Differentially Private Federated Learning with Laplacian Smoothing

Zhicong Liang[a], Bao Wang[b], Quanquan Gu[c], Stanley Osher[d], Yuan Yao[a,*]

[a]*Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China*
[b]*Department of Mathematics and Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, UT 84112, United States*
[c]*Department of Computer Science, University of California, Los Angeles, CA 90095, United States*
[d]*Department of Mathematics, University of California, Los Angeles, CA 90095, United States*

Abstract

Federated learning aims to protect data privacy by collaboratively learning a model without sharing private data among users. However, an adversary may still be able to infer the private training data by attacking the released model. Differential privacy provides a statistical protection against such attacks at the price of significantly degrading the accuracy or utility of the trained models. In this paper, we investigate a utility enhancement scheme based on Laplacian smoothing for differentially private federated learning (DP-Fed-LS), to improve the statistical precision of parameter aggregation with injected Gaussian noise without losing privacy budget. Our key observation is that the aggregated gradients in federated learning often enjoy a type of smoothness, *i.e.* sparsity in a graph Fourier basis with polynomial decays of Fourier coefficients as frequency grows, which can be exploited by the Laplacian smoothing efficiently. Under a prescribed differential privacy budget, convergence error bounds with tight rates are provided for DP-Fed-LS with uniform subsampling of heterogeneous **non-iid** data, revealing possible utility improvement of Laplacian smoothing in effective dimensionality and variance reduction, among others. Experiments over MNIST, SVHN, and Shakespeare datasets show that the proposed method can improve model accuracy with DP-guarantee and membership privacy under both uniform and Poisson

*Corresponding author

Email addresses: zliangak@connect.ust.hk (Zhicong Liang), bwang@math.utah.edu (Bao Wang), qgu@cs.ucla.edu (Quanquan Gu), sjo@math.ucla.edu (Stanley Osher), yuany@ust.hk (Yuan Yao)

subsampling mechanisms.

Keywords: Differential Privacy, federated learning, Laplacian smoothing

1. Introduction

In recent years, we have already witnessed the great success of machine learning algorithms in handling large-scale and high-dimensional data (He et al., 2016; Devlin et al., 2018; Silver et al., 2016; Berner et al., 2019; Senior et al., 2020). Most of these models are trained in a centralized manner by gathering all data into a single database. However, in applications like mobile keyboard development (Hard et al., 2018), speech recognition (Jiang et al., 2021), and autonomous driving (Nguyen et al., 2022), sensitive data are distributed in the devices of users, who are not willing to share their own data with others. Federated learning (FL), proposed in (McMahan et al., 2017), provides a solution that data owners can collaboratively learn a useful model without disclosing their private data. In FL, a server, and multiple data owners, referred to as clients, are involved in maintaining a global model. They no longer share the private data but the updated models trained on these data.

In some cases, however, federated learning is not sufficient to protect the sensitive data by simply decoupling the model training from the direct access to the raw training data (Shokri et al., 2017; Fredrikson et al., 2014, 2015). Information about raw data can still be identified from a well-trained model. In some extreme cases, a neural network can even memorize the whole training set with its huge number of parameters. For example, an adversary may infer the presence of particular records in training (Shokri et al., 2017) or even recover the identity (e.g. face images) in the training set by attacking the released model (Fredrikson et al., 2015, 2014). Differential privacy (DP) provides us with a solution to defend against these threats (Dwork and Nissim, 2004; Dwork et al., 2006). DP guarantees privacy in a statistical way that the well-trained models are not sensitive to the change of an individual record in the training set. This task is usually fulfilled by adding noise, calibrated to the model's sensitivity, to the outputs or the updates.

One major deficiency of DP lies in its potential significant degradation of the utility of the models due to the noise injection. Laplacian smoothing (LS) has recently been shown to be a good choice for reducing noise in noisy gradient, e.g. in stochastic gradient descent (SGD) (Osher et al., 2022), and thus promising for utility improvement in machine learning with DP (Wang et al., 2020). However,

due to the heterogeneity of client data distributions in federated learning, it remains open how to apply Laplacian smoothing effectively to such settings.

To fill in this gap, we develop in this paper a framework of exploiting Laplacian smoothing to improve the utility of the differentially private federated learning (DP-Fed) while maintaining the same DP budget.

1.1. Main Contributions

The major contributions of our work are summarized as follows.

- Our key observation in federated learning is that the *federated average of gradients* are often smooth or sparse in Fourier basis with polynomial decays. If we can capture the smooth or sparse signal in Fourier basis with corresponding low-pass filters, then we can reduce the variance and get a better estimate. Therefore Laplacian smoothing of the *federated average of gradients* is introduced to the differentially private federated learning, that can reduce variance with improved estimates of such gradients. With the aid of 1-D Fast Fourier Transform (FFT), such a Laplacian smoothing can be efficiently computed on the server. We denote the proposed algorithm as DP-Fed-LS.

- Convergence bounds under heterogeneous data distributions are developed for DP-Fed-LS in strongly-convex, general-convex, and non-convex settings under our differential privacy budget bounds. We show how Laplacian smoothing can help reduce the true dimension factor d in the differential privacy error term to an effective dimension $d_\sigma \leq d$, which helps alleviate the degeneration introduced by DP. The rates on convergence and communication complexity match those on federated learning without DP (Karimireddy et al., 2020), while our results extend to include the effect of differential privacy and Laplacian smoothing; as well as our rates match the ones of empirical risk minimization (ERM) via SGD with differential privacy in a centralized setting (Bassily et al., 2019; Wang et al., 2019a). See Table 1 for a comparison.

- The utility of Laplacian smoothing in DP-Fed is demonstrated by training a logistic regression model over MNIST, a convolutional neural network (CNN) over extended SVHN, in an **iid** fashion, and a long short-term memory (LSTM) model over Shakespeare dataset in a **non-iid** setting. These experiments show that DP-Fed-LS improves accuracy while providing at

least the same DP-guarantees and membership privacy as DP-Fed with two subsampling mechanisms across different datasets.

Method	μ strongly-convex	non-convex
DP-SGD (Bassily et al., 2014)	$\frac{d\log^3(N/\delta)}{\mu N^2\varepsilon^2}$	–
DP-SVRG (Wang et al., 2017)†	$\frac{d\log(N)\log(1/\delta)}{\mu N^2\varepsilon^2}$	$\frac{\sqrt{d\log(1/\delta)}}{N\varepsilon}$
DP-SRM (Wang et al., 2019a)	–	$\frac{\sqrt{d\log(1/\delta)}}{N\tilde{n}\varepsilon}$
DP-SGD-LS (Wang et al., 2020)	–	$\frac{\sqrt{\tilde{d}_\sigma\log(1/\delta)}}{N\varepsilon}$
DP-Fed-LS*	$\frac{d_\sigma\log(S)\log(1/\delta)}{\mu N^2\varepsilon^2}$	$\frac{\sqrt{d_\sigma\log(1/\delta)}}{N\varepsilon}$
Fed-Avg (Li et al., 2019)†	$\frac{\varsigma^2(0)}{\mu^2 NK\epsilon}+\frac{G^2K}{\mu^2\varepsilon}$	–
Fed-Avg (Khaled et al., 2020)†	$\frac{\varsigma^2(0)+G^2}{\mu NK\epsilon}+\frac{\varsigma(0)+G}{\mu\sqrt{\epsilon}}+\frac{NB}{\mu}$ ‡	–
Fed-Avg (Karimireddy et al., 2020)	$\frac{\varsigma^2(0)}{\mu SK\epsilon}+\frac{(1-\tau)G^2}{\mu S\epsilon}+\frac{G}{\mu\sqrt{\epsilon}}+\frac{B^2}{\mu}$ ‡	$\frac{\varsigma^2(0)}{SK\epsilon^2}+\frac{(1-\tau)G^2}{S\epsilon^2}+\frac{G}{\epsilon^{3/2}}+\frac{B^2}{\epsilon}$
DP-Fed-LS	$\frac{\varsigma^2(\sigma)}{\mu SK\epsilon}+\frac{(1-\tau)G^2}{\mu S\epsilon}+\frac{(1+4\sigma)G}{\mu\sqrt{\epsilon}}+\frac{(1+4\sigma)^2B^2}{\mu}+\frac{d_\sigma L\nu_1^2}{\mu S^2\epsilon}$ ‡	$\frac{\varsigma^2(\sigma)}{SK\epsilon^2}+\frac{(1-\tau)G^2}{S\epsilon^2}+\frac{(1+4\sigma)G}{\epsilon^{3/2}}+\frac{(1+4\sigma)^2B^2}{\epsilon}+\frac{d_\sigma L^2\nu_1^2}{S^2\epsilon^2}$

Table 1: Utility guarantee of (ε,δ)-DP (upper part) and rate of communication round needed to achieve ϵ accuracy (lower part) for μ strongly-convex and non-convex optimization problems. ‡ denotes that logarithmic factors are ignored here. See Appendix D for more details. † denotes that no client subsampling is used. In full participation scenarios, $\tau=1$ and $\log(S)=\log(N)$. $*$ after DP-Fed-LS further denotes the specific setting of **iid** ($G=0$) with $K\gg 1$. The effective dimension $d_\sigma=\sum_{i=1}^d\Lambda_i$ and $\tilde{d}_\sigma=\sum_{i=1}^d\Lambda_i^2$, where $\Lambda_i\leq 1$ is the eigenvalue of $\mathbf{A}_\sigma^{-1}$. For centralized settings (Bassily et al., 2014; Wang et al., 2017, 2020), N denotes the number of data points, while in federated learning, N denotes the number of clients. DP-SRM (Wang et al., 2019a) is a distributed setting where N and $\tilde{n}$ denote the number of clients and number of samples owned by each client, respectively. They consider data-level DP while we consider user-level DP.

1.2. Background and Related Works

Risk of Federated Learning. Despite its decoupling of training from direct access to raw data, federated learning may suffer from the risk of privacy leakage by

unintentionally allowing malicious clients to participate in the training (Hitaj et al., 2017; Melis et al., 2019; Zhu et al., 2019). In particular, model poisoning attacks are introduced in (Bagdasaryan et al., 2020; Bhagoji et al., 2019). Even though we can ensure the training is private, the released model may also leak sensitive information about the training data. Fredrikson et al. (2014, 2015) introduce the model inversion attack that can infer sensitive features or even recover the input given a model. Membership inference attacks can determine whether a record is in the training set by leveraging the ubiquitous overfitting of machine learning models (Shokri et al., 2017; Yeom et al., 2018; Sablayrolles et al., 2019). In these cases, simply decoupling the training from direct access to private data is insufficient to guarantee data privacy.

Differential Privacy. Differential privacy comes as a solution for privacy protection. Gradient perturbation (Bassily et al., 2014; Abadi et al., 2016) receives lots of recent attention in ML applications since it admits the public training process and ensures DP guarantee even for a non-convex objective. Papernot et al. (2017, 2018) propose PATE that bridges the target model and training data by multiple teacher models. Mironov (2017) proposes a natural relaxation of DP based on Rényi divergence (RDP), which allows tighter analysis of composite heterogeneous mechanisms. Wang et al. (2019b) provide a tight numerical upper bound on RDP parameters for randomized mechanism with uniform subsampling. Furthermore, they extend their bound to the case of Poisson subsampling (Zhu and Wang, 2019), which is the same as the one in (Mironov et al., 2019). Our differential privacy guarantees are based on these two numerical results, and we derive new closed-form bounds which are more precise or tighter than previous works (Wang et al., 2019a; Mironov et al., 2019; Bun et al., 2018).

Differential Privacy in Distributed Settings. DP has been applied in many distributed learning scenarios. Pathak et al. (2010) propose the first DP training protocol in distributed setting. Jayaraman et al. (2018) reduce the noise needed in (Pathak et al., 2010) by firstly training DP local models and then performing naive aggregation. Zhang et al. (2019) propose to decouple the feature extraction from the training process, where clients only need to extract features with frozen pre-trained convolutional layers and perturb them with Laplace noise. However, this method needs to introduce extra edge servers besides the central server in the standard federated learning.

Geyer et al. (2017) and McMahan et al. (2018b) consider a similar problem setting as this paper, which applies the Gaussian mechanism in federated learning to ensure DP. However, Geyer et al. (2017) only train models over MNIST, with repetition of the data across different clients, which is unrealistic in applications.

McMahan et al. (2018a) use moment accountant in (Mironov et al., 2019; Zhu and Wang, 2019), and show that given a sufficiently large number of clients ($\sim$ 760K in their example), their models suffer no utility degradation. However, in many scenarios, one has to deal with a much smaller number of clients, which will induce a large noise level with the same DP constraint, significantly reducing the utility of the models. This motivates us to leverage Laplacian smoothing to mitigate the utility degradation due to DP, broadening its scope of application, and we further provide convergence bounds and evaluate the membership privacy of our method by the membership inference attack, comparing with (Geyer et al., 2017; McMahan et al., 2018a).

1.3. Paper Organization

This paper is organized as follows. Section 2 presents our proposed algorithm of differentially private federated learning with Laplacian smoothing (DP-Fed-LS), and demonstrates our key observation and motivation to apply LS. In Section 3, we characterize the privacy budget such that our algorithm satisfies (ε, δ)-differential privacy guarantee. In Section 4, we provide a convergence analysis for DP-Fed-LS, which characterizes the influences on the optimization error and communication complexity of differential privacy and Laplacian smoothing with heterogeneous data. In Section 5, we demonstrate the utility of DP-Fed-LS with three applications, *i.e.* MNIST, SVHN, and Shakespeare (**non-iid**) datasets. Conclusion is given in Section 6. Appendices collect all the remaining proofs and empirical results in this paper.

Reproducible source codes can be downloaded at:

`https://github.com/zliangak/dp-fed-ls`.

2. Differentially Private Federated Learning with Laplacian Smoothing

In this section, we formulate the basic scheme of private (noisy) federated learning with Laplacian smoothing. Consider the following distributed optimization model,

$$\min_{w} f(w) \coloneqq \frac{1}{N}\sum_{j=1}^{N} f_j(w), w \in \mathbb{R}^d$$

where f_j represent the loss function of client j, and N is the number of clients. Here $f_j(w) = \mathbb{E}_{x_j} f_j(w, x_j, y_j)$, where $\mathbb{E}_{x_j}$ is the expectation over the dataset of the j-th client.

We propose *differentially private federated learning with Laplacian smoothing* (DP-Fed-LS), which is summarized in Algorithm 1, to solve the above optimization problem. In each communication round t, the server distributes the global model w^t to a selected subset out of N total clients. These selected (active) clients will perform K steps mini-batch SGD to update the models on their private data, and send back the model update $\Delta_j^{t,K}$s, from which the server will aggregate and yield a new global model w^{t+1}. This process will be repeated until the global model converges. We call a setting **iid** if data of different clients are sampled from the same distribution independently. Otherwise, if data from different clients are independent but not identically distributed, we will call it **non-iid** setting (McMahan et al., 2017; Li et al., 2019). In **non-iid** setting, the data from each client will depend on the her/his characteristic and lack representativeness for the whole population.

In each update of the mini-batch SGD, we bound the local model $w_j^{t,i}, i \in [K]$ within a $\mathcal{L}$-ball ($\mathcal{L} > 0$) centering around w^t by clipping: $\text{clip}(v, \mathcal{L}) \leftarrow v/\max(1, \|v\|_2/\mathcal{L})$. In each round, we regard the aggregation of locally-trained models as the *federated average of gradients*, where we add calibrated Gaussian noise $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \nu^2\mathbf{I})$ to guarantee DP. Then we apply Laplacian smoothing with a smoothing factor σ on the noisy aggregated *federated average of gradients* (Eq. (*) in Algorithm 1), to stabilize the training while preserving DP based on the post-processing lemma (Proposition 2.1 of (Dwork and Roth, 2014)). It will reduces to DP-Fed if $\mathbf{A}_\sigma = \mathbf{I}$, *e.g.* $\sigma = 0$.

It is worth to note that Laplacian smoothing is only applied to the global update in federated average, while in the local update, the general mini-batch SGD is applied. One may wonder if we apply the same Laplacian smoothing on both local and global updates like (Wang et al., 2020). However, the empirical performance of such a proposal will become worse in federated learning, because the noise scales in local and global updates are significantly different, especially with heterogeneous **non-iid** data over clients. Therefore, it is difficult to find a unified Laplacian smoothing that can achieve a good trade-off between bias and variance (as discussed below Proposition 1) for both updates. As we shall see in the below, smoothness or sparsity still holds with the *federated average of gradients*, despite the heterogeneity over clients.

2.1. Laplacian Smoothing

To understand the Laplacian smoothing in DP-Fed-LS, consider the following general iteration:

$$w^{t+1} = w^t - \eta \mathbf{A}_\sigma^{-1} \nabla f(w^t, x_{i_t}, y_{i_t}), \tag{1}$$

Algorithm 1 Differentially-Private Federated Learning with Laplacian Smoothing (DP-Fed-LS)

parameters:
activate client fraction $\tau \in (0, 1]$
total communication round T
clipping parameter $\mathcal{L}$
local and global learning rate η_l, η_g
noise level ν

function CLIENTUPDATE(j, w^t)
$w_j^{t,0} \leftarrow w^t$
for $i = 0$ **to** $K - 1$ **do**
$g_j(w_j^{t,i}) \leftarrow$ mini-batch gradient
$w_j^{t,i+1} \leftarrow w^t + \text{CLIP}\big(w_j^{t,i} - \eta_l g_j(w_j^{t,i}) - w^t, \mathcal{L}\big)$
end for
return $\Delta_j^t \leftarrow w_j^{t,K} - w^t$
function CLIP($v, \mathcal{L}$) return $v/\max(1, \|v\|_2/\mathcal{L})$

Server executes:
initialize w^0
for $t = 0$ **to** $T - 1$ **do**
$\mathcal{S}_t \leftarrow$ (a random subset of clients selected by uniform or Poisson subsampling with ratio τ)
$S \leftarrow |\mathcal{S}_t|$
for client $j \in \mathcal{S}_t$ **in parallel do**
$\Delta_j^t \leftarrow$ CLIENTUPDATE(j, w^t)
end for
$\Delta^t \leftarrow \frac{\eta_g}{S} \mathbf{A}_\sigma^{-1}\big(\sum_{j=1}^S \Delta_j^t + \mathcal{N}(\mathbf{0}, \nu^2 \mathbf{I})\big)(*)$
$w^{t+1} \leftarrow w^t + \Delta^t$
end for
Output $\bar{w}^T = \sum_{t=0}^T a_t w^t / (\sum a_t)$, for some $a_t \geq 0$.

where η is the learning rate and $f(w, x_{i_t}, y_{i_t})$ is the loss of a given model with parameter w on the training data $\{x_{i_t}, y_{i_t}\}$. In Laplacian smoothing (Osher et al., 2022), we let $\mathbf{A}_\sigma = \mathbf{I} + \sigma \mathbf{L}$, where $\mathbf{L} \in \mathbb{R}^{d \times d}$ is the 1-dimensional Laplacian matrix of a cycle graph, *i.e.* $\mathbf{A}_\sigma$ a circulant matrix whose first row is $(1 + 2\sigma, -\sigma, 0, \cdots, 0, -\sigma)$ with $\sigma \geq 0$ being a constant. When $\sigma = 0$, Laplacian smoothing stochastic gradient descent reduces to SGD.

Laplacian smoothing can be effectively implemented by using the fast Fourier transform. To be specific, for any 1-D signal v (a flattened layer of $\nabla f(w^t, x_{i_t}, y_{i_t})$ in our case), we would like to calculate $u = \mathbf{A}_\sigma^{-1} v$. Since $v = \mathbf{A}_\sigma u = u - \sigma d * u$, where $d = [-2, 1, 0, ..., 0, 1]^T$ and $*$ denotes the convolutional operator. We have the following equality by exploiting the 1-D fast Fourier transform (FFT)

$$\text{fft}(v) = \text{fft}(u) \cdot \big(1 - \sigma \cdot \text{fft}(d)\big), \tag{2}$$

where $\cdot$ is point-wise multiplication. In other words, the Laplacian matrix $\mathbf{L}$ has eigenvectors defined by the Fourier basis, which diagonalizes convolutions via 1-D fast Fourier transform. Going back to Eq. (2), we solve u by applying the inverse

Fourier transform

$$u = \text{ifft}\Big(\frac{\text{fft}(v)}{1 - \sigma \cdot \text{fft}(d)}\Big).$$

The motivation behind Laplacian smoothing lies in that if the target parameter v is smooth under Fourier basis, then when it is contaminated by Gaussian noise, *i.e.* $\tilde{v} = v + \mathbf{n}$, $v \in \mathbb{R}^d$, $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \nu^2\mathbf{I})$, a smooth approximation of $\tilde{v}$ is helpful to reduce the noise. The Laplacian smoothing estimate is defined by

$$\hat{v}_{LS} := \arg\min_u \|u - \tilde{v}\|_2^2 + \sigma\|\nabla u\|_2^2, \tag{3}$$

where ∇ is a 1-dimensional gradient operator such that $\mathbf{L} = \nabla^{\mathbf{T}}\nabla$. It satisfies $\mathbf{A}_\sigma \hat{v}_{LS} = \tilde{v} = v + \mathbf{n}$. The following proposition characterizes the prediction error of Laplacian smoothing estimate $\hat{v}_{LS}$.

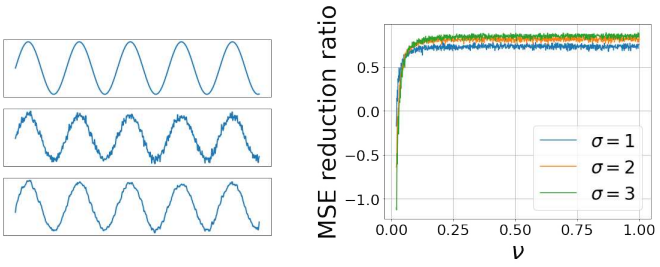
Proposition 1 (Bias-Variance decomposition)**.** *Let the graph Laplacian have eigen decomposition $\Delta\mathbf{e}_i = \lambda_i\mathbf{e}_i$ with eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_d$ and the first eigenvector $\mathbf{e_1} = \mathbf{1}/\sqrt{d}$. Then the mean square error (risk) of estimate $\hat{v}_{LS}$ admits the following decomposition,*

$$\begin{aligned}\mathcal{R}(\hat{v}_{LS}) := \mathbb{E}\|\hat{v}_{LS} - v\|_2^2 &= \|(\mathbf{I} - \mathbf{A}_\sigma^\dagger)v\|_2^2 + \mathbb{E}\|\mathbf{A}_\sigma^\dagger\mathbf{n}\|_2^2 \\ &= \sum_i \frac{\sigma^2\lambda_i^2}{(1+\sigma\lambda_i)^2}\langle v, \mathbf{e}_i\rangle^2 + \sum_i \frac{\nu^2}{(1+\sigma\lambda_i)^2},\end{aligned}$$

where the first term is called the **bias** *and the second term is called the* **variance**.

