FEDERATED LEARNING WITH LOCAL DIFFERENTIAL PRIVACY: TRADE-OFFS BETWEEN PRIVACY, UTILITY, AND COMMUNICATION

Muah Kim[⋆]*, Onur Günlü*[⋆]*, and Rafael F. Schaefer*[†]

[⋆]Information Theory and Applications Chair
Technische Universität Berlin
Berlin, Germany
{muah.kim, guenlue}@tu-berlin.de

[†]Lehrstuhl für Nachrichtentechnik/Kryptographie und Sicherheit
Universität Siegen
Siegen, Germany
rafael.schaefer@uni-siegen.de

ABSTRACT

Federated learning (FL) allows to train a massive amount of data privately due to its decentralized structure. Stochastic gradient descent (SGD) is commonly used for FL due to its good empirical performance, but sensitive user information can still be inferred from weight updates shared during FL iterations. We consider Gaussian mechanisms to preserve local differential privacy (LDP) of user data in the FL model with SGD. The trade-offs between user privacy, global utility, and transmission rate are proved by defining appropriate metrics for FL with LDP. Compared to existing results, the query sensitivity used in LDP is defined as a variable and a tighter privacy accounting method is applied. The proposed utility bound allows heterogeneous parameters over all users. Our bounds characterize how much utility decreases and transmission rate increases if a stronger privacy regime is targeted. Furthermore, given a target privacy level, our results guarantee a significantly larger utility and a smaller transmission rate as compared to existing privacy accounting methods.

***Index Terms*—** federated learning (FL), local differential privacy (LDP), stochastic gradient descent (SGD), Gaussian randomization, composition theorems.

1. INTRODUCTION

Differential privacy (DP) is widely used due to its strong privacy guarantees. DP tackles the privacy leakage about single data belonging to an individual in a dataset when some information from the dataset is publicly available. Common DP mechanisms add an independent random noise component to available data to provide privacy, which can be provided by using local sources of randomness such as physical unclonable functions (PUFs) [1]. Applied to machine learning (ML), preserving DP of a training dataset is studied, e.g., in [2–6]. Among various ML models, federated learning (FL) is a promising option due to its decentralized structure [7–9]. Since user data are not collected by an aggregator in FL, local differential privacy (LDP) [10] of an FL model is studied in the literature to guarantee individual user privacy, e.g., for a wireless multiple-access channel [11], by using splitting/shuffling [12], dimension selection [13], with experimental evaluations [14], and with a communication-efficient algorithm [15]. Due to the iterative process in ML algorithms, the violation of LDP after multiple rounds of weight updates needs to be addressed. It can be done by using privacy accounting methods of DP such as the sequential composition theorem (SC) [16], the advanced composition theorem (AC1) [5, 17], an improved advanced composition theorem (AC2) [18], and the moments accountant (MA) [3]. Unlike the composition theorems that directly compose the DP parameters, the MA approach circumvents the composition by converting DP into Rényi-differential privacy (RDP), whose composition has a simple linear form. The MA is shown to outperform the AC1 and the AC2 when a Gaussian noise mechanism is used [3]. The MA is further improved by using the optimal conversion from parameters of RDP to DP [19]. An FL model with stochastic gradient descent (FedSGD) with LDP is not yet investigated with a comprehensive theoretical analysis that considers privacy, utility, and communication jointly. This paper focuses on trade-offs between privacy, utility, and transmission rate, where LDP is provided to the FedSGD model by using a Gaussian mechanism. In [11], the trade-offs between those three metrics are analyzed for the non-stochastic gradient descent algorithm for learning. Most of related studies consider the SC [12–14], the AC1 [11], and the MA [15] for privacy accounting, all of which can be improved by using the privacy accounting method proposed in [19].

One main contribution of this paper is the privacy analysis for the FedSGD model by using an enhanced MA suggested in [19]. Furthermore, we propose a generic utility metric that considers the query sensitivity as a varying parameter, unlike in the literature, and we provide a lower bound on the utility metric. Our utility bound considers system heterogeneity by allowing users to have distinct dataset sizes, data sampling probabilities, and target privacy levels. The transmission rate

This work was supported in part by the German Federal Ministry of Education and Research (BMBF) within the national initiative for "*Post Shannon Communication (NewCom)*" under the Grant 16KIS1004 and in part by the German Research Foundation (DFG) under the Grant SCHA 1944/7-1.

is considered as the differential entropy of the noisy gradients for lossless communications. We illustrate significant gains from our bounds in terms of the required noise power, the utility metric, and the transmission rate as compared to the existing methods.

