathbf{v}_i^{(t)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle\right] = \frac{1}{K}\left[\sum_{i\in A^{(t)}}\mathbb{E}_{A^{(t)}}\langle\mathbf{v}_i^{(t)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle\right] \tag{23}$$

$$= \frac{1}{K}\left[K\mathbb{E}_{A^{(t)}}\langle\mathbf{v}_i^{(t)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle\right] \qquad \text{(for an arbitrary } i\in A^{(t)})$$

$$= \mathbb{E}_{A^{(t)}}\left[\langle\mathbf{v}_i^{(t)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle\right] = \sum_{i=1}^N p_i\langle\mathbf{v}_i^{(t)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle = \langle\sum_{i=1}^N p_i\mathbf{v}_i^{(t)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle \tag{24}$$

□

Bounding the variance. Bounding the variance is at the core of our proof of convergence. Recall the definition of γ_t in Assumption 3.1 which plays a pivotal role in bounding the variance.

Proof of Lemma 3.4. Recall $\Delta^{(t)} = \frac{1}{|A^{(t)}|}\sum_{i\in A^{(t)}} \mathbf{v}_i^{(t)}$ and $|A^{(t)}| = K$ and let $\bar{\mathbf{v}}^{(t)} = \sum_{i=1}^N p_i \mathbf{v}_i^{(t)}$. By Lemma 3.3 we have that $\mathbb{E}_{A^{(t)}}\left[\langle \Delta^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle\right] = \langle \sum_{i=1}^N p_i \mathbf{v}_i^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle$. Furthermore, our client sampling samples K clients independently and with replacement. Therefore,

$$\mathrm{Var}\left(\langle \Delta^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle\right) = \frac{1}{K}\mathbb{E}_{A^{(t)}}\left[\left(\langle \mathbf{s}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle - \langle \bar{\mathbf{v}}^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle\right)^2\right] \tag{25}$$

where $\mathbf{s}$ corresponds to an arbitrary client in $A^{(t)}$. Note that both terms $\langle \mathbf{s}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle$ and $\langle \bar{\mathbf{v}}^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle$ are nonnegative. We provide an upper bound on the absolute value of $|\langle \mathbf{s}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle - \langle \bar{\mathbf{v}}^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle|$ by considering two cases.
Case 1. In the first case we have $\langle \mathbf{s}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle \geq \langle \bar{\mathbf{v}}^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle$. Therefore,

$$|\langle \mathbf{s}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle - \langle \bar{\mathbf{v}}^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle| \tag{26}$$

$$= \langle \mathbf{s}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle - \langle \bar{\mathbf{v}}^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle = \sum_{i=1}^N p_i \langle \mathbf{s}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle - \langle \sum_{i=1}^N p_i \mathbf{v}_i^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle \tag{27}$$

$$\leq \sum_{i=1}^N p_i \langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle - \langle \sum_{i=1}^N p_i \mathbf{v}_i^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle = \sum_{i=1}^N p_i \langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) - \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle \tag{28}$$

$$\leq \sum_{i=1}^N p_i r\gamma_t = r\gamma_t \tag{29}$$

where in the last inequality we used the fact that for each $\mathbf{v}_i^{(t)}$ we have $\mathbf{v}_i^{(t)} \in \{0,1\}^n$ and $|\mathbf{v}_i^{(t)}|_1 = r$, and Assumption 3.1 together yield $|\langle \mathbf{v}_i, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) - \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle| \leq r\gamma_t$.

Case 2. In the first case we have $\langle \mathbf{s}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle \leq \langle \bar{\mathbf{v}}^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle$. Therefore,

$$|\langle \mathbf{s}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle - \langle \bar{\mathbf{v}}^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle| = \langle \bar{\mathbf{v}}^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle - \langle \mathbf{s}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle \tag{30}$$

$$= \langle \bar{\mathbf{v}}^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle - \sum_{i=1}^N p_i \langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle - \langle \mathbf{s}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle + \sum_{i=1}^N p_i \langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle \tag{31}$$

$$= \sum_{i=1}^N p_i \langle \mathbf{v}_i^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle - \sum_{i=1}^N p_i \langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle + \left(\sum_{i=1}^N p_i \langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle - \langle \mathbf{s}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle\right) \tag{32}$$

$$\leq \sum_{i=1}^N p_i \left|\langle \mathbf{v}_i^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)}) - \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle\right| + \sum_{i=1}^N p_i \left(\langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle - \langle \mathbf{s}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle\right) \tag{33}$$

$$\leq r\gamma + \sum_{i=1}^N p_i \left(\langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle - \langle \mathbf{s}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle\right) \tag{34}$$

where in the last inequality we used the same argument as the **Case 1**. Now it is left to provide an upper bound for $\langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x})\rangle - \langle \mathbf{s}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle$, which by definition is nonnegative.

$$\langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle - \langle \mathbf{s}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle \tag{35}$$

$$= \langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t))})\rangle - \langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_s(\mathbf{x}^{(t)})\rangle + \langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_s(\mathbf{x}^{(t)})\rangle - \langle \mathbf{s}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle \tag{36}$$

$$= \langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) - \nabla \widehat{f}_s(\mathbf{x}^{(t)})\rangle + \langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_s(\mathbf{x}^{(t)})\rangle - \langle \mathbf{s}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle \tag{37}$$

$$\leq \langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) - \nabla \widehat{f}_s(\mathbf{x}^{(t)})\rangle + \langle \mathbf{s}, \nabla \widehat{f}_s(\mathbf{x}^{(t)})\rangle - \langle \mathbf{s}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle \tag{38}$$

$$= \langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) - \nabla \widehat{f}_s(\mathbf{x}^{(t)}) \rangle + \langle \mathbf{s}, \nabla \widehat{f}_s(\mathbf{x}^{(t)}) - \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle \tag{39}$$

By Assumption 3.1 we know that $|\nabla \widehat{f}_s(\mathbf{x}^{(t)}) - \nabla \widehat{F}(\mathbf{x}^{(t)})|_\infty \leq \gamma_t$ and $|\nabla \widehat{f}_i(\mathbf{x}^{(t)}) - \nabla \widehat{F}(\mathbf{x}^{(t)})|_\infty \leq \gamma_t$. Therefore, $|\nabla \widehat{f}_s(\mathbf{x}^{(t)}) - \nabla \widehat{f}_i(\mathbf{x}^{(t)})|_\infty \leq 2\gamma_t$. Knowing that both $\mathbf{v}_i^{(t)}, \mathbf{s} \in \{0,1\}^n$ and $|\mathbf{v}_i^{(t)}|_1 = |\mathbf{s}|_1 = r$ yields

$$\langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle - \langle \mathbf{s}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle \tag{40}$$

$$\leq \langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) - \nabla \widehat{f}_s(\mathbf{x}^{(t)}) \rangle + \langle \mathbf{s}, \nabla \widehat{f}_s(\mathbf{x}^{(t)}) - \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle \leq 4r\gamma_t \tag{41}$$

Putting together the above inequality and the inequality in equation 34 we obtain the following upper bound for **Case 2**

$$|\langle \mathbf{s}, \nabla \widehat{F}(\mathbf{x}^{(t)}) \rangle - \langle \bar{\mathbf{v}}^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)}) \rangle| \leq 5r\gamma_t.$$

Provided the upper bounds in both cases we have

$$\mathrm{Var}\left(\langle \Delta^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)}) \rangle\right) = \frac{1}{K} \mathbb{E}_{A^{(t)}} \left[\left(\langle \mathbf{s}, \nabla \widehat{F}(\mathbf{x}^{(t)}) \rangle - \langle \bar{\mathbf{v}}^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)}) \rangle \right)^2 \right] \leq \frac{36 r^2 \gamma_t^2}{K}. \tag{42}$$

□

B.2. Putting Everything Together: Proof of Theorem 3.5

We now are ready to prove Theorem 3.5.

