VOTING-BASED APPROACHES FOR DIFFERENTIALLY PRIVATE FEDERATED LEARNING

Yuqing Zhu[1,2], **Xiang Yu**[2], **Yi-Hsuan Tsai**[2], **Francesco Pittaluga**[2], **Masoud Faraki**[2], **Manmohan chandraker**[2,3] and **Yu-Xiang Wang**[1]
[1]University of California, Santa Barbara
[2]NEC Laboratories America
[3]University of California, San Diego
{yuqingzhu,yuxiangw}@ucsb.edu
{xiangyu,ytsai,francescopittaluga,mfaraki,manu}@nec-labs.com

ABSTRACT

While federated learning (FL) enables distributed agents to collaboratively train a centralized model without sharing data with each other, it fails to protect users against inference attacks that mine private information from the centralized model. Thus, facilitating federated learning methods with differential privacy (DPFL) becomes attractive. Existing algorithms based on privately aggregating clipped gradients require many rounds of communication, which may not converge, and cannot scale up to large-capacity models due to explicit dimension-dependence in its added noise. In this paper, we adopt the knowledge transfer model of private learning pioneered by Papernot et al. (2017; 2018) and extend their algorithm *PATE*, as well as the recent alternative *PrivateKNN* (Zhu et al., 2020) to the federated learning setting. The key difference is that our method privately aggregates the *labels* from the agents in a *voting scheme*, instead of aggregating the *gradients*, hence avoiding the dimension dependence and achieving significant savings in communication cost. Theoretically, we show that when the margins of the voting scores are large, the agents enjoy exponentially higher accuracy and stronger (data-dependent) differential privacy guarantees on both agent-level and instance-level. Extensive experiments show that our approach significantly improves the privacy-utility trade-off over the current state-of-the-art in DPFL.

1 INTRODUCTION

With increasing ethical and legal concerns on leveraging private data, federated learning (McMahan et al., 2017) (FL) has emerged as a paradigm that allows agents to collaboratively train a centralized model without sharing local data. In this work, we consider two typical settings of federated learning: (1) Local agents are in large number, i.e., learning user behavior over many mobile devices (Hard et al., 2018). (2) Local agents are in small number with sufficient instances, i.e., learning a health related model across multiple hospitals without sharing patients' data (Huang et al., 2019).

When implemented using secure multi-party computation (SMC) (Bonawitz et al., 2017), federated learning eliminates the need for any agent to share its local data. However, it does not protect the agents or their users from inference attacks that combine the learned model with side information. Extensive studies have established that these attacks could lead to blatant reconstruction of the proprietary datasets (Dinur & Nissim, 2003) and identification of individuals (a legal liability for the participating agents) (Shokri et al., 2017). Motivated by this challenge, there had been a number of recent efforts (Truex et al., 2019b; Geyer et al., 2017; McMahan et al., 2018) in developing federated learning methods with differential privacy (DP), which is a well-established definition of privacy that provably prevents such attacks.

Among the efforts, DP-FedAvg (Geyer et al., 2017; McMahan et al., 2018) extends the NoisySGD method (Song et al., 2013; Abadi et al., 2016) to the federated learning setting by adding Gaussian noise to the clipped accumulated gradient. The recent state-of-the-art DP-FedSGD (Truex et al., 2019b) is under the same framework but with per-sample gradient clipping. A notable limitation

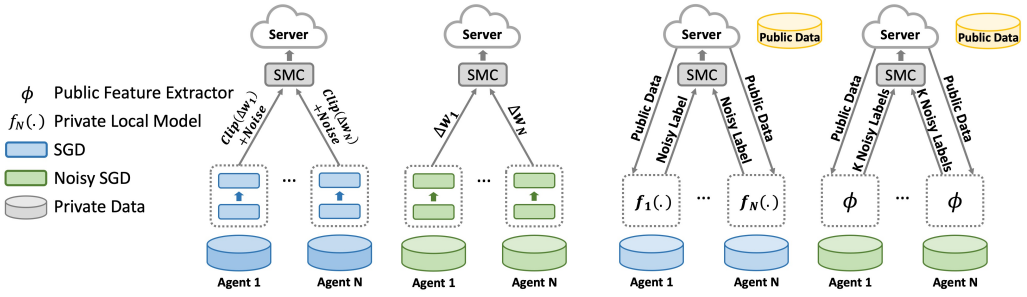


Figure 1: DP-FedAvg and *PATE-FL* are used for agent-level DP. DP-FedSGD and *Private-kNN-FL* are used for instance-level DP.

for these gradient-based methods is that they require clipping the magnitude of gradients to τ and adding noise proportional to τ to *every coordinate* of the shared global model with d parameters. The clipping and perturbation steps introduce either large bias (when τ is small) or large variance (when τ is large), which interferes the SGD convergence and makes it hard to scale up to large-capacity models. In Sec. 3, we concretely demonstrate these limitations with examples and theory. Particularly, we show that the FedAvg may fail to decrease the loss function together with gradient clipping, and DP-FedAvg requires many outer-loop iterations (i.e., many rounds of communication to synchronize model parameters) to converge under differential privacy.

To avoid the gradient clipping, we propose to conduct the aggregation over the label space, as shown to be an effective approach in standard (non-federated) learning settings, i.e., voting-based model-agnostic approaches (Papernot et al., 2017; 2018; Zhu et al., 2020). To achieve it, we relax the traditional federated learning setting to allow unlabeled public data at the server side. We also consider a more complete scenario for federated learning, where there are a large number of local agents or a limited number of local agents. The agent-level privacy as introduced in DP-FedAvg, works seamlessly with our setting having many agents. However, when there are few agents, hiding each data belonging to one specific agent becomes burdensome or unnecessary. To this end, we provide a more complete privacy notion, i.e., agent-level and instance-level. Under each of the setting, we theoretically and empirically show that the proposed label aggregation method effectively removes the sensitivity issue caused by gradient clipping or noise addition, and achieves favorable privacy-utility trade-off compared to other DPFL algorithms.

Our contributions are summarized as the following:

1. We propose a voting-based DPFL framework via label aggregation with theoretical privacy guarantees, demonstrating the advantages over gradient aggregation based DPFL methods.
2. We formalize DPFL under two notions: agent-level DP and instance-level DP regarding a large or limited amount of agents, where specifically, instance-level DP is a more appropriate DP notion when learning across limited agents.
3. We conduct theoretical analysis and show our improvement over DP-FedAvg, on gradient estimation, convergence rate, communication cost and dependence on network complexity.
4. Extensive evaluation demonstrates that our method improves the privacy-utility trade-off over randomized gradient-based approaches in both agent-level and instance-level cases.

The rest of the paper is organized as: in Sec. 2, we introduce the differential privacy settings for federate learning. In Sec. 3, we analyze in detail about the challenges for gradient-based differentially private federated learning. In Sec. 4, we demonstrate the proposed method under our newly proposed DP notions. In Sec. 5, extensive empirical experiments are presented.

2 Preliminary

In this section, we start with introducing the typical notations of federated learning and differential privacy. Then, two randomized gradient-based baselines, DP-FedAvg and DP-FedSGD, are introduced as the DPFL background.

