Echo of Neighbors: Privacy Amplification for Personalized Private Federated Learning with Shuffle Model

Yixuan Liu,[1,2,3] **Suyun Zhao,** [1,2,3] **Li Xiong** [4] **Yuhan Liu** [1,2,3] **Hong Chen** [1,2,3*]

[1] Key Laboratory of Data Engineering and Knowledge Engineering of Ministry of Education, Renmin University of China
[2] Engineering Research Center of Ministry of Education on Database and BI
[3] Information School, Renmin University of China, Beijing, China
[4] Department of Computer Science, Emory University
{liuyixuan, zhaosuyun, liuyh2019, chong}@ruc.edu.cn, lxiong@emory.edu,

Abstract

Federated Learning, as a popular paradigm for collaborative training, is vulnerable against privacy attacks. Different privacy levels regarding users' attitudes need to be satisfied locally, while a strict privacy guarantee for the global model is also required centrally. Personalized Local Differential Privacy (PLDP) is suitable for preserving users' varying local privacy, yet only provides a central privacy guarantee equivalent to the worst-case local privacy level. Thus, achieving strong central privacy as well as personalized local privacy with a utility-promising model is a challenging problem. In this work, a general framework (APES) is built up to strengthen model privacy under personalized local privacy by leveraging the privacy amplification effect of the shuffle model. To tighten the privacy bound, we quantify the heterogeneous contributions to the central privacy user by user. The contributions are characterized by the ability of generating "echos" from the perturbation of each user, which is carefully measured by proposed methods Neighbor Divergence and Clip-Laplace Mechanism. Furthermore, we propose a refined framework (S-APES) with the post-sparsification technique to reduce privacy loss in high-dimension scenarios. To the best of our knowledge, the impact of shuffling on personalized local privacy is considered for the first time. We provide a strong privacy amplification effect, and the bound is tighter than the baseline result based on existing methods for uniform local privacy. Experiments demonstrate that our frameworks ensure comparable or higher accuracy for the global model.

1 Introduction

Federated Learning (FL) (McMahan et al. 2017) is an emerging machine learning paradigm that allows multiple clients to train a global model collaboratively while keeping the private raw data of each client locally. While not directly sharing private data, recent works indicate that FL by itself is insufficient to preserve privacy of users' data. By observing the global model or intermediate parameters during the training process, adversaries can infer the membership of users or even reconstruct training records (Fredrikson, Jha, and Ristenpart 2015; Zhu, Liu, and Han 2019; Nasr, Shokri, and Houmansadr 2019). These attacks can lead to severe data leakage, hence it is necessary to provide additional protection with strict privacy guarantees for both the global model and local parameters. Moreover, in practice, different local privacy levels may be desired depending on users' privacy preferences. A one-size-fits-all approach would either downgrade the model utility or sacrifice privacy protection for certain users. Thus, an open problem in FL is how to provide strong central privacy as well as personalized local privacy while maintaining model utility.

Methods	Personalization	FL Process	
		Local	Central
PLDP	✓	✓	Weak
Uni-Shuffle	✗	✓	✓
APES	✓	✓	✓
S-APES	✓	✓	Strong

Table 1: Comparison of related work. ✓denotes protected, ✗denotes unprotected.

Several recent works have attempted to address this problem. Personalized Local Differential Privacy (PLDP) protects both local gradients and the global model by perturbing gradients with heterogeneous parameters (Chen et al. 2016; Li et al. 2020; Shen, Xia, and Yu 2021; Yang, Wang, and Wang 2021). The central privacy of the global model is equivalent to the weakest local privacy. For achieving both strong central and local privacy, a potential solution is the shuffle model (Bittau et al. 2017). It amplifies central privacy by permuting data points randomly after local perturbation. However, existing studies on shuffle model only focus on the scenarios where local privacy requirements are assumed uniform (Uni-Shuffle for short) (Erlingsson et al. 2019; Balle et al. 2019; Girgis et al. 2021; Feldman, McMillan, and Talwar 2022). To the best of our knowledge, there is no work that provides both strong central privacy for the global model and personalized local privacy guarantees, while achieving strong utility of global model (cf. Tab.1).

To narrow this gap, we propose **APES**, a privacy **A**mplification framework for **PE**rsonalized private federated learning with **S**huffle model (cf. Fig. 1). APES gains a strong

*Corresponding author: Hong Chen, chong@ruc.edu.cn

privacy amplification effect. Unlike previous works that just permute data, both data points and privacy parameters are randomly shuffled in APES. Clip-Laplace Mechanism is also introduced to implement the framework without damaging model utility. In order to mitigate the privacy-loss explosion problem caused by high dimensions, we propose **S-APES** which improves **<u>APES</u>** with the post-**<u>S</u>**parsification. The basic idea is to select only informative dimensions of gradients after perturbation and pad the rest, which saves privacy cost.

To bound the privacy of APES and S-APES, we carefully quantify the obfuscation effects contributed by users with heterogeneous privacy parameters. First, inspired by Feldman, McMillan, and Talwar, the central privacy of a specific user is boosted by the rest of the users who generate "echos" of her with heterogeneous probabilities; next, to measure the probabilities, we propose Neighbor Divergence and Clip-Laplace Mechanism for limited output range and bounded divergence among distinct output distributions by users' local randomizers; then "echos" are transformed into certain form, and a tight privacy bound is derived.

Our main contributions are summarized as follows:

(i) We propose privacy amplification frameworks via shuffle model for personalized private federated learning. APES strikes a better balance between central privacy and model utility with Neighbor Divergence and Clip-Laplace Mechanism. Based on it, improved S-APES enhances the privacy for the high-dimension scene.

(ii) We provide theoretical analysis for both privacy and convergence bound of the proposed frameworks. To the best of our knowledge, the shuffling effect on personalized local differential privacy is considered for the first time and a strong privacy amplification effect is yielded. The central privacy bound is tighter than the bound derived by naïvely adopting existing methods for unified privacy.

(iii) Comprehensive experiments are conducted to confirm that APES and S-APES achieve comparable or higher accuracy for the global model with stronger central privacy compared to the state-of-the-art methods without downgrading personalized local privacy guarantee.

2 Preliminaries

In this section, we illustrate the privacy definition, shuffle model and several properties of differential privacy, all of which are prepared for the proposed methods.

2.1 Central and Local Differential Privacy

Differential privacy (DP) (Dwork, Roth et al. 2014) is a *de facto* standard that is widely accepted to preserve privacy in FL. The notion is typically built up in a central setting where a trusted server can access the raw data. Local differential privacy (LDP) (Erlingsson, Pihur, and Korolova 2014), on the other hand, offers users a stronger privacy guarantee for the settings without assumption of a trusted server.

Definition 1 (Differential Privacy) *For any $\epsilon, \delta \geq 0$, a randomized algorithm $M : \mathcal{D} \to \mathcal{Z}$ is (ϵ, δ)-differential privacy if for any neighboring datasets $D, D' \in \mathcal{D}$ and any subsets $S \subseteq \mathcal{Z}$,*

$$\Pr[M(D) \in S] \leq e^{\epsilon} \Pr[M(D') \in S] + \delta$$

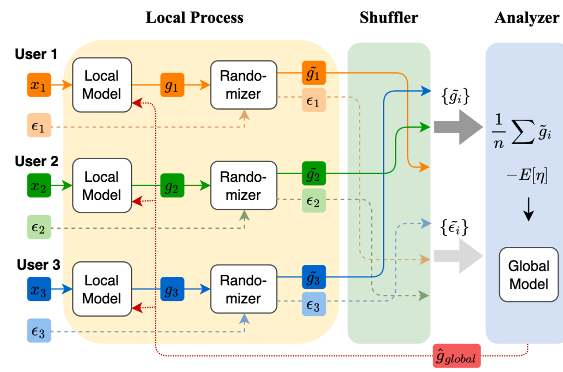

Figure 1: Procedure of APES. Gradients g_i trained by user data x_i are randomized locally, then privacy parameters ϵ_i and g_i are shuffled separately. Analyzer acts as the curator to aggregate and calibrate gradients $\tilde{g}_i$ for global model.

Definition 2 (Local Differential Privacy) *For any $\epsilon, \delta \geq 0$, an algorithm $M : \mathcal{D} \to \mathcal{Z}$ is (ϵ, δ)-local differential privacy if $\forall g, g' \in \mathcal{D}$ and $\forall z \in \mathcal{Z}$,*

$$\Pr[M(g) = z] \leq e^{\epsilon} \Pr[M(g') = z] + \delta$$

2.2 Shuffle-based Privacy

Shuffle model (Bittau et al. 2017) was proposed to strengthen central privacy while preserving local user privacy. Given n datapoints as the dataset $D = \{g_1, g_2, ..., g_n\}$, each $g_i \in D$ owned by user u_i is perturbed locally by a randomizer $M : \mathcal{D} \to \mathcal{Z}$ to ensure (ϵ^l, δ^l)-LDP before being sent to shuffler. Shuffler, a trusted third party, permutes and releases all the datapoints by algorithm $S : Z \to Z$ to analyzer. Untrusted analyzer aggregates all the datapoints. The process $P = S \circ M$ satisfies at least (ϵ^l, δ^l)-DP against analyzer (cf. Lemma 1). Recent works (Erlingsson et al. 2019; Balle et al. 2019; Girgis et al. 2021; Feldman, McMillan, and Talwar 2022) achieve a much stronger central privacy guarantee, which is considered as privacy amplification effect by shuffling. Among existing works, Feldman, McMillan, and Talwar provides a tight privacy upper bound for single-message summation. Take neighboring datasets D and D' that only differ at g_1 (or g'_1), any perturbed datapoint $\tilde{g}_i$ can be regarded as a sampling from the distribution of a specific perturbed point $\tilde{g}_1$ or $\tilde{g}'_1$ with probability $\exp(-\epsilon^l)$. By this observation, the privacy bound is yielded.

2.3 Privacy Tools

As a general technique to implement DP, Laplace Mechanism (Dwork, Roth et al. 2014) perturbs numerical values.

Definition 3 (Laplace Mechanism) *Given any function $f : \mathcal{D} \to \mathcal{Z}^d$ and neighboring datasets D and D', let $\Delta f = \max ||f(D) - f(D')||_1$ be the sensitivity function, Laplace mechanism $M(D) = f(D) + Y^d$ satisfies ϵ-DP, where Y^d is random variable $i.i.d$ drawn from distribution $Lap(0, \frac{\Delta f}{\epsilon})$.*

Composition theorems provide tight bounds for the algorithm combined with several DP blocks.

Lemma 1 (Parallel Composition) *(Yu et al. 2019) Given an* (ϵ_i, δ_i)*-DP algorithm* $M_i : \mathcal{D} \to \mathcal{Z}$ *for* $i \in [m]$*, a class of* $\{M_i\}_{i\in[m]}$ *on disjoint subsets of* D *is* $(\max \epsilon_i, \max \delta_i)$*-DP.*

Lemma 2 (Advanced Composition) *(Dwork, Roth et al. 2014) Given an* (ϵ_i, δ_i)*-DP algorithm* $M_i : \mathcal{D} \to \mathcal{Z}$ *for* $i \in [m]$*, the sequence of* $\{M_i\}_{i\in[m]}$ *on the same dataset* D *under m-fold composition is* $(\epsilon', \delta' + m\delta)$*-DP where* $\epsilon' = \epsilon\sqrt{2m \log 1/\delta'} + m\epsilon(e^\epsilon - 1)$.

No matter what dataset or query is adopted, any privacy mechanism can be reduced to a basic random response with the same privacy level (Kairouz, Oh, and Viswanath 2015).

Lemma 3 (Degraded Privacy) *For any* ϵ*-DP mechanism* M*, for* $X : \{x, \bar{x}\}$*,* $\exists \tilde{M}$ *dominates* M *where:*

$$\Pr[\tilde{M}(x) = z] = \begin{cases} \frac{e^\epsilon}{1+e^\epsilon}, & z = x \\ \frac{1}{1+e^\epsilon}, & z = \bar{x} \end{cases}$$

3 Proposed Methods

This section illustrates our methods for a strong privacy amplification effect. We first introduce Clip-Laplace Mechanism to implement the effect. Then two frameworks are proposed. APES is a general framework which shuffles both privacy parameters and gradients, the improved S-APES sparsifies dimensions without downgrading shuffling effect.