In the bias-variance decomposition of the risk above, if $\sigma = 0$, the risk becomes bias-free with variance $d\nu^2$; if $\sigma > 0$, bias is introduced while variance is reduced. The optimal choice of σ must depend on an optimal trade-off between the bias and variance in this case. When the true parameter v is smooth, in the sense that its projections $\langle v, \mathbf{e}_i\rangle \to 0$ rapidly as i increases, the introduction of bias can be much smaller compared to the reduction of variance, hence the mean squared error (risk) can be reduced with Laplacian smoothing. A bias-variance trade-off with similar idea for graph neural network can be found in (Nt and Maehara, 2019).

To illustrate the Proposition 1, in Figure 1, we show the efficacy of Laplacian smoothing. We consider a vector signal $y = \sin(x)$ where x is a vector of size 500, whose entries are evenly spaced over $[0, 30]$. We perturb it by Gaussian noise: $\tilde{y} = y + \mathbf{n}$ where $\mathbf{n} \sim \mathcal{N}(0, \nu^2)$. Then we get the Laplacian smoothing estimate $\hat{y}_{LS} := \arg\min_u \|u - \tilde{y}\|^2 + \sigma\|\nabla u\|^2$. From Figure 1 (a), we notice that $\hat{y}_{LS}$ can significantly smooth the noisy signal. Then in Figure 1 (b), we



(a) (b)

Figure 1: Efficacy of Laplacian smoothing. In (a), signals from top to bottom are $y = \sin(x)$, where x is a vector of size 500, whose entries are evenly spaced over [0, 30]. $\tilde{y} = y + \mathcal{N}(0, \nu^2)$ with $\nu = 0.1$ and $\hat{y}_{LS}$ with $\sigma = 1$. In (b), we compute the MSE reduction ratio of Laplacian smoothing estimator $(\text{MSE}(\tilde{y}) - \text{MSE}(\hat{y}_{LS}))/\text{MSE}(\tilde{y})$ along different noise level ν, where $\hat{y}_{LS} := \arg\min_u \|u - \tilde{y}\|^2 + \sigma\|\nabla u\|^2$.

compute the MSE reduction ratio of Laplacian smoothing estimator: $(\text{MSE}(\tilde{y}) - \text{MSE}(\hat{y}_{LS}))/\text{MSE}(\tilde{y})$ to demonstrate the efficacy of Laplacian smoothing. We see that, when the noise level ν is small, Laplacian smoothing will introduce higher MSE: the bias introduced by Laplacian smoothing is larger than its variance reduction. However, once the noise level increases, Laplacian smoothing will significantly reduce the MSE. The larger the σ is, the more MSE reduction achieved.

2.2. Sparsity of Aggregated Gradients in the Fourier Basis

To verify that the true signal v is smooth or sparse with respect to the Fourier basis, we show in Figure 2 the magnitudes distribution of CNN by layers in frequency domain of $v = \frac{1}{S}\sum_j \Delta_j^t$, in non-DP (noise free and no clipping) federated learning under the fast Fourier transform. We will firstly flatten the weights into a 1-D vector layer-wise, by the natural order of Pytorch, and then perform the FFT on them layer-by-layer. We use the experimental setting described in Section 5, where our CNN stacks two 5 × 5 convolutional layers with max-pooling, two fully-connected layers with 384 and 192 units, respectively, and a final softmax output layer. The patterns in four different training communication rounds ($t = 1, 50, 100, 200$) are shown. One can see that from the log-log plot, as the communication round and frequency grow, the magnitudes of Fourier coefficients demonstrate a power law decay with respect to the frequency, indicated by a linear envelope between $\log_{10}(\text{Magnitude})$ and $\log_{10}(\text{Frequency})$ when $\log_{10}(\text{Frequency})$ increases. In other words, it shows that the projections of magnitudes $\langle v, \mathbf{e}_i \rangle \to 0$ at a polynomial rate when the frequency in Fourier basis is large enough, supporting

the assumption above for variance reduction. In Figure G.10 in Appendix G, we also show that the frequency distribution of *federated average of gradients* is insensitive to different flattening orders and permutation of the output channel weight indices of convolutional layer. What's more we visualize that $\langle v, \mathbf{e}_i \rangle$ goes to 0 rapidly as i increases in Figure G.11 in Appendix G.

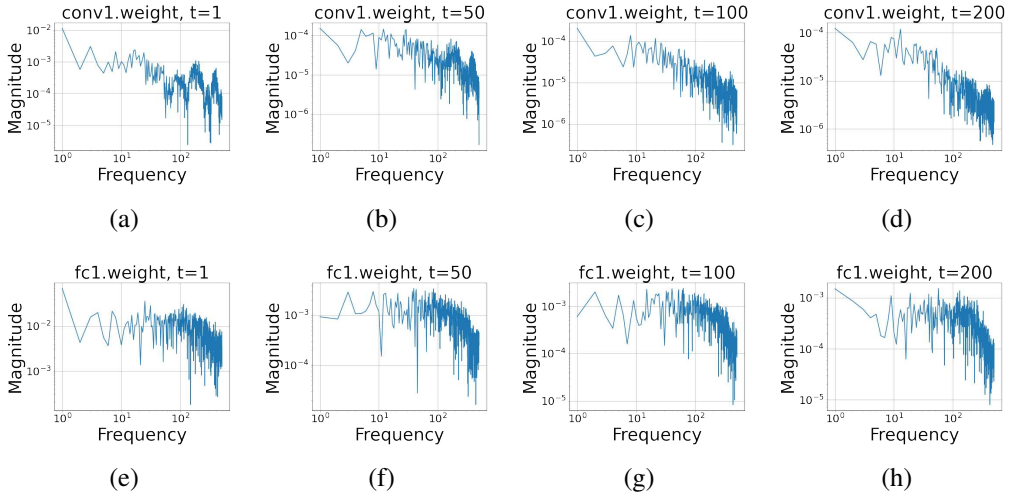


Figure 2: Frequency distribution of *federated average of gradients* $v = \frac{1}{S}\sum_j \Delta_j^t$ over different CNN layers and communication rounds t in non-DP federated learning, following experiment setting in Section 5. Here we use the first convolutional layer (conv1.weight) and the first fully-connected layer (fc1.weight) as an example.

2.3. Connection to Mirror Descent

Laplacian smoothing can also be viewed as a special case of mirror descent (Nemirovskij and Yudin, 1983) by setting $\Phi(w) = \frac{1}{2}\|w\|^2_{\mathbf{A}_\sigma}$. It can be also regarded as a case of natural gradient descent where the parameter space adopts a Riemannian metric $\|\cdot\|_{\mathbf{A}_\sigma}$ (Amari, 1998). In mirror descent, we update our parameters by

$$w^{t+1} = \arg\min_w \{\eta\langle \nabla f(w^t, x_{i_t}, y_{i_t}), w\rangle + D_h(w\|w^t)\},$$

where $D_h(y\|x) = \Phi(y) - \Phi(x) - \langle \nabla\Phi(x), y - x\rangle$ is the Bregman divergence. Setting the gradient at w^{t+1} to zero gives

$$\eta\nabla f(w^t, x_{i_t}, y_{i_t}) + \nabla\Phi(w^{t+1}) - \Phi(w^t) = 0.$$

Since $\nabla\Phi(w) = \mathbf{A}_\sigma w$, we have

$$w^{t+1} = w^t - \eta \mathbf{A}_\sigma^{-1} \nabla f(w^t, x_{i_t}, y_{i_t}),$$

which reduces to Eq. (1). From this point of view, Laplacian smoothing serves as a regularizer that constrains the difference between w^{t+1} and w^t to be smooth under the Fourier basis, which agrees with the discussion above.

In Appendix F, we demonstrate an additional classification example where Laplacian smoothing reaches improved estimates of smooth signals (parameters) against Gaussian noise. Among a variety of usages such as reducing the variance of SGD on-the-fly, escaping spurious minima, and improving generalization in training many machine learning models including neural networks (Osher et al., 2022; Wang et al., 2020), the Laplacian smoothing in this paper particularly improves the utility when Gaussian noise is injected to federated learning for privacy, which will be discussed in the following sections.

3. Differential Privacy Guarantee

In this section, we provide closed-form DP guarantees for differentially private federated learning, with or without LS, under both scenarios that active clients are sampled with uniform subsampling or with Poisson subsampling.

First of all, recall the definition of differential privacy and Rényi differential privacy (RDP).

Definition 1 ((ε,δ)-DP)**.** *(Dwork and Roth, 2014) A randomized mechanism $\mathcal{M}: \mathcal{D} \to \mathbb{R}^d$ satisfies (ε,δ)-DP if for any two adjacent datasets $D, D' \in \mathcal{D}$ differing by only one element, and any output subset $O \subseteq \mathbb{R}^d$, it holds that*

$$\mathbb{P}[\mathcal{M}(D) \in O] \leq e^\varepsilon \cdot \mathbb{P}[\mathcal{M}(D') \in O] + \delta.$$

Definition 2 ((α,ρ)-RDP)**.** *(Mironov, 2017) For $\alpha > 1$ and $\rho > 0$, a randomized mechanism $\mathcal{M}: \mathcal{D} \to \mathbb{R}^d$ satisfies (α,ρ)-Rényi DP, i.e. (α,ρ)-RDP, if for all adjacent datasets $D, D' \in \mathcal{D}$ differing by one element, it has*

$$D_\alpha\big(\mathcal{M}(D)||\mathcal{M}(D')\big) := \frac{1}{\alpha-1} \log \mathbb{E}\big(\mathcal{M}(D)/\mathcal{M}(D')\big)^\alpha \leq \rho.$$

Lemma 1 (From (α,ρ)-RDP to (ε,δ)-DP)**.** *(Mironov, 2017) If a randomized mechanism $\mathcal{M}: \mathcal{D} \to \mathbb{R}^d$ satisfies (α,ρ)-RDP, then $\mathcal{M}$ satisfies $(\rho+\log(1/\delta)/(\alpha-1),\delta)$-DP for all $\delta \in (0,1)$.*

In federated learning, we consider the user-level DP. So the terms **element** and **dataset** in the definition will refer to a single client, and a set of clients respectively in our scenario. There are two ways to construct a subset of active clients. The first one is uniform subsampling, *i.e.* in each communication round, a subset of fixed size $S = \tau \cdot N$ of clients are sampled uniformly. The second one is Poisson subsampling, which includes each client in the subset with probability τ independently. If we trace back to the definition, this subtle difference actually comes from the difference of how we construct the adjacent datasets D and D'. For uniform subsampling, D and D' are adjacent if and only if there exist two samples $a \in D$ and $b \in D'$ such that if we replace a in D with b, then D is identical with D' (Dwork and Roth, 2014). However, for Poisson subsampling, D and D' are said to be adjacent if $D \cup \{a\}$ or $D \backslash \{a\}$ is identical to D' for some sample a (Mironov et al., 2019; Zhu and Wang, 2019). This subtle difference results in two different parallel scenarios below.

Theorem 1 (Differential Privacy Guarantee for DP-Fed-LS with Uniform Subsampling)**.** *For any $\delta \in (0, 1)$, $\varepsilon > 0$, DP-Fed or DP-Fed-LS with uniform subsampling, satisfies (ε,δ)-DP when the variance of the injected Gaussian noise $\mathcal{N}(0, \nu^2)$ satisfies*

$$\nu \geq \frac{\tau \mathcal{L}}{\varepsilon} \sqrt{\frac{14T}{\lambda} \left(\frac{\log(1/\delta)}{1-\lambda} + \varepsilon \right)}, \tag{4}$$

if there exists $\lambda \in (0, 1)$ such that $\nu^2/4\mathcal{L}^2 \geq \frac{2}{3}$ and $\alpha - 1 \leq \frac{\nu^2}{6\mathcal{L}^2} \log(1/(\tau\alpha(1 + \nu^2/4\mathcal{L}^2)))$, where $\alpha = \log(1/\delta)/((1-\lambda)\varepsilon) + 1$.

Theorem 2 (Differential Privacy Guarantee for DP-Fed-LS with Poisson Subsampling)**.** *For any $\delta \in (0, 1)$, $\varepsilon > 0$, DP-Fed or DP-Fed-LS with Poisson subsampling, satisfies (ε,δ)-DP when its injected Gaussian noise $\mathcal{N}(0, \nu^2)$ is chosen to be*

$$\nu \geq \frac{\tau \mathcal{L}}{\varepsilon} \sqrt{\frac{2T}{\lambda} \left(\frac{\log(1/\delta)}{1-\lambda} + \varepsilon \right)}, \tag{5}$$

if there exists $\lambda \in (0, 1)$ such that $\nu^2/\mathcal{L}^2 \geq \frac{5}{9}$ and $\alpha - 1 \leq \frac{2\nu^2}{3\mathcal{L}^2} \log(1/(\tau\alpha(1 + \nu^2/\mathcal{L}^2)))$, where $\alpha = \log(1/\delta)/((1-\lambda)\varepsilon) + 1$.

Theorem 1 and Theorem 2 characterize the closed-form relationship between (ε, δ)-DP and the corresponding noise level ν, based on the numerical results in (Wang et al., 2019b; Zhu and Wang, 2019; Mironov et al., 2019). As we can see later, they will also serve as backbone theorems when we analyse the optimization

error bounds of DP-Fed-LS. The two conditions in the above theorems are used for inequality scaling. In practical implementation, we will do a grid search of $\lambda \in (0, 1)$ and select the one that gives the smallest lower bound of ν while satisfying both conditions. After that, we set ν to its lower bound.

Proofs of Theorem 1 and Theorem 2 are given in Appendix A and Appendix B . These closed-form bounds are of similar rates as the numerical moment accountant (Wang et al., 2019b; Zhu and Wang, 2019; Mironov et al., 2019) up to a constant (see Appendix E).

4. Convergence with Differential Privacy Guarantee

In this section, convergence and communication complexity bounds are provided for DP-Fed-LS in Algorithm 1 with uniform subsampling.

First of all, we state several commonly used assumptions adapted for the non-Euclidean geometry in Laplacian smoothing. In the following statements, the primal norm $\|\cdot\| = \|\cdot\|_{\mathbf{A}_\sigma}$ and its dual norm $\|\cdot\|_* = \|\cdot\|_{\mathbf{A}_\sigma^{-1}}$, where $\sigma = 0$ reduces to the case of Euclidean geometry.

Assumption 1 ((G, B)-BGD (Bounded Gradient Dissimilarity))**.** *There exist constants $G \geq 0$ and $B \geq 1$ such that*

$$\frac{1}{N}\sum_{j=1}^{N} \|\nabla f_j(w)\|_*^2 \leq G^2 + B^2\|\nabla f(w)\|_*^2, \ \forall w.$$

This assumption describes the heterogeneity of data clients with G. For **non-iid** data, $G > 0$, while $G = 0$ and $B = 1$ reduce to the **iid** case.

Assumption 2. *$f_1, ..., f_N$ are all β-smooth: for all u and v, $\|\nabla f_j(u) - \nabla f_j(v)\|_* \leq \beta\|u - v\|$.*

Assumption 3. *$f_1, ..., f_N$ are all μ-strongly convex:*

$$f_j(u) \geq f_j(v) + \langle u - v, \nabla f_j(v)\rangle + \frac{\mu}{2}\|u - v\|^2, \quad \textit{for all } u, v.$$

Assumption 4. *$f_1, ..., f_N$ are all convex:*

$$f_j(u) \geq f_j(v) + \langle u - v, \nabla f_j(v)\rangle, \quad \textit{for all } u, v.$$

Assumption 5. *Let $g_j(w)$ be a stochastic mini-batch gradient of client j. The variance of $g_j(w)$ under the dual norm in each device is bounded:*

$$\mathbb{E}\|g_j(w) - \nabla f_j(w)\|_*^2 \leq \varsigma_j^2(\sigma) \quad \textit{for all } j \in [N].$$

We further denote $\varsigma^2(\sigma) = \frac{1}{N}\sum_{j=1}^N \varsigma_j^2(\sigma)$.

Here for $\sigma = 0$, it reduces to the common assumption in federated learning (Karimireddy et al., 2020); for $\sigma > 0$, variance could be significantly reduced as the discussions in Section 2.

Assumption 6. *f_1, ..., f_N are all L-Lipschitz: $\|f_j(u) - f_j(v)\|_* \leq L\|u - v\|$ for all u, v.*

For simplicity, we use $\nu_{\mathcal{L}}$ to represent ν in Theorem 1 as a linear function of the clipping parameter $\mathcal{L}$. Then we have $\nu_{\mathcal{L}} = \mathcal{L}\nu_1$. We use $\tilde{\mathcal{O}}$ to denote asymptotic growth rate up to a logarithmic factor (including $\log K$, $\log S$), while $\mathcal{O}$ up to a constant.

Now we are ready to present the convergence guarantees for strongly-convex, general-convex and non-convex loss scenarios. For the non-convex scenario, the convergence guarantee is for stationary point and the error is measured by the expected norm of the gradient at the approximate stationary point.

Theorem 3 (Convergence Guarantees for DP-Fed-LS)**.** *Assuming the conditions in Theorem 1 hold, with $\log(1/\delta) \geq \varepsilon$ and a proper constant local and global update step sizes η_l and η_g. Let $\mathcal{L} = \eta_l K L$, $\eta_g \geq \sqrt{S}$, and communication round $T = \frac{\varepsilon^2 N^2}{C_0 L^2 S \log(1/\delta)}$, then DP-Fed-LS with uniform subsampling satisfies (ε, δ)-DP and the following error bounds, where the expectation is taken over the randomness in SGD in local client update, client selection and noise injection.*

- *μ **Strongly-Convex:** Under Assumption 1, 2, 3, 5, 6, it holds that*

$$\mathcal{E}(\bar{w}^T) = \mathbb{E}[f(\bar{w}^T)] - f(w^*) \leq \tilde{\mathcal{O}}\left(\frac{(\frac{\varsigma^2(\sigma)}{K} + (1-\tau)G^2 + d_\sigma)L^2 \log(1/\delta)}{\mu\varepsilon^2 N^2}\right),$$

- ***General-Convex:** Under Assumption 1, 2, 4, 5, 6, it holds that*

$$\mathcal{E}(\bar{w}^T) = \mathbb{E}[f(\bar{w}^T)] - f(w^*) \leq \mathcal{O}\left(\frac{\sqrt{(\frac{\varsigma^2(\sigma)}{K} + (1-\tau)G^2 + d_\sigma)D_\sigma L^2 \log(1/\delta)}}{\varepsilon N}\right),$$

- ***Non-Convex:*** *Under Assumption 1, 2, 5, 6, it holds that*

$$\mathcal{E}(\bar{w}^T) = \mathbb{E}\|\nabla f(\bar{w})\|^2_{\boldsymbol{A}_\sigma^{-1}} \leq \mathcal{O}\left(\frac{\sqrt{(\frac{\varsigma^2(\sigma)}{K} + (1-\tau)G^2 + d_\sigma)F_0\beta L^2 \log(1/\delta)}}{\varepsilon N}\right).$$

The effective dimension $d_\sigma = \sum_{i=1}^d \Lambda_i$, *where* $\Lambda_i = \frac{1}{1+2\sigma(1-\cos(2\pi i/d))} \leq 1$ *is the eigenvalue of* $\boldsymbol{A}_\sigma^{-1}$. *For an optimum* w^*, $D_\sigma = \|w^0 - w^*\|^2_{\boldsymbol{A}_\sigma}$, $F_0 = f(w^0) - f(w^*)$ *and* $C_0 = \frac{14}{\lambda}(1 + \frac{1}{1-\lambda})$.

The sketchy proof of Theorem 3 can be found in Section 4.1 with details in Appendix C. For simplicity, we set $\mathcal{L} = \eta_l K L$ to avoid the clipping effect, which can be left for future exploration.

In this theorem, dominant errors are introduced by the variance of stochastic gradients ($\varsigma^2(\sigma)/K$), heterogeneity of **non-iid** data ($(1-\tau)G^2$) and DP (d_σ), in comparison to the initial error. Among the three dominant errors, the variance term $\varsigma^2(\sigma)/K$ will diminish while the number of local iteration K grows large enough ($K \gg 1$). What's more, the heterogeneity term $(1-\tau)G^2$ will be reduced if subsampling ratio τ is high. Particularly, in **iid** ($G = 0$) or full-device participation ($\tau = 1$) setting, this term will vanish. Therefore the error term introduced by DP, of effective dimensionality d_σ, dominates the variance and heterogeneity terms in these scenarios, whose rates in Theorem 3 matches the optimal ones of ERM via SGD with differential privacy in centralized setting (Wang et al., 2020, 2017), as shown in the upper part of Table 1 [1]. In particular when $\sigma = 0$, the bounds above reduce to the standard DP-Fed setting. The benefit of introducing Laplacian smoothing ($\sigma > 0$) lies in the reduction of variance $\varsigma^2(\sigma)$ and the effective dimension $d_\sigma \leq d_0 = d$, although it might increase the initial error D_σ.

The following corollary provides the communication complexity of DP-Fed-LS in Algorithm 1 with uniform subsampling, with tight bounds on the number of communications T to reach an optimization error ϵ. Its proof is deferred to Appendix C.6.

Corollary 1 (Communication Complexity)**.** *Assuming the same conditions in Theorem 3, the communication complexity of DP-Fed-LS with uniform subsampling and fixed noise level* $\nu_\mathcal{L} = \mathcal{L}\nu_1$ *independent to* T *satisfies the following rates to reach an* ϵ*-optimality gap,*

[1] In Table 1, the term $\log(S)$ of DP-Fed-LS comes from the numerator of learning rate $\tilde{\eta}$ in Theorem 4 in Appendix, implicitly involved in $\tilde{\mathcal{O}}$.