Proof of Theorem 3.5. Recall that $\mathbf{x}^{(t+1)} \leftarrow \mathbf{x}^{(t)} + \eta \Delta^{(t)}$. Similar to the proof of Theorem 3.2 equation (14) we derive that

$$\widehat{F}(\mathbf{x}^{(t+1)}) \geq \widehat{F}(\mathbf{x}^{(t)}) + \eta \langle \Delta^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)}) \rangle - \frac{\eta^2 \cdot m_F r^2}{2} \tag{43}$$

Given Lemmas 3.3 and 3.4 and using Chebyshev's inequality, over the random choices of $A^{(t)}$ and for every $\alpha > 0$ we obtain

$$\mathbb{P}\left[\left| \langle \Delta^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)}) \rangle - \langle \sum_{i=1}^N p_i \mathbf{v}_i^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)}) \rangle \right| \leq \frac{6r\gamma_t/\sqrt{K}}{\alpha} \right] \tag{44}$$

$$\geq \mathbb{P}\left[\left| \langle \Delta^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)}) \rangle - \mathbb{E}_{A^{(t)}} \left[\langle \Delta^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)}) \rangle \right] \right| \leq \frac{\sqrt{\mathrm{Var}\left(\langle \Delta^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)}) \rangle \right)}}{\alpha} \right] \geq 1 - \alpha^2 \tag{45}$$

Given this, with probability at least $1 - \alpha^2$ and in the worst case it holds that

$$\widehat{F}(\mathbf{x}^{(t+1)}) \geq \widehat{F}(\mathbf{x}^{(t)}) + \eta \langle \Delta^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)}) \rangle - \frac{\eta^2 \cdot m_F r^2}{2} \tag{46}$$

$$\geq \widehat{F}(\mathbf{x}^{(t)}) + \eta \left(\langle \sum_{i=1}^N p_i \mathbf{v}_i^{(t)}, \nabla \widehat{F}(\mathbf{x}^{(t)}) \rangle - \frac{6r\gamma_t/\sqrt{K}}{\alpha} \right) - \frac{\eta^2 \cdot m_F r^2}{2} \tag{47}$$

$$\geq \widehat{F}(\mathbf{x}^{(t)}) + \eta \left(\sum_{i=1}^N p_i \langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle - r\gamma_t - \frac{6r\gamma_t/\sqrt{K}}{\alpha} \right) - \frac{\eta^2 \cdot m_F r^2}{2} \quad \text{(by Lemma A.3)}$$

$$\geq \widehat{F}(\mathbf{x}^{(t)}) + \eta \left(\mathtt{OPT} - \widehat{F}(\mathbf{x}^{(t)}) \right) - \eta \left(r\gamma_t + \frac{6r\gamma_t/\sqrt{K}}{\alpha} \right) - \frac{\eta^2 \cdot m_F r^2}{2} \quad \text{(by Corollary A.2)}$$

Now, by changing signs and adding $\mathtt{OPT}$ to both sides, we get

$$\mathtt{OPT} - \widehat{F}(\mathbf{x}^{(t+1)}) \leq (1 - \eta) \left(\mathtt{OPT} - \widehat{F}(\mathbf{x}^{(t)}) \right) + \eta \left(r\gamma_t + \frac{6r\gamma_t/\sqrt{K}}{\alpha} \right) + \frac{\eta^2 \cdot m_F r^2}{2} \tag{48}$$

Applying the same inequality inductively gives

$$\texttt{OPT} - \widehat{F}(\mathbf{x}^{(t+1)}) \leq (1-\eta)^{t+1}\left(\texttt{OPT} - \widehat{F}(\mathbf{0})\right) + \eta\left(r\sum_{t=0}^{T-1}(1-\eta)^{T-t-1}\gamma_t + \frac{6r\sum_{t=0}^{T-1}(1-\eta)^{T-t-1}\gamma_t}{\alpha\sqrt{K}}\right) \tag{49}$$

$$+ \frac{\sum_{t=0}^{T-1}(1-\eta)^{T-t-1}\eta^2\cdot m_F r^2}{2} \tag{50}$$

$$\leq (1-\eta)^{t+1}\texttt{OPT} + \eta\left(r\sum_{t=0}^{T-1}\gamma_t + \frac{6r\sum_{t=0}^{T-1}\gamma_t}{\alpha\sqrt{K}}\right) + \frac{T\eta^2\cdot m_F r^2}{2} \tag{51}$$

Taking the union bound over T steps and $\alpha = \sqrt{\frac{\delta}{T}}$, with probability at least $1 - T\cdot\alpha^2 = 1-\delta$, we get

$$\left(1-(1-\eta)^T\right)\texttt{OPT} \leq \widehat{F}(\mathbf{x}^{(T)}) + \eta\left(r\sum_{t=0}^{T-1}\gamma_t + \frac{6r\sum_{t=0}^{T-1}\gamma_t}{\sqrt{K\delta/T}}\right) + \frac{T\eta^2\cdot m_F r^2}{2} \tag{52}$$

Setting $\eta = 1/T$ yields

$$(1-1/\mathrm{e})\,\texttt{OPT} \leq \widehat{F}(\mathbf{x}^{(T)}) + \frac{1}{T}\left(r\sum_{t=0}^{T-1}\gamma_t + \frac{6r\sum_{t=0}^{T-1}\gamma_t}{\sqrt{K\delta/T}}\right) + \frac{m_F r^2}{2T} \tag{53}$$

$$= \widehat{F}(\mathbf{x}^{(T)}) + \frac{rD}{T} + \frac{6rD}{\sqrt{KT\delta}} + \frac{m_F r^2}{2T} \tag{54}$$

for $\sum_{t=0}^{T-1}\gamma_t = D$. Note that for any $\varepsilon \geq 0$ by setting $T = \max\{\frac{3rD}{\varepsilon}, \frac{3m_F r^2}{2\varepsilon}, \frac{324r^2D^2}{K\delta\varepsilon}\}$ we obtain

$$\left(1-\frac{1}{\mathrm{e}}\right)\texttt{OPT} \leq \widehat{F}(\mathbf{x}^{(T)}) + \varepsilon \tag{55}$$

□

C. Proof of Convergence for FEDCG+ (Theorem 4.3)

C.1. Bounding the variance

Unbiased client selection. Here we prove that our client selection is unbiased (Lemma 4.1). The proof is almost identical to the one for Lemma 3.3. We present a proof here for the sake of completeness.

