Personalization Improves Privacy–Accuracy Tradeoffs in Federated Learning

Alberto Bietti [1] **Chen-Yu Wei** [2] **Miroslav Dudík** [3] **John Langford** [3] **Zhiwei Steven Wu** [4]

Abstract

Large-scale machine learning systems often involve data distributed across a collection of users. Federated learning algorithms leverage this structure by communicating model updates to a central server, rather than entire datasets. In this paper, we study stochastic optimization algorithms for a personalized federated learning setting involving local and global models subject to user-level (joint) differential privacy. While learning a private global model induces a cost of privacy, local learning is perfectly private. We provide generalization guarantees showing that coordinating local learning with private centralized learning yields a generically useful and improved tradeoff between accuracy and privacy. We illustrate our theoretical results with experiments on synthetic and real-world datasets.

1. Introduction

Many modern applications of machine learning involve data from a large set of users. In such settings, both privacy considerations and bandwidth limits may require keeping each user's data on their device, instead communicating with a centralized server via shared model updates, a scenario commonly known as federated learning (Kairouz et al., 2021). When the data distribution varies across users, it is often beneficial to consider personalized models where parts of the model (*e.g.*, user-specific embeddings, or additive offsets) are local to each user, leading to updates that may be performed locally. Such personalized federated learning has been deployed at scale in a wide range of applications, including keyboard next-word prediction (Wang et al., 2019), automated speech recognition, and news personalization (Paulik et al., 2021).

In this paper, we focus on personalized federated learning algorithms based on stochastic optimization, which are the most widely used in this context (McMahan et al., 2017; Wang et al., 2021). Concretely, we consider stochastic optimization problems of the following form:

$$\min_{w,\theta_{1:N}} \left\{ f(w, \theta_{1:N}) := \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\xi \sim P_i}[f_i(w, \theta_i, \xi)] \right\}, \quad (1)$$

where N is the number of users, w a global parameter, $\theta_{1:N} = (\theta_1, \dots, \theta_N)$ a set of local parameters, and P_i denotes the sample distribution for user i. Similar objectives have been considered in other works (Agarwal et al., 2020; Hanzely et al., 2021). The personalization problem raises the question of how much learning should happen at the local vs global level: if the local models are expressive enough, local learning alone should suffice to learn good models with enough samples per user, yet if the data has some shared structure across users, a global model may help us learn more efficiently.

Although federated learning algorithms do not store user data in the central server, private user information may still be leaked in the resulting models, unless appropriate mechanisms are applied to guarantee privacy. In this work, we consider the notion of *user-level (joint) differential privacy* (DP) (Dwork et al., 2006; 2010a; Kearns et al., 2014), which ensures an adversary cannot reliably detect the presence or absence of all the data associated with a single user based on the output information. Such a notion has been used successfully in federated learning problems, with practical optimization algorithms that lead to useful privacy guarantees (McMahan et al., 2018; Hu et al., 2021).

Unfortunately, assuring such privacy may come at a cost to accuracy (see, for example, the lower bounds of (Bassily et al., 2014)). The key question we address is:

Can we leverage personalization to improve privacy–accuracy tradeoffs in federated learning?

The insight motivating our work is that in personalized federated learning, it is only the *global* portion of the optimization that needs to experience this drop in accuracy. A user's

[1] Center for Data Science, New York University [2] University of Southern California [3] Microsoft Research, New York [4] Carnegie Mellon University. Correspondence to: Alberto Bietti <alberto.bietti@nyu.edu>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

local model can compensate for the privacy-guaranteeing accuracy limitations of the global model. We formalize this by considering algorithms with a *personalization parameter* α that may vary the level of personalization from local learning ($\alpha = 0$) to global learning ($\alpha = \infty$). We then show generalization bounds on the objective 1 of the form

$$f(w_n, \theta_{1:N,n}) - f^* \leq C_{\text{stat}}(\alpha, n, N) + C_{\text{priv}}(\alpha, \epsilon, N), \quad (2)$$

where $f^* = \min_{w,\theta_{1:N}} f(w, \theta_{1:N})$ is the optimal risk, n is the number of observed samples per user and ϵ is the DP privacy parameter. Here, we expect the second term to vanish when $\alpha \to 0$, as local learning does not suffer from privacy. Crucially, this privacy cost does not depend on the number of samples per user n, while the first term, which captures statistical efficiency, generally decreases with n for any α. This emphasizes how adjusting the level of personalization through α can help improve generalization by adjusting the trade-off between these two terms.

Concretely, we provide precise guarantees of this form for simple federated private stochastic gradient algorithms, where the number of iterations corresponds to the number of samples per user n, and the personalization parameter α is given by the relative step-size between global and local updates. In particular, we show that α affects the complexity of learning by changing the geometry of the optimization: in problems that benefit from global models, small α makes learning more difficult but reduces the cost of privacy. We complement our theoretical results with experiments on synthetic and real-world federated learning datasets, which illustrate how varying the step-size ratio leads to improved trade-offs between accuracy and privacy.

2. Related Work

Model personalizaton. There are a variety of approaches for personalization in federated learning. In *local fine-tuning*, a global model is learnt by federated learning and then used as a warm start for on-device learning from the cache of local data (Wang et al., 2019; Paulik et al., 2021). This approach can be augmented with federated learning of hyperparameters (Wang et al., 2019; Jiang et al., 2019; Khodak et al., 2019) to obtain federated learning variants of meta-learning approaches like MAML (Finn et al., 2017) and Reptile (Nichol et al., 2018).

Another approach is to view personalization as a *multi-task learning problem*, and learn task-specific models with a regularization that forces parameters for similar tasks to be close, *e.g.*, in a Euclidean norm (Vanhaesebrouck et al., 2017; Smith et al., 2017; Arivazhagan et al., 2019; Mansour et al., 2020; Dinh et al., 2020; Shen et al., 2020; Huang et al., 2021; Marfoq et al., 2021; Singhal et al., 2021). Task similarity is typically expressed as a weighted undirected graph, a matrix, or a clustering. A special case of the multi-task approach is to learn a global model in addition to local models, which are regularized to be close to the global model (Mansour et al., 2020; Deng et al., 2020; Hanzely & Richtárik, 2020; Hanzely et al., 2021; Marfoq et al., 2021). This is closest to the approach here, which also separates global and local parameters. However, the approach here is more general, because we allow a broader range of modeling relationships between the global and local parameters, similar to federated residual learning (Agarwal et al., 2020). The statistical aspects of such personalization models were studied by Mansour et al. (2020); Agarwal et al. (2020), who in particular provide generalization guarantees for additive personalization models similar to ours. However, compared to the present paper, these works do not provide privacy guarantees nor study the effect of varying the level of personalization.

Privacy. The results here advance a recent line of work that provides formal user-level privacy guarantees for model personalizaton in federated learning. Similar to the prior works (Jain et al., 2021) and (Hu et al., 2021), we adopt the privacy formulation of *joint differential privacy* (JDP, Kearns et al., 2014), a variation of DP that is more suitable for the problem of model personalization than standard DP. (Jain et al., 2021) provides private algorithms that first learn a shared linear representation for all users, and allow each user to learn their local linear regression model over the lower-dimensional representation. This leads to a factorized model of personalization, which is different than ours and not handled by our theoretical assumptions due to non-convexity. (Hu et al., 2021) provides a private personalizaton algorithm through the mean-regularized multi-task learning objective without establishing its statistical rates. In contrast, we consider more general personalization schemes and provide statistical guarantees.

The results here are also related to other work in DP with a similar motivation to model personalization. (Li et al., 2020) studies meta-learning under DP. Their framework does not cover model personalization with a separate model for each user. (Noble et al., 2021) studies federated learning with DP with heterogeneous data across nodes, but they do not support personalization. (Bellet et al., 2018) studies fully decentralized DP algorithms for collaborative learning over a network instead of federated learning.

Finally, the notion of user-level privacy has also been adopted in prior works (McMahan et al., 2018; Levy et al., 2021), but these do not consider model personalization.

3. Preliminaries

In this section, we introduce the problem of personalized federated optimization, as well as the notion of user-level privacy that we consider.

3.1. Problem setting

We consider stochastic optimization problems of the form

$$\min_{w,\theta_{1:N}} \left\{ f(w,\theta_{1:N}) := \frac{1}{N}\sum_{i=1}^{N} f_i(w,\theta_i) \right\},$$

where N is the number of users, $w \in \mathbb{R}^{d_w}$ a global parameter, $\theta_{1:N} = (\theta_1,\dots,\theta_N) \in (\mathbb{R}^{d_\theta})^N$ a set of local parameters, and $f_i(w,\theta_i) := \mathbb{E}_{\xi\sim P_i}[f_i(w,\theta_i,\xi)]$ is the expected risk of user i, with random samples ξ drawn from an unknown user-specific distribution P_i.

While our algorithms may be run on arbitrary differentiable models, our analysis focuses on the convex setting, where $f_i(w,\theta_i,\xi_i)$ is jointly convex in (w,θ_i) for all ξ_i.

Additive model. An important special case is the additive model for supervised learning, where $d_w = d_\theta = d$, and

$$f_i(w,\theta_i,(x,y)) = \ell(y,(w+\theta_i)^\top x), \qquad (3)$$

where ℓ is a loss function and $\xi = (x,y)$ is a training sample.

As a running example, consider a movie recommendation app, which seeks to predict how each user i will rate any given movie. In this case, x corresponds to the features describing a movie (e.g., its genre, popularity, length, actors, etc.), and y is the user's rating of that movie. Following the federated protocol, the data about user activity stays on the device and only model parameters can be communicated to the server. In this case, only the information about the global parameter w is communicated.

The additive model makes the optimization problem (1) underdetermined, with many possible equivalent solutions obtained by adding any vector $v \in \mathbb{R}^d$ to w and then subtracting v from each θ_i, effectively "shifting" the predictive ability between global and local parameters. This is what allows the optimization algorithm to achieve different tradeoffs between statistical generalization (accuracy) and sharing of information across users (privacy).

To develop the intuition about this tradeoff, first consider the homogeneous scenario, where the optimal parameters $\theta_i^* \in \arg\min_\theta \mathbb{E}_{(x,y)\sim P_i}[\ell(y,\theta^\top x)]$ for each user are equal ($\theta_1^* = \cdots = \theta_N^*$). There are two extreme approaches: (i) **local learning**, where only θ_i are trained individually for each user, leading to poor sample complexity but perfect privacy (ii) **global learning**, where only w is trained, and we benefit from using samples from all users, but need communication with a centralized server and lose privacy.

In the more realistic heterogeneous scenario, when the θ_i^* are different but have some shared components, *e.g.*, only some coordinates differ across users, **joint learning** of w and $\theta_{1:N}$ can achieve greater accuracy (at the same number of samples) than both local and global learning, by leveraging more samples to estimate the shared components, while benefiting from user-specific data to personalize. Note that in this case, it is possible to further improve privacy, by allowing some shared components to be fitted entirely locally. Quantifying this improvement is the focus of our work.