- μ ***Strongly-Convex:***

$$T = \tilde{\mathcal{O}}\bigg(\frac{(1+4\sigma)^2\beta B^2}{\mu} + \frac{\varsigma^2(\sigma)}{\mu K S\epsilon} + \frac{d_\sigma L^2\nu_1^2}{\mu S^2\epsilon} + \frac{(1+4\sigma)\sqrt{\beta}G}{\mu\sqrt{\epsilon}} + (1-\tau)\frac{G^2}{\mu\epsilon S}\bigg),$$

- ***General-Convex:***

$$T = \mathcal{O}\bigg(\frac{(1+4\sigma)^2\beta B^2 D_\sigma}{\epsilon} + \frac{\varsigma^2(\sigma)D_\sigma}{KS\epsilon^2} + \frac{d_\sigma D_\sigma L^2\nu_1^2}{S^2\epsilon^2} + \frac{(1+4\sigma)\sqrt{\beta}D_\sigma G}{\epsilon^{3/2}} + (1-\tau)\frac{D_\sigma G^2}{\epsilon^2 S}\bigg),$$

- ***Non-Convex:***

$$T = \mathcal{O}\bigg(\frac{(1+4\sigma)^2\beta B^2 F_0}{\epsilon} + \frac{\varsigma^2(\sigma)\beta F_0}{KS\epsilon^2} + \frac{d_\sigma F_0 L^2\nu_1^2\beta}{S^2\epsilon^2} + \frac{(1+4\sigma)\beta F_0 G}{\epsilon^{3/2}} + (1-\tau)\frac{F_0\beta G^2}{\epsilon^2 S}\bigg).$$

In Corollary 1, we regard that ν_1 is a given constant independent to the communication round T, such that $\nu_{\mathcal{L}} = \mathcal{L}\nu_1$ and $\mathcal{L} = \eta_l KL$. In this case, if $\nu_1 \geq 8/3$ and $\alpha - 1 \leq \frac{\nu_1^2}{6}\ln\frac{1}{\tau\alpha(1+\nu_1^2/4)}$, then (ε, δ)-DP satisfying $T \leq \lambda\varepsilon^2\nu_1^2/14\tau^2\big(\frac{\log(1/\delta)}{1-\lambda} + \varepsilon\big)$ can be achieved for any $\lambda \in (0,1)$ (See Appendix D). Compared with the best known rates in federated average without DP (Karimireddy et al., 2020), the communication complexity in Corollary 1 involves an extra term for the injected noise ν_1 in DP, while other terms match the best known rates, which are tighter than others in literature (Yu et al., 2019; Khaled et al., 2020; Li et al., 2019) with the same (G, B)-BGD assumption, as shown in the lower part of Table 1.

In spite of the reduced variance and effective dimension mentioned above, LS might increase the initial error by a factor of $(1+4\sigma)^2$ and a non-dominant part of the heterogeneity term G by a factor of $(1+4\sigma)$. This is because in the local update, we only apply the general mini-batch SGD, while we apply LS on the global update. In this case, to bound the divergence of local parameters, we need to firstly transform the local update in norm $\|\cdot\|_{\mathbf{A}_\sigma}$ to its dual norm $\|\cdot\|_{\mathbf{A}_\sigma^{-1}}$ by the norm equivalence $\|\cdot\|_{\mathbf{A}_\sigma} \leq \frac{1}{\Lambda_{\min}^2}\|\cdot\|_{\mathbf{A}_\sigma^{-1}}$, where $\Lambda_{\min} \geq \frac{1}{1+4\sigma}$ is the smallest eigenvalue of $\mathbf{A}_\sigma^{-1}$. Therefore, we introduce an extra constant $(1+4\sigma)$ to the non-dominant part of heterogeneity term G. The factor $(1+4\sigma)^2$ in the initial error is a by-product of the learning rate adjusted to the local update divergence.

4.1. Error Decomposition and Sketchy Proof of Theorem 3

To prove Theorem 3, we establish the following **Meta Theorem** summarizing a decomposition of the optimization error into four components caused by initial error, heterogeneous clients data, stochastic gradient variance and differential privacy noise.

Meta Theorem. *There exists constant step size η_l and η_g, Gaussian noise ν, communication round T, and clipping parameter $\mathcal{L}$ such that DP-Fed-LS satisfies*

$$\mathcal{E}(\bar{w}^T) \leq \mathcal{E}_{init}(T) + \mathcal{E}_{hete}(T) + \mathcal{E}_{var}(T) + \mathcal{E}_{dp}(T) \tag{6}$$

where $\mathcal{E}(\bar{w}^T) = \mathbb{E}[f(\bar{w}^T)] - f(w^*)$ for strongly convex and general convex cases while $\mathcal{E}(\bar{w}^T) = \mathbb{E}\|\nabla f(\bar{w})\|_{\boldsymbol{A}_\sigma^{-1}}$ for non-convex case. $\mathcal{E}_{init}$ is the initial error, $\mathcal{E}_{hete}$ is introduced by the heterogeneity of clients' data, $\mathcal{E}_{var}$ accounts for the variance of stochastic gradients, and $\mathcal{E}_{dp}$ is due to privacy noise under Laplacian smoothing.

To see this, the following Theorem 4, 5 and 6 instantiate the Meta Theorem for three scenarios, *i.e.* strongly convex, convex, and non-convex cases of loss functions, respectively, whose detailed proof can be found in Appendix C.

Theorem 4 (μ Strongly-Convex)**.** *Under Assumption 1, 2, 3, 5, 6, $\eta_g \geq 1$, $a_t = (1-\mu\tilde{\eta}K/2)^{-t}$, $\tilde{\eta} \coloneqq \eta_l\eta_g = \min\left\{\frac{2\log(\max(e,\mu^2 T D_\sigma/H_\sigma))}{\mu KT}, \frac{\Lambda_{\min}^2}{8K\beta(1+B^2)}\right\}$, $\mathcal{L} = \eta_l KL$, and $T \geq \frac{1}{\mu\tilde{\eta}K}$, Algorithm 1 with uniform subsampling satisfies*

$$\mathcal{E}_{init} = 3\mu_\sigma \exp(-\mu\tilde{\eta}KT/2)D_\sigma \leq \tilde{\mathcal{O}}\left(\frac{H_\sigma}{\mu T}\right),$$

$$\mathcal{E}_{var} \leq 2\tilde{\eta}(1+\frac{S}{\eta_g^2})\frac{\varsigma^2(\sigma)}{S} \leq \tilde{\mathcal{O}}\left(\frac{(1+\frac{S}{\eta_g^2})\varsigma^2(\sigma)}{\mu KST}\right),$$

$$\mathcal{E}_{dp} \leq \frac{2\tilde{\eta}d_\sigma KL^2\nu_1^2}{S^2} \leq \tilde{\mathcal{O}}\left(\frac{d_\sigma L^2\nu_1^2}{\mu S^2 T}\right),$$

$$\mathcal{E}_{hete} \leq 24(1+4\sigma)^2\tilde{\eta}^2K^2\beta G^2 + 8\tilde{\eta}(1-\tau)\frac{K}{S}G^2 \leq \tilde{\mathcal{O}}\left(\frac{(1+4\sigma)^2\beta G^2}{\mu^2T^2} + \frac{G^2(1-\tau)}{\mu ST}\right),$$

where $H_\sigma = \left(\frac{1}{\eta_g^2 K} + \frac{1}{SK}\right)\varsigma^2(\sigma) + \frac{4}{S}(1-\tau)G^2 + \frac{L^2\nu_1^2 d_\sigma}{S^2}$, $D_\sigma = \|w^0 - w^*\|_{\boldsymbol{A}_\sigma}^2$, and the effective dimension $d_\sigma = \sum_{i=1}^d \Lambda_i$.

Theorem 5 (General-Convex)**.** *Under Assumption 1, 2, 4, 5, 6, $\eta_g \geq 1$, $a_t = 1/(T+1)$, and $\tilde{\eta} \coloneqq \eta_l\eta_g = \min\left\{\sqrt{\frac{D_\sigma}{H_\sigma TK^2}}, \sqrt[3]{\frac{D_\sigma}{Q_1 TK^3}}, \frac{\Lambda_{\min}^2}{8K\beta(1+B^2)}\right\}$, $\mathcal{L} = \eta_l KL$, Algorithm 1 with uniform subsampling satisfies*

$$\mathcal{E}_{init} = \frac{2}{\tilde{\eta}TK}D_\sigma \leq \frac{16\beta(1+B^2)(1+4\sigma)^2 D_\sigma}{T},$$

$$\mathcal{E}_{var} \leq 2\tilde{\eta}(1+\frac{S}{\eta_g^2})\frac{\varsigma^2(\sigma)}{S} \leq 2\sqrt{\frac{(1+\frac{S}{\eta_g^2})\varsigma^2(\sigma)D_\sigma}{KST}},$$

$$\mathcal{E}_{dp} \leq \frac{2\tilde{\eta}d_\sigma KL^2\nu_1^2}{S^2} \leq 2\sqrt{\frac{D_\sigma L^2\nu_1^2 d_\sigma}{S^2T}},$$

$$\mathcal{E}_{hete} \leq \underbrace{24(1+4\sigma)^2\beta G^2}_{Q_1}\tilde{\eta}^2K^2+8\tilde{\eta}(1-\tau)\frac{K}{S}G^2 \leq \sqrt[3]{\frac{24(1+4\sigma)^2\beta D_\sigma^2 G^2}{T^2}}+4\sqrt{\frac{(1-\tau)D_\sigma G^2}{ST}},$$

where $H_\sigma = \big(\frac{1}{\eta_g^2 K}+\frac{1}{SK}\big)\varsigma^2(\sigma)+\frac{4}{S}(1-\tau)G^2+\frac{L^2\nu_1^2 d_\sigma}{S^2}$, $D_\sigma = \|w^0-w^*\|_{\boldsymbol{A}_\sigma}^2$, *and the effective dimension* $d_\sigma = \sum_{i=1}^d \Lambda_i$.

Theorem 6 (Non-Convex). *Under Assumption 1, 2, 5, 6,* $\eta_g \geq 1$, $a_t = 1/(T+1)$, $\mathcal{L} = \eta_l KL$ *and setting* $\tilde{\eta} \coloneqq \eta_l\eta_g = \min\{\sqrt{\frac{F_0}{H_\sigma T\beta K^2}}, \sqrt[3]{\frac{F_0}{Q_2 TK^3}}, \frac{\Lambda_{\min}^2}{8K\beta(1+B^2)}\}$, *Algorithm 1 with uniform subsampling satisfies*

$$\mathcal{E}_{init} = \frac{8}{\tilde{\eta}TK}F_0 \leq \frac{64\beta(1+B^2)(1+4\sigma)^2F_0}{T},$$

$$\mathcal{E}_{var} \leq 4\tilde{\eta}\beta(1+\frac{S}{\eta_g^2})\frac{\varsigma^2(\sigma)}{S} \leq 4\sqrt{\frac{(1+\frac{S}{\eta_g^2})\varsigma^2(\sigma)F_0\beta}{KST}},$$

$$\mathcal{E}_{dp} \leq \frac{4\tilde{\eta}d_\sigma KL^2\beta\nu_1^2}{S^2} \leq 2\sqrt{\frac{F_0L^2\nu_1^2\beta d_\sigma}{S^2T}},$$

$$\mathcal{E}_{hete} \leq \underbrace{32(1+4\sigma)^2\beta^2G^2}_{Q_2}\tilde{\eta}^2K^2+16\tilde{\eta}(1-\tau)\beta\frac{K}{S}G^2 \leq \sqrt[3]{\frac{32(1+4\sigma)^2F_0^2G^2\beta^2}{T^2}}+8\sqrt{\frac{(1-\tau)F_0\beta G^2}{ST}},$$

where $H_\sigma = \big(\frac{1}{\eta_g^2 K}+\frac{1}{SK}\big)\varsigma^2(\sigma)+\frac{4}{S}(1-\tau)G^2+\frac{L^2\nu_1^2 d_\sigma}{S^2}$, $F_0 = f(w^0)-f(w^*)$, *and the effective dimension* $d_\sigma = \sum_{i=1}^d \Lambda_i$.

Finally, Theorem 3 follows from substituting ν_1 in Theorem 4, 5 and 6 by the one in Theorem 1. As we can see, in the non-DP setting, $\mathcal{E}_{dp}$ will reduce to 0 since we can set $\nu_1 = 0$. In this case, the benefit of introducing Laplacian smoothing ($\sigma > 0$) lies in the reduction of variance $\varsigma^2(\sigma)$. If we further take a full gradient descent in each client device, then $\mathcal{E}_{var}$ will becomes zero, too. In this case, an implicit benefit of Laplacian smoothing is that it allows us to take a larger step size with high probability and make training progress in shallow directions effectively,

which is detailed in Section 3 in Osher et al. (2022). In addition, if $G = 0$, which means that different clients have similar optima, then $\mathcal{E}_{hete}$ will also vanish. In this case, the federated learning setting will have $O(\frac{1}{T})$ convergence rate in convex and non-convex setting while linear convergence rate in strongly convex setting as gradient descent. One may notice that if we let $G = 0$ then sampling ratio τ no longer contributes to the errors. Actually, this is because in the initial error term $\mathcal{E}_{init}$, we absorb the τ-related term by $B^2\big(1 - \frac{S}{N}\big)\frac{1}{S} \leq 2B^2$ for simplicity.

5. Experimental Results

In this section, we show that Laplacian smoothing in DP-Fed-LS ($\sigma > 0$) improves the utility of plain DP-Fed ($\sigma = 0$) with varying ε and $\delta = 1/N^{1.1}$ (McMahan et al., 2018b) in (ε, δ)-DP on three benchmark classification tasks. In practice, we will firstly flatten the weights of a layer into a 1-D vector by the natural order in Pytorch, and then apply the Laplacian smoothing layer-wise. Please see Section 5.2 for details about the orders of flattening. The detailed settings, parameter tuning and other results are deferred to Appendix H.

Logistic regression with iid MNIST dataset. We train a differentially private federated logistic regression on the MNIST dataset (LeCun et al., 1998). MNIST is a dataset of 28×28 grayscale images of digit from 0 to 9, containing 60K training samples and 10K testing samples. We split 50K training samples into 1000/500 clients each containing 50/100 samples in an **iid** fashion (McMahan et al., 2017) for uniform/Poisson subsampling. We use 10K training samples for validation.

CNN with iid SVHN dataset. We train a differentially private federated CNN on the extended SVHN dataset (Netzer et al., 2011). SVHN is a dataset of 32×32 colored images of digits from 0 to 9, containing 73,257 training samples and 26,032 testing samples. We enlarge the training set with another 531,131 extended samples and split them into 2,000 clients each containing about 300 samples in an **iid** fashion (McMahan et al., 2017). We also split the testing set by 10K/16K for validation and testing. Our CNN stacks two 5×5 convolutional layers with max-pooling, two fully-connected layers with 384 and 192 units, respectively, and a final softmax output layer (about 3.4M parameters in total) (Papernot et al., 2017). We pretrain the model over the MNIST dataset to speed up the training without losing privacy guarantee.

LSTM with non-iid Shakespeare dataset. We train a differentially private LSTM on the Shakespeare dataset (Caldas et al., 2018; McMahan et al., 2017), which is built from all works of William Shakespeare, where each speaking role is considered as a client, whose local database consists of all her/his lines. This is

a **non-iid** setting. The full dataset contains 1,129 clients and 4,226,158 samples. Each sample consists of 80 successive characters and the task is to predict the next character. In our setting, we remove the clients that own less than 100 samples to stabilize training, which reduces the total client number to 975. We split the training, validation, and testing set chronologically, with fractions of 0.7, 0.1, 0.2. Our LSTM first embeds each input character into a 8-dimensional space, after which two LSTM layers are stacked, each have 256 nodes. The outputs will be then fed into a linear layer, of which the number of output nodes equals the number of distinct characters (Caldas et al., 2018; McMahan et al., 2017).

For demonstration purpose, we apply the privacy budget in Theorem 1 and 2 for the logistic regression. For CNN and LSTM, we apply the moment accountants in (Wang et al., 2019b)[2] and (Mironov et al., 2019)[3] for uniform subsampling and Poisson subsampling, respectively. For moment accountants, we should provide a noise multiplier z to control the noise level. Then we can compute the privacy budget with given communication round and subsampling ratio. For CNN and LSTM, the selected noise multiplier z are 2.4, 2.2, 2.0, 1.8 and 1.4, 1.2, 1.0, 0.8, respectively.

5.1. Improved Test Accuracy under the Same Privacy Budget

From Table 2, 3 and 4, we notice that DP-Fed-LS outperforms DP-Fed in almost all settings. The accuracy are reported based on 5 independent runs. In particular, when ε is small, the improvement of DP-Fed-LS is remarkably large. We show the average training curves over 5 runs in Figure 3, 4 and 5, where we find that DP-Fed-LS converges slower than DP-Fed in both subsampling scenarios. However, DP-Fed-LS will generalize better than DP-Fed at the later stage of the training. This phenomenon further validates our founding in Theorem 4-6, and the discussion after Theorem 3 and Corollary 1 that, Laplacian smoothing will introduce higher initial error and a non-dominate part the heterogeneity, but the model will finally benefit from the reduced effective dimensionality d_σ, which become dominant term in the later stage of the training.

[2] https://github.com/yuxiangw/autodp

[3] https://github.com/tensorflow/privacy/tree/master/tensorflow_privacy/privacy/analysis

	ε	6	7	8	9
Uniform	$\sigma = 0.0$	82.87 ± 0.97	84.67 ± 0.54	84.99 ± 1.01	85.41 ± 0.57
	$\sigma = 1.0$	$\mathbf{84.77 \pm 0.28}$	$\mathbf{85.90 \pm 0.22}$	86.32 ± 0.40	$\mathbf{86.63 \pm 0.80}$
	$\sigma = 2.0$	83.92 ± 0.62	85.43 ± 1.00	85.78 ± 0.49	86.22 ± 0.42
	$\sigma = 3.0$	84.73 ± 0.91	85.62 ± 0.58	$\mathbf{86.40 \pm 0.50}$	86.42 ± 0.54
	ε	6	7	8	9
Poisson	$\sigma = 0.0$	83.94 ± 0.28	85.45 ± 0.24	86.30 ± 0.74	86.53 ± 0.79
	$\sigma = 1.0$	$\mathbf{85.64 \pm 0.52}$	86.51 ± 0.46	86.61 ± 0.67	86.95 ± 0.56
	$\sigma = 2.0$	85.49 ± 0.28	86.34 ± 0.52	$\mathbf{86.79 \pm 0.28}$	$\mathbf{87.23 \pm 0.53}$
	$\sigma = 3.0$	85.52 ± 0.72	$\mathbf{86.51 \pm 0.28}$	86.63 ± 0.70	86.85 ± 0.32

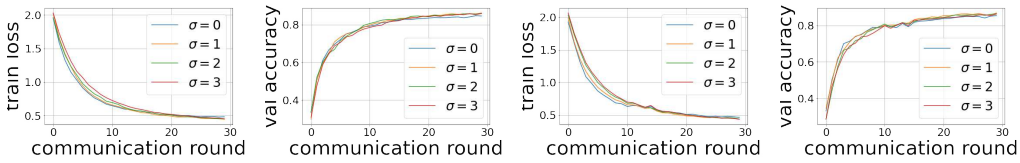
Table 2: Test accuracy of logistic regression on MNIST with DP-Fed ($\sigma = 0$) and DP-Fed-LS ($\sigma = 1, 2, 3$) under different $(\varepsilon, 1/1000^{1.1})$ and $(\varepsilon, 1/500^{1.1})$-DP guarantees for uniform and Poisson subsampling.

	ε	2.83	3.15	3.53	4.05
Uniform	$\sigma = 0.0$	76.14 ± 0.77	78.56 ± 0.55	80.28 ± 0.73	82.05 ± 0.38
	$\sigma = 0.5$	80.67 ± 0.57	$\mathbf{82.18 \pm 0.17}$	82.94 ± 0.35	84.17 ± 0.64
	$\sigma = 1.0$	80.82 ± 0.24	82.17 ± 0.36	$\mathbf{83.19 \pm 0.19}$	$\mathbf{84.40 \pm 0.33}$
	$\sigma = 1.5$	$\mathbf{81.02 \pm 0.47}$	81.38 ± 0.75	82.93 ± 0.23	83.51 ± 0.51
	ε	1.39	1.55	1.74	2.00
Poisson	$\sigma = 0.0$	75.93 ± 0.24	78.23 ± 0.85	79.94 ± 0.88	82.01 ± 0.42
	$\sigma = 0.5$	80.68 ± 0.30	81.86 ± 0.36	82.87 ± 0.39	$\mathbf{83.94 \pm 0.39}$
	$\sigma = 1.0$	$\mathbf{80.82 \pm 0.43}$	$\mathbf{82.10 \pm 0.23}$	$\mathbf{82.89 \pm 0.46}$	83.85 ± 0.37
	$\sigma = 1.5$	80.79 ± 0.34	81.80 ± 0.33	82.83 ± 0.61	83.74 ± 0.23

Table 3: Test accuracy of CNN on SVHN with DP-Fed ($\sigma = 0$) and DP-Fed-LS ($\sigma = 0.5, 1, 1.5$) under different $(\varepsilon, 1/2000^{1.1})$-DP guarantees and subsampling methods.

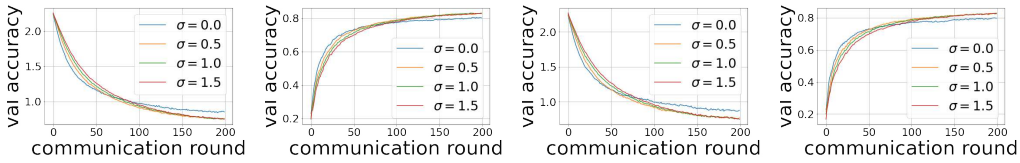
	ε	17.69	22.43	27.25	39.90
Uniform	$\sigma = 0.0$	38.79 ± 0.54	39.05 ± 0.18	41.48 ± 0.45	43.96 ± 0.20
	$\sigma = 0.5$	39.97 ± 0.58	41.44 ± 0.39	43.66 ± 0.69	$\mathbf{45.49 \pm 0.47}$
	$\sigma = 1.0$	40.36 ± 0.39	41.90 ± 0.29	$\mathbf{44.29 \pm 0.34}$	45.35 ± 0.27
	$\sigma = 1.5$	$\mathbf{40.76 \pm 0.48}$	$\mathbf{42.04 \pm 0.39}$	43.68 ± 0.43	44.91 ± 0.32
	ε	8.23	10.41	14.05	20.92
Poisson	$\sigma = 0.0$	38.58 ± 0.42	39.84 ± 0.25	41.49 ± 0.41	43.78 ± 0.42
	$\sigma = 0.5$	39.44 ± 0.47	40.87 ± 0.31	43.70 ± 0.55	$\mathbf{45.24 \pm 0.45}$
	$\sigma = 1.0$	$\mathbf{40.73 \pm 0.34}$	42.20 ± 0.28	$\mathbf{44.06 \pm 0.55}$	45.03 ± 0.11
	$\sigma = 1.5$	40.60 ± 0.34	$\mathbf{42.27 \pm 0.36}$	43.92 ± 0.26	45.06 ± 0.36

Table 4: Test accuracy of LSTM on Shakespeare with DP-Fed ($\sigma = 0$) and DP-Fed-LS ($\sigma = 0.5, 1, 1.5$) under different $(\varepsilon, 1/975^{1.1})$-DP guarantees and subsamplings.