Proof of Lemma 4.1. Recall that $\langle\widetilde{\Delta}^{(t+\tau)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle = \langle\frac{1}{|A^{(t)}|}\sum_{i\in A^{(t)}}\widetilde{\Delta}_i^{(t+\tau)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle$ where $A^{(t)}$ is a set of size K.

$$\mathbb{E}_{A^{(t)}}\left[\langle\widetilde{\Delta}^{(t+\tau)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle\right] = \mathbb{E}_{A^{(t)}}\left[\langle\frac{1}{|A^{(t)}|}\sum_{i\in A^{(t)}}\widetilde{\Delta}_i^{(t+\tau)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle\right] \tag{56}$$

$$(|A^{(t)}| = K)$$

$$= \frac{1}{K}\mathbb{E}_{A^{(t)}}\left[\langle\sum_{i\in A^{(t)}}\widetilde{\Delta}_i^{(t+\tau)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle\right] = \frac{1}{K}\left[\sum_{i\in A^{(t)}}\mathbb{E}_{A^{(t)}}\langle\widetilde{\Delta}_i^{(t+\tau)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle\right] \tag{57}$$

$$= \frac{1}{K}\left[K\mathbb{E}_{A^{(t)}}\langle\widetilde{\Delta}_i^{(t+\tau)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle\right] \qquad \text{(for an arbitrary } i\in A^{(t)})$$

$$= \mathbb{E}_{A^{(t)}}\langle\mathbf{v}_i^{(t)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle = \sum_{i=1}^{N}p_i\langle\widetilde{\Delta}_i^{(t+\tau)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle = \langle\sum_{i=1}^{N}p_i\widetilde{\Delta}_i^{(t+\tau)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle \tag{58}$$

□

Bounding the variance. Bounding the variance is at the core of our proof of convergence. Assumption 3.1 plays a pivotal role in bounding the variance. The main difficulty and difference between this proof and the proof of Lemma 3.4 is that firstly the variance is effected by the divergence caused by local steps, and secondly $\widetilde{\Delta}_i^{(t+\tau)}$ may not be integral vectors and it could potentially have $O(n)$ nonzero entries.

In order to handle the divergence caused by local steps we assume Lipschitzness. First, let us derive useful inequalities using the Lipschitzness condition.

Consequences of Lipschitzness. Consider t-th iteration and recall the definition of L_t i.e., L_t is such that for all $i \in [N]$ and $j \in [\tau]$ it holds

$$\left\|\nabla \widehat{f}_i(\mathbf{x}_i^{(t,j)}) - \nabla \widehat{f}_i(\mathbf{x}^{(t)})\right\| \leq L_t \|\mathbf{x}_i^{(t,j)} - \mathbf{x}^{(t)}\|$$

where $\mathbf{x}_i^{(t)} = \mathbf{x}^{(t)}$ and $\mathbf{x}_i^{(t,j)}$ are the local models for client i at time step t and $t+j$, respectively. We first bound the divergence of $\mathbf{x}_i^{(t,\tau)}$ from $\mathbf{x}^{(t)}$.

$$\left\|\mathbf{x}_i^{(t,\tau)} - \mathbf{x}^{(t)}\right\| = \left\|\frac{1}{\tau}\sum_{j=0}^{\tau-1} \widetilde{\mathbf{v}}_i^{(t,j)}\right\| \leq \frac{1}{\tau}\sum_{j=0}^{\tau-1}\left\|\widetilde{\mathbf{v}}_i^{(t,j)}\right\| \leq \sqrt{r} \tag{59}$$

Since $\widetilde{\mathbf{v}}_i^{(t,j)} \in \mathcal{P}$, the same upper bound holds for every $0 \leq j \leq \tau$; $\|\mathbf{x}_i^{(t,j)} - \mathbf{x}^{(t)}\| \leq \sqrt{r}$. Assuming L_t, for all $0 \leq j \leq \tau$,

$$\left\|\nabla \widehat{f}_i(\mathbf{x}_i^{(t,j)}) - \nabla \widehat{f}_i(\mathbf{x}^{(t)})\right\| \leq L\|\mathbf{x}_i^{(t,j)} - \mathbf{x}^{(t)}\| \leq L_t\sqrt{r} \tag{60}$$

Note $\zeta_i^{(t,j)}$ are sampled according to $\mathbf{x}_i^{(t,j)}$, and $\nabla \widetilde{f}_i(\mathbf{x}_i^{(t,j)}, \zeta_i^{(t,j)})$ estimates $\nabla \widehat{f}_i(\mathbf{x}_i^{(t,j)})$ within factor σ i.e., $\|\nabla \widehat{f}_i(\mathbf{x}_i^{(t,j)}) - \nabla \widetilde{f}_i(\mathbf{x}_i^{(t,j)}, \zeta_i^{(t,j)})\| \leq \sigma$. Hence, given this estimation and equation (60), for every $0 \leq j \leq \tau$ it holds that

$$\left\|\nabla \widetilde{f}_i(\mathbf{x}_i^{(t,j)}, \zeta_i^{(t,j)}) - \nabla \widehat{f}_i(\mathbf{x}^{(t)})\right\| \leq \sigma + L_t\sqrt{r} \tag{61}$$

Proof of Lemma 4.2. By Lemma 4.1 we know:

$$\mathbb{E}_{A^{(t)}}\left[\langle \widetilde{\Delta}^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle\right] = \langle \sum_{i=1}^{N} p_i \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle$$

where $\widetilde{\Delta}^{(t+\tau)} = \frac{1}{|A^{(t)}|}\sum_{i\in A^{(t)}} \widetilde{\Delta}_i^{(t+\tau)}$. Define $\overline{\Delta}^{(t+\tau)} = \sum_{i=1}^{N} p_i \widetilde{\Delta}_i^{(t+\tau)}$. Each client is selected to be in set $A^{(t)}$ independently and with replacement. Therefore,

$$\mathrm{Var}(\langle \widetilde{\Delta}^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle) = \frac{1}{K}\mathbb{E}_{A^{(t)}}\left[\left(\langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle - \langle \overline{\Delta}^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle\right)^2\right] \tag{62}$$

where $\widetilde{\Delta}_s^{(t+\tau)} = \mathbf{x}_s^{(t,\tau)} - \mathbf{x}_s^{(t,0)}$ corresponds to an arbitrary client in $A^{(t)}$. Note that both terms $\langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle$ and $\langle \overline{\Delta}^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle$ are nonnegative. We provide an upper bound on the absolute value of $\left|\langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle - \langle \overline{\Delta}^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle\right|$ by considering two cases.

Case 1. In the first case we have $\langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle \geq \langle \overline{\Delta}^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle$. Therefore,

$$\left|\langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle - \langle \overline{\Delta}^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle\right| \tag{63}$$

$$= \langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle - \langle \overline{\Delta}^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle \tag{64}$$

$$= \sum_{i=1}^{N} p_i \langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle - \langle \sum_{i=1}^{N} p_i \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle \tag{65}$$

Observe that $\widetilde{\Delta}_s^{(t+\tau)} = \frac{1}{\tau}\sum_{j=0}^{\tau-1}\widetilde{\mathbf{v}}_s^{(t,j)}$ with $\widetilde{\mathbf{v}}_s^{(t,j)} = \operatorname{argmax}_{\mathbf{v}\in P(M)}\langle \mathbf{v}, \nabla \widetilde{f}_s(\mathbf{x}_s^{(t,j)}, \zeta_s^{(t,j)})\rangle$.

