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 $\sigma$, clipping threshold $S$.
Initialize global model $\theta^0$
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 $\theta^0$
return update $\triangle_i{}^t = \theta - \theta^0$
```

Algorithm 5 DP-FedAvg (extend)

```
Input: Agent selection probability $q$, noise scale $\sigma$, clipping threshold $S$.
Initialize global model $\theta^0$
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 $\theta^0$
$\triangle_i{}^t = (\theta - \theta^0)/\max(1, \frac{||\theta - \theta^0||_2}{S})$
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 $\eta E G$ 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. *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.

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

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*

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

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. To reduce the privacy budget on global model training, we apply semi-supervised training on the Digit datasets, while for other datasets, the global model is trained using labeled data only.

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. For Office-Caltech and DomainNet, we split 70% data from the server domain as public available unlabeled data while the remaining 30% data is used for testing.

Details of Digit Datasets Evaluation 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.

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.

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

Table 4: DomainNet with seven classes

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

Table 5: Office-Caltech10

Office-Caltech Evaluation: Both AlexNet and Resnet50 are Imagenet pre-trained. For *Private-kNN-FL*, we instantiate the public feature extractor using the network backbone without the classifier layer. We set $\sigma = 15$ for *Private-kNN-FL* with AlexNet and $\sigma = 25$ for ResNet50. The privacy is calculated over all unlabeled data (T is the number of shared data).

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

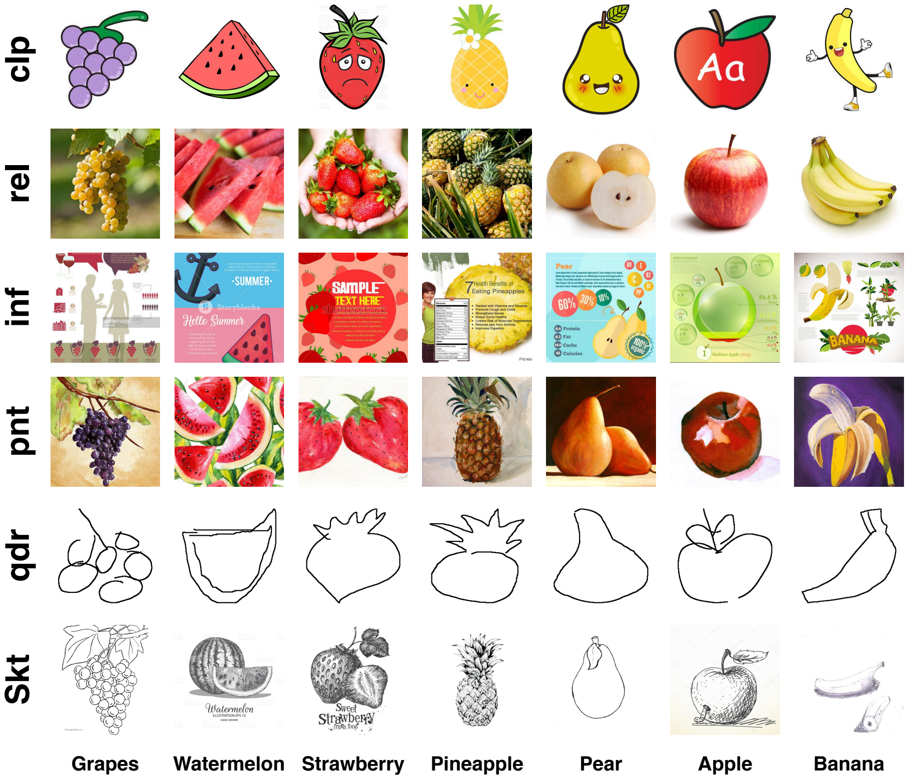

Figure 3: An overview of DomainNet dataset with seven selected fruit classes.