2.1 FEDERATED LEARNING

Federated learning (McMahan et al., 2017; Bonawitz et al., 2017; Mohassel & Zhang, 2017; Smith et al., 2017) is a distributed machine learning framework that allows clients to collaboratively train a global model without sharing local data. We consider N agents, each agent i has n_i data kept locally and privately from a party-specific domain distribution $\mathcal{D}_i$. C is the number of classes. The objective is to output a global model that performs well on the target (server) distribution. Most prior works consider the target distribution as a uniform distribution over the union of all local data, which is restrictive in practice. Here we consider an agnostic federated learning scenario (Mohri et al., 2019; Peng et al., 2019b), where the server distribution $\mathcal{D}_G$ can be different from all agent distributions. In light of this, we assume each agent has access to part of unlabeled server data drawn from the target distribution $\mathcal{D}_G$.

FedAvg (McMahan et al., 2017) is a vanilla federated learning algorithm that we consider as a non-DP baseline. In this algorithm, a fraction of agents is sampled at each communication round with a probability q. Each selected agent downloads the shared global model and improves it by learning from local data using E iterations of stochastic gradient descent (SGD). We denote this local update process as an inner loop. Only the gradient is sent to the server, where it is averaged with other selected agents' gradient to improve the global model. The global model is learned after T communication rounds, where each communication round is denoted as one outer loop.

2.2 DIFFERENTIAL PRIVACY FOR FEDERATED LEARNING

Differential privacy (Dwork et al., 2006) is a quantifiable and composable definition of privacy that provides provable guarantees against identification of individuals in a private dataset.

Definition 1. *A randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ with a domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies (ϵ, δ)-differential privacy, if for any two* adjacent *datasets $D, D' \in \mathcal{D}$ and for any subset of outputs $S \subseteq \mathcal{R}$, it holds that $\Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon}\Pr[\mathcal{M}(D') \in S] + \delta$.*

The definition applies to a variety of different granularity, depending on how the *adjacent* datasets are defined, i.e., if we are to protect whether one agent participates into training, the neighboring datasets are defined by adding or removing the entire local data within that agent. It is known as agent-level (user-level) differential privacy, which has been investigated in DP-FedAvg (Geyer et al., 2017; McMahan et al., 2018). Compared to FedAvg, DP-FedAvg (Figure 1) enforces clipping of per-agent model gradient to a threshold S and adds noise to the scaled gradient before it is averaged at the server. Note that this DP notion is favored when data samples within one agent reveal the same sensitive information, e.g., cell phone agents send the same message.

However, when there are only a few agents, hiding the entire dataset from one agent becomes difficult and inappropriate. We then consider the instance-level DP, where the adjacent dataset is defined by differing one single training example. This definition is consistent with the standard non-federated learning differential privacy (Abadi et al., 2016; Bassily et al., 2014; Chaudhuri et al., 2011). Model training with instance-level DP restricts the adversary's power in detecting a specific training instance's presence or absence. DP-FedSGD (Truex et al., 2019a; Peterson et al., 2019), one such state-of-the-art for the instance-level DP, performs NoisySGD (Abadi et al., 2016) for a fixed number of iterations at each agent. The gradient updates are averaged on each communication round at the server, as shown in Figure 1.

SMC is a cryptographic technique that securely aggregates local updates before the server receives it. While SMC does not have a differential privacy guarantee, it can be combined with DP to *amplify* the privacy guarantee (Bhowmick et al., 2018; Agarwal et al., 2018; Truex et al., 2019b) against attackers that eavesdrop what sent out by each agent. Specifically, if each party adds a small independent noise to the part they contribute, SMC ensures that the attackers are only able to observe the total, even if she taps the network messages and hacks into the server. In our experiment, we assume that the aggregation is conducted by SMC for all privacy-preserving algorithms that we consider.

3 CHALLENGES FOR GRADIENT-BASED FEDERATED LEARNING

In this section, we highlight the main challenges of the conventional DPFL frameworks in terms of accuracy, convergence and communication cost. For other challenges, we refer the readers to

a survey (Kairouz et al., 2019). The details of DP-FedAvg are summarized in appendix algorithm section.

3.1 Challenge 1: Biased Gradient Estimation

Recent works (Li et al., 2018) have shown that the FedAvg may not converge well under heterogeneity (e.g., non-identical distributions). Here, we provide a simple example with a linear loss function to show that the clipping step of DP-FedAvg may raise additional challenges.

Example 2 (Gradient clipping)**.** *Let $N = 2$, each agent i has a linear loss function $\ell_i(\theta) = \theta^T x_i$. Consider the special case when $x_1 = \tau + \alpha$ and $x_2 = -\tau$, where τ is the clipping threshold. Then the global update will be* 0.

3.2 Challenge 2: Slow Convergence

Recent works (Li et al., 2019; Wang et al., 2019) have investigated the convergence rate in FL methods. Here, we draw connections to DP-FedAvg's convergence rate and demonstrate that using many outer-loop iterations (T) could have a similar convergence issue under differential privacy.

When $E = 1$ in the local update (inner loop), the FedAvg algorithm is equivalent to SGD with distributed data, which requires many rounds of communication. The appeal of FedAvg is to set E to be larger so that each agent performs E iterations to update its own parameters before synchronizing the parameters to the global model, hence reducing the number of rounds in communication. However, setting $E > 1$ may not improve convergence at all.

Now, we take a closer look at the effect of increasing E in the case of piecewise linear functions. Let the objective functions of agents, $f_1, ..., f_N$ be piecewise linear (which implies that the global objective $F = \frac{1}{N}\sum_{i=1}^{N} f_i$ is piecewise linear) and G-Lipschitz. Let η be the learning rate for individual agents. In appendix convergence section, we establish that the effect of increasing E is essentially increasing the learning rate for a large family of optimization problems with piecewise linear objective functions. It is known that for the family of $G-$Lipschitz functions supported on a B-bounded domain, any Krylov-space method [1] has a rate of convergence that is lower bounded by $O(BG/\sqrt{T})$ (Nesterov, 2003, Section 3.2.1). This indicates that the variant of FedAvg that aggregates only the loss function part of the gradient or projects only when synchronizing requires $\Omega(1/\alpha^2)$ rounds of outer loop (i.e., communication), in order to converge to an α stationary point, i.e., increasing E does *not* help, even if no noise is added.

This also says that DP-FedAvg is essentially the same as *stochastic* subgradient method in almost all locations of a piecewise linear objective function with gradient noise being $\mathcal{N}(0, \sigma^2/NI_d)$. The additional noise in DP-FedAvg imposes more challenges to the convergence. If we plan to run T rounds and achieve (ϵ, δ)-DP, we need to choose $\sigma = \frac{\eta E G \sqrt{2T\log(1.25/\delta)}}{N\epsilon}$ (see, e.g., McMahan et al., 2018, Theorem 1). which results in a convergence rate upper bound of

$$\frac{GB(\sqrt{1 + \frac{2Td\log(1.25/\delta)}{N^2\epsilon^2}})}{\sqrt{T}} = O\left(\frac{GB}{\sqrt{T}} + \frac{\sqrt{d\log(1.25/\delta)}}{N\epsilon}\right),$$

for an optimal choice of the learning rate $E\eta$.