3.1 Clip-Laplace Mechanism

To make bounding privacy while maintaining model accuracy possible, it is necessary to introduce a mechanism for LDP with a finite and fixed output range. Existing works on this task provide non-fixed output ranges (Geng et al. 2018), or increase noise scale when ranges of input and output are not overlapped (Holohan et al. 2018; Croft, Sack, and Shi 2022). To address this issue, we introduce a variant of Laplace Mechanism, *Clip-Laplace*, which provides ϵ-DP for continuous real values with the same finite output ranges.

Definition 4 (Clip-Laplace Mechanism) *Given any function* $f : \mathcal{X} \to \mathcal{Y}^d$ *and sensitivity* $\Delta f = \max ||f(X) - f(X')||_1$ *for any neighboring datasets* X *and* X'*. Clip-Laplace Mechanism is* $M : \mathcal{Y}^d \to \mathcal{Z}^d$*. Each* $Z \in \mathcal{Z}^d$ *is a r.v. i.i.d. drawn from distribution* $CLap(f(x), \lambda, A)$ *of which the probability density function is defined as follows:*

$$p(z) = \begin{cases} \frac{1}{2\lambda S} \exp\left(-\frac{|z-f(x)|}{\lambda}\right), & -A \le z \le A \\ 0, & otherwise \end{cases}$$

where normalization factor $S = 1 - \frac{1}{2}\exp(\frac{-A+f(x)}{\lambda}) - \frac{1}{2}\exp(\frac{-A-f(x)}{\lambda})$ *and* $A \ge \Delta f/2$.

Theorem 1 *Clip Laplace mechanism preserves* ϵ*-LDP when the* $f(x) \in [-\Delta f/2, \Delta f/2]$*, and* $\lambda = \Delta f/\epsilon$.

The proof is provided in Appendix A.

Discussion. (i) When achieving the same level of ϵ-LDP, the variance of Clip-Laplacian outputs is smaller than classic Laplacian outputs. This property is based on the assumption of symmetric limited range of inputs (cf. Theorem 1), which is reasonable for many fields such as gradients aggregation, location statistics, financial analysis and so on. (ii) The Clip-Laplacian outputs are biased. A feasible solution for correction is to calibrate the outputs with the expectation, which can be estimated when privacy parameters are given.

3.2 APES Framework

We formalize **APES**, a privacy **A**mplification framework for **PE**rsonalized private federated learning with **S**huffle Model. The framework includes three procedures: local updating, shuffling and analyzing process with three parties separately. Convergence upper bound of APES is given at last.

Architecture Consider 3 parties: (i) n users, each holds a dataset X_i and a randomizer M_i satisfying ϵ^l_i-LDP. (ii) A shuffler with algorithm S. (iii) An analyzer, trains global model with shuffled messages. The process $P = S \circ M$ ensures (ϵ^c, ϵ^c)-DP for global model, where $M = (M_1, ..., M_n)$ with $\epsilon^l = (\epsilon^l_1, ..., \epsilon^l_n)$ for dimension level.

Basic Framework Algorithm 1 outlines the procedures of APES. We denote clip bound by C, learning rate by α and training epochs by T. Main procedures are as follows:

- *Local Updating*. Each user randomizes each dimension of model gradient g_i with ϵ^l_i by applying Clip-Laplace Mechanism. Both perturbed gradient $\tilde{g}_i$ and ϵ^l_i are sent to Shuffler. To keep the order of dimensions, dimension index k of $\tilde{g}_i$ is sent as well.
- *Shuffling Process*. Shuffler shuffles $\{\tilde{g}_i\}_{i\in[n]}$ within the same dimension, $\{\epsilon^l_i\}_{i\in[n]}$ is also permuted.
- *Analyzing Process*. Considering Clip-Laplace Mechanism is biased, the average gradient $\tilde{g}$ needs to be calibrated. We cannot calibrate $\tilde{g}_i$ one by one as the correspondence of ϵ^l_i and g_i is invisible to analyzer. Empirically, we observe that the value of $\tilde{g}$ is close to the value of $\mathbb{E}[\tilde{\tilde{g}}]$ (cf. Fig. 6 in Appendix C), where $\bar{g} = \frac{1}{n}\sum_{i=1}^n g_i$, $\mathbb{E}[\tilde{\tilde{g}}] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\tilde{\tilde{g}}_i]$ and $\tilde{\tilde{g}}_i \sim CLap(\bar{g}, 2C/\epsilon^l_i, C)$. Hence we can estimate the clean gradients $\bar{g}$ by approximating $\mathbb{E}[\tilde{g}]$ with $\mathbb{E}[\tilde{\tilde{g}}]$. Specifically, $\mathbb{E}[\tilde{g}]$ is estimated by $\tilde{g}$, and each term of $E[\tilde{\tilde{g}}]$ with ϵ^l_i is as follows:

$$\mathbb{E}[\tilde{\tilde{g}}_i] = \frac{(C + \lambda_i) \cdot (e_1 - e_2) + 2\bar{g}}{2 - e_1 - e_2} \tag{1}$$

where $e_1 = e^{\frac{-C-\bar{g}}{\lambda_i}}$, $e_2 = e^{\frac{-C+\bar{g}}{\lambda_i}}$ and $\lambda_i = 2C/\epsilon^l_i$.

Convergence Analysis To demonstrate the performance of global model under Clip-Laplace perturbation, we provide the upper bound of convergence of Algorithm 1 with the objective function $h(w; w^{(0)}) = F(w) + \frac{\mu}{2}||w - w^{(0)}||^2$. The regularization term $\frac{\mu}{2}||w - w^{(0)}||^2$ of h is introduced for the ease of calculation (Li et al. 2020).

Theorem 2 (Convergence Upper Bound) *After* T *aggregations, the expected decrease in the global loss function* $f(w) = \frac{1}{n}\sum_i F_i(w)$ *of APES is bounded as follows:*

$$\mathbb{E}[f(\tilde{w}^{(T)}) - f(w^*)] \le a_1^T(\mathbb{E}[f(\tilde{w}^{(0)})] - f(w^*)) + \frac{a_1^T - 1}{a_1 - 1}(O(a_2 C/\min(\epsilon^l_i)) + O(a_3 C^2/\min(\epsilon^l_i)^2))$$

where $a_1 = 1 + \frac{2\beta(\alpha B - 1)}{\mu} + \frac{2\beta L B(\alpha+1)}{\mu\bar{\mu}} + \frac{2\beta L B^2(1+\alpha)^2}{\bar{\mu}^2}$, $a_2 = L(\frac{1}{\mu} + \frac{BL(1+\alpha)}{\bar{\mu}})$, $a_3 = \frac{L}{2}$.

Algorithm 1: Basic Framework: APES

Input $T, \{(X_i, \epsilon_i^l)\}_{i\in[n]}, h(w), C, \alpha$.
Output model w
Analyzer initializes and broadcasts $w^{(0)}$.
for $t = 1, 2, ..., T$ **do**
  ▷ `Local Updating`
  for each user $i \in [n]$ **do**
    $w_i \leftarrow \tilde{w}^{(t)}$ ▷ Update local model
    $g_i \leftarrow \nabla_{w_i}\mathcal{L}(w_i, X_i)$
    $\bar{g}_i \leftarrow \text{Clip}(g_i, -C, C)$
    $\tilde{g}_i \leftarrow Randomize(\cdot)$ ▷ Local perturbation
    user i uploads $(\tilde{g}_i, \epsilon_i^l)$ to Shuffler
  ▷ `Shuffling Process`
  for each dimension $k \in [d]$ **do**
    generate permutation π_k over $[d]$
    $\{(\tilde{g}_{i,\pi_k(k)}, k)\}_{i\in[n]} \leftarrow \text{Shuffle}(\pi_k, \{\tilde{g}_{i,k}\}_{i\in[n]})$
  generate permutation π over $[n]$
  $\{\epsilon^l_{\pi(i)}\}_{i\in[n]} \leftarrow \text{Shuffle}(\pi, \{\epsilon_i^l\}_{i\in[n]})$ ▷ Shuffle ϵ
  Sends $\{\{(\tilde{g}_{i,\pi_k(k)}, k)\}_{i\in[n]}\}_{k\in[d]}$ and $\{\epsilon^l_{\pi(i)}\}_{i\in[n]}$
  ▷ `Analyzing Process`
  for each dimension $k \in [d]$ **do**
    $\tilde{g}_k \leftarrow \frac{1}{n}\sum_i \tilde{g}_{i,k}$ ▷ Aggregate by dimension
  $\hat{g} \leftarrow \text{Calibrate}(\tilde{g}, \{\epsilon_i\}_{i\in[n]})$
  $w^{(t+1)} \leftarrow w^{(t)} - \alpha\hat{g}$ and broadcast.
return $w^{(T)}$

The proof refers to Appendix B.

Discussion. The convergence upper bound increases as the bias and variance (the second and the third term) of Clip-Laplace perturbation grow, of which the influence is the same as classic Laplace Mechanism.

3.3 S-APES Framework

To strengthen privacy in the high-dimension scenario, we propose **S-APES** framework, which improves **APES** with post-**S**parsification technique.

Since gradients are usually high-dimensional, limiting the number of dimensions helps to save the privacy cost (Ye and Hu 2020; Duan, Ye, and Hu 2022). Selecting part of dimensions with large magnitude keeps majority of information (Aji and Heafield 2017) and reduces privacy loss, but needs extra protection since the selection itself is data-dependent process. To select informative dimensions without breaching privacy, we propose post-parsification technique.

Post Sparsification Algorithm 2 demonstrates the local process of S-APES with post-sparsification. Concretely, each user u_i is asked to select the largest b absolute values over d dimensions of $\tilde{g}_i$. To keep the selected dimension index private, the selection is executed after local perturbation. For avoiding the shuffling effect degradation caused by members reduction, user pads the rest of $(d - b)$ dimensions with perturbed 0. Denotes sparsification process by K, then the whole process of S-APES is defined as $P_s = S \circ K \circ M$.

Algorithm 2: $Randomize(\cdot)$ for S-APES

Input $\{(g_i, \epsilon_i^l)\}_{i\in[n]}, C$.
Output perturbed gradient $\tilde{g}_i$
  $\tilde{g}_i \leftarrow \text{CLap}(0, (d\Delta f)/\epsilon_i^l, C)$ ▷ Clip-Laplace perturbing
  $I_b \leftarrow \{k | k \in max(|\tilde{g}_{i,k}|_{k\in[d]})\}^b$ ▷ Post-top-b index set
  for each index $k \notin I_b$ **do**
    $\tilde{g}_{i,k} \leftarrow \text{CLap}(0, (d\Delta f)/\epsilon_i^l, C)$ ▷ Dummy padding
  return $\tilde{g}_i$

4 Privacy Analysis

In this section, we first derive a naïve privacy bound based on existing works, then show the local and central privacy bound of our frameworks. The sketch of privacy amplification effect analysis is provided at last.

4.1 Baseline Results

To analyze the privacy amplification effect of shuffling under personalized LDP, the most naïve way is applying existing shuffling bounds (Erlingsson et al. 2019; Balle et al. 2019; Girgis et al. 2021; Feldman, McMillan, and Talwar 2022) on heterogeneous local privacy budgets, i.e., ϵ_i^l, with classic Laplace Mechanism. However, different ϵ_i^l lead to different scales of the Laplacian distributions and their divergence may be infinite. As a result, the central privacy may be unbounded. Hence based on the previous work (Feldman, McMillan, and Talwar 2022) we can only approximate the true bound by using the same maximum ϵ_i^l for all users:

$$\epsilon^c \leq \ln(1+\frac{e^{\max(\epsilon_i^l)}-1}{e^{\max(\epsilon_i^l)}+1}(\frac{8(e^{\max(\epsilon_i^l)}\log(4/\delta))^{1/2}}{n^{1/2}}+\frac{8e^{\max(\epsilon_i^l)}}{n})) \tag{2}$$

Intuitively, the mixture distribution formed by all the users' output distributions w.r.t ϵ_i^l after shuffling is still "confused" to attacker. Hence it is possible to obtain an exact bound.

4.2 Main Results

Without loss of generality, we suppose two neighboring datasets $D = \{g_1, g_2, ..., g_n\}$ and $D' = \{g_1', g_2, ..., g_n\}$ that only differs at g_1 or g_1' of user u_1.

