First Provable Guarantees for Practical Private FL: Beyond Restrictive Assumptions

Egor Shulgin [1] Grigory Malinovsky [1] Sarit Khirirat [1] Peter Richtárik [1]

Abstract

Federated Learning (FL) enables collaborative training on decentralized data. Differential Privacy (DP) is crucial for FL, but current private methods often rely on unrealistic assumptions (e.g., bounded gradients or heterogeneity), hindering practical application. Existing works that relax these assumptions typically neglect practical FL mainstays like partial client participation or multiple local updates. We introduce Fed-α-NormEC, the first differentially private FL framework providing provable convergence and DP guarantees under standard assumptions while fully supporting these practical elements. Fed-α-NormEC integrates local updates (full and incremental gradient steps), separate server and client stepsizes, and, crucially, partial client participation—essential for real-world deployment and vital for privacy amplification. Our theoretical guarantees are corroborated by experiments on private deep learning tasks.

1. Introduction

Federated Learning (FL) [44; 33] has emerged as a widely adopted framework for collaboratively training machine learning models across multiple devices or organizations without centralized data collection. Despite its advantages, FL poses several unique challenges. One major issue is the communication bottleneck caused by unreliable and comparatively slow network connections between the server and the clients [6]. Another significant challenge is partial client participation, which arises from the practical infeasibility of involving all clients in every communication round. This is due to both the large scale of the client population and the intermittent availability of individual clients [8]. Furthermore, FL systems must address data heterogeneity, as local datasets across clients are typically diverse and not identically distributed [26; 50]. The increasing interest in FL has led to the development of specialized distributed optimization algorithms designed to improve communication efficiency, accommodate partial client participation, and address data heterogeneity [63; 25].

Although FL methods avoid the exchange of raw data by keeping it decentralized, this design alone does not guarantee complete privacy. Despite preventing direct data sharing, FL remains vulnerable to a range of privacy threats. Prior studies [4; 47] have demonstrated that sensitive information can be inferred from the shared model parameters, either by an untrusted central server or by adversaries performing inference attacks.

To ensure privacy and mitigate emerging risks, Differential Privacy (DP) [13] has become a standard framework for providing formal privacy guarantees in machine learning. It offers a principled way to limit the influence of any individual data on the model, thereby reducing the risk of information leakage during training and inference. Differential Privacy mechanisms are integrated into FL methods to provide formal privacy guarantees while supporting effective training in decentralized settings.

To mitigate the risk of information leakage, FL can be extended to ensure theoretical privacy guarantees via differential privacy (DP) [13]. DP is often enforced by a clipping operator that bounds gradient sensitivity, with DP noise added to the updates before communication. Gradient clipping assists with Differential Privacy, as in Differentially Private Stochastic Gradient Descent (DP-SGD) (Abadi et al., 2016), but it also introduces bias that can even prevent convergence [9; 32]. For instance, FedAvg with model clipping does not converge to the global optimal solution for solving a convex quadratic problem [67].

Convergence guarantees for distributed DP methods with clipping are often established under restrictive assumptions, such as bounded gradient norms [66; 37; 39] and/or bounded heterogeneity [51; 35], which may not hold in realistic, highly heterogeneous FL settings. These assumptions effectively downplay the impact of clipping bias, and to the best

[1] King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. Correspondence to: <egor.shulgin@kaust.edu.sa>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

of our knowledge, convergence guarantees remain unclear when this bias is not explicitly addressed.

To eliminate the bias caused by the clipping operator and enable convergence, Error Compensation (EC) [17], also known as the Error Feedback (EF, EF21) mechanism [61; 54], has been applied to methods that use gradient clipping [29]. This technique works by keeping track of the error and adding it back in future updates. While it helps ensure convergence of methods with clipping in the non-private setting, it does not provide convergence guarantees when DP noise is added in the private setting.

Recent works—such as Shulgin et al. [60] and Islamov et al. [24]—proposed the first methods that successfully combine sharp convergence rates with differential privacy guarantees. The former employs EF alongside local momentum, while the latter uses smoothed normalization techniques [65; 5], which serve as alternatives to gradient clipping for controlling sensitivity. Notably, smoothed normalization has been shown to be more robust to parameter choices compared to standard clipping with a tuned threshold.

Both methods leverage the EF mechanism to reduce bias and achieve strong convergence and privacy guarantees under standard assumptions, without requiring bounded gradients or restricted data heterogeneity. However, they are limited to the distributed optimization setting and do not support key Federated Learning features such as partial client participation or multiple local training steps – components that are essential in real-world FL applications. Overall, a formal theoretical analysis of private FL algorithms that include practical elements like partial participation and local training under standard assumptions remains largely unexplored.

Contributions. We describe our contributions below.

• **A practical method private Federated Learning.** We introduce Fed-α-NormEC —a federated learning algorithm that integrates smoothed normalization and the error feedback mechanism EF21 into clients' local updates. Unlike previous approaches, Fed-α-NormEC enables partial client participation and local training through multiple gradient-type steps. It also leverages separate server and local step sizes, offering flexibility in managing the effects of local updates and global aggregation. To reduce the need for full gradient computations, the algorithm incorporates a cyclic incremental gradient method.

• **Convergence guarantees for non-convex, smooth problems under standard assumptions** We establish the convergence of Fed-α-NormEC for minimizing non-convex, smooth objectives without relying on commonly imposed but restrictive assumptions such as bounded gradients or bounded heterogeneity. Our analysis encompasses both local gradient descent and incremental gradient updates. Notably, in the special case of full client participation with a single local gradient step, we recover the convergence guarantees of α-NormEC. For the more practical scenario involving multiple local steps, we provide— to the best of our knowledge—the first convergence analysis of differentially private federated learning methods incorporating local training. Furthermore, by introducing a server-side step size, we are able to disentangle the effects of data heterogeneity and server aggregation, leading to a clearer characterization of their individual contributions to the optimization error.

• **Differential privacy guarantees with amplification via partial participation.** We provide a privacy analysis of the proposed method for both single and multiple local update steps. Specifically, we consider an independent client sampling scheme, where each client participates in each round with probability p, independently of others. Our analysis shows that this partial participation setup enables significant reduction in differential privacy (DP) noise variance via privacy amplification through subsampling.

• **Empirical validations of Fed-α-NormEC on image classification.** We demonstrate the effectiveness of Fed-α-NormEC by applying it to the image classification task on the CIFAR-10 dataset using the ResNet20 architecture. Experiments highlight the impact of key algorithm parameters and client participation levels, corroborating our theoretical insights on convergence and privacy trade-offs. Notably, we show that partial participation, by leveraging privacy amplification, can achieve target accuracy with significantly improved communication efficiency compared to full participation, showcasing Fed-α-NormEC's utility for real-world private deep learning.

2. Related Works

Clipping. Two popular clipping operators for FL algorithms are per-sample clipping and per-update clipping. Per-sample clipping [38] bounds the norm of the local gradient being used to update the local model parameters on each client, and ensures example-level privacy [1]. Per-update clipping [16] limits the bound of the local model update, and preserves user-level privacy [67; 16], which provides stronger privacy guarantee than example-level privacy. The convergence of FL algorithms, such as FedAvg [44] and SCAFFOLD [26], with per-sample and/or per-update clipping was analyzed by [67; 51; 35; 38; 64]. In this paper, we leverage per-update smoothed normalization, introduced by Bu et al. [5] as an alternative to clipping, to design FL algorithms that accommodate local training and differential privacy.

Federated learning with clipping and privacy. A simple yet popular FL algorithm, FedAvg [44], has been adapted to provide differential privacy (DP) by clipping model updates and injecting random noise [45; 16; 62]. These DP-FedAvg algorithms were outperformed by DP-SCAFFOLD [51], a DP version of SCAFFOLD [26]. However, these existing results require restrictive assumptions that do not hold in practice, especially in deep neural network training, such as uniformly bounded stochastic noise [38; 11], bounded gradients [67; 37; 39; 66] (which effectively ignores the impact of clipping bias), and/or bounded heterogeneity [51; 35]. To the best of our knowledge, there has been a recent work by Das et al. [12] that provides convergence guarantees for DP-FedAvg without these restrictive assumptions, but their results are limited to convex, smooth problems and require a stepsize to depend on an inaccessible constant $\Delta_i := f_i(x^\star) - \min_{x\in\mathbb{R}^d} f_i(x)$, where $x^\star = \arg\min_{x\in\mathbb{R}^d} f(x)$. In this paper, we provide convergence guarantees for private FL algorithms with smoothed normalization and error feedback. In particular, our guarantees do not rely on the restrictive assumptions commonly found in prior work, and our theoretical stepsizes can be implemented in practice.

Communcation efficiency. The most common and natural way to reduce communication is by skipping rounds through the use of local updates, which has become a standard approach in federated learning. This strategy has been extensively studied [28; 40; 30; 18; 52]

Another common biased estimator, besides clipping and normalization, is compression, which improves communication efficiency by reducing message size. The convergence of FL algorithms with compression—such as FedAvg [20], local gradient descent [27; 59], and fixed-point methods [10]—has been studied, but typically under the assumption of unbiased compression. While biased compression of local updates has been explored [19], it often requires integration with other techniques for effective gradient tracking. To our knowledge, no FL method to date uses biased compression to both address data heterogeneity and enhance communication efficiency.

Server and local stepsizes. The use of separate server and local stepsizes has been shown to be crucial in federated learning [7; 53; 42]. This separation provides greater flexibility in optimization. The local stepsize helps mitigate the impact of data heterogeneity and controls the variance from local updates [43], while the global (server-side) stepsize manages the aggregation process and stabilizes extrapolation during model updates [36].

Random reshuffling. Random reshuffling, a without-replacement sampling strategy, is widely used in SGD and often outperforms sampling with replacement. Its convergence properties have been extensively studied [48; 21; 57; 65], including in FL settings [49; 55; 41]. Other without-replacement strategies include Shuffle-Once [56] and Incremental Gradient methods [3; 31]. In this work, Fed-α-NormEC can be extended to support Incremental Gradient updates, partial participation, and differential privacy with provable convergence guarantees.

Error feedback. Error feedback, also known as error compensation, has proven effective in enhancing the convergence of distributed gradient algorithms with compressed communication, leading to faster convergence and improved solution accuracy. Popular error feedback mechanisms include EF14 [58], EF21 [54], EF21-SGDM [14], EControl [15], and EFSkip [2]. Beyond compression, error feedback has been adapted by substituting compression with other operators. For instance, EF21 has inspired the development of Clip21 [29] (using clipping instead of compression) and α-NormEC [60] (employing smoothed normalization). In this paper, we contribute by adapting α-NormEC to the FL setting, resulting in Fed-α-NormEC.

3. Preliminaries

Federated optimization problem. Consider an FL setting with the server being connected with M clients over the network. Each client $i \in [1, M]$ has a private dataset. The objective is to determine the vector of model parameters $x \in \mathbb{R}^d$ that solves the following optimization problem:

$$\min_{x\in\mathbb{R}^d} f(x) = \frac{1}{M}\sum_{i=1}^{M} f_i(x), \ f_i(x) := \frac{1}{N}\sum_{j=1}^{N} f_{i,j}(x). \quad (1)$$

Here, $f_{i,j}(x)$ is the loss of the model parameterized by x on training data $j \in [1, N]$ of client $i \in [1, M]$. Also, we assume that the objective functions f, f_i, and $f_{i,j}$ satisfy the following conditions.

Assumption 1. *Consider Problem (1). Assume that each individual function $f_{i,j}(x)$ is L-smooth and bounded below by $f_{i,j}^{\inf} > -\infty$; that each local function $f_i(x)$ is bounded below by $f_i^{\inf} > -\infty$; and that the global objective $f(x)$ is bounded below by $f^{\inf} > -\infty$.*

DP-FedAvg. The simplest FL algorithm for solving Problem (1) is DP-FedAvg [46]. The algorithm contains two steps: local model updating on each client and model aggregation on the server. The server updates the next global model vector x^{k+1} via:

$$x^{k+1} = x^k - \frac{\eta}{B}\left[\sum_{i\in S^k} \Psi(x^k - \mathcal{T}_i(x^k)) + z_i^k\right],$$

where S^k is the subset of $[1, M]$ with size $B \le M$, $\Psi(\cdot)$ is a bounding operation such as clipping or normalization, $\mathcal{T}_i(x^k)$ is the local update performed by client i based on the current global model x^k and its private data associated with the local function $f_i(\cdot)$, and $z_i^k \in \mathbb{R}^d$ is the DP noise. Since $\Psi(\cdot)$ constrains the magnitude of the model update

$\mathcal{T}_i(x^k) - x^k$, we can calibrate the variance of the DP noise z_i^k proportionally to this bound to achieve the desired privacy guarantees. Moreover, the fact that only a subset of B clients communicate with the server in each round leads to a significant reduction in the required noise variance due to the privacy amplification effect of subsampling.

Bias from Clipping or Normalization. Clipping and normalization inherently introduce bias, causing DP-FedAvg to generally not converge, even without the addition of DP noise. For instance, Zhang et al. [67] demonstrates that FedAvg with model clipping fails to converge to the global optimum when solving a convex quadratic problem. Existing analyses of DP-FedAvg often circumvent the impact of this clipping bias by assuming bounded gradients [67; 37; 39; 66]. Acknowledging this limitation, a recent work by [12] attempts to analyze the convergence of DP-FedAvg without relying on the bounded gradient assumption. However, their findings are restricted to convex and smooth problems and necessitate a step size that depends on the inaccessible constant.

4. Fed-α-NormEC

Now, we describe Fed-α-NormEC for solving federated optimization under privacy and communication constraints. The method operates in communication rounds indexed by $k = 0, 1, \dots, K$. At each round, the server broadcasts the current global model x^k to a subset of participating clients. Each selected client then performs a local update based on the received model using a designated operator $\mathcal{T}_i(x^k)$, which may involve gradient descent or other iterative refinement procedures. In addition, each client computes its local memory vector v_i^k, which captures information from previous updates and enables the use of error feedback techniques. This vector is then used to construct the local update that the client sends back to the server. The memory vector is updated according to the following rule:

$$v_i^{k+1} = v_i^k + \beta \text{Norm}_\alpha \left(\frac{x^t - \mathcal{T}_i(x^k)}{\gamma} - v_i^k \right),$$

where $\beta > 0$ controls the update of error compensation and $\gamma > 0$ is a local stepsize associated with local update operator $\mathcal{T}_i(x)$. Note that smoothed normalization operator is defined as $\text{Norm}_\alpha(g) := \frac{1}{\alpha + \|g\|} g$, for some $\alpha \geq 0$. This operator ensures bounded sensitivity of the client update as $\| \text{Norm}_\alpha(g)\| \leq 1$ for any $g \in \mathbb{R}$.
Each client sends its update $\hat{\Delta}_i^k$ to the server with a fixed probability p, independently across clients. The update $\hat{\Delta}_i^k$ from the i^{th} client is defined as $\hat{\Delta}_i^k := q_i^k \text{Norm}_\alpha \left(\frac{x^t - \mathcal{T}_i(x^k)}{\gamma} - v_i^k \right)$ in the case of non-private training, or as $\hat{\Delta}_i^k := q_i^k \left(\text{Norm}_\alpha \left(\frac{x^t - \mathcal{T}_i(x^k)}{\gamma} - v_i^k \right) + z_i^k \right)$ in the private setting. Here, q_i^k is equal to $1/p$ with probability p, and 0 otherwise, modeling partial client participation. The noise vector z_i^k, to ensure differential privacy, is sampled from a Gaussian distribution with zero mean and variance σ_{DP}^2.

Algorithm 1 (DP-)Fed-α-NormEC

```
Input: Tuning parameters $\gamma > 0$, $\beta > 0$, and $\eta \in (0,1)$; normalization parameter $\alpha > 0$; initialized vectors $x^0, v_i^0 \in \mathbb{R}^d$ for $i \in [1, M]$ and $\hat{v}^0 = \frac{1}{M}\sum_{i=1}^M v_i^0$; local fixed-point operators $\mathcal{T}_i(\cdot)$; probability of transmitting the client's local vector to the server $p \in [0,1]$; Gaussian noise with zero mean and $\sigma_{\text{DP}}^2$-variance $z_i^k \in \mathbb{R}^d$.
for each iteration $k = 0, 1, \dots, K$ do
    for each client $i = 1, 2, \dots, M$ in parallel do
        Compute local updating $\mathcal{T}_i(x^k)$
        Compute $\Delta_i^k = \text{Norm}_\alpha\left(\frac{x^k - \mathcal{T}_i(x^k)}{\gamma} - v_i^k\right)$
        Update $v_i^{k+1} = v_i^k + \beta\Delta_i^k$
        Choose $q_i^k = 1/p$ with prob. $p$ and 0 otherwise
        Non-private setting: Transmit $\hat{\Delta}_i^k = q_i^k \Delta_i^k$
        Private setting: Transmit $\hat{\Delta}_i^k = q_i^k(\Delta_i^k + z_i^k)$
    end for
    Server computes $\hat{v}^{k+1} = \hat{v}^k + \frac{\beta}{M}\sum_{i=1}^M \hat{\Delta}_i^k$
    Server updates $x^{k+1} = x^k - \frac{\eta}{\|\hat{v}^{k+1}\|}\left(\hat{v}^{k+1}\right)$
end for
Output: $x^{K+1}$
```

Next, the server aggregates the normalized local update vectors received from the clients and computes the global memory vector $\hat{v}^k$ and the server updates the global model x^{k+1} using the normalized step as follows:

$$\hat{v}^{k+1} = \hat{v}^k + \frac{\beta}{M}\sum_{i=1}^M \hat{\Delta}_i^k, \quad x^{k+1} = x^k - \frac{\eta}{\|\hat{v}^{k+1}\|}\hat{v}^{k+1},$$

where $\eta > 0$ is the server-side stepsize. The full description is presented in Algorithm 1.

Now, we provide the convergence result for Fed-α-NormEC that incorporates multiple local gradient descent (GD) steps and partial participation in a differentially private setting.