Therefore, using the Lipschitzness condition we get the following. In what follows let $\mathbf{d}_1 = L_t\sqrt{r}\mathbf{1}$ and $\mathbf{d}_2 = \sigma\mathbf{1}$ be vectors of length n where every components are $L_t\sqrt{r}$ and σ, respectively.

$$\langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}_i^{(t)})\rangle = \frac{1}{\tau}\sum_{j=0}^{\tau-1}\langle \widetilde{\mathbf{v}}_s^{(t,j)}, \nabla \widehat{f}_i(\mathbf{x}_i^{(t)})\rangle \tag{66}$$

$$\leq \frac{1}{\tau}\sum_{j=0}^{\tau-1}\langle \widetilde{\mathbf{v}}_s^{(t,j)}, \nabla \widetilde{f}_i(\mathbf{x}_i^{(t,j)}) + \mathbf{d}_1 + \mathbf{d}_2\rangle \qquad \text{(by equation (61))}$$

$$\leq \frac{1}{\tau}\sum_{j=0}^{\tau-1}\langle \widetilde{\mathbf{v}}_i^{(t,j)}, \nabla \widetilde{f}_i(\mathbf{x}_i^{(t,j)}) + \mathbf{d}_1 + \mathbf{d}_2\rangle \qquad \text{(by definition of } \widetilde{\mathbf{v}}_i^{(t,j)})$$

$$\leq \frac{1}{\tau}\sum_{j=0}^{\tau-1}\langle \widetilde{\mathbf{v}}_i^{(t,j)}, \nabla \widehat{f}_i(\mathbf{x}_i^{(t)})\rangle + \frac{1}{\tau}\sum_{j=0}^{\tau-1}\langle \widetilde{\mathbf{v}}_i^{(t,j)}, \mathbf{d}_1 + \mathbf{d}_2\rangle \tag{67}$$

$$\leq \frac{1}{\tau}\sum_{j=0}^{\tau-1}\langle \widetilde{\mathbf{v}}_i^{(t,j)}, \nabla \widehat{f}_i(\mathbf{x}_i^{(t)})\rangle + (\sigma r + L_t r^{1.5}) \tag{68}$$

$$= \langle \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}_i^{(t)})\rangle + (\sigma r + L_t r^{1.5}) \tag{69}$$

$$= \langle \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle + (\sigma r + L_t r^{1.5}) \qquad (\mathbf{x}_i^{(t)} = \mathbf{x}^{(t)})$$

Therefore, for **Case 1** we get

$$\sum_{i=1}^{N} p_i \langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle - \langle \sum_{i=1}^{N} p_i \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle \tag{70}$$

$$\leq \sum_{i=1}^{N} p_i \langle \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle - \langle \sum_{i=1}^{N} p_i \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle + (\sigma r + L_t r^{1.5}) \tag{71}$$

$$= \sum_{i=1}^{N} p_i \langle \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}) - \nabla \widehat{F}(\mathbf{x})\rangle + (\sigma r + L_t r^{1.5}) \tag{72}$$

$$= \sum_{i=1}^{N} p_i \left(\frac{1}{\tau}\sum_{j=0}^{\tau-1}\langle \widetilde{\mathbf{v}}_i^{(t,j)}, \nabla \widehat{f}_i(\mathbf{x}) - \nabla \widehat{F}(\mathbf{x})\rangle \right) + (\sigma r + L_t r^{1.5}) \tag{73}$$

$$\leq \sum_{i=1}^{N} p_i \left(\frac{1}{\tau}\sum_{j=0}^{\tau-1} r\gamma_t \right) + (\sigma r + L_t r^{1.5}) \qquad \text{(Assumption 3.1)}$$

$$\leq r\gamma_t + (\sigma r + L_t r^{1.5}) \tag{74}$$

Note that in the above we used the fact that for each $\widetilde{\mathbf{v}}_i^{(t,j)}$ we have $\widetilde{\mathbf{v}}_i^{(t,j)} \in \{0,1\}^n$ and $|\widetilde{\mathbf{v}}_i^{(t,j)}|_1 = r$.

Case 2. In this case we have $\langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle \leq \langle \overline{\Delta}^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle$. Therefore,

$$|\langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle - \langle \overline{\Delta}^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle| \tag{75}$$

$$= \langle \overline{\Delta}^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle - \langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle \tag{76}$$

$$= \langle \overline{\Delta}^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle - \sum_{i=1}^{N} p_i \langle \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle - \langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)})\rangle \tag{77}$$

$$+ \sum_{i=1}^{N} p_i \langle \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)})\rangle \tag{78}$$

$$\leq \sum_{i=1}^{N} p_i \left| \langle \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)}) - \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle \right| \tag{79}$$

$$+ \sum_{i=1}^{N} p_i \left(\langle \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle - \langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle \right) \tag{80}$$

$$\leq \sum_{i=1}^{N} \frac{p_i}{\tau} \sum_{j=0}^{\tau-1} \left| \langle \widetilde{\mathbf{v}}_i^{(t,j)}, \nabla \widehat{F}(\mathbf{x}^{(t)}) - \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle \right| \tag{81}$$

$$+ \sum_{i=1}^{N} p_i \left(\langle \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle - \langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle \right) \tag{82}$$

$$\leq r\gamma_t + \sum_{i=1}^{N} p_i \left(\langle \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle - \langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle \right) \tag{83}$$

where in the last inequality we used the same argument by noting $\widetilde{\mathbf{v}}_i^{(t,j)} \in \{0,1\}^n$ and $|\widetilde{\mathbf{v}}_i^{(t,j)}|_1 = r$. Now it is left to provide an upper bound for $\left| \langle \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle - \langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle \right|$.

$$\left| \langle \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle - \langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle \right| \tag{84}$$

$$= \left| \langle \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle - \langle \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{f}_s(\mathbf{x}^{(t)}) \rangle + \langle \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{f}_s(\mathbf{x}^{(t)}) \rangle - \langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle \right| \tag{85}$$

$$= \left| \langle \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) - \nabla \widehat{f}_s(\mathbf{x}^{(t)}) \rangle + \langle \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{f}_s(\mathbf{x}^{(t)}) \rangle - \langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle \right| \tag{86}$$

$$\leq \left| \langle \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) - \nabla \widehat{f}_s(\mathbf{x}^{(t)}) \rangle \right| + \left| \langle \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{f}_s(\mathbf{x}^{(t)}) \rangle - \langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle \right| \tag{87}$$

