ACCELERATING DIFFERENTIALLY PRIVATE FEDERATED LEARNING VIA ADAPTIVE EXTRAPOLATION

Shokichi Takakura, Seng Pei Liew & Satoshi Hasegawa
LY Corporation
{stakakur,sliew,satoshi.hasegawa}@lycorp.co.jp

ABSTRACT

DP-FedAvg is one of the most popular algorithms for federated learning (FL) with differential privacy (DP), but it is known to suffer from the slow convergence in the presence of heterogeneity among clients' data. Most of the existing methods to accelerate DP-FL require 1) additional hyperparameters or 2) additional computational cost for clients. This is not desirable since 1) hyperparameter tuning is computationally expensive and data-dependent choice of hyperparameters raises the risk of privacy leakage, and 2) clients are often resource-constrained. To address this issue, we propose DP-FedEXP, which adaptively selects the global step size based on the diversity of the local updates without requiring any additional hyperparameters or client computational cost. We show that DP-FedEXP provably accelerates the convergence of DP-FedAvg and it empirically outperforms existing methods tailored for DP-FL.

1 INTRODUCTION

Federated learning (FL) (Konečný et al., 2017) with differential privacy (DP) (Dwork et al., 2006) has been intensively studied due to the growing concern for privacy in the field of machine learning. A practical approach to incorporate DP to the FL framework is DP-FedAvg (McMahan et al., 2017). Unfortunately, (DP-)FedAvg has been known to suffer from slow convergence in the presence of data heterogeneity across clients. This issue is known as *the client drift error* (Karimireddy et al., 2019).

To deal with data heterogeneity, a line of work has studied variance reduction techniques in (non-private) FL setting (Karimireddy et al., 2020a;b; Mitra et al., 2021). Extending the above techniques to the DP setting, DP-SCAFFOLD (Noble et al., 2022) has been proposed and shown to achieve improved convergence guarantee. Although the above methods enjoy theoretically favorable properties, they require clients to be stateful and additional computational cost in clients. This is impractical since clients are often resource-constrained.

Another line of work has sought to accelerate the convergence of (DP-)FedAvg by regarding the local updates as pseudo-gradients and updating the global model using global optimization algorithms such as Adam (Kingma & Ba, 2015) with additional hyperparameters such as global step size (Reddi et al., 2021). Although the performance crucially relies on the choice of the hyperparameters, it is difficult to obtain the optimal hyperparameters in the DP settings since hyperparameter tuning on sensitive data leads to additional privacy leakage Papernot & Steinke (2021). Furthermore, it is highly costly in practice to tune the hyperparameters in the FL setting, since the data is distributed across clients.

To develop an effective and practical DP-FL algorithm, we pose the following question:

Can DP-FL be accelerated under heterogeneity of client data without any additional hyperparameters and computational cost for clients?

In this paper, to address the above question, we propose DP-FedEXP by incorporating FedEXP (Jhunjhunwala et al., 2023), which adaptively determines the global step size to the heterogeneity of the local updates, into the DP-FL framework in a non-trivial way. Specifically, we consider the two different scenarios of DP: Local Differential Privacy (LDP) and Central Differential Privacy

(CDP). We found that the step size formula for FedEXP cannot be directly extended in both cases. Thus, we carefully design the step size formula for LDP and CDP and develop a simple but effective framework to accelerate the convergence of existing DP-FL algorithms. We would like to emphasize that our proposed method is *orthogonal* to existing works which try to accelerate (DP-)FL by modifying the local training procedure (Li et al., 2020; Karimireddy et al., 2020b; Noble et al., 2022; Shi et al., 2023) and thus, it can be combined with them to further improve the performance.

Our contribution can be summarized as follows:

- We propose LDP-FedEXP and CDP-FedEXP with simple but effective parameter-free step size rules in DP-FL.
- We provide formal differential privacy guarantee and convergence guarantees for general non-convex objectives. We prove that the proposed method provably accelerates the convergence in the presence of data heterogeneity.
- In the numerical experiments, we show that DP-FedEXP outperforms existing algorithms in utility while preserving the privacy guarantee.

2 Problem Settings and Preliminaries

2.1 Problem Settings

Federated Learning We consider the following optimization problem with M clients:

$$\min_{w \in \mathbb{R}^d} F(w) := \frac{1}{M} \sum_{i=1}^{M} f_i(w), \tag{1}$$

where $w \in \mathbb{R}^d$, M is the number of clients and $f_i(w) := \frac{1}{|\mathcal{D}_i|} \sum_{d_i \in \mathcal{D}_i} l(w, d_i)$ is the loss function of the i-th client computed on a loss function l and the local dataset $\mathcal{D}_i$.

Differential Privacy In this paper, we consider two scenarios of differential privacy: Central Differential Privacy (CDP) and Local Differential Privacy (LDP) . In the CDP setting, we assume that the central server is trusted while we do not assume any trusted server in the LDP setting. Here, we provide the formal definitions of (ε, δ)-CDP and (ε, δ)-LDP.

Definition 2.1 (Central Differential Privacy Dwork et al. (2014))**.** Let $\mathcal{X}$ be the set of all possible client datasets. A central randomized mechanism $\mathcal{Q} : \mathcal{X}^M \to \mathcal{Y}$ satisfies (ε, δ)-CDP if for any two neighboring inputs $x, x' \in \mathcal{X}^M$, which differ in one client dataset, we have

$$\forall S \subset \mathcal{Y} : \Pr[\mathcal{Q}(x) \in S] \leq e^{\varepsilon} \Pr[\mathcal{Q}(x') \in S] + \delta.$$

Definition 2.2 (Local Differential Privacy Kasiviswanathan et al. (2011))**.** Let $\mathcal{X}$ be the set of all possible client datasets. A local randomized mechanism $\mathcal{R} : \mathcal{X} \to \mathcal{Y}$ satisfies (ε, δ)-LDP if for any two inputs $x, x' \in \mathcal{X}$, we have

$$\forall S \subset \mathcal{Y} : \Pr[\mathcal{R}(x) \in S] \leq e^{\varepsilon} \Pr[\mathcal{R}(x') \in S] + \delta.$$

If $\delta = 0$, $\mathcal{R}$ is called to satisfy *pure differential privacy*.

2.2 DP-FedAvg

At round t in DP-FedAvg (McMahan et al., 2017), the server sends the global model $w^{(t-1)}$ to all clients. Then, each client performs τ steps of local training $w_i^{(t-1,0)} := w^{(t-1)}, w_i^{(t-1,k)} := w_i^{(t-1,k-1)} - \eta_l \nabla f_i(w_i^{(t-1,k-1)})$ $(k = 1 \dots \tau)$ using (stochastic) gradient descent with step size η_l and computes the local update $\tilde{\Delta}_i^{(t)} := w_i^{(t-1,\tau)} - w^{(t-1)}$. To bound the sensitivity of the local updates, each client i applies clipping to their local update $\Delta_i^{(t)} := \min\{C/\|\tilde{\Delta}_i^{(t)}\|, 1\} \cdot \tilde{\Delta}_i^{(t)}$ with threshold $C > 0$. Then, each client sends the central server the local update $\Delta_i^{(t)}$ in the CDP setting and the randomized update $c_i^{(t)} = \Delta_i^{(t)} + \varepsilon_i^{(t)}$ $(\varepsilon_i^{(t)} \sim \mathcal{N}(0, \sigma^2))$ in the LDP setting. The central

server aggregates the local updates as follows:

$$\begin{cases} \bar{c}^{(t)} := \frac{1}{M}\sum_{i=1}^{M} c_i^{(t)} & \text{(LDP setting)}, \\ \bar{c}^{(t)} := \frac{1}{M}\sum_{i=1}^{M} \Delta_i^{(t)} + \varepsilon^{(t)} \quad (\varepsilon^{(t)} \sim \mathcal{N}(0, \sigma^2/M)) & \text{(CDP setting)}. \end{cases}$$

Here, we consider Gaussian mechanism as a local randomizer in the LDP setting but our proposed framework can be applied to PrivUnit (Bhowmick et al., 2018), which is known to satisfy pure differential privacy and achieve the asymptotically optimal trade-off between privacy and utility Bhowmick et al. (2018); Asi et al. (2022). See Appendix D for details.

In DP-FedAvg, the server updates the global model by just adding the averaged local update as $w^{(t+1)} = w^{(t)} + \bar{c}^{(t)}$. To accelerate the convergence, several works deal with the noisy local updates as pseudo-gradients and update the global model as $w^{(t+1)} = w^{(t)} + \eta_g \bar{c}^{(t)}$, where η_g is a global step size (Reddi et al., 2021; Noble et al., 2022). To ensure the convergence, η_g should be chosen carefully. However, it is difficult in practice to tune such a hyperparameter with formal DP guarantee since hyperparameter tuning is computationally expensive and requires additional privacy budget (Papernot & Steinke, 2021). To fill the gap between the theory and practice, it is desirable to determine the step size *in an adaptive manner*.

2.3 FedEXP

In the context of non-DP federated learning, FedEXP (Jhunjhunwala et al., 2023) and FedEXProx (Li et al., 2024) have been proposed to determine the global step size adaptively to the heterogeneity of the local updates. Following the adaptive step size rule of POCS (Pierra, 1984), they define the global step size as

$$\eta_g^{(t)} := \frac{\frac{1}{M}\sum_{i=1}^{M}\left\|\Delta_i^{(t)}\right\|^2}{\left\|\bar{\Delta}^{(t)}\right\|^2}, \tag{2}$$

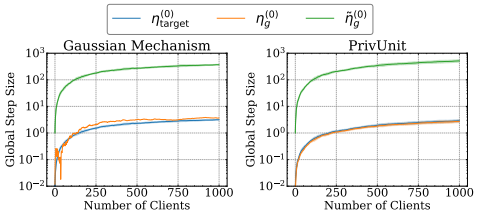

Figure 1: The adaptive step size $\eta_g^{(0)}$ at initialization in the LDP setting. Our proposed step size is close to the target step size $\eta_{\text{target}}^{(0)}$ for both Gaussian mechanism and PrivUnit case.

where $\bar{\Delta}^{(t)} = \frac{1}{M}\sum_{i=1}^{M}\Delta_i^{(t)}$ is the average of the local updates. Here, we follow the formula in Li et al. (2024) and omit the coefficient $1/2$ and a small constant added to the denominator, which appear in Jhunjhunwala et al. (2023) since the convergence analysis in Jhunjhunwala et al. (2023) does not require these factors. In the case of $\tau = 1$, the above formula is reduced to $\frac{\frac{1}{M}\sum_{i=1}^{M}\left\|\nabla f_i(w^{(t)})\right\|^2}{\left\|\nabla F(w^{(t)})\right\|^2}$, which is known as a measure of the heterogeneity among clients (Haddadpour & Mahdavi, 2019; Wang et al., 2020). Thus, FedEXP adaptively determines the global step size based on the diversity of the clients' data. Although FedEXP has been shown to accelerate the convergence in the non-private setting, it is still unclear how to extend the algorithm to the DP setting.