The above bound is tight for stochastic subgradient methods, and in fact also information-theoretically optimal. The $GB/\sqrt{T}$ part of upper bound matches the information-theoretical lower bound for all methods that have access to T-calls of stochastic subgradient oracle (Agarwal et al., 2009, Theorem 1), while the second matches the information-theoretical lower bound for all (ϵ, δ)-differentially private methods on the agent level (Bassily et al., 2014, Theorem 5.3). That is, the first term indicates that there must be many rounds of communications, while the second term says that the dependence in ambient dimension d is unavoidable for DP-FedAvg. Clearly, our method also has such a dependence *in the worst case*, but it is easier for our approach to adapt to the structure that exists in the data (i.e., high consensus among voting), as we will illustrate later. In contrast, it has larger impact on DP-FedAvg, since it needs to explicitly add noises with variance $\Omega(d)$.

[1] One that outputs a solution in the subspace spanned by a sequence of subgradients.

Algorithm 1 *PATE-FL*

Input: Noise σ, global data $\mathcal{D}_G$, Q query

for i in N clients **do**
  Train local model f_i using $\mathcal{D}_i$
end for
for $t = 0, 1, ..., Q$, pick $x_t \in \mathcal{D}_G$ **do**
  for each agent i in $1, ..., N$ **do**
    $\tilde{f}_i(x_t) = f_i(x_t) + \mathcal{N}(0, \frac{\sigma^2}{N} I_C)$.
  end for
  $\tilde{y}_t = \arg\max_{y \in \{1,...,C\}} [\sum_{i=1}^{N} \tilde{f}_i(x_t)]_y$
end for
Train a global model θ using $(x_t, \tilde{y}_t)_{t=1}^{Q}$

Algorithm 2 *Private-kNN-FL*

Input: Noise σ, global data $\mathcal{D}_G$, Q query

for $t = 0, 1, ..., Q$, pick $x_t \in \mathcal{D}_G$ **do**
  for each agent i in $1, ..., N$ **do**
    Apply ϕ on $\mathcal{D}_i$ and x_t
    $y_1, ..., y_k \leftarrow$ top-k closest labels
    $\tilde{f}_i(x_t) = \frac{1}{k}(\sum_{j=1}^{k} y_j) + \mathcal{N}(0, \frac{\sigma^2}{N} I_C)$
  end for
  $\tilde{y}_t = \arg\max_{y \in \{1,...,C\}} [\sum_{i=1}^{N} \tilde{f}_i(x_t)]_y$
end for
Train a global model θ using $(x_t, \tilde{y}_t)_{t=1}^{Q}$

Another observation is that when N is small, no DP method with reasonable ϵ, δ parameters is able to achieve high accuracy. This partially motivates us to consider the other regime that deals with instance-level DP.

3.3 Other Challenges

Expensive Communication Cost: Up-stream communication cost (Konečnỳ et al., 2016), i.e., total transmitted updates from local agent to server, is another key concern in FL. For FedAvg, our convergence analysis suggests that increasing E does not speed up the convergence. A high communication cost is expected till the model converges. CpSGD (Agarwal et al., 2018) is another DPFL method, aiming at reducing the communication cost by gradient quantization with binomial noise. However, sampling from binomial distribution can be difficult on devices, which prevents it from being practical in real-world scenarios.

Network Complexity: DP-FedAvg requires to clip gradient magnitude to τ at each coordinate in parameters, which is hard to scale up to large models, as the noise level increases proportional to the network capacity. To address this issue, recent works apply delicate clipping strategies (McMahan et al., 2018; Geyer et al., 2017) and reduce data dimension with PCA (Abadi et al., 2016). In this work, we propose to avoid such dimension dependence and empirically investigate how network architecture affects performance in various DPFL approaches.

4 Algorithm

We assume there are unlabeled data drawn from $\mathcal{D}_G$ at the server, which is public and accessible from any agent. The goal is to design an (ϵ, δ)-DP algorithm (either on the agent-level or instance-level) that outputs pseudo-labels for a subset of server's unlabeled data. Then a global model is trained in a semi-supervised way, using pseudo-labeled and unlabeled data.

PATE-FL In *PATE-FL* (Algorithm 1), each agent i trains a local "teacher" model f_i using its own private local data. For each "student" query x_t, every agent adds Gaussian Noise to her prediction (i.e., C-dim histogram), aggregates their noisy predictions via SMC and the label with the most votes is returned to the server as the "pseudo-label" of x_t. Similar to the original PATE, the idea behind the privacy guarantee is that by adding or removing any instance, it can *change* at most one agent's prediction. The same argument also naturally applies to *adding or removing one agent*. In fact we gain a factor of 2 in the stronger agent-level DP due to a smaller sensitivity (see the proof for details)! Another important difference is that in the original PATE, the teachers are trained on random splits of the data, while in our case, the agents are naturally present with different distributions. We propose to optionally use domain adaptation techniques to mitigate these differences when training the "teachers".

Private-kNN-FL Next we present how the teachers f_i is constructed in *Private-kNN-FL* method (see Algorithm 2). Each agent has a data-independent feature extractor ϕ. For every unlabeled query x_t, agent i finds the k_i nearest neighbor to x_t from its local data by measuring their Euclidean distance in the feature space $\mathcal{R}^{d_\phi}$ and $f_i(x_t)$ outputs the frequency vector of the votes for these

nearest neighbors. Subsequently, $f_i(x_t)$ from all agents are privately aggregated with the argmax of the noisy voting scores returned to the server.

Different from the original Private-kNN (Zhu et al., 2020), we apply kNN on each agent's local data instead of the entire private dataset. This distinction allows us to receive up to kN neighbors while bounding the contribution of individual agents by k. Comparing to PATE-FL, this approach enjoys a stronger instance-level DP guarantee since the sensitivity from adding or removing one instance is a factor of $k/2$ times smaller than that of the agent-level.

4.1 Privacy Analysis

We provide our privacy analysis based on Renyi differential privacy (RDP) (Mironov, 2017). RDP inherits and generalizes the information-theoretical properties of DP, and has been used for privacy analysis in DP-FedAvg. We defer the background about RDP, its connection to DP and all proofs of our technical results to the appendix RDP section.

Theorem 3 (Privacy guarantee)**.** *Let PATE-FL and Private-kNN-FL answer Q queries with noise scale σ. For agent-level protection, both algorithms guarantee $(\alpha, Q\alpha/(2\sigma^2))$-RDP for all $\alpha \geq 1$. For instance-level protection, PATE-FL and Private-kNN-FL obey $(\alpha, Q\alpha/\sigma^2)$ and $(\alpha, Q\alpha/(k\sigma^2))$-RDP respectively.*

This theorem says that both algorithms achieve agent-level and instance-level differential privacy. With the same noise injection to the agent's output, *Private-kNN-FL* enjoys a *stronger* instance-level DP (by a factor of $k/2$) compared to its agent-level guarantee, while *PATE-FL*'s instance-level DP is *weaker* by a factor of 2.

Improved accuracy and privacy with large margin: Let $f_1, ..., f_N : \mathcal{X} \to \triangle^{C-1}$ where $\triangle^{C-1}$ denotes the probability simplex — the soft-label space. Note that both algorithms we propose can be viewed as voting of these local agents, which output a probability distribution in $\triangle^{C-1}$. First, let us define the margin parameter $\gamma(x)$ that measures the difference between the largest and second largest coordinate of $\frac{1}{N}\sum_{i=1}^N f_i(x)$.