In the movie recommendation example, if only global learning is performed, the prediction $w^\top x$ can use overall popularity statistics, but the resulting prediction is the same for all users, without any personalization. On the other hand, if only local learning is performed, although the system can fully personalize and provide full privacy, the quality of recommendation is limited by the number of movies the user watched before. With joint learning, the system can capture both global trends through w and adapt to each user's preference through θ_i.

3.2. User-level (joint) differential privacy

We aim to provide *user-level* privacy (Dwork et al., 2010a), which ensures that an adverary cannot detect the presence or absence of *all of the data associated with a single user* when given the output of the algorithm. To achieve this goal, we design algorithms using the *billboard model* (Hsu et al., 2014) of *differential privacy* (DP) (Dwork et al., 2006). Let us first revisit the definition of DP, which informally requires that changing any single user's data cannot change the algorithm's output by much.

Definition 3.1 (User-level DP; Dwork et al., 2006; 2010a)**.** A randomized mechanism M is (ϵ,δ)-differential privacy (DP) if for all pairs of data sets D, D' that differ by a single user's data and all events E in the output range,

$$\Pr[M(D)\in E] \le e^\epsilon \Pr[M(D')\in E] + \delta.$$

Billboard model. In the billboard model, a server computes aggregate information subject to the constraint of DP and shares the information as public messages with all N users. Then, based on the public messages and their own private data, each user computes their own personalized model. The billboard model is particularly compatible with algorithms in the federated setting, where the DP messages are typically noisy summary statistics about the users' local models (Jain et al., 2021; Hu et al., 2021). (Hsu et al., 2014) shows that algorithms under the billboard model provide an extremely strong privacy guarantee, known as *joint differential privacy* (JDP, Kearns et al., 2014).

Joint differential privacy (JDP). Let D_i denote the collection of samples associated with each user i. Two datasets D and D' are called i-neighbors if they only differ by user i's private data. For any mechanism M, we denote $M_{-i}(D)$ as the output information to all other users except user i.

Algorithm 1 Personalized-Private-SGD (PPSGD)

```
Input: $\eta$: step-size, $\alpha$: global/local ratio,
    $\sigma_\zeta$: privacy noise level, $C$: clipping parameter.
Initialize $w_0 = \theta_0 = 0$.
for $t = 1$ to $n$ do
    for all clients $i$ in parallel do
        Sample data $\xi_{i,t} \sim P_i$
        Compute $g^t_{\theta,i} = \nabla_\theta f_i(w_{t-1}, \theta_{i,t-1}, \xi_{i,t})$
            $g^t_{w,i} = \nabla_w f_i(w_{t-1}, \theta_{i,t-1}, \xi_{i,t})$
        Update $\theta_{i,t} = \theta_{i,t-1} - \frac{\eta}{N} g^t_{\theta,i}$
        Clip gradient: $\tilde{g}^t_{w,i} = g^t_{w,i} / \max(1, \frac{\|g^t_{w,i}\|}{C})$
        Send $\tilde{g}^t_{w,i}$ to the server
    end for
    Sample $\zeta_t \sim \mathcal{N}(0, \sigma_\zeta^2 I_{d_w})$
    Update $w_t = w_{t-1} - \alpha\eta(\frac{1}{N}\sum_{i=1}^N \tilde{g}^t_{w,i} + \zeta_t)$
end for
```

Definition 3.2 (Joint-Differential Privacy, JDP)**.** An algorithm M is (ϵ, δ)-jointly differentially private, written as (ϵ, δ)-JDP, if for all i, all i-neighboring datasets D, D', and all events E in the output space to all other users except i,

$$\Pr[M_{-i}(D) \in E] \leq e^\epsilon \Pr[M_{-i}(D') \in E] + \delta.$$

In the setting of model personalization, JDP implies that even if all of users except i collude, potentially pooling their private data and local models together, user i's private data are still protected, so long as i does not reveal their own model.

4. Main Algorithm and Analysis

Our main algorithm, shown in Algorithm 1, is a personalized version of distributed SGD (Dekel et al., 2012), with a *personalization parameter* α that controls the relative step-size between global and local updates.

At each round t, each user i samples a fresh datapoint $\xi_{i,t}$ (this could also be a mini-batch), updates its local model θ_i, and sends the gradient with respect to w to the central server, which aggregates gradients of all users before updating the global model w. The total number of rounds n thus corresponds to the number of samples per user. In order to guarantee user-level privacy, we clip each user's gradients $g^t_{w,i}$ before aggregation, and add Gaussian noise, following common practice (Abadi et al., 2016; McMahan et al., 2018; Hu et al., 2021; Chen et al., 2020). The choice of α affects the degree by which the algorithm favors local learning over global learning, with $\alpha = 0$ forcing local-only updates.

Privacy analysis. We first establish the formal user-level JDP guarantee of Algorithm 1.

Theorem 4.1 (Privacy)**.** *Suppose we set the noise parameter*

$$\sigma_\zeta \geq c \frac{C\sqrt{n \log(1/\delta)}}{N\epsilon}, \tag{4}$$

for an absolute constant c, then Algorithm 1 satisfies (ϵ, δ)-JDP in the billboard model.

We defer the full proof to Appendix B.1. At a high level, the proof first shows that the aggregate information released by the server (the sequence of global model updates) satisfies (ϵ, δ)-DP. Since the sequence of global models are sufficient statistics for each user to identify their personalized model θ_i, the JDP guarantee follows from the billboard lemma (see Lemma B.1).

Generalization analysis. Our generalization analysis relies on the following assumptions. We begin with an assumption about minimizers of f, which relies on the following norm for a vector $z = (w, \theta_{1:N})$ that captures the geometry induced by the personalization parameter α:

$$\|z\|^2_\alpha := \frac{1}{\alpha}\|w\|^2 + \|\theta_{1:N}\|^2. \tag{5}$$

Assumption 4.2 (Minimizers)**.** f admits a minimizer z^* with finite norm $\|z^*\|_\alpha$.

In the case of local learning ($\alpha = 0$), this implies z^* must have no global component. For joint learning, different minimizers might exist, and our bounds scale with the minimal norm $\|z^*\|_\alpha$ among all such minimizers.

Assumption 4.3 (Convexity and smoothness)**.** For all i and P_i-almost every ξ, the function $(w, \theta_i) \mapsto f_i(w, \theta_i, \xi)$ is jointly convex and L-smooth (its gradients are L-Lipschitz).

Note that this implies that $f(z)$ is jointly convex in $z = (w, \theta_{1:N})$. If we consider the example $f_i(w, \theta_i, (x, y)) = \frac{1}{2}(y - (w + \theta_i)^\top x)^2$, and assume $\|x\| \leq R$ almost surely, it is easy to verify that the assumption holds with $L = 2R^2$.

We will also make the following boundedness assumption on gradients, which is commonly made in the context of private stochastic optimization (*e.g.*, Bassily et al., 2014; Feldman et al., 2020), and simplifies our analysis by avoiding the need to study the effect of clipping on optimization.

Assumption 4.4 (Bounded gradients)**.** For all i, w, θ_i, and P_i-almost every ξ, we have $\|\nabla_w f_i(w, \theta_i, \xi)\| \leq G$.

This assumptions avoid the need to clip gradients when $C \geq G$, and our analysis hereafter assumes $C = G$. We note that G may be large in some cases, growing with the norm of optimal parameters, thus smaller values of C may often be beneficial in practice for better privacy guarantees.

Gradient variances. We consider the following variance quantities, which we assume to be finite, and which are obtained by considering gradients at a given minimizer z^*:

$$\begin{aligned}\sigma_{w,i}^2 &= \mathbb{E}_\xi \left\|\nabla_w f_i(w^*, \theta_i^*, \xi) - \nabla_w f_i(w^*, \theta_i^*)\right\|^2 \\ \sigma_{\theta,i}^2 &= \mathbb{E}_\xi \left\|\nabla_\theta f_i(w^*, \theta_i^*, \xi)\right\|^2 \\ \bar{\sigma}_w^2 &= \frac{1}{N}\sum_i \sigma_{w,i}^2, \qquad \bar{\sigma}_\theta^2 = \frac{1}{N}\sum_i \sigma_{\theta,i}^2.\end{aligned}$$

As an example, note that in a simple additive model (3) with squared loss and additive label noise of variance τ^2, we have $\sigma_{w,i}^2 = \sigma_{\theta,i}^2 = \tau^2 \operatorname{Tr}(\mathbb{E}_{P_i}[xx^\top])$, recovering standard statistical quantities. If the algorithm relies on mini-batches of size m instead of single datapoints, we may then replace $\sigma_{w,i}^2$ and $\sigma_{\theta,i}^2$ by $\sigma_{w,i}^2/m$ and $\sigma_{\theta,i}^2/m$, respectively, by averaging gradients over the mini-batches.

We now provide our main result on the convergence rate of Algorithm 1, which also yields a generalization bound on the excess risk. The proof is in Appendix B.2.

Theorem 4.5 (Generalization)**.** *Under Assumptions 4.2, 4.3, and 4.4, let $z^* = (w^*, \theta_{1:N}{}^*)$ be any minimizer of f, $L_\alpha := L\max(\alpha, \frac{1}{N})$, and*

$$\sigma_{tot,\alpha}^2 := \frac{\alpha\bar{\sigma}_w^2 + \bar{\sigma}_\theta^2}{N} + \alpha d_w \sigma_\zeta^2. \tag{6}$$

With $\eta = \min\{\frac{1}{4L_\alpha}, \frac{\|z^*\|_\alpha}{\sqrt{n}\sigma_{tot,\alpha}}\}$ and $C = G$, Algorithm 1 satisfies

$$\mathbb{E}[f(\bar{z}_n) - f(z^*)] \leq \frac{4L_\alpha\|z^*\|_\alpha^2}{n} + 3\frac{\sigma_{tot,\alpha}\|z^*\|_\alpha}{\sqrt{n}}, \tag{7}$$

with $\bar{z}_n = \frac{1}{n}\sum_{t=0}^{n-1} z_t$. In particular, with σ_ζ as in (4), hiding absolute constants, we have the following generalization bound:

$$\mathbb{E}[f(\bar{z}_n) - f(z^*)] \lesssim \tag{8}$$

$$\frac{L_\alpha\|z^*\|_\alpha^2}{n} + \|z^*\|_\alpha\sqrt{\frac{\alpha\bar{\sigma}_w^2 + \bar{\sigma}_\theta^2}{Nn}} + \|z^*\|_\alpha\sqrt{\frac{\alpha d_w G^2\log(\frac{1}{\delta})}{N^2\epsilon^2}} \tag{9}$$

The generalization bound takes the form (2), with a cost of privacy that does not depend on the number of samples per user n. As is common in the analysis of SGD, our bound displays a bias term that decays as $1/n$, and a variance term decaying as $1/\sqrt{n}$, controlled by the gradient variance $\sigma_{tot,\alpha}^2$. Ignoring the privacy noise, given that we use N samples at each round, this variance scales as $1/N$, leading to an overall asymptotic rate of $1/\sqrt{Nn}$, where Nn is the total sample size. Using minibatches of size m would further improve this to $1/\sqrt{Nnm}$, where the total number of samples is now Nnm.