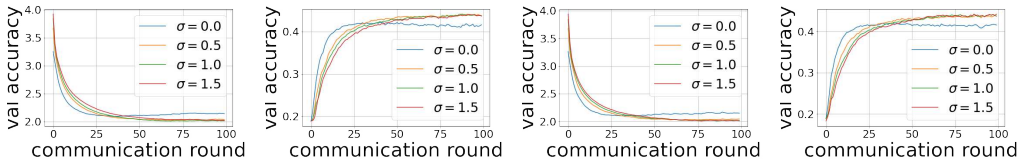
(a) (b) (c) (d)

Figure 3: Training curves of logistic regression on MNIST with DP-Fed ($\sigma = 0$), DP-Fed-LS ($\sigma = 1, 2, 3$). (a) and (b): training loss and validation accuracy with uniform subsampling and $(7, 1/1000^{1.1)}$-DP. (c) and (d): training loss and validation accuracy with Poisson subsampling and $(7, 1/500^{1.1})$-DP.



(a) (b) (c) (d)

Figure 4: Training curves of CNN on SVHN with DP-Fed ($\sigma = 0$), DP-Fed-LS ($\sigma = 0.5, 1, 1.5$). (a) and (b): training loss and validation accuracy with uniform subsampling and $(3.53, 1/2000^{1.1})$-DP. (c) and (d): training loss and validation accuracy with Poisson subsampling and $(1.74, 1/2000^{1.1})$-DP.



(a) (b) (c) (d)

Figure 5: Training curves of LSTM on Shakespeare dataset with DP-Fed ($\sigma = 0$), DP-Fed-LS ($\sigma = 0.5, 1, 1.5$). (a) and (b): training loss and validation accuracy with uniform subsampling and $(27.25, 1/975^{1.1})$-DP. (c) and (d): training loss and validation accuracy with Poisson subsampling, and $(14.05, 1/975^{1.1})$-DP.

5.2. *Stability under Large Noise, Learning Rate and Different Orders of Parameter Flattening*

In Figure 6, we show the training curves where relatively large noise multipliers z are applied with Poisson subsampling and different local learning rates η_l. Here our CNNs are trained for 1 run from scratch without pretraining. When the noise levels are large, the training curves fluctuate a lot. In these extreme cases, DP-Fed-LS outperforms DP-Fed by a large margin with the same learning rate selection. For example, when $z = 3.5$ and $\eta_l = 0.1$ or $\eta_l = 0.125$, validation accuracy of DP-Fed starts to drop at the 25th epoch while DP-Fed-LS with the same learning rate can still converge. Overall speaking, DP-Fed-LS is more stable against large noise levels and the change of local learning rate than DP-Fed.

Before we apply fast Fourier transform, we will firstly flatten the model's parameters into a 1-dimensional vector layer-wise. In this case, the order of the flattening matters. For example, the weight of a convolutional layer is ordered by output channel, input channel, width and height ("OIWH") by default in Pytorch. One can apply another ordering, like "OIHW" or "OHWI" for flattening. In Table 5, we demonstrate that DP-Fed-LS is insensitive to the order of parameter flattening and consistently performs better than DP-Fed. The accuracy is reported based on 5 independent runs.

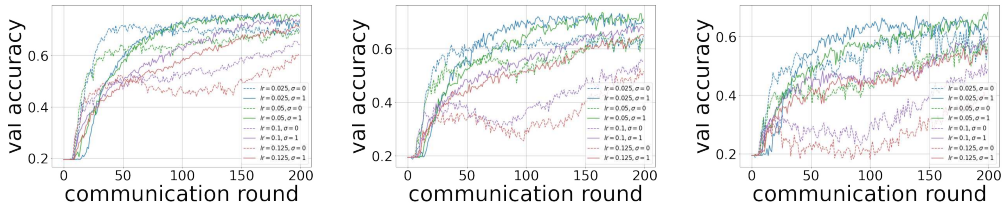


Figure 6: Training curves of CNN on SVHN where large noise levels are applied, with Poisson subsampling and different local learning rates η_l. From left tor right, noise multiplier $z = 3.0, 3.5$ and 4.0. For DP-Fed-LS, we set $\sigma = 1$.

5.3. *Comparison with Other Denoising Estimators*

In Table 6, we show the results of another two adaptive denoising estimators: the James-Stein estimator (JS) and the soft-thresholding estimator (TH), which have been shown to be useful for high dimensional parameter estimation and number release (Balle and Wang, 2018), comparing with other denoising estimators (Barak

	Order	OIWH	OIHW	OWHI	OHWI
Uniform	$\sigma = 0.5$	82.94 ± 0.35	82.91 ± 0.29	83.25 ± 0.45	83.20 ± 0.18
	$\sigma = 1.0$	83.19 ± 0.19	82.92 ± 0.39	83.66 ± 0.40	83.54 ± 0.48
	$\sigma = 1.5$	82.93 ± 0.23	82.50 ± 0.46	83.34 ± 0.50	83.10 ± 0.33
	Order	OIWH	OIHW	OWHI	OHWI
Poisson	$\sigma = 0.5$	82.87 ± 0.39	82.82 ± 0.07	83.13 ± 0.41	82.87 ± 0.61
	$\sigma = 1.0$	82.89 ± 0.46	83.19 ± 0.16	83.11 ± 0.42	83.31 ± 0.64
	$\sigma = 1.5$	82.83 ± 0.61	82.45 ± 0.64	82.82 ± 0.37	83.20 ± 0.37

Table 5: Test accuracy of CNN on SVHN with DP-Fed-LS ($\sigma = 0.5, 1, 1.5$), with different unfolding ordering of convolutional kernel, under $(3.53, 1/2000^{1.1})$-DP guarantees along with uniform subsampling, and $(1.74, 1/2000^{1.1})$-DP guarantees along with Poisson subsampling. "O", "I", "W", and "H" represent for output channel, input channel, width, and height. OIWH is the default setting. The accuracy for pure DP-Fed is 79.89 ± 0.48 and 80.28 ± 0.66 respectively.

et al., 2007; Hay et al., 2009; Williams and McSherry, 2010; Bernstein et al., 2017). As mentioned in (Balle and Wang, 2018), thanks to the fact that we know the parameter ν exactly, both JS and TH estimators are completely free of tuning parameters. As we can see in Table 6, neither of these two estimators performs well in our scenario, compared with Laplacian smoothing in Table 3, indicating that high dimensional sparsity assumption (Balle and Wang, 2018) does not hold here on *federated average of gradients*.

	ε	2.83	3.15	3.53	4.05
Uniform	JS	34.85 ± 1.56	35.92 ± 1.45	37.55 ± 2.00	38.65 ± 1.48
	TH	17.86 ± 1.89	18.22 ± 1.80	18.37 ± 0.81	18.96 ± 1.24
	ε	1.39	1.55	1.74	2.00
Poisson	JS	35.61 ± 1.98	36.25 ± 1.55	37.01 ± 1.02	38.81 ± 2.01
	TH	17.05 ± 2.44	16.55 ± 1.94	17.84 ± 1.69	18.17 ± 1.49

Table 6: Test accuracy of CNN on SVHN James-Stein (JS) and soft-thresholding (TH) estimators (Balle and Wang, 2018) under different $(\varepsilon, 1/2000^{1.1})$-DP guarantees and subsampling methods (the same settings as the ones we used in Table 3).

5.4. Membership Inference Attack

Membership privacy is a simple yet quite practical notion of privacy (Shokri et al., 2017; Yeom et al., 2018; Sablayrolles et al., 2019). Given a model θ and sample z, *membership inference attack* is to infer the probability that a sample z belongs to the training dataset (Sablayrolles et al., 2019). Specifically, a test set $\mathcal{T} = \{(z_i, m_i)\}$, is constructed with samples from both training data $(m_i = 1)$ and

hold-out data ($m_i = 0$), where the prior probability $P(m = 1|\mathcal{T}) = \rho_{\mathcal{T}}$. Then a successful membership attack can increase the excess probability using knowledge of model θ, $P(m(z) = 1|\theta, z \in \mathcal{T}) - \rho_{\mathcal{T}}$. In (Sablayrolles et al., 2019), Sablayrolles et al. define (ε, δ)-membership privacy, and show that (ε, δ)-membership privacy can guarantee an upper bound $P(m(z) = 1|\theta, z \in \mathcal{T}) - \rho_{\mathcal{T}} \leq \frac{e\varepsilon}{4} + \delta$.

To evaluate the membership information leakage of models, *threshold attack* (Yeom et al., 2018) is adopted in our experiment. It is widely used as a metric to evaluate membership privacy (Jayaraman and Evans, 2019; Wu et al., 2020; Yeom et al., 2018). It bases on the intuition that a sample with relatively small loss is more likely to belong to the training set, due to the more or less overfitting of ML models. Specifically, the test set $\mathcal{T}$ consists of both training and hold-out data of equal size (thus $\rho_{\mathcal{T}} = 0.5$). Given a sample $z = (x, y)$ and a model $M_w(x)$, we calculate the loss $\ell(y, M_w(x))$. Then we select a threshold t: if $\ell \leq t$, we regard this sample in the training set; otherwise, it belongs to the hold-out set. As the threshold varies over all possible values in $(0, \mathcal{U})$, where $\mathcal{U}$ is the upper bound for ℓ, the area under ROC curve (AUC) is used to measure the information leakage. In perfectly-private situation, the AUC should be 0.5, indicating that the adversary could not infer whether a given sample belongs to the training set or not. The larger the AUC, the more membership information leaks.

Improved membership privacy. We follow the CNN setup here while we only split 64K data into 500 clients and set $\tau = 0.2$ for training (Jayaraman and Evans, 2019; Shokri et al., 2017; Yeom et al., 2018). Our test set $\mathcal{T}$ for membership inference attack includes 10K training data and 10K testing data of SVHN. In Figure 7, we show the AUC values of threshold attack against different models. We observe that Non-DP model actually suffers high risk of membership leakage. In addition, applying DP can significantly lower the risk. Comparing with DP-Fed, DP-Fed-LS may even further improve the membership privacy.

6. Conclusion

In this paper, based on the observation that in federated learning the average of gradients is often sparse or smooth under a cyclic graph Fourier basis, Laplacian smoothing is introduced for the variance reduction of the noisy *federated average of gradients* to improve the generalization accuracy with the same differential privacy guarantee. Privacy bounds in closed-form are given under uniform or Poisson subsampling mechanisms. Optimization error bound is decomposed into four components caused by heterogeneous data distribution of clients, stochastic gradient variance, differential privacy, and a non-dominant initialization, which sheds

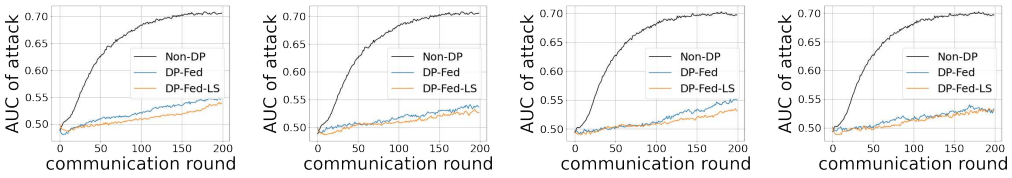


Figure 7: AUC of threshold attack of different model on SVHN. (a) and (b): uniform subsampling with noise multiplier $z = 1.8$ and $z = 2.2$ for DP models. (c) and (d): Poisson subsampling with $z = 1.8$ and $z = 2.2$. The larger the AUC, the more membership information leakages. For DP-Fed-LS, the LS parameter $\sigma = 1$.

light on the theoretical understanding of the effectiveness of Laplacian smoothing. Experimental results show that DP-Fed-LS outperforms DP-Fed in both **iid** and **non-iid** settings, regarding accuracy and membership privacy, demonstrating its potential in practical applications.

Acknowledgements

Zhicong Liang was supported by Hong Kong PhD Fellowship Scheme by the Hong Kong Research Grants Council (HKRGC). The research of Bao Wang was sponsored by NSF grants DMS-2152762, DMS-2208361, DMS-2219956, and the DOE grant DE-SC0023490. Quanquan Gu was supported by NSF SaTC-1717950. The research of Stanley Osher was supported in part by AFOSR MURI FA9550-18-502, ONR N00014-18-1-2527, N00014-18-20-1-2093, and N00014-20-1-2787. The research of Yuan Yao was supported in part by National Natural Science Foundation of China / Research Grants Council Joint Research Scheme Grant HKUST635/20, Hong Kong Research Grant Council (HKRGC) Grant 16308321, 16303817, ITF UIM/390, as well as awards from Tencent AI Lab, Si Family Foundation, and Microsoft Research-Asia. This research made use of the computing resources of the X-GPU cluster supported by the Hong Kong Research Grant Council Collaborative Research Fund: C6021-19EF.

Appendix A. Proof of Theorem 1

We firstly introduce the notation of ℓ_2-sensitivity and some lemmas for future reference.

Definition 3 (ℓ_2-Sensitivity)**.** *For any given function $f(\cdot)$, the ℓ_2-sensitivity of f is defined by*

$$\Delta(f) = \max_{\|D-D'\|_1=1} \|f(D) - f(D')\|_2,$$

where $\|D - D'\|_1 = 1$ means the data sets D and D' differ in only one entry.

Lemma 2 (Composition Theorem of RDP)**.** *(Mironov, 2017) If k randomized mechanisms $\mathcal{M}_i : \mathcal{D} \to \mathbb{R}^d$, for $i \in [k]$, satisfy (α, ρ_i)-RDP, then their composition denoted as $\big(\mathcal{M}_1(D), \ldots, \mathcal{M}_k(D)\big)$ satisfies $(\alpha, \sum_{i=1}^k \rho_i)$-RDP. Moreover, the input of the i-th mechanism can be based on outputs of the previous $(i-1)$ mechanisms.*

Lemma 3 (RDP for Uniform Subsampling)**.** *Gaussian mechanism $\mathcal{M} = f(D) + \mathcal{N}(0, \nu^2)$ applied on a subset of samples drawn uniformly without replacement with probability τ satisfies $(\alpha, 3.5\tau^2\alpha/\nu^2)$-RDP given $\nu^2 \geq \frac{2}{3}$ and $\alpha - 1 \leq \frac{2}{3}\nu^2 \ln\big(1/\alpha\tau(1+\nu^2)\big)$, where the sensitivity of f is 1.*

Remark 1. *Comparing with the result $(\alpha, 5\tau^2\alpha/\nu^2)$ in (Wang et al., 2019a), and $(\alpha, 6\tau^2\alpha/\nu^2)$ in (Bun et al., 2018), Lemma 3 provides a tighter bound while relaxing their requirement on ν^2 that $\nu^2 \geq 1.5$ and $\nu^2 \geq 5$ respectively. The proof of Lemma 3 is deferred to Appendix A.1.*

Here we are going to provide privacy upper bound for FedAvg (DP-Fed). We drop the superscript K from $w_j^{t,K}$ for simplicity, then

$$w^{t+1} = w^t + \frac{\eta_g}{S}\left(\sum_{j\in M_t} w_j^t - S \cdot w^t + \mathbf{n}\right), \tag{A.1}$$

Similarly, for the one with Laplacian Smoothing (DP-Fed-LS), it becomes

$$\tilde{w}^{t+1} = \tilde{w}^t + \frac{\eta_g}{S} A_\sigma^{-1}\left(\sum_{j\in M_t} \tilde{w}_j^t - S \cdot \tilde{w}^t + \mathbf{n}\right), \tag{A.2}$$

where $\mathbf{n} \sim \mathcal{N}(0, \nu^2 I)$, and w_j^t is the updated model from client j, based on the previous global model w^t.

Proof. In the following, we will show that the Gaussian noise $\mathcal{N}(0, \nu^2)$ in Eq. (A.1) for each coordinate of $\mathbf{n}$, the output of DP-Fed, $\mathbf{w}$, after T iteration is (ε,δ)-DP. We drop the superscript K from $w_j^{t,K}$ for simplicity.

Let us consider the mechanism $\mathcal{M}_t = \frac{1}{S}\sum_{j=1}^{K} w_j^t - w^t + \frac{1}{S}\mathbf{n}$ with query $\mathbf{q}_t = \frac{1}{S}\sum_{j=1}^{N} w_j^t - w^t$ and its subsampled version $\hat{\mathcal{M}}_t = \frac{1}{S}\sum_{j\in M_t} w_j^t - w^t + \frac{1}{S}\mathbf{n}$. Define the query noise $\mathbf{n}_q = \mathbf{n}/S$ whose variance is $\nu_q^2 := \nu^2/S^2$. We will firstly evaluate the sensitivity of w_j^t. For each local iteration

$$w_j^t \leftarrow w_j^t - \eta_l g(w_j^t)$$

$$w_j^t \leftarrow w^t + \text{clip}\big(w_j^t - w^t, \mathcal{L}\big),$$

where $\text{clip}(v, \mathcal{L}) \leftarrow v/\max(1, \|v\|_2/\mathcal{L})$. All the local output $\Delta_j^t \leftarrow w_j^t - w^t$ will be inside the l_2-norm ball centering around w^t with radius $\mathcal{L}$. We have l_2-sensitivity of $\mathbf{q}_t$ as $\Delta(\mathbf{q}) = \|w_j^t - w_j^{t'}\|_2/S \leq 2\mathcal{L}/S$.

According to (Mironov, 2017), if we add noise with variance,

$$\nu^2 = S^2\nu_q^2 = \frac{14\tau^2\alpha T\mathcal{L}^2}{\lambda\varepsilon}, \tag{A.3}$$

the mechanism $\mathcal{M}_t$ will satisfy $(\alpha, \alpha\Delta^2(\mathbf{q})/(2\nu_q^2)) = (\alpha, \lambda\varepsilon/7\tau^2 T)$-RDP. By Lemma 3, $\hat{\mathcal{M}}_t$ will satisfy (α,$\lambda\varepsilon/T$)-RDP provided that $\nu_q^2/\Delta^2(\mathbf{q}) = \nu^2/(S^2\Delta^2(\mathbf{q})) \geq \frac{2}{3}$ and $\alpha - 1 \leq \frac{2\nu_q^2}{3\Delta^2(\mathbf{q})}\log\big(1/\tau\alpha(1 + \nu_q^2/\Delta^2(\mathbf{q}))\big)$. By post-processing theorem, $\tilde{\mathcal{M}}_t = A_\sigma^{-1}\big(\frac{1}{S}\sum_{j\in M_t} w_j^t - w^t + \frac{1}{S}\mathbf{n}\big)$ will also satisfy $(\alpha, \lambda\varepsilon/T)$-RDP.

Let $\alpha = \log(1/\delta)/((1-\lambda)\varepsilon) + 1$, we obtain that $\hat{\mathcal{M}}_t$ (and $\tilde{\mathcal{M}}_t$) satisfies $(\log(1/\delta)/(1-\lambda)\varepsilon + 1, \lambda\varepsilon/T)$-RDP as long as the following inequalities hold

$$\frac{\nu_q^2}{\Delta^2(\mathbf{q})} = \frac{\nu^2}{S^2\Delta^2(\mathbf{q})} = \frac{\nu^2}{4\mathcal{L}^2} \geq \frac{2}{3} \tag{A.4}$$

and

$$\alpha - 1 \leq \frac{\nu^2}{6\mathcal{L}^2}\ln\frac{1}{\tau\alpha(1 + \nu^2/4\mathcal{L}^2)}. \tag{A.5}$$

Therefore, according to Lemma 2, we have w^t (and $\tilde{w}^t$) satisfies $(\log(1/\delta)/(1-\lambda)\varepsilon + 1, \lambda t\varepsilon/T)$-RDP. Finally, by Lemma 1, we have w^t (and $\tilde{w}^t$) satisfies $(\lambda t\varepsilon/T + (1-\lambda)\varepsilon, \delta)$-DP. Thus, the output of DP-Fed (and DP-Fed-LS), w (and $\tilde{w}$), is (ε,δ)-DP. □

Appendix A.1. Proof of Lemma 3

Proof. This proof is inspired by Lemma 3.7 of (Wang et al., 2019a), while we relax their requirement and get a tighter bound. According to Theorem 9 in (Wang et al., 2019b), Gaussian mechanism applied on a subset of size $S = \tau \cdot N$, whose samples are drawn uniformly satisfies (α, ρ')-RDP, where

$$\rho'(\alpha) \leq \frac{1}{\alpha - 1} \log \left(1 + \tau^2 \binom{\alpha}{2} \min \left\{ 4(e^{\rho(2)} - 1), 2e^{\rho(2)} \right\} + \sum_{j=3}^{\alpha} \tau^j \binom{\alpha}{j} 2e^{(j-1)\rho(j)} \right)$$

where $\rho(j) = j/2\nu^2$. As mentioned in (Wang et al., 2019b), the dominant part in the summation on the right hand side arises from the term $\min \left\{ 4(e^{\rho(2)} - 1), 2e^{\rho(2)} \right\}$ when ν^2 is relatively large. We will bound this term as a whole instead of bounding it firstly by $4(e^{\rho(2)} - 1)$ (Wang et al., 2019a). For $\nu^2 \geq \frac{2}{3}$, we have

$$\min \left\{ 4(e^{\rho(2)} - 1), 2e^{\rho(2)} \right\} = \min \left\{ 4(e^{1/\nu^2} - 1), 2e^{1/\nu^2} \right\} \leq 6/\nu^2. \tag{A.6}$$

For the term summing from $j = 3$ to α, we have

$$\begin{aligned}
\sum_{j=3}^{\alpha} \tau^j \binom{\alpha}{j} 2e^{(j-1)\rho(j)} &= \sum_{j=3}^{\alpha} \tau^j \binom{\alpha}{j} 2e^{\frac{(j-1)j}{2\nu^2}} \leq \sum_{j=3}^{\alpha} \tau^j \frac{\alpha^j}{j!} 2e^{\frac{(j-1)j}{2\nu^2}} \\
&\leq \sum_{j=3}^{\alpha} \tau^j \frac{\alpha^j}{3!} 2e^{\frac{(\alpha-1)j}{2\nu^2}} = \tau^2 \frac{\alpha^2}{3} \sum_{j=3}^{\alpha} \tau^{j-2} \alpha^{j-2} e^{\frac{(\alpha-1)j}{2\nu^2}} \\
&\leq \tau^2 \binom{\alpha}{2} \sum_{j=3}^{\alpha} \tau^{j-2} \alpha^{j-2} e^{\frac{(\alpha-1)j}{2\nu^2}} \\
&\leq \tau^2 \binom{\alpha}{2} \frac{\tau \alpha e^{\frac{3(\alpha-1)}{2\nu^2}}}{1 - \tau \alpha e^{\frac{\alpha-1}{2\nu^2}}} \leq \tau^2 \binom{\alpha}{2} \frac{\tau \alpha e^{\frac{3(\alpha-1)}{2\nu^2}}}{1 - \tau \alpha e^{\frac{3(\alpha-1)}{2\nu^2}}},
\end{aligned} \tag{A.7}$$

where the first inequality follows from the fact that $\binom{\alpha}{j} \leq \frac{\alpha^j}{j!}$, and the last inequality follows from the condition that $\tau \alpha \exp(\alpha - 1)/(2\nu^2) < 1$. In this case, given that

$$\alpha - 1 \leq \frac{2}{3} \nu^2 \ln \frac{1}{\tau \alpha (1 + \nu^2)}, \tag{A.8}$$

we have

$$\sum_{j=3}^{\alpha} \tau^j \binom{\alpha}{j} 2e^{(j-1)\rho(j)} \leq \tau^2 \binom{\alpha}{2} \frac{1}{\nu^2} \tag{A.9}$$