3 Proposed Method: DP-FedEXP

In this section, we propose DP-FedEXP (LDP-FedEXP and CDP-FedEXP), which extend FedEXP to the LDP and CDP setting in a non-trivial way.

3.1 LDP-FedEXP

3.1.1 Naive Implementation of FedEXP with Noisy Updates

In the setting of LDP, the server can only access the noisy updates $c_i^{(t)}$. Extending equation 2 to the DP setting naively, we obtain the following formula:

$$\tilde{\eta}_g^{(t)} := \frac{\frac{1}{M}\sum_{i=1}^{M}\left\|c_i^{(t)}\right\|^2}{\left\|\bar{c}^{(t)}\right\|^2}. \tag{3}$$

Unfortunately, Fig. 1 shows that $\tilde{\eta}_g^{(t)}$ is extremely large and causes instability in the training process.

To investigate the reason of this phenomenon, let us evaluate the expectation of the numerator in the above formula. We have $\mathbb{E}[\frac{1}{M}\sum_{i=1}^M \|c_i^{(t)}\|^2] = \frac{1}{M}\sum_{i=1}^M \|\Delta_i^{(t)}\|^2 + d\sigma^2$. Since the noise scale σ is relatively large in the LDP setting, the noise term $d\sigma^2$ dominates the numerator. Furthermore, since the noise term does not depend on the number of clients M, increasing the number of clients does not help to stabilize the training.

3.1.2 Step Size Formula for Gaussian Mechanism

To develop a practical step size rule in the DP setting, let us consider the following *approximate projection condition*:

$$\frac{1}{M}\sum_{i=1}^M \left\| w_i^{(t,\tau)} - w^* \right\|^2 = (1-\alpha)\left\| w^{(t)} - w^* \right\|^2, \tag{4}$$

for some $0 \leq \alpha \leq 1$ (Jhunjhunwala et al., 2023), where w^* is a optimal solution of problem equation 1. Intuitively, this condition implies that the parameters of the local models are closer to the optimal solution on average after τ steps of local training. Under the above condition, the distance between updated model and the optimal model is evaluated as

$$\left\| w^{(t+1)} - w^* \right\|^2 \simeq (1-\alpha\eta_g)\left\| w^{(t)} - w^* \right\|^2 - \eta_g \frac{1}{M}\sum_{i=1}^M \left\| \Delta_i^{(t)} \right\|^2 + \eta_g^2 \left\| \bar{c}^{(t)} \right\|^2,$$

for sufficiently large d with high-probability. Here, we ignore the effect of clipping for simplicity. See Lemma C.4 for the detailed derivation. To ensure that the distance between the global model and the optimal model decreases for any $\|w^{(t)} - w^*\|^2$, we need to set the global step size as

$$\eta_g \leq \eta_{\text{target}}^{(t)} := \frac{\frac{1}{M}\sum_{i=1}^M \left\| \Delta_i^{(t)} \right\|^2}{\left\| \bar{c}^{(t)} \right\|^2} \tag{5}$$

but we cannot compute $\eta_{\text{target}}^{(t)}$ since the server cannot access $\Delta_i^{(t)}$ directly. Instead of the exact calculation of $\frac{1}{M}\sum_{i=1}^M \|\Delta_i^{(t)}\|^2$, we propose to use its unbiased estimator $\frac{1}{M}\sum_{i=1}^M \|c_i^{(t)}\|^2 - d\sigma^2$. That is, the global step size for LDP-FedEXP is given by

$$\eta_g^{(t)} := \max\left\{ 1, \frac{\frac{1}{M}\sum_{i=1}^M \left\| c_i^{(t)} \right\|^2 - d\sigma^2}{\left\| \bar{c}^{(t)} \right\|^2} \right\}. \tag{6}$$

Here, we take the maximum of 1 and the bias-corrected step size to ensure the acceleration of the convergence. As shown in Fig. 1, $\eta_g^{(t)}$ is close to $\eta_{\text{target}}^{(t)}$ for large M. Using the above formula, LDP-FedEXP updates the global model as $w^{(t+1)} := w^{(t)} + \eta_g^{(t)}\bar{c}^{(t)}$. We show the entire training process in Algorithm 1.

3.2 CDP-FedEXP

In the CDP setting, the server can calculate equation 5 but it does not satisfy DP. Since $\|\bar{c}^{(t)}\|$ can be arbitrarily small and the sensitivity of $\eta_{\text{target}}^{(t)}$ is not bounded, we cannot apply Gaussian mechanism to Eq. equation 5 directly. Thus, we propose the following formula:

$$\eta_g^{(t)} := \max\left\{ 1, \frac{\frac{1}{M}\sum_{i=1}^M \left\| \Delta_i^{(t)} \right\|^2 + \xi^{(t)}}{\left\| \bar{c}^{(t)} \right\|^2} \right\}, \tag{7}$$

where $\xi^{(t)}$ follows $\mathcal{N}(0, \sigma_\xi^2)$. We show the entire training process in Algorithm 2.

4 THEORETICAL ANALYSIS

In this section, we show that the proposed methods provably accelerate the DP-FedAvg while maintaining the privacy guarantee.

4.1 PRIVACY

Here, we provide the formal privacy guarantee of LDP-FedEXP and CDP-FedEXP.

Proposition 4.1 (LDP case)**.** *LDP-FedEXP satisfies the same privacy guarantee as DP-FedAvg in the LDP setting. That is, the local computation at each client in LDP-FedEXP with Gaussian mechanism satisfies (ε, δ)-LDP, where $\rho = 2C^2/\sigma^2$ and $\varepsilon = \alpha\rho + \log(1/\delta)/(\alpha - 1)$ for any $\delta \in (0, 1)$ and $\alpha \in (1, \infty)$.*

Proposition 4.2 (CDP case)**.** *The entire training process of CDP-FedEXP satisfies (ε, δ)-CDP, where $\rho = 2C^2T/M\sigma^2, \rho_\xi = C^4T/2M^2\sigma_\xi^2$ and $\varepsilon = \alpha(\rho + \rho_\xi) + \log(1/\delta)/(\alpha - 1)$ for any $\delta \in (0, 1)$ and $\alpha \in (1, \infty)$.*

See Appendix E for details. For LDP case, the privacy guarantee of LDP-FedEXP is the same as that of LDP-FedAvg since we use the same mechanism for the local computation. For CDP case, additional privacy budget $\alpha\rho_\xi$ is required for privatizing the numerator in the step size formula. However, as shown in the utility analysis, it is sufficient to set $\sigma_\xi = d\sigma^2/M$ and we have $\rho_\xi = C^4T/2d^2\sigma^4 = O(\rho^2M^2/Td^2)$. Thus, the additional privacy budget consumption is negligible if $\rho = O(1)$ and $T \cdot d^2 \gg M^2$, which is a common setting in modern deep learning tasks.

4.2 UTILITY

In this section, we prove the convergence guarantee of DP-FedEXP for general non-convex objectives. Here, we require the following standard assumptions:

Assumption 4.3 (Smoothness and Lipschitz continuity)**.** Each client loss function f_i is L-smooth and G-Lipschitz continuous, where $L, G > 0$ are constants. That is, for any $w, w' \in \mathbb{R}^d$, we have $\|\nabla f_i(w) - \nabla f_i(w')\| \leq L\|w - w'\|$ and $\|\nabla f_i(w)\| \leq G$.

Assumption 4.4 (Bounded gradient diversity)**.** For any $w \in \mathbb{R}^d$, the diversity of the gradients is bounded as $\frac{1}{M}\sum_{i=1}^M \|\nabla f_i(w) - \nabla F(w)\|^2 \leq {\sigma_g}^2$, where σ_g^2 is a constant.

Under the above assumptions, we have the following results.

Theorem 4.5 (LDP case)**.** *Assume that Assumptions 4.3 and 4.4 hold. Let $F^* = \min_w F(w)$ and $C = \eta_l\tau G$. Then, for any $\eta_l = \Theta(1/(L\tau)) < 1/(24L\tau)$ and the sequence $\{w^{(t)}\}_{t=1}^T$ generated by LDP-FedEXP satisfies*

$$\min_{t\in[T]} \left\| \nabla F(w^{(t)}) \right\|^2 \leq O\Bigg(\underbrace{\frac{F(w^0) - F^*}{\sum_{t=1}^T \eta_g^{(t)}\eta_l\tau}}_{T_1} + \underbrace{\eta_l^2 L^2\tau^2\sigma_g^2}_{T_2} + \underbrace{\eta_l L\tau\sigma_g^2}_{T_3} + \underbrace{\frac{L\sigma^2q^2}{\eta_l\tau}\left[\frac{d}{M} + \sqrt{\frac{d}{M}}\right]}_{T_4^{gauss} := \textit{privacy error}} \Bigg)$$

with probability at least $1 - Te^{-c\cdot q^2}$ for any $q \in [1, \sqrt{M}]$, where c is a numerical constant.

Theorem 4.6 (CDP case)**.** *Assume that Assumptions 4.3 and 4.4 hold. Let $F^* = \min_w F(w), \sigma_\xi = d\sigma^2/M$, and $C = \eta_l\tau G$. Then, for any $\eta_l = \Theta(1/(L\tau)) < 1/(24L\tau)$, the sequence $\{w^{(t)}\}_{t=1}^T$ generated by CDP-FedEXP satisfies*

$$\min_{t\in[T]} \left\| \nabla F(w^{(t)}) \right\|^2 \leq T_1 + T_2 + T_3 + \underbrace{O\left(\frac{L\sigma^2q^2}{\eta_l\tau} \cdot \frac{d}{M}\right)}_{T_4^{cdp} := \textit{privacy error}}$$

with probability at least $1 - Te^{-c\cdot q^2}$ for $q \in [1, \sqrt{M}]$, where c is a numerical constant

See Appendix F for the proof. The difficulty of the proof lies in the correlation between the global step size $\eta_g^{(t)}$ and the noisy update $\bar{c}^{(t)}$ as discussed in previous works (Jhunjhunwala et al., 2023; Li et al., 2024). Since the step size $\eta_g^{(t)}$ depends on noisy update $\bar{c}^{(t)}$ in a complicated way, we need to carefully evaluate the error terms from DP noise.

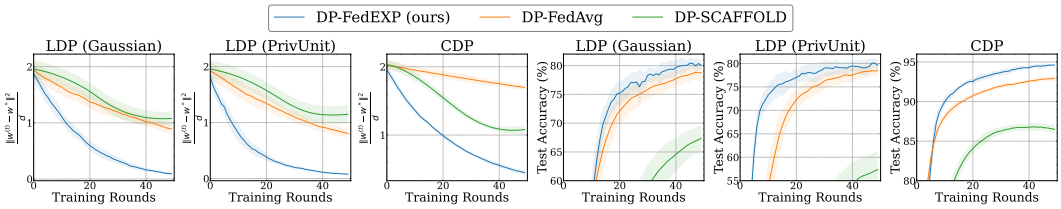

Figure 2: The distance to the optimal solution for the synthetic dataset (left) and test accuracy for the MNIST dataset (right). DP-FedEXP consistently outperforms baseline algorithms.