Theorem 3 (Local Bound) *Given $\epsilon_l = (\epsilon_1^l, ..., \epsilon_n^l)$, the local process $M = (M_1, ..., M_n)$ of APES on d-dimension gradients satisfies ϵ_i^l-LDP in dimension level, $d\epsilon_i^l$-LDP in user level for each user u_i.*

Discussion. Our frameworks achieve personalized LDP for each user. This comes from Theorem 1.

Theorem 4 (Central Upper bound) *Let $i, j \in [n]$, $\delta_s \in [0,1]$, $\sum_{i=2}^n\sum_{j=1}^n \frac{p_{ij}}{n} \geq 16\ln(4/\delta_s)$, $P = S\circ M$ of APES satisfies (ϵ^c, δ^c)-DP where $\delta^c \leq \frac{e^{\epsilon_j^l}-1}{e^{\epsilon_j^l}+1}\delta_s$,*

$$\epsilon^c \leq \ln(1+\frac{e^{\max(\epsilon_j^l)}-1}{e^{\max(\epsilon_j^l)}+1}(\frac{8(\ln(4/\delta_s))^{1/2}}{(\sum_{i=2}^n\sum_{j=1}^n\frac{p_{ij}}{n})^{1/2}}+\frac{8}{\sum_{i=2}^n\sum_{j=1}^n\frac{p_{ij}}{n}}))$$

when $\sum_{i=2}^n\sum_{j=1}^n \frac{p_{ij}}{n} \geq 16\ln(4/\delta_s)$, $\delta_s \in [0,1]$ and $p_{ij} = \frac{\epsilon_i^l}{\epsilon_j^l}\cdot\frac{1-e^{-\epsilon_j^l}}{1-e^{-\epsilon_i^l}}\cdot e^{-\max(\epsilon_i^l,\epsilon_j^l)}$.

Discussion. APES gains a strong central privacy for dimension level. Theorem 4 indicates most users are provided with much stricter central privacy as ϵ^c than their local privacy ϵ_i^l. A sketch of the proof is provided in section 4.3.

Proposition 1 (User Level Central Bound) *With $\delta'^{uc} > 0$ and $0 < b < d$, the process $P_s = S \circ K \circ M$ of S-APES with b-dimension sparsification is $(\epsilon^{uc}, \delta^{uc})$-DP where $\epsilon^{uc} = \epsilon^c\sqrt{4b\ln(1/\delta^{uc})} + 2b\epsilon^c(\exp(\epsilon^c) - 1)$ and $\delta^{uc} = \delta'^{uc} + 2b\delta^c$.*

Discussion. S-APES achieves the same dimension-level ϵ^c as APES. Considering dimensions of a gradient are not independent and extracting b dimensions leads to $2b$ sensitivity, we derive the user-level privacy amplification effect by composition theorems. Note that ϵ^{uc} grows linearly with b, which implies privacy loss reduces when fewer dimensions are uploaded by post-sparsification.

4.3 EoN: Privacy Amplification Analysis

To analyze privacy of proposed frameworks, we first introduce Neighbor Divergence, then present the sketch of Echo of Neighbors (EoN) analysis for privacy amplification effect.

Neighbor Divergence We introduce *Neighbor Divergence* for characterizing how well a user's output distribution closes the gap between itself and other users' distributions. Concretely, it defines the distance among output distributions of local randomizers of users with heterogeneous privacy budgets and different raw datapoints.

Definition 5 (Neighbor Divergence) *Consider any g_0, $g_1 \in \mathcal{D}$ and randomizers M_i, M_j satisfying ϵ_i, ϵ_j-LDP separately. Let $\mu_i^{(0)}$ and $\mu_j^{(1)}$ be distributions of $M_i(g_0)$ and $M_j(g_1)$ respectively, $U_i^{(0)} \sim \mu_i^{(0)}$, $U_j^{(1)} \sim \mu_j^{(1)}$, the neighbor divergence between μ_i and μ_j' is defined as:*

$$D_N(\mu_i^{(0)}||\mu_j^{(1)}) = \max_{S \subseteq Supp(U_i^{(0)})}[\ln \frac{\Pr[U_i^{(0)} \in S]}{\Pr[U_j^{(1)} \in S]}]$$

In particular, the neighbor divergence under Clip-Laplace Mechanism is demonstrated as follows.

Lemma 4 *Let $f(x) \in [-C, C]$, $\lambda = \Delta f/\epsilon$ and $\Delta f = 2C$, the neighbor divergence between distribution $\mu_i^{(0)}$ and $\mu_j^{(1)}$ under Clip-Laplace Mechanism is $D_N(\mu_i^{(0)}||\mu_j^{(1)}) \leq \ln(\alpha\frac{\epsilon_i}{\epsilon_j}e^{(\frac{(\epsilon_i+\epsilon_j)}{2} + \frac{A|\epsilon_i - \epsilon_j|}{2C})})$. Specifically, $D_N(\mu_i^{(0)}||\mu_j^{(1)}) \leq \ln(\frac{\epsilon_i}{\epsilon_j} \cdot \frac{1-e^{-\epsilon_j}}{1-e^{-\epsilon_i}} \cdot e^{\max(\epsilon_i, \epsilon_j)})$ when $A = C$. α denotes $\frac{(1-\frac{1}{2}\exp(\frac{\epsilon_j(-A+C)}{2C}) - \frac{1}{2}\exp(\frac{\epsilon_j(-A-C)}{2C}))}{(1-\frac{1}{2}\exp(\frac{\epsilon_i(-A+C)}{2C}) - \frac{1}{2}\exp(\frac{\epsilon_i(-A-C)}{2C}))}$*

A sketch of EoN Analysis We analyze the central privacy bound in Theorem 4 with the observation of *Echos of Neighbors*. There are three main steps: (i) After shuffling, output distributions of the rest of the users are converted into the same distribution from u_1 which can be seen as "echos" by neighbor divergence. (ii) Then all the "echos" are transformed into certain distributions which disentangle from different ϵ_i^l by degraded privacy. These distributions form a mixed distribution. (iii) Finally, we measure the divergence between the mixed distributions on D and D'.

Step (i). Recall that LDP mechanism $M_i : \mathcal{Y} \to \mathcal{Z}$ satisfying ϵ_i^l-LDP for any $i \in [n]$. Based on neighbor divergence, for any $\mu_i^{(s)}$ and $\mu_j^{(t)}$ by local randomizers we have $p_{ij} \leq \mu_i^{(s)}/\mu_j^{(t)}$ where $p_{ij} = e^{-D_N(\mu_j^{(t)}||\mu_i^{(s)})}$ by Lemma 4. Specifically, for any user's distribution $\mu_i^{(i)}$ on $g_i \in D\backslash\{g_1, g_1'\}$, "echo" $\mu_j^{(1)}$(or $\mu'^{(1)}_j$) from u_1 with g_1 (or g_1') is generated as follows:

$$\mu_i^{(i)} = \frac{p_{ij}}{2}\mu_j^{(1)} + \frac{p_{ij}}{2}\mu'^{(1)}_j + (1 - p_{ij})\gamma_i^{(i)} \tag{3}$$

The distribution $\gamma_i^{(i)} = \mu_i^{(i)} - p_{ij}/2 \cdot (\mu_j^{(1)} + \mu'^{(1)}_j)/(1 - p_{ij})$. The idea is inspired by a prior work (Feldman, McMillan, and Talwar 2022). Consider the situation that both g and ϵ^l are shuffled, the correspondence between g_i and ϵ_i^l is broken. An adversary cannot decide which ϵ_j^l is used for perturbing g_1, hence any value in $\{\epsilon_i^l\}$ is possible. Based on it we derive a general bound, then consider the worst-case situation with the largest ϵ_j on g_1 for the upper bound at step (iii).

Step (ii). Except for u_1, the mixed distributions of multiple $\mu_j^{(1)}$ with different ϵ_j from $n-1$ users is still hard to bound. Hence, with the help of degraded privacy (cf. Lemma 3) we transform $(\mu_j^{(1)} + \mu'^{(1)}_j)$ into $(\rho^{(1)} + \rho'^{(1)})$ for any $j \in [n]$ to disentangle ϵ_j from $\mu_j^{(1)}$.

Lemma 5 (Transformation) *Let $\rho^{(1)}$ and $\rho'^{(1)}$ denotes the distribution of a function $\mathcal{G} : g_1 \to \mathcal{Z}$ and $\mathcal{G}' : g_1' \to \mathcal{Z}$ respectively, $\mu_i^{(i)}$ be the distribution of $M_i(g_i)$, and $\gamma_i^{(i)}$ be the rest part of $\mu_i^{(i)}$ except $\rho^{(1)}$ and $\rho'^{(1)}$, then $\mu_i^{(i)}$ is mapped as follows.*

$$\mu_i^{(i)} = \frac{1}{n}\sum_{j=1}^{n}(\frac{p_{ij}}{2}\rho^{(1)} + \frac{p_{ij}}{2}\rho'^{(1)} + (1 - p_{ij})\gamma_i^{(i)}) \tag{4}$$

where $p_{ij} = \exp(-D_N(\mu_j^{(1)}||\mu_i^{(i)}))$.

Proof *By Lemma 3, we have $\mu_j^{(1)} = (e^\epsilon/(1+e^\epsilon))\rho^{(1)} + (1/(1+e^\epsilon))\rho'^{(1)}$ and $\mu'^{(1)}_j = (1/(1+e^\epsilon))\rho^{(1)} + (e^\epsilon/(1+e^\epsilon))\rho'^{(1)}$. The influence of ϵ_j on $\mu_i^{(i)}$ is isolated:*

$$\mu_i^{(i)} = \frac{1}{n}\sum_{j=1}^{n}(\frac{p_{ij}}{2}\mu_j^{(1)} + \frac{p_{ij}}{2}\mu'^{(1)}_j + (1 - p_{ij})\gamma_i^{(s)})$$

$$= \frac{1}{n}\sum_{j=1}^{n}(\frac{p_{ij}}{2}\rho^{(1)} + \frac{p_{ij}}{2}\rho'^{(1)} + (1 - p_{ij})\gamma_i^{(s)})$$

Step (iii). Now we can bound the divergence of the transformed distributions on D and D'.

Lemma 6 (Generalized Central Bound) *Let $i, j \in [n]$, $\delta_s \in [0, 1]$, $\sum_{i=2}^{n}\sum_{j=1}^{n}\frac{p_{ij}}{n} \geq 16\ln(4/\delta_s)$, $P = S \circ M$ of APES on D and D' is (ϵ^c, δ^c)-distinguishable where $\delta^c \leq \frac{e^{\epsilon_j}-1}{e^{\epsilon_j}+1}\delta_s$,*

$$\epsilon_c \leq \ln(1 + \frac{e^{\epsilon^*} - 1}{e^{\epsilon^*} + 1}(\frac{8(\ln(4/\delta_s))^{1/2}}{(\sum\limits_{i=2}^{n}\sum\limits_{j=1}^{n}\frac{p_{ij}}{n})^{1/2}} + \frac{8}{\sum\limits_{i=2}^{n}\sum\limits_{j=1}^{n}\frac{p_{ij}}{n}}))$$

Proof *By Lemma 5, any output distribution* $\mu_i^{(i)}$ *can be mapped into* $\rho^{(1)}$ *or* $\rho'^{(1)}$ *with probability* $p_{ij}/2n$*, into* $\gamma_i^{(i)}$ *with* $(1-p_{ij})/n$*. Consider* $(n-1)$ *outputs of users, we get a set of mapping distributions including* $n(n-1)$ *elements.*

With any $T \subseteq [n(n-1)]$*, we define an mapping event* $U_T = \{u_1, ...u_{n(n-1)}\}$ *where*

$$u_t = \begin{cases} \rho^{(1)} \text{or } \rho'^{(1)}, & t \in T \\ \gamma_t^{(t)}, & t \in [n(n-1)] \backslash T \end{cases}$$

The effect of γ_i *is removed in process* P *under the same* U_T*:*

$$\frac{\Pr[P(D) = \mathbf{z}]}{\Pr[P(D') = \mathbf{z}]} = \frac{\Pr[U_T]\Pr[P(D) = \mathbf{z}|U_T]}{\Pr[U_T]\Pr[P(D') = \mathbf{z}|U_T]} = \frac{\Pr[P(D_T) = \mathbf{z}]}{\Pr[P(D'_T) = \mathbf{z}]}$$