Local vs global learning. Note that when multiple minimizers z^* exist, one may choose the one with minimal norm $\|z^*\|_\alpha$. If we consider the additive model (3), we may always consider a minimizer of the form $z^* = (0, \theta_{1:N}^*)$. Then $\|z^*\|_\alpha$ is bounded even for $\alpha = 0$ (**local learning**), which yields an excess risk of order

$$\frac{L\sum_i\|\theta_i^*\|^2}{Nn} + \sqrt{\frac{\bar{\sigma}_\theta^2\sum_i\|\theta_i^*\|^2}{Nn}}.$$

In particular, there is no cost for privacy. If we further assume $\theta_i^* = v^*$ for all i, this becomes

$$\frac{L\|v^*\|^2}{n} + \sqrt{\frac{\bar{\sigma}_\theta^2\|v^*\|^2}{n}}.$$

We see slow convergence that does not improve with N, which is expected given that no information is shared across users. In this setting, we may instead consider a different minimizer $z^* = (v^*, 0)$, for which taking the limit $\alpha \to \infty$ (**global learning**) yields the excess risk of order

$$\frac{L\|v^*\|^2}{n} + \|v^*\|\sqrt{\frac{\bar{\sigma}_w^2}{Nn}} + \|v^*\|\sqrt{\frac{d_w G^2\log(\frac{1}{\delta})}{N^2\epsilon^2}}.$$

The variance term now decays faster for large N, but has an additional privacy term, which decreases quickly with N but does not improve with large n and may be quite large, particularly in high dimensions, consistent with lower bounds for differentially private optimization (Bassily et al., 2014).

Characterizing benefits of personalization. Depending on the number of samples per user n, it may be helpful to choose varying levels of personalization α to adjust this tradeoff, from large α for small n to small α for large n when the privacy cost dominates the bound. The next lemma quantifies this tradeoff for the additive model with homogeneous users (*i.e.*, $\theta_1^* = \cdots = \theta_N^*$):

Lemma 4.6 (Threshold on n for personalization benefits)**.** *Assume $(v, 0)$ is a minimizer of f, with f an additive model of the form (3). For any α, the minimizer with minimal α-norm is given by $(w, \theta_{1:N})$ with $w = \frac{\alpha N}{\alpha N+1}v$ and $\theta_i = \frac{1}{\alpha N+1}v$. Assuming $\bar{\sigma}_w = \bar{\sigma}_\theta = \sigma$, the variance term in the excess risk (9) with n samples per user is monotonic in α, taking the form:*

$$\|v\|\sqrt{\frac{N}{\alpha N+1}\left(\frac{(\alpha+1)\sigma^2}{Nn} + \frac{\alpha d_w G^2\log(1/\delta)}{N^2\epsilon^2}\right)}.$$

This is non-decreasing with α if and only if

$$n \gtrsim \frac{N(N-1)\sigma^2\epsilon^2}{d_w G^2\log(1/\delta)}. \tag{10}$$

Thus, if we ignore the bias term in (9), this suggests that when (10) holds, using local learning (smaller α) should help improve generalization, while if the reverse inequality holds, global learning should be preferred. Eq. (10) suggests that this threshold on the user sample size n scales quadratically with the number of users N, linearly with the privacy level ϵ, and inversely with the dimension d_w. When the optimal user parameters are different, this transition would likely happen for smaller n, as we expect local models to be useful even at small sample sizes regardless of privacy.

Improving the bias term. We remark that the bias term decreases as $1/n$ in both local and global learning scenarios described above, and does not improve with the number of users. As in standard SGD (*e.g.*, Dekel et al., 2012), this term decreases with the number of rounds, and may thus decrease more quickly with the number of samples if more communication rounds are performed for the same total number of samples. In Section 5, we show that sampling users at each round can help improve this term.

Comparison to Local SGD. A common approach for federated optimization is the local SGD (or federated averaging) algorithm (McMahan et al., 2017), which performs multiple local steps before communicating with the server. This is in contrast to our method, which more closely resembles mini-batch SGD. We note that despite its practical success, the known theoretical guarantees of local SGD typically do not improve on mini-batch SGD except for specific settings (Woodworth et al., 2020). Our study therefore focuses on understanding personalization in the SGD setting, but we note that extending our analysis to local SGD is an interesting direction for future work.

5. Heterogeneous Sample Sizes, User Sampling

In this section, we study a variant of Algorithm 1 where we sample some of the users uniformly at each round, with possibly different minibatch sizes for each user. This reflects the fact that users may have different amounts of data, and that not all clients may be available at each round.

Let q be the probability of sampling any given user at each round, $m_i \geq 1$ be the mini-batch size for user i, and define $M = \sum_i m_i$ and $m_{\max} = \max_i m_i$. Algorithm 2 then optimizes the objective

$$f^m(w, \theta_{1:N}) := \sum_{i=1}^{N} \frac{m_i}{M} f_i(w, \theta_i). \qquad (11)$$

Namely, the algorithm optimizes each user's performance proportionally to their amount of samples. We denote the number of rounds by T here, since it no longer corresponds to the total number of samples per user n. Note that the expected total number of samples observed for user i after T rounds is now given by $n_i = Tqm_i$.

Algorithm 2 PPSGD with client sampling

```
Input: $q$: client sampling probability,
       $m_i$: minibatch sizes, $\eta$: step size,
       $\alpha$: global/local ratio, $\sigma_\zeta$: privacy noise level,
       $C$: clipping parameter.
Initialize $w_0 = \theta_0 = 0$.
for $t = 1$ to $T$ do
    Sample $b_{i,t} \sim Ber(q)$
    for all clients $i$ with $b_{i,t} = 1$ in parallel do
        Sample a minibatch $\{\xi_{i,t}^{(k)}\}_{k=1}^{m_i} \sim P_i^{\otimes m_i}$
        Compute $g_{\theta,i}^t = \sum_{k=1}^{m_i} \nabla_\theta f_i(w_{t-1}, \theta_{i,t-1}, \xi_{i,t}^{(k)})$
                $g_{w,i}^t = \sum_{k=1}^{m_i} \nabla_w f_i(w_{t-1}, \theta_{i,t-1}, \xi_{i,t}^{(k)})$
        Update $\theta_{i,t} = \theta_{i,t-1} - \frac{\eta}{qM} g_{\theta,i}^t$
        Clip gradient $\tilde{g}_{w,i}^t = g_{w,i}^t / \max(1, \frac{\|g_{w,i}^t\|}{C})$
        Send $\tilde{g}_{w,i}^t$ to the server
    end for
    Sample $\zeta_t \sim \mathcal{N}(0, \sigma_\zeta^2 I_{d_w})$
    Update $w_t = w_{t-1} - \alpha\eta(\frac{1}{qM} \sum_{i:b_{i,t}=1} \tilde{g}_{w,i}^t + \zeta_t)$
end for
```

The following result provides a privacy guarantee for Algorithm 2. Compared to Theorem 4.1, it also leverages a standard argument of privacy amplification by subsampling.

Theorem 5.1 (Privacy with client sampling)**.** *There exist absolute constants c_1, c_2 such that for any $\epsilon < c_1 q^2 T$, if we take*

$$\sigma_\zeta \geq c_2 \frac{C\sqrt{T \log(1/\delta)}}{M\epsilon}, \qquad (12)$$

then Algorithm 2 satisfies (ϵ, δ)-JDP.

Note that compared to the uniform sampling case, the clipping parameter C may need to be larger since we are clipping the sum of up to $m_{\max}$ gradients. In our generalization analysis hereafter, we thus assume $C = Gm_{\max}$, so that clipping is not needed under Assumption 4.4. We now give an optimization and generalization guarantee.

Theorem 5.2 (Generalization with client sampling)**.** *Under Assumptions 4.2, 4.3, and 4.4, let $z^* = (w^*, \theta_{1:N}{}^*)$ be any minimizer of f^m, $L_{m,\alpha} = L \max\{\alpha + \frac{\alpha m_{\max}}{M}, \frac{m_{\max}}{M}\}$, and*

$$\sigma_{m,\alpha}^2 := \frac{\alpha\bar{\sigma}_{w,m}^2 + \alpha\tilde{\sigma}_{w,m}^2 + \bar{\sigma}_{\theta,m}^2}{qM} + \alpha d_w \sigma_\zeta^2, \qquad (13)$$

where $\bar{\sigma}_{w,m}^2 := \frac{1}{M}\sum_i m_i \sigma_{w,i}^2$ and $\bar{\sigma}_{\theta,m}^2 := \frac{1}{M}\sum_i m_i \sigma_{\theta,i}^2$ and $\tilde{\sigma}_{w,m}^2 := \frac{1}{M}\sum_i q(1-q) m_i^2 \|\nabla_w f_i(w^*, \theta_i^*)\|^2$.

With $\eta = \min\{\frac{1}{4L_{m,\alpha}}, \frac{\|z^*\|_\alpha}{\sqrt{T}\sigma_{m,\alpha}}\}$ and $C = G_m := Gm_{\max}$, Algorithm 2 satisfies

$$\mathbb{E}[f^m(\bar{z}_T) - f^m(z^*)] \leq \frac{4L_{m,\alpha}\|z^*\|_\alpha^2}{T} + 3\frac{\sigma_{m,\alpha}\|z^*\|_\alpha}{\sqrt{T}},$$

with $\bar{z}_T = \frac{1}{T}\sum_{t=0}^{T-1} z_t$. Taking σ_ζ as in (12), we have:

$$\mathbb{E}[f^m(\bar{z}_T) - f^m(z^*)] \lesssim \frac{L_{m,\alpha}\|z^*\|_\alpha^2}{T} + \tag{14}$$
$$+ \|z^*\|_\alpha \sqrt{\frac{\alpha\bar{\sigma}_{w,m}^2 + \alpha\tilde{\sigma}_{w,m}^2 + \bar{\sigma}_{\theta,m}^2}{qMT} + \frac{\alpha d_w G_m^2 \log(\frac{1}{\delta})}{M^2\epsilon^2}}$$

Note that the privacy cost now scales with $G_m/M = Gm_{\max}/M$ instead of G/N in Theorem 4.5. Note that we have $Gm_{\max}/M \geq G/N$, with equality if and only if all the m_i are equal. This highlights the additional cost of privacy with heterogeneous sample sizes, particularly when some users have much more data than others. Compared to Theorem 4.5, there is also an additional variance term $\tilde{\sigma}_{w,m}$ induced by the randomness of client sampling. Note that $\tilde{\sigma}_{w,m}$ vanishes when (i) $q = 1$ (*i.e.*, no sampling), or (ii) $\nabla_w f_i(w^*, \theta_i^*) = 0$ for all i, which always holds for the additive model (3).