Theorem 1 (Fed-α-NormEC with local GD steps)**.** *Consider Fed-α-NormEC for solving Problem (1) where Assumption 1 holds. Let $\mathcal{T}_i(x^k) = x^k - \frac{\gamma}{T}\sum_{j=0}^{T-1}\nabla f_i(x_i^{k,j})$, where the sequence $\{x_i^{k,j}\}$ is generated by $x_i^{k,j+1} = x_i^{k,j} - \frac{\gamma}{T}\nabla f_i(x_i^{k,j})$, for $j = 0, 1, \ldots, T-1$, given that $x_i^{k,0} = x^k$. Furthermore, let $\beta, \alpha > 0$ be chosen such that $\frac{\beta}{\alpha+R} < 1$ with $R = \max_{i\in[1,M]}\left\|v_i^0 - \frac{x^0 - \mathcal{T}_i(x^0)}{\gamma}\right\|$. If $\eta\gamma \le \frac{1}{K+1}\frac{\Delta^{\inf}}{4L\sqrt{2L}}$, $0 < \eta \le \frac{\gamma}{2}\frac{\beta R}{\alpha+R}$, and $0 < \gamma \le \frac{1}{2L}$, then*

$$\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \le \frac{3}{K+1}\frac{f(x^0)-f^{\inf}}{\eta} + 2R$$

$$+ 2\sqrt{\frac{\beta^2 B}{M}(K+1)} + \gamma \cdot \mathbb{I}_{T\neq 1}\left[8L\sqrt{2L}\sqrt{\Delta^{\inf}}\right] + \eta \cdot \frac{L}{2},$$

for $B = 2\frac{(p-1)^2}{p} + 2\frac{\sigma_{\mathrm{DP}}^2}{p}$, $\Delta^{\inf} = f^{\inf} - \frac{1}{M}\sum_{i=1}^M f_i^{\inf} > 0$.

From Theorem 1, Fed-α-NormEC with multiple local GD steps achieves sub-linear convergence, with additive constants arising from smoothed normalization R, partial participation and private noise $B = 2\frac{(p-1)^2}{p} + \frac{2\sigma_{\mathrm{DP}}^2}{p}$, and data heterogeneity $\Delta^{\inf}$. Our result applies under partial participation, unlike Shulgin et al. [60], which is limited to full participation. Moreover, it accommodates local steps without requiring bounded heterogeneity assumptions, unlike Noble et al. [51]; Li et al. [35].

Fed-α-NormEC with One Local Step. From Theorem 1, we analyze Fed-α-NormEC with a single local step, i.e. with $\mathcal{T}_i(x) = x - \gamma\nabla f_i(x)$, for $i \in [1, M]$, to investigate the impact of smoothed normalization, partial participation, and privacy noise on the convergence.

Full participation and non-private setting. In a full participation and non-private setting, Fed-α-NormEC achieves the convergence according to Theorem 1 with $T = 1$, $p = 1$, and $\sigma_{\mathrm{DP}} = 0$, thus yielding two constant terms, $\mathbb{I}_{T\neq 1}\left[8L\sqrt{2L}\sqrt{\Delta^{\inf}}\right]$ and B, vanish. Therefore, the convergence bound consists of only three terms: $\frac{3}{K+1}\frac{f(x^0)-f^{\inf}}{\eta} + 2R + \frac{\eta L}{2}$. Under this scenario, the convergence bound of Fed-α-NormEC recovers that of α-NormEC as its special case. Similarly to Corollary 1 of Shulgin et al. [60] for α-NormEC, the convergence rate of Fed-α-NormEC with properly tuned hyperparameters η, β, R almost matches that of standard gradient descent at the $\mathcal{O}\left(\frac{1}{\sqrt{K+1}}\right)$ in the gradient norm.

Partial participation and private setting. In a partial participation and private setting, Fed-α-NormEC attains the convergence according to Theorem 1 with $T = 1$. If σ_{DP} is a constant, then proper choices of hyperparameters η, β must be fine-tuned to ensure the convergence of Fed-α-NormEC, as shown below:

Corollary 1. *Consider Fed-α-NormEC for solving Problem (1) under the same setting as Theorem 1. Let $T = 1$ and $N = 0$ (one local GD step). If $v_i^0 \in \mathbb{R}^d$ is chosen such that $\gamma = \frac{1}{2L}$, $\max_{i\in[1,M]}\left\|\frac{x^0-\mathcal{T}_i(x^0)}{\gamma} - v_i^0\right\| = \frac{D_1}{(K+1)^{1/6}}$ with $D_1 > 0$, and $\beta = \frac{D_2}{(K+1)^{2/3}}$ with $D_2 > 0$, and $\eta \le \frac{LD_1D_2}{2(\alpha+D_1)(K+1)^{5/6}}$, then*

$$\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \le \frac{A_1}{(K+1)^{1/6}} + \frac{A_2}{(K+1)^{5/6}},$$

where $A_1 = \frac{f(x^0)-f^{\inf}}{\eta_0} + 2D_1 + 2D_2\sqrt{\frac{2p(1-1/p)^2}{M} + \frac{2\sigma_{DP}^2}{p}}$, $A_2 = \frac{L\eta_0}{2}$, and $\eta_0 = \frac{LD_1D_2}{2(\alpha+D_1)}$.

Corollary 1 establishes the convergence of Fed-α-NormEC in a partial participation and private setting, where the variance σ_{DP} is a constant. In contrast to α-NormEC, the convergence bound for Fed-α-NormEC contains an extra term due to client sampling from partial participation $B = 2\frac{(p-1)^2}{p} + 2\sigma_{\mathrm{DP}}^2/p$. Reducing p decreases bandwidth of participating clients to communicate at each round at the price of a larger error term B. In practice, p is not typically fixed but varying according to the availability of clients at each round. Furthermore, if we assume full client participationg, i.e. $p = 1$, then Fed-α-NormEC achieves the same $\mathcal{O}\left(\frac{1}{(K+1)^{1/6}}\right)$ convergence rate as α-NormEC in the full participation and private setting, where σ_{DP} is a constant.

DP utility bound with privacy amplification. Fed-α-NormEC satisfies (ϵ, δ)-DP and achieves the utility guarantee by setting the standard deviation of the DP noise according to Abadi et al. [1]. We set $\sigma_{\mathrm{DP}} = c \cdot p\sqrt{(K+1)\log(1/\delta)}\epsilon^{-1}$ for some constant $c > 0$ and $0 < p \le 1$. Notably, σ_{DP} exhibits a reduced dependency on p thanks to the amplification effect of subsampling in the local privacy setting. The utility guarantee is given below:

Corollary 2. *Consider Fed-α-NormEC for solving Problem (1) under the same setting as Theorem 1. Let $T = 1$ (one local GD step), let $\sigma_{\mathrm{DP}} = cp\sqrt{(K+1)\log(1/\delta)}/\epsilon$ with $c > 0$, and let $p = \frac{\hat{B}}{M}$ for $\hat{B} \in [1, M]$. If $\beta = \frac{\hat{\beta}}{K+1}$ with $\hat{\beta} = \sqrt{\frac{3(f(x^0)-f^{\inf})}{\gamma}}\sqrt[4]{\frac{M}{B_2}}$, $\gamma < \frac{\Delta^{\inf}(\alpha+R)}{\sqrt{2L}\hat{\beta}R}$ $\alpha = R = \mathcal{O}\left(\sqrt[4]{d}\frac{\sqrt{f(x^0)-f^{\inf}}}{\sqrt{\gamma}}\sqrt[4]{\frac{B_2}{M}}\right)$ with $B_2 = 2c^2\frac{\hat{B}}{M}\frac{\log(1/\delta)}{\epsilon^2}$, and $\eta = \frac{1}{K+1}\frac{\gamma}{2}\frac{\hat{\beta}R}{\alpha+R}$, then*

$$\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \le \mathcal{O}\left(\Delta\sqrt[4]{\frac{d\hat{B}}{M^2}\frac{\log(1/\delta)}{\epsilon^2}}\right),$$

where $\Delta = \max(\alpha, 2)\sqrt{L}\sqrt{f(x^0) - f^{\inf}}$.

Corollary 2 establishes the utility bound of Fed-α-NormEC in the partial participation and private setting. By setting $p = \hat{B}/M$, where $\hat{B} \in [1, M]$ denotes the number of clients sampled at each round, Fed-α-NormEC achieves a utility bound of $\mathcal{O}\left(\Delta\sqrt[4]{d\cdot\frac{B}{M^2}\cdot\frac{\log(1/\delta)}{\epsilon^2}}\right)$, which improves upon the $\mathcal{O}\left(\Delta\sqrt[4]{d\cdot\frac{1}{M}\cdot\frac{\log(1/\delta)}{\epsilon^2}}\right)$ utility bound of α-NormEC. This improvement arises due to privacy amplification via subsampling induced by partial participation. Finally, when $p = 1$ (i.e., under full participation), Fed-α-NormEC recovers the same utility bound as α-NormEC.

Extension to Fed-α-NormEC with multiple local steps. We can extend our findings for Fed-α-NormEC to incorporate both local gradient descent (GD) steps and local incremental gradient (IG) method steps. Detailed information is available in Appendix C.

5. Experiments

We evaluate the performance of Fed-α-NormEC on solving a non-convex optimization task involving deep neural network training. Following the experimental setup from prior work [60] common for DP training, we use the CIFAR-10 dataset [34] and the ResNet20 architecture [22]. Detailed settings and additional results are provided in the Appendix. We analyze the performance of Fed-α-NormEC in the differentially private setting by setting the variance of added noise at $p\beta\sqrt{K\log(1/\delta)}\epsilon^{-1}$ for $\epsilon = 8, \delta = 10^{-5}$ and vary β to simulate different privacy levels. The step size γ is tuned for every combination of parameters p and β. The behavior of test accuracy is shown in Figures 1 and 2, with the corresponding training loss depicted in Figure 5.

The convergence behavior of Fed-α-NormEC as a function of communication rounds is depicted in Figure 1. The plots illustrate performance for Full ($p = 1.0$, solid lines) and Partial client participation ($p = 0.25$, dotted lines) across three settings for the hyperparameter β. The choice of β markedly influences performance. Empirically, $\beta = 0.01$ (orange lines) consistently delivers the best results, achieving the lowest training loss and highest test accuracy for both full and partial participation. For instance, with full participation, $\beta = 0.01$ leads to approximately 70% test accuracy, while $\beta = 0.1$ (green lines) results in the poorest performance (around 55-60% accuracy). Our theory (Theorem 1) supports this sensitivity, as β influences both error feedback and the DP noise term (since $\sigma_{\rm DP} \propto p\beta$). The convergence bound includes a term $\sqrt{\beta^2 B(K+1)/M}$, implying an optimal β balances error compensation and noise.

Per communication round, Full participation ($p = 1.0$) outperforms Partial participation ($p = 0.25$) for a fixed β. This is consistent with Theorem 1: the client sampling variance component of B ($(p-1)^2/p$) is zero for $p = 1$ but positive for $p = 0.25$. Although the DP noise contribution to B ($\sigma_{\rm DP}^2/p \propto p\beta^2$) is smaller for $p = 0.25$, the client sampling variance appears more dominant in round-wise performance. These results underscore the trade-offs in selecting β and the impact of client participation on round-wise performance.

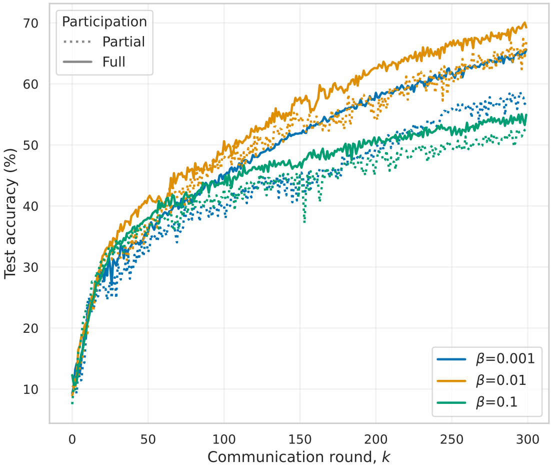


Figure 1: Convergence of Fed-α-NormEC under Full [solid] and Partial participation [dotted] for $p = 0.25$.

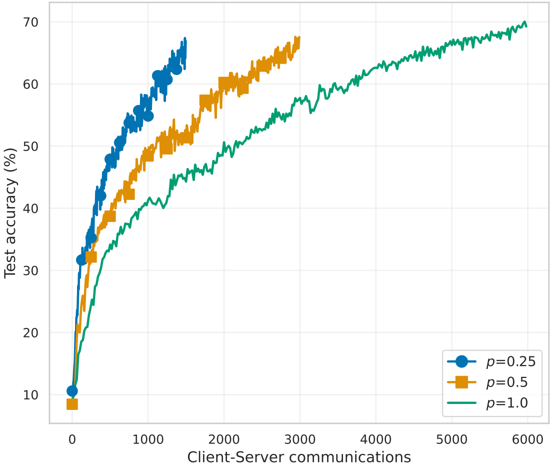


Figure 2: Fed-α-NormEC under varying participation rates; x-axis shows total client-to-server transmissions.

Figure 2 further analyzes Fed-α-NormEC's performance against the total number of client-server communications (i.e., $k \times p \times M$). This visualization offers direct insights into communication efficiency. Notably, configurations with smaller client participation probabilities ($p = 0.25$ and $p = 0.5$) achieve target performance levels with significantly fewer total client-server transmissions compared to full participation ($p = 1.0$). For instance, to reach approximately 65% test accuracy, $p = 0.25$ (blue circles) requires about 1200 total communications, whereas $p = 1.0$ (green line) needs nearly 4500.

6. Acknowledgements

The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST): i) KAUST Baseline Research Scheme, ii) Center of Excellence for Generative AI, under award number 5940, iii) SDAIA-KAUST Center of Excellence in Artificial Intelligence and Data Science.

References

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016. (Cited on pages 2 and 5)

[2] Bao, H., Chen, P., Sun, Y., and Li, Z. EFSkip: A new error feedback with linear speedup for compressed federated learning with arbitrary data heterogeneity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 15489–15497, 2025. (Cited on page 3)

[3] Bertsekas, D. P. et al. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning*, 2010 (1-38):3, 2011. (Cited on page 3)

[4] Boenisch, F., Dziedzic, A., Schuster, R., Shamsabadi, A. S., Shumailov, I., and Papernot, N. When the curious abandon honesty: Federated learning is not private. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, pp. 175–199. IEEE, 2023. (Cited on page 1)

[5] Bu, Z., Wang, Y.-X., Zha, S., and Karypis, G. Automatic clipping: Differentially private deep learning made easier and stronger. *Advances in Neural Information Processing Systems*, 36:41727–41764, 2023. (Cited on page 2)

[6] Caldas, S., Konečny, J., McMahan, H. B., and Talwalkar, A. Expanding the reach of federated learning by reducing client resource requirements. *arXiv preprint arXiv:1812.07210*, 2018. (Cited on page 1)

[7] Charles, Z. and Konečnỳ, J. On the outsized importance of learning rates in local update methods. *arXiv preprint arXiv:2007.00878*, 2020. (Cited on page 3)

[8] Chen, W., Horvath, S., and Richtarik, P. Optimal client sampling for federated learning. *arXiv preprint arXiv:2010.13723*, 2020. (Cited on page 1)

[9] Chen, X., Wu, S. Z., and Hong, M. Understanding gradient clipping in private SGD: A geometric perspective. *Advances in Neural Information Processing Systems*, 33:13773–13782, 2020. (Cited on page 1)

[10] Chraibi, S., Khaled, A., Kovalev, D., Richtárik, P., Salim, A., and Takáč, M. Distributed fixed point methods with compressed iterates. *arXiv preprint arXiv:1912.09925*, 2019. (Cited on page 3)

[11] Crawshaw, M., Bao, Y., and Liu, M. Episode: Episodic gradient clipping with periodic resampled corrections for federated learning with heterogeneous data. *arXiv preprint arXiv:2302.07155*, 2023. (Cited on page 3)

[12] Das, R., Hashemi, A., Sanghavi, S., and Dhillon, I. S. On the convergence of differentially private federated learning on non-lipschitz objectives, and with normalized client updates. *arXiv preprint arXiv:2106.07094*, 2021. (Cited on pages 3 and 4)

[13] Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014. (Cited on page 1)

[14] Fatkhullin, I., Tyurin, A., and Richtárik, P. Momentum provably improves error feedback! *Advances in Neural Information Processing Systems*, 36:76444–76495, 2023. (Cited on page 3)

[15] Gao, Y., Islamov, R., and Stich, S. EControl: Fast distributed optimization with compression and error control. *arXiv preprint arXiv:2311.05645*, 2023. (Cited on page 3)

[16] Geyer, R. C., Klein, T., and Nabi, M. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017. (Cited on pages 2 and 3)

[17] Gorbunov, E., Kovalev, D., Makarenko, D., and Richtárik, P. Linearly converging error compensated sgd. *Advances in Neural Information Processing Systems*, 33:20889–20900, 2020. (Cited on page 2)

[18] Gorbunov, E., Hanzely, F., and Richtárik, P. Local sgd: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics*, pp. 3556–3564. PMLR, 2021. (Cited on page 3)

[19] Gruntkowska, K., Tyurin, A., and Richtárik, P. EF21-P and friends: Improved theoretical communication complexity for distributed optimization with bidirectional compression. In *International Conference on Machine Learning*, pp. 11761–11807. PMLR, 2023. (Cited on page 3)

[20] Haddadpour, F., Kamani, M. M., Mokhtari, A., and Mahdavi, M. Federated learning with compression: Unified analysis and sharp guarantees. In *International Conference on Artificial Intelligence and Statistics*, pp. 2350–2358. PMLR, 2021. (Cited on page 3)

[21] Haochen, J. and Sra, S. Random shuffling beats SGD after finite epochs. In *International Conference on Machine Learning*, pp. 2624–2633. PMLR, 2019. (Cited on page 3)

[22] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. (Cited on page 6)

[23] Idelbayev, Y. Proper ResNet implementation for CIFAR10/CIFAR100 in PyTorch. `https://github.com/akamaster/pytorch_resnet_cifar10`. Accessed: 2024-12-31. (Cited on page 12)

[24] Islamov, R., Horvath, S., Lucchi, A., Richtarik, P., and Gorbunov, E. Double momentum and error feedback for clipping with fast rates and differential privacy. *arXiv preprint arXiv:2502.11682*, 2025. (Cited on page 2)