2. SYSTEM MODEL

2.1. Federated SGD (FedSGD)

Consider that the FedSGD method consists of a central server and K users. The K users are assumed to have their own neural networks with the same structure. At each time step t such that $t \in \{1, 2, \ldots, T\}$, the server distributes the aggregated weight values $\mathbf{w}^{(t)} \in \mathbb{R}^d$ to all users and the K users' networks are initialized with those weight values. Then, user k randomly samples a dataset $\mathcal{J}_k^{(t)}$ from its whole dataset $\mathcal{D}_k$ with probability q_k and calculates the gradient $\mathbf{g}_k^{(t)}$ from $\mathcal{J}_k^{(t)}$. In particular, we use a loss function $\ell(\mathbf{w}^{(t)}, x)$ defined for each data sample x and given weights $\mathbf{w}^{(t)}$, and we represent the local (per user) loss $\mathcal{L}_k(\mathbf{w}^{(t)}, \mathcal{J}_k^{(t)}) \in \mathbb{R}$ as

$$\mathcal{L}_k(\mathbf{w}^{(t)}, \mathcal{J}_k^{(t)}) \coloneqq \frac{1}{|\mathcal{J}_k^{(t)}|} \sum_{x \in \mathcal{J}_k^{(t)}} \ell(\mathbf{w}^{(t)}, x) \tag{1}$$

where $|\mathcal{J}_k^{(t)}|$ denotes the size of a set $\mathcal{J}_k^{(t)}$. Suppose the difference between the true loss and an empirical loss can be made negligible by using a sufficiently large dataset. Each user's local gradient $\mathbf{g}_k^{(t)}$ is then represented as

$$\mathbf{g}_k^{(t)}(\mathcal{J}_k^{(t)}) = \nabla_{\mathbf{w}} \mathcal{L}_k(\mathbf{w}^{(t)}, \mathcal{J}_k^{(t)}) = \frac{1}{|\mathcal{J}_k^{(t)}|} \sum_{x \in \mathcal{J}_k^{(t)}} \nabla_{\mathbf{w}} \ell(\mathbf{w}^{(t)}, x)$$

where $\nabla_{\mathbf{w}}$ represents the gradient along the weight vector $\mathbf{w} = (w_1, w_2, \ldots, w_d)$, i.e.,$\nabla_{\mathbf{w}} = (\frac{\partial}{\partial w_1}, \frac{\partial}{\partial w_2}, \ldots, \frac{\partial}{\partial w_d})$. Let $\|\mathbf{x}\|$ denote the ℓ_2-norm of a d-dimensional vector $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, i.e., $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^d x_i^2}$. Assume that G is the maximum ℓ_2-norm value of all possible gradients for any given weight vector $\mathbf{w}_k$ and sampled dataset $\mathcal{J}_k$, i.e., $G = \sup_{\mathbf{w}_k \in \mathbb{R}^d, \mathcal{J}_k \in \mathcal{D}_k} \mathbb{E}[\|\mathbf{g}_k(\mathcal{J}_k)\|]$. Each user clips its local gradient by a clipping threshold value $C \in (0, G]$ as

$$\bar{\mathbf{g}}_k^{(t)} = \mathbf{g}_k^{(t)} \Big/ \max\left\{1, \left\|\mathbf{g}_k^{(t)}\right\| / C\right\}. \tag{2}$$

The clipped gradients $\{\bar{\mathbf{g}}_k^{(t)}\}_{k=1}^K$ are sent to the server and aggregated to obtain the updated weight vector $\mathbf{w}^{(t+1)}$. Suppose we use a learning rate η_t, then the weight update is

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta_t \cdot \sum_{k=1}^K \frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|} \bar{\mathbf{g}}_k^{(t)} \tag{3}$$

where $\mathcal{J}^{(t)} = \cup_{k=1}^K \mathcal{J}_k^{(t)}$. The global loss of the FL system is

$$\begin{aligned}\mathcal{L}(\mathbf{w}^{(t)}, \mathcal{J}^{(t)}) &= \sum_{x \in \mathcal{J}^{(t)}} \frac{1}{|\mathcal{J}^{(t)}|} \ell(\mathbf{w}^{(t)}, x) \\ &= \sum_{k=1}^K \frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|} \mathcal{L}_k(\mathbf{w}^{(t)}, \mathcal{J}_k^{(t)}) \end{aligned} \tag{4}$$

which can be obtained by taking a weighted sum of the local losses defined in (1). Thus, the weighted average of local gradients is equivalent to the gradient of the global loss calculated with the whole sampled data, i.e., we have

$$\begin{aligned}\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}^{(t)}, \mathcal{J}^{(t)}) &= \frac{1}{|\mathcal{J}^{(t)}|} \sum_{x \in \mathcal{J}^{(t)}} \nabla_{\mathbf{w}} \ell(\mathbf{w}^{(t)}, x) \\ &= \sum_{k=1}^K \frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|} \sum_{x \in \mathcal{J}_k^{(t)}} \frac{1}{|\mathcal{J}_k^{(t)}|} \nabla_{\mathbf{w}} \ell(\mathbf{w}^{(t)}, x) \\ &= \sum_{k=1}^K \frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|} \mathbf{g}_k^{(t)}(\mathbf{w}^{(t)}, \mathcal{J}_k^{(t)}). \end{aligned} \tag{5}$$

This implies that such an FL model results in the same weight update as the centralized model, so one global loss optimization problem can be divided into multiple local problems [7].