When all users are sampled ($q = 1$) and $m_i = 1$ for each user, we have $M = N$ and $\tilde{\sigma}_{w,m} = 0$, and we recover the same bound as in Theorem 4.5, where $T = n$ corresponds to the number of samples per user. In this case the bias term decreases slowly as $1/n$. In contrast, if $m_i = 1$ and $q = 1/N$, *i.e.*, we sample one user per round on average, then $qM = 1$, and T now corresponds to the total number of samples Nn. The bias term now decreases as $1/Nn$, while the variance decreases at the same $1/\sqrt{Nn}$ rate. Of course, this comes at the cost of more frequent communication rounds, since each round only uses the data of a single user.

Optimizing average user performance. In some cases, we may have clients with heterogeneous amounts of data, but still want to optimize the average performance f in (1), rather than the weighted average f^m in (11). This can be achieved by replacing $g_{\theta,i}^t$ and $g_{w,i}^t$ by averages over the minibatch instead of sums in Algorithm 2. This leads to the following guarantee on the excess risk (see Appendix B.6):

$$\mathbb{E}[f(\bar{z}_T) - f(z^*)] \lesssim \tag{15}$$
$$\frac{L_\alpha\|z^*\|_\alpha^2}{T} + \|z^*\|_\alpha \sqrt{\frac{\alpha\bar{\sigma}_w^2 + \bar{\sigma}_\theta^2}{qNm_{\min}T} + \frac{\alpha d_w G^2 \log(\frac{1}{\delta})}{N^2\epsilon^2}}. \tag{16}$$

Note that the privacy cost is similar to the case with homogeneous data in Section 4, rather than the larger cost of Theorem 5.2, since we are now treating each user equally. Nevertheless, the variance term displays an effective total sample size of $n_{\text{tot}} = qNm_{\min}T$, which is smaller than in Theorem 5.2, where $n_{\text{tot}} = qMT$ corresponds to the expected total number of samples, unless all the m_i are equal. This highlights that this improvement in privacy comes at a cost in statistical efficiency.

6. Experiments

In this section, we present numerical experiments that illustrate our theoretical guarantees on both synthetic and real-world federated learning datasets. Our code is available at https://github.com/albietz/ppsgd.

Experiment setup. We consider linear additive models of the form (3) with the squared loss. For classification datasets, we use a one-versus-all setup. Unless otherwise mentioned, we run our algorithms for one pass over the data, simulating the online setting considered by our theory. We sample a fixed number of users (denoted Q) at each iteration, chosen randomly without replacement, and consider minibatches of size $m = 10$ for each user. Following standard practice, we compute privacy guarantees ϵ at each iteration using the moments accountant (Abadi et al., 2016), with a fixed $\delta = 10^{-4}$, after applying Gaussian noise of variance $\sigma_\zeta^2 = \sigma^2 C^2$, where C is a clipping parameter and σ a noise multiplier. For each run, we consider a fixed step-size η, personalization parameter α, clipping parameter C, and noise multiplier σ, chosen from a grid (see Appendix A). In order to assess the best achievable performance, we optimize the learning rate separately at each number of iterations reported in our figures.

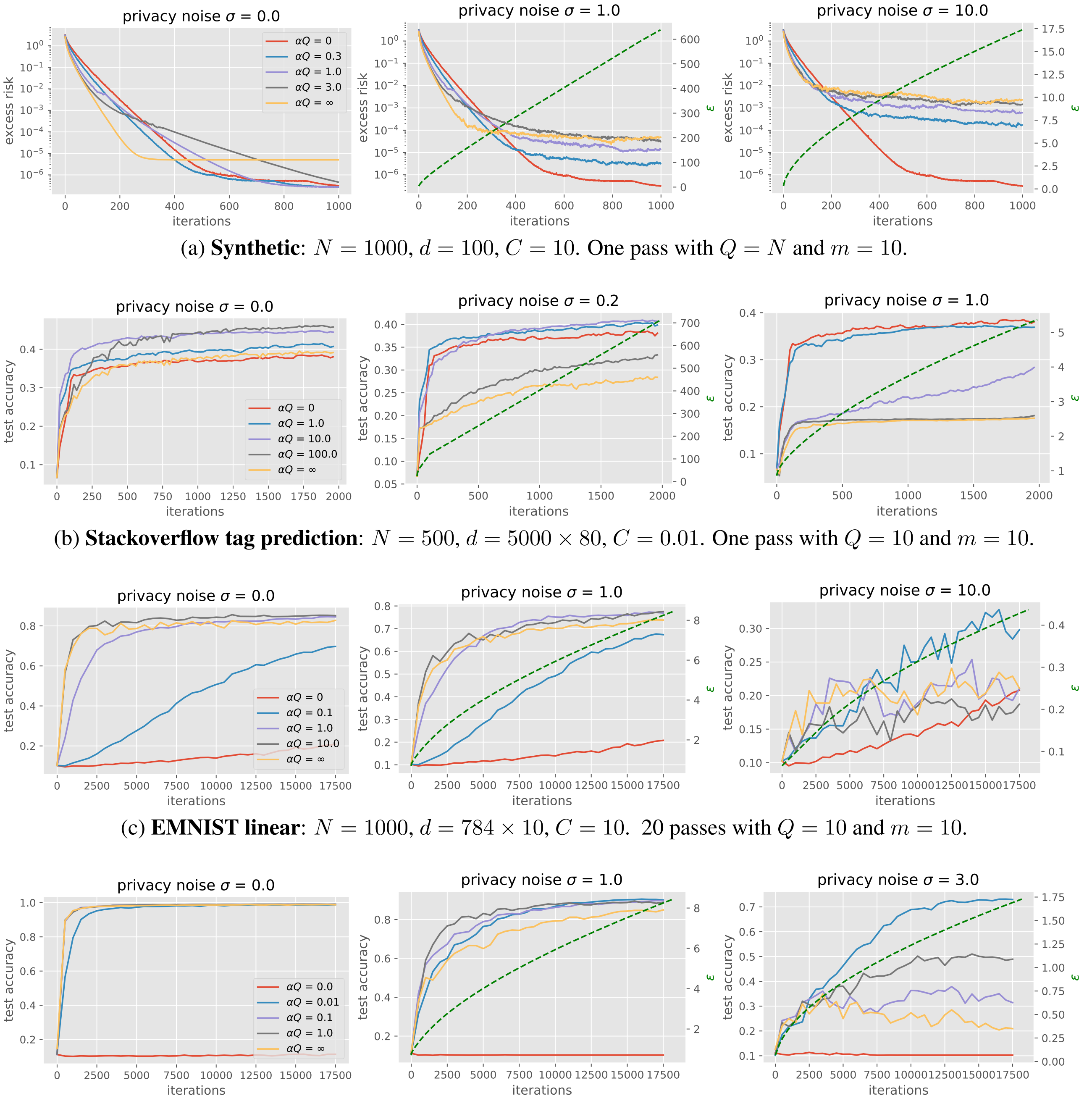
Synthetic dataset. We begin with a synthetic regression example, where each user's data is generated from a ground truth linear model, and multiple coordinates of the parameters are shared across users. We consider $N = 1000$ users and data in $d = 100$ dimensions. The parameters θ_i^* are generated as follows from a random $\theta_0 \sim \mathcal{N}(0, 10^2 I_d)$:

$$[\theta_i^*]_{1:95} = [\theta_0]_{1:95}$$
$$[\theta_i^*]_{96:100} = [\theta_0]_{96:100} + \delta_i, \quad \delta_i \sim \mathcal{N}(0, 0.01^2 I),$$

while the samples (x_i, y_i) for user i are generated by the following process:

$$x_i \sim \mathcal{N}(0, \Sigma), \qquad \Sigma = \operatorname{diag}\{\lambda_k\}_{k=1}^d, \lambda_k = 1/k$$
$$y_i = \theta_i^{*\top} x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \tau^2).$$

With this setup, the excess risk can be computed in closed form, and is given by $\frac{1}{N}\sum_i \|w + \theta_i - \theta_i^*\|_\Sigma^2$, and is shown in Figure 1a. We observe that global learning ($\alpha = \infty$) enjoys faster convergence at initial iterations in all cases, thanks to better data efficiency, but eventually stops improving, even without privacy noise, since minimizing risk requires personalization in this model. Small choices of α lead to better performance in later iterations, and with larger levels of privacy noise, full local learning ($\alpha = 0$) dominates joint learning quite early in the learning process, after about 200 iterations, *i.e.*, 2000 samples per user.



(a) **Synthetic**: $N = 1000$, $d = 100$, $C = 10$. One pass with $Q = N$ and $m = 10$.

(b) **Stackoverflow tag prediction**: $N = 500$, $d = 5000 \times 80$, $C = 0.01$. One pass with $Q = 10$ and $m = 10$.

(c) **EMNIST linear**: $N = 1000$, $d = 784 \times 10$, $C = 10$. 20 passes with $Q = 10$ and $m = 10$.

(d) **EMNIST CNN**: $N = 1000$, $d = 784 \times 10$, $C = 0.1$. 20 passes with $Q = 10$ and $m = 10$.

Figure 1. Test performance of PP-SGD on the synthetic, Stackoverflow, and EMNIST datasets. Each plot shows curves for different levels of personalization α for fixed privacy noise σ and clipping parameter C. In the private setting, the green dashed line displays the privacy guarantee ϵ at iteration T for all curves in each plot (except $\alpha = 0$ with is always fully private). Q is the number of sampled users and m the number of samples per user in each round. The choice $\alpha Q = 1$ equalizes the variance of global and local updates due to samples.

Stackoverflow tag prediction. We now consider a subset of the Stackoverflow dataset[1] consisting of $N = 500$ users, each with about 300–500 training and 10–100 test documents in a bag-of-words format of dimension 5000. The task is to predict Stackoverflow tags from documents that consist of user questions, and we use the 80 most popular tags as labels.