[25] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021. (Cited on page 1)

[26] Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020. (Cited on pages 1, 2, and 3)

[27] Khaled, A. and Richtárik, P. Gradient descent with compressed iterates. *arXiv preprint arXiv:1909.04716*, 2019. (Cited on page 3)

[28] Khaled, A., Mishchenko, K., and Richtárik, P. Tighter theory for local sgd on identical and heterogeneous data. In *International conference on artificial intelligence and statistics*, pp. 4519–4529. PMLR, 2020. (Cited on page 3)

[29] Khirirat, S., Gorbunov, E., Horváth, S., Islamov, R., Karray, F., and Richtárik, P. Clip21: Error feedback for gradient clipping. *arXiv preprint arXiv:2305.18929*, 2023. (Cited on pages 2, 3, and 13)

[30] Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. A unified theory of decentralized sgd with changing topology and local updates. In *International conference on machine learning*, pp. 5381–5393. PMLR, 2020. (Cited on page 3)

[31] Koloskova, A., Doikov, N., Stich, S. U., and Jaggi, M. On convergence of incremental gradient for non-convex smooth functions. *arXiv preprint arXiv:2305.19259*, 2023. (Cited on page 3)

[32] Koloskova, A., Hendrikx, H., and Stich, S. U. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, pp. 17343–17363. PMLR, 2023. (Cited on page 1)

[33] Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016. (Cited on page 1)

[34] Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, 2009. (Cited on page 6)

[35] Li, B., Jiang, X., Schmidt, M. N., Alstrøm, T. S., and Stich, S. U. An improved analysis of per-sample and per-update clipping in federated learning. In *The Twelfth International Conference on Learning Representations*, 2024. (Cited on pages 1, 2, 3, 5, and 12)

[36] Li, H., Acharya, K., and Richtárik, P. The power of extrapolation in federated learning. *arXiv preprint arXiv:2405.13766*, 2024. (Cited on page 3)

[37] Li, Z., Zhao, H., Li, B., and Chi, Y. SoteriaFL: A unified framework for private federated learning with communication compression. *Advances in Neural Information Processing Systems*, 35:4285–4300, 2022. (Cited on pages 1, 3, and 4)

[38] Liu, M., Zhuang, Z., Lei, Y., and Liao, C. A communication-efficient distributed gradient clipping algorithm for training deep neural networks. *Advances in Neural Information Processing Systems*, 35:26204–26217, 2022. (Cited on pages 2 and 3)

[39] Lowy, A., Ghafelebashi, A., and Razaviyayn, M. Private non-convex federated learning without a trusted server. In *International Conference on Artificial Intelligence and Statistics*, pp. 5749–5786. PMLR, 2023. (Cited on pages 1, 3, and 4)

[40] Malinovskiy, G., Kovalev, D., Gasanov, E., Condat, L., and Richtarik, P. From local sgd to local fixed-point

methods for federated learning. In *International Conference on Machine Learning*, pp. 6692–6701. PMLR, 2020. (Cited on page 3)

[41] Malinovsky, G. and Richtárik, P. Federated random reshuffling with compression and variance reduction. *arXiv preprint arXiv:2205.03914*, 2022. (Cited on page 3)

[42] Malinovsky, G., Horváth, S., Burlachenko, K., and Richtárik, P. Federated learning with regularized client participation. *arXiv preprint arXiv:2302.03662*, 2023. (Cited on page 3)

[43] Malinovsky, G., Mishchenko, K., and Richtárik, P. Server-side stepsizes and sampling without replacement provably help in federated optimization. In *Proceedings of the 4th International Workshop on Distributed Machine Learning*, pp. 85–104, 2023. (Cited on pages 3 and 29)

[44] McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017. (Cited on pages 1 and 2)

[45] McMahan, H. B., Ramage, D., Talwar, K., and Zhang, L. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017. (Cited on page 3)

[46] McMahan, H. B., Ramage, D., Talwar, K., and Zhang, L. Learning differentially private recurrent language models. In *International Conference on Learning Representations*, 2018. (Cited on page 3)

[47] Melis, L., Song, C., De Cristofaro, E., and Shmatikov, V. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 691–706, 2019. (Cited on page 1)

[48] Mishchenko, K., Khaled, A., and Richtárik, P. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33:17309–17320, 2020. (Cited on page 3)

[49] Mishchenko, K., Khaled, A., and Richtárik, P. Proximal and federated random reshuffling. In *International Conference on Machine Learning*, pp. 15718–15749. PMLR, 2022. (Cited on page 3)

[50] Mishchenko, K., Malinovsky, G., Stich, S., and Richtárik, P. Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally! In *International Conference on Machine Learning*, pp. 15750–15769. PMLR, 2022. (Cited on page 1)

[51] Noble, M., Bellet, A., and Dieuleveut, A. Differentially private federated learning on heterogeneous data. In *International conference on artificial intelligence and statistics*, pp. 10110–10145. PMLR, 2022. (Cited on pages 1, 2, 3, 5, and 12)

[52] Patel, K. K., Glasgow, M., Zindari, A., Wang, L., Stich, S. U., Cheng, Z., Joshi, N., and Srebro, N. The limits and potentials of local sgd for distributed heterogeneous learning with intermittent communication. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 4115–4157. PMLR, 2024. (Cited on page 3)

[53] Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečnỳ, J., Kumar, S., and McMahan, H. B. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020. (Cited on page 3)

[54] Richtárik, P., Sokolov, I., and Fatkhullin, I. EF21: a new, simpler, theoretically better, and practically faster error feedback. *Advances in Neural Information Processing Systems*, 34:4384–4396, 2021. (Cited on pages 2 and 3)

[55] Sadiev, A., Malinovsky, G., Gorbunov, E., Sokolov, I., Khaled, A., Burlachenko, K., and Richtárik, P. Federated optimization algorithms with random reshuffling and gradient compression. *arXiv preprint arXiv:2206.07021*, 2022. (Cited on page 3)

[56] Safran, I. and Shamir, O. How good is SGD with random shuffling? In *Conference on Learning Theory*, pp. 3250–3284. PMLR, 2020. (Cited on page 3)

[57] Safran, I. and Shamir, O. Random shuffling beats SGD only after many epochs on ill-conditioned problems. *Advances in Neural Information Processing Systems*, 34:15151–15161, 2021. (Cited on page 3)

[58] Seide, F., Fu, H., Droppo, J., Li, G., and Yu, D. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Interspeech*, volume 2014, pp. 1058–1062. Singapore, 2014. (Cited on page 3)

[59] Shulgin, E. and Richtárik, P. Shifted compression framework: Generalizations and improvements. In *Uncertainty in Artificial Intelligence*, pp. 1813–1823. PMLR, 2022. (Cited on page 3)

[60] Shulgin, E., Khirirat, S., and Richtárik, P. Smoothed normalization for efficient distributed private optimization. *arXiv preprint arXiv:2502.13482*, 2025. (Cited on pages 2, 3, 5, 6, 12, 13, and 15)

[61] Stich, S. U. and Karimireddy, S. P. The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019. (Cited on page 2)

[62] Triastcyn, A. and Faltings, B. Federated learning with bayesian differential privacy. In *2019 IEEE International Conference on Big Data (Big Data)*, pp. 2587–2596. IEEE, 2019. (Cited on page 3)

[63] Wang, J., Charles, Z., Xu, Z., Joshi, G., McMahan, H. B., Al-Shedivat, M., Andrew, G., Avestimehr, S., Daly, K., Data, D., et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021. (Cited on page 1)

[64] Wang, L., Jayaraman, B., Evans, D., and Gu, Q. Efficient privacy-preserving stochastic nonconvex optimization. In *Uncertainty in Artificial Intelligence*, pp. 2203–2213. PMLR, 2023. (Cited on page 2)

[65] Yun, C., Sra, S., and Jadbabaie, A. Can single-shuffle SGD be better than reshuffling SGD and GD? *arXiv preprint arXiv:2103.07079*, 2021. (Cited on pages 2 and 3)

[66] Zhang, X., Fang, M., Liu, J., and Zhu, Z. Private and communication-efficient edge learning: A sparse differential gaussian-masking distributed sgd approach. In *Proceedings of the Twenty-First International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pp. 261–270, 2020. (Cited on pages 1, 3, and 4)

[67] Zhang, X., Chen, X., Hong, M., Wu, Z. S., and Yi, J. Understanding clipping for federated learning: Convergence and client-level differential privacy. In *International Conference on Machine Learning, ICML 2022*, 2022. (Cited on pages 1, 2, 3, and 4)

Contents

A. Conclusion

This paper presented Fed-α-NormEC, the first differentially private federated learning algorithm to offer provable convergence for nonconvex, smooth problems without resorting to unrealistic assumptions such as bounded gradients or heterogeneity. Fed-α-NormEC uniquely combines smoothed normalization and error compensation with essential practical FL components: local updates, distinct server/client learning rates, partial client participation (vital for privacy amplification), and DP noise. Our contributions pave the way for more reliable and deployable private FL systems. Finally, we verify the effectiveness of Fed-α-NormEC by experiments on private deep neural network training.

B. Notations

We use $[a, b]$ for the set $\{a, a+1, \ldots, b\}$ for integers a, b such that $a \leq b$, $\mathrm{E}\left[u\right]$ for the expectation of a random variable u, and $f(x) = \mathcal{O}(g(x))$ if $f(x) \leq Ag(x)$ for some $A > 0$ for functions $f, g : \mathbb{R}^d \to \mathbb{R}$. Finally, for vectors $x, y \in \mathbb{R}^d$, $\langle x, y\rangle$ denotes their inner product, and $\|x\|$ denotes the Euclidean norm of x.

C. Fed-α-NormEC with Multiple Local Steps

In this section, we present the convergence of Fed-α-NormEC with multiple local steps in a partial participation and private setting.

Local GD steps. We obtain the convergence of Fed-α-NormEC with local GD steps in Theorem 1. The convergence bound comprises an error term due to data heterogeneity $\mathbb{I}_{T\neq 1}\left[8L\sqrt{2L}\cdot\sqrt{\Delta^{\inf}}\right]$ where $\Delta^{\inf} = f^{\inf} - \frac{1}{M}\sum_{i=1}^{M} f_i^{\inf}$. Our theorem does not assume bounded heterogeneity that is imposed by Noble et al. [51]; Li et al. [35]. Notably, if all clients share the same infimum, i.e., $f_1^{\inf} = f_2^{\inf} = \ldots = f_M^{\inf}$, this data heterogeneity error term vanishes. Furthermore, this error term is proportional to the local step size γ, due to the presence of a separate server update and distinct server- and client-side step sizes. These theoretical results highlight that the less heterogeneous the client data is, the more effective Fed-α-NormEC becomes.

Local IG steps. To avoid full gradient computations in the clients, we also introduce a variant of Fed-α-NormEC that uses cyclic incremental gradient (IG) steps. In particular, for each client, local updates are performed using gradient steps of the individual loss functions $f_{i,j}$ for each client, applied in a cyclic manner over the local dataset. The local fixed-point operators $\mathcal{T}_i(\cdot)$ are defined as $\mathcal{T}_i(x^k) = x^k - \gamma\cdot\frac{1}{N}\sum_{j=0}^{N-1}\nabla f_{i,j}(x_i^{k,j})$. Here, we focus on the deterministic version of the algorithm, avoiding high-probability analyses that are typically required for methods involving clipping or normalization. Generalization to random reshuffling and arbitrary numbers of epochs is left for future work. Further note that using cyclic incremental gradient updates introduces an additional error term of $\gamma\cdot 4L\sqrt{2L}\cdot\sqrt{\frac{1}{M}\sum_{i=1}^{M}\Delta_i^{\inf}}$, where $\Delta_i^{\inf} = f^{\inf} - \frac{1}{N}\sum_{j=1}^{N} f_{i,j}^{\inf}$. This error vanishes if all functions $f_{i,j}$ share the same infimum $f_i^{\inf}$, in which case we recover the previous result for the local GD setting.

A more detailed discussion of convergence and privacy for the method with local steps, along with formal statements of the theorems, is presented in the supplementary materials.

D. Additional experiments and details

Additional details. All methods are run using a constant learning rate, without auxiliary techniques such as learning rate schedules, warm-up phases, or weight decay. The CIFAR-10 dataset is partitioned into 90% for training and 10% for testing. Training samples are randomly shuffled and evenly distributed across $n = 20$ workers, each using a local batch size of 32. We use a fixed random seed (42) to ensure reproducibility. Our implementation builds upon the publicly available GitHub repository of Idelbayev [23], and all experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU.

We use a fixed smoothed normalization parameter $\alpha = 0.01$, as it was shown to have an insignificant effect on convergence [60]. Server normalization (Line 12 in Algorithm 1) is not used, as omitting it empirically improves final performance [60]. All methods are evaluated across the following hyperparameter combinations: step size $\gamma \in \{0.001, 0.01, 0.1\}$ and sensitivity threshold $\beta \in \{0.001, 0.01, 0.1\}$. We analyze the performance of Fed-α-NormEC in the differentially private setting by training the model for 300 communication rounds.

D.1. Fed-α-NormEC vs FedAvg

We compare the performance of our Algorithm 1 (Fed-α-NormEC) with the standard FedAvg approach, as defined in Section 3:

$$x^{k+1} = x^k - \frac{\eta}{B}\left[\sum_{i\in S^k}\Psi(x^k - \mathcal{T}_i(x^k)) + z_i^k\right],$$

where Ψ is the smoothed normalization operator, $\mathcal{T}_i(x) = x - \gamma\nabla f_i(x)$ is the local gradient mapping, $\eta = \gamma$, and $p = 1$ in the Differentially Private (DP) setting. We follow the same experimental setup as described in Section 5.

Figure 3 presents the convergence of training loss and test accuracy for both methods across different values of the sensitivity parameter β. The results demonstrate that the Error Compensation (EC) mechanism in Fed-α-NormEC consistently accelerates convergence and improves test accuracy compared to FedAvg, across all privacy levels (i.e., all tested values of β). Notably, Fed-α-NormEC achieves its best performance for $\beta = 0.01$, which aligns with the findings in Section 5.

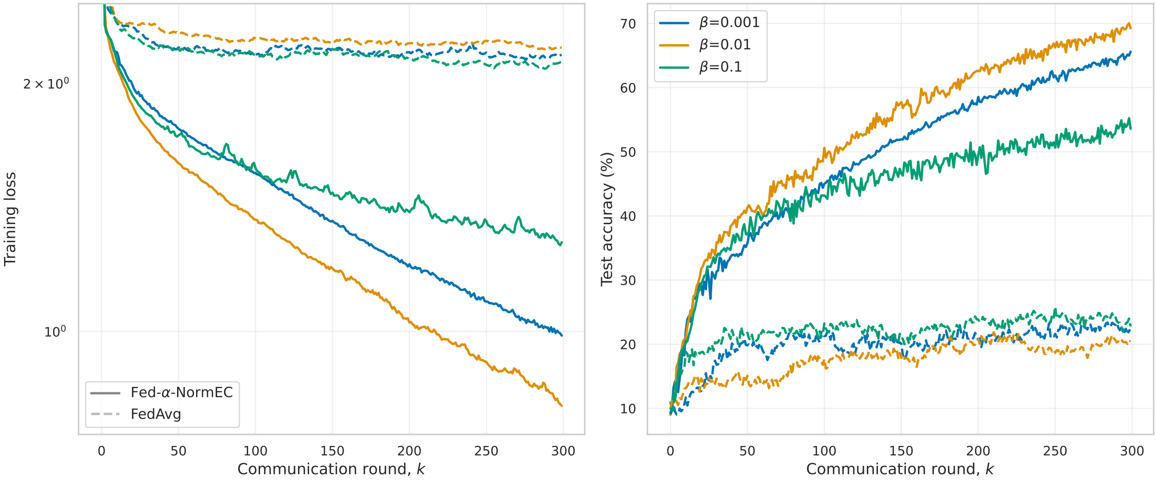
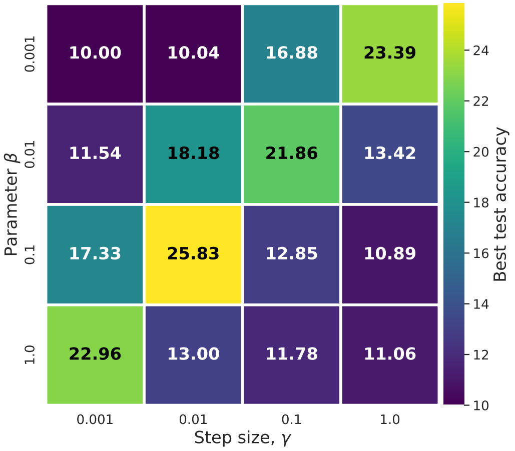


Figure 3: Error Compensation (EC) provides significant benefits across various β values.

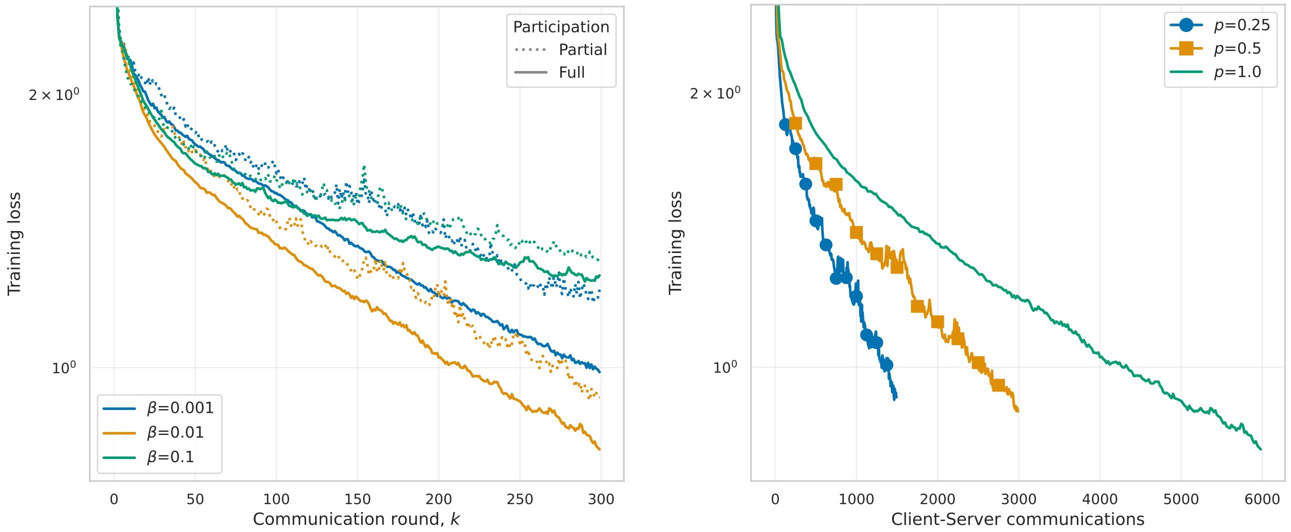
To further analyze the effect of hyperparameters, Figure 4 shows the highest test accuracy achieved by FedAvg for each (β, γ) pair. The optimal performance for FedAvg is observed at $\beta = 0.1$, while the best results are generally found along the diagonal, where the product $\beta \cdot \gamma = 0.001$.