Comparison with FedEXP The above theorems imply that the error of DP-FedEXP are decomposed into four terms: initialization error T_1, client drift error T_2, global variance T_3, and privacy error T_4. As shown in Theorem 2 from Jhunjhunwala et al. (2023), the error of FedEXP is given by $T_1 + T_2 + T_3$. Thus, the DP noise only affects the privacy error term T_4, which vanishes as the number of clients M goes to infinity.

Comparison with DP-FedAvg The error of DP-FedAvg is given by $O\left(\frac{F(w^{(0)})-F^*}{T\eta_l\tau}\right) + O(\eta_l^2 L^2\tau^2\sigma_g^2) + O(\frac{L\sigma^2}{\eta_l\tau}\cdot\frac{d}{M})$ for both LDP and CDP cases (Zhang et al., 2022). The initialization error term $O\left(\frac{F(w^{(0)})-F^*}{T\eta_l\tau}\right)$ is always larger than that of LDP-FedEXP and CDP-FedEXP since $\eta_g^{(t)} \geq 1$ for any t. Thus, DP-FedEXP provably accelerate the convergence of DP-FedAvg in both LDP and CDP setting. For the privacy error term T_4, LDP-FedEXP with the Gaussian mechanism has the additional term of order $\sqrt{d/M}$. This can be reduced to $1/\sqrt{M}$ by adopting PrivUnit as a local randomizer as shown in Section D. In contrast, for the CDP case, CDP-FedEXP achieves the same privacy error as DP-FedAvg by setting $\sigma_\xi = d\sigma^2/M$.

5 Numerical Experiments

In this section, we evaluate the performance of DP-FedEXP on synthetic and real datasets. For the synthetic experiment, we consider a linear regression problem, where clients share the common minimizer following Jhunjhunwala et al. (2023). For the realistic experiment, we consider the image classification task on the MNIST dataset (LeCun, 1998). We compare our proposed method with the baseline algorithms such as DP-FedAvg and DP-SCAFFOLD. For fair comparison, we have tuned the clipping threshold C and the local learning rate η_l for each method via grid search. See Appendix G for the detailed setup and additional results.

DP-FedEXP consistently outperforms baselines Fig. 2 illustrates the distance to the optimum w^* for the synthetic experiment and the test accuracy for the MNIST experiment. In both experiments, we can see that DP-FedEXP effectively accelerates DP-FedAvg. In addition, as shown in Table 1, our proposed methods achieve the same privacy guarantee as DP-FedAvg in the LDP setting and the additional privacy budget in the CDP setting is negligible. Furthermore, DP-FedEXP consistently outperforms DP-SCAFFOLD. In our setup, DP-SCAFFOLD does not improve the performance compared to DP-FedAvg except for the case of CDP in the synthetic experiment. One possible reason is that DP-SCAFFOLD in Noble et al. (2022) is designed for sample-level DP and the noise scale for client-level DP is much larger than that for sample-level DP.

Table 1: Comparison of the privacy budget ε ($\delta = 10^{-5}$) for DP-FedEXP and DP-FedAvg.

Problem setting	DP-FedEXP	DP-FedAvg
LDP (Gaussian)	15.659	15.659
LDP (PrivUnit)	6	6
CDP (Synthetic)	15.647	15.258
CDP (MNIST)	15.261	15.258

6 Conclusion

In this study, we have pursued a practical federated learning framework with formal privacy guarantee. To this end, we have proposed DP-FedEXP for both LDP and CDP settings, which adaptively selects the global step size in DP-FL with respect to the heterogeneity of the local updates. Our

proposed framework does not require any additional hyperparameters, additional communication cost or additional computational cost at clients. Then, we have proved differential privacy guarantee and provided the convergence analysis of our proposed methods. We have shown that DP-FedEXP provably accelerates DP-FedAvg while maintaining the privacy guarantee. Finally, we have shown that our proposed methods outperform existing DP-FL algorithms in the numerical experiments.

REFERENCES

Galen Andrew, Om Thakkar, Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17455–17466, 2021.

Anonymous. Towards hyperparameter-free optimization with differential privacy. In *Submitted to The Thirteenth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=2kGKsyhtvh`. under review.

Hilal Asi, Vitaly Feldman, and Kunal Talwar. Optimal algorithms for mean estimation under local differential privacy. In *International Conference on Machine Learning*, pp. 1046–1056, 2022.

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.

Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Automatic clipping: Differentially private deep learning made easier and stronger. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 41727–41764, 2023.

Ameya Daigavane, Gagan Madan, Aditya Sinha, Abhradeep Guha Thakurta, Gaurav Aggarwal, and Prateek Jain. Node-level differentially private graph neural networks. In *ICLR 2022 Workshop on PAIR^2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*, 2022.

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, 2006.

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

Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. In *Advances in Neural Information Processing Systems*, volume 34, pp. 11631–11642, 2021.

David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

Farzin Haddadpour and Mehrdad Mahdavi. On the convergence of local descent methods in federated learning. *arXiv preprint arXiv:1910.14425*, 2019.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Divyansh Jhunjhunwala, Shiqiang Wang, and Gauri Joshi. Fedexp: Speeding up federated averaging via extrapolation. In *International Conference on Learning Representations*, 2023.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes. In *International Conference on Machine Learning*, pp. 3252–3261, 2019.

Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv preprint arXiv:2008.03606*, 2020a.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143, 2020b.

Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015.

Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency, 2017. URL `https://arxiv.org/abs/1610.05492`.

Yann LeCun. The mnist database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*, 1998.

Hanmin Li, Kirill Acharya, and Peter Richtárik. The power of extrapolation in federated learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine learning and systems*, volume 2, pp. 429–450, 2020.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*, 2021.

Jingcheng Liu and Kunal Talwar. Private selection from private candidates. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pp. 298–309, 2019.

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017.

Ilya Mironov. Rényi differential privacy. In *IEEE 30th Computer Security Foundations symposium*, pp. 263–275, 2017.

Aritra Mitra, Rayana Jaafar, George J Pappas, and Hamed Hassani. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14606–14619, 2021.

Shubhankar Mohapatra, Sajin Sasy, Xi He, Gautam Kamath, and Om Thakkar. The role of adaptive optimizers for honest private hyperparameter selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 7806–7813, 2022.

Maxence Noble, Aurélien Bellet, and Aymeric Dieuleveut. Differentially private federated learning on heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pp. 10110–10145, 2022.

Nicolas Papernot and Thomas Steinke. Hyperparameter tuning with renyi differential privacy. *arXiv preprint arXiv:2110.03620*, 2021.

Guy Pierra. Decomposition through formalization in a product space. *Mathematical Programming*, 28(1):96–115, 1984.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

Yifan Shi, Yingqi Liu, Kang Wei, Li Shen, Xueqian Wang, and Dacheng Tao. Make landscape flatter in differentially private federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24552–24562, 2023.

Qiaoyue Tang, Frederick Shpilevskiy, and Mathias Lécuyer. DP-AdamBC: Your DP-Adam Is Actually DP-SGD (Unless You Apply Bias Correction). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15276–15283, 2024.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Hua Wang, Sheng Gao, Huanyu Zhang, Weijie J Su, and Milan Shen. Dp-hypo: an adaptive private hyperparameter optimization framework. *arXiv preprint arXiv:2306.05734*, 2023.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in neural information processing systems*, volume 33, pp. 7611–7623, 2020.

Xinwei Zhang, Xiangyi Chen, Mingyi Hong, Zhiwei Steven Wu, and Jinfeng Yi. Understanding clipping for federated learning: Convergence and client-level differential privacy. In *International Conference on Machine Learning*, 2022.

A Other Related Work

Adapive Optimization Algorithms with DP Inspired by the success of adaptive optimization algorithms such as Adam (Kingma & Ba, 2015) in the non-private setting, their DP variants have been utilized in various fields (Li et al., 2021; Daigavane et al., 2022). However, despite their success in the non-private setting, their DP variants often suffer from the slow convergence. Tang et al. (2024) have found that the bias from DP noise degrades the performance of DP-Adam and proposed DP-AdamBC, which removes the bias in the second moment estimation of Adam update. This implies that it is not straightforward to extend adaptive methods in the non-private setting to the DP setting. Note that the above attempts are mainly focused on the centralized setting and it is still unclear how to incorporate the adaptivity to the heterogeneity of the client data into DP-FL algorithms.

Hyperparameter Tuning with DP In the most of the existing works, the privacy leakage from hyperparameter tuning is ignored. However, as discussed in Papernot & Steinke (2021), hyperparameters can raise the privacy risks of memorizing the training data. Thus, to ensure the formal privacy guarantee, we should audit the privacy leakage from the entire training process including hyperparameter tuning. Several works (Liu & Talwar, 2019; Wang et al., 2023; Papernot & Steinke, 2021; Mohapatra et al., 2022) have proposed to privatize hyperparameter tuning by consuming additional privacy budget. However, these methods often result in much weaker privacy guarantees unless larger DP noise is used. For example, Papernot & Steinke (2021) have reported that the privacy parameter can be doubled or even tripled by accounting the privacy leakage from hyperparameter tuning. Furthermore, it is prohibitively expensive or even infeasible to conduct hyperparameter tuning with distributed data in the FL setting.

Hyperparameter-Free DP Optimization A line of work has investigated adaptive methods to select hyperparameters for DP optimization algorithms (Andrew et al., 2021; Bu et al., 2023; Anonymous, 2024). For example, Adaptive clipping (Andrew et al., 2021) selects clipping threshold in DP-FL by estimating a quantile of the update norm with a negligible amount of privacy budget. Furthermore, Anonymous (2024) have proposed a hyperparameter-free algorithm for DP optimization in the centralized setting. However, to the best of our knowledge, there is no work that provides hyperparameter-free step size rule to deal with the heterogeneity of the client data for DP-FL.