2.2. Local differential privacy (LDP)

FL guarantees a certain level of privacy since the users do not directly send their data to the central server publicly [7]. However, a certain amount of information can still be inferred from the shared information of the local networks, so a privacy mechanism is still necessary to protect user data. We consider LDP to guarantee individual user privacy. A mechanism $\mathcal{M}_k$ is $(\epsilon_k, \delta_k)-$LDP w.r.t. the user k's dataset $\mathcal{D}_k$, if any two neighboring datasets $D \sim D' \subseteq \mathcal{D}_k$ satisfy for any $\mathcal{S} \subseteq \text{range}(\mathcal{M}_k)$ that

$$\begin{aligned}&\Pr\left[\mathcal{M}_{k,D}(\mathbf{g}_k) \in \mathcal{S}\right] \\ &\qquad \le e^{\epsilon_k} \cdot \Pr\left[\mathcal{M}_{k,D'}(\mathbf{g}_k) \in \mathcal{S}\right] + \delta_k. \end{aligned} \tag{6}$$

We assume that the local gradients $\mathbf{g}_k^{(t)}$ are clipped and then LDP is satisfied by adding a Gaussian noise component $\mathbf{Z}_k$. Suppose the Gaussian noise variance of each dimension is proportional to C^2, i.e., $\mathbf{Z}_k \sim \mathcal{N}(\mathbf{0}, C^2\sigma_k^2\mathbf{I}_d)$ for some $\sigma_k^2 > 0$, where $\mathbf{I}_d$ is the $d \times d$ identity matrix. Denote the noisy gradients as $\tilde{\mathbf{g}}_k^{(t)}$, so the Gaussian LDP mechanism can be represented as $\tilde{\mathbf{g}}_k^{(t)} = \mathcal{M}_k(\bar{\mathbf{g}}_k^{(t)}) = \bar{\mathbf{g}}_k^{(t)} + \mathbf{Z}_k \sim \mathcal{N}(\bar{\mathbf{g}}_k^{(t)}, C^2\sigma_k^2\mathbf{I}_d)$. With such an LDP mechanism, the weight update equation (3) can be used by replacing $\bar{\mathbf{g}}_k^{(t)}$ with $\tilde{\mathbf{g}}_k^{(t)}$ for all $t = 1, 2, \ldots, T$.

3. TRADE-OFFS BETWEEN PRIVACY, UTILITY, AND TRANSMISSION RATE

We next characterize the Gaussian noise variance required to guarantee a target LDP level after T rounds of weight updates

for FL with LDP, i.e., *T-fold composition*. Furthermore, we analyze the effects of the Gaussian noise on the utility and the transmission rate, given a target LDP level. With an example we illustrate that a tighter privacy composition bound can yield a significantly larger utility and smaller transmission rate for the same target LDP level.

3.1. Theoretical Analysis

We define utility $\mathcal{U}(T)$ after T iterations as the multiplicative inverse of the convergence rate, i.e., we have

$$\mathcal{U}(T) = \frac{1}{\mathbb{E}[\mathcal{L}(\mathbf{w}^{(T)}, \mathcal{J}^{(T)})] - \mathcal{L}(\mathbf{w}^*)} \tag{7}$$

where $\mathbf{w}^*$ is the optimal weight vector that minimizes the global loss, i.e., $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \mathcal{L}(\mathbf{w}, \cup_{k=1}^K \mathcal{D}_k)$. This utility metric is used instead of accuracy to track the learning performance analytically. Consider that the transmission of a noisy gradient is lossless, and define the user k's transmission rate $R_{\text{tr},k}$ by the differential entropy of its noisy gradient, i.e., $R_{\text{tr},k} = h(\tilde{\mathbf{g}}_k^{(t)})$ for $t = 1, 2, \ldots, T$. The transmission rate is measured by a differential entropy term for simplicity, which can be extended by allowing distortion. We remark that the transmission rate can be further reduced with quantization, as suggested, e.g., in [20].

The following theorem provides the trade-offs between LDP parameters, utility, and the transmission rate for the assumed FL model; see Appendix A for its proof.

Theorem 1. *User k's Gaussian mechanism $\mathcal{M}_k$, where $k = 1, 2, \ldots, K$, with noise variance of each dimension $C^2\sigma_k^2$ is (ϵ_k, δ_k)-LDP after T rounds of weight updates for*

$$\epsilon_k > 2\log(\delta_k^{-1}) \max(\delta_k, \frac{1}{\sigma_k^2 \ln(\frac{1}{q_k\sigma_k})}), \tag{8}$$

$$q_k < \frac{1}{16\sigma_k}, \qquad \textit{and} \qquad \sigma_k \geq 1 \tag{9}$$

if we have

$$\sigma_k^2 \geq \frac{4q_k^2 T}{1-q_k}\left[\frac{2}{\epsilon_k^2}\log\frac{1}{\delta_k} + \frac{1}{\epsilon_k} - \frac{2}{\epsilon_k^2}\left(\log(2\log\delta_k^{-1}) + 1 - \log\epsilon_k\right)\right] + \mathcal{O}\left(\frac{\log^2(\log\delta_k^{-1})}{\log\delta_k^{-1}}\right). \tag{10}$$

For a μ-smooth and λ-strongly convex loss $\mathcal{L}(\mathbf{w}, \mathcal{S})$ with respect to a d-dimensional weight vector $\mathbf{w} \in \mathbb{R}^d$ given an arbitrary subset $\mathcal{S}$ of $\mathcal{D}$ such that $\mathcal{S} \subseteq \mathcal{D}$ and for a learning rate $\eta_t = \frac{G}{C\lambda t}$, the utility of the noisy FedSGD model after T iterations is bounded as

$$\mathcal{U}(T) \geq \frac{\lambda^2 T}{\mu G^2} \min\left\{\frac{1}{2}, \frac{1}{1+d\sigma^2}\right\} \tag{11}$$

where $\sigma^2 = \frac{\sum_{k=1}^K (|D_k| q_k \sigma_k)^2}{(\sum_{k=1}^K |D_k| q_k)^2}$, and G is the maximum value of the gradient. The transmission rate $R_{tr,k}$ of user k with the noise power of each dimension $C^2\sigma_k^2$ can be bounded as

$$R_{tr,k} \leq d\log(2\pi e C^2 \sigma_k / \sqrt{d}). \tag{12}$$

Table 1: Noise variance lower bounds used for comparisons.