Then we define $T_0 \subseteq T$*, any element* $u_t \in U_T$ *is mapped to* $\rho^{(1)}$ *for* $t \in T_0$*, to* $\rho'^{(1)}$ *for* $t \in T \backslash T_0$*. Put aside the randomness on* g_1 *and* g'_1 *for now, to reach the mixed output* $\mathbf{z}$ *with the same number of* $\rho^{(1)}$ *or* $\rho'^{(1)}$*, mapping event* T_0 *on* D *and* T'_0 *on* D' *should be different as* $|T'_0| - |T_0| = 1$*. Recall that* $|T| \sim \sum_{i=2}^n \sum_{j=1}^n \text{Bern}(p_{ij}/n)$ *and* $|T_0| \sim \text{Bin}(1/2, |T|)$ *according to Lemma 5, we can bound the divergence between* $P(D_T)$ *and* $P(D'_T)$ *as follows:*

$$\frac{\Pr[P(D_T) = \mathbf{z}]}{\Pr[P(D'_T) = \mathbf{z}]} = \frac{\Pr[P(D_{T,T_0}) = \{\rho_1 \in T_0, \rho'_1 \in T \backslash T_0\}]}{\Pr[P(D'_{T,T'_0}) = \{\rho_1 \in T'_0, \rho'_1 \in T \backslash T'_0\}]}$$

$$\cdot \frac{\Pr[U_{T_0}]}{\Pr[U'_{T'_0}]} = \frac{\binom{|T|}{|T_0|}(\frac{1}{2})^{|T_0|}(\frac{1}{2})^{|T|-|T_0|}}{\binom{|T|}{|T'_0|}(\frac{1}{2})^{|T'_0|}(\frac{1}{2})^{|T|-|T'_0|}} = \frac{|T_0|+1}{|T|-|T_0|} \quad (5)$$

With Chernoff bound and Hoeffding's inequality, when $\sum_{i=2}^n \sum_{j=1}^n \frac{p_{ij}}{n} \geq 16\ln(4/\delta_s)$*, Eq. (5) is bounded as* $\frac{|T_0|+1}{|T|-|T_0|} \leq \ln(1 + \frac{8(\ln(4/\delta_s))^{1/2}}{(\sum_{i=2}^n \sum_{j=1}^n \frac{p_{ij}}{n})^{1/2}} + \frac{8}{\sum_{i=2}^n \sum_{j=1}^n \frac{p_{ij}}{n}})$.

At last, we consider the randomness on g_1 *and* g'_1 *with certain privacy budget* ϵ^**, the rest of the proof follows existing work (Feldman, McMillan, and Talwar 2022) and the general bound is proved. The full proof of Lemma 6 is provided to Appendix A.*

From the analysis above, it is realized that which ϵ^* adopted by g_1 or g'_1 is crucial for the bound. For the worst case that $\epsilon^* = \max(\epsilon_j)$ for $j \in [n]$, the divergence is upper bounded as Theorem 4. The proof refers to Appendix A.

5 Experiments

We conduct comprehensive experiments on APES and S-APES with the public dataset and various privacy settings.

5.1 Experiment Settings

Dataset and Implementation QMNIST (Yadav and Bottou 2019) is an extended version of MNIST dataset (LeCun et al. 1998), which consists of 120,000 28-by-28-pixel images with 10 classes. We set users as n=10,000 and partition the dataset evenly for users. The frameworks are evaluated with a Logistic Regression model with d=7850. All the experiments are implemented on a workstation with an Intel(R) Xeon(R) E5-2640 v4 CPU at 2.40GHz and a NVIDIA Tesla P40 GPU running on Ubuntu.

Baselines We compare the proposed methods with the following schemes. (i) Baseline frameworks include: **Non-Private**: FedProx (Li et al. 2020) without privacy protection. **LDP-Min**: all users adopt $\min \epsilon_i^l$ as privacy budget compulsively, which preserves privacy of all the users. **PLDP**: FedProx with personalized LDP. **UniS**: FedProx with shuffle model under personalized LDP, the privacy bound refers to Eq. (2). All the baseline frameworks exploit classic Laplace Mechanism as local randomizer. (ii) Baseline bounds of privacy amplification effect include: the numerical generic result of **BBGN'19** (Balle et al. 2019), the nemurical result of **FMT'22** (Feldman, McMillan, and Talwar 2022), the upper bound of **GDDTK'21** (Girgis et al. 2021) and **Erlingsson'19** (Erlingsson et al. 2019).

Parameter Selection We stimulate the personalized privacy preference ϵ^l for several situations as Tab. 2. δ^s for shuffling is set to 10^{-8} and δ^{uc} after dimension composition is 3.6×10^{-5}, smaller than $1/n$.

5.2 Experiment Results

We first show the effectiveness of the total frameworks, then confirm the privacy amplification effect, Clip-Laplace, and post-sparsification adopted in frameworks separately.

Effectiveness of frameworks Tab. 3 demonstrates that our frameworks achieve stronger central privacy with comparable or higher utility under the same personalized LDP. we compare the model accuracy and central privacy budgets of one epoch under Uniform2. (i) APES gains stricter privacy and the highest accuracy than baseline private frameworks. Dimension-level ϵ^c and user-level ϵ^{uc} reduce by more than 21% compared to UniS and PLDP. LDP-min gets tighter bound, yet the model performs poorly. There is no baseline framework achieves both better sides. The performance of APES benefits from privacy amplification effect of EoN and Clip-Laplace perturbation. (ii) S-APES provides the same ϵ^c as APES and further enhances user-level privacy. ϵ^{uc} diminishes by 55.6%, 66.7%, 99.6% compared to APES, UniS, and PLDP separately, It is noticed that local ϵ^{ul} also drops by dimension reduction. Though S-APES sacrifices accuracy of APES by 1.8%, it is still higher than baselines. The post-sparsification in S-APES substantially boosts privacy with this tolerable utility reduction.

Fig. 3 confirms that our frameworks perform well on multiple distributions and ranges of ϵ^l locally (cf. Tab. 2). (i) Accuracy of APES and S-APES is higher than UniS for the most settings. An exception is in Gauss1 LDP, which implies that S-APES may bot be appropriate for small ϵ^l. Too much perturbation strengthens the privacy, but makes selecting informative dimensions harder. (ii) APES performs more stable than UniS for different ϵ^l. A reasonable deduction is that outputs of Clip-Laplace is not as sensitive as Laplace to varying parameters, which is verified in Fig. 9 in Appendix.

[1] Since the true central privacy under classic Laplace Mechanisms with varied ϵ_i^l is unbounded, ϵ^c of UniS in Tab. 3 is best considered as an approximation when the ϵ_i^l of different users are very similar to each other.

Name	Distribution of $\epsilon^l = (\epsilon_1^l, ..., \epsilon_n^l)$	Clip range
Uniform1	$\mathcal{U}(0.05, 0.5)$	$[0.05, 0.5]$
Uniform2	$\mathcal{U}(0.05, 1)$	$[0.05, 1]$
Gauss1	$\mathcal{N}(0.1, 1)$	$[0.05, 0.5]$
Gauss2	$\mathcal{N}(0.2, 1)$	$[0.05, 1]$
MixGauss1	$\mathcal{N}(0.1, 1)$ 90%, $\mathcal{N}(0.5, 1)$ 10%	$[0.05, 0.5]$
MixGauss2	$\mathcal{N}(0.2, 1)$ 90%, $\mathcal{N}(1, 1)$ 10%	$[0.05, 1]$

Table 2: Distributions of Personalized LDP Budgets $\epsilon^l . \mathcal{U}, \mathcal{N}$ represents Uniform and Gaussian Distribution respectively. Clip range $[a, b]$ denotes any value g outside the range $[a, b]$ is clipped by $\max(a, g)$ or $\min(b, g)$.

Frameworks	ϵ^{ul}	ϵ^c	ϵ^{uc}	Accuracy
Non-Private	∞	∞	∞	**84.35%**
LDP-Min	392.5	0.05	40.1	56.11%
PLDP	$392.5 \sim 7850$	1	7850	77.54%
UniS	$392.5 \sim 7850$	0.069[1]	76.9	77.54%
APES	$392.5 \sim 7850$	0.057	57.6	**79.67%**
S-APES	$78.5 \sim 1570$	0.057	**25.6**	78.14%

Table 3: Privacy and Utility under Unifrom2 LDP. ϵ^{ul}: local user level, ϵ^c: central dimension level, ϵ^{uc}: central user level privacy budgets.

Privacy Amplification Effect In Fig. 2, we provide numerical evaluations for privacy amplification effect under fixed personalized LDP settings. Given dimension-level local privacy $\epsilon^l \in [0.05, 1]$, we observe following results: (i) our bounds achieve the strongest central privacy with the smallest value of ϵ^c compared to baseline bounds under the same n. The bound gets sharper especially when ϵ^l concentrates on smaller values. E.g., for the same range that $\epsilon^l \in [0.05, 1]$, most ϵ_i^l in Gauss2 are smaller than ϵ_i^l in Uniform2, which leads to lower ϵ^c. This effect comes from the EoN analysis, by which privacy contribution of each local perturbation is taken into consideration. (ii) The amplification effect gets stronger when more datapoints are shuffled, as more randomness is introduced for obfuscation. When n grows, almost all the privacy bounds ϵ^c reduce. Moreover, Fig. 7 and 8 in Appendix C demonstrate EoN gives a more obvious amplification effect when the range of ϵ^l gets larger.

Stability of Clip-Laplace Mechanism Fig. 4 shows a relatively mild impact of clip bound C on Clip-Laplace perturbation. We compare model accuracy by adopting Clip-Laplace Mechanism (CLap for short) and classic Laplace mechanism (Lap for short) in APES separately. Overall, the highest accuracy is obtained with CLap when $C = 0.1$. CLap performs well especially for large C, while Lap is only good at small C. It implies CLap may be suitable for perturbing gradients with larger norms. Fig. 9 in Appendix C explores why CLap adapts to varying parameters. The variance of CLap is more stable compared to Lap for the same level of LDP when C changes. As a price of low variance resulting from the limited output range, a larger bias is introduced (cf. Fig. 10 in Appendix C).

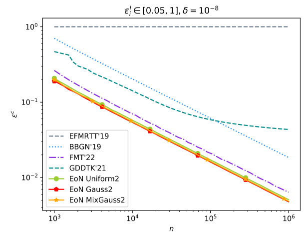

Figure 2: Privacy Bounds

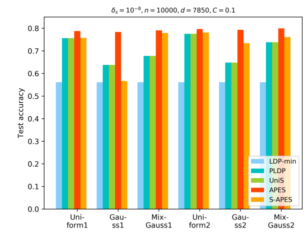

Figure 3: Impact of ϵ^l

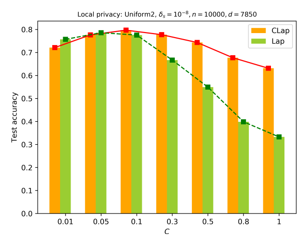

Figure 4: Impact of C

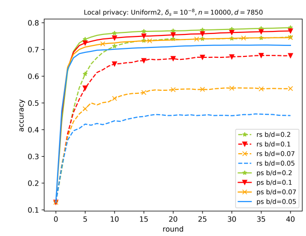

Figure 5: Impact of *ps*

Performance of Post-Sparsification We evaluate the parameters and the effectiveness of post-sparsification technique (*ps* for short) with Uniform2. (i) The trade-off between accuracy and privacy of *ps* is discussed above, while the knob is sparsification ratio b/d. In Fig. 5, the model with *ps* achieves almost optimal accuracy as APES when $b/d = 0.2$, hence only smaller ratios are evaluated. As b/d grows, fewer dimensions are uploaded and the accuracy falls. In return, privacy cost is saved. (ii) Then we observe that *ps* is more effective than random-sparification (*rs* for short) which randomly select dimensions with b/d. Specifically, model accuracy based on *rs* is lower than *ps* for all the ratios, and dramatically drops when b/d gets larger, as fewer informative dimensions are uploaded by *rs*.