B Detailed Procedure of the Proposed Algorithms

Algorithm 1 LDP-FedEXP

Input: initial $w^{(0)}$, clipping threshold C, number of rounds T
Output: final $w^{(T)}$
for $t = 1$ **to** T **do**
  Server sends $w^{(t-1)}$ to all clients
  for client $i = 1$ **to** M **do**
    $\tilde{\Delta}_i^{(t)} \leftarrow \text{localupdate}(w^{(t-1)}, \mathcal{D}_i)$
    $\Delta_i^{(t)} \leftarrow \min\{C/\|\tilde{\Delta}_i^{(t)}\|, 1\} \cdot \tilde{\Delta}_i^{(t)}$
    $c_i^{(t)} \leftarrow \text{LocalRandomizer}(\Delta_i^{(t)})$
    Client i sends $c_i^{(t)}$ to server
  end for
  Aggregate local updates:
  $\bar{c}^{(t)} \leftarrow \frac{1}{M}\sum_{i=1}^{M} c_i^{(t)}$
  Compute global step size $\eta_g^{(t)}$ as in equation 6 or equation 8.
  Update global model with $w^{(t)} \leftarrow w^{(t-1)} + \eta_g^{(t)}\bar{c}^{(t)}$
end for

Algorithm 2 CDP-FedEXP

Input: initial $w^{(0)}$, clipping threshold C, noise scale σ, number of rounds T
Output: final $w^{(T)}$
for $t = 1$ **to** T **do**
  Server sends $w^{(t-1)}$ to all clients
  for user $i = 1$ **to** M **do**
    $\tilde{\Delta}_i^{(t)} \leftarrow \text{localupdate}(w^{(t-1)}, \mathcal{D}_i)$
    $\Delta_i^{(t)} \leftarrow \min\{C/\|\tilde{\Delta}_i^{(t)}\|, 1\} \cdot \tilde{\Delta}_i^{(t)}$
    Client i sends $\Delta_i^{(t)}$ to server
  end for
  Aggregate local updates and add noise:
  $\bar{c}^{(t)} \leftarrow \frac{1}{M}\sum_{i=1}^{M} \Delta_i^{(t)} + \varepsilon^{(t)} \quad (\varepsilon^{(t)} \sim \mathcal{N}(0, \sigma^2/M))$
  Compute global step size $\eta_g^{(t)}$ as in equation 7.
  Update global model with $w^{(t)} \leftarrow w^{(t-1)} + \eta_g^{(t)}\bar{c}^{(t)}$
end for

C Auxiliary Results

Lemma C.1 (Gaussian tail bound)**.** *Let X be a random variable following the Gaussian distribution $\mathcal{N}(0, \sigma^2)$. Then, for any $q > 0$, we have*

$$X \leq \sigma q \quad \textit{with probability at least } 1 - e^{-q^2/2}.$$

Proof. From the Hoeffding bound Wainwright (2019), we obtain

$$\text{Prob}\left(X > t\right) \leq e^{-t^2/2\sigma^2}.$$

Setting $t = \sigma q$ completes the proof. □

Lemma C.2 (Tail bound for norm of Gaussian)**.** *Let $x_i \in \mathbb{R}^d$ be a random variable following the Gaussian distribution $\mathcal{N}(0, \sigma^2 I_d)$. Then, for any $q \geq 1$, we have*

$$\frac{1}{n}\sum_{i=1}^{n}\|x_i\|^2 - d\sigma^2 \leq \sqrt{\frac{d}{n}}\sigma^2 \cdot q^2 \quad \textit{with probability at least } 1 - e^{-q^2/8}.$$

Proof. It is sufficient to consider the case of $\sigma^2 = 1$ by scaling x_i with $1/\sigma$. Since $Z_i := \|x_i\|^2$ follows the χ^2-distribution with d degrees of freedom, we have

$$\begin{aligned}\mathbb{E}\left[e^{\lambda(Z_i-d)}\right] &= e^{-d\lambda}\cdot\left[\int e^{\lambda X^2}\frac{1}{\sqrt{2\pi}}e^{-X^2/2}\mathrm{d}X\right]^d \\ &= e^{-d\lambda}\cdot\left[\frac{1}{\sqrt{1-2\lambda}}\right]^d \\ &\le e^{-2d\lambda^2} \quad \text{for any } |\lambda|\le 1/4.\end{aligned}$$

Thus, $\sum_{i=1}^n Z_i$ is subexponential random variable with parameters $(\nu^2, b) = (4dn, 4)$ and satisfies

$$\mathrm{Prob}\left(\sum_{i=1}^n Z_i - dn \ge t\right) \le \begin{cases}\exp\left(-\frac{t^2}{8dn}\right) & \text{for } t\in(0,dn),\\ \exp\left(-\frac{t}{8}\right) & \text{otherwise.}\end{cases}$$

Setting $t = q^2\cdot\sqrt{dn}$, we obtain

$$\begin{aligned}\mathrm{Prob}\left(\frac{1}{n}\sum_{i=1}^n Z_i - d \ge \sqrt{\frac{d}{n}}\cdot q^2\right) &\le \begin{cases}\exp\left(-q^4/8\right) & \text{for } t\in(0,\sqrt{dn}),\\ \exp\left(-\frac{q^2}{8}\right) & \text{otherwise.}\end{cases} \\ &\le \exp\left(-\frac{q^2}{8}\right) \quad \text{for any } q\ge 1.\end{aligned}$$

This completes the proof. □

Lemma C.3 (Vector Bernstein Inequality)**.** *Let $x_1,\dots,x_n\in\mathbb{R}^d$ be independent zero-mean random variables. Assume that $\|x_i\|\le R$ almost surely for any i. Then, for any $q\in[0,\sqrt{n}]$, we have*

$$\mathrm{Prob}\left(\left\|\frac{1}{n}\sum_{i=1}^n x_i\right\| \ge \frac{R(1+q)}{\sqrt{n}}\right) \le \exp\left(-\frac{q^2}{4}\right).$$

Proof. Let $V = \sum_{i=1}^n \mathbb{E}\left[\|x_i\|^2\right]$. Note that $V\le nR^2$ since $\|x_i\|\le R$ almost surely. Then, Theorem 12 in Gross (2011) implies

$$\mathrm{Prob}\left(\left\|\sum_{i=1}^n x_i\right\| \ge \sqrt{n}R + t\right) \le \mathrm{Prob}\left(\left\|\sum_{i=1}^n x_i\right\| \ge \sqrt{V} + t\right) \le \exp\left(-\frac{t^2}{4V}\right)$$

for any $t\in[0,V/R]$. Setting $t=\sqrt{n}Rq$, we obtained

$$\begin{aligned}\mathrm{Prob}\left(\left\|\frac{1}{n}\sum_{i=1}^n x_i\right\| \ge \frac{R(1+q)}{\sqrt{n}}\right) = \mathrm{Prob}\left(\left\|\sum_{i=1}^n x_i\right\| \ge \sqrt{n}R(1+q)\right) &\le \exp\left(-\frac{nR^2q^2}{4V}\right) \\ &\le \exp\left(-\frac{q^2}{4}\right)\end{aligned}$$

for any $q\in[0,\sqrt{n}]$. □

Lemma C.4. *Assume that the generalized approximate projection condition Eq. equation 4 holds. Then, for any $\eta_g > 0$, we have*

$$\begin{aligned}\left\|w^{(t+1)} - w^*\right\|^2 = (1-\alpha\eta_g)\left\|w^{(t)} - w^*\right\|^2 - \eta_g\frac{1}{M}\sum_{i=1}^M\left\|\Delta_i^{(t)}\right\|^2 \\ + \eta_g^2\left\|\bar{c}^{(t)}\right\|^2 + O\left(\frac{\eta_g\cdot\sqrt{\frac{d}{M}\sigma^2}\cdot\left\|w^{(t)}-w^*\right\|}{\sqrt{d}}\cdot q\right),\end{aligned}$$

with probability at least $1-e^{-q^2/2}$ for any $q>0$.

Proof. From the generalized approximate projection condition Eq. equation 4, we have

$$\frac{1}{M}\sum_{i=1}^{n}\left\|w^{(t)}+\Delta_i^{(t)}-w^*\right\|^2=\left\|w^{(t)}-w^*\right\|^2+\frac{2}{M}\sum_{i=1}^{M}\langle w^{(t)}-w^*,\Delta_i^{(t)}\rangle+\frac{1}{M}\sum_{i=1}^{n}\left\|\Delta_i^{(t)}\right\|^2$$
$$=(1-\alpha)\left\|w^{(t)}-w^*\right\|^2.$$

This implies

$$\frac{2}{M}\sum_{i=1}^{M}\langle w^{(t)}-w^*,\Delta_i^{(t)}\rangle=-\alpha\left\|w^{(t)}-w^*\right\|^2-\frac{1}{M}\sum_{i=1}^{M}\left\|\Delta_i^{(t)}\right\|^2.$$

Substituting the above equation, we obtain

$$\begin{aligned}\left\|w^{(t)}+\eta_g c^{(t)}-w^*\right\|^2&=\left\|w^{(t)}-w^*\right\|^2+\frac{2\eta_g}{M}\sum_{i=1}^{M}\langle\Delta_i^{(t)},w^{(t)}-w^*\rangle\\&\quad+2\eta_g\langle\bar{\varepsilon}^{(t)},w^{(t)}-w^*\rangle+\eta_g^2\left\|\bar{c}^{(t)}\right\|^2\\&=(1-\alpha\eta_g)\left\|w^{(t)}-w^*\right\|^2-\frac{\eta_g}{M}\sum_{i=1}^{M}\left\|\Delta_i^{(t)}\right\|^2\\&\quad+\eta_g^2\left\|c^{(t)}\right\|^2+O\left(\frac{\eta_g\sigma\left\|w^{(t)}-w^*\right\|}{\sqrt{M}}\cdot q\right),\end{aligned}$$

with probability at least $1-e^{-q^2/2}$ for any $q>0$. Here, we used the fact that $2\eta_g\langle\bar{\varepsilon}^{(t)},w^{(t)}-w^*\rangle$ follows $\mathcal{N}(0,\eta_g^2\sigma^2\left\|w^{(t)}-w^*\right\|^2/M)$ and Lemma C.1. This completes the proof. □

D Extention to PrivUnit

D.1 Brief review of PrivUnit

Here, we briefly explain PrivUnit and ScalarDP algorithms proposed by Bhowmick et al. (2018). In this paper, we follow the procedure in Bhowmick et al. (2018) and privatize the norm and the direction of the local update separately. That is, we randomize the local update $\Delta_i^{(t)}$ as $c_i^{(t)}=\hat{r}_i^{(t)}\cdot z_i^{(t)}$, where $z_i^{(t)}:=\mathrm{PrivUnit}\Big(\Delta_i^{(t)}/\|\Delta_i^{(t)}\|;\varepsilon_0,\varepsilon_1\Big)$, $\hat{r}_i^{(t)}:=\mathrm{ScalarDP}\Big(\|\Delta_i^{(t)}\|;\varepsilon_2\Big)$, and $\varepsilon_0,\varepsilon_1,\varepsilon_2$ are privacy parameters. Here, $\mathrm{PrivUnit}$ privatizes the direction and $\mathrm{ScalarDP}$ privatizes the norm. See Algorithm 3 and 4 for the detailed procedure. As shown in Lemma D.1, $c_i^{(t)}$ is an unbiased estimator of $\Delta_i^{(t)}$ and its variance is bounded by $O(dC^2\cdot(\frac{1}{\varepsilon_1}\vee\frac{1}{(e^{\varepsilon_1}-1)^2}))$ if $\varepsilon_1\in(0,d)$ and $\varepsilon_2=\Omega(1)$. We define $\sigma^2:=C^2\cdot(\frac{1}{\varepsilon_1}\vee\frac{1}{(e^{\varepsilon_1}-1)^2})$ for the PrivUnit case to ensure the consistency in the notation with the Gaussian mechanism case, where the variance of $c_i^{(t)}$ is given by $d\sigma^2$. We provide the detailed description of the algorithms in Algorithm 3 and 4.