Lemma 4. *Conditioning on the teachers, for each public data point x, the noise added to each coordinate of $\frac{1}{N}\sum_{i=1}^N f_i(x)$ is drawn from $\mathcal{N}(0, \sigma^2/N^2)$, then with probability $\geq 1 - C\exp\{-N^2\gamma(x)^2/8\sigma^2\}$, the privately released label matches the majority vote without noise.*

The proof (in Appendix) is a straightforward application of Gaussian tail bounds and a union bound over C coordinates. This lemma implies that for all public data point x such that $\gamma(x) \geq \frac{2\sqrt{2\log(C/\delta)}}{N}$, the output label matches noiseless majority votes with probability at least $1 - \delta$.

Next we show that for those data point x such that $\gamma(x)$ is large, the privacy loss for releasing $\arg\max_j[\frac{1}{N}\sum_{i=1}^N f_i(x)]_j$ is exponentially smaller.

Theorem 5. *For each public data point x, the mechanism that releases $\arg\max_j[\frac{1}{N}\sum_{i=1}^N f_i(x) + \mathcal{N}(0, (\sigma^2/N^2)I_C)]_j$ obeys (α, ϵ)-data-dependent-RDP, where*

$$\epsilon \leq Ce^{-\frac{N^2\gamma(x)^2}{8\sigma^2}} + \frac{1}{\alpha - 1}\log\left(1 + e^{\frac{(2\alpha-1)\sigma^2}{2s^2} - \frac{N^2\gamma(x)^2}{16\sigma^2} + \log C}\right),$$

where $s = 1$ for PATE-FL, and $s = 1/k$ for Private-KNN-FL.

This bound implies that when the margin of the voting scores is large, the agents enjoy exponentially stronger (data-dependent) differential privacy guarantees in both agent-level and instance-level. In other words, our proposed methods avoid the dependence on model dimension d that are inherited in DP-FedAvg and can release models for free privacy cost when a high consensus among votes from local agents.

4.2 Communication Cost

Finally, regarding the communication issue, our proposed methods are *parallel* as each agent work independently without any synchronization. Overall, we reduce the up-stream communication cost from $d \cdot T$ floats (model size times T rounds) to $C \cdot Q$ floats in one round.

Datasets	# Agents	Methods	Accuracy	$\epsilon \downarrow$
SVHN, MNIST $\rightarrow$ USPS	200	FedAvg	$87.6 \pm 0.1\%$	-
		DP-FedAvg	$76.3 \pm 0.3\%$	3.7
		PATE-FL (Ours)	$83.8 \pm 0.2\%$	3.6
		PATE-FL+DA (Ours)	$92.5 \pm 0.2\%$	2.8
CelebA	300	FedAvg	$84.9 \pm 0.1\%$	-
		DP-FedAvg	$83.2 \pm 0.1\%$	4.0
		PATE-FL (Ours)	$85.0 \pm 0.1\%$	4.0

Table 1: **Agent-level DP Evaluation.** We compare the state-of-the-art DPFL methods with ours on the Digit and CelebA datasets. For (ϵ, δ)-DP setting, we set $\delta = 10^{-3}$ across all the methods.

5 Experimental Results

We verify our *PATE-FL* for agent-level DP on Digit (LeCun et al., 1998; Netzer et al., 2011) and CelebA (Liu et al., 2015). Then, we evaluate *Private-kNN-FL* on Office-Caltech10 (Gong et al., 2012) and DomainNet (Peng et al., 2019a) for instance-level DP. Five independent rounds of experiments are conducted to report mean accuracy and its standard deviation. To reduce the privacy budget on global model training, we apply semi-supervised training on Digit datasets, while for other datasets, the global model is trained using labeled data only. We defer the experimental details to appendix.

5.1 Evaluation on Agent-level DP

Digit Datasets Evaluation: MNIST, SVHN and USPS together as Digit datasets, is a controlled setting to mimic the real case, where distribution of agent-to-server or agent-to-agent can be different. We simulate 140 agents using SVHN with 3000 records each and 60 agents using MNIST with 1000 records each. USPS serves as unlabeled public data, and 3000 records can be accessed by the local agents.

PATE-FL+DA refers to our *PATE-FL* framework, with each agent model trained using the domain adaptation (DA) technique (Ganin et al., 2016). We set the noise scale $\sigma = 25$ for *PATE-FL* and $\sigma = 30$ for *PATE-FL+DA*. The noise is set larger for *PATE-FL+DA* because there is a stronger consensus among agent predictions, allowing larger noise level without sacrificing accuracy. Both *PATE-FL* and *PATE-FL+DA* privately pseudo label 500 USPS data. Following PATE (Papernot et al., 2018), a semi-supervised model is trained using both labeled and pseudo-labeled data via virtual adversarial training (VAT) (Miyato et al., 2018). For DP-FedAvg, we clip the local update at each communication round to $S = 0.08$ and set the noise scale as $\sigma = 0.06$. At each communication round, we randomly sample agents with probability $q = 0.05$. We apply ImageNet (Deng et al., 2009) pre-trained AlexNet (Krizhevsky et al., 2012) for all Digit experiments.

In Table 1, our methods *PATE-FL* and *PATE-FL+DA* are compared to private and non-private baselines. We observe: (1) When the privacy cost ϵ of DP-FedAvg and *PATE-FL* is close, our method significantly improves the accuracy from 76.3% to 83.8%. (2) The further improved accuracy 92.5% of *PATE-FL+DA* demonstrates that our framework can orthogonally benefit from DA techniques, where it is highly uncertain yet for the gradient-based methods.

CelebA Dataset Evaluation: CelebA is a 220k face attribute dataset with 40 attributes defined. 300 agents are designed with partitioned training data. We split 600 unlabeled data at server, and the rest 59,400 images are for testing. Detailed settings are referred to appendix. Consistent to Digits dataset, our method achieves clear performance gain by 1.8% compared to DP-FedAvg while maintaining the same privacy cost.

5.2 Evaluation on Instance-level DP

When agents are few, preserving privacy across agents becomes hard and meaningless. We then focus on preserving each instance's privacy, a.k.a instance-level DP. FedAvg is non-private baseline.

Office-Caltech Evaluation: Office-Caltech consists of data from four domains: Caltech (C), Amazon (A), Webcam(W) and DSLR (D). We pick one domain as server each time and the rest ones are for local agents (e.g., in $A, C, D \rightarrow W$, Webcam is treated as the server). We split 70% data from

Network	Methods	$A, C, D \rightarrow W$ (Acc.)	$\epsilon \downarrow$	$A, C, W \rightarrow D$(Acc.)	$\epsilon \downarrow$
AlexNet	FedAvg	$90.5 \pm 0.1\%$	-	$96.8 \pm 0.1\%$	-
	DP-FedAvg	$28.1 \pm 0.7\%$	46.6	$48.2 \pm 0.8\%$	47.1
	DP-FedSGD	$32.6 \pm 0.9\%$	4.1	$48.3 \pm 0.9\%$	4.0
	DP-FedSGD	$75.2 \pm 0.5\%$	12.4	$83.7 \pm 0.6\%$	7.9
	Private-kNN-FL (Ours)	$75.4 \pm 0.3\%$	3.9	$84.3 \pm 0.3\%$	3.7
ResNet50	FedAvg	$96.5 \pm 0.1\%$	-	$97.8 \pm 0.1\%$	-
	DP-FedSGD	$25.8 \pm 0.6\%$	4.0	$42.7 \pm 0.5\%$	3.9
	Private-kNN-FL (Ours)	$86.3 \pm 0.4\%$	2.8	$91.9 \pm 0.2\%$	2.0