6 Conclusion

This work focuses on personalized private federated learning. To balance privacy and utility, we propose privacy amplification frameworks with shuffle model under personalized LDP. Comprehensive evaluations on the public dataset confirm that our frameworks improve central privacy by reducing ϵ^C up to 66% compared to existing work with comparable or higher accuracy.

In the future, we intend to extend the work for several directions. First, we will explore the effectiveness of the work for larger models as sharing more parameters requires higher standards for both LDP performance and communication efficiency. Future improvements on sparsification techniques may alleviate the concern. Second, it may be possible to adapt the work to non-IID data distribution settings. Further efforts in optimizing personalized models without exposing local privacy parameters is needed, and more elaborate calibration for skewed gradients due to non-IID data may also be required.

Acknowledgments

We would like to thank all the anonymous reviewers for their time and efforts on our manuscript, their insightful comments and valuable suggestions helped us shape the final draft. Our work is supported by National Natural Science Foundation of China (62072460, 62076245, 62172424,62276270), Beijing Natural Science Foundation (4212022), National Science Foundation (NSF) CNS-2124104, CNS-2125530, CNS-1952192, IIS-1838200, National Institute of Health (NIH) R01LM013712, R01ES033241, UL1TR002378, Cisco Research University Award #2738379, Fundamental Research Funds for the Central Universities, and Research Funds of Renmin University of China (21XNH180).

References

Aji, A. F.; and Heafield, K. 2017. Sparse communication for distributed gradient descent. *arXiv preprint arXiv:1704.05021*.

Balle, B.; Bell, J.; Gascón, A.; and Nissim, K. 2019. The privacy blanket of the shuffle model. In *Annual International Cryptology Conference*, 638–667. Springer.

Bittau, A.; Erlingsson, Ú.; Maniatis, P.; Mironov, I.; Raghunathan, A.; Lie, D.; Rudominer, M.; Kode, U.; Tinnes, J.; and Seefeld, B. 2017. Prochlo: Strong privacy for analytics in the crowd. In *Proceedings of the 26th symposium on operating systems principles*, 441–459.

Chen, R.; Li, H.; Qin, A. K.; Kasiviswanathan, S. P.; and Jin, H. 2016. Private spatial data aggregation in the local setting. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, 289–300. IEEE.

Croft, W.; Sack, J.-R.; and Shi, W. 2022. Differential privacy via a truncated and normalized laplace mechanism. *Journal of Computer Science and Technology*, 37(2): 369–388.

Duan, J.; Ye, Q.; and Hu, H. 2022. Utility Analysis and Enhancement of LDP Mechanisms in High-Dimensional Space. In *2022 IEEE 38nd International Conference on Data Engineering (ICDE)*, 407–419. IEEE.

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

Erlingsson, Ú.; Feldman, V.; Mironov, I.; Raghunathan, A.; Talwar, K.; and Thakurta, A. 2019. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2468–2479. SIAM.

Erlingsson, Ú.; Pihur, V.; and Korolova, A. 2014. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, 1054–1067.

Feldman, V.; McMillan, A.; and Talwar, K. 2022. Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, 954–964. IEEE.

Fredrikson, M.; Jha, S.; and Ristenpart, T. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 1322–1333.

Geng, Q.; Ding, W.; Guo, R.; and Kumar, S. 2018. Truncated Laplacian mechanism for approximate differential privacy. *arXiv preprint arXiv:1810.00877*.

Girgis, A. M.; Data, D.; Diggavi, S.; Suresh, A. T.; and Kairouz, P. 2021. On the renyi differential privacy of the shuffle model. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2321–2341.

Holohan, N.; Antonatos, S.; Braghin, S.; and Mac Aonghusa, P. 2018. The bounded laplace mechanism in differential privacy. *arXiv preprint arXiv:1808.10410*.

Kairouz, P.; Oh, S.; and Viswanath, P. 2015. The composition theorem for differential privacy. In *International conference on machine learning*, 1376–1385. PMLR.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2: 429–450.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Nasr, M.; Shokri, R.; and Houmansadr, A. 2019. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, 739–753. IEEE.

Schmidt, M.; and Roux, N. L. 2013. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*.

Shen, Z.; Xia, Z.; and Yu, P. 2021. Pldp: Personalized local differential privacy for multidimensional data aggregation. *Security and Communication Networks*, 2021.

Wei, K.; Li, J.; Ding, M.; Ma, C.; Yang, H. H.; Farokhi, F.; Jin, S.; Quek, T. Q.; and Poor, H. V. 2020. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15: 3454–3469.

Yadav, C.; and Bottou, L. 2019. Cold case: The lost mnist digits. *Advances in neural information processing systems*, 32.

Yang, G.; Wang, S.; and Wang, H. 2021. Federated learning with personalized local differential privacy. In *2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS)*, 484–489. IEEE.

Ye, Q.; and Hu, H. 2020. Local differential privacy: Tools, challenges, and opportunities. In *International Conference on Web Information Systems Engineering*, 13–23. Springer.

Yu, L.; Liu, L.; Pu, C.; Gursoy, M. E.; and Truex, S. 2019. Differentially private model publishing for deep learning. In *2019 IEEE symposium on security and privacy (SP)*, 332–349. IEEE.

Zhu, L.; Liu, Z.; and Han, S. 2019. Deep leakage from gradients. *Advances in neural information processing systems*, 32.

Appendix

This Appendix includes: (i) Section A: omitted proofs of privacy theorems. (ii) Section B: omitted proofs of convergence analysis. (iii) Section C: additional experiment results.

A Privacy Analysis

In this section, we temporarily mix the symbol ϵ_i^l with ϵ_i for briefness.

Proof of Theorem 5 Let x and $x' \in \mathcal{D}^d$ be such that $||x - x'|| \leq 1$, $f(x)$ and $f(x') \in [-\Delta f/2, \Delta f/2]$, $f(x') = f(x) + a$.

$$\frac{\Pr[M(x) = z]}{\Pr[M(x') = z]}$$

$$= \prod_{k=1}^{d} (\frac{1/S \cdot \exp(-|f(x)_k - z_k|/\lambda)}{1/S' \cdot \exp(-|f(x')_k - z_k|/\lambda)})$$

$$= \prod_{k=1}^{d} (\frac{1 - \frac{1}{2}\exp(\frac{\epsilon(-A+f(x)_k+a_k)}{\Delta f}) - \frac{1}{2}\exp(\frac{\epsilon(-A-f(x)_k-a_k)}{\Delta f})}{1 - \frac{1}{2}\exp(\frac{\epsilon(-A+f(x)_k)}{\Delta f}) - \frac{1}{2}\exp(\frac{\epsilon(-A-f(x)_k)}{\Delta f})}$$

$$\cdot \exp(\frac{\epsilon|f(x)_k - f(x')_k|}{\Delta f})) \tag{6}$$

$$\leq \exp(\frac{\epsilon \cdot ||\Delta f/2 - (-\Delta f/2)||_1}{\Delta f})) \tag{7}$$

$$\leq \exp(\epsilon)$$

The inequation (7) is derived based on $\frac{\partial l}{\partial f(x)} \leq 0$ and $\frac{\partial l}{\partial f(x)} \geq 0$ where l denotes the right hand of Eq. (6).

Proof of Lemma 4 Let g_0 and $g_1 \in \mathcal{D}^d$ be such that $||g_0 - g_1|| \leq 1$, $f(g_0)$ and $f(g_1) \in [-C, C]$ and $\Delta f = 2C$, we have

$$D_N(\mu_i^{(0)}||\mu_j^{(1)}) = \frac{\Pr[M_i(g_0) = z]}{\Pr[M_j(g_1) = z]}$$

$$= \prod_{k=1}^{d} (\frac{1/(b_i S_i^{(0)}) \cdot \exp(-|f(g_0)_k - z_k|/b_i)}{1/(b_j S_j^{(1)}) \cdot \exp(-|f(g_1)_k - z_k|/b_j)})$$

$$\leq \prod_{k=1}^{d} (\frac{1 - \frac{1}{2}\exp(\frac{\epsilon_j(-A+f(g_1)_k)}{\Delta f}) - \frac{1}{2}\exp(\frac{\epsilon_j(-A-f(g_1)_k)}{\Delta f})}{1 - \frac{1}{2}\exp(\frac{\epsilon_i(-A+f(g_0)_k)}{\Delta f}) - \frac{1}{2}\exp(\frac{\epsilon_i(-A-f(g_0)_k)}{\Delta f})}$$

$$\cdot \exp(\frac{|\epsilon_j f(g_1)_k - \epsilon_i f(g_0)_k| + |z_k(\epsilon_i - \epsilon_j)|}{\Delta f}) \cdot \frac{\epsilon_i}{\epsilon_j}) \tag{8}$$

When $\epsilon_j f(g_1)_k \geq \epsilon_i f(g_0)_k$, we define $l(u) = (1 - \frac{1}{2}\exp(\frac{\epsilon(-A+u)}{\Delta f}) - \frac{1}{2}\exp(\frac{\epsilon(-A-u)}{\Delta f})) \cdot \exp(\frac{\epsilon u}{\Delta f})$, the maximum and minimum of $l(u)$ is $l(C)$ and $l(-C)$ separately when $C \leq A$. Then the right hand of Eq. (8) is bounded by:

$$D_N(\mu_i||\mu_j') \leq \frac{\epsilon_i^{(0)}}{\epsilon_j^{(1)}} \cdot \frac{1 - \frac{1}{2}\exp(\frac{\epsilon_j(-A+C)}{2C}) - \frac{1}{2}\exp(\frac{\epsilon_j(-A-C)}{2C})}{1 - \frac{1}{2}\exp(\frac{\epsilon_i(-A+C)}{2C}) - \frac{1}{2}\exp(\frac{\epsilon_i(-A-C)}{2C})}$$

$$\cdot \exp(\frac{(\epsilon_i + \epsilon_j)}{2} + \frac{A|\epsilon_i - \epsilon_j|}{2C})$$

In particular, when $A = C$, the bound is tight as follows:

$$D_N(\mu_i||\mu_j') \leq \frac{\epsilon_i}{\epsilon_j} \cdot \exp(\frac{|\epsilon_i - \epsilon_j|}{2}) \cdot \frac{\exp(\epsilon_j/2) - \exp(-\epsilon_j/2)}{\exp(\epsilon_i/2) - \exp(-3\epsilon_i/2)}$$

$$\leq \frac{\epsilon_i}{\epsilon_j} \cdot \frac{1 - e^{-\epsilon_j}}{1 - e^{-\epsilon_i}} \cdot e^{\max(\epsilon_i, \epsilon_j)} \tag{9}$$

Similarly, we get the same bound as Eq. (9) when $\epsilon_j f(g_1)_k < \epsilon_i f(g_0)_k$. Thus, the proof is completed. This Lemma allows us to transform echos in the following part.

Proof of Lemma 6 From the proof in main body, we have the intermediate result as Eq. (5), which is the following equation:

$$\frac{\Pr[P(D_T) = \mathbf{z}]}{\Pr[P(D_T') = \mathbf{z}]} = \frac{|T_0| + 1}{|T| - |T_0|} \tag{10}$$

Recall that $|T| \sim \sum_{i=2}^n \sum_{j=1}^n \mathrm{Bern}(p_{ij}/n)$ and $|T_0| \sim \mathrm{Bin}(1/2, |T|)$, by Chernoff bound and Hoeffding's inequality , T and T_0 are concentrated to certain values. Specifically, when $\sum_{i=2}^n \sum_{j=1}^n \frac{p_{ij}}{n} \geq 3\ln(4/\delta_s)$,

$$|T - \sum_{i=2}^{n} \sum_{j=1}^{n} \frac{p_{ij}}{n}| \leq (3\ln(4/\delta) \sum_{i=2}^{n} \sum_{j=1}^{n} \frac{p_{ij}}{n})^{\frac{1}{2}} \tag{11}$$

$$|T_0 - T/2| \leq (T/2\ln(4/\delta))^{\frac{1}{2}} \tag{12}$$

Based on Eq. (11) and (12), when $\sum_{i=2}^n \sum_{j=1}^n \frac{p_{ij}}{n} \geq 16\ln(4/\delta_s)$, the following equation is established with the probability $(1 - \delta_s)$.

$$\frac{|T_0| + 1}{|T| - |T_0|} \leq \frac{|T|/2 + (|T|/2\ln(4/\delta_s))^{\frac{1}{2}} + 1}{|T|/2 - (|T|/2\ln(4/\delta_s))^{\frac{1}{2}}}$$

$$\leq 1 + \frac{8(\ln(4/\delta_s))^{\frac{1}{2}}}{(\sum_{i=2}^{n} \sum_{j=1}^{n} \frac{p_{ij}}{n})^{\frac{1}{2}}} + \frac{8}{\sum_{i=2}^{n} \sum_{j=1}^{n} \frac{p_{ij}}{n}} = e^{\epsilon_0} \tag{13}$$

Hence the divergence between $P(D_T)$ and $P(D_T')$ is bounded without considering the randomness of user 1.