Parameter β \ Step size, γ	0.001	0.01	0.1	1.0
0.001	10.00	10.04	16.88	23.39
0.01	11.54	18.18	21.86	13.42
0.1	17.33	25.83	12.85	10.89
1.0	22.96	13.00	11.78	11.06

Figure 4: The highest test accuracy achieved by FedAvg for different β and γ parameters.

Importantly, prior work [29; 60] has shown that FedAvg with clipping or normalization may fail to converge in certain settings, whereas Fed-α-NormEC remains robust and convergent. Our results further support this observation, highlighting the effectiveness of the Error Compensation mechanism in improving both convergence speed and final accuracy, especially in privacy-constrained federated learning scenarios.



(a) Convergence of Fed-α-NormEC under Full [solid] and Partial participation [dotted] for $p = 0.25$.

(b) Fed-α-NormEC under varying participation rates; x-axis shows total client-to-server transmissions.

Figure 5: Training loss convergence of Fed-α-NormEC corresponding to the test accuracy plots shown in the main text.

E. Useful Lemmas

We introduce useful lemmas for our convergence analysis.

First, Lemma 1 establishes the bounds for $\left\|\frac{x^k-\mathcal{T}_i(x^k)}{\gamma}-v_i^k\right\|$ and $\left\|\frac{x^k-\mathcal{T}_i(x^k)}{\gamma}-v_i^{k+1}\right\|$, two quantities that will be applied in the induction proof to establish the first convergence step of Fed-α-NormEC.

Lemma 1. *Let $v_i^k \in \mathbb{R}^d$ be governed by*

$$v_i^{k+1} = v_i^k + \beta \text{Norm}_\alpha\left(\frac{x^k-\mathcal{T}_i(x^k)}{\gamma} - v_i^k\right), \textit{ for } i \in [1,M] \textit{ and } k \geq 0,$$

and let the fixed-point operator $\mathcal{T}_i(\cdot)$ satisfy

$$\|\mathcal{T}_i(x) - \mathcal{T}_i(y)\| \leq \rho \|x-y\|, \textit{ for } \rho > 0 \textit{ and } x, y \in \mathbb{R}^d.$$

If $\left\|\frac{x^k-\mathcal{T}_i(x^k)}{\gamma}-v_i^k\right\| \leq C$ for some $C > 0$, $\left\|x^{k+1}-x^k\right\| \leq \eta$, $\frac{\beta}{\alpha+C} < 1$, and $\eta \leq \frac{\gamma\beta C}{(1+\rho)(\alpha+C)}$, then $\left\|\frac{x^{k+1}-\mathcal{T}_i(x^{k+1})}{\gamma}-v_i^{k+1}\right\| \leq C$ and $\left\|\frac{x^k-\mathcal{T}_i(x^k)}{\gamma}-v_i^{k+1}\right\| \leq C$.

Proof. From the definition of the Euclidean norm,

$$\begin{aligned} \left\|P_i(x^{k+1}) - v_i^{k+1}\right\| &\overset{v_i^{k+1}}{=} \left\|P_i(x^{k+1}) - v_i^k - \beta N_\alpha(P_i(x^k) - v_i^k)\right\| \\ &\leq \left\|P_i(x^{k+1}) - P_i(x^k)\right\| + \left\|P_i(x^k) - v_i^k - \text{Norm}_\alpha\left(P_i(x^k) - v_i^k\right)\right\|, \end{aligned}$$

where $P_i(x) = (x - \mathcal{T}_i(x))/\gamma$.

Next, by the triangle inequality and by the fact that $\|\mathcal{T}_i(x) - \mathcal{T}_i(y)\| \leq \rho\|x-y\|$ for $\rho > 0$ and $x, y \in \mathbb{R}^d$, we bound the first term:

$$\begin{aligned}\left\|P_i(x^{k+1}) - P_i(x^k)\right\| &= \left\|\frac{x^{k+1} - \mathcal{T}_i(x^{k+1})}{\gamma} - \frac{x^k - \mathcal{T}_i(x^k)}{\gamma}\right\| \\ &\leq \frac{1}{\gamma}\left(\left\|x^{k+1} - x^k\right\| + \left\|\mathcal{T}_i(x^k) - \mathcal{T}_i(x^{k+1})\right\|\right) \\ &\leq \frac{1}{\gamma}(1+\rho)\left\|x^{k+1} - x^k\right\|.\end{aligned}$$

Therefore,

$$\left\|P_i(x^{k+1}) - v_i^{k+1}\right\| \leq \frac{1}{\gamma}(1+\rho)\left\|x^{k+1} - x^k\right\| + \left\|P_i(x^k) - v_i^k - \mathrm{Norm}_\alpha\left(P_i(x^k) - v_i^k\right)\right\|.$$

Next, from Lemma 1 of [60], we can bound the second term:

$$\left\|P_i(x^k) - v_i^k - \mathrm{Norm}_\alpha\left(P_i(x^k) - v_i^k\right)\right\| \leq \left|1 - \frac{\beta}{\alpha + \left\|\frac{x^k - \mathcal{T}_i(x^k)}{\gamma} - v_i^k\right\|}\right| \left\|\frac{x^k - \mathcal{T}_i(x^k)}{\gamma} - v_i^k\right\|.$$

If $\left\|\frac{x^k - \mathcal{T}_i(x^k)}{\gamma} - v_i^k\right\| \leq C$ for some $C > 0$, and $\frac{\beta}{\alpha + C} < 1$, then

$$\begin{aligned}\left\|P_i(x^k) - v_i^k - \mathrm{Norm}_\alpha\left(P_i(x^k) - v_i^k\right)\right\| &\leq \left|1 - \frac{\beta}{\alpha + C}\right| C \\ &\leq \left(1 - \frac{\beta}{\alpha + C}\right) C.\end{aligned}$$

Hence, we obtain

$$\left\|P_i(x^{k+1}) - v_i^{k+1}\right\| \leq \frac{1}{\gamma}(1+\rho)\left\|x^{k+1} - x^k\right\| + \left(1 - \frac{\beta}{\alpha + C}\right) C.$$

If $\left\|x^{k+1} - x^k\right\| \leq \eta$, then

$$\left\|P_i(x^{k+1}) - v_i^{k+1}\right\| \leq \frac{1}{\gamma}(1+\rho)\eta + \left(1 - \frac{\beta}{\alpha + C}\right) C.$$

If $\eta \leq \frac{\gamma}{1+\rho}\frac{\beta C}{(\alpha + C)}$, then $\left\|P_i(x^{k+1}) - v_i^{k+1}\right\| \leq C$. Furthermore, we can show that

$$\begin{aligned}\left\|P_i(x^k) - v_i^{k+1}\right\| &\overset{v_i^{k+1}}{=} \left\|P_i(x^k) - v_i^k - \beta\mathrm{Norm}_\alpha\left(P_i(x^k) - v_i^k\right)\right\| \\ &\overset{\text{Lemma 1 of [60]}}{\leq} \left|1 - \frac{\beta}{\alpha + \left\|\frac{x^k - \mathcal{T}_i(x^k)}{\gamma} - v_i^k\right\|}\right| \left\|\frac{x^k - \mathcal{T}_i(x^k)}{\gamma} - v_i^k\right\| \\ &\overset{\beta/(\alpha+C)<1}{\leq} \left(1 - \frac{\beta}{\alpha + C}\right) C \\ &\leq C.\end{aligned}$$

□

Next, Lemma 2 bounds $\left\|e^k\right\|$ under the recursion of $e^{k+1} = e^k + \beta\frac{1}{M}\sum_{i=1}^M z_i^k$, where z_i^k is the random vector, and by utilizing Lemma 2, we obtain Lemma 3, which bounds $\left\|\frac{1}{M}\sum_{i=1}^M v_i^k - \hat{v}^k\right\|$, the quantity that will be applied to conclude the convergence of Fed-α-NormEC.

Lemma 2. *Let $e^k \in \mathbb{R}^d$ be governed by*

$$e^{k+1} = e^k + \beta z^k, \quad \text{for } 0 \leq k \leq K,$$

where $z^k = \frac{1}{M}\sum_{i=1}^M z_i^k$ and each $z_i^k \in \mathbb{R}^d$ is an independent random vector satisfying

$$\mathrm{E}\left[z_i^k\right] = 0, \quad \text{and} \quad \mathrm{E}\left[\left\|z_i^k\right\|^2\right] \leq \sigma^2.$$

Then,

$$\mathrm{E}\left[\left\|e^{k+1}\right\|\right] \leq \mathrm{E}\left[\left\|e^0\right\|\right] + \sqrt{\frac{\beta^2(K+1)\sigma^2}{M}}.$$

Proof. By applying the recursion of e^{k+1} recursively,

$$e^{k+1} = e^0 + \beta\sum_{l=0}^k z^l.$$

From the definition of the Euclidean norm, and next by the triangle inequality and by taking the expectation,

$$\mathrm{E}\left[\left\|e^{k+1}\right\|\right] \leq \mathrm{E}\left[\left\|e^0\right\|\right] + \mathrm{E}\left[\left\|\beta\sum_{l=0}^k z^l\right\|\right].$$

By Jensen's inequality,

$$\begin{aligned}
\mathrm{E}\left[\left\|e^{k+1}\right\|\right] &\leq \mathrm{E}\left[\left\|e^0\right\|\right] + \sqrt{\mathrm{E}\left[\left\|\beta\sum_{l=0}^k z^l\right\|^2\right]} \\
&= \mathrm{E}\left[\left\|e^0\right\|\right] + \sqrt{\beta^2\sum_{l=0}^k \mathrm{E}\left[\left\|z^l\right\|^2\right] + \beta^2\sum_{i\neq j}\mathrm{E}\left[\langle z^i, z^j\rangle\right]}.
\end{aligned}$$

Since z_i^k is independent of one another, we obtain $\mathrm{E}\left[\left\langle z^i, z^j\right\rangle\right] = 0$ for $i \neq j$, and $\mathrm{E}\left[\left\|z^k\right\|^2\right] = \frac{1}{M}\sum_{i=1}^M \mathrm{E}\left[\left\|z_i^k\right\|^2\right] \leq \sigma^2/n$. Therefore,

$$\mathrm{E}\left[\left\|e^{k+1}\right\|\right] \leq \mathrm{E}\left[\left\|e^0\right\|\right] + \sqrt{\beta^2\frac{(K+1)\sigma^2}{M}}.$$

□

Lemma 3. *Consider Fed-α-NormEC with any local updating operator $\mathcal{T}_i(\cdot)$ for solving Problem (1), where Assumption 1 holds. Then,*

$$\mathrm{E}\left[\left\|\frac{1}{M}\sum_{i=1}^M v_i^{k+1} - \hat{v}^{k+1}\right\|\right] \leq \sqrt{\frac{\beta^2 B}{M}(K+1)},$$

where $B = 2p(1-1/p)^2 + 2(1-p) + 2\sigma_{\mathrm{DP}}^2/p$.

Proof. Define $e^k := \frac{1}{M}\sum_{i=1}^M v_i^k - \hat{v}^k$. Then,

$$\begin{aligned} e^{k+1} &= \frac{1}{M}\sum_{i=1}^M v_i^{k+1} - \hat{v}^{k+1} \\ &\overset{v_i^{k+1}, \hat{v}^{k+1}}{=} \frac{1}{M}\sum_{i=1}^M v_i^k - \hat{v}^k + \beta n^k \\ &= e^k + \beta n^k, \end{aligned}$$

where $n^k = \frac{1}{M}\sum_{i=1}^M n_i^k$ and $n_i^k = (1-q_i^k)\mathrm{Norm}_\alpha\left(\frac{x^k - \mathcal{T}_i(x^k)}{\gamma} - v_i^k\right) - q_i^k z_i^k$.

Next, as q_i^k and z_i^k are independent random vectors, n_i^k is also independent of one another, and satisfies $\mathrm{E}\left[n_i^k\right] = 0$ and

$$\begin{aligned} \mathrm{E}\left[\left\|n_i^k\right\|^2\right] &= \mathrm{E}\left[\left\|(1-q_i^k)\mathrm{Norm}_\alpha\left(\frac{x^k - \mathcal{T}_i(x^k)}{\gamma} - v_i^k\right) - q_i^k z_i^k\right\|^2\right] \\ &\leq 2\mathrm{E}\left[(1-q_i^k)^2\left\|\mathrm{Norm}_\alpha\left(\frac{x^k - \mathcal{T}_i(x^k)}{\gamma} - v_i^k\right)\right\|^2\right] + 2\mathrm{E}\left[(q_i^k)^2\left\|z_i^k\right\|^2\right] \\ &\overset{\|\mathrm{Norm}_\alpha(\cdot)\|\leq 1}{\leq} 2\mathrm{E}\left[(1-q_i^k)^2\right] + 2\mathrm{E}\left[(q_i^k)^2\left\|z_i^k\right\|^2\right] \\ &\overset{q_i^k \text{ and } z_i^k \text{ are independent}}{=} 2\mathrm{E}\left[(1-q_i^k)^2\right] + 2\mathrm{E}\left[(q_i^k)^2\right]\mathrm{E}\left[\left\|z_i^k\right\|^2\right] \\ &\leq 2p(1-1/p)^2 + 2(1-p) + 2p/p^2 \cdot \sigma_{\mathrm{DP}}^2. \end{aligned}$$

Therefore, from Lemma 2 with $z^k = n^k$ and $z_i^k = n_i^k$, we obtain

$$\mathrm{E}\left[\left\|e^{k+1}\right\|\right] \leq \mathrm{E}\left[\left\|e^0\right\|\right] + \sqrt{\frac{\beta^2(K+1)\cdot B}{M}},$$

where $B = 2p(1-1/p)^2 + 2(1-p) + 2\sigma_{\mathrm{DP}}^2/p$. Finally, since $\hat{v}^0 = \frac{1}{n}\sum_{i=1}^n v_i^0$, we obtain $\left\|e^0\right\| = 0$, and complete the proof. □

Finally, Lemma 4 provides the descent inequality for $f(x^k) - f^{\inf}$ in normalized gradient descent. From these established descent inequalities, and Lemma 5 derives the sublinear convergence up to constants.

Lemma 4. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be lower-bounded by $f^{\inf} > -\infty$ and L-smoooth, and let $x^k \in \mathbb{R}^d$ be governed by*

$$x^{k+1} = x^k - \gamma\frac{G^k}{\|G^k\|},$$

where $\gamma > 0$. Then,

$$f(x^{k+1}) - f^{\inf} \leq f(x^k) - f^{\inf} - \gamma\left\|\nabla f(x^k)\right\| + 2\gamma\left\|\nabla f(x^k) - G^k\right\| + \frac{L\gamma^2}{2}.$$

Proof. By the lower-bound and smoothess of $f(\cdot)$, and by the definition of x^{k+1},

$$\begin{aligned} f(x^{k+1}) - f^{\inf} &\leq f(x^k) - f^{\inf} - \frac{\gamma}{\|G^k\|}\left\langle \nabla f(x^k), G^k \right\rangle + \frac{L\gamma^2}{2} \\ &\leq f(x^k) - f^{\inf} - \frac{\gamma}{\|G^k\|}\left\langle G^k, G^k \right\rangle + \frac{\gamma}{\|G^k\|}\left\langle G^k - \nabla f(x^k), G^k \right\rangle + \frac{L\gamma^2}{2} \\ &= f(x^k) - f^{\inf} - \gamma\left\|G^k\right\| + \frac{\gamma}{\|G^k\|}\left\langle G^k - \nabla f(x^k), G^k \right\rangle + \frac{L\gamma^2}{2}. \end{aligned}$$

By Cauchy-Scwartz inequality, i.e. $\langle x, y\rangle \leq \|x\|\,\|y\|$ for $x, y \in \mathbb{R}^d$,

$$f(x^{k+1}) - f^{\inf} \quad \leq \quad f(x^k) - f^{\inf} - \gamma \left\|G^k\right\| + \gamma \left\|\nabla f(x^k) - G^k\right\| + \frac{L\gamma^2}{2}.$$

Finally, by the triangle inequality,

$$f(x^{k+1}) - f^{\inf} \quad \leq \quad f(x^k) - f^{\inf} - \gamma \left\|\nabla f(x^k)\right\| + 2\gamma \left\|\nabla f(x^k) - G^k\right\| + \frac{L\gamma^2}{2}.$$

□

Lemma 5. *Let $\{V^k\}$, $\{W^k\}$ be non-negative sequences satisfying*

$$V^{k+1} \leq (1 + b_1\gamma^2)V^k - b_2\gamma W^k + b_3\gamma.$$

Then,

$$\min_{k\in[0,K]} W^k \leq \frac{\exp(b_1\gamma^2(K+1))}{K+1}\frac{V^0}{b_2\gamma} + \frac{b_3}{b_2}.$$

Proof. Define $w^k := \frac{w^{k+1}}{1+b_1\gamma^2}$ for all $k \geq 0$. Then,

$$\begin{aligned} w^k W^k \quad &\leq \quad \frac{w^k(1+b_1\gamma^2)V^k}{b_2\gamma} - \frac{w^k V^{k+1}}{b_2\gamma} + \frac{b_3}{b_2} \\ &= \quad \frac{w^{k-1}V^k - w^k V^{k+1}}{b_2\gamma} + \frac{b_3}{b_2}. \end{aligned}$$

By summing the inequality over $k = 0, 1, \ldots, K$,

$$\begin{aligned} \sum_{k=0}^{K} w^k W^k \qquad &\leq \qquad \frac{w^{-1}V^0 - w^K V^{K+1}}{b_2\gamma(K+1)} + \frac{b_3}{b_2}\sum_{k=0}^{K} w^k \\ &\overset{w^k, V^k \geq 0}{\leq} \quad \frac{w^{-1}V^0}{b_2\gamma(K+1)} + \frac{b_3}{b_2}\sum_{k=0}^{K} w^k. \end{aligned}$$

Therefore,

$$\begin{aligned} \min_{k\in[0,K]} W^k \quad &\leq \quad \frac{1}{\sum_{k=0}^{K} w^k}\sum_{k=0}^{K} w^k W^k \\ &\leq \quad \frac{w^{-1}V^0}{b_2\gamma(K+1)\sum_{k=0}^{K} w^k} + \frac{b_3}{b_2}. \end{aligned}$$

Next, since

$$\begin{aligned} \sum_{k=0}^{K} w^k \quad &\geq \quad (K+1)\min_{k\in[0,K]} w^k \\ &= \quad (K+1)w^{K+1} \\ &= \quad \frac{(K+1)w^{-1}}{(1+b_1\gamma^2)^{K+1}}, \end{aligned}$$

we get

$$\min_{k\in[0,K]} W^k \leq \frac{(1+b_1\gamma^2)^{K+1}V^0}{b_2\gamma(K+1)} + \frac{b_3}{b_2}.$$

Finally, since $1 + x \leq \exp(x)$, we have $(1+b_1\gamma^2)^{K+1} \leq \exp(b_1\gamma^2(K+1))$. Hence, we obtain the final result.