Table 2: **Instance-level DP on Office-Caltech using different backbones.**

	Clipart (Acc.)	$\epsilon \downarrow$	Painting (Acc.)	$\epsilon \downarrow$	Real (Acc.)	$\epsilon \downarrow$
FedAvg	$81.8 \pm 0.2\%$	-	$72.8 \pm 0.2\%$	-	$82.0 \pm 0.3\%$	-
DP-FedSGD	$44.2 \pm 0.2\%$	4.4	$42.6 \pm 0.3\%$	4.6	$39.1 \pm 0.6\%$	4.3
DP-FedSGD	$55.6 \pm 0.2\%$	11.6	$60.0 \pm 0.6\%$	14.6	$55.1 \pm 0.6\%$	11.9
Private-kNN-FL (Ours)	$55.8 \pm 0.6\%$	4.4	$61.2 \pm 0.8\%$	4.7	$55.5 \pm 0.7\%$	4.2

Table 3: **Instance-level DP on DomainNet.** We compare our method with DP-FedSGD and the non-private baseline FedAvg. Total number of local agents is 5. We set $\delta = 10^{-4}$.

the server domain as public available unlabeled data while the remaining 30% data is used for testing. For *Private-kNN-FL*, we instantiate the data-independent feature extractor using the network backbone without the classifier layer. Both AlexNet and Resnet50 are Imagenet pre-trained. We set $\sigma = 15$ for *Private-kNN-FL* with AlexNet and $\sigma = 25$ for ResNet50. We use all unlabeled data for privacy-preserving queries.

We observe in Table 2, DP-FedSGD degrades when backbone changes from light load AlexNet to heavy load ResNet50, while ours is improved by 10%. It is because larger model capacity leads to more sensitive response to gradient clipping or noise injection. In contrast, our *Private-kNN-FL* avoids the gradient operation by label aggregation and can still benefit from the larger model capacity. Again, our method achieves consistently better utility-privacy trade-off as maintaining same privacy cost and can achieve significantly better utility, or maintaining same utility and can achieve much low privacy cost.

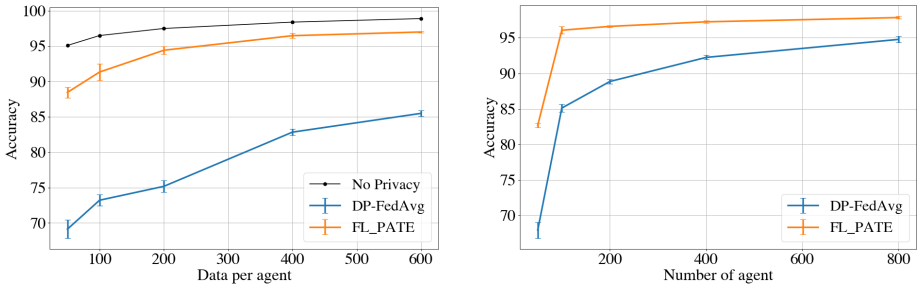
DomainNet Evaluation: DomainNet contains 0.6 million images of 345 categories, ranging from six domains: Clipart, Painting, Real, Quickdraw, Infograph and Sketch. As a challenging dataset even for non-private setting (Peng et al., 2019b), we only consider seven fruit classes (apple, banana, grapes, strawberry, watermelon, pear, pineapple) for demonstration. Large domain shift exists between infograph/quickdraw and other domains (Peng et al., 2019b). Thus we only report results on cases where servers are chosen from Clipart, Painting and Real. Five domain data are assigned to five agents respectively. 70% of the left domain data is split for server and 30% rest for testing.

Table 3 compares our *Private-kNN-FL* method with DP-FedSGD. We observe that when the privacy cost ϵ is aligned close, our method outperforms DP-FedSGD by more than 10% in accuracy gain across all the three cases. When the accuracy is aligned close, our method saves more than 60% privacy cost, showing consistent advantage over DP-FedSGD.

5.3 Ablation Study

In this section, we investigate the agent-level privacy-utility trade-off with respect to the number of agents and the volume of local data. MNIST is utilized for generality and simplicity. We randomly pick 1000 testing data as the unlabeled server data and the remaining 9000 data for testing. We adopt the model structure proposed in (Abadi et al., 2016) for both of our methods.

Effect of *Data per Agent*: We fix the number of agent to 100 and range the number of data per agent from $\{50, 100, 200, 400, 600\}$. By only relaxing the "data per agent" factor, we fairly tune the other privacy parameters for all the methods to its maximized performance. In Figure 2 (a), as "data per agent" increases, all the methods improves as the overall dataset volume increases. Our method achieves consistently higher accuracy over DP-FedAvg. The failure cases for both methods are when "data per agent" is below 50, which cannot ensure the well-trained local agent models. Label aggregation over such weak local models results in failure or sub-optimal performance.



(a) Effect on the amount of data per agent (b) Effect on the number of agent

Figure 2: Ablation study on amount of data per agent and number of agent. The non-private model, FedAvg, is served as the performance upper bound.

Effect of *Number of Agents*: In Figure 2 (b), we vary $N \in \{50, 100, 200, 400, 800\}$ and set overall privacy budget fixed as $\epsilon = 5, \delta = 10^{-3}$. Following (Geyer et al., 2017), each agent has exactly 600 data, where data samples are duplicated when $N \in \{200, 400, 800\}$. We conduct grid search for each method to obtain optimal hyper-parameters. Our method shows clear performance advantage over DP-FedAvg. We also see DP-FedAvg gradually approaches our method as the number of agents increases.

6 Conclusions

In this work, we propose voting-based approaches for differentially private federated learning (DPFL) under two privacy regimes: agent-level and instance-level. We substantially investigate the real-world challenges of DPFL and demonstrate the advantages of our methods over gradient aggregation-based DPFL methods on utility, convergence, reliance on network capacity, and communication cost. Extensive empirical evaluation shows that our methods improve the privacy-utility trade-off in both privacy regimes.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016. 1, 3, 5, 8, 15

Alekh Agarwal, Martin J Wainwright, Peter L Bartlett, and Pradeep K Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, pp. 1–9, 2009. 4

Naman Agarwal, Ananda Theertha Suresh, Felix Xinnan X Yu, Sanjiv Kumar, and Brendan McMahan. cpsgd: Communication-efficient and differentially-private distributed sgd. In *Advances in Neural Information Processing Systems*, pp. 7564–7575, 2018. 3, 5

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, pp. 464–473, 2014. 3, 4

Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*, 2018. 3

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191, 2017. 1, 3

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3):1069–1109, 2011. 3

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009. 7

Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 202–210, 2003. 1

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006. 3

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 7

Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017. 1, 3, 5, 9, 12, 13, 15

Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2066–2073. IEEE, 2012. 7

Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018. 1

Li Huang, Andrew L Shea, Huining Qian, Aditya Masurkar, Hao Deng, and Dianbo Liu. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *Journal of biomedical informatics*, 99:103291, 2019. 1

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International Conference on Machine Learning (ICML-15)*, 2015. 12

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019. 4

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016. 5

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012. 7

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 7

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018. 4

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019. 4

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015. 7

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017. 1, 3

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. *ICLR*, 2018. 1, 3, 4, 5

Ilya Mironov. Rényi differential privacy. In *Computer Security Foundations Symposium (CSF), 2017 IEEE 30th*, pp. 263–275. IEEE, 2017. 6, 12