Next, we take the randomness of the mechanism on g_1 and g_1' with certain privacy budget ϵ_j into consideration, according to previous work (Kairouz, Oh, and Viswanath 2015; Feldman, McMillan, and Talwar 2022). Formally, we assume G (or G') as the index of the mapping element from $\tilde{g}_1$ (or $\tilde{g}_1'$), and also apply degraded privacy (cf. Lemma 3) on $\tilde{g}_1$ and $\tilde{g}_1'$. The mapping event on the $\tilde{g}_1$ (or $\tilde{g}_1'$) is defined as $U_G = \{\rho^{(1)} \text{ for } p_1, {\rho'}^{(1)} \text{ for } 1 - p_1\}$ and $U_G' = \{\rho^{(1)} \text{ for } p_1', {\rho'}^{(1)} \text{ for } 1 - p_1'\}$, where $p_1 = e^{\epsilon^*}/(1 + e^{\epsilon^*})$ and $p_1' = 1/(1 + e^{\epsilon^*})$.

Then we consider the mapping event U_s from $(n - 1)$ users of process P except for $\tilde{g}_1$ (or $\tilde{g}_1'$). To reach the mixed output **z** with the same number of $\rho^{(1)}$ or ${\rho'}^{(1)}$, U_s needs to satisfy the following situation:

$$U_s = \{U_T \text{ with } p_1, U_T' \text{ with } 1 - p_1\}$$
$$U_s' = \{U_T \text{ with } p_1', U_T' \text{ with } 1 - p_1', \}$$

which can be written as:

$$U_s = \{U_T \cup U_T' \text{ with } 1 - p_1, U_T \text{ with } 2p_1 - 1\}$$
$$U_s' = \{U_T \cup U_T' \text{ with } p_1', U_T' \text{ with } 1 - 2p_1'\}$$

Hence we can bound the divergence between U_s and U_s':

$$\frac{\Pr[P(D) = \mathbf{z}] - \delta^c}{\Pr[P(D') = \mathbf{z}]} = \frac{\Pr[U_G]\Pr[U_S|U_G] - \delta^c}{\Pr[U_{G'}']\Pr[U_S'|U_{G'}']}$$

$$= \frac{p_1(2(1 - p_1)(\frac{1}{2})\Pr[U_T \cup U_T'] + (2p_1 - 1)\Pr[U_T]) - \delta^c}{(1 - p_1')((2p_1'(\frac{1}{2})\Pr[U_T \cup U_T'] + (1 - 2p_1')\Pr[U_T']))} \tag{14}$$

For convenience, we define the probabilities as $A_1 = (\frac{1}{2})\Pr[U_T \cup U_T']$, $A_2 = \Pr[U_T]$, $A_2' = \Pr[U_T']$, $2(1-p_1) = 2p_1' = (1-p)$ and $(2p_1 - 1) = (1-2p_1') = p$. Consider the bound by Eq. (13) that $A_2 \le e^{\epsilon_0} A_2' + \delta_s$, and $A_2 \le e^{\epsilon_0} A_1 + \delta_s$ according to hockey-stick divergence, we can rewrite Eq. (14) as

$$
\begin{aligned}
&\frac{\Pr[P(D)=\mathbf{z}]-\delta^c}{\Pr[P(D')=\mathbf{z}]} = \frac{p_1((1-p)A_1+pA_2)-p\delta_s}{(1-p_1')((1-p)A_1+pA_2')} \\
&\le \frac{p_1((1-p)A_1+p(\min\{A_1,A_2'\}+e^{\epsilon_0}\min\{A_1,A_2'\}))-p\delta_s}{(1-p_1')((1-p)A_1+pA_2')} \\
&\le \frac{p_1((1-p)A_1+p(A_2'+e^{\epsilon_0}((1-p)A_1+pA_2')))-p\delta_s}{(1-p_1')((1-p)A_1+pA_2')} \\
&\le \frac{(1+p(e^{\epsilon_0}-1))((1-p)A_1+pA_2'))}{(1-p)A_1+pA_2'} \\
&= 1+p(e^{\epsilon_0}-1) \\
&= 1+\frac{e^{\epsilon^*}-1}{e^{\epsilon^*}+1}\left(\frac{8(\ln(4/\delta_s))^{1/2}}{(\sum\limits_{i=2}^{n}\sum\limits_{j=1}^{n}\frac{p_{ij}}{n})^{1/2}}+\frac{8}{\sum\limits_{i=2}^{n}\sum\limits_{j=1}^{n}\frac{p_{ij}}{n}}\right)
\end{aligned}
$$

As $\epsilon^c \le \ln(\frac{\Pr[P(D)=\mathbf{z}]-\delta^c}{\Pr[P(D')=\mathbf{z}]})$, the general bound of process P in APES is proved, and $\delta^c \le \frac{e^{\epsilon^*}-1}{e^{\epsilon^*}+1}\delta_s$.

Proof of Theorem 4 For the worst case, user 1 adpots $\max(\epsilon_j)$ as her privacy budget ϵ^*, which leads to an upper bound ϵ^c derived from Lemma 6. Therefore, $P(D)$ and $P(D')$ are (ϵ^c, δ^c)-DP:

$$
\epsilon^c \le \ln\left(1+\frac{e^{\max(\epsilon_j)}-1}{e^{\max(\epsilon_j)}+1}\left(\frac{8(\ln(4/\delta_s))^{1/2}}{(\sum\limits_{i=2}^{n}\sum\limits_{j=1}^{n}\frac{p_{ij}}{n})^{1/2}}+\frac{8}{\sum\limits_{i=2}^{n}\sum\limits_{j=1}^{n}\frac{p_{ij}}{n}}\right)\right)
$$

when $\sum_{i=2}^{n}\sum_{j=1}^{n}\frac{p_{ij}}{n} \ge 16\ln(4/\delta_s)$, $\delta_s \in [0,1]$ and $\delta^c \le \frac{e^{\max(\epsilon_j)}-1}{e^{\max(\epsilon_j)}+1}\delta_s$. Given Lemma 4, p_{ij} equals $(\frac{\epsilon_i}{\epsilon_j}\cdot\frac{1-e^{-\epsilon_j}}{1-e^{-\epsilon_i}}\cdot e^{-\max(\epsilon_i,\epsilon_j)})$.

B Convergence Analysis

To give the final convergence upper bound of the frameworks after T aggregations, we make the following analysis: (i) Show the assumptions on loss function. (ii) Derive a general form for APES with general perturbation. (iii) Consider the perturbation of Clip-Lalpalce Mechanism in bound. (iv) Bring the effect of calibration into the bound.

Step (i) Assumptions To analyze the convergence of APES, we first make following assumptions of loss function (Schmidt and Roux 2013; Li et al. 2020; Wei et al. 2020).

1. The functions $\{F_i\}$ of users u_i for $i \in [n]$ are non-convex, L-Lipschitz smooth, and there exists $L > 0$, such that $\nabla^2 F_i \succeq L\mathcal{I}$ with $\bar{\mu} := L + \mu > 0$.
2. The functions $\{F_i\}$ satisfy β-Polyak-Lojasiewicz(PL) condition, which implies $f(\tilde{w}) - f(w^*) \le \frac{1}{2\beta}\|\nabla f(\tilde{w})\|^2$, and w^* is the optimal parameters for loss function.
3. $f(\cdot)$ satisfies l-Lipschitz continuous condition.
4. w^* is the optimal solution for minimum objective function: $w^* = \min_w h(w; w_t)$. We assume $h(w; w_t)$ is α-close to minimum function: $\|\nabla h(w^*; w_t)\| \le \alpha\|\nabla h(w; w_t)\|$ where $h(w; w_t) = F_i(w) + \frac{\mu}{2}\|w - w_t\|^2$.

We also measure the dissimilarity between users in Definition (6), which is also assumed in previous works that we refers above.

Definition 6 (User Dissimilarity) *The local loss functions F_i of users are B-locally dissimilar at w if $\mathbb{E}_i[\|\nabla F_i(w)\|^2] \le \|\nabla f(w)\|^2 B^2$. Here $B(w) = \sqrt{\frac{\mathbb{E}_i[\|\nabla F_i(w)\|^2]}{\|\nabla f(w)\|^2}}$ for $\|\nabla f(w)\| \neq 0$.*

Step (ii) Proof of General Form We give the general for of convergence upper bound by yielding the bound in one aggregation and inducting to T aggregations.

Lemma 7 (Bound of Single Aggregation) *For a global function $f(w) = \mathbb{E}[F_i(w)]$ in analyzer, the expectation of difference of $f(w)$ for a single aggregation between t-th and $(t+1)$-th round is bounded as follows:*

$$
\begin{aligned}
&\mathbb{E}[f(\tilde{w}^{(t+1)}) - f(\tilde{w}^{(t)})] \\
&\le (\frac{\alpha B - 1}{\mu} + \frac{LB(\alpha+1)}{\mu\bar{\mu}} + \frac{LB^2(1+\alpha)^2}{2\bar{\mu}^2})\mathbb{E}[\|\nabla f(\tilde{w}^{(t)})\|^2] \\
&+ (\frac{1}{\mu} + \frac{BL(1+\alpha)}{\bar{\mu}})\mathbb{E}[\|\nabla f(\tilde{w}^{(t)})\|\|\eta^{(t+1)}\|] + \frac{L}{2}\mathbb{E}[\|\eta^{(t+1)}\|^2]
\end{aligned}
$$

where η denotes the perturbation term introduced by DP mechanisms, and $\eta^{(t)} = \frac{1}{n}\sum_i(\tilde{w_i^{(t)}} - w_i^{(t)})$.

Proof *According to the assumption 1, F_i is L-Lipschitz smooth,*

$$
\begin{aligned}
f(\tilde{w}^{(t+1)}) &\le f(\tilde{w}^{(t)}) + <\nabla f(\tilde{w}^{(t)}), \tilde{w}^{(t+1)} - \tilde{w}^{(t)}> \\
&+ \frac{L}{2}\|\tilde{w}^{(t+1)} - \tilde{w}^{(t)}\|^2
\end{aligned} \tag{15}
$$

where $\tilde{w}^{(t+1)} = \mathbb{E}[F_i(\tilde{w}^{(t+1)})]$, $\tilde{w}^{(t)} = \frac{1}{\eta}\sum_{i=1}^{n} w_i^{(t+1)} + \eta^{(t+1)}$ and $\eta^{(t)} = \sum_{i=1}^{n}\eta_i^{(t)}$. To bound $f(\tilde{w}^{(t+1)}) - f(\tilde{w}^{(t)})$, we have to bound $\|\tilde{w}^{(t+1)} - \tilde{w}^{(t)}\|$ and $(\tilde{w}^{(t+1)} - \tilde{w}^{(t)})$ separately.

First, we focus on $\|\tilde{w}^{(t+1)} - \tilde{w}^{(t)}\|$.