By Assumption 3.1 we know that $|\nabla \widehat{f}_s(\mathbf{x}^{(t)}) - \nabla \widehat{F}(\mathbf{x}^{(t)})|_\infty \leq \gamma_t$ and $|\nabla \widehat{f}_i(\mathbf{x}^{(t)}) - \nabla \widehat{F}(\mathbf{x}^{(t)})|_\infty \leq \gamma_t$. Therefore, $|\nabla \widehat{f}_s(\mathbf{x}^{(t)}) - \nabla \widehat{f}_i(\mathbf{x}^{(t)})|_\infty \leq 2\gamma_t$. This is used to upper bound the first term in (87). In order to bound the second term in (87) we appeal to equation (69) in **Case 1**. (Note that there are two cases two consider here because of the absolute value, however the argument is similar and we argue about one case.) Let $\mathbf{1}$ be the all one vector of length n, then

$$\left| \langle \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) - \nabla \widehat{f}_s(\mathbf{x}^{(t)}) \rangle \right| + \left| \langle \widetilde{\Delta}_i^{(t+\tau)}, \nabla \widehat{f}_s(\mathbf{x}^{(t)}) \rangle - \langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle \right| \tag{88}$$

$$\leq \frac{1}{\tau} \sum_{j=0}^{\tau-1} \left| \langle \widetilde{\mathbf{v}}_i^{(t,j)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) - \nabla \widehat{f}_s(\mathbf{x}^{(t)}) \rangle \right| + \left| \langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{f}_s(\mathbf{x}^{(t)}) \rangle - \langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle \right| \tag{89}$$

$$+ (\sigma r + L_t r^{1.5}) \tag{90}$$

$$\leq \frac{1}{\tau} \sum_{j=0}^{\tau-1} \left| \langle \widetilde{\mathbf{v}}_i^{(t,j)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) - \nabla \widehat{f}_s(\mathbf{x}^{(t)}) \rangle \right| + \frac{1}{\tau} \sum_{j=0}^{\tau-1} \left| \langle \widetilde{\mathbf{v}}_s^{(t,j)}, \nabla \widehat{f}_s(\mathbf{x}^{(t)}) - \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle \right| \tag{91}$$

$$+ (\sigma r + L_t r^{1.5}) \tag{92}$$

$$\leq \frac{1}{\tau} \sum_{j=0}^{\tau-1} \langle \widetilde{\mathbf{v}}_i^{(t,j)}, 2\gamma_t \mathbf{1} \rangle + \frac{1}{\tau} \sum_{j=0}^{\tau-1} \langle \widetilde{\mathbf{v}}_s^{(t,j)}, 2\gamma_t \mathbf{1} \rangle + (\sigma r + L_t r^{1.5}) \tag{93}$$

$$\leq 4r\gamma_t + (\sigma r + L_t r^{1.5}) \tag{94}$$

where in the last inequality we used the fact that each $\widetilde{\mathbf{v}}_i^{(t,j)}, \widetilde{\mathbf{v}}_s^{(t,j)} \in \{0,1\}^n$ and $|\widetilde{\mathbf{v}}_i^{(t,j)}|_1 = |\widetilde{\mathbf{v}}_s^{(t,j)}|_1 = r$.

Putting together the above inequality and the inequality in equation 83 we obtain the following upper bound for **Case 2**

$$\left| \langle \widetilde{\Delta}_s^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)}) \rangle - \langle \overline{\Delta}^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)}) \rangle \right| \leq 5r\gamma_t + (\sigma r + L_t r^{1.5}).$$

Finally, we are at the place where we can present our upper bound for the variance

$$\mathrm{Var}(\langle \widetilde{\Delta}^{(t+\tau)}, \nabla \widehat{F}(\mathbf{x}^{(t)}) \rangle) \leq \frac{1}{K} \mathbb{E}_{A^{(t)}} \left[\left(6r\gamma + 2(\sigma r + L r^{1.5}) \right)^2 \right]$$

$$= \frac{1}{K}\left(6r\gamma_t + 2(\sigma r + L_t r^{1.5})\right)^2. \qquad \square$$

C.2. Proof of Theorem 4.3

Convergence of FEDCG+. While at the high level this proof is similar to the previous convergence proofs in Theorems 3.2, 3.5, it still requires taking care of the error caused due to the local steps. (This is an additive error term that cannot be controlled by sampling more clients at each round.)

Proof of Theorem 4.3. For any $t \in \mathcal{I}_\tau$ we analyze the difference between $\widehat{F}(\mathbf{x}^{(t)})$ and $\widehat{F}(\mathbf{x}^{(t+\tau)})$. Recall that $\mathbf{x}^{(t+\tau)} \leftarrow \mathbf{x}^{(t)} + \eta\widetilde{\Delta}^{(t+\tau)}$ where η is the server's learning rate and $\widetilde{\Delta}^{(t+\tau)} = \frac{1}{|A^{(t)}|}\sum_{i\in A^{(t)}} \widetilde{\Delta}_i^{(t+\tau)} = \frac{1}{\tau|A^{(t)}|}\sum_{i\in A^{(t)}}\sum_{j=0}^{\tau-1} \widetilde{\mathbf{v}}_i^{(t,j)}$. Similar to the proof of Theorem 3.2 equation (14) we derive that

$$\widehat{F}(\mathbf{x}^{(t+\tau)}) - \widehat{F}(\mathbf{x}^{(t)}) \geq \eta\langle\widetilde{\Delta}^{(t+\tau)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle - \frac{\eta^2\cdot m_F r^2}{2} \tag{95}$$

Let $\overline{\Delta}^{(t+\tau)} = \sum_{i=1}^N p_i\widetilde{\Delta}_i^{(t+\tau)}$ and by Lemma 4.1 we have that $\mathbb{E}_{A^{(t)}}\left[\eta\langle\widetilde{\Delta}^{(t+\tau)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle\right] = \eta\langle\overline{\Delta}^{(t+\tau)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle$. Given Lemma 4.2 and using Chebyshev's inequality, over the random choices of $A^{(t)}$ and for every $\alpha > 0$ and $\chi^2 = \frac{1}{K}\left(6r\gamma_t + 2(\sigma r + L_t r^{1.5})\right)^2$ we obtain

$$\mathbb{P}\left[\left|\langle\widetilde{\Delta}^{(t+\tau)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle - \langle\overline{\Delta}^{(t+\tau)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle\right| \leq \frac{\chi}{\alpha}\right] \geq 1-\alpha^2 \tag{96}$$

Given this and (95), with probability at least $1-\alpha^2$ and in the worst case it holds that

$$\widehat{F}(\mathbf{x}^{(t+\tau)}) \geq \widehat{F}(\mathbf{x}^{(t)}) + \eta\langle\widetilde{\Delta}^{(t+\tau)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle - \frac{\eta^2\cdot m_F r^2}{2} \tag{97}$$

$$\geq \widehat{F}(\mathbf{x}^{(t)}) + \eta\left(\langle\overline{\Delta}^{(t+\tau)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle - \frac{\chi}{\alpha}\right) - \frac{\eta^2\cdot m_F r^2}{2} \tag{98}$$

Claim C.1. *Let γ_t be as in Assumption 3.1, we have $\langle\overline{\Delta}^{(t+\tau)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle \geq \sum_{i=1}^N p_i\langle\widetilde{\Delta}_i^{(t+\tau)}, \nabla\widehat{f}_i(\mathbf{x}^{(t)})\rangle - r\gamma_t$.*

Proof. Proof is similar to the proof of Lemma A.3.