Composition Method	Lower Bound of σ_k^2
Proposed	(10)
MA [3]	$\frac{4q_k^2 T}{1-q_k}\Big(\frac{2}{\epsilon_k^2}\log\frac{1}{\delta_k} + \frac{1}{\epsilon_k} + \mathcal{O}(\log\delta_k^{-1})\Big)$
AC1 [5, 17]	(13)
AC2 [18]	$\frac{4q_k^2}{1-q_k}\frac{8T\log(e + \frac{\epsilon_k}{\delta_k})}{\epsilon_k^2}$

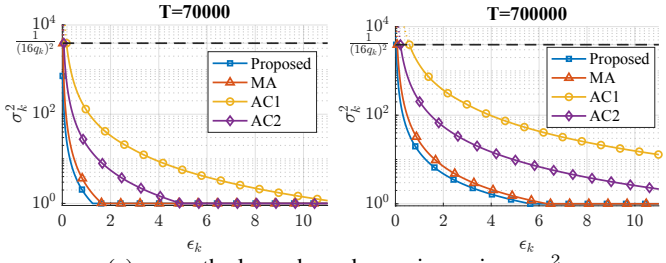
Proof Sketch. The noise variance bound (10) follows by extending [19, Theorem 5], where a generalized RDP-DP conversion is applied, to allow random sampling from each dataset with probability q_k and a query sensitivity of $2C$ by using the proof of [3, Lemma 3]. The utility bound (11) is obtained by extending [21, Lemma 1] with an additional assumption of noisy SGD algorithm on the FL model. In particular, we introduce an adaptive learning rate η_t that depends on the clipping threshold C to bound the utility when the gradients are aggregated from K users, where the noisy gradient of each user is obtained by randomly sampling the data, clipping the gradients, and adding Gaussian noise to the clipped gradients. It is followed by the transmission rate bound derived from the upper bound on the differential entropy when the random vector's covariance is upper-bounded. □

Theorem 1 illustrates the trade-offs between three metrics. If, e.g., the noise variance σ^2 is increased to guarantee stronger privacy, the utility is lower bounded by a smaller value and the transmission rate upper bound increases. For a sufficiently small δ_k, the term in big $\mathcal{O}$ notation in (10) becomes negligible as compared to the other terms. In that case, the noise variance σ^2 is increasing linearly with T. With this choice of noise variance $\sigma^2 \propto T$, the denominator of the utility bound (11) increases linearly with T. Thus, the utility bound converges to a constant even if T tends to infinity. This implies that with a Gaussian noise mechanism used for LDP, the utility lower bound can be finite even when T tends to infinity, and the achievement of the minimum loss $\mathcal{L}(\mathbf{w}^*)$ is not guaranteed. For this case, the maximum difference between the achieved loss and the optimal loss, which is not necessarily zero, is upper bounded.

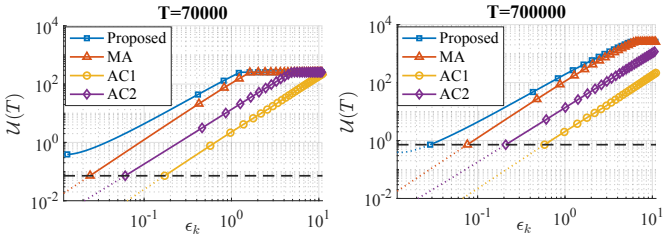
3.2. Local Privacy Accounting Method Comparisons

We compare the bounds in Theorem 1 with the results obtained from the MA, AC1, and AC2. We do not consider the SC for comparison since the AC1 is known to outperform the SC method. Table 1 lists the noise variance bounds of those composition methods when they are applied to the assumed model with a data sampling probability q_k and a clipping threshold $C > 0$, i.e., the query sensitivity is $2C$.

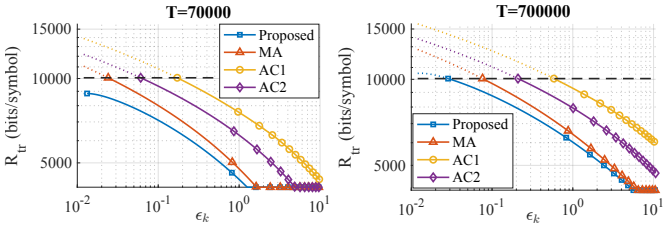
We obtain the noise bound with the AC1 by using its implicit solution. Consider that each weight update round is



(a) ϵ_k vs. the lower bound on noise variance σ_k^2.



(b) ϵ_k vs. the lower bound on utility $\mathcal{U}(T)$.



(c) ϵ_k vs. the upper bound on the transmission rate $R_{\text{tr},k}$.