Lemma D.1. *For $\varepsilon_0,\varepsilon_1,\varepsilon_2\in[0,d]$, $c=\mathrm{PrivUnit}(\Delta/\|\Delta\|;\varepsilon_0,\varepsilon_1)\cdot\mathrm{ScalarDP}(\|\Delta\|;\varepsilon_2)$ is an unbiased estimator of Δ if $\|\Delta\|\le C$. That is, $E[c]=\Delta$. Moreover, c satisfies $(\varepsilon_0+\varepsilon_1+\varepsilon_2)$-DP.*

Proof. See Proposition 3 and Lemma 4.1 in Bhowmick et al. (2018) for the proof. □

D.2 Step Size Formula for PrivUnit

Here, we provide the step size rule for PrivUnit. Let $\hat{r}_i^{(t)}=\mathrm{ScalarDP}(\Delta_i^{(t)};\varepsilon_2)$ and $z_i^{(t)}=\mathrm{PrivUnit}(\Delta_i^{(t)}/\|\Delta_i^{(t)}\|;\varepsilon_0,\varepsilon_1)$. Note that $c_i^{(t)}=\hat{r}_i^{(t)}\cdot z_i^{(t)}$. Since $\|z_i\|=1/m$, where $m>0$ is a constant, we can calculate $|\hat{r}_i^{(t)}|$ as $m\cdot\|c_i^{(t)}\|$. Furthermore, since $\hat{r}_i^{(t)}$ takes discrete values, we can reconstruct $\hat{r}_i^{(t)}$ from $|\hat{r}_i^{(t)}|$ except for special choices of privacy parameter ε_2. However, as

Algorithm 3 PrivUnit

Input: $u \in \mathbb{S}^{d-1}, \varepsilon_0, \varepsilon_1 > 0$
Output: Randomized vector $Z \in \mathbb{R}^d$
$p \leftarrow \frac{e^{\varepsilon_0}}{1+e^{\varepsilon_0}}$
Select γ such that

$$\gamma \leq \frac{e^{\varepsilon_1}-1}{e^{\varepsilon_1}+1}\sqrt{\frac{\pi}{2(d-1)}},$$

or

$$\varepsilon_1 \geq \frac{1}{2}\log d + \log 6 - \frac{d-1}{2}\log(1-\gamma^2) + \log\gamma \text{ and } \gamma \geq \sqrt{\frac{2}{d}}$$

Draw random vector V according to the following distribution:

$$V \leftarrow \begin{cases} \text{uniform on } \{v \in \mathbb{S}^{d-1} \mid \langle v, u\rangle \geq \gamma\} & \text{w.p. } \gamma, \\ \text{uniform on } \{v \in \mathbb{S}^{d-1} \mid \langle v, u\rangle < \gamma\} & \text{otherwise.} \end{cases}$$

$\alpha \leftarrow \frac{d-1}{2}, \tau = \frac{1+\gamma}{2}$, and

$$m \leftarrow \frac{(1-\gamma^2)^\alpha}{2^{d-2}(d-1)}\left[\frac{p}{B(\alpha,\alpha) - B(\tau;\alpha,\alpha)} - \frac{1-p}{B(\tau;\alpha,\alpha)}\right]$$

Rescale V as $Z \leftarrow \frac{1}{m}\cdot V$

Algorithm 4 ScalarDP

Input: magnitude $r \in [0, C]$, privacy parameter $\varepsilon_2 > 0$
Output: Randomized magnitude $\hat{r}$
$k \leftarrow e^{\lceil \varepsilon_2/3 \rceil}$
$r_{\max} \leftarrow C$
Sample $J \in \{0, \ldots, k\}$ according to the following distribution:

$$J \leftarrow \begin{cases} \lfloor kr/r_{\max} \rfloor & \text{w.p. } \lceil kr/r_{\max} \rceil - kr/r_{\max}, \\ \lceil kr/r_{\max} \rceil & \text{otherwise.} \end{cases}$$

Draw randomized response $\hat{J}$ according to the following distribution:

$$\hat{J} \leftarrow \begin{cases} J & \text{w.p. } \frac{e^{\varepsilon_2}}{e^{\varepsilon_2}+k}, \\ \text{uniform on } \{0, \ldots, k\}\backslash\{J\} & \text{otherwise.} \end{cases}$$

Debias $\hat{r}$ as $\hat{r} \leftarrow a(\hat{J} - b)$, where $a = \left(\frac{e^{\varepsilon_2}+k}{e^{\varepsilon_2}-1}\right)\frac{r_{\max}}{k}$ and $b = \frac{k(k+1)}{2(e^{\varepsilon_2}+k)}$

shown in Bhowmick et al. (2018), the variance of the noisy update is not constant and depends on the norm of the original update in a complicated way. Thus, it is not straightforward to develop an unbiased estimator of $\|\Delta_i^{(t)}\|^2$. To deal with this issue, we utilize the following upper bound of the variance of PrivUnit:

$$\mathbb{E}\left[\left(\hat{r}_i^{(t)} - r_i^{(t)}\right)^2\right] \leq c_1\left(r_i^{(t)}\right)^2 + c_2 r_i^{(t)} + c_3,$$

where $r_i^{(t)} = \|\Delta_i^{(t)}\|$, and c_1, c_2, c_3 are constants defined in Algorithm 5. Based on the above upper bound, we propose the following formula for the step size:

$$\eta_g^{(t)} = \max\left\{1, \frac{\frac{1}{M}\sum_{i=1}^M \hat{s}_i}{\left\|\bar{c}^{(t)}\right\|^2}\right\}, \tag{8}$$

where $\hat{s}_i = \frac{(\hat{r}_i^{(t)})^2 - c_2\hat{r}_i^{(t)} - c_3}{1+c_1}$. See Algorithm 5 for the detailed procedure. Here, $\frac{1}{M}\sum_{i=1}^M \hat{s}_i$ is not

Algorithm 5 Norm Estimation for PrivUnit

Input: Noisy update $c := \mathrm{PrivUnit}(\Delta/\|\Delta\|; \varepsilon_0, \varepsilon_1) \cdot \mathrm{ScalarDP}(\|\Delta\|; \varepsilon_2)$
Output: Estimated value $\hat{s}$ of $\|\Delta\|^2$
Set $a, b, k > 0$ as in Algorithm 4 and m as in Algorithm 3
$\tilde{r} \leftarrow m \cdot \|c\|$, $\tilde{J} \leftarrow \tilde{r}/a + b$.
if $\tilde{J} \in \mathbb{Z}$ **then** $\hat{r} \leftarrow \tilde{r}$ **else** $\hat{r} \leftarrow -\tilde{r}$
$\hat{s} \leftarrow \frac{1}{1+c_1}(\hat{r}^2 - c_2\hat{r} - c_3)$,
where $c_1 = \frac{k+1}{e^{\varepsilon_2}-1}, c_2 = -c_1 C, c_3 = (c_1+1)\frac{C^2}{4k^2} + c_1 C^2\left[\frac{(2k+1)(e^{\varepsilon_2}+k)}{6k(e^{\varepsilon_2}-1)} - \frac{k+1}{4(e^{\varepsilon_2}-1)}\right]$.

an unbiased estimator of $\frac{1}{M}\sum_{i=1}^M \|\Delta_i^{(t)}\|^2$ but it satisfies

$$\mathbb{E}\left[\frac{1}{M}\sum_{i=1}^M \hat{s}_i\right] \le \frac{1}{M}\sum_{i=1}^M \left\|\Delta_i^{(t)}\right\|^2.$$

This property is sufficient to prove the convergence guarantee in Theorem D.2. In addition, as shown in Fig. 1, the step size formula in equation 8 accurately estimates $\eta_{\mathrm{target}}^{(t)}$.

Theorem D.2. *Assume that Assumptions 4.3 and 4.4 hold. Let $F^* = \min_w F(w)$ and $C = \eta_l \tau G$. Then, for any $\eta_l = \Theta(1/(L\tau)) < 1/(24L\tau)$ and the sequence $\{w^{(t)}\}_{t=1}^T$ generated by LDP-FedEXP with PrivUnit for $\varepsilon_1, \varepsilon_2 = \Theta(1)$ satisfies*

$$\min_{t\in[T]}\left\|\nabla F(w^{(t)})\right\|^2 \le T_1 + T_2 + T_3 + \underbrace{O\left(\frac{L\sigma^2 q^2}{\eta_l \tau}\left[\frac{d}{M} + \sqrt{\frac{1}{M}}\right]\right)}_{T_4^{privunit} := \textit{privacy error}}$$

with probability at least $1 - Te^{-c\cdot q^2}$ for any $q \in [1, \sqrt{M}]$, where c is a numerical constant.

See Appendix F for the proof.

In the following, we prove some properties of PrivUnit and norm estimation procedure in Algorithm 5 for the convergence analysis.

Lemma D.3. *Assume that $\frac{k(k+1)}{e^{\varepsilon_2}+k} \notin \mathbb{Z}$. Then, the estimated value $\hat{s}$ computed by Algorithm 5 satisfies $E[\hat{s}] \le r^2$.*

Proof. First, we show that $\hat{r} = \mathrm{ScalarDP}(\|\Delta\|)$. From the definition of $c = \mathrm{PrivUnit}(\Delta/\|\Delta\|) \cdot \mathrm{ScalarDP}(\|\Delta\|)$ and $\|\mathrm{PrivUnit}(\Delta/\|\Delta\|)\| = 1/m$, we have $\tilde{r} = |\mathrm{ScalarDP}(\|\Delta\|)|$. If $\mathrm{ScalarDP}(\|\Delta\|) < 0$ and $\tilde{J} \in \mathbb{Z}$, $\mathrm{ScalarDP}(\|\Delta\|) = -\tilde{r}$ and $\hat{J} = \mathrm{ScalarDP}(\|\Delta\|)/a + b = -\tilde{r}/a + b \in \mathbb{Z}$. This implies $\hat{J} + \tilde{J} = 2b = \frac{k(k+1)}{e^{\varepsilon}+k} \in \mathbb{Z}$, which contradicts the assumption. Thus, $\tilde{J} \notin \mathbb{Z}$ and $\hat{r} = -\tilde{r} = \mathrm{ScalarDP}(\|\Delta\|)$ if $\mathrm{ScalarDP}(\|\Delta\|) < 0$. On the other hand, if $\mathrm{ScalarDP}(\|\Delta\|) \ge 0$, $\tilde{J} = \tilde{r}/a + b = \mathrm{ScalarDP}(\|\Delta\|)/a + b \in \mathbb{Z}$ and $\hat{r} = \tilde{r} = \mathrm{ScalarDP}(\|\Delta\|)$. Combining the above arguments, we have $\hat{r} = \mathrm{ScalarDP}(\|\Delta\|)$.