$$\sum_{i=1}^N p_i\langle\widetilde{\Delta}_i^{(t+\tau)}, \nabla\widehat{f}_i(\mathbf{x}^{(t)})\rangle - \langle\overline{\Delta}^{(t+\tau)}, \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle = \sum_{i=1}^N p_i\langle\widetilde{\Delta}_i^{(t+\tau)}, \nabla\widehat{f}_i(\mathbf{x}^{(t)}) - \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle$$
$$= \sum_{i=1}^N\sum_{j=0}^{\tau-1}\frac{p_i}{\tau}\langle\widetilde{\mathbf{v}}_i^{(t,j)}, \nabla\widehat{f}_i(\mathbf{x}^{(t)}) - \nabla\widehat{F}(\mathbf{x}^{(t)})\rangle \leq \sum_{i=1}^N\sum_{j=0}^{\tau-1}\frac{p_i}{\tau}r\gamma_t = r\gamma_t \qquad \square$$

Therefore,

$$\widehat{F}(\mathbf{x}^{(t+\tau)}) \geq \widehat{F}(\mathbf{x}^{(t)}) + \eta(\sum_{i=1}^N p_i\langle\widetilde{\Delta}_i^{(t+\tau)}, \nabla\widehat{f}_i(\mathbf{x}^{(t)})\rangle - r\gamma_t - \frac{\chi}{\alpha}) - \frac{\eta^2\cdot m_F r^2}{2} \tag{99}$$

Claim C.2. $\langle\widetilde{\Delta}_i^{(t+\tau)}, \nabla\widehat{f}_i(\mathbf{x}^{(t)})\rangle \geq \langle\mathbf{v}_i^{(t)}, \nabla\widehat{f}_i(\mathbf{x}^{(t)})\rangle - 2(\sigma r + L_t r^{1.5})$.

Hence, by Claim C.2 and equation (99) we have

$$\widehat{F}(\mathbf{x}^{(t+\tau)}) \geq \widehat{F}(\mathbf{x}^{(t)}) + \eta(\sum_{i=1}^N p_i\langle\widetilde{\Delta}_i^{(t+\tau)}, \nabla\widehat{f}_i(\mathbf{x}^{(t)})\rangle - r\gamma_t - \frac{\chi}{\alpha} - 2(\sigma r + L_t r^{1.5})) - \frac{\eta^2\cdot m_F r^2}{2} \tag{100}$$

Now, by changing signs and adding $\texttt{OPT}$ to both sides, we get

$$\texttt{OPT} - \widehat{F}(\mathbf{x}^{(t+\tau)}) \leq (1-\eta)\left(\texttt{OPT} - \widehat{F}(\mathbf{x}^{(t)})\right) + \eta\left(r\gamma_t + \frac{\chi}{\alpha} + 2(\sigma r + L_t r^{1.5})\right) + \frac{\eta^2 \cdot m_F r^2}{2} \tag{101}$$

Applying the same inequality inductively gives and using $1 > 1-\eta$

$$\texttt{OPT} - \widehat{F}(\mathbf{x}^{(t+\tau)}) \leq (1-\eta)^{t+1}\left(\texttt{OPT} - \widehat{F}(\mathbf{0})\right) \tag{102}$$

$$+ \eta\Big(r\sum_{t\in I_\tau}\gamma_t + \frac{6r\sum_{t\in I_\tau}\gamma_t + 2(\sigma r T/\tau + \sum_{t\in I_\tau} L_t r^{1.5})}{\alpha\sqrt{K}} + 2(\sigma r T/\tau + r^{1.5}\sum_{t\in I_\tau} L_t)\Big) + \frac{T\eta^2\cdot m_F r^2}{2} \tag{103}$$

$$= (1-\eta)^{t+1}\texttt{OPT} \tag{104}$$

$$+ \eta\Big(r\sum_{t\in I_\tau}\gamma_t + \frac{6r\sum_{t\in I_\tau}\gamma_t + 2(\sigma r T/\tau + \sum_{t\in I_\tau} L_t r^{1.5})}{\alpha\sqrt{K}} + 2(\sigma r T/\tau + r^{1.5}\sum_{t\in I_\tau} L_t)\Big) + \frac{T\eta^2\cdot m_F r^2}{2} \tag{105}$$

Taking the union bound over T/τ steps and $\alpha = \sqrt{\frac{\delta\tau}{T}}$, with probability at least $1 - T\cdot\alpha^2/\tau = 1-\delta$, we get

$$\left(1-(1-\eta)^{T/\tau}\right)\texttt{OPT} \leq \widehat{F}(\mathbf{x}^{(T/\tau)}) \tag{106}$$

$$+ \eta\left(r\sum_{t\in I_\tau}\gamma_t + \frac{6r\sum_{t\in I_\tau}\gamma_t + 2(\sigma r T/\tau + \sum_{t\in I_\tau} L_t r^{1.5})}{\sqrt{K\tau\delta/T}} + 2(\sigma r T/\tau + r^{1.5}\sum_{t\in I_\tau} L_t)\right) \tag{107}$$

$$+ \frac{T\eta^2\cdot m_F r^2}{2\tau} \tag{108}$$

Setting $\eta = \tau/T$ yields

$$(1-1/\mathrm{e})\,\texttt{OPT} \leq \widehat{F}(\mathbf{x}^{(T/\tau)}) + \frac{\tau}{T}\left(r\sum_{t\in I_\tau}\gamma_t + \frac{6r\sum_{t\in I_\tau}\gamma_t + 2(\sigma r T/\tau + \sum_{t\in I_\tau} L_t r^{1.5})}{\sqrt{K\tau\delta/T}} + 2(\sigma r T/\tau + r^{1.5}\sum_{t\in I_\tau} L_t)\right) \tag{109}$$

$$+ \frac{\tau\cdot m_F r^2}{2T} \tag{110}$$

$$= \widehat{F}(\mathbf{x}^{(T/\tau)}) + \frac{\tau r D}{T} + \frac{\sqrt{\tau}(6rD + 2r^{1.5}Q)}{\sqrt{KT\delta}} + 2\sigma r\left(\frac{\sqrt{\tau}}{\sqrt{KT\delta}} + 1\right) + \frac{2\tau r^{1.5}Q}{T} + \frac{\tau\cdot m_F r^2}{2T} \tag{111}$$

□

Proof of Claim C.2

Proof of Claim C.2. Recall the definitions of $\widetilde{\Delta}_i^{(t+\tau)}$ and $\mathbf{v}_i^{(t)}$. For $\widetilde{\Delta}_i^{(t+\tau)} = \frac{1}{\tau}\sum_{j=0}^{\tau-1}\widetilde{\mathbf{v}}_i^{(t,j)}$ where for each j we have $\widetilde{\mathbf{v}}_i^{(t,j)} = \mathrm{argmax}_{\mathbf{y}\in\mathcal{P}}\langle \mathbf{y}, \nabla\widetilde{f}_i(\mathbf{x}_i^{(t,j)})\rangle$. Furthermore, $\mathbf{v}_i^{(t)} = \mathrm{argmax}_{\mathbf{y}\in\mathcal{P}}\langle \mathbf{y}, \nabla\widehat{f}_i(\mathbf{x}_i^{(t)})\rangle$. We have,

$$\langle \mathbf{v}_i^{(t)}, \nabla\widehat{f}_i(\mathbf{x}^{(t)})\rangle - \langle \widetilde{\Delta}_i^{(t+\tau)}, \nabla\widehat{f}_i(\mathbf{x}^{(t)})\rangle = \frac{1}{\tau}\sum_{j=0}^{\tau-1}(\langle \mathbf{v}_i^{(t)}, \nabla\widehat{f}_i(\mathbf{x}^{(t)})\rangle - \langle \widetilde{\mathbf{v}}_i^{(t,j)}, \nabla\widehat{f}_i(\mathbf{x}^{(t)})\rangle) \tag{112}$$