Combining the results in Eq. (A.6) and Eq. (A.9), we have

$$\rho'(\alpha) \le \frac{1}{\alpha-1}\log\left(1+\binom{\alpha}{2}\frac{6\tau^2}{\nu^2}+\binom{\alpha}{2}\frac{\tau^2}{\nu^2}\right)$$
$$\le \frac{1}{\alpha-1}\tau^2\binom{\alpha}{2}\frac{7}{\nu^2} = 3.5\alpha\tau^2/\nu^2.$$

Condition $\tau\alpha\exp(\alpha-1)/(2\nu^2) < 1$ directly follows from Eq.(A.8). □

Appendix B. Proof of Theorem 2

Lemma 4 (RDP for Poisson Subsampling)**.** *Gaussian mechanism $\mathcal{M} = f(D) + \mathcal{N}(0,\nu^2)$ applied to a subset that includes each data point independently with probability τ satisfies $(\alpha, 2\tau^2\alpha/\nu^2)$-RDP given $\nu^2 \ge \frac{5}{9}$ and $\alpha - 1 \le \frac{2}{3}\nu^2\log\big(1/\alpha\tau(1+\nu^2)\big)$, where the sensitivity of f is 1.*

Remark 2. *The bound in Lemma 4 matches the bound of $(\alpha, 2\alpha\tau^2/\nu^2)$-DP in (Mironov et al., 2019). However, we relax the requirement that $\nu \ge 4$ used in (Mironov et al., 2019), and simplify the multiple requirements over α that $1 < \alpha \le \frac{\nu^2 C}{2} - 2\ln\nu$ and $\alpha \le \frac{\nu^2C^2/2 - \ln 5 - 2\ln\nu}{C+\ln(\tau\alpha)+1/(2\nu^2)}$, where $C = \ln\big(1+\frac{1}{\tau(\alpha-1)}\big)$, to only one requirement. This makes our closed-form privacy bound in Theorem 2 below more concise and easier to implement. The proof of Lemma 4 is deferred to Section Appendix B.1.*

Proof. The proof is identical to proof of Theorem 1 except that we use Lemma 4 instead of Lemma 3. According to the definition of Poisson subsampling, we have l_2-sensitivity of $\mathbf{q}_t$ as $\Delta(\mathbf{q}) \le \|w_j^{t'}\|_2/S \le \mathcal{L}/S$. We start from the Eq. (A.3) in the proof of Theorem 1. If we add noise with variance

$$\nu^2 = S^2\nu_q^2 = \frac{2\tau^2\alpha T\mathcal{L}^2}{\lambda\varepsilon}, \tag{B.1}$$

the mechanism $\mathcal{M}_t$ will satisfy $(\alpha, \alpha\Delta^2(\mathbf{q})/(2\nu_q^2)) = (\alpha, \frac{\lambda\varepsilon}{4\tau^2T})$-RDP. According to Lemma 4, $\hat{\mathcal{M}}_t$ will satisfy $(\alpha,\lambda\varepsilon/T)$-RDP provided that

$$\frac{\nu_q^2}{\Delta^2(\mathbf{q})} = \frac{\nu^2}{S^2\Delta^2(\mathbf{q})} = \frac{\nu^2}{\mathcal{L}^2} \ge \frac{5}{9}, \tag{B.2}$$

and

$$\alpha - 1 \le \frac{2\nu^2}{3\mathcal{L}^2}\ln\frac{1}{\tau\alpha(1+\nu^2/\mathcal{L}^2)}. \tag{B.3}$$

By post-processing theorem, $\tilde{\mathcal{M}}_t = A_\sigma^{-1}\big(\frac{1}{S}\sum_{j\in M_t} w_j^t - w^t + \frac{1}{S}\mathbf{n}\big)$ will also satisfy $(\alpha, \lambda\varepsilon/T)$-RDP. Let $\alpha = \log(1/\delta)/((1-\lambda)\varepsilon) + 1$, we obtain that $\hat{\mathcal{M}}_t$ (and $\tilde{\mathcal{M}}_t$) satisfies $(\log(1/\delta)/(1-\lambda)\varepsilon + 1, \lambda\varepsilon/T)$-RDP. Therefore, according to Lemma 2, we have w^t (and $\tilde{w}^t$) satisfies $(\log(1/\delta)/(1-\lambda)\varepsilon + 1, \lambda t\varepsilon/T)$-RDP. Finally, by Lemma 1, we have w^t (and $\tilde{w}^t$) satisfies $(\lambda t\varepsilon/T + (1-\lambda)\varepsilon, \delta)$-DP. Thus, the output of DP-Fed (and DP-Fed-LS), w (and $\tilde{w}$), is (ε,δ)-DP. □

Appendix B.1. Proof of Lemma 4

Proof. According to (Mironov et al., 2019; Zhu and Wang, 2019), Gaussian mechanism applied on a subset where samples are included into the subset with probability ratio τ independently satisfies (α, ρ')-RDP, where

$$\rho'(\alpha) \le \frac{1}{\alpha-1}\log\left((\alpha\tau-\tau+1)(1-\tau)^{\alpha-1} + \binom{\alpha}{2}(1-\tau)^{\alpha-2}\tau^2 e^{\rho(2)} + \sum_{j=3}^{\alpha}\binom{\alpha}{j}(1-\tau)^{\alpha-j}\tau^j e^{(j-1)\rho(j)}\right)$$

where $\rho(j) = j/2\nu^2$.

We notice that, when σ is relatively large, the sum in right-hand side will be dominated by the first two terms. For the first term, we have

$$(\alpha\tau-\tau+1)(1-\tau)^{\alpha-1} \le \frac{\alpha\tau-\tau+1}{1+(\alpha-1)\tau} = 1, \tag{B.4}$$

where the first inequality follows from the inequality that

$$(1+x)^n \le \frac{1}{1-nx} \text{ for } x \in [-1, 0], n \in \mathbb{N}.$$

For the second term, we have

$$\tau^2\binom{\alpha}{2}(1-\tau)^{\alpha-2}e^{\frac{1}{\nu^2}} \le \tau^2\binom{\alpha}{2}e^{\frac{1}{\nu^2}} \le \tau^2\binom{\alpha}{2}\frac{7}{2\nu^2} \tag{B.5}$$

given that $\nu^2 \ge \frac{5}{9}$. The summation from $j = 3$ to α follows Eq. (A.9) given that

$$\alpha - 1 \le \frac{2}{3}\nu^2 \ln\frac{1}{\tau\alpha(1+\nu^2)}. \tag{B.6}$$

Combining Eq. (B.4), (B.5) and (A.9), we have

$$\rho'(\alpha) \leq \frac{1}{\alpha-1}\log\left(1+\tau^2\binom{\alpha}{2}\frac{7}{2\nu^2}+\tau^2\binom{\alpha}{2}\frac{1}{2\nu^2}\right) \leq \tau^2\alpha\frac{4}{2\nu^2} = 2\alpha\tau^2/\nu^2. \tag{B.7}$$

□

Appendix C. Proof of Theorem 3

Theorem 3 follows from substituting ν_1 in Theorem 4, 5 and 6 by the one in Theorem 1. For completeness, we wrap it into a corollary below.

Corollary 2. *Assume that* $\log(1/\delta) \geq \epsilon$, $\eta_g \geq \sqrt{S}$, *the conditions on* a_t, $\mathcal{L}$ *and* $\tilde{\eta}$ *in Theorem 4, 5 and 6, as well as the assumptions in Theorem 1. Algorithm 1 with uniform subsampling satisfies* (ε, δ)*-DP and the following optimization error bounds.*

- μ ***Strongly-Convex:*** *Select* $T = \frac{\varepsilon^2 N^2}{C_0 L^2 S \log(1/\delta)}$ *with* $T \geq \frac{1}{\mu_\sigma \tilde{\eta} K}$ *where* $\tilde{\eta}$ *follows from Theorem 4. Then*

$$\mathbb{E}[f(\bar{w}^T)] - f(w^*) \leq \tilde{\mathcal{O}}\left(\frac{(\varsigma^2(\sigma)/K + (1-\tau)G^2 + d_\sigma)L^2\log(1/\delta)}{\mu\varepsilon^2N^2}\right).$$

- ***General-Convex:*** *Set* $T = \frac{\varepsilon^2 N^2}{C_0 L^2 S \log(1/\delta)}$. *Then*

$$\mathbb{E}[f(\bar{w}^T)] - f(w^*) \leq \frac{\sqrt{(\varsigma^2(\sigma)/K + 4(1-\tau)G^2 + d_\sigma)D_\sigma L^2\log(1/\delta)}}{\varepsilon N}.$$

- ***Non-Convex:*** *Set* $T = \frac{\varepsilon^2 N^2}{C_0 L^2 S \log(1/\delta)}$. *Then*

$$\mathbb{E}\|\nabla f(\bar{w}^T)\|^2_{\boldsymbol{A}_\sigma^{-1}} \leq \frac{\sqrt{(\varsigma^2(\sigma)/K + 4(1-\tau)G^2 + d_\sigma)F_0\beta L^2\log(1/\delta)}}{\varepsilon N}.$$

The proof of Corollary 2 is deferred to Appendix C.7.

Appendix C.1. Proof of Lemmas

We firstly provide some useful Lemmas.

Lemma 5 (Noise reduction of Laplacian smoothing)**.** *Consider Gaussian noise $\boldsymbol{n}(\mathcal{L}) \sim \mathcal{N}(0, \nu_{\mathcal{L}}^2 \boldsymbol{I})$, where $\nu_{\mathcal{L}}$ is the noise level scaled by $\mathcal{L}$, i.e. $\nu_{\mathcal{L}} = \mathcal{L}\nu_1$. We have*

$$\mathbb{E}\|\boldsymbol{n}(\mathcal{L})\|_{\boldsymbol{A}_\sigma^{-1}}^2 \leq \mathcal{L}^2 \nu_1^2 d_\sigma$$

where $d_\sigma \coloneqq d\zeta_\sigma$ and $\zeta_\sigma \coloneqq \frac{1}{d}\sum_{i=i}^{d} \frac{1}{1+2\sigma-2\sigma\cos(2\pi i/d)}$.

Proof of Lemma 5. The proof is inspired by the proof of Lemma 4 in (Wang et al., 2020). Let the eigenvalue decomposition of $\mathbf{A}_\sigma^{-1}$ be $\mathbf{A}_\sigma^{-1} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where $\mathbf{\Lambda}$ is a diagonal matrix with $\Lambda_i = \frac{1}{1-2\sigma-2\sigma\cos(2\pi i/d)}$, we have

$$\begin{aligned}
\mathbb{E}\|\mathbf{n}(\mathcal{L})\|_{\mathbf{A}_\sigma^{-1}}^2 &= \mathbb{E}[\mathrm{Tr}(\mathbf{n}^\top \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top \mathbf{n})] \\
&= \mathrm{Tr}(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top \mathbb{E}[\mathbf{n}\mathbf{n}^\top]) \\
&= \nu_{\mathcal{L}}^2 \mathrm{Tr}(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top) \\
&= \nu_{\mathcal{L}}^2 \sum_{i=i}^{d} \frac{1}{1+2\sigma-2\sigma\cos(2\pi i/d)} \\
&= \mathcal{L}^2 \nu_1^2 d_\sigma
\end{aligned}$$

where $\zeta_\sigma = \frac{1}{d}\sum_{i=1}^{d} \Lambda_i$. □

Lemma 6 (Bounding the divergence of local parameters)**.** *Following convexity, Assumption 1, 2, 5, we have*

$$\begin{aligned}
&\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{K} \mathbb{E}\|w^t - w_j^{t,i}\|_{\boldsymbol{A}_\sigma}^2 \\
&\leq \frac{1}{\Lambda_{\min}^2}\left(4K^3\eta_l^2 G^2 + 4K^3\eta_l^2 B^2 \|\nabla f(w^t)\|_{\boldsymbol{A}_\sigma^{-1}}^2 + 2K^2\eta_l^2\varsigma^2(\sigma)\right) \\
&\leq \frac{1}{\Lambda_{\min}^2}\left(4K^3\eta_l^2 G^2 + 8K^3\eta_l^2 B^2 \beta\big(f(w^t) - f(w^*)\big) + 2K^2\eta_l^2\varsigma^2(\sigma)\right)
\end{aligned}$$

where $w_j^{t,i}$ denote the model of client j in i-th iteration of the t-th communication round.

Proof of Lemma 6. The proof of is inspired by Lemma 8 in (Karimireddy et al., 2020), while we consider the $\mathbf{A}_\sigma$ norm. Recall that the local update made on client j is $w_j^{t,i} = w_j^{t,i-1} + \eta_l \nabla f_j(w_j^{t,i-1}, x_j^{i-1})$. When $i = 0$, the $w_j^{t,i}$ will just equal w^t. For $i \geq 1$, we have:

$$
\begin{aligned}
&\mathbb{E}\|w_j^{t,i} - w^t\|_{\mathbf{A}_\sigma}^2 \\
&\quad = \mathbb{E}\|w_j^{t,i-1} - w^t - \eta_l g_j(w_j^{t,i-1})\|_{\mathbf{A}_\sigma}^2 \\
&\quad = \mathbb{E}\|w_j^{t,i-1} - w^t - \eta_l \nabla f_j(w_j^{t,i-1})\|_{\mathbf{A}_\sigma}^2 + \eta_l^2 \mathbb{E}\|g_j(w_j^{t,i-1}) - \nabla f_j(w_j^{t,i-1})\|_{\mathbf{A}_\sigma}^2 \\
&\quad \leq \mathbb{E}\|w_j^{t,i-1} - w^t - \eta_l \nabla f_j(w_j^{t,i-1})\|_{\mathbf{A}_\sigma}^2 + \frac{\eta_l^2}{\Lambda_{\min}^2} \mathbb{E}\|g_j(w_j^{t,i-1}) - \nabla f_j(w_j^{t,i-1})\|_{\mathbf{A}_\sigma^{-1}}^2 \\
&\quad \leq \mathbb{E}\|w_j^{t,i-1} - w^t - \eta_l \nabla f_j(w_j^{t,i-1})\|_{\mathbf{A}_\sigma}^2 + \frac{\eta_l^2}{\Lambda_{\min}^2} \varsigma_j^2(\sigma) \\
&\quad \leq \left(1 - \frac{1}{K-1}\right) \mathbb{E}\|w_j^{t,i-1} - w^t\|_{\mathbf{A}_\sigma}^2 + K\eta_l^2 \|\nabla f_j(w_j^{t,i-1})\|_{\mathbf{A}_\sigma}^2 + \frac{\eta_l^2}{\Lambda_{\min}^2} \varsigma_j^2(\sigma) \\
&\quad \leq \left(1 - \frac{1}{K-1}\right) \mathbb{E}\|w_j^{t,i-1} - w^t\|_{\mathbf{A}_\sigma}^2 + \frac{2K\eta_l^2}{\Lambda_{\min}^2} \eta_l^2 \mathbb{E}\|\nabla f_j(w_j^{t,i-1}) - \nabla f_j(w^t)\|_{\mathbf{A}_\sigma^{-1}}^2 \\
&\qquad + \frac{2K\eta_l^2}{\Lambda_{\min}^2} \|\nabla f_j(w^t)\|_{\mathbf{A}_\sigma^{-1}}^2 + \frac{\eta_l^2}{\Lambda_{\min}^2} \varsigma_j^2(\sigma) \\
&\quad \leq \left(1 - \frac{1}{K-1} + \frac{2K\eta_l^2\beta^2}{\Lambda_{\min}^2}\right) \mathbb{E}\|w_j^{t,i-1} - w^t\|_{\mathbf{A}_\sigma}^2 + \frac{2K\eta_l^2}{\Lambda_{\min}^2} \|\nabla f_j(w^t)\|_{\mathbf{A}_\sigma^{-1}}^2 + \frac{\eta_l^2}{\Lambda_{\min}^2} \varsigma_j^2(\sigma) \\
&\quad \leq \left(1 - \frac{1}{2(K-1)}\right) \mathbb{E}\|w_j^{t,i-1} - w^t\|_{\mathbf{A}_\sigma}^2 + \frac{2K\eta_l^2}{\Lambda_{\min}^2} \|\nabla f_j(w^t)\|_{\mathbf{A}_\sigma^{-1}}^2 + \frac{\eta_l^2}{\Lambda_{\min}^2} \varsigma_j^2(\sigma)
\end{aligned}
$$

where the last inequality comes from the assumption that $\eta_l \leq \frac{\Lambda_{\min}}{2K\beta}$. Unrolling the recursion above, we have

$$
\begin{aligned}
\mathbb{E}\|w_j^{t,i} - w^t\|_{\mathbf{A}_\sigma}^2 &\leq \frac{1}{\Lambda_{\min}^2} \sum_{k=0}^{i} \left(2K\eta_l^2 \|\nabla f_j(w^t)\|_{\mathbf{A}_\sigma^{-1}}^2 + \eta_l^2 \varsigma_j^2(\sigma)\right) \left(1 - \frac{1}{2(K-1)}\right)^k \\
&\leq \frac{2K}{\Lambda_{\min}^2} \left(2K\eta_l^2 \|\nabla f_j(w^t)\|_{\mathbf{A}_\sigma^{-1}}^2 + \eta_l^2 \varsigma_j^2(\sigma)\right)
\end{aligned}
$$

where the last step is due to

$$
\begin{aligned}
\sum_{k=0}^{i} \left(1 - \frac{1}{2(K-1)}\right)^k &= \frac{1 - \left(1 - \frac{1}{2(K-1)}\right)^{i+1}}{1 - \left(1 - \frac{1}{2(K-1)}\right)} \\
&\leq \frac{1}{1 - \left(1 - \frac{1}{2(K-1)}\right)} \leq 2(K-1) \leq 2K.
\end{aligned}
$$

Taking average over i and j, and considering Assumption 1, we have

$$\begin{aligned}
&\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\mathbb{E}\|w^t - w_j^{t,i}\|_{\mathbf{A}_\sigma}^2 \\
&\leq \frac{1}{\Lambda_{\min}^2}\left(\frac{1}{N}\sum_{j=1}^{N}4K^3\eta_l^2\|\nabla f_j(w^t)\|_{\mathbf{A}_\sigma^{-1}}^2 + 2K^2\eta_l^2\varsigma^2(\sigma)\right) \\
&\leq \frac{1}{\Lambda_{\min}^2}\left(4K^3\eta_l^2G^2 + 4K^3\eta_l^2B^2\|\nabla f(w^t)\|_{\mathbf{A}_\sigma^{-1}}^2 + 2K^2\eta_l^2\varsigma^2(\sigma)\right) \\
&\leq \frac{1}{\Lambda_{\min}^2}\left(4K^3\eta_l^2G^2 + 8K^3\eta_l^2B^2\beta\big(f(w^t) - f(w^*)\big) + 2K^2\eta_l^2\varsigma^2(\sigma)\right),
\end{aligned}$$

which completes the proof. □

Lemma 7 (Perturbed Strongly Convexity (Karimireddy et al., 2020))**.** *TThe following holds for any β-smoothness and μ-strongly convex function h, and for any x, y, z in the domain of h:*

$$\langle\nabla h(x), z - y\rangle \geq h(z) - h(y) + \frac{\mu}{4}\|y - z\| - \beta\|z - x\|^2.$$

Lemma 8 (Subsampling Variance)**.** *(Lemma B.1 in (Lei and Jordan, 2017)) Given a vector space $\mathcal{X} \in \mathbb{R}^d$ with norm $\|\cdot\|$, we consider a dataset $x_1, x_2, ..., x_N \in \mathcal{X}$. We select a subset $\mathcal{S}$ with size S from the given dataset without replacement. The subsampling mechanism can be uniform subsampling or Poisson subsampling. The variance of the subset's average can be bounded by the following upper bound:*

$$\mathbb{E}\left\|\frac{1}{S}\sum_{j\in\mathcal{S}}x_j - \bar{x}\right\|^2 = \frac{1}{S}\left(1 - \frac{S-1}{N-1}\right)\frac{1}{N}\sum_{j=1}^{N}\|x_j - \bar{x}\|^2,$$

Appendix C.2. Setup

Before the proof of the main theorem, we denote the sever update in round t as Δ^t, which can be expressed as:

$$\Delta^t = -\frac{\tilde{\eta}}{S}\sum_{j=1}^{S}\sum_{i=1}^{K}\mathbf{A}_\sigma^{-1}\nabla g_j(w_j^{t,i}) + \mathbf{A}_\sigma^{-1}\eta_g\frac{\mathbf{n}(\mathcal{L})}{S}$$

$$\text{and}\quad \mathbb{E}[\Delta^t] = -\frac{\tilde{\eta}}{S}\sum_{j=1}^{S}\sum_{i=1}^{K}\mathbb{E}\mathbf{A}_\sigma^{-1}\nabla f_j(w_j^{t,i}),$$

where $\tilde{\eta} = \eta_l \eta_g$, $\mathbf{n}(\mathcal{L}) \sim \mathcal{N}(0, \nu_{\mathcal{L}}^2 \mathbf{I})$, and $\nu_{\mathcal{L}}$ is the noise level as a proportional function of the clipping parameter $\mathcal{L}$. We get $\nu_{\mathcal{L}} = \mathcal{L}\nu_1$ in Theorem 1 with clipping parameter $\mathcal{L}$.