Fig. 1: The noise variance σ_k^2 and utility $\mathcal{U}(T)$ lower bounds vs. ϵ_k for $T = 7\times10^4, 7\times10^5$ and with parameters $\delta_k = 10^{-4}, q_k = 10^{-3}$ for all $k = 1, 2, \ldots, 100, d = 10^4, \mu = 1, \lambda = 1, C = 1$, and $G = 5$.

$(\epsilon_0, \delta_0)-$LDP that results in $(\epsilon_k, \delta_k)-$LDP after T iterations, which satisfies $\epsilon_k = \sqrt{2T\ln(\tilde{\delta}^{-1})}\epsilon_0 + T\epsilon_0(e^{\epsilon_0} - 1)$ and $\delta_k = T\delta_0 + \tilde{\delta}$ by [5, Theorem 3.20]. We obtain (ϵ_0, δ_0) from given (ϵ_k, δ_k) by choosing $\tilde{\delta} = 10^{-5}$ and plug it into the noise bound of a Gaussian mechanism from [5, Theorem 3.22] as

$$\sigma_k^2 \geq \frac{4q_k^2}{1-q_k}\frac{2}{\epsilon_0}\log\left(\frac{4}{5\delta_0}\right). \tag{13}$$

The factor $\frac{4q_k^2}{1-q_k}$ of the noise bounds results from using the query sensitivity of $2C$ and randomly sampled datasets with probability q_k. For evaluations, we assume a homogeneous system consisting of $K = 100$ users with the same parameters $q_k, \epsilon_k, \delta_k$, and σ_k. The aggregated noise variance σ^2 is then obtained as $\sigma^2 = \frac{1}{K}\sigma_k^2$. We choose the minimum possible σ_k^2 for each method and calculate the utility bound by plugging the noise variance σ^2 into (11) with $\delta_k = 10^{-4}, q_k = 10^{-3}, d = 10^4, \mu = 1, \lambda = 1, C = 1$, and $G = 5$.

The noise bound, the resulting utility, and the transmission rate are illustrated in Fig. 1 for parameters that satisfy corresponding constraints of each composition method for the number of iterations T of 7×10^4 and 7×10^5. The obtained noise variance bounds are valid in the region such that $\sigma_k^2 < 1/(16q_k)^2 = 3906.25$, represented by the black dashed lines in Fig. 1a. This condition also limits the utility bounds at $\{0.0072, 0.0717\}$ for $T = \{7, 70\}\times10^4$, respectively, and the transmission rate bound at 9.69×10^3 (bits/symbol), which correspond to the black dashed lines in Fig. 1b and Fig. 1c, respectively. The noise bounds in Fig. 1a decrease to 1 and then stay constant as the target privacy level ϵ_k increases. These conditions are from [3, Lemma 3], which are required to measure the effect of randomly sampled data on the noise variance. The utility bound and the transmission rate bound might be improved if the noise variance bound does not require the condition $\sigma_k \geq 1$. We remark that ϵ values shown in [19, Fig. 3] seem to be unfortunately wrong; furthermore, [19, Theorem 5] seems to lack a condition on ϵ that follows from [3, Lemma 3], which imposes a condition on α used in [19, Eq. (80)]. The violation of this condition makes the range of T values considered in [19, Fig. 3] invalid.

The proposed method requires the smallest noise variance, followed by MA, AC2, and AC1, for the same ϵ_k for every T considered in Fig. 1a. The required noise variance of every method increases with T because more iterations deteriorate individual privacy and a larger amount of noise is necessary to meet the same target privacy. Fig. 1b shows the utility bounds obtained with σ_k^2 values from Fig. 1a. The utility bound curves stay constant after they reach their maximum values. The maximum value of the utility bound increases with T as the weight vector gradually converges to the optimum at every iteration even if the noise variance also increases. Thus, the utility bound is not degraded by the noise variance if the target privacy ϵ_k is large enough and the noise variance σ_k^2 is small enough for homogeneous users. If we target $\epsilon_k = 0.3$ after $T = 7\times10^4$ iterations, for example, the guaranteed utility bounds are at $25.83, 10.79, 0.22$, and 1.40 when the proposed bound, MA, AC1, and AC2 are used, respectively. The corresponding transmission rate bounds are at $\{5.81, 6.44, 9.26, 7.91\}\times10^3$ (bits/symbol). Hence, a significantly larger utility bound and a smaller transmission rate can be achieved for the same privacy constraint ϵ_k with a tighter privacy composition bound.

4. CONCLUSION

Trade-offs between privacy, utility, and transmission rate of a FedSGD model with a Gaussian LDP mechanism were proved. We provided a noise variance bound that guarantees a given LDP level after multiple rounds of weight updates by using a tight composition theorem. The proposed utility bound allows distinct parameters for all users and allows the gradients to be clipped and noisy. The noise variance required is illustrated to be significantly smaller than the ones obtained by using existing privacy composition methods MA, AC1, and AC2. Similarly, our bounds lead to a significantly larger utility and a smaller transmission rate. In future work, we will illustrate gains from our bounds as compared to existing methods for large available datasets used for FL.

5. REFERENCES

[1] O. Günlü, *Key Agreement with Physical Unclonable Functions and Biometric Identifiers*, Ph.D. thesis, TU Munich, Germany, Nov. 2018, published by Dr. Hut Verlag in Feb. 2019.