On one hand, both $\nabla\widehat{f}_i(\mathbf{x}_i^{(t)})$ and $\nabla\widetilde{f}_i(\mathbf{x}_i^{(t,j)})$ are nonnegative vectors and each $\widetilde{\mathbf{v}}_i^{(t,j)}$ and $\mathbf{v}_i^{(t)}$ corresponds to a maximum-weight independent set in the matroid, with respect to the gradient vectors, and they can be found easily by a greedy algorithm. On the other hand, equation (61) tells us $\left\|\nabla\widetilde{f}_i(\mathbf{x}_i^{(t,j)}, \zeta_i^{(t,j)}) - \nabla\widehat{f}_i(\mathbf{x}^{(t)})\right\| \leq \sigma + L_t\sqrt{r}$.

Let $A = \{e \mid \mathbf{v}_i^{(t)}(e) = 1\}$ be the set of indices where $\mathbf{v}_i^{(t)}$ is one, similarly define $B = \{e \mid \widetilde{\mathbf{v}}_i^{(t,j)}(e) = 1\}$. Then, by definition and equation (61) we get:

$$\langle \mathbf{v}_i^{(t)}, \nabla \widehat{f}_i(\mathbf{x}^{(t)}) \rangle = \sum_{a \in A} \nabla \widehat{f}_i(\mathbf{x}^{(t)})(a) \tag{113}$$

$$\geq \sum_{b \in B} \nabla \widehat{f}_i(\mathbf{x}^{(t)})(b) \geq \sum_{b \in B} \nabla \widetilde{f}_i(\mathbf{x}_i^{(t,j)}, \zeta_i^{(t,j)})(b) - r(\sigma + L_t\sqrt{r}) \tag{114}$$

Similarly, we have

$$\langle \widetilde{\mathbf{v}}_i^{(t,j)}, \nabla \widetilde{f}_i(\mathbf{x}_i^{(t,j)}, \zeta_i^{(t,j)}) \rangle = \sum_{b \in B} \nabla \widetilde{f}_i(\mathbf{x}_i^{(t,j)}, \zeta_i^{(t,j)})(b) \tag{115}$$

$$\geq \sum_{a \in A} \nabla \widetilde{f}_i(\mathbf{x}_i^{(t,j)}, \zeta_i^{(t,j)})(a) \geq \sum_{a \in A} \nabla \widehat{f}_i(\mathbf{x}^{(t)})(a) - r(\sigma + L_t\sqrt{r}). \tag{116}$$

Putting the above two together gives us the desired bound. □

C.3. Gradient Estimation

In terms of computation cost on clients' devices, we point out that the definition of multilinear extension involves summing over all subsets S of E. There are $2^{|E|}$ such subsets, thus even computing $\widehat{g}(\mathbf{x})$ for a single $\mathbf{x}$ could take exponential time. However, we can randomly sample m subsets $R_1, \ldots, R_m$ of E according to $\mathbf{x}$. Then a simple application of Chernoff's bound shows for any multilinear extension $\widehat{g}$ and $\mathbf{x}$, $\left|\frac{1}{m}\sum_{i=1}^{m} g(R_i) - \widehat{g}(\mathbf{x})\right| \leq \sigma \max_S g(S)$ (Vondrák, 2008; Călinescu et al., 2011) with probability at least $1 - \mathrm{e}^{-m\sigma^2/4}$. Observe that $\frac{\partial \widehat{g}}{\partial \mathbf{x}(e)} = \mathbb{E}[g(R \cup \{e\})] - \mathbb{E}[g(R)]$ where R is a random subset of $E \setminus \{e\}$ sampled according to $\mathbf{x}$. Hence, by a similar argument, with m random samples, we can compute an σ-approximation of $\nabla \widehat{g}(\mathbf{x})$ with $1 - \mathrm{e}^{-m\sigma^2/4}$ probability. By suitably choosing m as in Algorithm 2, we can assume the gradients are estimated within $(1 \pm \sigma)$ accuracy with high probability.

The above discussion yields:

Lemma C.3. *Let $\sigma > 0$ be an error for gradient estimation and set $m = O(\log(1/\delta)/\sigma^2)$ for $\delta > 0$. Let $\zeta_i^{(t,j)} = \{R_i^{(t,j,1)}, \ldots, R_i^{(t,j,m)}\}$ be subsets of E that are sampled independently according to i-th client's local model $\mathbf{x}_i^{(t,j)}$. Let $\nabla \widetilde{f}_i(\mathbf{x}_i^{(t,j)}, \zeta_i^{(t,j)})$ denote the* stochastic *gradient for the i-th client that approximates $\nabla \widehat{f}_i(\mathbf{x}_i^{(t,j)})$. Then with $1-\delta$ probability we have $\|\nabla \widehat{f}_i(\mathbf{x}_i^{(t,j)}) - \nabla \widetilde{f}_i(\mathbf{x}_i^{(t,j)}, \zeta_i^{(t,j)})\| \leq \sigma$.*

D. Discrete Algorithm in Federated Setting: Examples

We consider two well-studied problems, namely `Facility Location` and `Maximum Coverage` problems and discuss the details of how to compute importance factors in federated setting efficiently and prove that the expected number of clients participating in each round is small for these two problems.

D.1. Facility Location Problem

Let $\mathcal{C}$ be a set of N clients and E be a set of facilities with $|E| = n$. For $c : \mathcal{C} \times E \to \mathbb{R}$ let the i-th client's score function over a subset of facilities be $f_i(A) = \max_{j \in A} c(i, j)$. The objective for `Max Facility Location` is

$$\max_{S \subseteq E, |S| \leq k} \left\{ F(A) = \sum_{i=1}^{N} \max_{j \in A} c(i, j) \right\}$$

For each client i the importance factor is $\max_{j \in E} \frac{c(i,j)}{F(\{j\})}$.

In many applications computing importance factors is straightforward. Let us elaborate on this with an real-world application.