[1] https://www.tensorflow.org/federated/api_docs/python/tff/simulation/datasets/stackoverflow/load_data

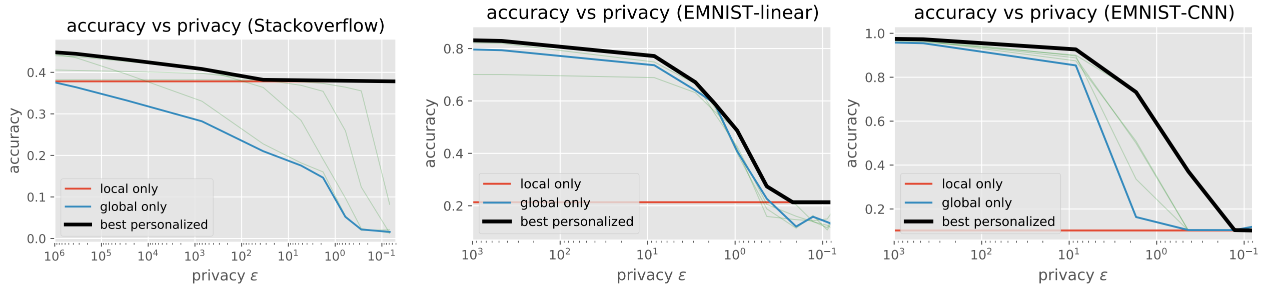


Figure 2. Accuracy–privacy tradeoffs on Stackoverflow and EMNIST at the end of training. The red, blue, and light green curves are obtained by varying the amount of privacy noise σ while optimizing the learning rate with local, global, and intermediate models, respectively. The black curve also optimizes over the level of personalization, which improves the tradeoff.

In Figure 1b, we show the top-1 accuracy on the test set, while running our method for a single pass over the data, sampling $Q = 10$ users and $m = 10$ documents per user at each iteration. We observe that without privacy, the best performance is obtained with an intermediate amount of personalization, $\alpha = 10$, suggesting that local models are helpful for generalization on this dataset. With privacy noise, however, models with large α quickly degrade in performance, and small α performs best for achieving reasonable privacy levels. Overall, personalization plays a key role for improving the privacy–accuracy tradeoff on this dataset, as shown in Figure 2.

Federated EMNIST. Finally, we consider a subset of the federated EMNIST digit classification dataset[2] consisting of $N = 1000$ users with about 100 training and 10 test samples per user. Here, a single pass does not lead to good performance, suggesting this may be a hard learning problem that benefits from multiple passes (Pillaud-Vivien et al., 2018). We thus run our method for 20 epochs, sampling 10 users and 10 images per user at each iteration, and show the resulting test accuracy in Figure 1c. On this dataset, we see that local learning does generally poorly, perhaps because there is a lot of shared structure in images across users that benefits significantly from global learning. Nevertheless, there is still a benefit to using small α for achieving good privacy guarantees, and personalization still plays a role in improving the privacy–accuracy tradeoff (albeit less pronounced than on Stackoverflow, see Figure 2).

In Figure 1d, we show similar plots on federated EMNIST for a 4-layer CNN with a shared global representation $\Phi_W(x)$ and an additive model at the output layer, leading to a loss of the form

$$f_i((w, W), \theta_i, (x, y)) = \ell(y, (w + \theta_i)^\top \Phi_W(x)).$$

We see that similar conclusions hold in this non-convex scenario, namely, small α is helpful for achieving better privacy guarantees, while large α leads to better accuracy for poor privacy levels. Here the local-only models ($\alpha = 0$) perform very poorly, since the global representation has frozen random weights, highlighting that learning a useful representation in this model requires a privacy cost.

7. Discussion and Conclusion

In this paper, we studied personalized private federated optimization algorithms, providing generalization guarantees in the convex setting that highlight the role of personalization and privacy. We show both theoretically and empirically that adjusting the amount of personalization is always beneficial for improving the accuracy–privacy tradeoff, by providing a way to attenuate the cost of privacy through local learning at large sample sizes. Promising future directions include extending our analysis to different models of personalization, *e.g.*, based on neural networks, and to provide adaptive algorithms which may dynamically adjust the level of personalization in an online fashion.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016.

Agarwal, A., Langford, J., and Wei, C.-Y. Federated residual learning. *arXiv preprint arXiv:2003.12880*, 2020.

Arivazhagan, M. G., Aggarwal, V., Singh, A. K., and Choudhary, S. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

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

[2] https://www.tensorflow.org/federated/api_docs/python/tff/simulation/datasets/emnist/load_data

Bellet, A., Guerraoui, R., Taziki, M., and Tommasi, M. Personalized and private peer-to-peer machine learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Chen, X., Wu, Z. S., and Hong, M. Understanding gradient clipping in private SGD: A geometric perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Dekel, O., Gilad-Bachrach, R., Shamir, O., and Xiao, L. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(1), 2012.

Deng, Y., Kamani, M. M., and Mahdavi, M. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

Dinh, C. T., Tran, N., and Nguyen, J. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, 2006.

Dwork, C., Naor, M., Pitassi, T., and Rothblum, G. N. Differential privacy under continual observation. In *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, 2010a.

Dwork, C., Rothblum, G. N., and Vadhan, S. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, 2010b.

Feldman, V., Koren, T., and Talwar, K. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, 2020.

Finn, C., Abbeel, P., , and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Hanzely, F. and Richtárik, P. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*, 2020.

Hanzely, F., Zhao, B., and Kolar, M. Personalized federated learning: A unified framework and universal optimization techniques. *arXiv preprint arXiv:2102.09743*, 2021.

Hsu, J., Huang, Z., Roth, A., Roughgarden, T., and Wu, Z. S. Private matchings and allocations. In *Symposium on Theory of Computing, STOC*, 2014.

Hu, S., Wu, Z. S., and Smith, V. Private multi-task learning: Formulation and applications to federated learning. *arXiv preprint arXiv:2108.12978*, 2021.

Huang, Y., Chu, L., Zhou, Z., Wang, L., Liu, J., Pei, J., and Zhang, Y. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Jain, P., Rush, J., Smith, A., Song, S., and Guha Thakurta, A. Differentially private model personalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Jiang, Y., Konečnỳ, J., Rush, K., and Kannan, S. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*, 2019.

Kairouz, P., McMahan, H. B., et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021.

Kearns, M., Pai, M. M., Roth, A., and Ullman, J. Mechanism design in large games: Incentives and privacy. *American Economic Review*, 104(5):431–35, May 2014.

Khodak, M., Balcan, M.-F. F., and Talwalkar, A. S. Adaptive gradient-based meta-learning methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Levy, D., Sun, Z., Amin, K., Kale, S., Kulesza, A., Mohri, M., and Suresh, A. T. Learning with user-level privacy. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Li, J., Khodak, M., Caldas, S., and Talwalkar, A. Differentially private meta-learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

Mansour, Y., Mohri, M., Ro, J., and Suresh, A. T. Three approaches for personalization with applications to federated learning, 2020.

Marfoq, O., Neglia, G., Bellet, A., Kameni, L., and Vidal, R. Federated multi-task learning under a mixture of distributions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

McMahan, H. B., Ramage, D., Talwar, K., and Zhang, L. Learning differentially private recurrent language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

Nichol, A., Achiam, J., and Schulman, J. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Noble, M., Bellet, A., and Dieuleveut, A. Differentially private federated learning on heterogeneous data. *arXiv preprint arXiv:2111.09278*, 2021.

Paulik, M., Seigel, M., Mason, H., Telaar, D., Kluivers, J., van Dalen, R., Lau, C. W., Carlson, L., Granqvist, F., Vandevelde, C., et al. Federated evaluation and tuning for on-device personalization: System design & applications. *arXiv preprint arXiv:2102.08503*, 2021.

Pillaud-Vivien, L., Rudi, A., and Bach, F. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Shen, T., Zhang, J., Jia, X., Zhang, F., Huang, G., Zhou, P., Kuang, K., Wu, F., and Wu, C. Federated mutual learning. *arXiv preprint arXiv:2006.16765*, 2020.

Singhal, K., Sidahmed, H., Garrett, Z., Wu, S., Rush, J., and Prakash, S. Federated reconstruction: Partially local federated learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Smith, V., Chiang, C.-K., Sanjabi, M., and Talwalkar, A. Federated multi-task learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Vanhaesebrouck, P., Bellet, A., and Tommasi, M. Decentralized collaborative learning of personalized models over networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

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

Wang, K., Mathews, R., Kiddon, C., Eichner, H., Beaufays, F., and Ramage, D. Federated evaluation of on-device personalization. *arXiv preprint arXiv:1910.10252*, 2019.

Woodworth, B., Patel, K. K., Stich, S., Dai, Z., Bullins, B., Mcmahan, B., Shamir, O., and Srebro, N. Is local sgd better than minibatch sgd? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

A. Experiment details

We provide the hyperparameter grids for each dataset below. Our experiments always optimize the step-size at any fixed iteration. To obtain the plots in Figure 2, we also optimize over α to obtain the "best personalized" curve, while varying over a finer grid of noise levels, provided below.

Synthetic dataset:

- step-size η: [0.01, 0.02, 0.05, 0.1, 0.2, 0.4, 0.7, 1., 1.2, 1.5, 1.8],
- personalization level αQ: [0, 0.1, 0.3, 1, 3, 10, 30, 100, ∞],
- noise multiplier σ: [0, 0.1, 0.3, 1, 3, 10, 30, 100, 300, 1000].

Stackoverflow:

- step-size η: [2, 5, 10, 20, 50, 100, 150, 200, 250],
- personalization level αQ: [0, 0.1, 0.3, 1, 3, 10, 30, 100, ∞],
- noise multiplier σ: [0, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 100].

EMNIST-linear:

- step-size η: [0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01],
- personalization level αQ: [0, 0.1, 0.3, 1, 3, 10, 30, 100, ∞],
- noise multiplier σ: [0, 0.1, 0.3, 1, 2, 3, 5, 10, 20, 30, 50, 100].

EMNIST-CNN:

- step-size η: [0.05, 0.1, 0.2, 0.5, 1., 2., 5.],
- personalization level αQ: [0, 0.1, 0.3, 1, 3, 10, 30, 100, ∞],
- noise multiplier σ: [0, 0.1, 0.3, 1., 3., 10., 30., 100.].

B. Proofs of Main Results

For proving the privacy guarantees, we make use of the following billboard lemma.

Lemma B.1 (Billboard Lemma (Hsu et al., 2014))**.** *Suppose a message broadcasting mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{W}$ is (ϵ, δ)-differentially private. Consider any set of functions: $g_i : \mathcal{D}_i \times \mathcal{W} \to \mathcal{W}'$. The composition $\{g_i(\Pi_i \mathcal{D}, \mathcal{M}(\mathcal{D}))\}$ is (ϵ, δ)-joint differentially private, where $\Pi_i : \mathcal{D} \to \mathcal{D}_i$ is the projection of $\mathcal{D}$ onto $\mathcal{D}_i$.*

B.1. Proof of Theorem 4.1 (Privacy for PP-SGD)

Proof. To prove the JDP guarantee, by the billboard lemma (Lemma B.1), it is sufficient to show that the global model updates satisfy standard DP w.r.t. any changes to a single user's data.

With the assumption that gradients are uniformly bounded by G, each update of the global models has ℓ_2 sensitivity bounded by $\alpha\eta/N$ w.r.t. any change to a single user's data. If we take

$$\sigma_\zeta \geq c\frac{G\sqrt{n\log(1/\delta)}}{N\epsilon},$$

for some absolute constant c, then by the Gaussian mechanism, each update is $(\frac{\epsilon}{\sqrt{n\log(1/\delta)}}, \delta)$-differentially private, and by the advanced composition theorem (Dwork et al., 2010b), the algorithm that outputs global models is (ϵ, δ)-differentially private after n steps.