[2] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Privacy aware learning," *J. ACM*, vol. 61, no. 6, pp. 1–57, Dec. 2014.

[3] M. Abadi, A. Chu, I. Goodfellow, H B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *ACM Conf. Comput. Commun. Security*, Vienna, Austria, Oct. 2016, pp. 308–318.

[4] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *J. Mach. Learn. Research*, vol. 12, no. 3, Nov. 2011.

[5] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comp. Sci.*, vol. 9, no. 3-4, pp. 211–407, 2014.

[6] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *ACM Conf. Comput. Ccommun. Security*, Denver, CO, Oct. 2015, pp. 1310–1321.

[7] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Int. Conf. Artif. Intell. Statist.*, Ft. Lauderdale, FL, Apr. 2017, pp. 1273–1282.

[8] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," [Online]. Available: arxiv.org/abs/1712.07557, Dec. 2017.

[9] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q.S. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3454–3469, Apr. 2020.

[10] B. Ding, J. Kulkarni, and S. Yekhanin, "Collecting telemetry data privately," in *Adv. Neural Inf. Process. Syst.*, Long Beach, CA, Dec. 2017, pp. 3571–3580.

[11] M Seif, R. Tandon, and M. Li, "Wireless federated learning with local differential privacy," [Online]. Available: arxiv.org/abs/2002.05151, Feb. 2020.

[12] L. Sun, J. Qian, X. Chen, and P. S. Yu, "LDP-FL: Practical private aggregation in federated learning with local differential privacy," [Online]. Available: arxiv.org/abs/2007.15789, July 2020.

[13] R. Liu, Y. Cao, M. Yoshikawa, and H. Chen, "FedSel: Federated SGD under local differential privacy with top-k dimension selection," [Online]. Available: arxiv.org/abs/2003.10637, Mar. 2020.

[14] S. Truex, L. Liu, K. H. Chow, M. E. Gursoy, and W. Wei, "LDP-Fed: Federated learning with local differential privacy," in *ACM Int. Workshop Edge Syst., Analytics Netw.*, Crete, Greece, Apr. 2020, pp. 61–66.

[15] L. Wang, R. Jia, and D. Song, "D2P-Fed: Differentially private federated learning with efficient communication," [Online]. Available: arxiv.org/pdf/2006.13039, Oct. 2020.

[16] C. Dwork and J. Lei, "Differential privacy and robust statistics," in *ACM Symp. Theory Comput.*, Bethesda, MD, May 2009, pp. 371–380.

[17] C. Dwork, G. N. Rothblum, and S. Vadhan, "Boosting and differential privacy," in *IEEE Annual Symp. Found. Comput. Sci.*, Las Vegas, NV, Oct. 2010, pp. 51–60.

[18] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," in *Int. Conf. Mach. Learn.*, Lille, France, July 2015, pp. 1376–1385.

[19] S. Asoodeh, J. Liao, F. P. Calmon, O. Kosut, and L. Sankar, "A better bound gives a hundred rounds: Enhanced privacy guarantees via f-divergences," [Online]. Available: arxiv.org/abs/2001.05990, Jan. 2020.

[20] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, "UVeQFed: Universal vector quantization for federated learning," [Online]. Available: arxiv.org/abs/2006.03262, July 2020.

[21] A. Rakhlin, O. Shamir, and K. Sridharan, "Making gradient descent optimal for strongly convex stochastic optimization," [Online]. Available: arxiv.org/pdf/1109.5647, Dec. 2012.

A. PROOF OF THEOREM 1

Clipping gradient norms with C makes the query sensitivity to be at most $2C$ because for arbitrary clipped gradients $\bar{\mathbf{g}}_1$ and $\bar{\mathbf{g}}_2$ the following inequality holds:

$$\max_{\forall \bar{\mathbf{g}}_1, \bar{\mathbf{g}}_2} \sqrt{\sum_{i\in[d]} (\bar{g}_{1,i} - \bar{g}_{2,i})^2} = \max_{\forall \bar{\mathbf{g}}_1} \sqrt{\sum_{i\in[d]} (\bar{g}_{1,i} - (-\bar{g}_{1,i}))^2} \leq 2C.$$

The proof of (10) follows mainly from [19, Theorem 5], which provides the lower bound of noise variance when the query sensitivity is 1 and the same dataset is repeatedly used at every iteration. [19, Theorem 5] is obtained by using MA through the following steps:

- Compute $\gamma(\alpha)$ such that the Gaussian mechanism is $(\alpha, \gamma(\alpha))-$RDP for a given α.
- Apply the linear composition of RDP [3], i.e., the model is $(\alpha, T\gamma(\alpha))-$RDP after T iterations.
- Convert $(\alpha, T\gamma(\alpha))-$RDP into $(\epsilon(\alpha, \delta), \delta)-$DP for a given δ.