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 7

Payman Mohassel and Yupeng Zhang. Secureml: A system for scalable privacy-preserving machine learning. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 19–38. IEEE, 2017. 3

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. *arXiv preprint arXiv:1902.00146*, 2019. 3

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2003. 4

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 7

Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *International Conference on Learning Representations (ICLR-17)*, 2017. 1, 2

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. In *International Conference on Learning Representations (ICLR-18)*, 2018. 1, 2, 7, 14

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1406–1415, 2019a. 7

Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko. Federated adversarial domain adaptation. *arXiv preprint arXiv:1911.02054*, 2019b. 3, 8

Daniel Peterson, Pallika Kanani, and Virendra J Marathe. Private federated learning with domain adaptation. *arXiv preprint arXiv:1912.06733*, 2019. 3

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (SP)*, pp. 3–18. IEEE, 2017. 1

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, pp. 4424–4434, 2017. 3

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *Conference on Signal and Information Processing*, 2013. 1

Stacey Truex, Nathalie Baracaldo, Ali Anwar, Thomas Steinke, Heiko Ludwig, Rui Zhang, and Yi Zhou. A hybrid approach to privacy-preserving federated learning. In *AISec*, 2019a. 3

Stacey Truex, Nathalie Baracaldo, Ali Anwar, Thomas Steinke, Heiko Ludwig, Rui Zhang, and Yi Zhou. A hybrid approach to privacy-preserving federated learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, pp. 1–11, 2019b. 1, 3

Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K Leung, Christian Makaya, Ting He, and Kevin Chan. Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications*, 37(6):1205–1221, 2019. 4

Yuqing Zhu, Xiang Yu, Manmohan Chandraker, and Yu-Xiang Wang. Private-knn: Practical differential privacy for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 6, 15

A Definitions

Definition 6. *A function ℓ is Lipschitz continuous with constant $G > 0$, if*

$$|\ell(x) - \ell(y)| \leq G||x - y||_2$$

for all x, y.

B Other properties of differential privacy

Definition 7 (Renyi Differential Privacy (Mironov, 2017))**.** *We say a randomized algorithm $\mathcal{M}$ is $(\alpha, \epsilon(\alpha))$-RDP with order $\alpha \geq 1$ if for neighboring datasets D, D',*

$$\mathbb{D}_\alpha(\mathcal{M}(D)||\mathcal{M}(D')) := \frac{1}{\alpha - 1} \log \mathbb{E}_{o\sim\mathcal{M}(D')}\left[\left(\frac{\Pr[\mathcal{M}(D) = o]}{\Pr[\mathcal{M}(D') = o]}\right)^\alpha\right] \leq \epsilon(\alpha).$$

RDP inherits and generalizes the information-theoretical properties of DP.

Lemma 8 (Selected Properties of RDP (Mironov, 2017))**.** *If $\mathcal{M}$ obey $\epsilon_\mathcal{M}(\cdot)$-RDP, then*

1. *[Indistinguishability] For any measurable set $S \subset Range(\mathcal{M})$, and any neighboring D, D'*

$$e^{-\epsilon(\alpha)}\Pr[\mathcal{M}(D') \in S]^{\frac{\alpha}{\alpha-1}} \leq \Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon(\alpha)}\Pr[\mathcal{M}(D') \in S]^{\frac{\alpha-1}{\alpha}}.$$

2. *[Post-processing] For all function f, $\epsilon_{f\circ\mathcal{M}}(\cdot) \leq \epsilon_\mathcal{M}(\cdot)$.*

3. *[Composition] $\epsilon_{(\mathcal{M}_1,\mathcal{M}_2)}(\cdot) = \epsilon_{\mathcal{M}_1}(\cdot) + \epsilon_{\mathcal{M}_2}(\cdot)$.*

This composition rule often allows for tighter calculations of (ϵ, δ)-DP for the composed mechanism than the strong composition theorem in (Kairouz et al., 2015). Moreover, we can covert RDP to (ϵ, δ)-DP for any $\delta > 0$ using:

Lemma 9 (From RDP to DP)**.** *If a randomized algorithm $\mathcal{M}$ satisfies $(\alpha, \epsilon(\alpha))$-RDP, then $\mathcal{M}$ also satisfies $(\epsilon(\alpha) + \frac{\log(1/\delta)}{\alpha-1}, \delta)$-DP for any $\delta \in (0, 1)$.*

C DP-FedAvg Algorithms

In this section, we provide two kinds of DP-FedAvg algorithms. Algorithm 3 is from (Geyer et al., 2017), where noise is added at the server. To prevent the adversary from tapping the network messages, we extend Algorithm 3 to Algorithm 5. Both algorithms ensure the same agent-level DP guarantees. When we refer to DP-FedAvg, it corresponds to the version in Algorithm 5.

D More Discussions of Challenges for Gradient-Based FL

Proposition 10. Let the objective function of agents $f_1, ..., f_N$ obeys that f_i is piecewise linear (which implies that the global objective $F = \frac{1}{N}\sum_{i=1}^N f_i$ is piecewise linear) and G-Lipschitz. Let η be the learning rate taken by individual agents. Then the outer loop FedAvg update is equivalent to $\theta^+ = \theta - E\eta g$ for some $g \in \mathbb{R}^d$, where (a) $g = \nabla F(\theta)$ if θ is in the ν interior of the linear region of $f_1, ..., f_N$ and $E < \nu/(\eta G)$; (2) g is a Clarke-subgradient [2] of F at θ, if θ is on the boundary of at least two linear regions and at least ν away in Euclidean distance from another boundary and $E < \nu/(\eta G)$; (c) otherwise, we have that $\|g - \nabla F(\theta)\|_2 \leq E\eta G$. Moreover, statement (c) is true even if we drop the piecewise linear assumption.

Proof. For the Statement (a), observe that for all θ' such that $\|\theta' - \theta\| \leq \nu$ neighborhood, we have that $\nabla f_i(\theta') = \nabla f_i(\theta)$. When $E < \nu/(\eta G)$, the cumulative gradients of agent i is equal to $E\nabla f_i(\theta)$. For Statement (b), notice that the Clarke subdifferential at θ is the convex hull of the

[2] Clarke-subgradient is a generalization of the subgradient to non-convex functions. It reduces to the standard (Moreau) subgradient when F is convex.

Algorithm 3 Standard DP-FedAvg (Geyer et al., 2017)

Input: Agent selection probability $q \in (0,1)$, noise scale σ, clipping threshold S.

Initialize global model θ^0
for $t = 0, 1, 2, ..., T$ **do**
  $m_t \leftarrow$ Sample agents with q
  for each agent i in parallel **do**
    $\triangle_i^t = \text{LocalUpdate}(i, \theta^t, t)$
  end for
  $\triangle^t = \sum_{i=0}^{m_t} \left(\triangle_i{}^t / max(1, \frac{||\triangle_i{}^t||_2}{S}) \right)$
  $\theta^{t+1} = \theta^t + \frac{1}{m_t}(\triangle^t + \mathcal{N}(0, \sigma^2 S^2))$
end for

Algorithm 4 LocalUpdate(i, θ^0, t)

$\theta \leftarrow \theta^0$
$\theta \leftarrow E$ iterations SGD from θ^0
return update $\triangle_i{}^t = \theta - \theta^0$

Algorithm 5 DP-FedAvg (extend)

Input: Agent selection probability q, noise scale σ, clipping threshold S.