From assumption 3 with α-clossness we have

$$
\|\nabla h_k(w^{(t+1)}; \tilde{w}^{(t)})\| \le \alpha\|\nabla F_k(\tilde{w}^{(t)})\| \tag{16}
$$

Since $\nabla^2 h_k(w; \tilde{w}^{(t)}) = \bar{\mu} > 0$, h_k is $\bar{\mu}$-strong convex. Let $w_k^* = \arg\min_w h_k(w; \tilde{w}^{(t)})$, we have

$$
\begin{aligned}
&\bar{\mu}\|w_i^{(t+1)^*} - w_i^{(t+1)}\| \\
&\le \|\nabla h_i(w^{(t+1)^*}; \tilde{w}^{(t)}) - \nabla h_i(w^{(t+1)}; \tilde{w}^{(t)})\| \\
&= \|0 - \nabla h_i(w^{(t+1)}; \tilde{w}^{(t)})\| = \alpha\|\nabla F_i(\tilde{w}^{(t)})\|
\end{aligned} \tag{17}
$$

From the strong convexity of h_k, we also know that

$$
\begin{aligned}
&\bar{\mu}\|w_i^{(t+1)^*} - \tilde{w}_i^{(t)}\| \\
&\le \|\nabla h_i(w^{(t+1)^*}; \tilde{w}^{(t)}) - \nabla h_i(\tilde{w}^{(t)}; \tilde{w}^{(t)})\| \\
&= \|0 - \nabla h_i(\tilde{w}^{(t)}; \tilde{w}^{(t)})\| \\
&= \|\nabla F_i(\tilde{w}^{(t)}) + \mu(\tilde{w}^{(t)} - \tilde{w}^{(t)})\| \\
&= \|\nabla F_i(\tilde{w}^{(t)})\|
\end{aligned} \tag{18}
$$

Based on triangle inequality, B-user dissimilarity and Eq. (15)(17)(18), $\|\tilde{w}^{(t+1)} - \tilde{w}^{(t)}\|$ *is bounded:*

$$\begin{aligned}
&\|\tilde{w}^{(t+1)} - \tilde{w}^{(t)}\| \\
&\le \|\tilde{w}^{(t+1)} - \tilde{w}^{(t+1)^*}\| + \|\tilde{w}^{(t+1)^*} - w^{(t)}\| + \|\eta^{(t+1)}\| \\
&\le \mathbb{E}_i[\|w_i^{(t+1)} - w_i^{(t+1)^*}\| + \|w_i^{(t+1)^*} - w_i^{(t)}\|] + \|\eta^{(t+1)}\| \\
&\le \frac{1+\alpha}{\bar{\mu}}\mathbb{E}_i[\|\nabla F_i(\tilde{w}^{(t)})\|] + \|\eta^{(t+1)}\| \qquad (19)
\end{aligned}$$

Then, let us bound $(\tilde{w}^{(t+1)} - \tilde{w}^{(t)})$*. Differentiate* h_i*, we obtain*

$$\begin{aligned}
&\tilde{w}^{(t+1)} - \tilde{w}^{(t)} = w^{(t+1)} - \tilde{w}^{(t)} + \eta^{(t+1)} \\
&\le \frac{1}{\mu}(\mathbb{E}[\nabla h_i(w^{(t+1)}; \tilde{w}^{(t)}) - \nabla F_i(w^{(t+1)})]) + \eta^{(t+1)} \\
&= \frac{1}{\mu}(\mathbb{E}[\nabla h_i(w^{(t+1)}; \tilde{w}^{(t)}) - \nabla F_i(w^{(t+1)}) + \nabla F_i(\tilde{w}^{(t)}) \\
&- \nabla F_i(\tilde{w}^{(t)}))] + \eta^{(t+1)} \qquad (20)
\end{aligned}$$

By triangle inequality and Eq. (16)(19)*, Lemma 6 and the L-Lipschitz condition on* F_i*, We can bound the part* $(\nabla h_i(w^{(t+1)}; \tilde{w}^{(t)}) - \nabla F_i(w^{(t+1)}) + \nabla F_i(\tilde{w}^{(t)}))$ *of Eq.*(20)*.*

$$\begin{aligned}
&\|(\mathbb{E}[\nabla h_i(w^{(t+1)}; \tilde{w}^{(t)}) - \nabla F_i(w^{(t+1)}) + \nabla F_i(\tilde{w}^{(t)})]\| \\
&\le \mathbb{E}[\|\nabla h_i(w^{(t+1)}; \tilde{w}^{(t)})\| + \|\nabla F_i(w^{(t+1)}) - \nabla F_i(\tilde{w}^{(t)})\|] \\
&= (\alpha B + \frac{LB(1+\alpha)}{\bar{\mu}})\mathbb{E}[\|\nabla F_i(w^{(t+1)})\|] \\
&= (\alpha B + \frac{LB(1+\alpha)}{\bar{\mu}})\|f(\tilde{w}^{(t)})\| \qquad (21)
\end{aligned}$$

At last, substitute Eq. (19)*,* (20)*, and* (21) *into* (15)*, Lemma 7 is proved.*

$$\begin{aligned}
&\mathbb{E}[f(\tilde{w}^{(t+1)}) - f(\tilde{w}^{(t)})] \\
&\le (\frac{\alpha B - 1}{\mu} + \frac{LB(\alpha+1)}{\mu\bar{\mu}} + \frac{LB^2(1+\alpha)^2}{2\bar{\mu}^2})\mathbb{E}[\|\nabla f(\tilde{w}^{(t)})\|^2] \\
&+ (\frac{1}{\mu} + \frac{BL(1+\alpha)}{\bar{\mu}})\mathbb{E}[\|\nabla f(\tilde{w}^{(t)})\|\|\eta^{(t+1)}\|] \\
&+ \frac{L}{2}\mathbb{E}[\|\eta^{(t+1)}\|^2] \qquad (22)
\end{aligned}$$

The proof is completed.

By lemma 7, we can derive the convergence upper bound for T aggregations.

Lemma 8 (General Form of Convergence Upper Bound) *The expected decrease in the global loss function* $f(w) = \frac{1}{n}\sum_i F_i(w)$ *after T aggregations is bounded as follows:*

$$\begin{aligned}
&\mathbb{E}[f(\tilde{w}^{(T)}) - f(w^*)] \le a_1^T\mathbb{E}[f(\tilde{w}^{(0)}) - f(w^*)] \\
&+ \frac{a_1^T - 1}{a_1 - 1}(a_2\mathbb{E}[\|\eta\|] + a_3\mathbb{E}[\|\eta\|^2]) \qquad (23)
\end{aligned}$$

where $a_1 = 1 + \frac{2\beta(\alpha B-1)}{\mu} + \frac{2\beta LB(\alpha+1)}{\mu\bar{\mu}} + \frac{2\beta LB^2(1+\alpha)^2}{\bar{\mu}^2}, a_2 = l(\frac{1}{\mu} + \frac{BL(1+\alpha)}{\bar{\mu}}), a_3 = \frac{L}{2}$*.*

Proof *To bound* $\mathbb{E}[f(\tilde{w}^{(T)}) - f(w^*)]$*, we first transform it as follows:*

$$\begin{aligned}
&\mathbb{E}[f(\tilde{w}^{(T)}) - f(w^*) + f(w^*) - f(\tilde{w}^{(t)})] \\
&\le (\frac{\alpha B - 1}{\mu} + \frac{LB(\alpha+1)}{\mu\bar{\mu}} + \frac{LB^2(1+\alpha)^2}{2\bar{\mu}^2})\mathbb{E}[\|\nabla f(\tilde{w}^{(t)})\|^2] \\
&+ (\frac{1}{\mu} + \frac{BL(1+\alpha)}{\bar{\mu}})\mathbb{E}[\|\nabla f(\tilde{w}^{(t)})\|\|\eta^{(t+1)}\|] + \frac{L}{2}\mathbb{E}[\|\eta^{(t+1)}\|^2]
\end{aligned}$$

From assumption 3 and 4, we have $\mathbb{E}[f(\tilde{w}^{(t)}) - f(x^*)] \le \frac{1}{2\beta}\|\nabla f(\tilde{w}^{(t)})\|^2$ *and* $\|\nabla f(\cdot)\| \le l$*. Substract* $f(w^*)$ *from Eq.*(22) *in both sides, we have*

$$\begin{aligned}
&\mathbb{E}[f(\tilde{w}^{(t)}) - f(w^*)] \\
&\le (1 + \frac{2\beta(\alpha B - 1)}{\mu} + \frac{2\beta LB(\alpha+1)}{\mu\bar{\mu}} + \frac{2\beta LB^2(1+\alpha)^2}{\bar{\mu}^2}) \\
&\cdot \mathbb{E}[f(\tilde{w}^{(t)}) - f(w^*)] + l(\frac{1}{\mu} + \frac{BL(1+\alpha)}{\bar{\mu}})\mathbb{E}[\|\eta^{(t+1)}\|] \\
&+ \frac{L}{2}\mathbb{E}[\|\eta^{(t+1)}\|^2] \qquad (24)
\end{aligned}$$

Considering expectations of perturbation term η *is the same in each epoch, we define* $\mathbb{E}[\|\eta^{(t)}\|] = \mathbb{E}[\|\eta\|]$ *for* $t \in [0, T]$*. By Eq.* (24) *we have*

$$\begin{aligned}
&\mathbb{E}[f(\tilde{w}^{(t)}) - f(w^*)] \le a_1\mathbb{E}[f(\tilde{w}^{(t)}) - f(w^*)] + a_2\mathbb{E}[\|\eta^{(t+1)}\|] \\
&+ a_3\mathbb{E}[\|\eta^{(t+1)}\|^2] \qquad (25)
\end{aligned}$$

where $a_1 = 1 + \frac{2\beta(\alpha B-1)}{\mu} + \frac{2\beta LB(\alpha+1)}{\mu\bar{\mu}} + \frac{2\beta LB^2(1+\alpha)^2}{\bar{\mu}^2}, a_2 = l(\frac{1}{\mu} + \frac{Bh(\alpha+1)}{\bar{\mu}}), a_3 = \frac{L}{2}$*.*

The bound of T aggregations is inducted with Eq.(25)*.*

$$\begin{aligned}
&\mathbb{E}[f(\tilde{w}^{(T)}) - f(w^*)] \\
&\le a_1(...(a_1(a_1(\mathbb{E}[f(\tilde{w}^{(0)}) - f(w^*)])))) \\
&+ (a_1^0 + a_1^1 + ... + a_1^{T-1})(a_2\mathbb{E}[\|\eta\|] + a_3\mathbb{E}[\|\eta\|^2]) \\
&= a_1^T\mathbb{E}[f(\tilde{w}^{(0)}) - f(w^*)] + \sum_{t=0}^{T-1} a_1^t(a_2\mathbb{E}[\|\eta\|] + a_3\mathbb{E}[\|\eta\|^2]) \\
&= a_1^T\mathbb{E}[f(\tilde{w}^{(0)}) - f(w^*)] + \frac{a_1^T - 1}{a_1 - 1}(a_2\mathbb{E}[\|\eta\|] + a_3\mathbb{E}[\|\eta\|^2])
\end{aligned}$$

Theorem 8 is proved.

Step (iii) Proof after CLap perturbation The specific perturbation bias and variance introduced by Clip-Laplace Mechanism is considered in this section. We have $\tilde{g}_i \sim CLap(g_i, b_i, C)$ where $b_i = \Delta f/\epsilon_i$, and $g_i = \nabla f(x_i; w)$. Thus, local perturbed term η_i equals $\tilde{g}_i - g_i$, and aggregated perturbed term η is $\frac{1}{n}\sum_i \eta_i$.