Movie recommendation system. Consider a movie recommendation application (Stan et al., 2017) where each client i has a user-specific utility function f_i to evaluate sets of movies. The global task is to find a set of k movies that are most

Algorithm 4 Computing importance factors for `Facility Location`

Input: Ground set E
Let $\mathcal{O}$ be a vector of length n. {intention: $\mathcal{O}[i] = F(\{i\})$.}
for each client i in parallel **do**
  Compute $\mathcal{O}_i = [f_i(\{1\}), \dots, f_i(\{|E|\})]$
  Send $\mathcal{O}_i$ back to the secure aggregator.
end for
SecAgg: $\mathcal{O} = \sum_{i\in[N]} \mathcal{O}_i$
Server sends $\mathcal{O}$ to all clients.
for each client i in parallel **do**
  Compute $w_i = \max_{j\in E} \frac{c(i,j)}{\mathcal{O}[j]}$
end for

satisfactory to *all* the clients. An example is the MovieLens dataset consisting of 1 million ratings by $N = 6041$ clients for $|E| = n = 4000$ movies. It is in the interest of clients that we respect their privacy and they are reluctant to share their rating with a central server and other clients. We consider a well motivated objective function. Let $r_{i,j}$ denote the rating of client i for movie j (if such a rating does not exist set $r_{i,j} = 0$). We associate to each client i a facility location objective function $f_i(S) = \max_{j\in S} r(i,j)$ where $S \subseteq E$ is a subset of movies. The servers objective is $\max_{S\subseteq E, |S|\leq k} \frac{1}{N} \sum f_i(S)$. In this example, the average rating of each movie is publicly available. That is $F(\{j\}) = 1/N \sum_{i=1}^{N} f_i(\{j\})$ for each movie j is publicly available. Hence, it is straightforward for each client to compute its own importance factor.

Computing importance factors in federated setting. It is straightforward to see each client can compute its corresponding importance factor in a federated setting and using a secure aggregator without sharing its data with other clients. Each client i sends a vector $(c(i,1), \dots, c(i,n))$ to the server and by simply summing up these vectors the server has a histogram over facilities. This histogram is then broadcasts to the clients where they can compute their own importance factor; see Algorithm 4). Furthermore, Algorithm 4 requires only two communication rounds.

Theorem D.1. *In Algorithm 4, every clients correctly computes its own importance factor. Moreover, Algorithm 4 has only two communication rounds and during each round each client requires only $O(n)$ local function evaluations.*

Having w_i on hand we can execute Algorithm 3 for `Max Facility Location` problem. Note that, in this problem we are dealing with a uniform matroid of rank k.

Theorem D.2. *Suppose clients' importance factors are computed using Algorithm 4 and let $\varepsilon \in (0,1)$. Algorithm 3 after k communication rounds returns a set S of size k such that with probability at least $1 - 1/n$*

$$(1 - 1/\mathrm{e} - \varepsilon)\mathit{OPT} \leq F(S)$$

Moreover, the expected number of clients participating during each round is $\tilde{O}(kn^2/\varepsilon^2)$.

Proof. The approximation guarantee follows from Theorem 5.1. The expected number of clients participating in each round of Algorithm 3 is $\tilde{O}(kn^2/\varepsilon^2)$. This is because

$$\sum_{i=1}^{n} \kappa_i \leq \kappa \sum_{i=1}^{n} w_i \leq \tilde{O}(kn/\varepsilon^2) \sum_{i=1}^{N} w_i = \tilde{O}(kn/\varepsilon^2) \sum_{i=1}^{N} \max_{j\in E} \frac{c(i,j)}{F(\{j\})} \leq \tilde{O}(kn/\varepsilon^2) \sum_{j=1}^{|E|} \frac{\sum_{i\in I} c(i,j)}{F(\{j\})}$$

$$= \tilde{O}(kn/\varepsilon^2) \sum_{j=1}^{|E|} \frac{F(\{j\})}{F(\{j\})} = \tilde{O}(kn^2/\varepsilon^2)$$

□

D.2. Maximum Coverage Problem

Let $\mathcal{C} = \{C_1, \dots, C_N\}$ be a set of clients and $E = \{G_1, \dots, G_n\}$ be a family of sets where each $G_i \subseteq \mathcal{C}$ is a group of clients. Given a positive integer k, in the `Max Coverage` problem the objective is to select at most k groups of clients

Algorithm 5 Computing importance factors for `Max Coverage`

Let $\mathcal{O}$ be a vector of length n. {intention: $\mathcal{O}[i] = |G_i|$.}
for each client i in parallel **do**
  Compute vector $\mathcal{O}_i \in \{0,1\}^n$

$$\begin{cases} \mathcal{O}_i[a] = 1 & \text{if } C_i \in G_a \\ \mathcal{O}_i[a] = 0 & \text{otherwise} \end{cases}$$

  Send $\mathcal{O}_i$ back to the secure aggregator.
end for
Secure aggregator computes $\mathcal{O} = \sum_{i\in[N]} \mathcal{O}_i$ and sends it to the server.
Server sends $\mathcal{O}$ to all clients.
for each client i in parallel **do**
  Compute $w_i = \max\limits_{C_i \in G_a} \frac{1}{|\mathcal{O}[a]|}$
end for

from E such that the maximum number of clients are covered, i.e., the union of the selected groups has maximal size. One can formulate this problem as follows. For every $i \in [N]$ and $A \subseteq [n]$ define $f_i(A)$ as

$$f_i(A) = \begin{cases} 1 & \text{if there exists } a \in A \text{ such that } C_i \in G_a, \\ 0 & \text{otherwise.} \end{cases}$$

Note that f_i's are monotone and submodular. Furthermore, define $F(A) = \sum_{i\in[N]} f_i(A)$ which is monotone and submodular as well. Now the `Max Coverage` problem is equivalent to

$$\max_{A\subseteq[n], |A|\le k} \left\{ F(A) = \sum_{i\in[N]} f_i(A) \right\} \tag{117}$$

For each client C_i, its importance factor w_i is $\max\limits_{G_a \in E, C_i \in G_a} \frac{1}{|G_a|}$.

Computing importance factors in federated setting. Having a histogram over the group sizes suffices for computing the importance factors. Similar to `Facility Location` the importance factors can be computed in federated setting by having the membership histograms over the groups; see Algorithm 5.

Theorem D.3. *In Algorithm 5, every clients correctly computes its own importance factor without revealing which groups they belong to. Moreover, Algorithm 5 has only two communication rounds.*

Remark D.4. We point out that a simple algorithm where at each round clients share their membership with the server using SecAgg protocols is applicable here. However, such algorithm requires full client participation at each round. This can be resolved by sampling sufficiently many clients at each round.

Having w_i, we can execute Algorithm 3 for `Max Coverage` problem. Note that, in this problem we are dealing with a uniform matroid of rank k.

Theorem D.5. *Suppose clients' importance factors are computed using Algorithm 5 and let $\varepsilon \in (0,1)$. Algorithm 3 after k communication rounds returns a set $S \subseteq E$ of size k such that with probability at least $1 - 1/n$*

$$(1 - 1/\mathrm{e} - \varepsilon)\mathit{OPT} \le F(S)$$

Moreover, the expected number of clients participating during each round is $\tilde{O}(kn^2/\varepsilon^2)$.

Proof. The approximation guarantee follows from Theorem 5.1. The expected number of clients participating in each round of Algorithm 3 is $\tilde{O}(kn^2/\varepsilon^2)$. This is because

$$\sum_{i=1}^{n} \kappa_i \le \kappa \sum_{i=1}^{n} w_i \le \tilde{O}(kn/\varepsilon^2) \sum_{i=1}^{N} w_i = \tilde{O}(kn/\varepsilon^2) \sum_{i=1}^{N} \max_{G_a \in E, C_i \in G_a} \frac{1}{|G_a|} \le \tilde{O}(kn/\varepsilon^2) \sum_{j=1}^{|E|} \frac{|G_j|}{|G_j|} = \tilde{O}(kn^2/\varepsilon^2) \quad \square$$