Initialize global model θ^0
for $t = 0, 1, 2, ..., T$ **do**
  $m_t \leftarrow$ Sample agents with q
  for each agent i in parallel **do**
    $\triangle_i^t = \text{NoisyUpdate}(i, \theta^t, t, \sigma, m_t)$
  end for
  $\triangle^t = \sum_{i=0}^{m_t} \triangle_i{}^t$
  $\theta^{t+1} = \theta^t + \frac{1}{m_t}\triangle^t$
end for

Algorithm 6 NoisyUpdate$(i, \theta^0, t, \sigma, m_t)$

$\theta \leftarrow \theta^0$
$\theta \leftarrow E$ iterations SGD from θ^0
$\triangle_i{}^t = (\theta - \theta^0) / \max(1, \frac{||\theta - \theta^0||_2}{S})$
return update $\triangle_i{}^t + \mathcal{N}(0, \sigma^2 S^2 / m_t)$

one-sided gradient, thus as we move along the negative gradient direction in the inner loop, we enter and remains in the linear region. Thus the update direction is

$$\frac{1}{N}\left(\sum_{i \text{ s.t. } f_i \text{ is differentiable at } \theta} E\eta \nabla f_i(\theta) + \sum_{i \text{ s.t. } f_i \text{ is not differentiable at } \theta} \eta g_i + (E-1)\nabla f_i(\theta - \eta g_i) \right)$$

for all g_i such that it is a Clarke-subgradient of f_i it can be written as a convex combination. The proof is complete by observing that the $1/N \sum_i$ is also a convex combination and by multiplying and dividing by E. Statement (c) is a straightforward application of the Lipschitz property which says that E steps can at most get you away for ηEG and clearly piecewise linear assumption is not required. □

This proposition says that in almost all θ, increasing E has the effect of increasing the learning rate of the subgradient "descent" method for piecewise linear objective functions; and increasing the learning rate of an approximate gradient method in general for Lipschitz objective functions. It is known that for the family of $G-$Lipschitz function supported on a B-bounded domain, any Krylov-space method [3] has a rate of convergence that is lower bounded by $O(BG/\sqrt{T})$ if running for T iterations. A close inspection of the lower bound construction reveals that the worst-case problem is $\min_{\theta \in \mathbb{R}^T} \max_i \theta_i + \|\theta\|^2$, namely, a regularized piecewise linear function. This is saying that the variant of FedAvg that aggregates only the loss-function part of the gradient or projects only when synchronizing essentially requires $\Omega(1/\alpha^2)$ rounds of outer loop iterations (thus communication) in order to converge to an α stationary point, i.e., increasing E does *not* help, even if no noise is added.

E Data-dependent Privacy Analysis

E.1 Privacy Analysis

Theorem 11 (Restatement of Theorem 3)**.** *Let PATE-FL and Private-kNN-FL answer Q queries with noise scale σ. For agent-level protection, both algorithms guarantee $(\alpha, Q\alpha/(2\sigma^2))$-RDP for all $\alpha \geq 1$. For instance-level protection, PATE-FL and Private-kNN-FL obey $(\alpha, Q\alpha/\sigma^2)$ and $(\alpha, Q\alpha/(k\sigma^2))$-RDP respectively.*

[3] One that outputs a solution in the subspace spanned by a sequence of subgradients.

Proof. In *PATE-FL*, for query x, by the independence of the noise added, the noisy sum is identically distributed to $\sum_{i=1}^{N} f_i(x) + \mathcal{N}(0, \sigma^2)$. Adding or removing one data instance from will change $\sum_{i=1}^{N} f_i(x)$ by at most $\sqrt{2}$ in L2. The Gaussian mechanism thus satisfies $(\alpha, \alpha s^2/2\sigma^2)$-RDP on the instance-level for all $\alpha \geq 1$ with an L2-sensitivity $s = \sqrt{2}$. This is identical to the analysis in the original PATE (Papernot et al., 2018).

For the agent-level, the L2 and L1 sensitivities are both 1 for adding or removing one agent.

In *Private-kNN-FL*, the noisy sum is identically distributed to $\frac{1}{k}\sum_{i=1}^{N}\sum_{j=1}^{k} y_{i,j} + \mathcal{N}(0, \sigma^2)$. The change of adding or removing one agent will change the sum by at most 1, which implies the same L2 sensitivity and same agent-level protection as *PATE-FL*. The $L2$-sensitivity from adding or removing one instance, on the other hand changes the score by at most $\sqrt{2/k}$ in L2 due to that the instance being replaced by another instance, this leads to an an improved instance-level DP that reduces ϵ by a factor of $\sqrt{\frac{k}{2}}$.

The overall RDP guarantee follows by the composition over Q queries. The approximate-DP guarantee follows from the standard RDP to DP conversion formula $\epsilon(\alpha) + \frac{\log(1/\delta)}{\alpha - 1}$ and optimally choosing α. □

E.2 Improved accuracy and privacy with large margin

Let $f_1, ..., f_N : \mathcal{X} \to \triangle^{C-1}$ where $\triangle^{C-1}$ denotes the probability simplex — the soft-label space. Note that both algorithms we propose can be viewed as voting of these teachers which outputs a probability distribution in $\triangle^{C-1}$. First let us define the margin parameter $\gamma(x)$ which measures the difference between the largest and second largest coordinate of $\frac{1}{N}\sum_{i=1}^{N} f_i(x)$.

Lemma 12 (Restatement of Lemma 4)**.** *Conditioning on the teachers, for each public data point x, the noise added to each coordinate is drawn from $\mathcal{N}(0, \sigma^2/N^2)$, then with probability $\geq 1 - C\exp\{-N^2\gamma(x)^2/8\sigma^2\}$, the privately released label matches the majority vote without adding noise.*

Proof. The proof is a straightforward application of Gaussian tail bounds and a union bound over C coordinates. Specifically, $\mathbb{P}[Z_{j^*} < -\gamma(x)/2] \leq e^{-\frac{N^2\gamma(x)^2}{8\sigma^2}}$ for the argmax j^*. For $j \neq j^*$, $\mathbb{P}[Z_j > \gamma(x)/2] \leq e^{-\frac{N^2\gamma(x)^2}{8\sigma^2}}$. By a union bound over all coordinates C, we get that there perturbation from the boundedness is smaller than $\gamma(x)/2$, which implies correct release of the majority votes. □

This lemma implies that for all public data point x such that $\gamma(x) \geq \frac{2\sqrt{2\log(C/\delta)}}{N}$, the output label matches noiseless majority votes with probability exponentially close to 1.

Next we show that for those data point x such that $\gamma(x)$ is large, the privacy loss for releasing $\arg\max_j [\frac{1}{N}\sum_{i=1}^{N} f_i(x)]_j$ is exponentially smaller. The result is based on the following privacy amplification lemma that is a simplification of Theorem 6 in the appendix of (Papernot et al., 2018).