□

F. Multiple Local GD Steps

We derive the convergence theorem of Fed-α-NormEC using multiple local gradient descent (GD) steps (Theorem 1).

F.1. Key Lemmas

We begin by introducing key lemmas for analyzing Fed-α-NormEC using multiple local GD steps. Lemma 6 bounds $\frac{1}{M}\sum_{i=1}^{M}\|\nabla f_i(x)\|$, while Lemma 7 proves the properties of local GD steps.

Lemma 6. *Let f be bounded from below by $f^{\inf} > -\infty$, and each f_i be bounded from below by $f_i^{\inf} > -\infty$ and L-smooth. Then,*

$$\frac{1}{M}\sum_{i=1}^{M}\|\nabla f_i(x)\| \leq \sqrt{\frac{2L}{\Delta^{\inf}}}[f(x)-f^{\inf}] + \sqrt{2L\Delta^{\inf}},$$

where $\Delta^{\inf} = f^{\inf} - \frac{1}{M}\sum_{i=1}^{M} f_i^{\inf} > 0$.

Proof. Let f be bounded from below by $f^{\inf} > -\infty$, and each f_i be bounded from below by $f_i^{\inf} > -\infty$ and L-smooth. Then,

$$\|\nabla f_i(x)\|^2 \leq 2L[f_i(x) - f_i^{\inf}].$$

Therefore,

$$\frac{1}{M}\sum_{i=1}^{M}\|\nabla f_i(x)\|^2 \leq A[f(x)-f^{\inf}] + B,$$

where $A = 2L$, $B = 2L\Delta^{\inf}$, and $\Delta^{\inf} = f^{\inf} - \frac{1}{M}\sum_{i=1}^{M} f_i^{\inf} > 0$. Thus, we obtain

$$\begin{aligned}
\frac{1}{M}\sum_{i=1}^{M}\|\nabla f_i(x)\| &\overset{\text{Jensen's inequality}}{\leq} \sqrt{\frac{1}{M}\sum_{i=1}^{M}\|\nabla f_i(x)\|^2} \\
&\leq \sqrt{A[f(x)-f^{\inf}]+B} \\
&= \frac{A[f(x)-f^{\inf}]+B}{\sqrt{A[f(x)-f^{\inf}]+B}} \\
&\overset{f(x)\geq f^{\inf}}{\leq} \frac{A}{\sqrt{B}}[f(x)-f^{\inf}] + \sqrt{B}.
\end{aligned}$$

□

Lemma 7. *Let each f_i be L-smooth, and let $\mathcal{T}_i(x^k) = x^k - \frac{\gamma}{T}\sum_{j=0}^{T-1}\nabla f_i(x_i^{k,j})$, where the sequence $\{x_i^{k,l}\}$ is generated by*

$$x_i^{k,l+1} = x_i^{k,l} - \frac{\gamma}{T}\nabla f_i(x_i^{k,l}), \quad \textit{for} \quad l = 0, 1, \ldots, T-1,$$

given that $x_i^{k,0} = x^k$. If $\gamma \le \frac{1}{2L}$, and $\left\|x^{k+1} - x^k\right\| \le \eta$ with $\eta > 0$, then

1. $x_i^{k,l} = x^k - \frac{\gamma}{T}\sum_{j=0}^{l-1}\nabla f_i(x_i^{k,l})$.
2. $\frac{1}{T}\sum_{j=0}^{T-1}\left\|x_i^{k+1,j} - x_i^{k,j}\right\| \le 2\eta$.
3. $\frac{1}{T}\sum_{j=0}^{T-1}\left\|x^k - x_i^{k,j}\right\| \le 2\gamma\left\|\nabla f_i(x^k)\right\|$.
4. $\left\|\mathcal{T}_i(x^{k+1}) - \mathcal{T}_i(x^k)\right\| \le 2\eta$.
5. $\left\|(x^k - \gamma\nabla f_i(x^k)) - \mathcal{T}_i(x^k)\right\| \le 2L\gamma^2\left\|\nabla f_i(x^k)\right\|$.

Proof. We prove the first statement by recursively applying the equation for $x_i^{k,j+1}$ for $j = 0, 1, \ldots, l-1$.

Next, we prove the second statement. From the definition of the Euclidean norm, by the triangle inequality, and by the L-smoothness of $f_i(\cdot)$,

$$\begin{aligned}
\left\|x_i^{k+1,l} - x_i^{k,l}\right\| &\overset{x_i^{k,j}}{=} \left\|x^{k+1} - x^k - \frac{\gamma}{T}\sum_{j=0}^{l-1}(\nabla f_i(x_i^{k+1,j}) - \nabla f_i(x_i^{k,j}))\right\| \\
&\le \left\|x^{k+1} - x^k\right\| + \frac{\gamma}{T}\sum_{j=0}^{l-1}\left\|\nabla f_i(x_i^{k+1,j}) - \nabla f_i(x_i^{k,j})\right\| \\
&\le \left\|x^{k+1} - x^k\right\| + \frac{L\gamma}{T}\sum_{j=0}^{l-1}\left\|x_i^{k+1,j} - x_i^{k,j}\right\|.
\end{aligned}$$

If $\left\|x^{k+1} - x^k\right\| \le \eta$ with $\eta > 0$, then

$$\begin{aligned}
\left\|x_i^{k+1,l} - x_i^{k,l}\right\| &\le \eta + \frac{L\gamma}{T}\sum_{j=0}^{l-1}\left\|x_i^{k+1,j} - x_i^{k,j}\right\| \\
&\overset{l\le T}{\le} \eta + \frac{L\gamma}{T}\sum_{j=0}^{T-1}\left\|x_i^{k+1,j} - x_i^{k,j}\right\|
\end{aligned}$$

Therefore,

$$\sum_{j=0}^{T-1}\left\|x_i^{k+1,j} - x_i^{k,j}\right\| \le \eta T + L\gamma\sum_{j=0}^{T-1}\left\|x_i^{k+1,j} - x_i^{k,j}\right\|.$$

If $\gamma \le \frac{1}{2L}$, then $L\gamma \le 1/2$, and

$$\frac{1}{T}\sum_{j=0}^{T-1}\left\|x_i^{k+1,j} - x_i^{k,j}\right\| \le 2\eta.$$

Next, we prove the third statement. From the definition of the Euclidean norm, and of $x_i^{k,l}$ from the first statement,

$$\begin{aligned}\left\|x^k - x_i^{k,j}\right\| &= \left\|\frac{\gamma}{T}\sum_{j=0}^{l-1}\nabla f_i(x_i^{k,j})\right\| \\ &= \left\|\frac{\gamma}{T}\sum_{j=0}^{l-1}[\nabla f_i(x_i^{k,j}) - \nabla f_i(x^k) + \nabla f_i(x^k)]\right\|.\end{aligned}$$

By the triangle inequality, and by the L-smoothness of $f_i(\cdot)$,

$$\begin{aligned}\left\|x^k - x_i^{k,j}\right\| &\leq \frac{\gamma}{T}\sum_{j=0}^{l-1}\left\|\nabla f_i(x_i^{k,j}) - \nabla f_i(x^k)\right\| + \frac{\gamma}{T}\sum_{j=0}^{l-1}\left\|\nabla f_i(x^k)\right\| \\ &\leq \frac{L\gamma}{T}\sum_{j=0}^{l-1}\left\|x_i^{k,j} - x^k\right\| + \frac{\gamma}{T}\sum_{j=0}^{l-1}\left\|\nabla f_i(x^k)\right\|.\end{aligned}$$

By the fact that $l \leq T$ and that $\|x\| \geq 0$ for $x \in \mathbb{R}^d$,

$$\left\|x^k - x_i^{k,j}\right\| \leq \frac{L\gamma}{T}\sum_{j=0}^{T-1}\left\|x_i^{k,j} - x^k\right\| + \gamma\left\|\nabla f_i(x^k)\right\|.$$

Therefore,

$$\sum_{j=0}^{T-1}\left\|x^k - x_i^{k,j}\right\| \leq L\gamma\sum_{j=0}^{T-1}\left\|x_i^{k,j} - x^k\right\| + \gamma T\left\|\nabla f_i(x^k)\right\|.$$

If $\gamma \leq \frac{1}{2L}$, then $L\gamma \leq 1/2$, and

$$\sum_{j=0}^{T-1}\left\|x^k - x_i^{k,j}\right\| \leq 2\gamma T\left\|\nabla f_i(x^k)\right\|.$$

Next, we prove the fourth statement. From the definition of $\mathcal{T}_i(x^k)$,

$$\left\|\mathcal{T}_i(x^{k+1}) - \mathcal{T}_i(x^k)\right\| = \left\|x^{k+1} - x^k - \frac{\gamma}{T}\sum_{j=0}^{l-1}[\nabla f_i(x_i^{k,l+1}) - \nabla f_i(x_i^{k,l})]\right\|.$$

By the triangle inequality, and by the L-smoothness of $f_i(\cdot)$,

$$\begin{aligned}\left\|\mathcal{T}_i(x^{k+1}) - \mathcal{T}_i(x^k)\right\| &\leq \left\|x^{k+1} - x^k\right\| + \frac{\gamma}{T}\sum_{j=0}^{l-1}\left\|\nabla f_i(x_i^{k,l+1}) - \nabla f_i(x_i^{k,l})\right\| \\ &\leq \left\|x^{k+1} - x^k\right\| + \frac{L\gamma}{T}\sum_{j=0}^{l-1}\left\|x_i^{k,l+1} - x_i^{k,l}\right\|.\end{aligned}$$

By the fact that $\left\|x^{k+1} - x^k\right\| \leq \eta$, that $l \leq T$, and that $\sum_{j=0}^{T-1}\left\|x_i^{k+1,j} - x_i^{k,j}\right\| \leq 2\eta T$,

$$\left\|\mathcal{T}_i(x^{k+1}) - \mathcal{T}_i(x^k)\right\| \leq \eta + L\gamma \cdot 2\eta \overset{L\gamma \leq 1/2}{\leq} 2\eta.$$

Finally, we prove the fifth statement. From the definition of $\mathcal{T}_i(x^k)$,

$$\left\|(x^k - \gamma\nabla f_i(x^k)) - \mathcal{T}_i(x^k)\right\| = \left\|\left(x^k - \frac{\gamma}{T}\sum_{l=0}^{T-1}\nabla f_i(x^k)\right) - \left(x^k - \frac{\gamma}{T}\sum_{l=0}^{T-1}\nabla f_i(x_i^{k,l})\right)\right\|.$$

By the triangle inequality, the L-smoothness of $f_i(\cdot)$, and the fact that

$$\sum_{j=0}^{T} \left\| x^k - x_i^{k,j} \right\| \leq 2\gamma T \left\| \nabla f_i(x^k) \right\|,$$

we obtain

$$\begin{aligned} \left\| (x^k - \gamma \nabla f_i(x^k)) - \mathcal{T}_i(x^k) \right\| &\leq \frac{\gamma}{T} \sum_{l=0}^{T-1} \left\| \nabla f_i(x^k) - \nabla f_i(x_i^{k,l}) \right\| \\ &\leq \frac{L\gamma}{T} \sum_{l=0}^{T-1} \left\| x^k - x_i^{k,l} \right\| \\ &\leq 2L\gamma^2 \left\| \nabla f_i(x^k) \right\|. \end{aligned}$$

□

F.2. Proof of Theorem 1

Now we are ready to prove the convergence rate of Fed-α-NormEC using multiple local GD steps.

Theorem (Fed-α-NormEC with local GD steps)**.** Consider Fed-α-NormEC for solving Problem (1) where Assumption 1 holds. Let $\mathcal{T}_i(x^k) = x^k - \gamma \frac{1}{T} \sum_{j=0}^{T-1} \nabla f_i(x_i^{k,j})$, where the sequence $\{x_i^{k,j}\}$ is generated by $x_i^{k,j+1} = x_i^{k,j} - \frac{\gamma}{T} \nabla f_i(x_i^{k,j})$, for $j = 0, 1, \ldots, T-1$, given that $x_i^{k,0} = x^k$. Furthermore, let $\beta, \alpha > 0$ be chosen such that $\frac{\beta}{\alpha + R} < 1$ with $R = \max_{i \in [1,M]} \left\| v_i^0 - \frac{x^0 - \mathcal{T}_i(x^0)}{\gamma} \right\|$. If $\eta\gamma \leq \frac{1}{K+1} \frac{\Delta^{\inf}}{4L\sqrt{2L}}$, $0 < \eta \leq \frac{\gamma}{3} \frac{\beta R}{\alpha + R}$, and $0 < \gamma \leq \frac{1}{2L}$, then

$$\begin{aligned} \min_{k \in [0,K]} \mathrm{E}\left[\left\| \nabla f(x^k) \right\| \right] \leq & \frac{3}{K+1} \frac{f(x^0) - f^{\inf}}{\eta} + 2R + 2\sqrt{\frac{\beta^2 B}{M}(K+1)} \\ & + \gamma \cdot \mathbb{I}_{T \neq 1} \left[8L\sqrt{2L}\sqrt{\Delta^{\inf}} \right] + \eta \cdot \frac{L}{2}, \end{aligned}$$

where $B = 2p(1 - 1/p)^2 + 2(1-p) + 2\sigma_{\mathrm{DP}}^2/p$, and $\Delta^{\inf} = f^{\inf} - \frac{1}{M}\sum_{i=1}^{M} f_i^{\inf} > 0$.

Proof. We prove the result in the following steps.

Step 1) Bound $\left\| v_i^k - \frac{x^k - \mathcal{T}_i(x^k)}{\gamma} \right\|$ by induction, and bound $\left\| v_i^{k+1} - \frac{x^k - \mathcal{T}_i(x^k)}{\gamma} \right\|$. We prove that $\left\| v_i^k - \frac{x^k - \mathcal{T}_i(x^k)}{\gamma} \right\| \leq \max_{i \in [1,M]} \left\| v_i^0 - \frac{x^0 - \mathcal{T}_i(x^0)}{\gamma} \right\|$ by induction. It is trivial to show the condition when $k = 0$. Next, suppose that $\left\| v_i^k - \frac{x^k - \mathcal{T}_i(x^k)}{\gamma} \right\| \leq \max_{i \in [1,M]} \left\| v_i^0 - \frac{x^0 - \mathcal{T}_i(x^0)}{\gamma} \right\|$ holds. From Lemma 7, $\mathcal{T}_i(x^k) = x^k - \frac{\gamma}{T} \sum_{j=0}^{T-1} \nabla f_i(x_i^{k,j})$ satisfies

$$\left\| \mathcal{T}_i(x^{k+1}) - \mathcal{T}_i(x^k) \right\| \leq 2\eta.$$

Therefore, from Lemma 1 with $\rho = 2$, $C = R = \max_{i \in [1,M]} \left\| v_i^0 - \frac{x^0 - \mathcal{T}_i(x^0)}{\gamma} \right\|$, we can prove that by choosing $\frac{\beta}{\alpha + R} < 1$ and $\eta \leq \frac{\gamma \beta R}{(1+\rho)(\alpha+R)}$, $\left\| v_i^{k+1} - \frac{x^{k+1} - \mathcal{T}_i(x^{k+1})}{\gamma} \right\| \leq R$. We complete the induction proof.

Next, from Lemma 1, $\left\| v_i^{k+1} - \frac{x^k - \mathcal{T}_i(x^k)}{\gamma} \right\| \leq \max_{i \in [1,M]} \left\| v_i^0 - \frac{x^0 - \mathcal{T}_i(x^0)}{\gamma} \right\|$.