Next, we show that $E[\hat{s}] \le r^2$. As shown in Bhowmick et al. (2018), the variance of $\hat{r}$ is bounded as follows:

$$\begin{aligned}\mathrm{Var}\hat{r} &\le \frac{k+1}{e^{\varepsilon_2}-1}\left[r^2 + \frac{r_{\max}^2}{4k^2} - r r_{\max} + \frac{(2k+1)(e^{\varepsilon_2}+k)r_{\max}^2}{6k(e^{\varepsilon_2}-1)} - \frac{(k+1)r_{\max}^2}{4(e^{\varepsilon_2}-1)}\right] + \frac{r_{\max}^2}{4k^2} \\ &= c_1 r^2 + c_2 r + c_3.\end{aligned}$$

Thus, we have

$$\begin{aligned}\mathbb{E}\left[\hat{s}\right] &= \mathbb{E}\left[\frac{1}{1+c_1}(\hat{r}^2 - c_2\hat{r} - c_3)\right] \\ &= \frac{1}{1+c_2}\left(r^2 + \mathrm{Var}\hat{r} - c_2 r - c_3\right) \\ &\leq \frac{1}{1+c_2}\left(r^2 + c_1 r^2 + c_2 r + c_3 - c_2 r - c_3\right) \\ &= r^2.\end{aligned}$$

This completes the proof. □

Lemma D.4 (Properties of PrivUnit and ScalarDP)**.** *Assume that* $\varepsilon_1 \in [0, d]$. *Then,* $z = \mathrm{PrivUnit}(u/\|u\|)$ *and* $\hat{r} = \mathrm{ScalarDP}(\|u\|)$ *satisfy*

$$\|z\|^2 = O\left(\frac{d}{\varepsilon_1} \vee \frac{d}{(e^{\varepsilon_1}-1)^2}\right),$$
$$|\hat{r}| = O\left(\frac{e^{\varepsilon_2}}{e^{\varepsilon_2}-1} \cdot C\right),$$

with probability 1.

Proof. The first inequality follows from Proposition 4 in Bhowmick et al. (2018).

From the definition of $\hat{r}$, we have $|\hat{r}| \leq a\left|\hat{J} - b\right| \leq a(k+b)$. Substituting, $k = \lceil e^{\varepsilon_2/3} \rceil$, $a = \frac{e^{\varepsilon_2}+k}{e^{\varepsilon_2}-1}\frac{C}{k}$ and $b = \frac{k(k+1)}{2(e^{\varepsilon_2}+k)}$, we obtain the second inequality. □

Lemma D.5 (Tail bounds for PrivUnit)**.** *Let* $z_i = \mathrm{PrivUnit}(u_i/\|u_i\|)$ *and* $\hat{r}_i = \mathrm{ScalarDP}(\|u_i\|)$ *for* $u_i \in \mathbb{R}^d$ $(\|u_i\| \leq C)$ *with* $\varepsilon_1, \varepsilon_2 = O(1)$. *Then, for any* $v_i \in \mathbb{R}^d$, *we have*

$$\frac{1}{M}\sum_{i=1}^{M}\langle \hat{r}_i \cdot z_i - u_i, v_i\rangle = O\left(\sqrt{\frac{C^2 d \sum_{i=1}^{M}\|v_i\|^2}{M^2}} \cdot q\right),$$
$$\left\|\frac{1}{M}\sum_{i=1}^{M}(\hat{r}_i \cdot z_i - u_i)\right\|^2 = O\left(\frac{dC^2(1+q^2)}{M}\right),$$
$$\frac{1}{M}\sum_{i=1}^{M}\hat{s}_i - \frac{1}{M}\sum_{i=1}^{M}\|\Delta_i\|^2 = O\left(C^2\sqrt{\frac{1}{M}} \cdot q\right),$$

with probability at least $1 - e^{-q^2/4}$ *for any* $q \in (0, \sqrt{M}]$.

Proof. From Lemma D.4 and D.1, we have $|\langle \hat{r}_i \cdot z_i - u_i, v_i\rangle| \leq \|\hat{r}_i z_i - u_i\|\|v_i\| = O(\sqrt{d}C\|v_i\|)$ and $\mathbb{E}\left[\langle \hat{r}_i \cdot z_i - u_i, v_i\rangle\right] = 0$. Thus, from the Hoeffding inequality, we have

$$\frac{1}{M}\sum_{i=1}^{M}\langle \hat{r}_i \cdot z_i - u_i, v_i\rangle = O\left(\sqrt{\frac{dC^2 \sum_{i=1}^{M}\|v_i\|^2}{M^2}} \cdot q\right),$$

with probability at least $1 - 2e^{-2q^2}$ for any $q > 0$.

For the second inequality, Lemma D.4 and D.1 imply $\|\hat{r}_i \cdot z_i - u_i\| = O(\sqrt{d}C)$ and $\mathbb{E}\left[\hat{r}_i \cdot z_i - u_i\right] = 0$. Thus, using the vector Bernstein inequality in Lemma C.3, we have

$$\left\|\frac{1}{M}\sum_{i=1}^{M}(\hat{r}_i \cdot z_i - u_i)\right\| = O\left(\sqrt{\frac{d}{M}}C(1+q)\right),$$

with probability at least $1 - e^{-q^2/4}$ for $q \in (0, \sqrt{M})$. This yields

$$\left\| \frac{1}{M} \sum_{i=1}^{M} (\hat{r}_i \cdot z_i - u_i) \right\|^2 = O\left(\frac{dC^2(1+q^2)}{M} \right).$$

For the third inequality, from the definition of $\hat{s}_i$ and Lemma D.4, we have

$$|\hat{s}_i| = \left| \frac{1}{1+c_1} (\hat{r}^2 - c_2 \hat{r} - c_3) \right| = O(C^2).$$

Thus, from the Hoeffding inequality, we have

$$\frac{1}{M} \sum_{i=1}^{M} \hat{s}_i - \frac{1}{M} \sum_{i=1}^{M} \|\Delta_i\|^2 \leq \frac{1}{M} \sum_{i=1}^{M} \hat{s}_i - \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}\left[\hat{s}_i\right] = O\left(C^2 q \sqrt{\frac{1}{M}} \right),$$

with probability at least $1 - e^{-q^2/2}$ for any $q > 0$. For the first inequality, we used Lemma D.3. □

E Proofs for Section 4.1

To tightly audit the privacy leakage of the Gaussian mechanism, we adopt the Rényi Differential Privacy (RDP) Mironov (2017).

Definition E.1 (RDP)**.** For any $\alpha \in (1, \infty)$ and any $\varepsilon > 0$, a mechanism $M : \mathcal{X} \to \mathcal{Y}$ is said to be (local) $(\alpha, \varepsilon) - RDP$ if for any inputs $x, x' \in \mathcal{X}$,

$$D_\alpha(M(x) \mid M(x')) := \frac{1}{\alpha - 1} \log \mathbb{E}_{\theta \sim M(x')} \left[\left(\frac{M(x)(\theta)}{M(x')(\theta)} \right)^\alpha \right] \leq \varepsilon.$$

LDP case Since the l^2-sensitivity of the local computation at each step is bounded by $2C$, as shown in Mironov (2017), Gaussian mechanism is $(\alpha, \alpha\rho)$-RDP, where $\rho = 2C^2/\sigma^2$

The RDP bound can be converted into the (ϵ, δ)-DP bound using the following lemma:

Lemma E.2 (Mironov (2017))**.** *Let M be (α, ε)-RDP for $\alpha \in (1, \infty)$. Then, M is $(\epsilon + \log(1/\delta)/(\alpha - 1), \delta)$-DP for every $\delta \in (0, 1)$.*

Applying this lemma, we obtain the result for the Gaussian mechanism.

CDP case The l^2-sensitivity of $\bar{\Delta}^{(t)}$ and $\frac{1}{M} \sum_{i=1}^{M} \left\| \Delta_i^{(t)} \right\|^2$ are bounded by $2C/M$ and C^2/M, respectively. Thus, $\bar{c}^{(t)}$ and $\frac{1}{M} \sum_{i=1}^{M} \left\| \Delta_i^{(t)} \right\|^2 + \xi^{(t)}$ satisfies $(\alpha, 2\alpha C^2/M\sigma^2)$-RDP and $(\alpha, \frac{\alpha C^4}{2M^2\sigma_\xi^2})$-RDP, respectively. Then, the entire training process with T iterations satisfy $(\alpha, \alpha(\rho + \rho_\xi))$-RDP, where $\rho = 2C^2T/M\sigma^2$, $\rho_\xi = C^4T/2M^2\sigma_\xi^2$. Applying Lemma E.2 yields Proposition 4.2.

F Proof for Theorem 4.5 and 4.6

To simplify the notation, let

$$h_i^{(t)} := -\Delta_i^{(t)}/(\eta_l \tau) = \frac{1}{\tau} \sum_{k=0}^{\tau-1} \nabla F_i(w_i^{(t,k)}),$$

$$\bar{h}^{(t)} := -\bar{\Delta}^{(t)}/(\eta_l \tau) = \frac{1}{M} \sum h_i^{(t)},$$

$$\bar{\epsilon}^{(t)} := -(\bar{c}^{(t)} - \bar{\Delta}^{(t)})/(\eta_l \tau)$$

$$\delta_s^{(t)} := \begin{cases} \frac{1}{M} \sum_{i=1}^{M} \left\| c_i^{(t)} \right\|^2 - d\sigma^2 - \frac{1}{M} \sum_{i=1}^{M} \left\| \Delta_i^{(t)} \right\|^2 & \text{for LDP-FedEXP with Gaussian,} \\ \frac{1}{M} \sum_{i=1}^{M} \hat{s}_i^{(t)} - \frac{1}{M} \sum_{i=1}^{M} \left\| \Delta_i^{(t)} \right\|^2 & \text{for LDP-FedEXP with PrivUnit,} \\ \xi^{(t)} & \text{for CDP-FedEXP.} \end{cases}$$

Then, the global step size $\eta_g^{(t)}$ is given by

$$\eta_g^{(t)} = \max\left\{1, \frac{\frac{1}{M}\sum_{i=1}^{M}\left\|h_i^{(t)}\right\|^2 + \delta_s^{(t)}/(\eta_l\tau)^2}{\left\|\bar{h}^{(t)} + \bar{\epsilon}^{(t)}\right\|^2}\right\}. \tag{9}$$

From the smoothness of F, $F(w^{(t+1)})$ satisfies the following:

$$\begin{aligned}
F(w^{(t+1)}) - F(w^{(t)}) &\leq -\eta_g\eta_l\tau\langle\nabla F(w^{(t)}), \bar{h}^{(t)} + \bar{\epsilon}^{(t)}\rangle + \frac{(\eta_g^{(t)})^2\eta_l^2\tau^2 L}{2}\|\bar{h}^{(t)} + \bar{\epsilon}^{(t)}\|^2, \\
&\leq -\eta_g\eta_l\tau\Bigg[\langle\nabla F(w^{(t)}), \bar{h}^{(t)} + \bar{\epsilon}^{(t)}\rangle \\
&\quad - \frac{\eta_l\tau L}{2}\max\left\{\frac{1}{M}\sum_{i=1}^{M}\left\|h_i^{(t)}\right\|^2 + \delta_s^{(t)}/(\eta_l\tau)^2, \left\|\bar{h}^{(t)} + \bar{\epsilon}^{(t)}\right\|^2\right\}\Bigg].
\end{aligned} \tag{10}$$

Here, the second inequality follows from Eq. equation 9.