As compared to [19, Theorem 5], Theorem 1 considers a randomly sampled dataset due to the SGD and also considers a fixed but general query sensitivity $2C$. We first obtain the $\gamma_k(\alpha_k)$ parameter for the Gaussian mechanism of user k for an arbitrary integer α_k when the dataset is obtained by applying random sampling with probability q_k. We then obtain the noise variance σ_k^2 bound by using MA. To avoid confusion, we denote the RDP cost of [19, Theorem 5] by $\gamma_0(\alpha) = \frac{\alpha_k}{2\sigma_k^2}$ and that of Theorem 1 by $\gamma_k(\alpha_k)$. We obtain $\gamma_k(\alpha_k)$ by using the same method used in the proof of [3, Lemma 3].

Suppose we use a Gaussian mechanism $\mathcal{M}_k(\mathcal{D}_k) = \bar{\mathbf{g}}_k(\mathcal{D}_k) + \mathcal{N}(\mathbf{0}, C^2\sigma_k^2\mathbf{I}_d)$ for a given dataset $\mathcal{D}_k$ and the corresponding clipped gradient $\bar{\mathbf{g}}_k$. Consider a neighboring dataset $\mathcal{D}_k'$ to $\mathcal{D}_k$ that only differs by a single data D_n, i.e., $\mathcal{D}_k = \mathcal{D}_k' \cup \{D_n\}$ without loss of generality. Instead of considering the worst case that $\bar{\mathbf{g}}_k(\mathcal{D}_k) = -\bar{\mathbf{g}}_k(\mathcal{D}_k')$ and $\|\bar{\mathbf{g}}_k(\mathcal{D}_k)\| = C$, we assume $\bar{\mathbf{g}}_k(\mathcal{D}_k') = \mathbf{0}$ and $\bar{\mathbf{g}}_k(\mathcal{D}_k) = 2C\mathbf{e}_1$ and analyze the Rényi divergence of two perturbed gradients without loss of generality. This assumption makes the problem one-dimensional because the neighboring datasets $\mathcal{D}_k$ and $\mathcal{D}_k'$ result in the gradients that have different elements only in the first dimension. Let μ_0 denote the PDF of $\mathcal{N}(0, C^2\sigma_k^2)$ and μ_1 denote the PDF of $\mathcal{N}(2C, C^2\sigma_k^2)$. The noisy gradients of the Gaussian mechanism can be represented in one dimension as

$$\mathcal{M}_k(\mathcal{D}_k') \sim \mu_0 \tag{14}$$
$$\mathcal{M}_k(\mathcal{D}_k) \sim \mu \coloneqq (1-q_k)\mu_0 + q_k\mu_1. \tag{15}$$

To observe γ_k for given μ, μ_0, and an integer α_k, we start with the definition of RDP:

$$\frac{1}{\alpha_k}\mathbb{E}_{z\sim\mu}[(\mu(z)/\mu_0(z))^{\alpha_k}] \leq \gamma_k \tag{16}$$
$$\frac{1}{\alpha_k}\mathbb{E}_{z\sim\mu_0}[(\mu_0(z)/\mu(z))^{\alpha_k}] \leq \gamma_k. \tag{17}$$

Two inequalities can be shown by the same method, so we show only the second inequality here. Changing the probability that the expectation is taken over from μ_0 to μ and using the binomial expansion, we have

$$\begin{aligned}
&\mathbb{E}_{z\sim\mu_0}[(\mu_0(z)/\mu(z))^{\alpha_k}] = \mathbb{E}_{z\sim\mu}[(\mu_0(z)/\mu(z))^{\alpha_k+1}] \\
&= \mathbb{E}_{z\sim\mu}[(1+(\mu_0(z)-\mu(z))/\mu(z))^{\alpha_k+1}] \\
&= \sum_{i=0}^{\alpha_k+1}\binom{\alpha_k+1}{i}\mathbb{E}_{z\sim\mu}[((\mu_0(z)-\mu(z))/\mu(z))^i].
\end{aligned} \tag{18}$$

The first term of the summation coming from $i=0$ simply becomes 1. The second term when $i=1$ is 0 by simple calculus. The third term with $i=2$ can be bounded as

$$\begin{aligned}
&\mathbb{E}_{z\sim\mu}[((\mu_0(z)-\mu(z))/\mu(z))^2] \\
&= \mathbb{E}_{z\sim\mu}[((q_k\mu_0(z)-q_k\mu_1(z))/\mu(z))^2] \\
&= q_k^2\int_{-\infty}^{\infty}(\mu_0(z)-\mu_1(z))^2/\mu(z)dz \\
&\leq \frac{q_k^2}{1-q_k}\int_{-\infty}^{\infty}(\mu_0(z)-\mu_1(z))^2/\mu_0(z)dz \\
&= \frac{q_k^2}{1-q_k}\mathbb{E}_{z\sim\mu_0}[((\mu_0(z)-\mu_1(z))/\mu_0(z))^2].
\end{aligned} \tag{19}$$

The expected value of the above term can be further simplified and bounded as below:

$$\begin{aligned}
&\mathbb{E}_{z\sim\mu_0}[((\mu_0(z)-\mu_1(z))/\mu_0(z))^2] \\
&= \mathbb{E}_{z\sim\mu_0}\left[\left(1-\exp\left(\frac{-z^2+4Cz-4C^2+z^2}{2C^2\sigma_k^2}\right)\right)^2\right] \\
&= 1-2\mathbb{E}_{z\sim\mu_0}\left[\exp\left(\frac{4Cz-4C^2}{2C^2\sigma_k^2}\right)\right] \\
&\quad +\mathbb{E}_{z\sim\mu_0}\left[\exp\left(\frac{8Cz-8C^2}{2C^2\sigma_k^2}\right)\right] \\
&= 1-2+\exp\left(\frac{8C^2}{2C^2\sigma_k^2}\right) \\
&= \exp\left(\frac{4}{\sigma_k^2}\right)-1 = 4/\sigma_k^2+\mathcal{O}\left(\frac{1}{\sigma_k^4}\right).
\end{aligned} \tag{20}$$