Step 2) Bound $f(x^k) - f^{\inf}$. From Lemma 4 with $G^k = \hat{v}^{k+1}$,

$$\begin{aligned} f(x^{k+1}) - f^{\inf} &\leq f(x^k) - f^{\inf} - \eta \left\|\nabla f(x^k)\right\| + 2\eta \left\|\nabla f(x^k) - \hat{v}^{k+1}\right\| + \frac{L\eta^2}{2} \\ &\overset{\text{triangle inequality}}{\leq} f(x^k) - f^{\inf} - \eta \left\|\nabla f(x^k)\right\| + 2\eta \left\|\nabla f(x^k) - v^{k+1}\right\| \\ &\quad +2\eta \left\|\hat{v}^{k+1} - v^{k+1}\right\| + \frac{L\eta^2}{2}, \end{aligned}$$

where $v^{k+1} = \frac{1}{M}\sum_{i=1}^M v_i^{k+1}$. Next, since

$$\begin{aligned} \left\|\nabla f(x^k) - v^{k+1}\right\| &= \left\|\nabla f(x^k) - \frac{1}{M}\sum_{i=1}^M v_i^{k+1}\right\| \\ &\overset{\text{triangle inequality}}{\leq} \frac{1}{M}\sum_{i=1}^M \left\|v_i^{k+1} - \nabla f_i(x^k)\right\| \\ &\overset{\text{triangle inequality}}{\leq} \frac{1}{M}\sum_{i=1}^M \left\|v_i^{k+1} - \frac{x^k - \mathcal{T}_i(x^k)}{\gamma}\right\| \\ &\quad +\frac{1}{M}\sum_{i=1}^M \left\|\frac{x^k - \mathcal{T}_i(x^k)}{\gamma} - \nabla f_i(x^k)\right\|, \end{aligned}$$

where $\mathcal{T}_i(x^k) = x^k - \frac{\gamma}{T}\sum_{j=0}^{T-1} \nabla f_i(x_i^{k,j})$, we get

$$\left\|\nabla f(x^k) - v^{k+1}\right\| \leq \frac{1}{M}\sum_{i=1}^M \left\|v_i^{k+1} - \frac{x^k - \mathcal{T}_i(x^k)}{\gamma}\right\| + \frac{1}{\gamma}\frac{1}{M}\sum_{i=1}^M \left\|x^k - \mathcal{T}_i(x^k) - \gamma\nabla f_i(x^k)\right\|.$$

Plugging the upperbound for $\left\|\nabla f(x^k) - v^{k+1}\right\|$ into the main inequality in $f(x^k) - f^{\inf}$, we obtain

$$\begin{aligned} f(x^{k+1}) - f^{\inf} &\leq f(x^k) - f^{\inf} - \eta \left\|\nabla f(x^k)\right\| + 2\eta\frac{1}{M}\sum_{i=1}^M \left\|v_i^{k+1} - \frac{x^k - \mathcal{T}_i(x^k)}{\gamma}\right\| \\ &\quad +\frac{2\eta}{\gamma}\frac{1}{M}\sum_{i=1}^M \left\|(x^k - \gamma\nabla f_i(x^k)) - \mathcal{T}_i(x^k)\right\| + 2\eta\left\|\hat{v}^{k+1} - v^{k+1}\right\| + \frac{L\eta^2}{2}. \end{aligned}$$

By the fact that $\left\|v_i^{k+1} - \frac{x^k - \mathcal{T}_i(x^k)}{\gamma}\right\| \leq R$ from Step 1),

$$\begin{aligned} f(x^{k+1}) - f^{\inf} &\leq f(x^k) - f^{\inf} - \eta \left\|\nabla f(x^k)\right\| + 2\eta R \\ &\quad +\frac{2\eta}{\gamma}\frac{1}{M}\sum_{i=1}^M \left\|(x^k - \gamma\nabla f_i(x^k)) - \mathcal{T}_i(x^k)\right\| + 2\eta\left\|\hat{v}^{k+1} - v^{k+1}\right\| + \frac{L\eta^2}{2}. \end{aligned}$$

To complete the proof, we consider two possible cases for $\mathcal{T}_i(x^k)$: 1) when $T = 1$ and 2) when $T \neq 1$.

Case 1) $\mathcal{T}_i(x^k)$ with $T = 1$. When $\mathcal{T}_i(x^k)$ with $T = 1$, $\left\|(x^k - \gamma\nabla f_i(x^k)) - \mathcal{T}_i(x^k)\right\| = 0$, and

$$f(x^{k+1}) - f^{\inf} \leq f(x^k) - f^{\inf} - \eta \left\|\nabla f(x^k)\right\| + 2\eta R + 2\eta\left\|\hat{v}^{k+1} - v^{k+1}\right\| + \frac{L\eta^2}{2}.$$

Case 2) $\mathcal{T}_i(x^k)$ with $T > 1$. When $\mathcal{T}_i(x^k)$ with $T > 1$, from Lemma 7,

$$\begin{aligned} f(x^{k+1}) - f^{\inf} &\leq f(x^k) - f^{\inf} - \eta \left\|\nabla f(x^k)\right\| + 2\eta R \\ &\quad +4L\gamma\eta\frac{1}{M}\sum_{i=1}^M \left\|\nabla f_i(x^k)\right\| + 2\eta\left\|\hat{v}^{k+1} - v^{k+1}\right\| + \frac{L\eta^2}{2}. \end{aligned}$$

Therefore, from two cases, we obtain the descent inequality,

$$\begin{aligned} f(x^{k+1}) - f^{\inf} \quad &\leq \quad f(x^k) - f^{\inf} - \eta \left\|\nabla f(x^k)\right\| + 2\eta R \\ &\quad +4L\gamma\eta \frac{1}{M}\sum_{i=1}^{M} \left\|\nabla f_i(x^k)\right\| + 2\eta \left\|\hat{v}^{k+1} - v^{k+1}\right\| + \frac{L\eta^2}{2}. \end{aligned}$$

Next, from Lemma 6,

$$\begin{aligned} f(x^{k+1}) - f^{\inf} \quad &\leq \quad \left(1 + \frac{4L\sqrt{2L}}{\sqrt{\Delta^{\inf}}}\gamma\eta\right)(f(x^k) - f^{\inf}) - \eta \left\|\nabla f(x^k)\right\| + 2\eta R \\ &\quad +4L\sqrt{2L}\gamma\eta\sqrt{\Delta^{\inf}} + 2\eta \left\|\hat{v}^{k+1} - v^{k+1}\right\| + \frac{L\eta^2}{2}. \end{aligned}$$

Since

$$\begin{aligned} \mathrm{E}\left[\left\|\hat{v}^{k+1} - v^{k+1}\right\|\right] \quad &\leq \quad \frac{1}{\gamma}\mathrm{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M} v_i^{k+1} - \hat{v}^{k+1}\right\|\right] \\ &\overset{\text{Lemma 3}}{\leq} \quad \frac{1}{\gamma}\sqrt{\frac{\beta^2 B}{M}(K+1)}, \end{aligned}$$

by taking the expectation,

$$\begin{aligned} \mathrm{E}\left[f(x^{k+1}) - f^{\inf}\right] \quad &\leq \quad \left(1 + \frac{4L\sqrt{2L}}{\sqrt{\Delta^{\inf}}}\gamma\eta\right)\mathrm{E}\left[f(x^k) - f^{\inf}\right] - \eta\mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] + 2\eta R \\ &\quad +8L\sqrt{2L}\gamma\eta\sqrt{\Delta^{\inf}} + 2\eta\sqrt{\frac{\beta^2 B}{M}(K+1)} + \frac{L\eta^2}{2}. \end{aligned}$$

By applying Lemma 5 with $\eta\gamma \leq \frac{1}{K+1}\frac{\Delta^{\inf}}{4L\sqrt{2L}}$ and using the fact $(1 + \eta\gamma\frac{4L\sqrt{2L}}{\Delta^{\inf}})^{K+1} \leq \exp(\eta\gamma\frac{4L\sqrt{2L}}{\Delta^{\inf}}(K+1)) \leq \exp(1) \leq 3$ we finalize the proof.

□

F.3. Corollaries for Fed-α-NormEC with multiple local GD steps from Theorem 1

Corollary 3 (Convergence bound for Fed-α-NormEC with multiple local GD steps)**.** *Consider* Fed-α-NormEC *for solving Problem* (1) *under the same setting as Theorem 1. Let $T > 1$ (multiple local GD steps). If $\gamma = \frac{1}{2L(K+1)^{1/8}}$, $v_i^0 \in \mathbb{R}^d$ is chosen such that $\max_{i\in[1,M]}\left\|\frac{x^0 - \mathcal{T}_i(x^0)}{\gamma} - v_i^0\right\| = \frac{D_1}{(K+1)^{1/8}}$ with $D_1 > 0$, and $\beta = \frac{D_2}{(K+1)^{5/8}}$ with $D_2 > 0$, and $\eta = \frac{\hat{\eta}}{(K+1)^{7/8}}$ with $\hat{\eta} = \min\left(\frac{\Delta^{\inf}}{2\sqrt{2L}}, \frac{D_1 D_2}{4L(\alpha + D_1)}\right)$, then*

$$\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \leq \frac{A_1}{(K+1)^{1/8}} + \frac{A_2}{(K+1)^{7/8}},$$

where $A_1 = 3\frac{f(x^0) - f^{\inf}}{\hat{\eta}} + 2D_1 + \frac{2\sqrt{B}D_2}{\sqrt{M}} + 4\sqrt{2L}\sqrt{\Delta^{\inf}}$ and $A_2 = \hat{\eta}L/2$.

Proof. Let $T > 1$. Then, from Theorem 1,

$$\begin{aligned} \min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \leq &\frac{3}{K+1}\frac{f(x^0) - f^{\inf}}{\eta} + 2R + 2\sqrt{\frac{\beta^2 B}{M}(K+1)} + \eta \cdot \frac{L}{2} \\ &+ \gamma \cdot \left[8L\sqrt{2L}\sqrt{\Delta^{\inf}}\right], \end{aligned}$$

where $B = 2p(1 - 1/p)^2 + 2(1 - p) + 2\sigma_{\mathrm{DP}}^2/p$.

Next, suppose that

- $\gamma = \frac{1}{2L(K+1)^{1/8}}$ to guarantee that $\gamma \le 1/(2L)$
- $v_i^0 \in \mathbb{R}^d$ such that $\max_{i\in[1,M]} \left\| \frac{x^0 - \mathcal{T}_i(x^0)}{\gamma} - v_i^0 \right\| = R = \frac{D_1}{(K+1)^{1/8}}$ with $D_1 > 0$
- $\beta = \frac{D_2}{(K+1)^{5/8}}$ with $D_2 > 0$.

Then, we choose $\eta = \frac{\hat{\eta}}{(K+1)^{7/8}}$ with $\hat{\eta} = \min\left(\frac{\Delta^{\inf}}{2\sqrt{2L}}, \frac{D_1 D_2}{4L(\alpha+D_1)}\right)$ to ensure that $\eta\gamma \le \frac{1}{K+1}\frac{\Delta^{\inf}}{4L\sqrt{2L}}$ and $\eta \le \frac{\gamma}{2}\frac{\beta R}{\alpha+R}$. Therefore,

$$\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \le \frac{A_1}{(K+1)^{1/8}} + \frac{A_2}{(K+1)^{7/8}},$$

where $A_1 = 3\frac{f(x^0)-f^{\inf}}{\hat{\eta}} + 2D_1 + \frac{2\sqrt{B}D_2}{\sqrt{M}} + 4\sqrt{2L}\sqrt{\Delta^{\inf}}$ and $A_2 = \hat{\eta}L/2$.

□

Corollary 4 (Utility bound for Fed-α-NormEC with multiple local GD steps)**.** *Consider* Fed-α-NormEC *for solving Problem* (1) *under the same setting as Theorem 1. Let $T > 1$ (multiple local GD steps), let $\sigma_{\mathrm{DP}} = c\frac{p\sqrt{(K+1)\log(1/\delta)}}{\epsilon}$ with $c > 0$ (privacy with subsampling amplification), and let $p = \frac{\hat{B}}{M}$ for $\hat{B} \in [1, M]$ (client subsampling). If $\beta = \frac{\hat{\beta}}{K+1}$ with $\hat{\beta} = \sqrt{\frac{3(f(x^0)-f^{\inf})}{\gamma}}\sqrt[4]{\frac{M}{B_2}}$, $\gamma < \frac{\Delta^{\inf}(\alpha+R)}{\sqrt{2L}\hat{\beta}R}$, $\alpha = R = \mathcal{O}\left(\sqrt[4]{d}\frac{\sqrt{f(x^0)-f^{\inf}}}{\sqrt{\gamma}}\sqrt[4]{\frac{B_2}{M}}\right)$ with $B_2 = 2c^2\frac{\hat{B}}{M}\frac{\log(1/\delta)}{\epsilon^2}$, and $\eta = \frac{1}{K+1}\frac{\gamma}{2}\frac{\hat{\beta}R}{\alpha+R}$, then*

$$\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \le \mathcal{O}\left(\Delta\sqrt[4]{\frac{d\hat{B}}{M^2}\frac{\log(1/\delta)}{\epsilon^2}} + \sqrt{L}\sqrt{\Delta^{\inf}}\right),$$

where $\Delta = \max(\alpha, 2)\sqrt{L}\sqrt{f(x^0) - f^{\inf}}$.

Proof. Let $T > 1$. Then, from Theorem 1,

$$\begin{aligned}\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \le& \frac{3}{K+1}\frac{f(x^0)-f^{\inf}}{\eta} + 2R + 2\sqrt{\frac{\beta^2 B}{M}(K+1)} + \eta\cdot\frac{L}{2} \\ &+ \gamma\cdot\left[8L\sqrt{2L}\sqrt{\Delta^{\inf}}\right],\end{aligned}$$

where $B = 2p(1-1/p)^2 + 2(1-p) + 2\sigma_{\mathrm{DP}}^2/p$.

Also, let $\sigma_{\mathrm{DP}} = c\frac{p\sqrt{(K+1)\log(1/\delta)}}{\epsilon}$ with $c > 0$, and let $p = \frac{\hat{B}}{M}$ for $\hat{B} \in [1, M]$ is the number of clients being sampled on each round. Then, $B = \frac{2\hat{B}}{M}\left(1 - \frac{M}{\hat{B}}\right)^2 + 2\left(1 - \frac{\hat{B}}{M}\right) + 2\frac{c\sqrt{K+1}\log(1/\delta)}{\epsilon}$, and

$$\begin{aligned}\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \le& \frac{3}{K+1}\frac{f(x^0)-f^{\inf}}{\eta} + 2R + 2\beta\sqrt{\frac{B_1}{M}(K+1)} + 2\beta\sqrt{\frac{B_2}{M}(K+1)} + \eta\cdot\frac{L}{2} \\ &+ \gamma\cdot\left[8L\sqrt{2L}\sqrt{\Delta^{\inf}}\right],\end{aligned}$$

where $B_1 = \frac{2\hat{B}}{M}\left[\left(1 - \frac{M}{\hat{B}}\right)^2 + \frac{M}{\hat{B}} - 1\right]$ and $B_2 = 2c^2\frac{\hat{B}}{M}\frac{\log(1/\delta)}{\epsilon^2}$.

If $\beta = \frac{\hat{\beta}}{K+1}$ with $\hat{\beta} > 0$, then

$$\begin{aligned}\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \le& \frac{3}{K+1}\frac{f(x^0)-f^{\inf}}{\eta} + 2R + 2\hat{\beta}\sqrt{\frac{B_1}{M(K+1)}} + 2\hat{\beta}\sqrt{\frac{B_2}{M}} + \eta\cdot\frac{L}{2} \\ &+ \gamma\cdot\left[8L\sqrt{2L}\sqrt{\Delta^{\inf}}\right].\end{aligned}$$

Since $\beta = \frac{\hat{\beta}}{K+1}$, we obtain

$$\eta \leq \frac{1}{K+1} \min\left(\frac{\Delta^{\inf}}{2\sqrt{2L}}, \frac{\gamma}{2}\frac{\hat{\beta}R}{\alpha+R}\right).$$

If $\Delta^{\inf} > \frac{\gamma\sqrt{2L}\hat{\beta}R}{\alpha+R}$, then

$$\eta \leq \frac{1}{K+1}\frac{\gamma}{2}\frac{\hat{\beta}R}{\alpha+R}.$$

If $\eta = \frac{1}{K+1}\frac{\gamma}{2}\frac{\hat{\beta}R}{\alpha+R}$, then

$$\begin{aligned}\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \leq & \frac{6\alpha(f(x^0)-f^{\inf})}{\gamma\hat{\beta}R} + \frac{6(f(x^0)-f^{\inf})}{\gamma\hat{\beta}} + 2R + 2\hat{\beta}\sqrt{\frac{B_2}{M}} \\ & + 2\hat{\beta}\sqrt{\frac{B_1}{M(K+1)}} + \frac{1}{K+1}\cdot\frac{\gamma L\hat{\beta}R}{4(\alpha+R)} \\ & + \gamma\cdot\left[8L\sqrt{2L}\sqrt{\Delta^{\inf}}\right].\end{aligned}$$

If $\hat{\beta} = \sqrt{\frac{3(f(x^0)-f^{\inf})}{\gamma}}\sqrt[4]{\frac{M}{B_2}}$, then

$$\begin{aligned}\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \leq & \frac{2\sqrt{3}\alpha\sqrt{f(x^0)-f^{\inf}}}{\sqrt{\gamma}R}\sqrt[4]{\frac{B_2}{M}} + \frac{4\sqrt{3}\sqrt{f(x^0)-f^{\inf}}}{\sqrt{\gamma}}\sqrt[4]{\frac{B_2}{M}} + 2R \\ & + 2\hat{\beta}\sqrt{\frac{B_1}{M(K+1)}} + \frac{1}{K+1}\cdot\frac{\gamma L\hat{\beta}R}{4(\alpha+R)} + \gamma\cdot\left[8L\sqrt{2L}\sqrt{\Delta^{\inf}}\right].\end{aligned}$$

If $\alpha = R = \mathcal{O}\left(\sqrt[4]{d}\frac{\sqrt{f(x^0)-f^{\inf}}}{\sqrt{\gamma}}\sqrt[4]{\frac{B_2}{M}}\right)$, then

$$\begin{aligned}\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \leq & \mathcal{O}\left(\Delta\frac{\sqrt{f(x^0)-f^{\inf}}}{\sqrt{\gamma}}\sqrt[4]{d\frac{B_2}{M}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{K+1}}\right) + \mathcal{O}\left(\frac{1}{K+1}\right) \\ & + \gamma\cdot\left[8L\sqrt{2L}\sqrt{\Delta^{\inf}}\right] \\ \leq & \mathcal{O}\left(\Delta\frac{\sqrt{f(x^0)-f^{\inf}}}{\sqrt{\gamma}}\sqrt[4]{d\frac{B_2}{M}} + \gamma\cdot\left[8L\sqrt{2L}\sqrt{\Delta^{\inf}}\right]\right) \\ & + \mathcal{O}\left(\frac{1}{\sqrt{K+1}}\right) + \mathcal{O}\left(\frac{1}{K+1}\right),\end{aligned}$$

where $\Delta = 2\sqrt{3}\max(\alpha, 2)$. Finally, if $\gamma = 1/(2L)$, then we complete the proof.