For the right-hand side of Eq. equation 10, we have

$$\begin{aligned}
\langle\nabla F(w^{(t)}), \bar{h}^{(t)} + \bar{\epsilon}^{(t)}\rangle &= \langle\nabla F(w^{(t)}), \bar{h}^{(t)}\rangle + \langle\nabla F(w^{(t)}), \bar{\epsilon}^{(t)}\rangle \\
&= \frac{1}{2}\left(\left\|\nabla F(w^{(t)})\right\|^2 + \left\|\bar{h}^{(t)}\right\|^2 - \left\|\nabla F(w^{(t)}) - \bar{h}^{(t)}\right\|^2\right) + \langle\nabla F(w^{(t)}), \bar{\epsilon}^{(t)}\rangle \\
&\geq \frac{1}{2}\left\|\nabla F(w^{(t)})\right\|^2 - \frac{1}{2}\left\|\nabla F(w^{(t)}) - \bar{h}^{(t)}\right\|^2 - \left\|\nabla F(w^{(t)})\right\|\left\|\bar{\epsilon}^{(t)}\right\| \\
&\geq \frac{1}{2}\left\|\nabla F(w^{(t)})\right\|^2 - \frac{1}{2}\left\|\nabla F(w^{(t)}) - \bar{h}^{(t)}\right\|^2 \\
&\quad - \frac{1}{2}\left(\frac{1}{2}\left\|\nabla F(w^{(t)})\right\|^2 + 2\left\|\bar{\epsilon}^{(t)}\right\|^2\right) \\
&\geq \frac{1}{4}\left\|\nabla F(w^{(t)})\right\|^2 - \frac{1}{2M}\sum_{i=1}^{M}\left\|\nabla F_i(w^{(t)}) - h_i^{(t)}\right\|^2 - \left\|\bar{\epsilon}^{(t)}\right\|^2, \\
\left\|\bar{h}^{(t)} + \bar{\epsilon}^{(t)}\right\|^2 &\leq 2\left\|\bar{h}^{(t)}\right\|^2 + 2\left\|\bar{\epsilon}^{(t)}\right\|^2, \\
&\leq \frac{2}{M}\sum_{i=1}^{M}\left\|h_i^{(t)}\right\|^2 + 2\left\|\bar{\epsilon}^{(t)}\right\|^2.
\end{aligned}$$

Substituting the above inequalities into Eq. equation 10, we have

$$\begin{aligned}
F(w^{(t+1)}) - F(w^{(t)}) &\leq -\eta_g\eta_l\tau\Bigg[\frac{1}{4}\left\|\nabla F(w^{(t)})\right\|^2 - \frac{1}{2M}\sum\left\|\nabla F_i(w^{(t)}) - h_i^{(t)}\right\|^2 - \left\|\bar{\epsilon}^{(t)}\right\|^2 \\
&\quad - \frac{\eta_l\tau L}{2}\max\left\{\frac{1}{M}\sum_{i=1}^{M}\left\|h_i^{(t)}\right\|^2 + \delta_s^{(t)}/(\eta_l\tau)^2, \frac{2}{M}\sum_{i=1}^{M}\left\|h_i^{(t)}\right\|^2 + 2\left\|\bar{\epsilon}^{(t)}\right\|^2\right\}\Bigg] \qquad (11) \\
&\leq -\eta_g\eta_l\tau\Bigg[\frac{1}{4}\left\|\nabla F(w^{(t)})\right\|^2 - \frac{1}{2M}\sum\left\|\nabla F_i(w^{(t)}) - h_i^{(t)}\right\|^2 - \eta_l\tau L\cdot\underbrace{\frac{1}{M}\sum_{i=1}^{M}\left\|h_i^{(t)}\right\|^2}_{:=R} \\
&\quad - \underbrace{\left(\left\|\bar{\epsilon}^{(t)}\right\|^2 + \frac{\eta_l\tau L}{2}\max\left\{\frac{\delta_s^{(t)}}{(\eta_l\tau)^2} - \frac{1}{M}\sum_{i=1}^{M}\left\|h_i^{(t)}\right\|, 2\left\|\bar{\epsilon}^{(t)}\right\|^2\right\}\right)}_{:=T_4}\Bigg]. \qquad (12)
\end{aligned}$$

As in the proof of Theorem 2 in Jhunjhunwala et al. (2023), we have

$$\begin{aligned}
R &\leq \frac{1}{M}\sum \left\| h_i^{(t)} \right\|^2 \\
&\leq \frac{1}{M}\sum \left\| h_i^{(t)} - \nabla f_i(w^{(t)}) + \nabla f_i(w^{(t)}) - \nabla F(w^{(t)}) + \nabla F(w^{(t)}) \right\|^2 \\
&\leq \frac{3}{M}\sum \left(\left\| h_i^{(t)} - \nabla f_i(w^{(t)}) \right\|^2 + \left\| \nabla f_i(w^{(t)}) - \nabla F(w^{(t)}) \right\|^2 + \left\| \nabla F(w^{(t)}) \right\|^2 \right) \\
&\leq \frac{3}{M}\sum_{i=1}^{M} \left\| h_i^{(t)} - \nabla F_i(w^{(t)}) \right\|^2 + 3\left\| \nabla F(w^{(t)}) \right\|^2 + O(\sigma_g^2).
\end{aligned}$$

Substituting R into Eq. equation 12, we arrive at

$$\begin{aligned}
&F(w^{(t+1)} - F(w^{(t)})) \\
&\leq -\eta_g^{(t)}\eta_l\tau \left[\frac{1}{4}\left\| \nabla F(w^{(t)}) \right\|^2 - \frac{1}{2M}\sum \left\| \nabla F_i(w^{(t)}) - h_i^{(t)} \right\|^2 - \eta_l\tau L \cdot R - T_4 \right] \\
&\leq -\eta_g^{(t)}\eta_l\tau \left[\frac{1}{4}\left\| \nabla F(w^{(t)}) \right\|^2 - \frac{1}{2M}\sum \left\| \nabla F_i(w^{(t)}) - h_i^{(t)} \right\|^2 - \underbrace{O(\eta_l\tau L\sigma_g^2)}_{:=T_3} - T_4 \right. \\
&\quad \left. - \eta_l\tau L \cdot \left(\frac{3}{M}\sum_{i=1}^{M} \left\| h_i^{(t)} - \nabla F_i(w^{(t)}) \right\|^2 + 3\left\| \nabla F(w^{(t)}) \right\|^2 \right) \right] \\
&\leq -\eta_g^{(t)}\eta_l\tau \left[\frac{1}{8}\left\| \nabla F(w^{(t)}) \right\|^2 - \frac{\eta_l\tau L}{M}\sum_{i=1}^{M} \left\| \nabla F_i(w^{(t)}) - h_i^{(t)} \right\|^2 - T_3 - T_4 \right] \\
&\leq -\eta_g^{(t)}\eta_l\tau \left[\frac{1}{8}\left\| \nabla F(w^{(t)}) \right\|^2 - \underbrace{O\big(\eta_l^2\tau^2L^2\sigma_g^2\big)}_{T_2} - T_3 - T_4 \right].
\end{aligned}$$

Here, we used $\eta_l \leq 1/(24\tau L)$ and Lemma 7 in Jhunjhunwala et al. (2023).

Averaging over T iterations, we have

$$\frac{\sum \eta_g^{(t)} \left\| \nabla F(w^{(t)}) \right\|^2}{\sum \eta_g^{(t)}} \leq O\left(\frac{(F(w^{(0)}) - F^*)}{\sum \eta_g^{(t)}\eta_l\tau} + T_2 + T_3 + T_4 \right),$$

which implies

$$\min \left\| \nabla F(w^{(t)}) \right\|^2 \quad \leq O\left(\frac{F(w^0) - F^*}{\sum \eta_g^{(t)}\eta_l\tau} + T_2 + T_3 + T_4 \right).$$

The remaining task is to evaluate T_4. Recall that T_4 is defined as

$$\begin{aligned}
T_4 &= \left\| \bar{\epsilon}^{(t)} \right\|^2 + \frac{\eta_l\tau L}{2} \max\left\{ \frac{\delta_s^{(t)}}{(\eta_l\tau)^2} - \frac{1}{M}\sum_{i=1}^{M} \left\| h_i^{(t)} \right\|, 2\left\| \bar{\epsilon}^{(t)} \right\|^2 \right\} \\
&\leq (1 + \eta_l\tau L)\left\| \bar{\epsilon}^{(t)} \right\|^2 + \frac{L}{\eta_l\tau}\left(\delta_s^{(t)} - \frac{1}{M}\sum_{i=1}^{M} \left\| \Delta_i^{(t)} \right\|^2 \right).
\end{aligned}$$