Let the eigenvalue decomposition of $\mathbf{A}_\sigma^{-1}$ be $\mathbf{A}_\sigma^{-1} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where $\mathbf{\Lambda} = \text{diag}(\Lambda_i)$ is a diagonal matrix with

$$\Lambda_i = \frac{1}{1 + 2\sigma(1 - \cos(2\pi i/d))},$$

and denote the smallest eigenvalue of $\mathbf{A}_\sigma^{-1}$ by

$$\Lambda_{\min} = \min_{1 \le i \le d} \frac{1}{1 + 2\sigma(1 - \cos(2\pi i/d))} \ge \frac{1}{1 + 4\sigma}.$$

Appendix C.3. Proof of Theorem 4

Proof. We start from the total update of a communication round

$$\begin{aligned}
&\mathbb{E}\|w^{t+1} - w^*\|_{\mathbf{A}_\sigma}^2 \\
&\quad = \mathbb{E}\left\|w^t + \Delta^t - w^*\right\|_{\mathbf{A}_\sigma}^2 \\
&\quad = \mathbb{E}\left\|w^t - w^* - \frac{\tilde{\eta}}{S}\sum_{j=1}^{S}\sum_{i=1}^{K}\mathbf{A}_\sigma^{-1} g_j(w_j^{t,i}) + \mathbf{A}_\sigma^{-1}\eta_g^2 \frac{\mathbf{n}(\mathcal{L})}{S}\right\|_{\mathbf{A}_\sigma}^2 \\
&\quad \le \mathbb{E}\|w^t - w^*\|_{\mathbf{A}_\sigma}^2 \underbrace{- \frac{2\tilde{\eta}}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\langle w^t - w^*, \nabla f_j(w_j^{t,i})\rangle}_{A_1} + \underbrace{\eta_g^2 \mathbb{E}\left\|\frac{\mathbf{n}(\mathcal{L})}{S}\right\|_{\mathbf{A}_\sigma^{-1}}^2}_{A_2} \\
&\qquad + \underbrace{\tilde{\eta}^2 \mathbb{E}\left\|\frac{1}{S}\sum_{j=1}^{S}\sum_{i=1}^{K} g_j(w_j^{t,i})\right\|_{\mathbf{A}_\sigma^{-1}}^2}_{A_3}
\end{aligned} \tag{C.1}$$

As for A_1, we apply Lemma 7, we have

$$\begin{aligned}
A_1 &= \frac{2\tilde{\eta}}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\langle w^* - w^t, \nabla f_j(w_j^{t,i})\rangle \\
&\le \frac{2\tilde{\eta}}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\left(f_j(w^*) - f_j(w^t) + \beta\|w_j^{t,i} - w^t\|_{\mathbf{A}_\sigma}^2 - \frac{\mu}{4}\|w^t - w^*\|_{\mathbf{A}_\sigma}^2\right) \\
&\le -2\tilde{\eta}K\big(f(w^t) - f(w^*)\big) + \frac{2\tilde{\eta}\beta}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\|w_j^{t,i} - w^t\|_{\mathbf{A}_\sigma}^2 - \frac{\tilde{\eta}\mu K}{2}\|w^t - w^*\|_{\mathbf{A}_\sigma}^2
\end{aligned}$$

As for A_3, by the equation $\mathbb{E}X^2 = (\mathbb{E}X)^2 + \mathbb{E}(X - \mathbb{E}X)^2$, we have

$$A_3 \leq \underbrace{\tilde{\eta}^2\mathbb{E}\left\|\frac{1}{S}\sum_{j=1}^{S}\sum_{i=1}^{K}\nabla f_j(w_j^{t,i})\right\|_{\mathbf{A}_\sigma^{-1}}^2}_{B_1} + \underbrace{\tilde{\eta}^2\mathbb{E}\left\|\frac{1}{S}\sum_{j=1}^{S}\sum_{i=1}^{K}\left(g_j(w_j^{t,i}) - \nabla f_j(w_j^{t,i})\right)\right\|_{\mathbf{A}_\sigma^{-1}}^2}_{B_2} \tag{C.2}$$

For B_1, we have

$$\begin{aligned}
B_1 &= \tilde{\eta}^2\mathbb{E}\left\|\frac{1}{S}\sum_{j=1}^{S}\sum_{i=1}^{K}\left(\nabla f(w_j^{t,i}) - \nabla f_j(w^t) + \nabla f_j(w^t)\right)\right\|_{\mathbf{A}_\sigma^{-1}}^2 \\
&\leq 2\tilde{\eta}^2\mathbb{E}\left\|\frac{1}{S}\sum_{j=1}^{S}\sum_{i=1}^{K}\left(\nabla f(w_j^{t,i}) - \nabla f_j(w^t)\right)\right\|_{\mathbf{A}_\sigma^{-1}}^2 + 2\tilde{\eta}^2K^2\mathbb{E}\left\|\frac{1}{S}\sum_{j=1}^{S}\nabla f_j(w^t)\right\|_{\mathbf{A}_\sigma^{-1}}^2 \\
&\leq \frac{2\tilde{\eta}^2K}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\mathbb{E}\|\nabla f(w_j^{t,i}) - \nabla f(w^t)\|_{\mathbf{A}_\sigma^{-1}}^2 \\
&\qquad + 2\tilde{\eta}^2K^2\mathbb{E}\left\|\frac{1}{S}\sum_{j=1}^{S}\nabla f_j(w^t) - \nabla f(w^t) + \nabla f(w^t)\right\|_{\mathbf{A}_\sigma^{-1}}^2 \\
&\leq \frac{2\tilde{\eta}^2K\beta^2}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\mathbb{E}\|w_j^{t,i} - w^t\|_{\mathbf{A}_\sigma}^2 + 2\tilde{\eta}^2K^2\|\nabla f(w^t)\|_{\mathbf{A}_\sigma^{-1}}^2 \\
&\qquad + 2\tilde{\eta}^2K^2\mathbb{E}\left\|\frac{1}{S}\sum_{j=1}^{S}\nabla f_j(w^t) - \nabla f(w^t)\right\|_{\mathbf{A}_\sigma^{-1}}^2 \\
&\leq \frac{2\tilde{\eta}K\beta^2}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\mathbb{E}\|w_j^{t,i} - w^t\|_{\mathbf{A}_\sigma}^2 + 2\tilde{\eta}^2K^2\|\nabla f(w^t)\|_{\mathbf{A}_\sigma^{-1}}^2 \\
&\qquad + 4\tilde{\eta}^2K^2\left(1 - \frac{S}{N}\right)\frac{1}{S}\left(G^2 + B^2\|\nabla f(w^t)\|_{\mathbf{A}_\sigma^{-1}}^2\right) \\
&\leq \frac{2\tilde{\eta}^2K\beta^2}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\mathbb{E}\|w_j^{t,i} - w^t\|_{\mathbf{A}_\sigma}^2 \\
&\quad + 8\tilde{\eta}^2K^2(1 + B^2)\beta\big(f(w^t) - f(w^*)\big) + \frac{4\tilde{\eta}^2K^2}{S}\left(1 - \frac{S}{N}\right)G^2
\end{aligned} \tag{C.3}$$

where the second last inequality comes from Assumption 1 and Lemma 8. When we apply Lemma 8, we set $x_j = f_j(w^t)$ and $\bar{x} = \frac{1}{N}\sum_{j=1}^{N} f_j(w^t)$. What's more,

we use the inequality $1-\frac{S-1}{N-1} \leq 2\big(1-\frac{S}{N}\big)$. The last inequality comes from Assumption 2. As for B_2, we apply Assumption 5, then we have

$$\begin{aligned} B_2 &\leq \tilde{\eta}^2 \mathbb{E}\left\| \frac{1}{S}\sum_{j=1}^{S}\sum_{i=1}^{K}\big(g_j(w_j^{t,i}) - \nabla f_j(w_j^{t,i})\big)\right\|_{\mathbf{A}_\sigma^{-1}}^2 \\ &\leq \tilde{\eta}^2 \mathbb{E}\frac{1}{S^2}\sum_{j=1}^{S}\sum_{i=1}^{K}\|g_j(w_j^{t,i}) - \nabla f_j(w_j^{t,i})\|_{\mathbf{A}_\sigma^{-1}}^2 \\ &\leq \frac{\tilde{\eta}^2 K}{S}\varsigma^2(\sigma) \end{aligned} \tag{C.4}$$

By Lemma 5 and the assumption that $\eta_l \leq \frac{1}{8K\beta\eta_g(1+B^2)}$, whence $\eta_l\eta_g K\beta \leq \frac{1}{2}$, we have

$$\begin{aligned} \mathbb{E}\|w^{t+1}-w^*\|_{\mathbf{A}_\sigma}^2 \leq& \left(1-\frac{\tilde{\eta}\mu K}{2}\right)\mathbb{E}\|w^t-w^*\|_{\mathbf{A}_\sigma}^2 - \tilde{\eta}K\big(f(w^t)-f(w^*)\big) \\ &\underbrace{+ 3\tilde{\eta}\beta\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\mathbb{E}\|w_j^{t,i}-w^t\|_{\mathbf{A}_\sigma}^2}_{C} + \frac{\tilde{\eta}^2 K}{S}\varsigma^2(\sigma) + \eta_g^2\frac{\eta_l^2K^2L^2\nu_1^2 d_\sigma}{S^2} \\ &+ \frac{4\tilde{\eta}^2K^2}{S}\left(1-\frac{S}{N}\right)G^2 \end{aligned}$$

According to Lemma 6, and the assumption $\eta_g \geq 1$ ($\tilde{\eta} = \eta_g\eta_l \geq \eta_l$) and $\eta_l \leq \frac{\Lambda_{\min}^2}{8K\beta\eta_g(1+B^2)}$, for C, we have

$$\begin{aligned} &3\tilde{\eta}\beta\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\mathbb{E}\|w_j^{t,i}-w^t\|_{\mathbf{A}_\sigma}^2 \\ &\leq \frac{12\tilde{\eta}\eta_l^2K^3\beta G^2}{\Lambda_{\min}^2} + \frac{6\tilde{\eta}\eta_l^2\beta K^2\,\varsigma^2(\sigma)}{\Lambda_{\min}^2} + \frac{24\tilde{\eta}\eta_l^2K^3B^2\beta^2}{\Lambda_{\min}^2}\big(f(w^t)-f(w^*)\big) \\ &\leq \frac{12\tilde{\eta}^3K^3\beta G^2}{\Lambda_{\min}^2} + \frac{\tilde{\eta}^2K}{\eta_g^2}\varsigma^2(\sigma) + \frac{1}{2}\tilde{\eta}K\big(f(w^t)-f(w^*)\big) \end{aligned} \tag{C.5}$$

In this case,

$$\begin{aligned} \mathbb{E}\|w^{t+1}-w^*\|_{\mathbf{A}_\sigma}^2 \leq& \left(1-\frac{\tilde{\eta}\mu K}{2}\right)\mathbb{E}\|w^t-w^*\|_{\mathbf{A}_\sigma}^2 - \frac{1}{2}\tilde{\eta}K\big(f(w^t)-f(w^*)\big) \\ &+ \tilde{\eta}^2K^2\left(\big(\frac{1}{\eta_g^2K}+\frac{1}{SK}\big)\varsigma^2(\sigma) + \frac{L^2\nu_1^2d_\sigma}{S^2} + \frac{4}{S}\left(1-\frac{S}{N}\right)G^2 + \frac{12}{\Lambda_{\min}^2}\tilde{\eta}K\beta G^2\right) \end{aligned}$$

Reorganizing the terms, we have

$$f(w^t) - f(w^*) \leq \frac{2}{\tilde{\eta}K}\left(1 - \frac{\tilde{\eta}\mu K}{2}\right)\mathbb{E}\|w^t - w^*\|^2_{\mathbf{A}_\sigma} - \frac{2}{\tilde{\eta}K}\mathbb{E}\|w^{t+1} - w^*\|^2_{\mathbf{A}_\sigma} + 2\tilde{\eta}K\left(\left(\frac{1}{K\eta_g^2} + \frac{1}{SK}\right)\varsigma^2(\sigma) + \frac{L^2\nu_1^2 d_\sigma}{S^2} + \frac{4}{S}\left(1 - \frac{S}{N}\right)G^2 + \frac{12}{\Lambda_{\min}^2}\tilde{\eta}K\beta G^2\right) \tag{C.6}$$

By averaging using weights $a_t = q^{-t}$ where $q \triangleq \left(1 - \frac{\mu\tilde{\eta}K}{2}\right)$, we have

$$\sum_{t=0}^{T} a_t\big(\mathbb{E}[f(w^t)] - f(w^*)\big) \leq \frac{2}{\tilde{\eta}K}\|w^0 - w^*\|^2_{\mathbf{A}_\sigma} + \sum_{t=0}^{T} 2a_t\tilde{\eta}K\left(\left(\frac{1}{\eta_g^2 K} + \frac{1}{SK}\right)\varsigma^2(\sigma) + \frac{L^2\nu_1^2 d_\sigma}{S^2} + \frac{4}{S}\left(1 - \frac{S}{N}\right)G^2 + \frac{12}{\Lambda_{\min}^2}\tilde{\eta}K\beta G^2\right)$$

Diving by $\sum_{t=0}^{T} a_t$, we have

$$\mathbb{E}[f(\bar{w}^T)] - f(w^*) \leq \frac{2}{\tilde{\eta}K\sum_{t=0}^{T} a_t}\|w^0 - w^*\|^2_{\mathbf{A}_\sigma} + 2\tilde{\eta}K\left(\left(\frac{1}{\eta_g^2 K} + \frac{1}{SK}\right)\varsigma^2(\sigma) + \frac{L^2\nu_1^2 d_\sigma}{S^2} + \frac{4}{S}\left(1 - \frac{S}{N}\right)G^2 + \frac{12}{\Lambda_{\min}^2}\tilde{\eta}K\beta G^2\right)$$

Now we consider $\tilde{\eta}K\sum_{t=0}^{T} a_t = \tilde{\eta}K\sum_{t=0}^{T} q^{-t}$. Since we assume that $T \geq \frac{1}{\mu\tilde{\eta}K}$, we have

$$\tilde{\eta}K\sum_{t=0}^{T} a_t = \tilde{\eta}Kq^{-T}\sum_{t=0}^{T} q^t = \tilde{\eta}Kq^{-T}\frac{1 - (1 - \mu\tilde{\eta}K/2)^{T+1}}{\mu\tilde{\eta}K/2} \geq \frac{2q^{-T}}{3\mu}.$$

So

$$\frac{1}{\tilde{\eta}K\sum_{t=0}^{T} a_t} \leq \frac{3}{2}\mu q^T = \frac{3}{2}\mu\left(1 - \frac{\mu\tilde{\eta}K}{2}\right)^T \leq \frac{3}{2}\mu\exp(-\mu\tilde{\eta}KT/2)$$

In this case,

$$\mathbb{E}[f(\bar{w}^T)] - f(w^*) \leq 3\mu\exp(-\mu\tilde{\eta}KT/2)\underbrace{\|w^0 - w^*\|^2_{\mathbf{A}_\sigma}}_{D_\sigma} + \frac{24}{\Lambda_{\min}^2}\tilde{\eta}^2K^2\beta G^2 + 2\tilde{\eta}K\underbrace{\left(\left(\frac{1}{\eta_g^2 K} + \frac{1}{SK}\right)\varsigma^2(\sigma) + \frac{4}{S}\left(1 - \frac{S}{N}\right)G^2 + \frac{L^2\nu_1^2 d_\sigma}{S^2}\right)}_{H_\sigma} \tag{C.7}$$

Here we discuss two situations:

- If $\frac{1}{\mu KT} \leq \frac{\Lambda_{\min}^2}{8K\beta(1+B^2)} \leq \frac{2\log(\max(e,\mu^2 TD_\sigma/H_\sigma)))}{\mu KT}$, we choose $\tilde{\eta} = \frac{\Lambda_{\min}^2}{8K\beta(1+B^2)}$, then

$$\begin{aligned}
&\mathbb{E}[f(\bar{w}^T)] - f(w^*) \\
&\quad \leq 3\mu D_\sigma \exp\left(- \frac{\mu T}{16\beta(1+4\sigma)^2(1+B^2)}\right) + \tilde{O}\left(\frac{(1+4\sigma)^2\beta G^2}{\mu^2 T^2}\right) + \tilde{O}\left(\frac{H_\sigma}{\mu T}\right) \\
&\quad \leq \tilde{O}\left(\frac{H_\sigma}{\mu T}\right).
\end{aligned}$$

  where we use $\tilde{\eta}K \leq \frac{2\log(\max(e,\mu^2 RD_\sigma/H_\sigma)))}{\mu T} = \tilde{O}(1/(\mu T))$

- If $\frac{1}{\mu KT} \leq \frac{2\log(\max(e,\mu^2 TD_\sigma/H_\sigma)))}{\mu KT} \leq \frac{\Lambda_{\min}^2}{8K\beta(1+B^2)}$, we choose $\tilde{\eta} = \frac{2\log(\max(e,\mu^2 TD_\sigma/H_\sigma)))}{\mu KT}$, then

$$\begin{aligned}
&\mathbb{E}[f(\bar{w}^T)] - f(w^*) \\
&\quad \leq 3\mu D_\sigma \exp(-\log(\max(e, \mu^2 TD_\sigma/H_\sigma))) + \tilde{O}\left(\frac{H_\sigma}{\mu T}\right) + \tilde{O}\left(\frac{(1+4\sigma)^2\beta G^2}{\mu^2 T^2}\right) \\
&\quad \leq \tilde{O}\left(\frac{H_\sigma}{\mu T}\right)
\end{aligned}$$

  In this case, we choose $\tilde{\eta} = \min\left\{\frac{2\log(\max(e,\mu^2 TD_\sigma/H_\sigma)))}{\mu KT}, \frac{\Lambda_{\min}^2}{8K\beta(1+B^2)}\right\}$ $\left(T \geq \frac{8\beta(1+B^2)}{\mu\Lambda_{\min}^2}\right)$. Then

$$\mathbb{E}[f(\bar{w}^T)] - f(w^*) \leq \tilde{\mathcal{O}}\left(\frac{1}{\mu T}\left(\left(\frac{1}{\eta_g^2 K} + \frac{1}{SK}\right)\varsigma^2(\sigma) + \frac{4}{S}\left(1 - \frac{S}{N}\right)G^2 + \frac{L^2\nu_1^2 d_\sigma}{S^2}\right)\right)$$

which completes the proof. □

Appendix C.4. Proof of Theorem 5

Proof. We start from Eq. (C.6) and set $\mu = 0$ for general-convex case:

$$\begin{aligned}
f(w^t) - f(w^*) \leq \frac{2}{\tilde{\eta}K}\mathbb{E}\|w^t - w^*\|_{\mathbf{A}_\sigma}^2 - \frac{2}{\tilde{\eta}K}\mathbb{E}\|w^{t+1} - w^*\|_{\mathbf{A}_\sigma}^2 \\
+ 2\tilde{\eta}K\left(\left(\frac{1}{\eta_g^2 K} + \frac{1}{SK}\right)\varsigma^2(\sigma) + \frac{L^2\nu_1^2 d_\sigma}{S^2} + \frac{4}{S}\left(1 - \frac{S}{N}\right)G^2 + \frac{12}{\Lambda_{\min}^2}\tilde{\eta}K\beta G^2\right)
\end{aligned} \tag{C.8}$$

Summing the above inequality from $t = 0$ to $t = T$ and taking average , we have

$$\mathbb{E}\big[f(\bar{w}^t)] - f(w^*) \leq \frac{2}{\tilde{\eta}TK}\underbrace{\|w^0 - w^*\|^2_{\mathbf{A}_\sigma}}_{D_\sigma} + \underbrace{\frac{24}{\Lambda^2_{\min}}\beta G^2}_{Q_1}\tilde{\eta}^2K^2 + 2\tilde{\eta}K\underbrace{\Big(\big(\frac{1}{\eta_g^2K} + \frac{1}{SK}\big)\varsigma^2(\sigma) + \frac{L^2\nu_1^2 d_\sigma}{S^2} + \frac{4}{S}\Big(1 - \frac{S}{N}\Big)G^2\Big)}_{H_\sigma} \tag{C.9}$$

We set $\tilde{\eta}_{\max} = \frac{\Lambda^2_{\min}}{8K\beta(1+B^2)}$. Here we consider two situations:

- If $\tilde{\eta}^2_{\max} \leq \frac{D_\sigma}{H_\sigma TK^2}$ and $\tilde{\eta}^3_{\max} \leq \frac{D_\sigma}{Q_1TK^3}$, we set $\tilde{\eta} = \tilde{\eta}_{\max}$, then

$$\mathbb{E}[f(\bar{w}^T)] - f(w^*) \leq \frac{16\beta(1+B^2)(1+4\sigma)^2D_\sigma}{T} + 2\sqrt{\frac{D_\sigma H_\sigma}{T}} + \sqrt[3]{\frac{24(1+4\sigma)^2D_\sigma^2G^2\beta}{T^2}}$$

- If $\tilde{\eta}^2_{\max} \geq \frac{D_\sigma}{H_\sigma TK^2}$ or $\tilde{\eta}^3_{\max} \geq \frac{D_\sigma}{Q_1TK^3}$, we set $\tilde{\eta} = \min\left\{\sqrt{\frac{D_\sigma}{H_\sigma TK^2}}, \sqrt[3]{\frac{D_\sigma}{Q_1TK^3}}\right\}$, then

$$\mathbb{E}[f(\bar{w}^T)] - f(w^*) \leq 4\sqrt{\frac{D_\sigma H_\sigma}{T}} + \sqrt[3]{\frac{24(1+4\sigma)^2D_\sigma^2G^2\beta}{T^2}}$$

In conclusion, if we set $\tilde{\eta} = \min\left\{\sqrt{\frac{D_\sigma}{H_\sigma TK^2}}, \sqrt[3]{\frac{D_\sigma}{Q_1TK^3}}, \frac{\Lambda^2_{\min}}{8K\beta(1+B^2)}\right\}$, we have

$$\mathbb{E}[f(\bar{w}^T)] - f(w^*) \leq \frac{16\beta(1+B^2)(1+4\sigma)^2D_\sigma}{T} + 4\sqrt{\frac{D_\sigma H_\sigma}{T}} + \sqrt[3]{\frac{24(1+4\sigma)^2D_\sigma^2G^2\beta}{T^2}}$$

which completes the proof. □

Appendix C.5. Proof of Theorem 6

Proof of Theorem 6. According to the smoothness of f, we have

$$\begin{aligned}
f(w^{t+1}) &\leq f(w^t) + \langle \nabla f(w^t), w^{t+1} - w^t \rangle + \frac{\beta}{2}\|w^{t+1} - w^t\|_{\mathbf{A}_\sigma}^2 \\
&\leq f(w^t) - \left\langle \nabla f(w^t), \frac{\tilde{\eta}}{S}\sum_{j=1}^{S}\sum_{i=1}^{K}\mathbf{A}_\sigma^{-1} g_j(w_j^{t,i}) \right\rangle + \left\langle \nabla f(w^t), \mathbf{A}_\sigma^{-1}\eta_g \frac{\mathbf{n}(\mathcal{L})}{S} \right\rangle \\
&\quad + \frac{\beta}{2}\left(\left\| \frac{\tilde{\eta}}{S}\sum_{j=1}^{S}\sum_{i=1}^{K}\mathbf{A}_\sigma^{-1} g_j(w_j^{t,i}) \right\|_{\mathbf{A}_\sigma}^2 + \eta_g^2 \frac{\|\mathbf{A}_\sigma^{-1}\mathbf{n}(\mathcal{L})\|_{\mathbf{A}_\sigma}^2}{S^2} \right. \\
&\qquad \left. + 2\left\langle \frac{\tilde{\eta}}{S}\sum_{j=1}^{S}\sum_{i=1}^{K}\mathbf{A}_\sigma^{-1} g_j(w_j^{t,i}), \eta_g \frac{\mathbf{n}(\mathcal{L})}{S} \right\rangle \right)
\end{aligned}$$