Lemma 9 (Bound with Clip-Laplacian perturbation) *By personalized privacy level* $\epsilon^l = (\epsilon_1^l, ..., \epsilon_n^l)$*, the convergence bound of APES with Clip-laplacian noises is given as:*

$$\begin{aligned}
&\mathbb{E}[f(\tilde{w}^{(T)}) - f(w^*)] \le a_1^T\Delta \\
&+ \frac{a_1^T - 1}{a_1 - 1}(O(a_2C/\min(\epsilon_i^l)) + O(a_3C^2/\min(\epsilon_i^l)^2))
\end{aligned}$$

where $\mathbb{E}[f(\tilde{w}^{(0)}) - f(w^*)]$ *is denoted by* Δ*.*

Proof *Consider local perturbation for each user* u_i*,*

$$\mathbb{E}[\eta_i^2] = \frac{1}{2S}((b_i^2 - (g_i + C + b_i)^2)e^{\frac{-C-g_i}{b_i}} + (b_i^2 - (-g_i + C + b_i)^2)e^{\frac{-C+g_i}{b_i}}) + 2b_i^2 \leq \frac{(-4C^2 - 4b_iC)e^{\frac{-2C}{b_i}}}{1 - e^{\frac{-2C}{b_i}}} + 2b_i^2$$

We can obtain the bound of $\mathbb{E}[\|\eta\|^2]$.

$$\mathbb{E}[\|\eta\|^2] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}\eta_i^2] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\eta_i^2] \leq \mathbb{E}[\eta_i^2]_{max} = \frac{-4C^2 - 8C/\min(\epsilon_i^l)}{e^{\min(\epsilon_i^l)} - 1} + \frac{8C^2}{\min(\epsilon_i^l)^2} \quad (26)$$

Then $\mathbb{E}[\|\eta\|]$ *is also bounded as:*

$$\mathbb{E}[\|\eta\|]) = (\mathbb{E}[\|\eta\|^2]) - \mathbb{D}[\|\eta\|]))^{\frac{1}{2}} \leq (\mathbb{E}[\|\eta\|^2]))^{\frac{1}{2}} = (\frac{-4C^2 - 8C/\min(\epsilon_i^l)}{e^{\min(\epsilon_i^l)} - 1} + \frac{8C^2}{\min(\epsilon_i^l)^2})^{\frac{1}{2}} \quad (27)$$

Substituting equation (26) and (27) into (23), the convergence bound is provided:

$$\mathbb{E}[f(\tilde{w}^{(T)}) - f(w^*)] \leq a_1^T \mathbb{E}[f(\tilde{w}^{(0)}) - f(w^*)] + \frac{a_1^T - 1}{a_1 - 1}(a2(\frac{-4C^2 - 8C/\min(\epsilon_i^l)}{e^{\min(\epsilon_i^l)} - 1} + \frac{2C^2}{\min(\epsilon_i^l)^2})^{\frac{1}{2}}) + a3(\frac{-4C^2 - 8C/\min(\epsilon_i^l)}{e^{\min(\epsilon_i^l)} - 1} + \frac{8C^2}{\min(\epsilon_i^l)^2})))$$

The proof is completed.

Step (iv) Proof after Calibration Besides local perturbation, analyzer will also calibrate the noisy gradients centrally, which increases the convergence rate empirically. We provide the theoretical bound in this section.

Lemma 10 (Bound after Calibration) *After Clip Laplacian perturbation with* ϵ^l *and Calibration in server, the convergence rate of APES is bounded as:*

$$\mathbb{E}[f(\tilde{w}^{(T)}) - f(w^*)] \leq a_1^T \Delta + \frac{a_1^T - 1}{a_1 - 1}(O(a_2 C/\min(\epsilon_i^l)) + O(a_3 C^2/\min(\epsilon_i^l)^2))$$

Proof *In calibration, we estimate* $\bar{g}$ *by approximating* $\mathbb{E}[\tilde{g}]$ *with* $\mathbb{E}[\tilde{g}_i]$ *for* $i \in [n]$*. The bias introduced by* $\tilde{g}$ *is removed with estimated* $\bar{g}$ *mostly, we denote the residual bias for each user by* $\Delta\eta_i$*, which equals to* $\eta_i - \eta(\bar{g})$*. Note that* $\eta(\bar{g}) \in [-C, C]$*, the norm of* $\Delta\eta_i$ *is bounded:*

$$\mathbb{E}[\Delta\eta_i^2] = \mathbb{E}[\eta_i^2] + \mathbb{E}[\eta_i^2] - 2\mathbb{E}[\eta_i]\mathbb{E}[\eta_i] \leq \frac{(-8C^2 - 8b_iC)e^{\frac{-2C}{b_i}}}{1 - e^{\frac{-2C}{b_i}}} + \frac{(-8C^2 + 8b_iC)e^{\frac{-4C}{b_i}} - 8b_iCe^{\frac{-2C}{b_i}}}{(1 - e^{\frac{-2C}{b_i}})^2} + 6b_i^2$$

Therefore we obtain the bound of $\mathbb{E}[\|\Delta\eta_i\|^2]$.

$$\mathbb{E}[\|\Delta\eta_i\|^2] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}\Delta\eta_i^2] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\Delta\eta_i^2] \leq \mathbb{E}[\Delta\eta_i^2]_{max} = -\frac{8C^2}{e^{\min(\epsilon_i^l)} - 1} + \frac{8C^2}{(e^{\min(\epsilon_i^l)} - 1)^2} + \frac{24C^2}{\min(\epsilon_i^l)^2} \quad (28)$$

Then $\mathbb{E}[\|\Delta\eta_i\|]$ *is also bounded as:*

$$\mathbb{E}[\|\Delta\eta_i\|]) = (\mathbb{E}[\|\eta\|^2]) - \mathbb{D}[\|\eta\|]))^{\frac{1}{2}} \leq (\mathbb{E}[\|\eta\|^2]))^{\frac{1}{2}} = (-\frac{8C^2}{e^{\min(\epsilon_i^l)} - 1} + \frac{8C^2}{(e^{\min(\epsilon_i^l)} - 1)^2} + \frac{24C^2}{\min(\epsilon_i^l)^2})^{\frac{1}{2}} \quad (29)$$

Substituting Eq. (28) *and* (29) *into (23), the bound is provided:*

$$\mathbb{E}[f(\tilde{w}^{(T)}) - f(w^*)] \leq a_1^T \mathbb{E}[f(\tilde{w}^{(0)}) - f(w^*)] + \frac{a_1^T - 1}{a_1 - 1}(a2(-\frac{8C^2}{e^{\min(\epsilon_i^l)} - 1} + \frac{8C^2}{(e^{\min(\epsilon_i^l)} - 1)^2} + \frac{24C^2}{\min(\epsilon_i^l)^2})^{\frac{1}{2}}) + a3(-\frac{8C^2}{e^{\min(\epsilon_i^l)} - 1} + \frac{8C^2}{(e^{\min(\epsilon_i^l)} - 1)^2} + \frac{24C^2}{\min(\epsilon_i^l)^2})$$

Ideally, when estimated gradient $\bar{g}$ *is accurate, we have a better bound as follows:*

$$\mathbb{E}[f(\tilde{w}^{(T)}) - f(w^*)] \leq a_1^T \mathbb{E}[f(\tilde{w}^{(0)}) - f(w^*)] + \frac{(a_1^T - 1)a_2}{a_1 - 1} \cdot (\frac{-8C^2 - \frac{32C^2}{\min(\epsilon_i^l)}}{e^{\min(\epsilon_i^l)} - 1} - \frac{8C^2}{(e^{\min(\epsilon_i^l)} - 1)^2} + \frac{4C^2}{\min(\epsilon_i^l)^2})^{\frac{1}{2}} + a3(\frac{-8C^2 - \frac{32C^2}{\min(\epsilon_i^l)}}{e^{\min(\epsilon_i^l)} - 1} - \frac{8C^2}{(e^{\min(\epsilon_i^l)} - 1)^2} + \frac{4C^2}{\min(\epsilon_i^l)^2}) \leq a_1^T \Delta + \frac{a_1^T - 1}{a_1 - 1}(O(a_2 C/\min(\epsilon_i^l)) + O(a_3 C^2/\min(\epsilon_i^l)^2))$$

The proof is completed.

Given Lemma 10, the upper bound of convergence of the whole APES framework is obtained, as demonstrated in Theorem 2.

C Experiment Supplements

In this section, we provide error evaluation for calibration during analyzing process of APES and S-APES (Fig. 6), additional comparison on privacy amplification effects (Fig. 7, 8), and the output evaluation of Clip-Laplace Mechanism (Fig. 9, 10). At last, the model accuracy and ϵ^{uc} for all the LDP settings is demonstrated in Tab. 4.

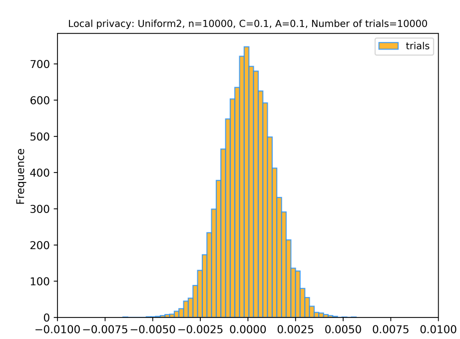

Figure 6: Error of approximation for calibrating noisy gradients. Most values of $(\tilde{g} - \mathbb{E}[\tilde{g}])$ concentrate around zero.

Framework	Uniform1		Gauss1		MixGauss1		Uniform2		Gauss2		MixGauss2	
	ϵ^{uc}	acc	ϵ^{uc}	acc	ϵ^{uc}	acc	ϵ^{uc}	acc	ϵ^{uc}	acc	ϵ^{uc}	acc
Non-Private	∞	**84.35%**	∞	**84.35%**	∞	**84.35%**	∞	**84.35%**	∞	**84.35%**	∞	**84.35%**
LDP-min	40.1	56.11%	40.1	56.11%	40.1	56.11%	40.1	56.11%	40.1	56.11%	40.1	56.11%
PLDP	3925	75.64%	3925	63.77%	3925	67.75%	7850	77.54%	7850	64.81%	7850	73.84%
UniS	17.4	75.64%	17.4	63.77%	17.5	67.75%	79.6	77.54%	76.9	64.81%	76.9	73.84%
APES	15.5	**78.77%**	15.3	**78.33%**	15.4	**79.06%**	57.6	**79.67%**	52.7	**79.29%**	53.7	**79.97%**
S-APES	**8.5**	75.7%	**8.5**	56.64%	**8.5**	77.01%	**25.6**	78.14%	**23.6**	73.4%	**23.9**	76.12%

Table 4: Comparison with baseline frameworks on user-level central privacy budget and accuracy under the all the personalized privacy budget settings. $\delta^{uc} = 3.6 \times 10^{-5}$.

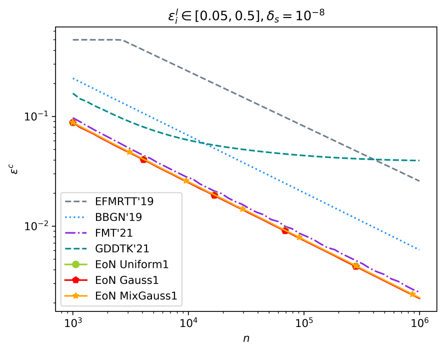

Figure 7: Privacy Bounds with smaller privacy budget range. The difference between EoN and baselines becomes smaller, but absolute value of ϵ^c of EoN is lower than larger range.

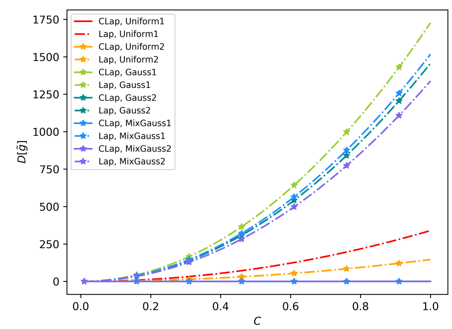

Figure 9: Variance of outputs of CLap on different C and ϵ^l. Smaller variance leads makes CLap more stable for varying parameters, which benefits from limited output ranges .

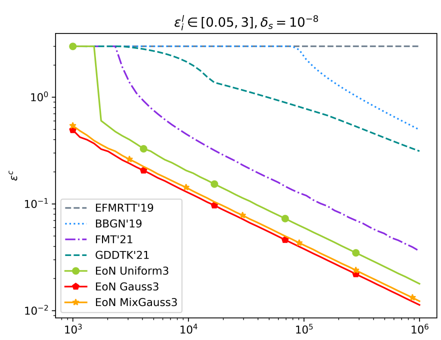

Figure 8: Privacy Bounds with larger privacy budget range. EoN gains more obvious amplification effect. Uniform3 denotes $\mathcal{U}(0.05, 3)$, Gauss3 denotes $\mathcal{N}(0.5, 1)$, MixGauss3 denotes 90% $\mathcal{N}(0.5, 1)$ and 10% $\mathcal{N}(3, 1)$. All the clip range of ϵ^l is in $[0.05, 3]$.

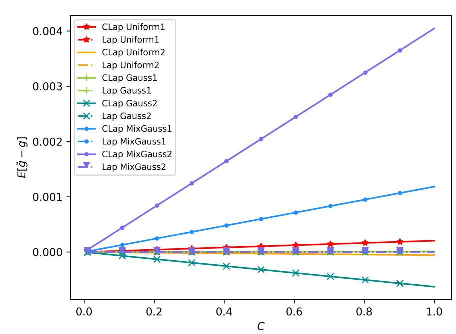

Figure 10: Bias of CLap outputs on different C and ϵ^l. Bias of CLap perturbation is larger than Lap perturbation, hence calibration is necessary.