For LDP-FedEXP with Gaussian mechanism, Lemma C.1 and C.2 yield

$$\left\|\bar{\epsilon}^{(t)}\right\|^2 \leq \frac{d}{(\eta_l\tau)^2}\cdot\left[1+q^2\right]\frac{\sigma^2}{M} = O\bigg(\frac{q^2}{(\eta_l\tau)^2}\frac{d\sigma^2}{M}\bigg),$$

$$\frac{1}{M}\sum_{i=1}^{M}\left\|\varepsilon_i^{(t)}\right\|^2 = d\cdot\left[1+\frac{q^2}{\sqrt{Md}}\right]\sigma^2$$

$$\frac{1}{M}\sum_{i=1}^{M}\langle\Delta_i^{(t)},\varepsilon_i^{(t)}\rangle \leq q\cdot\left(\frac{\sigma}{M}\sqrt{\sum_{i=1}^{M}\left\|\Delta_i^{(t)}\right\|^2}\right)$$

$$\leq \frac{1}{2M}\sum_{i=1}^{M}\left\|\Delta_i^{(t)}\right\|^2 + \frac{q^2\sigma^2}{2M},$$

with probability $1-Te^{-c\cdot q^2}$ for $q\in[1,\sqrt{M}]$, where c is a numerical constant. Here, we used the union bound over $t=1,\ldots,T$. Then, we obtain

$$\begin{aligned}
\delta_s^{(t)} - \frac{1}{M}\sum_{i=1}^{M}\left\|\Delta_i^{(t)}\right\| &= \frac{1}{M}\sum_{i=1}^{M}\left\|c_i^{(t)}\right\|^2 - d\sigma^2 - \frac{2}{M}\sum_{i=1}^{M}\left\|\Delta_i^{(t)}\right\|^2 \\
&= \frac{1}{M}\sum_{i=1}^{M}\left\|\Delta_i^{(t)}+\varepsilon_i^{(t)}\right\|^2 - d\sigma^2 - \frac{2}{M}\sum_{i=1}^{M}\left\|\Delta_i^{(t)}\right\|^2 \\
&= \frac{1}{M}\sum_{i=1}^{M}\left\|\varepsilon_i^{(t)}\right\|^2 - d\sigma^2 + \frac{2}{M}\sum_{i=1}^{M}\langle\Delta_i^{(t)},\varepsilon_i^{(t)}\rangle - \frac{1}{M}\sum_{i=1}^{M}\left\|\Delta_i^{(t)}\right\|^2 \\
&= q^2\cdot\sqrt{\frac{d}{M}}\sigma^2 + \frac{q^2\sigma^2}{M}.
\end{aligned}$$

Substituting these concentration inequalities, we obtain

$$\begin{aligned}
T_4 &= O\bigg((1+\eta_l\tau L)\frac{q^2}{(\eta_l\tau)^2}\frac{d\sigma^2}{M} + \frac{L}{\eta_l\tau}\bigg(q^2\cdot\sqrt{\frac{d}{M}}\sigma^2 + \frac{q^2\sigma^2}{M}\bigg)\bigg) \\
&= O\bigg(\frac{L\sigma^2q^2}{\eta_l\tau}\bigg[\frac{d}{M}+\sqrt{\frac{d}{M}}\bigg]\bigg),
\end{aligned}$$

since $q\geq 1$ and $\eta_l = \Theta(1/L\tau)$.

For LDP-FedEXP with PrivUnit, Lemma D.5 yields

$$\delta_s^{(t)} = \frac{1}{M}\sum_{i=1}^{M}\hat{s}_i^{(t)} = O(C^2q\sqrt{\frac{1}{M}}),$$

$$\left\|\bar{\epsilon}^{(t)}\right\|^2 = O\bigg(\frac{dC^2(1+q^2)}{M(\eta_l\tau)^2}\bigg)$$

with probability $1-Te^{-c\cdot q^2}$ for $q\in[1,\sqrt{M}]$, where c is a numerical constant. Substituting these concentration inequalities, we obtain

$$\begin{aligned}
T_4 &= O\bigg((1+\eta_l\tau L)\frac{dC^2(1+q^2)}{M(\eta_l\tau)^2}\bigg) + O\bigg(\frac{L}{\eta_l\tau}C^2q\sqrt{\frac{1}{M}}\bigg) \\
&= O\bigg(\frac{LC^2q^2}{\eta_l\tau}\bigg[\frac{d}{M}+\sqrt{\frac{1}{M}}\bigg]\bigg) \\
&= O\bigg(\frac{L\sigma^2q^2}{\eta_l\tau}\bigg[\frac{d}{M}+\sqrt{\frac{1}{M}}\bigg]\bigg).
\end{aligned}$$

For CDP-FedEXP, we have

$$\delta_s^{(t)} = \xi_i^{(t)} = O(q\sigma_\xi),$$
$$\left\|\bar{\epsilon}^{(t)}\right\|^2 = O\left(\frac{q}{(\eta_l\tau)^2}\frac{d\sigma^2}{M}\right),$$

with probability $1 - Te^{-c\cdot q^2}$ for $q \in [1, \sqrt{M}]$, where c is a numerical constant. Substituting these concentration inequalities, we obtain

$$\begin{aligned} T_4 &= O\left((1+\eta_l\tau L)\frac{q}{(\eta_l\tau)^2}\frac{d\sigma^2}{M} + \frac{L}{\eta_l\tau}q\sigma_\xi\right) \\ &= O\left(\frac{L\sigma^2q^2}{\eta_l\tau}\frac{d}{M}\right). \end{aligned}$$

G Supplementary Material for Numerical Experiments

Here, we provide additional details and results for the numerical experiments in Section 5.

G.1 Detailed Setup

Common Setup In both experiments, we run the training for $T = 50$ rounds and set $\sigma = 5 \cdot C/\sqrt{M}, \sigma_\xi = d\sigma^2/M$ for the CDP case, $\sigma = 0.7 \cdot C$ for the LDP (Gaussian) case, and $\varepsilon_0 = \varepsilon_1 = \varepsilon_2 = 2$ for the LDP (PrivUnit) case. Following Jhunjhunwala et al. (2023), we set the final model as the average of the last 2 iterates to mitigate the effect of oscillating behavior of DP-FedEXP. For privacy analysis, we utilized the numerical composition (Gopi et al., 2021) to tightly audit the privacy leakage.

Synthetic Experiment Setup First, we generate the target vector $w^* \in \mathbb{R}^d$ according to the standard normal distribution, which is shared among all clients. Then, we generate the local dataset following a similar procedure in Li et al. (2020); Jhunjhunwala et al. (2023) with $M = 1000$. In this experiment, we set $\tau = 20$. For the CDP setting, we set $d = 500$ while $d = 100$ for the LDP setting since the noise level of LDP is much larger than that of CDP.

Realistic Experiment Setup We divide the training data into $M = 1000$ clients according to Dirichlet distribution with $\alpha = 0.3$, following the procedure in Hsu et al. (2019). In this experiment, we set $\tau = 10$. For the CDP setting, we use a simple convolutional neural network (CNN) model with two convolutional layers and two fully connected layers. For LDP setting, we use a small CNN model with two convolutional layers and one fully connected layer.

Hyperparameter Tuning We tuned the hyper parameters (local learning rate η_l and clipping threshold C) via grid search and select the best hyperparameters which maximize the test accuracy for the realistic dataset or minimize the training loss for the synthetic dataset averaged over the last 5 rounds. In the synthetic experiment, the grid for η_l is $\{0.01, 0.03, 0.1, 0.3, 1\}$ and for C is $\{0.1, 0.3, 1, 3, 10\}$. In the realistic experiment, the grid for η_l is $\{0.0001, 0.0003, 0.001, 0.003, 0.01\}$ and for C is $\{0.1, 0.3, 1, 3, 10\}$. We summarize the best performing hyperparameters in Table 2.

Synthetic Dataset In principle, we follow a similar procedure in Li et al. (2020); Jhunjhunwala et al. (2023). First, we generate the true model w^* by sampling from the standard normal distribution. Then, we generate vectors $x_i \in \mathbb{R}^d$ according to $x_i \sim \mathcal{N}(m_i, I_d)$, where $m_i \sim \mathcal{N}(u_i, 1), u_i \sim \mathcal{N}(0, 0.1)$. The client objective is defined as $f_i(w) := \left\|x_i^\top w - y_i\right\|^2$, where $y_i = x_i^\top w^*$.

Table 2: Best hyperparameters selected via grid search for DP-FedEXP, DP-FedAvg, and DP-SCAFFOLD.

		FedEXP		FedAvg		SCAFFOLD	
Dataset	DP type	η_l	C	η_l	C	η_l	C
Synthetic	LDP (Gaussian)	0.003	0.3	0.003	3	0.003	0.3
	LDP (PrivUnit)	0.003	1	0.003	3	0.003	0.3
	CDP	0.001	0.3	0.003	3	0.001	1
MNIST	LDP (Gaussian)	0.03	0.1	0.03	0.3	0.1	0.1
	LDP (PrivUnit)	0.03	0.3	0.03	0.3	0.03	0.1
	CDP	0.1	0.3	0.1	1	0.1	0.3

Model Architectures We summarize the architectures of the models used in the MNIST experiments in Table 3.

Table 3: Model architectures used in the experiments.

Setting	Model Architecture
CDP	Convolutional layer (4 filters, 4x4) Convolutional layer (8 filters, 4x4) Fully connected layer ($128 \to 32$) ReLU activation Fully connected layer ($32 \to 10$) Softmax activation
LDP	Convolutional layer (2 filters, 4x4) Convolutional layer (1 filters, 4x4) Fully connected layer ($16 \to 10$) Softmax activation

G.2 Additional Results

Here, we provide additional results omitted in the main text due to space constraints.

Adaptivity in Global Step Size Fig. 3 plots the global step size $\eta_g^{(t)}$ of each algorithm. Interestingly, in the synthetic experiment, the global step size of DP-FedEXP decreases as the training progresses. This enables to speed up the training process and to mitigate the effect of the DP noise on the converged model at the same time. This phenomenon clearly demonstrates the advantage of the adaptive step size in DP-FL.

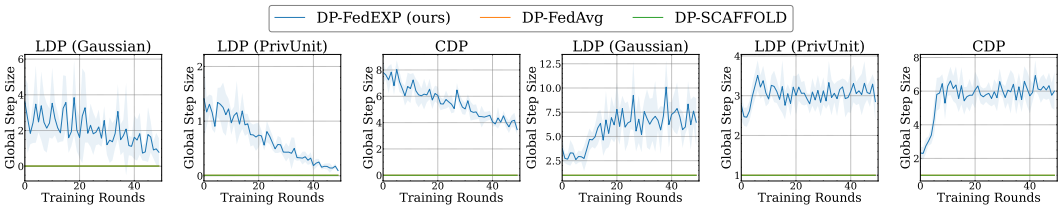

Figure 3: Global step sizes for the synthetic dataset (left) and the MNIST dataset (right).

Additional Results for the MNIST Dataset To evaluate the performance of the model at the end of the training process, we report the test accuracy averaged over the last 5 rounds in Table 4. Our proposed DP-FedEXP comprehensively outperforms the baselines in all settings.

Table 4: Test accuracy of algorithms on the MNIST dataset averaged over the last 5 rounds. Mean (standard deviation) over 5 runs with different random seeds is reported.

DP Type	DP-FedEXP	DP-FedAvg	DP-SCAFFOLD
LDP (Gaussian)	**80.24** (0.94)	78.69 (1.26)	66.89 (2.29)
LDP (PrivUnit)	**79.65** (1.23)	78.40 (1.18)	56.83 (3.95)
CDP	**94.57** (0.19)	92.88 (0.29)	86.61 (0.52)