When considering both local and global models as outputs of the algorithm, combining the above with the billboard lemma yields the desired JDP guarantee.

□

B.2. Proof of Theorem 4.5 (Generalization for PP-SGD)

Proof. Note that the algorithm updates be written jointly on $z_t = (w_t, \theta_t)$ as

$$z_t = z_{t-1} - \eta H^{-1} g_t, \tag{17}$$

$$\text{with } g_t = \begin{pmatrix} \frac{1}{N}\sum_{i=1}^N \nabla_w f_i(w_{t-1}, \theta_{i,t-1}, \xi_{i,t}) + \zeta_t \\ \frac{1}{N}\nabla_\theta f_1(w_{t-1}, \theta_{1,t-1}, \xi_{1,t}) \\ \vdots \\ \frac{1}{N}\nabla_\theta f_N(w_{t-1}, \theta_{N,t-1}, \xi_{N,t}) \end{pmatrix}, \tag{18}$$

where H is a pre-conditioner of the form

$$H = \begin{pmatrix} \frac{1}{\alpha} I_{d_w} & 0 \\ 0 & I_{N d_\theta} \end{pmatrix}. \tag{19}$$

Note that H satisties $\|z\|_H^2 = \langle Hz, z\rangle = \|z\|_\alpha^2$, as defined in (5).

Denote by $\mathcal{F}_t$ the sigma algebra spanned by random variables up to time t. Note that we have $\mathbb{E}[g_t|\mathcal{F}_{t-1}] = \nabla f(z_{t-1})$. We also define

$$g_t^* = \begin{pmatrix} \frac{1}{N}\sum_{i=1}^N \nabla_w f_i(w^*, \theta_i^*, \xi_{i,t}) + \zeta_t \\ \frac{1}{N}\nabla_\theta f_1(w^*, \theta_1^*, \xi_{1,t}) \\ \vdots \\ \frac{1}{N}\nabla_\theta f_N(w^*, \theta_N^*, \xi_{N,t}) \end{pmatrix}, \tag{20}$$

which satisfies $\mathbb{E}[g_t^*] = \nabla f(z^*) = 0$. We have

$$\begin{aligned} \mathbb{E}[\|z_t - z^*\|_H^2|\mathcal{F}_{t-1}] &= \|z_{t-1} - z^*\|_H^2 - 2\eta\langle H\,\mathbb{E}[H^{-1}g_t|\mathcal{F}_{t-1}], z_{t-1} - z^*\rangle + \eta^2\,\mathbb{E}[\|H^{-1}g_t\|_H^2|\mathcal{F}_{t-1}] \\ &= \|z_{t-1} - z^*\|_H^2 - 2\eta\langle \nabla f(z_{t-1}), z_{t-1} - z^*\rangle + \eta^2\,\mathbb{E}[\|g_t\|_{H^{-1}}^2|\mathcal{F}_{t-1}] \\ &\leq \|z_{t-1} - z^*\|_H^2 - 2\eta(f(z_{t-1}) - f(z^*)) + 2\eta^2\,\mathbb{E}[\|g_t - g_t^*\|_{H^{-1}}^2|\mathcal{F}_{t-1}] + 2\eta^2\,\mathbb{E}[\|g_t^*\|_{H^{-1}}^2] \quad (21) \\ &\leq \|z_{t-1} - z^*\|_H^2 - 2\eta(1 - 2\eta L_\alpha)(f(z_{t-1}) - f(z^*)) + 2\eta^2\sigma_{tot,\alpha}^2, \quad (22) \end{aligned}$$

where the first inequality follows by convexity of f and using $\|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$, while the second uses Lemma B.2 below to bound $\mathbb{E}[\|g_t - g_t^*\|_{H^{-1}}^2|\mathcal{F}_{t-1}]$, as well as the relation

$$\begin{aligned} \mathbb{E}[\|g_t^*\|_{H^{-1}}^2] &= \alpha\,\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \nabla_w f_i(w^*, \theta_i^*, \xi_t) + \zeta_t\right\|^2\right] + \sum_{i=1}^N \mathbb{E}\left[\left\|\frac{1}{N}\nabla_\theta f_i(w^*, \theta_i^*, \xi_t)\right\|^2\right] \\ &= \frac{\alpha\bar{\sigma}_w^2 + \bar{\sigma}_\theta}{N} + \alpha d_w \sigma_\zeta^2 = \sigma_{tot,\alpha}^2. \end{aligned}$$

Note that this quantity matches the definition in (6).

Assuming $\eta \leq 1/4L_\alpha$, and taking total expectations, (22) yields

$$\mathbb{E}[\|z_{t-1} - z^*\|_H^2] \leq \mathbb{E}[\|z_{t-1} - z^*\|_H^2] - \eta\,\mathbb{E}[f(z_{t-1}) - f(z^*)] + 2\eta^2\sigma_{tot,\alpha}^2.$$

Summing this inequality from $t = 0$ to $t = n$, we obtain

$$\mathbb{E}[f(\bar{z}_n) - f(z^*)] \leq \frac{\|z_0 - z^*\|_\alpha^2}{\eta n} + 2\eta\sigma_{tot,\alpha}^2. \tag{23}$$

Taking $\eta = \min\{\frac{1}{4L_\alpha}, \frac{\|z^*\|_\alpha}{\sqrt{n}\sigma_{tot,\alpha}}\}$, with $z_0 = 0$ yields the desired bound.

□

Lemma B.2. *Let g_t and g_t^* be defined as in (18) and (20). We have*

$$\mathbb{E}[\|g_t - g_t^*\|_{H^{-1}}^2] \le 2L_\alpha(f(z_{t-1}) - f(z_t^*)), \tag{24}$$

with $L_\alpha := L\max(\alpha, \frac{1}{N})$.

Proof. Define $\tilde{f}_i(w, \theta_i) := f_i(w, \theta_i, \xi_{i,t})$ and $\tilde{f}(z) := \frac{1}{N}\sum_{i=1}^N \tilde{f}_i(w, \theta_i)$, which satisfies $\mathbb{E}[\tilde{f}(z)] = f(z)$ and $\mathbb{E}[\nabla\tilde{f}(z)] = \nabla f(z)$. By L-smoothness (Assumption 4.3), we have for every i,

$$\begin{aligned}&\|\nabla\tilde{f}_i(w_{t-1}, \theta_{i,t-1}) - \nabla\tilde{f}_i(w^*, \theta_i^*)\|^2\\ &\le 2L\left(\tilde{f}_i(w_{t-1}, \theta_{i,t-1}) - \tilde{f}_i(w^*, \theta_i^*) + \langle\nabla_w\tilde{f}_i(w^*, \theta_i^*), w_{t-1} - w^*\rangle + \langle\nabla_\theta\tilde{f}_i(w^*, \theta_i^*), \theta_{t-1} - \theta^*\rangle\right)\end{aligned}$$

Taking the average over i, we get

$$\frac{1}{N}\sum_{i=1}^N \|\nabla\tilde{f}_i(w_{t-1}, \theta_{i,t-1}) - \nabla\tilde{f}_i(w^*, \theta_i^*)\|^2 \le 2L\left(\tilde{f}(z_{t-1}) - \tilde{f}(z^*) + \langle\nabla\tilde{f}(z^*), z_{t-1} - z^*\rangle\right).$$

Thus,

$$\begin{aligned}&\|g_t - g_t^*\|_{H^{-1}}^2\\ &= \alpha\left\|\frac{1}{N}\sum_{i=1}^N\left(\nabla_w\tilde{f}_i(w_{t-1}, \theta_{i,t-1}) - \nabla_w\tilde{f}_i(w^*, \theta_i^*)\right)\right\|^2 + \sum_{i=1}^N\left\|\frac{1}{N}\left(\nabla_\theta\tilde{f}_i(w_{t-1}, \theta_{i,t-1}) - \nabla_\theta\tilde{f}_i(w^*, \theta_i^*)\right)\right\|^2\\ &\le \frac{\alpha}{N}\sum_{i=1}^N\left\|\nabla_w\tilde{f}_i(w_{t-1}, \theta_{i,t-1}) - \nabla_w\tilde{f}_i(w^*, \theta_i^*)\right\|^2 + \frac{1}{N^2}\sum_{i=1}^N\left\|\nabla_\theta\tilde{f}_i(w_{t-1}, \theta_{i,t-1}) - \nabla_\theta\tilde{f}_i(w^*, \theta_i^*)\right\|^2\\ &\qquad\qquad\text{(by Cauchy-Schwarz inequality)}\\ &\le \max\left(\alpha, \frac{1}{N}\right) \times \frac{1}{N}\sum_{i=1}^N\|\nabla\tilde{f}_i(w_{t-1}, \theta_{i,t-1}) - \nabla\tilde{f}_i(w^*, \theta_i^*)\|^2\\ &\le 2L_\alpha\left(\tilde{f}(z_{t-1}) - \tilde{f}(z^*) + \langle\nabla\tilde{f}(z^*), z_{t-1} - z^*\rangle\right).\end{aligned}$$

with $L_\alpha := L\max(\alpha, \frac{1}{N})$. Taking the expectation, we get

$$\mathbb{E}\left[\|g_t - g_t^*\|_{H^{-1}}^2|\mathcal{F}_{t-1}\right] \le 2L_\alpha\left(f(z_{t-1}) - f(z^*) + \langle\nabla f(z^*), z_{t-1} - z^*\rangle\right) = 2L_\alpha\left(f(z_{t-1}) - f(z^*)\right).$$

□

B.3. Proof of Lemma 4.6 (effect of α and critical sample size)

Proof. We would like to make $(w + \theta_i)^\top x = v^\top x$ for all x, while minimizing $\|(w, \theta_{1:N})\|_\alpha$. This is achieved by solving the following minimization problem:

$$\begin{aligned}\min_{w,\theta_{1:N}} \quad & \frac{1}{\alpha}\|w\|^2 + \sum_i\|\theta_i\|^2\\ s.t. \quad & w + \theta_i = v\end{aligned}$$

To solve w, we minimize $\frac{1}{\alpha}\|w\|^2 + \sum_i\|v - w\|^2 = \frac{1}{\alpha}\|w\|^2 + N\|v - w\|^2$, which gives $w = \frac{\alpha N}{\alpha N + 1}v$ and thus $\theta_i = \frac{1}{\alpha N + 1}v$ for all i.

Therefore,

$$\|z^*\|_\alpha = \sqrt{\frac{1}{\alpha}\left(\frac{\alpha N}{\alpha N+1}\right)^2\|v\|^2 + N\left(\frac{1}{\alpha N+1}\right)^2\|v\|^2} = \sqrt{\frac{N}{\alpha N+1}}\|v\|.$$

Plugging this into (9), we see that the variance term takes the form specified in the lemma statement.