By taking the expectation on both sides, we have

$$\begin{aligned}
\mathbb{E}[f(w^{t+1})] \leq f(w^t) &\underbrace{- \frac{\tilde{\eta}}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\langle \nabla f(w^t), \nabla f_j(w_j^{t,i}) \rangle_{\mathbf{A}_\sigma^{-1}}}_{A_1} + \underbrace{\frac{\beta\eta_g^2}{2S^2}\mathbb{E}\|\mathbf{n}(\mathcal{L})\|_{\mathbf{A}_\sigma^{-1}}^2}_{A_2} \\
&+ \underbrace{\frac{\tilde{\eta}^2\beta}{2}\mathbb{E}\left\| \frac{1}{S}\sum_{j=1}^{S}\sum_{i=1}^{K} g_j(w_j^{t,i}) \right\|_{\mathbf{A}_\sigma^{-1}}^2}_{A_3}
\end{aligned}$$

According to Eq (C.2), (C.3) and (C.4), we have

$$\begin{aligned}
A_3 &\leq 2\tilde{\eta}^2 K^2\beta(1+B^2)\|\nabla f(w^t)\|_{\mathbf{A}_\sigma^{-1}}^2 + 2\tilde{\eta}^2 K^2\beta\left(1 - \frac{S}{N}\right)\frac{1}{S}G^2 \\
&\quad + \frac{\tilde{\eta}^2 K\beta^3}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\mathbb{E}\|w_j^{t,i} - w^t\|_{\mathbf{A}_\sigma}^2 + \frac{\tilde{\eta}^2 K\beta}{2S}\varsigma^2(\sigma)
\end{aligned} \tag{C.10}$$

As for A_1, we apply the inequality $ab = \frac{1}{2}[(b-a)^2 - a^2] - \frac{1}{2}b^2 \geq \frac{1}{2}[a^2 - (b-a)^2]$, we have

$$\begin{aligned}
A_1 &\leq -\frac{\tilde{\eta}}{2N}\sum_{j=1}^{N}\sum_{i=1}^{K}\left[\|\nabla f(w^t)\|_{\mathbf{A}_\sigma^{-1}}^2 - \|\nabla f_j(w_j^{t,i}) - \nabla f(w^t)\|_{\mathbf{A}_\sigma^{-1}}^2 \right] \\
&\leq -\frac{\tilde{\eta}K}{2}\|\nabla f(w^t)\|_{\mathbf{A}_\sigma^{-1}}^2 + \frac{\tilde{\eta}\beta^2}{2N}\sum_{j=1}^{N}\sum_{i=1}^{K}\|w_j^{t,i} - w^t\|_{\mathbf{A}_\sigma}^2
\end{aligned}$$

According to Lemma 5, we have

$$A_2 \leq \frac{\tilde{\eta}^2 K^2 L^2 \beta \nu_1^2 d_\sigma}{2S^2}$$

Summing up A_1, A_2 and A_3, and using the inequality $\eta_l \leq \frac{\Lambda_{\min}^2}{8K\beta\eta_g(B^2+1)}$, where $\frac{1}{1+4\sigma} \leq \Lambda_{\min} \leq 1$ is the smallest eigenvalue of $\mathbf{A}_\sigma^{-1}$, we have

$$\begin{aligned}
\mathbb{E}[f(w^{t+1})] &\leq f(w^t) - \tilde{\eta}K\left(\frac{1}{2} - 2\tilde{\eta}K\beta(1+B^2)\right)\|\nabla f(w^t)\|^2_{\mathbf{A}_\sigma^{-1}} + \frac{\tilde{\eta}^2 K\beta}{2S}\varsigma^2(\sigma) \\
&\quad + \frac{\tilde{\eta}^2 K^2 L^2 \beta \nu_1^2 d_\sigma}{2S^2} + 2\tilde{\eta}^2 K^2 \beta\left(1 - \frac{S}{N}\right)\frac{1}{S}G^2 \\
&\quad + \tilde{\eta}\beta^2\left(\frac{1}{2} + \tilde{\eta}K\beta\right)\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\|w_j^{t,i} - w^t\|^2_{\mathbf{A}_\sigma} \\
&\leq f(w^t) - \frac{\tilde{\eta}K}{4}\|\nabla f(w^t)\|^2_{\mathbf{A}_\sigma^{-1}} + \frac{\tilde{\eta}^2 K\beta}{2S}\varsigma^2(\sigma) + \frac{\tilde{\eta}^2 K^2 L^2 \beta \nu_1^2 d_\sigma}{2S^2} \\
&\quad + 2\tilde{\eta}^2 K^2 \beta\left(1 - \frac{S}{N}\right)\frac{1}{S}G^2 + \tilde{\eta}\beta^2\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\|w_j^{t,i} - w^t\|^2_{\mathbf{A}_\sigma}
\end{aligned}$$

According to Lemma 6 and the assumption that $\eta_g \geq 1$ ($\tilde{\eta} = \eta_l\eta_g \geq \eta_l$) and $\eta_l \leq \frac{\Lambda_{\min}^2}{8K\beta\eta_g(B^2+1)}$, we have

$$\begin{aligned}
&\tilde{\eta}\beta^2\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\|w_j^{t,i} - w^t\|^2_{\mathbf{A}_\sigma} \\
&\leq \frac{1}{\Lambda_{\min}^2}\left(4K^3\tilde{\eta}\eta_l^2\beta^2 G^2 + 4K^3\tilde{\eta}\eta_l^2\beta^2 B^2\|\nabla f(w^t)\|^2_{\mathbf{A}_\sigma^{-1}} + 2K^2\tilde{\eta}\eta_l^2\beta^2\varsigma^2(\sigma)\right) \\
&\leq \frac{1}{\Lambda_{\min}^2}\left(4K^3\tilde{\eta}^3\beta^2 G^2 + 4K^3\tilde{\eta}^3\beta^2 B^2\|\nabla f(w^t)\|^2_{\mathbf{A}_\sigma^{-1}} + 2K^2\tilde{\eta}\eta_l^2\beta^2\varsigma^2(\sigma)\right) \\
&\leq \frac{4}{\Lambda_{\min}^2}K^3\tilde{\eta}^3\beta^2 G^2 + \frac{\tilde{\eta}K}{16}\|\nabla f(w^t)\|^2_{\mathbf{A}_\sigma^{-1}} + \frac{\tilde{\eta}^2 K\beta}{2\eta_g^2}\varsigma^2(\sigma)
\end{aligned} \tag{C.11}$$

In this case, we have

$$\begin{aligned}\mathbb{E}[f(w^{t+1})] &\leq f(w^t) - \frac{\tilde{\eta}K}{8}\|\nabla f(w^t)\|^2_{\mathbf{A}_\sigma^{-1}} + \frac{\tilde{\eta}^2K^2L^2\beta\nu_1^2 d_\sigma}{2S^2} + 2\tilde{\eta}^2K^2\beta\left(1-\frac{S}{N}\right)\frac{1}{S}G^2 \\ &\quad + \frac{\tilde{\eta}^2K^2\beta}{2}\left(\frac{1}{\eta_g^2 K} + \frac{1}{SK}\right)\varsigma^2(\sigma) + \frac{4}{\Lambda_{\min}^2}K^3\tilde{\eta}^3\beta^2G^2 \\ &\leq f(w^t) - \frac{\tilde{\eta}K}{8}\|\nabla f(w^t)\|^2_{\mathbf{A}_\sigma^{-1}} + \frac{4}{\Lambda_{\min}^2}K^3\tilde{\eta}^3\beta^2G^2 \\ &\quad + \frac{\tilde{\eta}^2K^2\beta}{2}\left((\frac{1}{\eta_g^2 K} + \frac{1}{SK})\varsigma^2(\sigma) + \frac{4}{S}\left(1-\frac{S}{N}\right)G^2 + \frac{L^2\nu_1^2\tilde{d}_\sigma}{S^2}\right)\end{aligned}$$

Summing the above inequality from $t=0$ to $t=T$ and taking average , we have

$$\begin{aligned}\mathbb{E}\|\nabla f(\bar{w}^t)\|^2_{\mathbf{A}_\sigma^{-1}} &\leq \frac{8}{\tilde{\eta}KT}\big(\underbrace{f(w^0)-f(w^*)}_{F_0}\big) + \underbrace{\frac{32\beta^2G^2}{\Lambda_{\min}^2}}_{Q_2}K^2\tilde{\eta}^2 \\ &\quad + 4\tilde{\eta}K\beta\Bigg(\underbrace{(\frac{1}{\eta_g^2K}+\frac{1}{SK})\varsigma^2(\sigma) + \frac{4}{S}\left(1-\frac{S}{N}\right)G^2 + \frac{L^2\nu_1^2d_\sigma}{S^2}}_{H_\sigma}\Bigg)\end{aligned} \tag{C.12}$$

We set $\eta_{\max} = \frac{\Lambda_{\min}^2}{8K\beta(1+B^2)}$. Here we consider two situations:

- If $\tilde{\eta}_{\max}^2 \leq \frac{F_0}{H_\sigma\beta TK^2}$ and $\tilde{\eta}_{\max}^3 \leq \frac{F_0}{Q_2TK^3}$, we set $\tilde{\eta} = \tilde{\eta}_{\max}$, then

  $$\mathbb{E}\|\nabla f(\bar{w}^T)\|^2_{\mathbf{A}_\sigma^{-1}} \leq \frac{64\beta(1+B^2)(1+4\sigma)^2F_0}{T} + 4\sqrt{\frac{F_0H_\sigma\beta}{T}} + \sqrt[3]{\frac{32(1+4\sigma)^2F_0^2G^2\beta^2}{T^2}}$$

- If $\tilde{\eta}_{\max}^2 \geq \frac{F_0}{H_\sigma\beta TK^2}$ or $\tilde{\eta}_{\max}^3 \geq \frac{F_0}{Q_2TK^3}$, we set $\tilde{\eta} = \min\{\sqrt{\frac{F_0}{H_\sigma\beta TK^2}}, \sqrt[3]{\frac{F_0}{Q_2TK^3}}\}$, then

  $$\mathbb{E}\|\nabla f(\bar{w}^T)\|^2_{\mathbf{A}_\sigma^{-1}} \leq 12\sqrt{\frac{F_0H_\sigma\beta}{T}} + \sqrt[3]{\frac{32(1+4\sigma)^2F_0^2G^2\beta^2}{T^2}}$$

In conclusion, if we set $\tilde{\eta} = \min\{\sqrt{\frac{F_0}{H_\sigma T\beta K^2}}, \sqrt[3]{\frac{F_0}{Q_2TK^3}}, \frac{\Lambda_{\min}^2}{8K\beta(1+B^2)}\}$, we have

$$\mathbb{E}\|\nabla f(\bar{w}^T)\|^2_{\mathbf{A}_\sigma^{-1}} \leq \frac{64\beta(1+B^2)(1+4\sigma)^2F_0}{T} + 12\sqrt{\frac{F_0H_\sigma\beta}{T}} + \sqrt[3]{\frac{32(1+4\sigma)^2F_0^2G^2\beta^2}{T^2}}, \tag{C.13}$$

which completes the proof. □

Appendix C.6. Proof of Corollary 1

Proof. Corollary 1 is a direct result of Theorem 4, 5 and 6. Here we take the non-convex case for example. We recall Eq. (C.13) that

$$\mathbb{E}\|\nabla f(\bar{w}^T)\|^2_{\mathbf{A}_\sigma^{-1}} \leq \frac{64\beta(1+B^2)(1+4\sigma)^2F_0}{T} + 12\sqrt{\frac{F_0H_\sigma\beta}{T}} + \sqrt[3]{\frac{32(1+4\sigma)^2F_0^2G^2\beta^2}{T^2}}.$$

To bound the error by ϵ, we require that for the first and the last term in the above equation, we have

$$T = O\left(\frac{64\beta(1+B^2)(1+4\sigma)^2F_0}{\epsilon}\right) \quad \text{and} \quad T = O\left(\frac{(1+4\sigma)F_0G\beta}{\epsilon^{3/2}}\right).$$

For the middle term, we have

$$T = O\left(\frac{F_0H_\sigma\beta}{\epsilon^2}\right).$$

Plugging in $H_\sigma = \left(\frac{1}{\eta_g^2K} + \frac{1}{SK}\right)\varsigma^2(\sigma) + \frac{4}{S}\left(1-\frac{S}{N}\right)G^2 + \frac{L^2\nu_1^2d_\sigma}{S^2}$, we conclude the proof. □

Appendix C.7. Proof of Corollary 2

Proof of Corollary 2. We assume $\log(1/\delta) \geq \varepsilon$, then applying Theorem 1 with ν_1, and $\tau = \frac{S}{N}$, we have

$$\begin{aligned}
\frac{L^2d_\sigma}{S^2}\cdot\nu_1^2 &= \frac{L^2d_\sigma}{S^2}\cdot\frac{\tau^2}{\varepsilon^2}\frac{14T}{\lambda}\left(\frac{\log(1/\delta)}{1-\lambda}+\varepsilon\right), && \text{by Theorem 1}\\
&\leq \frac{L^2d_\sigma}{S^2}\cdot\frac{\tau^2}{\varepsilon^2}\frac{14T}{\lambda}\left(\frac{\log(1/\delta)}{1-\lambda}+\log(1/\delta)\right), && \text{by assumption } \log(1/\delta)\geq\varepsilon\\
&\leq \frac{L^2d_\sigma}{S^2}\cdot\frac{\tau^2T\log(1/\delta)}{\varepsilon^2}\cdot\underbrace{\frac{14}{\lambda}\left(\frac{1}{1-\lambda}+1\right)}_{C_0}\\
&= \frac{C_0L^2d_\sigma T\log(1/\delta)}{\varepsilon^2N^2}
\end{aligned}$$

- μ **Strongly-Convex:** following the proof of Theorem 4, we have

$$\mathbb{E}[f(\bar{w}^T)] - f(w^*) \leq \tilde{\mathcal{O}}\left(\frac{1}{\mu T}\left(\frac{(1+\frac{S}{\eta_g^2})\varsigma^2(\sigma)}{SK} + \frac{4(1-\tau)}{S}G^2 + \frac{d_\sigma C_0L^2T\log(1/\delta)}{\varepsilon^2N^2}\right)\right)$$

If we select $T = \frac{\varepsilon^2 N^2}{C_0 L^2 S \log(1/\delta)}$ with $T \geq \frac{1}{\mu_\sigma \tilde{\eta} K}$ where

$$\tilde{\eta} = \min\left\{\frac{2\log(\max(e, \mu^2 T D_\sigma / H_\sigma))}{\mu_\sigma K T}, \frac{\Lambda_{\min}^2}{8K\beta(1+B^2)}\right\},$$

and assume $\eta_g \geq \sqrt{S}$, then we have

$$\mathbb{E}[f(\bar{w}^T)] - f(w^*) \leq \tilde{\mathcal{O}}\left(\frac{(d_\sigma + \varsigma^2(\sigma)/K + (1-\tau)G^2)L^2\log(1/\delta)}{\mu\varepsilon^2 N^2}\right).$$

- **General-Convex:** Following Theorem 5, we have

$$\begin{aligned}
&\mathbb{E}[f(\bar{w}^T)] - f(w^*) \\
&\quad \leq \mathcal{O}\left(\sqrt{\frac{D_\sigma H_\sigma}{T}}\right) \\
&\quad = \mathcal{O}\left(\sqrt{\frac{D_\sigma}{T}\left(\frac{(1+\frac{S}{\eta_g^2})\varsigma^2(\sigma)}{SK} + \frac{4}{S}(1-\tau)G^2 + \frac{L^2\nu_1^2 d_\sigma}{S^2}\right)}\right) \\
&\quad = \mathcal{O}\left(\sqrt{\frac{D_\sigma}{T}\left(\frac{(1+\frac{S}{\eta_g^2})\varsigma^2(\sigma)}{SK} + \frac{4}{S}(1-\tau)G^2 + \frac{d_\sigma C_0 L^2 T\log(1/\delta)}{\varepsilon^2 N^2}\right)}\right)
\end{aligned}$$

If we set $T = \frac{\varepsilon^2 N^2}{C_0 L^2 S \log(1/\delta)}$ and assume that $\eta_g \geq \sqrt{S}$, then we have

$$\mathbb{E}[f(\bar{w}^T)] - f(w^*) \leq \frac{\sqrt{(\varsigma^2(\sigma)/K + 4(1-\tau)G^2 + d_\sigma)D_\sigma L^2\log(1/\delta)}}{\varepsilon N}.$$

- **Non-Convex:** Following Theorem 6, we have

$$\begin{aligned}
&\mathbb{E}\|\nabla f(\bar{w}^T)\|^2_{\mathbf{A}_\sigma^{-1}} \\
&\quad \leq \mathcal{O}\left(\sqrt{\frac{F_0 H_\sigma \beta}{T}}\right) \\
&\quad = \mathcal{O}\left(\sqrt{\frac{F_0\beta}{T}\left(\frac{(1+\frac{S}{\eta_g^2})\varsigma^2(\sigma)}{SK} + \frac{4}{S}(1-\tau)G^2 + \frac{L^2\nu_1^2\tilde{d}_\sigma}{S^2}\right)}\right) \\
&\quad = \mathcal{O}\left(\sqrt{\frac{F_0\beta}{T}\left(\frac{(1+\frac{S}{\eta_g^2})\varsigma^2(\sigma)}{SK} + \frac{4}{S}(1-\tau)G^2 + \frac{\tilde{d}_\sigma C_0 L^2 T\log(1/\delta)}{\varepsilon^2 N^2}\right)}\right)
\end{aligned}$$

If we set $T = \frac{\varepsilon^2 N^2}{C_0 L^2 S \log(1/\delta)}$ and assume that $\eta_g \geq \sqrt{S}$, then we have

$$\mathbb{E}\|\nabla f(\bar{w}^T)\|^2_{\mathbf{A}_\sigma^{-1}} \leq \frac{\sqrt{(\varsigma^2(\sigma)/K + 4(1-\tau)G^2 + d_\sigma)F_0 \beta L^2 \log(1/\delta)}}{\varepsilon N},$$

which completes the proof. □

Appendix D. Details about Table 1 and Corollary 1

In Table 1, for (Khaled et al., 2020), the $\log(T)$ term in denominators are ignored. For the communication complexity with strongly-convex condition for (Karimireddy et al., 2020) and DP-Fed-LS, the $\log S$ and $\log K$ terms in numerator are ignored.

For Corollary 1, given fixed noise level ν_1 and communication round T, we would like to determined what (ε, δ)-DP can be achieved. Following from Theorem 1, we know that to satisfy (ε, δ)-DP, we need

$$\nu \geq \frac{\tau \mathcal{L}}{\varepsilon} \sqrt{\frac{14T}{\lambda}\left(\frac{\log(1/\delta)}{1-\lambda} + \varepsilon\right)}, \tag{D.1}$$

and ν satisfying $\nu^2/4\mathcal{L}^2 \geq \frac{2}{3}$ and $\alpha - 1 \leq \frac{\nu^2}{6\mathcal{L}^2}\log(1/(\tau\alpha(1+\nu^2/4\mathcal{L}^2)))$ for some $\lambda \in (0,1)$, where $\alpha = \log(1/\delta)/((1-\lambda)\varepsilon) + 1$. In other words, we require

$$T \leq \frac{\lambda \varepsilon^2 \nu_1^2}{14\tau^2\left(\frac{\log(1/\delta)}{1-\lambda} + \varepsilon\right)}. \tag{D.2}$$

and

$$\frac{\nu_1^2 \mathcal{L}^2}{S^2 \Delta^2(\mathbf{q})} = \frac{\nu_1^2}{4} \geq \frac{2}{3} \tag{D.3}$$

and

$$\alpha - 1 \leq \frac{\nu_1^2}{6} \ln \frac{1}{\tau\alpha(1+\nu_1^2/4)}. \tag{D.4}$$

for $\nu_1 = \nu/\mathcal{L}$. In other words, if $\nu_1 \geq 8/3$ and $\alpha - 1 \leq \frac{\nu_1^2}{6}\ln\frac{1}{\tau\alpha(1+\nu_1^2/4)}$, then (ε, δ)-DP satisfying Eq (D.2) can be achieved for any $\lambda \in (0,1)$.

In Theorem 1, we select $\lambda \in (0,1)$ such that ν_1's lower bound can satisfy two inequalities Eq. (D.3) and (D.4). However, in Corollary 1, our first step is to fix the noise level ν_1 such that it directly satisfies Eq. (D.3) and (D.4). In this case, $\lambda \in (0,1)$ is a free parameter. One could select $\lambda \in (0,1)$ such that the upper bound for T is maximized.

Appendix E. Comparison of Theorem 1 and 2 with Moment Accountants

In this section, we show that our bounds provided in Theorem 1 and Theorem 2 are tight by comparing them with the numerical moment accountants in (Wang et al., 2019b) and (Zhu and Wang, 2019; Mironov et al., 2019) respectively. We consider two settings where $T = 30$, $\tau = 0.05$, $N = 500$ and $T = 200$, $\tau = 0.05$, $N = 2000$, which we uses for the experiment over MNIST and SVHN respectively. Firstly, one thing we need to notice is that, in Theorem 1 and 2, noise level ν is in nearly inverse proportion to ε when ε is small, where the first term under the square root in Eq. 4 and Eq. 5 become the major term. However, when ε is relatively large, like settings we use in our experiment, this relation changes. The slopes of the curves lie in $[-1, -1/2]$, at similar rates. Note that when we apply Theorem 1 and 2, we will firstly select λ satisfying all the proposed conditions by line search. Then we choose the one minimizing the lower bound of ν.

Figure E.8 a) and b) compare Theorem 1 with accountant in (Wang et al., 2019b) under the two settings above. We can notice that the two curves are almost parallel when ε is relatively large. For Theorem 2 and accountant in (Zhu and Wang, 2019; Mironov et al., 2019), (Figure E.8 d) and e)), we can notice that their curves are getting close when ε becomes large. If we further choose a large $T = 1000$ ($\tau = 0.05$ and $N = 2000$), these observations are more obvious, which is shown in Figure E.8 c) and f). It demonstrate that our closed-form bounds are tight and only differ from numerical moment accountant by a constant.

Appendix F. Laplacian Smoothing

In Figure F.9, we compare the evolution curves of Gradient Descent (GD) and Laplacian smoothing Gradient Descent (LSGD). We can notice that the curve (Figure F.9 (b)) of LSGD is much more smoother than the one of GD.

Appendix G. Smoothness of Aggregated Gradients

In the following Figure G.10, we show the frequency distribution of *federated average of gradients* over a convolutional layer after we permute the ordering (input channel (I), output channel (O), width (W) and height (H)) or the weight indices. It follows the same experiment setting as Figure 2, and OIWH is the natural ordering. As we can see, the frequency distribution is insensitive to the ordering and the weight indices permutations.