□

F.4. Proof of Corollary 1

Corollary (Convergence bound for Fed-α-NormEC with one local GD step)**.** *Consider* Fed-α-NormEC *for solving Problem* (1) *under the same setting as Theorem 1. Let $T = 1$ and $N = 0$ (one local GD step). If $\gamma = \frac{1}{2L}$, $v_i^0 \in \mathbb{R}^d$ is chosen such that $\max_{i\in[1,M]} \left\| \frac{x^0 - \mathcal{T}_i(x^0)}{\gamma} - v_i^0 \right\| = \frac{D_1}{(K+1)^{1/6}}$ with $D_1 > 0$, and $\beta = \frac{D_2}{(K+1)^{2/3}}$ with $D_2 > 0$, and $\eta = \frac{\hat{\eta}}{(K+1)^{5/6}}$ with $\hat{\eta} = \frac{D_1 D_2}{4L(\alpha + D_1)}$, then*

$$\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \leq \frac{A_1}{(K+1)^{1/6}} + \frac{A_2}{(K+1)^{5/6}},$$

where $A_1 = 3\frac{f(x^0) - f^{\inf}}{\hat{\eta}} + 2D_1 + \frac{2\sqrt{B}D_2}{\sqrt{M}}$ and $A_2 = \hat{\eta}L/2$.

Proof. Let $T = 1$. Then, from Theorem 1,

$$\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \leq \frac{3}{K+1}\frac{f(x^0) - f^{\inf}}{\eta} + 2R + 2\sqrt{\frac{\beta^2 B}{M}(K+1)} + \eta \cdot \frac{L}{2},$$

where $B = 2p(1 - 1/p)^2 + 2(1-p) + 2\sigma_{\mathrm{DP}}^2/p$.

Next, suppose that

- $\gamma = \frac{1}{2L}$
- $v_i^0 \in \mathbb{R}^d$ such that $\max_{i\in[1,M]} \left\| \frac{x^0 - \mathcal{T}_i(x^0)}{\gamma} - v_i^0 \right\| = R = \frac{D_1}{(K+1)^{1/6}}$ with $D_1 > 0$
- $\beta = \frac{D_2}{(K+1)^{2/3}}$ with $D_2 > 0$.

Then, we choose $\eta = \frac{\hat{\eta}}{(K+1)^{5/6}}$ with $\hat{\eta} = \frac{D_1 D_2}{4L(\alpha + D_1)}$ to ensure that $\eta \leq \frac{\gamma}{2}\frac{\beta R}{\alpha + R}$. Therefore,

$$\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \leq \frac{A_1}{(K+1)^{1/6}} + \frac{A_2}{(K+1)^{5/6}},$$

where $A_1 = 3\frac{f(x^0) - f^{\inf}}{\hat{\eta}} + 2D_1 + \frac{2\sqrt{B}D_2}{\sqrt{M}}$ and $A_2 = \hat{\eta}L/2$.

□

F.5. Proof of Corollary 2

Corollary (Utility bound for Fed-α-NormEC with one local GD step)**.** *Consider* Fed-α-NormEC *for solving Problem* (1) *under the same setting as Theorem 1. Let $T = 1$ (one local GD step), let $\sigma_{\mathrm{DP}} = c\frac{p\sqrt{(K+1)\log(1/\delta)}}{\epsilon}$ with $c > 0$ (privacy with subsampling amplification), and let $p = \frac{\hat{B}}{M}$ for $\hat{B} \in [1, M]$ (client subsampling). If $\beta = \frac{\hat{\beta}}{K+1}$ with $\hat{\beta} = \sqrt{\frac{3(f(x^0) - f^{\inf})}{\gamma}}\sqrt[4]{\frac{M}{B_2}}$, $\gamma < \frac{\Delta^{\inf}(\alpha + R)}{\sqrt{2L}\hat{\beta}R}$, $\alpha = R = \mathcal{O}\left(\sqrt[4]{d}\frac{\sqrt{f(x^0) - f^{\inf}}}{\sqrt{\gamma}}\sqrt[4]{\frac{B_2}{M}}\right)$ with $B_2 = 2c^2\frac{\hat{B}}{M}\frac{\log(1/\delta)}{\epsilon^2}$, and $\eta = \frac{1}{K+1}\frac{\gamma}{2}\frac{\hat{\beta}R}{\alpha + R}$, then*

$$\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \leq \mathcal{O}\left(\Delta\sqrt[4]{\frac{d\hat{B}}{M^2}\frac{\log(1/\delta)}{\epsilon^2}}\right),$$

where $\Delta = \max(\alpha, 2)\sqrt{L}\sqrt{f(x^0) - f^{\inf}}$.

Proof. Let $T = 1$. Then, from Theorem 1,

$$\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \leq \frac{3}{K+1}\frac{f(x^0) - f^{\inf}}{\eta} + 2R + 2\sqrt{\frac{\beta^2 B}{M}(K+1)} + \eta \cdot \frac{L}{2},$$

where $B = 2p(1-1/p)^2 + 2(1-p) + 2\sigma_{\rm DP}^2/p$.

Also, let $\sigma_{\rm DP} = c\frac{p\sqrt{(K+1)\log(1/\delta)}}{\epsilon}$ with $c > 0$, and let $p = \frac{\hat{B}}{M}$ for $\hat{B} \in [1, M]$ is the number of clients being sampled on each round. Then, $B = \frac{2\hat{B}}{M}\left(1 - \frac{M}{\hat{B}}\right)^2 + 2\left(1 - \frac{\hat{B}}{M}\right) + 2\frac{c\sqrt{K+1}\log(1/\delta)}{\epsilon}$, and

$$\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \leq \frac{3}{K+1}\frac{f(x^0) - f^{\inf}}{\eta} + 2R + 2\beta\sqrt{\frac{B_1}{M}(K+1)} + 2\beta\sqrt{\frac{B_2}{M}(K+1)} + \eta\cdot\frac{L}{2},$$

where $B_1 = \frac{2\hat{B}}{M}\left[\left(1 - \frac{M}{\hat{B}}\right)^2 + \frac{M}{\hat{B}} - 1\right]$ and $B_2 = 2c^2\frac{\hat{B}}{M}\frac{\log(1/\delta)}{\epsilon^2}$.

If $\beta = \frac{\hat{\beta}}{K+1}$ with $\hat{\beta} > 0$, then

$$\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \leq \frac{3}{K+1}\frac{f(x^0) - f^{\inf}}{\eta} + 2R + 2\hat{\beta}\sqrt{\frac{B_1}{M(K+1)}} + 2\hat{\beta}\sqrt{\frac{B_2}{M}} + \eta\cdot\frac{L}{2}.$$

Since $\beta = \frac{\hat{\beta}}{K+1}$, we obtain

$$\eta \leq \frac{1}{K+1}\min\left(\frac{\Delta^{\inf}}{2\sqrt{2L}}, \frac{\gamma}{2}\frac{\hat{\beta}R}{\alpha + R}\right).$$

If $\Delta^{\inf} > \frac{\gamma\sqrt{2L}\hat{\beta}R}{\alpha+R}$, then

$$\eta \leq \frac{1}{K+1}\frac{\gamma}{2}\frac{\hat{\beta}R}{\alpha+R}.$$

If $\eta = \frac{1}{K+1}\frac{\gamma}{2}\frac{\hat{\beta}R}{\alpha+R}$, then

$$\begin{aligned}\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \leq & \frac{6\alpha(f(x^0) - f^{\inf})}{\gamma\hat{\beta}R} + \frac{6(f(x^0) - f^{\inf})}{\gamma\hat{\beta}} + 2R + 2\hat{\beta}\sqrt{\frac{B_2}{M}} \\ & + 2\hat{\beta}\sqrt{\frac{B_1}{M(K+1)}} + \frac{1}{K+1}\cdot\frac{\gamma L\hat{\beta}R}{4(\alpha+R)}.\end{aligned}$$

If $\hat{\beta} = \sqrt{\frac{3(f(x^0) - f^{\inf})}{\gamma}}\sqrt[4]{\frac{M}{B_2}}$, then

$$\begin{aligned}\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \leq & \frac{2\sqrt{3}\alpha\sqrt{f(x^0) - f^{\inf}}}{\sqrt{\gamma}R}\sqrt[4]{\frac{B_2}{M}} + \frac{4\sqrt{3}\sqrt{f(x^0) - f^{\inf}}}{\sqrt{\gamma}}\sqrt[4]{\frac{B_2}{M}} + 2R \\ & + 2\hat{\beta}\sqrt{\frac{B_1}{M(K+1)}} + \frac{1}{K+1}\cdot\frac{\gamma L\hat{\beta}R}{4(\alpha+R)}.\end{aligned}$$

If $\alpha = R = \mathcal{O}\left(\sqrt[4]{d}\frac{\sqrt{f(x^0) - f^{\inf}}}{\sqrt{\gamma}}\sqrt[4]{\frac{B_2}{M}}\right)$, then

$$\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \leq \mathcal{O}\left(\Delta\frac{\sqrt{f(x^0) - f^{\inf}}}{\sqrt{\gamma}}\sqrt[4]{d\frac{B_2}{M}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{K+1}}\right) + \mathcal{O}\left(\frac{1}{K+1}\right),$$

where $\Delta = 2\sqrt{3}\max(\alpha, 2)$. Finally, if $\gamma = 1/(2L)$, then we complete the proof. □

G. Multiple Local IG steps

In this section, we derive the convergence theorem of Fed-α-NormEC with multiple local steps using the Incremental Gradient (IG) method. The IG method has the following update rule.

$$\mathcal{T}_i^{IG}(x^k) = x^k - \gamma \frac{1}{N} \sum_{j=0}^{N-1} \nabla f_{i,j}(x_i^{k,j}), \tag{2}$$

where $x_i^{k,j}$ is updated according to:

$$x_i^{k,j+1} = x_i^{k,j} - \frac{\gamma}{N} \nabla f_{i,j}(x_i^{k,j}) \quad \text{for} \quad j = 0, 1, \dots, T-1.$$

In the update rule of the IG method, the number of local steps is equal to the size of the local data set. This implies that each client performs local updates $\mathcal{T}_i^{IG}(\cdot)$ using their entire local dataset. Furthermore, the IG method employs a fixed, deterministic permutation for its cyclic updates, unlike the well-known Random Reshuffling method.

G.1. Key Lemmas

First, we introduce key lemmas for analyzing Fed-α-NormEC using multiple local IG steps. Lemma 8 bounds $\frac{1}{M}\sum_{i=1}^M \frac{1}{N}\sum_{j=0}^{N-1} \left\| x_i^{k,j} - x^k \right\|$ while Lemma 9 proves the properties of local IG steps.

Lemma 8. *Consider the local IG method updates in (2). Let f be bounded from below by $f^{\inf} > -\infty$, let each f_i be bounded from below by $f_i^{\inf} > -\infty$, and let each $f_{i,j}$ be bounded from below by $f_{i,j}^{\inf}$ and L-smooth. Then,*

$$\frac{1}{M}\sum_{i=1}^M \frac{1}{N}\sum_{j=0}^{N-1} \left\| x_i^{k,j} - x^k \right\| \le \frac{2\sqrt{2}L\gamma(f(x^k) - f^{\inf})}{\sqrt{L\Delta^{\inf}}} + \sqrt{2}\gamma\sqrt{L\Delta^{\inf}} + 2\gamma\sqrt{L\frac{1}{M}\sum_{i=1}^M \Delta_i^{\inf}},$$

where $\Delta^{\inf} = f^{\inf} - \frac{1}{M}\sum_{i=1}^M f_i^{\inf}$ and $\Delta_i^{\inf} = f^{\inf} - \frac{1}{N}\sum_{j=1}^N f_{i,j}^{\inf}$

Proof. Applying Lemma 6 from [43] for the local IG method updates in (2), we have

$$\frac{1}{M}\sum_{i=1}^M \frac{1}{N}\sum_{j=0}^{N-1} \left\| x_i^{k,j} - x^k \right\|^2 \le 4L\gamma^2 \left(f(x^k) - f^{\inf}\right) + 2\gamma^2 L\Delta^{\inf} + 2\gamma^2 L\frac{1}{M}\sum_{i=1}^M \Delta_i^{\inf}.$$

Next, by Jensen's inequality,

$$\begin{aligned}
\frac{1}{M}\sum_{i=1}^M \frac{1}{N}\sum_{j=0}^{N-1} \left\| x_i^{k,j} - x^k \right\| &\le \sqrt{\frac{1}{M}\sum_{i=1}^M \frac{1}{N}\sum_{j=0}^{N-1} \left\| x_i^{k,j} - x^k \right\|^2} \\
&\le \sqrt{4L\gamma^2 \left(f(x^k) - f^{\inf}\right) + 2\gamma^2 L\Delta^{\inf} + 2\gamma^2 L\frac{1}{M}\sum_{i=1}^M \Delta_i^{\inf}} \\
&\le \sqrt{4L\gamma^2 \left(f(x^k) - f^{\inf}\right) + 2\gamma^2 L\Delta^{\inf}} + \sqrt{2\gamma^2 L\frac{1}{M}\sum_{i=1}^M \Delta_i^{\inf}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{1}{M}\sum_{i=1}^{M}\frac{1}{N}\sum_{j=0}^{N-1}\left\|x_i^{k,j}-x^k\right\| &\le \frac{4L\gamma^2\left(f(x^k)-f^{\inf}\right)+2\gamma^2L\Delta^{\inf}}{\sqrt{4L\gamma^2\left(f(x^k)-f^{\inf}\right)+2\gamma^2L\Delta^{\inf}}}+2\gamma\sqrt{L\frac{1}{M}\sum_{i=1}^{M}\Delta_i^{\inf}}\\ &\le \frac{4L\gamma^2\left(f(x^k)-f^{\inf}\right)+2\gamma^2L\Delta^{\inf}}{\sqrt{2\gamma^2L\Delta^{\inf}}}+2\gamma\sqrt{L\frac{1}{M}\sum_{i=1}^{M}\Delta_i^{\inf}}\\ &\le \frac{2\sqrt{2}L\gamma(f(x^k)-f^{\inf})}{\sqrt{L\Delta^{\inf}}}+\sqrt{2}\gamma\sqrt{L\Delta^{\inf}}+2\gamma\sqrt{L\frac{1}{M}\sum_{i=1}^{M}\Delta_i^{\inf}}.\end{aligned}$$

□

Lemma 9. *Let each f_i be L-smooth, and let $\mathcal{T}_i(x^k)=x^k-\frac{\gamma}{N}\sum_{j=0}^{N-1}\nabla f_{i,j}(x_i^{k,j})$, where the sequence $\{x_i^{k,l}\}$ is generated by*

$$x_i^{k,l+1}=x_i^{k,l}-\frac{\gamma}{N}\nabla f_{i,j}(x_i^{k,l}),\quad \textit{for}\quad l=0,1,\ldots,N-1,$$

given that $x_i^{k,0}=x^k$. If $\gamma\le\frac{1}{2L}$, and $\left\|x^{k+1}-x^k\right\|\le\eta$ with $\eta>0$, then

1. $x_i^{k,l}=x^k-\frac{\gamma}{N}\sum_{j=0}^{l-1}\nabla f_{i,j}(x_i^{k,l})$.
2. $\frac{1}{N}\sum_{j=0}^{N-1}\left\|x_i^{k+1,j}-x_i^{k,j}\right\|\le 2\eta$.
3. $\left\|\mathcal{T}_i(x^{k+1})-\mathcal{T}_i(x^k)\right\|\le 2\eta$.
4. $\frac{1}{M}\sum_{i=1}^{M}\left\|\mathcal{T}_i(x^k)-\left(x^k-\gamma\nabla f_i(x^k)\right)\right\|\le\gamma L\frac{1}{M}\sum_{i=1}^{M}\frac{1}{N}\sum_{j=0}^{N-1}\left\|x_i^{k,j}-x^k\right\|$

Proof. The first statement derives from unrolling the recursion for $x_i^{k,j+1}$.