To see the effect of α to the variance term, we identify the condition for α, such that the following function is increasing in α:

$$\psi(\alpha) = \frac{a(\alpha+1)}{\alpha N+1} + \frac{b\alpha}{\alpha N+1}$$

for constants a, b. Its derivative can be calculated as follows:

$$\psi'(\alpha) = \frac{a(\alpha N+1) - a(\alpha+1)N + b(\alpha N+1) - b\alpha N}{(\alpha N+1)^2} = \frac{a(1-N)+b}{(\alpha N+1)^2}.$$

Thus, $\psi(\alpha)$ is increasing in α if and only if $a(N-1) \le b$. In our case, $a = \frac{\sigma^2}{n}$ and $b = c\frac{d_w G^2 \log(1/\delta)}{N\epsilon}$, and the condition $a(N-1) \le b$ is equivalent to

$$n \ge \frac{(N^2-N)\sigma^2\epsilon}{c d_w G^2 \log(1/\delta)}.$$

□

B.4. Proof of Theorem 5.1 (privacy of PPSGD with client sampling)

Proof. The proof follows the same lines as that of Proposition 4.1, but additionally leverages a standard argument of privacy amplification by subsampling to accomodate user sampling. Specifically, we use (Abadi et al., 2016, Theorem 1), which leverages Rényi differential privacy to provide an improved dependency on δ compared to the advanced composition theorem (Dwork et al., 2010b).

In order to apply (Abadi et al., 2016, Theorem 1), we rewrite the global updates as

$$w_t = w_{t-1} - \frac{\alpha\eta}{qM}\left(\sum_{i:b_{i,t}=1} \tilde{g}^t_{w,i} + qM\zeta_t\right).$$

This matches the form of the the updates in (Abadi et al., 2016), with total noise standard deviation $qM\sigma_\zeta$, which should equal σC in their notations. Then, by Abadi et al. (2016, Theorem 1), their condition on σ becomes:

$$\frac{\sigma_\zeta qM}{C} \ge c_2 \frac{q\sqrt{T\log(1/\delta)}}{\epsilon}.$$

guarantees that the global update mechanism after T iterations is (ϵ, δ)-DP. This corresponds to the condition in the statement. Combining this with the billboard lemma (Lemma B.1) yields the desired JDP guarantee on the full algorithm output.

□

B.5. Proof of Theorem 5.2 (Generalization of PP-SGD with client sampling)

Proof. Note that the algorithm updates be written jointly on $z_t = (w_t, \theta_t)$ as

$$z_t = z_{t-1} - \eta H^{-1} g_t, \tag{25}$$

$$\text{with } g_t = \frac{1}{qM}\begin{pmatrix} \sum_{i=1}^N b_{i,t}\sum_{k=1}^{m_i} \nabla_w f_i(w_{t-1}, \theta_{i,t-1}, \xi^{(k)}_{i,t}) + qM\zeta_t \\ b_{1,t}\sum_{k=1}^{m_1} \nabla_\theta f_1(w_{t-1}, \theta_{1,t-1}, \xi^{(k)}_{1,t}) \\ \vdots \\ b_{N,t}\sum_{k=1}^{m_N} \nabla_\theta f_N(w_{t-1}, \theta_{N,t-1}, \xi^{(k)}_{N,t}) \end{pmatrix}, \tag{26}$$

$$\tag{27}$$

with H as in (19). We have $\mathbb{E}[g_t|\mathcal{F}_{t-1}] = \nabla f^m(z_{t-1})$. We also define

$$g_t^* = \frac{1}{qM}\begin{pmatrix} \sum_{i=1}^N b_{i,t}\sum_{k=1}^{m_i} \nabla_w f_i(w^*, \theta_i^*, \xi_{i,t}^{(k)}) + qM\zeta_t \\ b_{1,t}\sum_{k=1}^{m_1} \nabla_\theta f_1(w^*, \theta_1^*, \xi_{1,t}^{(k)}) \\ \vdots \\ b_{N,t}\sum_{k=1}^{m_N} \nabla_\theta f_N(w^*, \theta_N^*, \xi_{N,t}^{(k)}) \end{pmatrix}, \tag{28}$$

which satisfies $\mathbb{E}[g_t^*] = \nabla f^m(z^*) = 0$. We have

$$\begin{aligned} \mathbb{E}[\|z_t - z^*\|_H^2|\mathcal{F}_{t-1}] &= \|z_{t-1} - z^*\|_H^2 - 2\eta\langle H\,\mathbb{E}[H^{-1}g_t|\mathcal{F}_{t-1}], z_{t-1} - z^*\rangle + \eta^2\,\mathbb{E}[\|H^{-1}g_t\|_H^2|\mathcal{F}_{t-1}] \\ &= \|z_{t-1} - z^*\|_H^2 - 2\eta\langle \nabla f^m(z_{t-1}), z_{t-1} - z^*\rangle + \eta^2\,\mathbb{E}[\|g_t\|_{H^{-1}}^2|\mathcal{F}_{t-1}] \\ &\le \|z_{t-1} - z^*\|_H^2 - 2\eta(f^m(z_{t-1}) - f^m(z^*)) + 2\eta^2\,\mathbb{E}[\|g_t - g_t^*\|_{H^{-1}}^2|\mathcal{F}_{t-1}] + 2\eta^2\,\mathbb{E}[\|g_t^*\|_{H^{-1}}^2]. \end{aligned} \tag{29}$$

By Lemma B.3,

$$\mathbb{E}[\|g_t - g_t^*\|_{H^{-1}}^2|\mathcal{F}_{t-1}] \le 2L_{m,\alpha}\left(f^m(z_{t-1}) - f^m(z^*)\right) \tag{30}$$

where $L_{m,\alpha} := L\max\left(\alpha + \frac{\alpha m_{\max}}{qM}, \frac{m_{\max}}{qM}\right)$, and

$$\begin{aligned} \mathbb{E}[\|g_t^*\|_{H^{-1}}^2] &= \alpha\,\mathbb{E}\left[\left\|\frac{1}{qM}\sum_{i=1}^N b_{i,t}\sum_{k=1}^{m_i}\nabla_w f_i(w^*, \theta_i^*, \xi_{i,t}^{(k)}) + \zeta_t\right\|^2\right] + \sum_{i=1}^N \mathbb{E}\left[\left\|\frac{b_{i,t}}{qM}\sum_{k=1}^{m_i}\nabla_\theta f_i(w^*, \theta_i^*, \xi_{i,t}^{(k)})\right\|^2\right] \\ &= \frac{\alpha}{q^2M^2}\sum_{i=1}^N \mathbb{E}\left[\left\|b_{i,t}\sum_{k=1}^{m_i}\nabla_w f_i(w^*, \theta_i^*, \xi_{i,t}^{(k)}) - qm_i\nabla_w f_i(w^*, \theta_i^*)\right\|^2\right] + \frac{1}{qM^2}\sum_{i=1}^N m_i\sigma_{\theta,i}^2 + \alpha d_w\sigma_\zeta^2 \\ &= \frac{\alpha}{q^2M^2}\sum_{i=1}^N \mathbb{E}\left[\left\|b_{i,t}\sum_{k=1}^{m_i}\nabla_w f_i(w^*, \theta_i^*, \xi_{i,t}^{(k)}) - b_{i,t}m_i\nabla_w f_i(w^*, \theta_i^*)\right\|^2\right] \qquad (31) \\ &\quad + \frac{\alpha}{q^2M^2}\sum_{i=1}^N \mathbb{E}\left[\|(b_{i,t} - q)m_i\nabla_w f_i(w^*, \theta_i^*)\|^2\right] + \frac{1}{qM^2}\sum_{i=1}^N m_i\sigma_{\theta,i}^2 + \alpha d_w\sigma_\zeta^2 \\ &\le \frac{\alpha}{qM^2}\sum_{i=1}^N m_i\sigma_{w,i}^2 + \frac{\alpha}{qM^2}\sum_{i=1}^N m_i^2\,\mathbb{E}[(b_{i,t} - q)^2]\,\|\nabla_w f_i(w^*, \theta_i^*)\|^2 + \frac{\bar\sigma_{\theta,q}^2}{qM} + \alpha d_w\sigma_\zeta^2 \\ &\le \frac{\alpha\bar\sigma_{w,m}^2 + \bar\sigma_{\theta,m}^2}{qM} + \alpha d_w\sigma_\zeta^2 + \frac{\alpha q(1-q)}{qM^2}\sum_{i=1}^N m_i^2\,\|\nabla_w f_i(w^*, \theta_i^*)\|^2 \qquad (32) \\ &= \sigma_{m,\alpha}^2. \end{aligned}$$

Assuming $\eta \le 1/4L_{m,\alpha}$, and taking total expectations, (29) yields

$$\mathbb{E}[\|z_{t-1} - z^*\|_H^2] \le \mathbb{E}[\|z_{t-1} - z^*\|_H^2] - \eta\,\mathbb{E}[f^m(z_{t-1}) - f^m(z^*)] + 2\eta^2\sigma_{m,\alpha}^2.$$

Summing this inequality from $t = 0$ to $t = T$, we obtain

$$\mathbb{E}[f^m(\bar z_T) - f^m(z^*)] \le \frac{\|z_0 - z^*\|_\alpha^2}{\eta T} + 2\eta\sigma_{m,\alpha}^2. \tag{33}$$

Taking $\eta = \min\{\frac{1}{4L_{m,\alpha}}, \frac{\|z^*\|_\alpha}{\sqrt{T}\sigma_{m,\alpha}}\}$, with $z_0 = 0$ yields the desired bound. □

Lemma B.3. *Let g_t and g_t^* be defined as in (26) and (28). We have*

$$\mathbb{E}[\|g_t - g_t^*\|_{H^{-1}}^2] \le 2L_{m,\alpha}\left(f^m(z_{t-1}) - f^m(z^*)\right) \tag{34}$$

with $L_{m,\alpha} := L\max\left(\alpha + \frac{\alpha m_{\max}}{qM}, \frac{m_{\max}}{qM}\right)$.