In Figure G.11, we plot the curves of $\langle v, \mathbf{e}_i \rangle$, which verifies that the original signal v is smooth in the sense that $\langle v, \mathbf{e}_i \rangle \to 0$ rapidly. Here v is the first

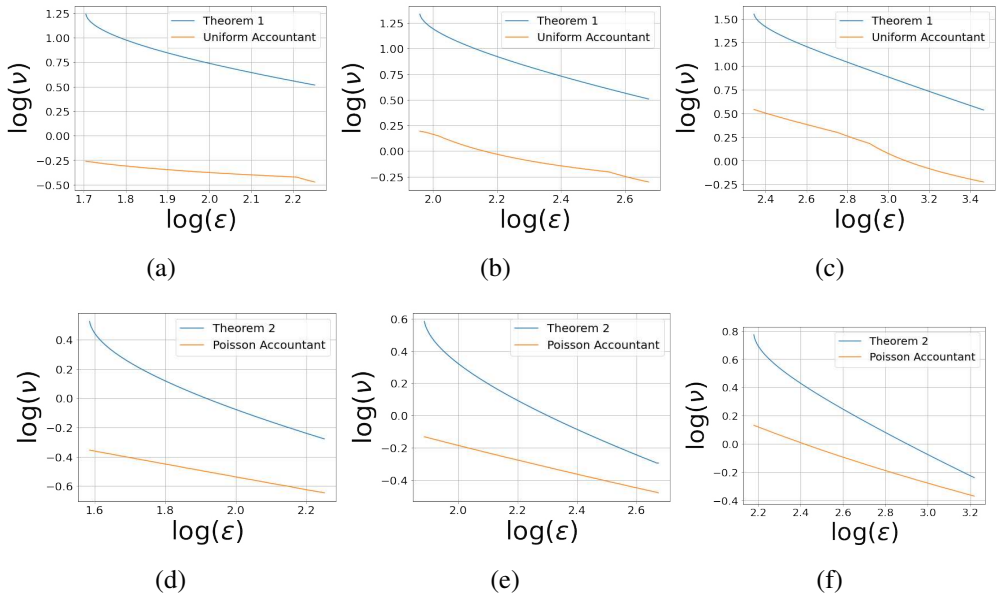
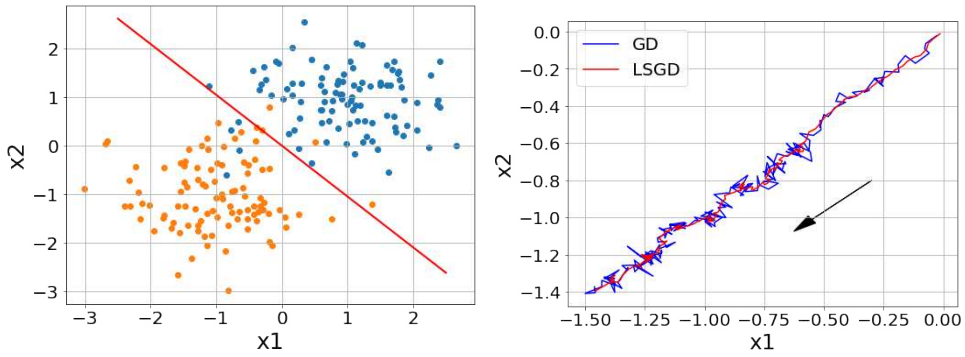


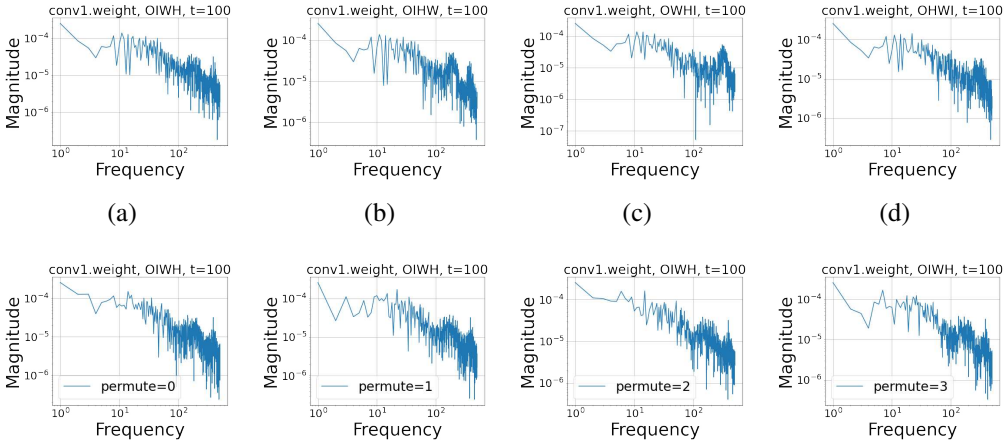
Figure E.8: Comparison of Theorem 1 and 2 with uniform (Wang et al., 2019b) and Poisson moment accountants (Zhu and Wang, 2019; Mironov et al., 2019). We can observe that Theorem 1 is nearly parallel to moment accountant (Wang et al., 2019b) and and 2 is close to moment accountant in (Zhu and Wang, 2019; Mironov et al., 2019) when ε becomes large. For example, in c), the slopes of least square regression for Theorem 1 and moment accountant are -0.80 and -0.73 respectively, while the intercepts are 3.31 and 2.29. It shows that the theoretical bound are of similar rates as numerical moment accountants and differ from moment accountants only by a small constant $e^{3.31-2.29} = 2.77$.

convolutional layer of *federated average of gradients* $\frac{1}{S}\sum_j \Delta_j^t$ of the CNN model defined in Section 5. Here $\mathbf{e}_i, i = 0, 1, ..., 100$ are the top 100 eigenvectors of the graph Laplacian corresponding to the largest eigenvalues.



(a) (b)

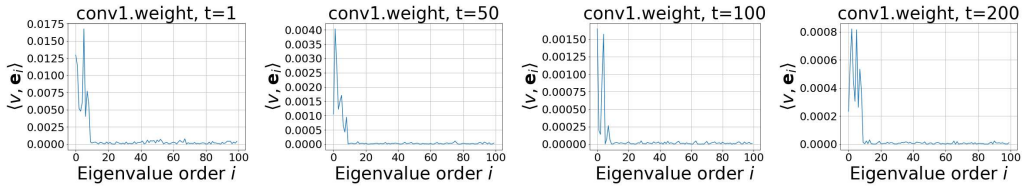
Figure F.9: Demonstration of Laplacian smoothing. We try to use a linear classifier $y = \text{sigmoid}(Wx)$ to separate data points from two distributions, *i.e.*, the blue points ($y = 0$) and the green points ($y = 1$) in (a). We use gradient descent (GD) and Laplacian smoothing gradient descent (LSGD with $\sigma = 1$) with binary cross entropy loss to fulfill this task. Here W is initialized as (0,0) and its perfect solution would be (c,c) for any $c < 0$. Gaussian noise with standard deviation of 0.3 is added on the gradients. Learning rate is set to be 0.1. In (b), we plot the evolution curves of W in 100 updates, where we can find that the curve of LSGD is much smoother than the one of GD.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure G.10: Frequency distribution of *federated average of gradients* $v = \frac{1}{S}\sum_j \Delta_j^t$ in non-DP federated learning, following experiment setting in Section 5. Here we use the first convolutional layer (conv1.weight) as an example. In (a)-(d), we permute the ordering of input channel (I), output channel (O), width (W) and height (H). In (e)-(h), we use the natural ordering but we permute the output channel weight indices of the convolutional layer with four independent permutations.



(a) (b) (c) (d)

Figure G.11: Projection $\langle v, \mathbf{e}_i \rangle$ of the first covolutional layer of *federated average of gradients* $\frac{1}{S}\sum_j \Delta_j^t$ in CNN, over different communication rounds t in non-DP federated learning, following experiment setting in Section 5. Here $\mathrm{e}_i, i = 0, 1, ..., 100$ are the top 100 eigenvectors of the graph Laplacian corresponding to the largest eigenvalues.

Appendix H. Detailed Experiment Settings and Other Results

In Table H.7, we list the default hyper-parameter for three classification models. To comply with traditional neural network training, we replace local iteration step K with local epoch E and denote the batch size as b. Here we use decay the local learning rate η_l by a factor of γ in each communication round.

	T	E	b	η_l	η_g	τ	$\mathcal{L}$	γ	N	weight decay
Logistic	30	5	10	0.1	1	0.05	0.4	0.99	1000/500	$4e-5$
CNN	200	10	64	0.1	1	0.05	0.3	0.99	2000	$4e-5$
LSTM	100	5	50	1.47	1	0.2	5	0.99	975	$4e-5$

Table H.7: Default parameters for logistic regression, CNN and LSTM

For all the tasks, we tune the hyper-parameters such that DP-Fed achieves the best validation accuracy, and then apply the same settings to DP-Fed-LS. During the parameter tuning, all results are reported based on 1 run. For example, the clipping parameter $\mathcal{L}$ is involved since a large one will induce too much noise while a small one will deteriorate training. In Table H.8, we show the result of different $\mathcal{L}$ for CNN, and we set the default $\mathcal{L}$ to 0.3. Other parameters are set as default in Table H.7. We further show the result of different local epoch E and local batch size b in Table H.9 for the CNN experiment. In Table H.10, we show the result of LSTM with different learning rates.

In Table H.11, we show the test accuracy of the curves in Figure 6 in Section 5.2. In Table H.12, we show the testing accuracy of the curves in Figure 7 in Section 5.4.

$\mathcal{L}$	0.1	**0.3**	0.5	0.7
$\sigma = 0.0$	77.55	84.14	82.75	81.54
$\sigma = 0.5$	75.76	85.23	84.79	82.33
$\sigma = 1.0$	73.73	85.24	84.70	82.35

Table H.8: Test Accuracy of CNN on SVHN with DP-Fed ($\sigma = 0$) and DP-Fed-LS ($\sigma = 0.5, 1$) under $(2.56, 1/2000^{1.1})$-DP guarantees ($z = 1.5$) and Poisson subsampling.

E, b	$E = 5, b = 32$	$E = 5, b = 64$	$E = 10, b = 32$	**E=10, b=64**
$\sigma = 0.0$	81.53	83.73	81.87	84.14
$\sigma = 0.5$	82.55	84.65	83.79	85.23
$\sigma = 1.0$	83.23	84.92	82.95	85.24

Table H.9: Test Accuracy of CNN on SVHN with DP-Fed ($\sigma = 0$) and DP-Fed-LS ($\sigma = 0.5, 1$) under $(2.56, 1/2000^{1.1})$-DP guarantees ($z = 1.5$) and Poisson subsampling.

$\mathcal{L}$	1.27	**1.47**	1.67	1.87
$\sigma = 0.0$	38.44	38.83	38.34	37.50
$\sigma = 0.5$	38.38	38.81	38.10	38.53
$\sigma = 1.0$	39.37	38.93	38.67	39.49

Table H.10: Test Accuracy of LSTM on Shakespeare with DP-Fed ($\sigma = 0$) and DP-Fed-LS ($\sigma = 0.5, 1$) under $(6.78, 1/975^{1.1})$-DP guarantees ($z = 1.6$) and Poisson subsampling.

		η_l	0.025	0.05	0.1	0.125
$z = 3.0$	$\varepsilon = 1.07$	$\sigma = 0.0$	72.14	70.11	64.88	59.99
		$\sigma = 1.0$	75.77	76.02	74.33	71.29
$z = 3.5$	$\varepsilon = 0.90$	$\sigma = 0.0$	66.46	66.14	56.48	51.98
		$\sigma = 1.0$	72.80	73.08	68.46	64.16
$z = 3.0$	$\varepsilon = 0.78$	$\sigma = 0.0$	63.51	58.06	51.16	41.12
		$\sigma = 1.0$	68.01	66.80	61.31	55.85

Table H.11: Test accuracy of CNN on SVHN with DP-Fed ($\sigma = 0.0$) and DP-Fed-LS ($\sigma = 1.0$) with Poisson subsampling, under different large noise level z and different local learning rate η_l in Figure 6 in Section 5.2.

	Uniform		Poisson	
	$z = 1.8$	$z = 2.2$	$z = 1.8$	$z = 2.2$
$\sigma = 0.0$	79.02	76.05	79.19	75.63
$\sigma = 1.0$	81.51	77.20	81.45	79.75

Table H.12: Test Accuracy of CNN on SVHN with DP-Fed ($\sigma = 0$) and DP-Fed-LS ($\sigma = 1$) with noise level $z = 1.8$ and $z = 2.2$ in Figure 7 in Section 5.4. The non-DP accuracy are 90.47 and 90.52 for uniform and Poisson subsampling respectively.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H., Mironov, I., Talwar, K., Zhang, L., 2016. Deep learning with differential privacy, in: 23rd ACM Conference on Computer and Communications Security (CCS 2016).

Amari, S.I., 1998. Natural gradient works efficiently in learning. Neural Computation 10, 251–276.

Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., Shmatikov, V., 2020. How to backdoor federated learning, in: International Conference on Artificial Intelligence and Statistics, PMLR. pp. 2938–2948.

Balle, B., Wang, Y.X., 2018. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising, in: International Conference on Machine Learning, PMLR. pp. 394–403.

Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., Talwar, K., 2007. Privacy, accuracy, and consistency too: a holistic solution to contingency table release, in: Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 273–282.

Bassily, R., Feldman, V., Talwar, K., Guha Thakurta, A., 2019. Private stochastic convex optimization with optimal rates. Advances in Neural Information Processing Systems 32, 11282–11291.

Bassily, R., Smith, A., Thakurta, A., 2014. Private empirical risk minimization: Efficient algorithms and tight error bounds, in: 2014 IEEE 55th Annual Symposium on Foundations of Computer Science, IEEE. pp. 464–473.

Berner, C., Brockman, G., Chan, B., Cheung, V., Debiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., et al., 2019. Dota 2 with large scale deep reinforcement learning. arXiv preprint arXiv:1912.06680 .

Bernstein, G., McKenna, R., Sun, T., Sheldon, D., Hay, M., Miklau, G., 2017. Differentially private learning of undirected graphical models using collective graphical models, in: International Conference on Machine Learning, PMLR. pp. 478–487.

Bhagoji, A.N., Chakraborty, S., Mittal, P., Calo, S., 2019. Analyzing federated learning through an adversarial lens, in: Proceedings of the 36th International Conference on Machine Learning, pp. 634–643.

Bun, M., Dwork, C., Rothblum, G.N., Steinke, T., 2018. Composable and versatile privacy via truncated cdp, in: Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, pp. 74–86.

Caldas, S., Wu, P., Li, T., Konečnỳ, J., McMahan, H.B., Smith, V., Talwalkar, A., 2018. Leaf: A benchmark for federated settings. arXiv preprint arXiv:1812.01097 .

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M., 2006. Our data, ourselves: Privacy via distributed noise generation, in: Annual International Conference on the Theory and Applications of Cryptographic Techniques, Springer. pp. 486–503.

Dwork, C., Nissim, K., 2004. Privacy-preserving datamining on vertically partitioned databases, in: Annual International Cryptology Conference, Springer. pp. 528–544.

Dwork, C., Roth, A., 2014. The algorithmic foundations of differential privacy. Foundations and trends in Theoretical Computer Science 9(3-4).

Fredrikson, M., Jha, S., Ristenpart, T., 2015. Model inversion attacks that exploit confidence information and basic countermeasures, in: 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS 2015).

Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., Ristenpart, T., 2014. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing, in: 23rd USENIX security symposium (USENIX Security 14), pp. 17–32.

Geyer, R.C., Klein, T., Nabi, M., 2017. Differentially private federated learning: A client level perspective. arXiv:1712.07557 .

Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., Ramage, D., 2018. Federated learning for mobile keyboard prediction. arXiv preprint arXiv:1811.03604 .

Hay, M., Li, C., Miklau, G., Jensen, D., 2009. Accurate estimation of the degree distribution of private networks, in: 2009 Ninth IEEE International Conference on Data Mining, IEEE. pp. 169–178.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Hitaj, B., Ateniese, G., Perez-Cruz, F., 2017. Deep models under the gan: information leakage from collaborative deep learning, in: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, ACM. pp. 603–618.

Jayaraman, B., Evans, D., 2019. Evaluating differentially private machine learning in practice, in: 28th USENIX Security Symposium (USENIX Security 19), pp. 1895–1912.

Jayaraman, B., Wang, L., Evans, D., Gu, Q., 2018. Distributed learning without distress: Privacy-preserving empirical risk minimization, in: Advances in Neural Information Processing Systems, pp. 6343–6354.

Jiang, D., Tan, C., Peng, J., Chen, C., Wu, X., Zhao, W., Song, Y., Tong, Y., Liu, C., Xu, Q., et al., 2021. A gdpr-compliant ecosystem for speech recognition with transfer, federated, and evolutionary learning. ACM Transactions on Intelligent Systems and Technology (TIST) 12, 1–19.

Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, S., Suresh, A.T., 2020. Scaffold: Stochastic controlled averaging for federated learning, in: International Conference on Machine Learning, PMLR. pp. 5132–5143.

Khaled, A., Mishchenko, K., Richtárik, P., 2020. Tighter theory for local sgd on identical and heterogeneous data, in: International Conference on Artificial Intelligence and Statistics, PMLR. pp. 4519–4529.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324.

Lei, L., Jordan, M., 2017. Less than a single pass: Stochastically controlled stochastic gradient, in: Artificial Intelligence and Statistics, pp. 148–156.

Li, X., Huang, K., Yang, W., Wang, S., Zhang, Z., 2019. On the convergence of fedavg on non-iid data, in: International Conference on Learning Representations.

McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A., 2017. Communication-efficient learning of deep networks from decentralized data, in: Artificial Intelligence and Statistics, PMLR. pp. 1273–1282.

McMahan, H.B., Andrew, G., Erlingsson, U., Chien, S., Mironov, I., Papernot, N., Kairouz, P., 2018a. A general approach to adding differential privacy to iterative training procedures. NeurIPS 2018 workshop on Privacy Preserving Machine Learnin .

McMahan, H.B., Ramage, D., Talwar, K., Zhang, L., 2018b. Learning differentially private recurrent language models. International Conference on Learning Representation .

Melis, L., Song, C., De Cristofaro, E., Shmatikov, V., 2019. Exploiting unintended feature leakage in collaborative learning, in: 2019 IEEE Symposium on Security and Privacy (SP), IEEE. pp. 691–706.

Mironov, I., 2017. Rényi differential privacy, in: 2017 IEEE 30th Computer Security Foundations Symposium (CSF), IEEE. pp. 263–275.

Mironov, I., Talwar, K., Zhang, L., 2019. Rényi differential privacy of the sampled gaussian mechanism. arXiv:1908.10530 .

Nemirovskij, A.S., Yudin, D.B., 1983. Problem complexity and method efficiency in optimization .

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y., 2011. Reading digits in natural images with unsupervised feature learning, in: NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011.

Nguyen, A., Do, T., Tran, M., Nguyen, B.X., Duong, C., Phan, T., Tjiputra, E., Tran, Q.D., 2022. Deep federated learning for autonomous driving, in: 33rd IEEE Intelligent Vehicles Symposium.

Nt, H., Maehara, T., 2019. Revisiting graph neural networks: All we have is low-pass filters. arXiv preprint arXiv:1905.09550 .

Osher, S., Wang, B., Yin, P., Luo, X., Barekat, F., Pham, M., Lin, A., 2022. Laplacian smoothing gradient descent. Research in the Mathematical Sciences 9, 55.

Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., Talwar, K., 2017. Semi-supervised knowledge transfer for deep learning from private training data. International Conference on Learning Representation .

Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., Erlingsson, Ú., 2018. Scalable private learning with pate. International Conference on Learning Representation .

Pathak, M., Rane, S., Raj, B., 2010. Multiparty differential privacy via aggregation of locally trained classifiers, in: Advances in Neural Information Processing Systems, pp. 1876–1884.

Sablayrolles, A., Douze, M., Schmid, C., Ollivier, Y., Jégou, H., 2019. White-box vs black-box: Bayes optimal strategies for membership inference, in: International Conference on Machine Learning, PMLR. pp. 5558–5567.

Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W., Bridgland, A., et al., 2020. Improved protein structure prediction using potentials from deep learning. Nature 577, 706–710.

Shokri, R., Stronati, M., Song, C., Shmatikov, V., 2017. Membership inference attacks against machine learning models. Proceedings of the 2017 IEEE Symposium on Security and Privacy .

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al., 2016. Mastering the game of go with deep neural networks and tree search. Nature 529, 484.

Wang, B., Gu, Q., Boedihardjo, M., Wang, L., Barekat, F., Osher, S.J., 2020. DP-LSSGD: A stochastic optimization method to lift the utility in privacy-preserving ERM, in: Proceedings of The First Mathematical and Scientific Machine Learning Conference, pp. 328–351.

Wang, D., Ye, M., Xu, J., 2017. Differentially private empirical risk minimization revisited: Faster and more general. Advances in Neural Information Processing Systems 30, 2722–2731.

Wang, L., Jayaraman, B., Evans, D., Gu, Q., 2019a. Efficient privacy-preserving stochastic nonconvex optimization. arXiv preprint arXiv:1910.13659 .

Wang, Y.X., Balle, B., Kasiviswanathan, S.P., 2019b. Subsampled Rényi differential privacy and analytical moments accountant, in: The 22nd International Conference on Artificial Intelligence and Statistics, PMLR. pp. 1226–1235.

Williams, O., McSherry, F., 2010. Probabilistic inference and differential privacy. Advances in Neural Information Processing Systems 23, 2451–2459.

Wu, B., Chen, C., Zhao, S., Chen, C., Yao, Y., Sun, G., Wang, L., Zhang, X., Zhou, J., 2020. Characterizing membership privacy in stochastic gradient Langevin dynamics, in: Thirty-Fourth AAAI conference on artificial intelligence.

Yeom, S., Giacomelli, I., Fredrikson, M., Jha, S., 2018. Privacy risk in machine learning: Analyzing the connection to overfitting, in: 2018 IEEE 31st Computer Security Foundations Symposium (CSF), IEEE. pp. 268–282.

Yu, H., Yang, S., Zhu, S., 2019. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 5693–5700.

Zhang, J., Wang, J., Zhao, Y., Chen, B., 2019. An efficient federated learning scheme with differential privacy in mobile edge computing, in: International Conference on Machine Learning and Intelligent Communications, Springer. pp. 538–550.

Zhu, L., Liu, Z., Han, S., 2019. Deep leakage from gradients, in: Advances in Neural Information Processing Systems.

Zhu, Y., Wang, Y.X., 2019. Poission subsampled Rényi differential privacy, in: International Conference on Machine Learning, pp. 7634–7642.