Lemma 13. *Let $\mathcal{M}$ satisfy $(2\alpha, \epsilon)$-RDP, and there is a singleton output that happens with probability $1 - q$ when $\mathcal{M}$ is applied to D. Then for any D' that is adjacent to D, Renyi-divergence*

$$D_\alpha(\mathcal{M}(D) \| \mathcal{M}(D')) \leq -\log(1-q) + \frac{1}{\alpha - 1}\log(1 + q^{1/2}(1-q)^{\alpha-1}e^{(\alpha-1)\epsilon}).$$

Proof. Let P, Q be the distribution of of $\mathcal{M}(D)$ and $\mathcal{M}(D')$ respectively and E be the event that the singleton output is selected.

$$\begin{aligned}
\mathbb{E}_Q[(dP/dQ)^\alpha] &= \mathbb{E}_Q[(dP/dQ)^\alpha | E]\mathbb{P}_Q[E] + \mathbb{E}_Q[(dP/dQ)^\alpha \mathbf{1}(E^c)] \\
&\leq (1-q)(\frac{1}{1-q})^\alpha + \sqrt{\mathbb{E}_Q[(dP/dQ)^{(2\alpha)}]}\sqrt{\mathbb{E}_Q[\mathbf{1}(E^c)^2]} \\
&\leq (1-q)^{-(\alpha-1)} + q^{1/2}e^{(2\alpha-1)\epsilon/2} = (1-q)^{-(\alpha-1)}\left(1 + (1-q)^{\alpha-1}q^{1/2}e^{\frac{2\alpha-1}{2}\epsilon}\right)
\end{aligned}$$

Splits	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Total
Total	938	1585	2274	3500	3282	1312	12891

Table 4: DomainNet with seven classes

The first part of the second line uses the fact that event E is a singleton with probability larger than $1-q$ under Q and the probability is always smaller than 1 under P. The second part of the second line follows from Cauchy-Schwartz inequality. The third line substitute the definition of $(2\alpha, \epsilon)$-RDP. Finally, the stated result follows by the definition of the Renyi divergence. □

Theorem 14 (Restatement of Theorem 5)**.** *The mechanism that releases* $\arg\max_j[\frac{1}{N}\sum_{i=1}^{N} f_i(x) + \mathcal{N}(0, (\sigma^2/N^2)I_C)]_j$ *obeys* (α, ϵ)*-data-dependent-RDP, where*

$$\epsilon \leq Ce^{-\frac{N^2\gamma(x)^2}{8\sigma^2}} + \frac{1}{\alpha-1}\log\left(1 + e^{\frac{(2\alpha-1)\sigma^2}{2s^2} - \frac{N^2\gamma(x)^2}{16\sigma^2} + \log C}\right),$$

where $s = 1$ *for PATE-FL, and* $s = 1/k$ *for Private-KNN-FL.*

Proof. The proof involves substituting $q = Ce^{-\frac{N^2\gamma(x)^2}{8\sigma^2}}$ from Lemma 4 into Lemma 13 and use the fact that $\mathcal{M}$ satisfies the RDP of a Gaussian mechanism from the RDP's post-processing lemma. The expression bound is simplified for readability using $-\log(1-x) < 2x$ for all $x > -0.5$ and that $(1-q)^{\alpha-1} \leq 1$. □

As we can see, when given teachers that are largely in consensus, the (data-dependent) privacy loss exponentially smaller.

F Datasets and Models

Here we provide full details on the datasets and models used in our experiments.

Hyperparameters. For DP-FedAvg, the hyperparameters include agent sampling probability q, the noise parameter σ, the clipping threshold S. We do a grid search on all hyperparameters, and observe $(S = 0.08, \sigma = 0.06)$ works best for the simple CNN (used in ablation study) and AlexNet. The choice of q depends on the number of agents and the task complexity. A smaller q implies a stronger privacy guarantee and a larger variance. We set $q = 0.05$ for Digit dataset and $q = 0.04$ for CelebA. The number of local iterations E is another consideration. We empirically observe $E = 20$ achieves beset trade-offs between privacy and accuracy. For all experiments, the learning rate is 0.015, and we decay the learning rate through communication rounds, which leads to better performance compared to the original implementation in (Geyer et al., 2017).

For DP-FedSGD, we train each local model using Noisy SGD (Abadi et al., 2016), where the privacy parameters include batch size, the clipping threshold S, and the noisy scale σ. After a grid search, we use a batch size of 16 for Caltech dataset and 32 for DomainNet. We set the clipping threshold S to 0.08 and tune the noisy scale based on a fixed privacy budget. To amplify the privacy guarantee of DP-SGD using SMC, we set the number of local iteration $E = 1$.

Dataset. We provide detailed information datasets here. For Office-Caltech and DomainNet-fruit, we provide the number of images in each domain. An overview of DomainNet with seven selected fruit classes is depicted in Figure 3.

Details of CelebA Datasets Evaluation For DP-FedAvg, we set $(S = 0.08, \sigma = 0.06, q = 0.04)$. Note that the global sensitivity depends on the number of attributes, in which we use the same clipping technique in (Zhu et al., 2020) to restrict each agent's prediction clipped to τ attributions. We set $\tau = 4, \sigma = 50$ for PATE-FL. We apply AlexNet for all methods in this evaluation.

DomainNet Evaluation: We set $\sigma = 35$ for *Private-kNN-FL* with ResNet50. Q is the number of shared data.

Splits	Amazon	Dslr	Webcam	Caltech	Total
Total	958	157	295	1123	2533

Table 5: Office-Caltech10

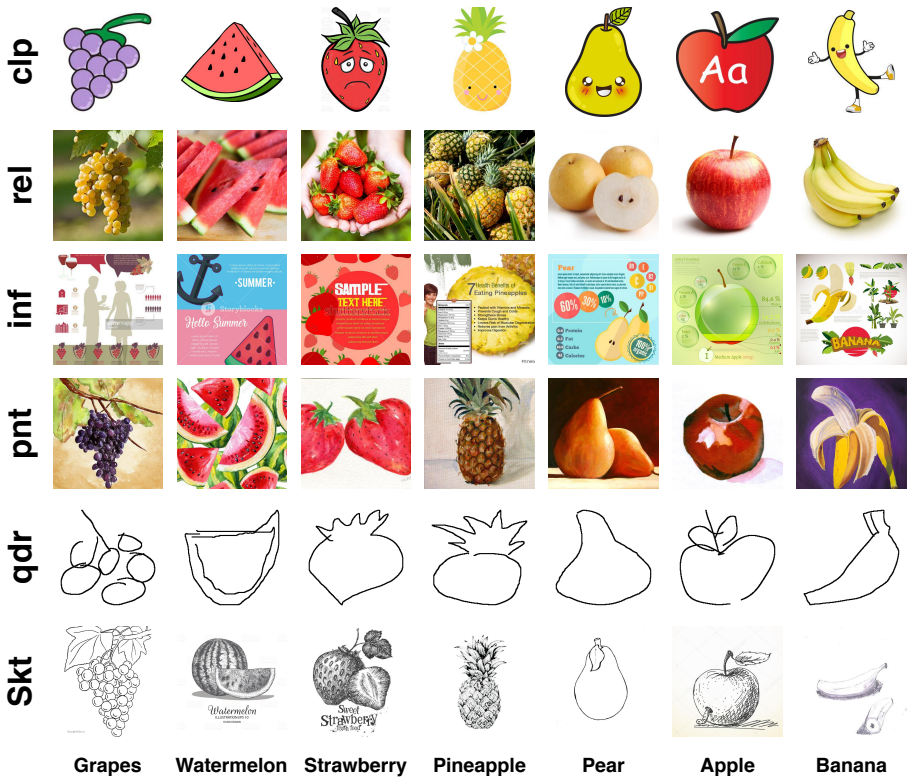


Figure 3: An overview of DomainNet dataset with seven selected fruit classes.