Proof. Define $\tilde{f}_i^{(k)}(w,\theta_i) := f_i(w,\theta_i,\xi_{i,t}^{(k)})$ and $\tilde{f}(z) := \frac{1}{qM}\sum_{i=1}^{N} b_{i,t}\sum_{k=1}^{m_i}\tilde{f}_i^{(k)}(w,\theta_i)$, which satisfies $\mathbb{E}[\tilde{f}(z)] = f^m(z)$ and $\mathbb{E}[\nabla\tilde{f}(z)] = \nabla f^m(z)$. By L-smoothness (Assumption 4.3), we have for every i,k,

$$\begin{aligned}
&\|\nabla\tilde{f}_i^{(k)}(w_{t-1},\theta_{i,t-1}) - \nabla\tilde{f}_i^{(k)}(w^*,\theta_i^*)\|^2 \\
&\le 2L\left(\tilde{f}_i^{(k)}(w_{t-1},\theta_{i,t-1}) - \tilde{f}_i^{(k)}(w^*,\theta_i^*) + \langle\nabla_w\tilde{f}_i^{(k)}(w^*,\theta_i^*), w_{t-1}-w^*\rangle + \langle\nabla_\theta\tilde{f}_i^{(k)}(w^*,\theta_i^*), \theta_{t-1}-\theta^*\rangle\right)
\end{aligned}$$

Taking the weighted sum over i,k with weights $\frac{b_{i,t}}{qM}$, we get

$$\frac{1}{qM}\sum_{i=1}^{N} b_{i,t}\sum_{k=1}^{m_i}\|\nabla\tilde{f}_i^{(k)}(w_{t-1},\theta_{i,t-1}) - \nabla\tilde{f}_i^{(k)}(w^*,\theta_i^*)\|^2 \le 2L\left(\tilde{f}(z_{t-1}) - \tilde{f}(z^*) + \langle\nabla\tilde{f}(z^*), z_{t-1}-z^*\rangle\right).$$

Taking the expectation on two sides, we further get

$$\begin{aligned}
\frac{1}{M}\sum_{i=1}^{N}\sum_{k=1}^{m_i}\mathbb{E}\left[\|\nabla\tilde{f}_i^{(k)}(w_{t-1},\theta_{i,t-1}) - \nabla\tilde{f}_i^{(k)}(w^*,\theta_i^*)\|^2\right] &\le 2L\left(f^m(z_{t-1}) - f^m(z^*) + \langle\nabla f^m(z^*), z_{t-1}-z^*\rangle\right) \\
&= 2L\left(f^m(z_{t-1}) - f^m(z^*)\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{E}\left[\|g_t - g_t^*\|_{H^{-1}}^2\right] \\
&= \alpha\,\mathbb{E}\left[\left\|\sum_{i=1}^{N}\frac{b_{i,t}}{qM}\sum_{k=1}^{m_i}\left(\nabla_w\tilde{f}_i^{(k)}(w_{t-1},\theta_{i,t-1}) - \nabla_w\tilde{f}_i^{(k)}(w^*,\theta_i^*)\right)\right\|^2\right] \\
&\quad + \sum_{i=1}^{N}\mathbb{E}\left[\left\|\frac{b_{i,t}}{qM}\sum_{k=1}^{m_i}\left(\nabla_\theta\tilde{f}_i^{(k)}(w_{t-1},\theta_{i,t-1}) - \nabla_\theta\tilde{f}_i^{(k)}(w^*,\theta_i^*)\right)\right\|^2\right] \\
&= \alpha\,\mathbb{E}\left[\left\|\sum_{i=1}^{N}\frac{1}{M}\sum_{k=1}^{m_i}\left(\nabla_w\tilde{f}_i^{(k)}(w_{t-1},\theta_{i,t-1}) - \nabla_w\tilde{f}_i^{(k)}(w^*,\theta_i^*)\right)\right\|^2\right] \\
&\quad + \alpha\,\mathbb{E}\left[\sum_{i=1}^{N}\frac{(b_{i,t}-q)^2}{q^2M^2}\left\|\sum_{k=1}^{m_i}\left(\nabla_w\tilde{f}_i^{(k)}(w_{t-1},\theta_{i,t-1}) - \nabla_w\tilde{f}_i^{(k)}(w^*,\theta_i^*)\right)\right\|^2\right] \\
&\quad + \sum_{i=1}^{N}\frac{1}{qM^2}\mathbb{E}\left[\left\|\sum_{k=1}^{m_i}\left(\nabla_\theta\tilde{f}_i^{(k)}(w_{t-1},\theta_{i,t-1}) - \nabla_\theta\tilde{f}_i^{(k)}(w^*,\theta_i^*)\right)\right\|^2\right] \\
&\le \alpha\sum_{i=1}^{N}\frac{1}{M}\sum_{k=1}^{m_i}\mathbb{E}\left[\left\|\left(\nabla_w\tilde{f}_i^{(k)}(w_{t-1},\theta_{i,t-1}) - \nabla_w\tilde{f}_i^{(k)}(w^*,\theta_i^*)\right)\right\|^2\right] \\
&\quad + \alpha\sum_{i=1}^{N}\frac{(1-q)m_i}{qM^2}\sum_{k=1}^{m_i}\mathbb{E}\left[\left\|\nabla_w\tilde{f}_i^{(k)}(w_{t-1},\theta_{i,t-1}) - \nabla_w\tilde{f}_i^{(k)}(w^*,\theta_i^*)\right\|^2\right] \\
&\quad + \sum_{i=1}^{N}\frac{m_i}{qM^2}\sum_{k=1}^{m_i}\mathbb{E}\left[\left\|\nabla_\theta\tilde{f}_i^{(k)}(w_{t-1},\theta_{i,t-1}) - \nabla_\theta\tilde{f}_i^{(k)}(w^*,\theta_i^*)\right\|^2\right] \\
&\le \max\left(\alpha + \frac{\alpha m_{\max}}{qM}, \frac{m_{\max}}{qM}\right)\times\frac{1}{M}\sum_{i=1}^{N}\sum_{k=1}^{m_i}\mathbb{E}\left[\left\|\nabla\tilde{f}_i^{(k)}(w_{t-1},\theta_{i,t-1}) - \nabla\tilde{f}_i^{(k)}(w^*,\theta_i^*)\right\|^2\right] \\
&\le \max\left(\alpha + \frac{\alpha m_{\max}}{qM}, \frac{m_{\max}}{qM}\right)\times 2L\left(f^m(z_{t-1}) - f^m(z^*)\right) \\
&= 2L_{m,\alpha}\left(f^m(z_{t-1}) - f^m(z^*)\right).
\end{aligned}$$

□

Algorithm 3 PPSGD with client sampling, average user performance

```
Input: $q$: client sampling probability,
       $m_i$: minibatch sizes, $\eta$: step size,
       $\alpha$: global/local ratio, $\sigma_\zeta$: privacy noise level,
       $C$: clipping parameter.
Initialize $w_0 = \theta_0 = 0$.
for $t = 1$ to $T$ do
    Sample $b_{i,t} \sim Ber(q)$
    for all clients $i$ with $b_{i,t} = 1$ in parallel do
        Sample a minibatch $\{\xi_{i,t}^{(k)}\}_{k=1}^{m_i} \sim P_i^{\otimes m_i}$
        Compute $g_{\theta,i}^t = \frac{1}{m_i}\sum_{k=1}^{m_i} \nabla_\theta f_i(w_{t-1}, \theta_{i,t-1}, \xi_{i,t}^{(k)})$
                $g_{w,i}^t = \frac{1}{m_i}\sum_{k=1}^{m_i} \nabla_w f_i(w_{t-1}, \theta_{i,t-1}, \xi_{i,t}^{(k)})$
        Update $\theta_{i,t} = \theta_{i,t-1} - \frac{\eta}{qN} g_{\theta,i}^t$
        Clip gradient $\tilde{g}_{w,i}^t = g_{w,i}^t / \max(1, \frac{\|g_{w,i}^t\|}{C})$
        Send $\tilde{g}_{w,i}^t$ to the server
    end for
    Sample $\zeta_t \sim \mathcal{N}(0, \sigma_\zeta^2 I_{d_w})$
    Update $w_t = w_{t-1} - \alpha\eta(\frac{1}{qN}\sum_{i:b_{i,t}=1} \tilde{g}_{w,i}^t + \zeta_t)$
end for
```

B.6. Optimizing average user risk with heterogeneous sample sizes

In this section, we study a variant of Algorithm 2 which optimizes the average risk f instead of the weighted risk f^m in (11), as described in Section 5. The algorithm is described in Algorithm 3, and we provide its generalization and privacy guarantees below.

Theorem B.4 (Generalization and privacy for Algorithm 3)**.** *Under Assumptions 4.2, 4.3, and 4.4, let* $z^* = (w^*, \theta_{1:N}{}^*)$ *be any minimizer of* f*,* $L_\alpha = L\max\{\alpha + \frac{\alpha}{N}, \frac{1}{N}\}$*, and*

$$\sigma_{m,\alpha}^2 := \frac{\alpha\bar{\sigma}_{w,m}^2 + \alpha\tilde{\sigma}_{w,m}^2 + \bar{\sigma}_{\theta,m}^2}{qN} + \alpha d_w \sigma_\zeta^2 \le \frac{\alpha\bar{\sigma}_w^2 + \bar{\sigma}_\theta^2}{qNm_{\min}} + \frac{\alpha\tilde{\sigma}_{w,m}^2}{qN} + \alpha d_w \sigma_\zeta^2, \tag{35}$$

where $\bar{\sigma}_{w,m}^2 := \frac{1}{N}\sum_i \frac{\sigma_{w,i}^2}{m_i}$ *and* $\bar{\sigma}_{\theta,m}^2 := \frac{1}{N}\sum_i \frac{\sigma_{\theta,i}^2}{m_i}$ *and* $\tilde{\sigma}_{w,m}^2 := \frac{1}{N}\sum_i q(1-q)\|\nabla_w f_i(w^*, \theta_i^*)\|^2$.

With $\eta = \min\{\frac{1}{4L_\alpha}, \frac{\|z^*\|_\alpha}{\sqrt{T}\sigma_{m,\alpha}}\}$ *and* $C = G$*, Algorithm 3 satisfies*

$$\mathbb{E}[f(\bar{z}_T) - f(z^*)] \le \frac{4L_\alpha\|z^*\|_\alpha^2}{T} + 3\frac{\sigma_{m,\alpha}\|z^*\|_\alpha}{\sqrt{T}},$$

with $\bar{z}_T = \frac{1}{T}\sum_{t=0}^{T-1} z_t$*.*

For c_1, c_2 *as in Theorem 5.1, and for any* $\epsilon < c_1 q^2 T$*, if we take*

$$\sigma_\zeta = c_2 \frac{C\sqrt{T\log(1/\delta)}}{N\epsilon},$$

then Algorithm 3 is (ϵ, δ)*-JDP, and the generalization guarantee becomes:*

$$\mathbb{E}[f(\bar{z}_T) - f(z^*)] \lesssim \frac{L_\alpha\|z^*\|_\alpha^2}{T} + \|z^*\|_\alpha\sqrt{\frac{\alpha\bar{\sigma}_{w,m}^2 + \alpha\tilde{\sigma}_{w,m}^2 + \bar{\sigma}_{\theta,m}^2}{qNT} + \frac{\alpha d_w G^2\log(\frac{1}{\delta})}{N^2\epsilon^2}}. \tag{36}$$

Proof. The proof can be obtained by replacing f_i by $\tilde{f}_i = \frac{M}{m_i N} f_i$ in the proofs of Theorem 5.2 and Theorem 5.1. □