Next, we prove the second statement. Let us consider

$$\begin{aligned}\left\|x_i^{k+1,j}-x_i^{k,j}\right\| &= \left\|x^{k+1}-\gamma\frac{1}{N}\sum_{l=0}^{j-1}\nabla f_{i,l}(x_i^{k+1,l})-\left(x^k-\gamma\frac{1}{N}\sum_{l=0}^{j-1}\nabla f_{i,l}(x_i^{k,l})\right)\right\|\\ &\le \|x^{k+1}-x^k\|+\gamma L\frac{1}{N}\sum_{l=0}^{j-1}\|x_i^{k+1,l}-x_i^{k,l}\|\\ &\le \|x^{k+1}-x^k\|+\gamma L\frac{1}{N}\sum_{j=0}^{N-1}\|x_i^{k+1,j}-x_i^{k,j}\|.\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{1}{N}\sum_{j=0}^{N-1}\left\|x_i^{k+1,j}-x_i^{k,j}\right\| &\le \frac{1}{N}\sum_{j=0}^{N-1}\left(\|x^{k+1}-x^k\|+\gamma\frac{1}{N}\sum_{j=0}^{N-1}\|x_i^{k+1,j}-x_i^{k,j}\|\right)\\ &\le \|x^{k+1}-x^k\|+\gamma L\frac{1}{N}\sum_{j=0}^{N-1}\|x_i^{k+1,j}-x_i^{k,j}\|.\end{aligned}$$

If $\gamma \leq \frac{1}{2L}$, then

$$\begin{aligned}
\frac{1}{N}\sum_{j=0}^{N-1}\left\|x_{i,j}^{k+1}-x_{i,j}^{k}\right\| &\leq \frac{1}{1-\gamma L}\|x^{k+1}-x^{k}\| \\
&\leq 2\|x^{k+1}-x^{k}\| \\
&= 2\eta.
\end{aligned}$$

Next, we prove the third statement. Let us consider

$$\begin{aligned}
\left\|\mathcal{T}_i(x^{k+1})-\mathcal{T}_i(x^{k})\right\| &= \left\|x^{k+1}-\gamma\frac{1}{N}\sum_{j=0}^{N-1}\nabla f_{i,j}(x_i^{k+1,j})-\left(x^{k}-\gamma\frac{1}{N}\sum_{j=0}^{N-1}\nabla f_{i,j}(x_i^{k,j})\right)\right\| \\
&\leq \|x^{k+1}-x^{k}\|+\gamma\frac{1}{N}\sum_{j=0}^{N-1}\left\|\nabla f_{i,j}(x_i^{k+1,j})-\nabla f_{i,j}(x_i^{k,j})\right\| \\
&\leq \|x^{k+1}-x^{k}\|+\gamma L\frac{1}{N}\sum_{j=0}^{N-1}\|x_i^{k+1,j}-x_i^{k,j}\|.
\end{aligned}$$

By the fact that $\left\|x^{k+1}-x^{k}\right\| \leq \eta$ and that $\gamma \leq \frac{1}{2L}$, and by the second statement,

$$\left\|\mathcal{T}_i(x^{k+1})-\mathcal{T}_i(x^{k})\right\| \leq 2\eta.$$

Finally, we prove the fourth statement. Let us consider

$$\begin{aligned}
\left\|\mathcal{T}_i(x^{k})-\left(x^{k}-\gamma\nabla f_i(x^{k})\right)\right\| &= \left\|x^{t}-\gamma\frac{1}{N}\sum_{j=0}^{N-1}\nabla f_{i,j}(x_i^{k,j})-(x^{t}-\gamma\nabla f_i(x^{k})\right\| \\
&= \left\|\gamma\left(\frac{1}{N}\sum_{i=0}^{N-1}\nabla f_{i,j}(x_i^{k,j})-\nabla f_i(x^{k})\right)\right\| \\
&= \left\|\gamma\left(\frac{1}{N}\sum_{i=0}^{N-1}\nabla f_{i,j}(x_i^{k,j})-\frac{1}{N}\sum_{i=0}^{N-1}\nabla f_{i,j}(x^{k})\right)\right\| \\
&= \gamma\frac{1}{N}\sum_{i=0}^{N-1}\left\|\nabla f_{i,j}(x_i^{k,j})-\nabla f_{i,j}(x^{k})\right\| \\
&\leq \gamma L\frac{1}{N}\sum_{i=0}^{N-1}\left\|x_i^{k,j}-x^{k}\right\|.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{1}{M}\sum_{i=1}^{M}\left\|\mathcal{T}_i(x^{k})-\left(x^{k}-\gamma\nabla f_i(x^{k})\right)\right\| &\leq \frac{1}{M}\sum_{i=1}^{M}\gamma L\frac{1}{N}\sum_{i=0}^{N-1}\left\|x_i^{k,j}-x^{k}\right\| \\
&\leq \gamma L\frac{1}{M}\sum_{i=1}^{M}\frac{1}{N}\sum_{j=0}^{N-1}\left\|x_i^{k,j}-x^{k}\right\|.
\end{aligned}$$

□

G.2. Convergence Theorem for Fed-α-NormEC with local IG steps

Theorem 2 (Fed-α-NormEC with local IG steps)**.** *Consider* Fed-α-NormEC *for solving Problem* (1) *where Assumption 1 holds. Let* $\mathcal{T}_i(x^k) = x^k - \gamma\frac{1}{N}\sum_{j=0}^{N-1}\nabla f_{i,j}(x_i^{k,j})$*, where the sequence* $\{x_i^{k,j}\}$ *is generated by* $x_i^{k,j+1} = x_i^{k,j} - \frac{\gamma}{N}\nabla f_{i,j}(x_i^{k,j})$ *for* $j = 0,1,\dots,T-1$*, given that* $x_i^{k,0} = x^k$*. Furthermore, let* $\beta,\alpha > 0$ *be chosen such that* $\frac{\beta}{\alpha+R} < 1$ *with* $R = \max_{i\in[1,M]}\left\|v_i^0 - \frac{x^0-\mathcal{T}_i(x^0)}{\gamma}\right\|$*. If* $\eta\gamma \leq \frac{1}{K+1}\frac{\Delta^{\inf}}{4L\sqrt{2L}}$*,* $0<\eta\leq\frac{\gamma}{3}\frac{\beta R}{\alpha+R}$*, and* $0<\gamma\leq\frac{1}{2L}$*, then*

$$\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \leq \frac{3}{K+1}\frac{f(x^0)-f^{\inf}}{\eta} + 2R + 2\sqrt{\frac{\beta^2 B}{M}(K+1)} + \gamma\cdot 8L\sqrt{2L}\sqrt{\Delta^{\inf}} + \gamma\cdot 4L\sqrt{2L}\sqrt{\frac{1}{M}\sum_{i=1}^M \Delta_i^{\inf}} + \eta\cdot\frac{L}{2},$$

where $B = 2p(1-1/p)^2 + 2(1-p) + 2\sigma_{\mathrm{DP}}^2/p$*, and* $\Delta^{\inf} = f^{\inf} - \frac{1}{M}\sum_{i=1}^M f_i^{\inf} > 0$*, and* $\Delta_i^{\inf} = f^{\inf} - \frac{1}{N}\sum_{j=1}^N f_{i,j}^{\inf} > 0$

Proof. We prove the result in the following steps.

Step 1) Bound $\left\|v_i^k - \frac{x^k-\mathcal{T}_i(x^k)}{\gamma}\right\|$ **by induction, and bound** $\left\|v_i^{k+1} - \frac{x^k-\mathcal{T}_i(x^k)}{\gamma}\right\|$**.** We prove $\left\|v_i^k - \frac{x^k-\mathcal{T}_i(x^k)}{\gamma}\right\| \leq \max_{i\in[1,M]}\left\|v_i^0 - \frac{x^0-\mathcal{T}_i(x^0)}{\gamma}\right\|$ by induction. We can easily show the condition when $k=0$. Next, let $\left\|v_i^k - \frac{x^k-\mathcal{T}_i(x^k)}{\gamma}\right\| \leq \max_{i\in[1,M]}\left\|v_i^0 - \frac{x^0-\mathcal{T}_i(x^0)}{\gamma}\right\|$. Then, from Lemma 9, $\mathcal{T}_i(x^k)$ satisfies

$$\left\|\mathcal{T}_i(x^{k+1}) - \mathcal{T}_i(x^k)\right\| \leq 2\eta.$$

Therefore, from Lemma 1 with $\rho = 2$, $C = R = \max_{i\in[1,M]}\left\|v_i^0 - \frac{x^0-\mathcal{T}_i(x^0)}{\gamma}\right\|$, we can prove that by choosing $\frac{\beta}{\alpha+R} < 1$ and $\eta \leq \frac{\gamma\beta R}{(1+\rho)(\alpha+R)}$, $\left\|v_i^{k+1} - \frac{x^{k+1}-\mathcal{T}_i(x^{k+1})}{\gamma}\right\| \leq R$. We complete the proof.

Next, from Lemma 1, $\left\|v_i^{k+1} - \frac{x^k-\mathcal{T}_i(x^k)}{\gamma}\right\| \leq \max_{i\in[1,M]}\left\|v_i^0 - \frac{x^0-\mathcal{T}_i(x^0)}{\gamma}\right\|$.

Step 2) Bound $f(x^k) - f^{\inf}$**.** From Lemma 4 with $G^k = \hat{v}^{k+1}$,

$$\begin{aligned} f(x^{k+1}) - f^{\inf} &\leq f(x^k) - f^{\inf} - \eta\left\|\nabla f(x^k)\right\| + 2\eta\left\|\nabla f(x^k) - \hat{v}^{k+1}\right\| + \frac{L\eta^2}{2} \\ &\overset{\text{triangle inequality}}{\leq} f(x^k) - f^{\inf} - \eta\left\|\nabla f(x^k)\right\| + 2\eta\left\|\nabla f(x^k) - v^{k+1}\right\| \\ &\quad +2\eta\left\|\hat{v}^{k+1} - v^{k+1}\right\| + \frac{L\eta^2}{2}, \end{aligned}$$

where $v^{k+1} = \frac{1}{M}\sum_{i=1}^M v_i^{k+1}$. Next, since

$$\begin{aligned} \left\|\nabla f(x^k) - v^{k+1}\right\| &= \left\|\nabla f(x^k) - \frac{1}{M}\sum_{i=1}^M v_i^{k+1}\right\| \\ &\overset{\text{triangle inequality}}{\leq} \frac{1}{M}\sum_{i=1}^M\left\|v_i^{k+1} - \nabla f_i(x^k)\right\| \\ &\overset{\text{triangle inequality}}{\leq} \frac{1}{M}\sum_{i=1}^M\left\|v_i^{k+1} - \frac{x^k-\mathcal{T}_i(x^k)}{\gamma}\right\| \\ &\quad +\frac{1}{M}\sum_{i=1}^M\left\|\frac{x^k-\mathcal{T}_i(x^k)}{\gamma} - \nabla f_i(x^k)\right\|, \end{aligned}$$

where $\mathcal{T}_i(x^k) = x^k - \gamma \frac{1}{N}\sum_{j=0}^{N-1} \nabla f_{i,j}(x_i^{k,j})$, we get

$$\left\|\nabla f(x^k) - v^{k+1}\right\| \leq \frac{1}{M}\sum_{i=1}^{M}\left\|v_i^{k+1} - \frac{x^k - \mathcal{T}_i(x^k)}{\gamma}\right\| + \frac{1}{\gamma}\frac{1}{M}\sum_{i=1}^{M}\left\|x^k - \mathcal{T}_i(x^k) - \gamma\nabla f_i(x^k)\right\|.$$

Plugging the upperbound for $\left\|\nabla f(x^k) - v^{k+1}\right\|$ into the main inequality in $f(x^k) - f^{\inf}$, we obtain

$$\begin{aligned} f(x^{k+1}) - f^{\inf} \quad &\leq \quad f(x^k) - f^{\inf} - \eta\left\|\nabla f(x^k)\right\| + 2\eta\frac{1}{M}\sum_{i=1}^{M}\left\|v_i^{k+1} - \frac{x^k - \mathcal{T}_i(x^k)}{\gamma}\right\| \\ &\quad +\frac{2\eta}{\gamma}\frac{1}{M}\sum_{i=1}^{M}\left\|(x^k - \gamma\nabla f_i(x^k)) - \mathcal{T}_i(x^k)\right\| + 2\eta\left\|\hat{v}^{k+1} - v^{k+1}\right\| + \frac{L\eta^2}{2}. \end{aligned}$$

By the fact that $\left\|v_i^{k+1} - \frac{x^k - \mathcal{T}_i(x^k)}{\gamma}\right\| \leq R$ from Step 1),

$$\begin{aligned} f(x^{k+1}) - f^{\inf} \quad &\leq \quad f(x^k) - f^{\inf} - \eta\left\|\nabla f(x^k)\right\| + 2\eta R \\ &\quad +\frac{2\eta}{\gamma}\frac{1}{M}\sum_{i=1}^{M}\left\|(x^k - \gamma\nabla f_i(x^k)) - \mathcal{T}_i(x^k)\right\| + 2\eta\left\|\hat{v}^{k+1} - v^{k+1}\right\| + \frac{L\eta^2}{2}. \end{aligned}$$

From Lemma 9,

$$\begin{aligned} f(x^{k+1}) - f^{\inf} \quad &\leq \quad f(x^k) - f^{\inf} - \eta\left\|\nabla f(x^k)\right\| + 2\eta R \\ &\quad +\frac{2\eta}{\gamma}\gamma L\frac{1}{M}\sum_{i=1}^{M}\frac{1}{N}\sum_{j=0}^{N-1}\left\|x_i^{k,j} - x^k\right\| + 2\eta\left\|\hat{v}^{k+1} - v^{k+1}\right\| + \frac{L\eta^2}{2}. \end{aligned}$$

Next, from Lemma 8,

$$\begin{aligned} f(x^{k+1}) - f^{\inf} \quad &\leq \quad \left(1 + \frac{4L\sqrt{2L}}{\sqrt{\Delta^{\inf}}}\gamma\eta\right)(f(x^k) - f^{\inf}) - \eta\left\|\nabla f(x^k)\right\| + 2\eta R \\ &\quad +4L\sqrt{2L}\gamma\eta\sqrt{\Delta^{\inf}} + 4L\sqrt{2L}\gamma\eta\sqrt{\frac{1}{M}\sum_{i=1}^{M}\Delta_i^{\inf}} \\ &\quad +2\eta\left\|\hat{v}^{k+1} - v^{k+1}\right\| + \frac{L\eta^2}{2}. \end{aligned}$$

Since

$$\begin{aligned} \mathrm{E}\left[\left\|\hat{v}^{k+1} - v^{k+1}\right\|\right] \quad &\leq \quad \frac{1}{\gamma}\mathrm{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}v_i^{k+1} - \hat{v}^{k+1}\right\|\right] \\ &\overset{\text{Lemma 3}}{\leq} \quad \frac{1}{\gamma}\sqrt{\frac{\beta^2 B}{M}(K+1)}, \end{aligned}$$

by taking the expectation,

$$\begin{aligned} \mathrm{E}\left[f(x^{k+1}) - f^{\inf}\right] \quad &\leq \quad \left(1 + \frac{4L\sqrt{2L}}{\sqrt{\Delta^{\inf}}}\gamma\eta\right)\mathrm{E}\left[f(x^k) - f^{\inf}\right] - \eta\mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] + 2\eta R \\ &\quad +8L\sqrt{2L}\gamma\eta\sqrt{\Delta^{\inf}} + 4L\sqrt{2L}\gamma\eta\sqrt{\frac{1}{M}\sum_{i=1}^{M}\Delta_i^{\inf}} \\ &\quad +2\eta\sqrt{\frac{\beta^2 B}{M}(K+1)} + \frac{L\eta^2}{2}. \end{aligned}$$

By applying Lemma 5 with $\eta\gamma \leq \frac{1}{K+1}\frac{\Delta^{\inf}}{4L\sqrt{2L}}$ and using the fact $(1+\eta\gamma\frac{4L\sqrt{2L}}{\Delta^{\inf}})^{K+1} \leq \exp(\eta\gamma\frac{4L\sqrt{2L}}{\Delta^{\inf}}(K+1)) \leq \exp(1) \leq 3$ we finalize the proof. □

G.3. Corollaries for Fed-α-NormEC with multiple local IG steps from Theorem 2

Corollary 5 (Convergence bound for Fed-α-NormEC with multiple local IG steps)**.** *Consider* Fed-α-NormEC *for solving Problem* (1) *under the same setting as Theorem* 2*. Let $T > 1$ (multiple local GD steps). If $\gamma = \frac{1}{2L(K+1)^{1/8}}$, $v_i^0 \in \mathbb{R}^d$ is chosen such that* $\max_{i\in[1,M]}\left\|\frac{x^0-\mathcal{T}_i(x^0)}{\gamma} - v_i^0\right\| = \frac{D_1}{(K+1)^{1/8}}$ *with $D_1 > 0$, and $\beta = \frac{D_2}{(K+1)^{5/8}}$ with $D_2 > 0$, and $\eta = \frac{\hat{\eta}}{(K+1)^{7/8}}$ with $\hat{\eta} = \min\left(\frac{\Delta^{\inf}}{2\sqrt{2L}}, \frac{D_1D_2}{4L(\alpha+D_1)}\right)$, then*

$$\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \leq \frac{A_1}{(K+1)^{1/8}} + \frac{A_2}{(K+1)^{7/8}},$$

where $A_1 = 3\frac{f(x^0)-f^{\inf}}{\hat{\eta}} + 2D_1 + \frac{2\sqrt{B}D_2}{\sqrt{M}} + 8\sqrt{2L}\sqrt{\Delta^{\inf}} + 4\sqrt{2L}\sqrt{\frac{1}{M}\sum\limits_{m=1}^{M}\Delta_i^{\inf}}$ *and* $A_2 = \hat{\eta}L/2$.

Proof. The proof is analogous to the proof of Corollary 3. □

Corollary 6 (Utility bound for Fed-α-NormEC with multiple local IG steps)**.** *Consider* Fed-α-NormEC *for solving Problem* (1) *under the same setting as Theorem* 2*. Let $T > 1$ (multiple local GD steps), let $\sigma_{\mathrm{DP}} = c\frac{p\sqrt{(K+1)\log(1/\delta)}}{\epsilon}$ with $c > 0$ (privacy with subsampling amplification), and let $p = \frac{\hat{B}}{M}$ for $\hat{B} \in [1, M]$ (client subsampling). If $\beta = \frac{\hat{\beta}}{K+1}$ with* $\hat{\beta} = \sqrt{\frac{3(f(x^0)-f^{\inf})}{\gamma}}\sqrt[4]{\frac{M}{B_2}}$, $\gamma < \frac{\Delta^{\inf}(\alpha+R)}{\sqrt{2L}\hat{\beta}R}$, $\alpha = R = \mathcal{O}\left(\sqrt[4]{d}\frac{\sqrt{f(x^0)-f^{\inf}}}{\sqrt{\gamma}}\sqrt[4]{\frac{B_2}{M}}\right)$ *with* $B_2 = 2c^2\frac{\hat{B}}{M}\frac{\log(1/\delta)}{\epsilon^2}$, *and* $\eta = \frac{1}{K+1}\frac{\gamma}{2}\frac{\hat{\beta}R}{\alpha+R}$, *then*

$$\min_{k\in[0,K]} \mathrm{E}\left[\left\|\nabla f(x^k)\right\|\right] \leq \mathcal{O}\left(\Delta\sqrt[4]{\frac{d\hat{B}}{M^2}\frac{\log(1/\delta)}{\epsilon^2}} + \sqrt{L}\sqrt{\Delta^{\inf}} + \sqrt{L}\sqrt{\frac{1}{M}\sum_{i=1}^{M}\Delta_i^{\inf}}\right),$$

where $\Delta = \max(\alpha, 2)\sqrt{L}\sqrt{f(x^0) - f^{\inf}}$.

Proof. The proof is analogous to the proof of Corollary 4. □