The third term with $i=2$ eventually becomes

$$\begin{aligned}
&\binom{\alpha_k+1}{2}\mathbb{E}_{z\sim\mu}\left[\left(\frac{\mu_0(z)-\mu(z)}{\mu(z)}\right)^2\right] \\
&\leq \frac{4\alpha_k(\alpha_k+1)q_k^2}{2(1-q_k)\sigma_k^2}+\mathcal{O}\left(\frac{q_k^2\alpha_k^2}{\sigma_k^4}\right).
\end{aligned} \tag{21}$$

[3, Lemma 3] shows that the other terms, i.e., $(i \geq 3)$ terms, are upper bounded by $\mathcal{O}\left(\frac{q_k^3\alpha_k^3}{\sigma_k^3}\right)$. Thus, $\gamma_k(\alpha_k) = \frac{2q_k^2}{(1-q_k)\sigma_k^2}(\alpha_k+1)+\mathcal{O}(\frac{q_k^3\alpha_k^2}{\sigma_k^3})$. Note that using $\gamma_k(\alpha_k)$ with σ_k^2 is equivalent to using $\gamma_0(\alpha_k)$ with $\frac{(1-q_k)\alpha_k}{4q_k^2(\alpha_k+1)}\sigma_k^2$ because

$$\gamma_k(\alpha_k) = \frac{2q_k^2(\alpha_k+1)}{(1-q_k)\sigma_k^2} = \frac{\alpha_k}{2\left(\frac{(1-q_k)\alpha_k}{4q_k^2(\alpha_k+1)}\sigma_k^2\right)}. \tag{22}$$

In accordance with an assumption that α_k is optimally chosen to be $\alpha_k = 2\log(\delta_k^{-1})/\epsilon_k$ from [19, Theorem 5], we obtain $\alpha_k/(\alpha_k+1) \approx 1$ when δ_k is sufficiently small. We can obtain the following inequality by directly using [19, Theorem 5] for

the noise variance associated with $\gamma_k(\alpha_k)$:

$$\frac{1-q_k}{4q_k^2}\sigma_k^2 \approx \frac{(1-q_k)\alpha_k}{4q_k^2(\alpha_k+1)}\sigma_k^2 \geq \frac{2T}{\epsilon_k^2}\log\frac{1}{\delta_k}+\frac{T}{\epsilon_k}-\frac{2T}{\epsilon_k^2}\left(\log(2\log\delta_k^{-1})+1-\log\epsilon_k\right) + \mathcal{O}\left(\frac{\log^2(\log\delta_k^{-1})}{\log\delta_k^{-1}}\right). \quad (23)$$

Dividing both hand sides by the coefficient of σ_k^2 completes the proof of (10).

Due to the assumption that $\mathcal{L}$ is λ-strongly convex, for all $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$ and any subgradient $\mathbf{g}$ of $\mathcal{L}$ at $\mathbf{w}$ we have

$$\mathcal{L}(\mathbf{w}') - \mathcal{L}(\mathbf{w}) \geq \langle \mathbf{g}, \mathbf{w}' - \mathbf{w}\rangle + \frac{\lambda}{2}\left\|\mathbf{w}' - \mathbf{w}\right\|^2 \quad (24)$$

so that we obtain

$$\langle \mathbf{g}_k^{(t)}, \mathbf{w}^{(t)} - \mathbf{w}^*\rangle = -\langle \mathbf{g}_k^{(t)}, \mathbf{w}^* - \mathbf{w}^{(t)}\rangle \geq \mathcal{L}(\mathbf{w}^{(t)}) - \mathcal{L}(\mathbf{w}^*) + \frac{\lambda}{2}\left\|\mathbf{w}^{(t)} - \mathbf{w}^*\right\|^2 \geq 0 \quad (25)$$

and

$$\mathcal{L}(\mathbf{w}^{(t)}) - \mathcal{L}(\mathbf{w}^*) \geq \langle \mathbf{g}^*, \mathbf{w}^{(t)} - \mathbf{w}^*\rangle + \frac{\lambda}{2}\left\|\mathbf{w}^{(t)} - \mathbf{w}^*\right\|^2 \geq \frac{\lambda}{2}\left\|\mathbf{w}^{(t)} - \mathbf{w}^*\right\|^2, \quad (26)$$

where $\mathbf{g}^*$ is the gradient at the optimum point $\mathbf{w}^*$ calculated with an arbitrary subset $\mathcal{S}$ of $\mathcal{D}$. The μ-smoothness condition gives

$$\mathbb{E}[\mathcal{L}(\mathbf{w}^{(t)}) - \mathcal{L}(\mathbf{w}^*)] \leq \mathbb{E}\left[\frac{\mu}{2}\left\|\mathbf{w}^{(t)} - \mathbf{w}^*\right\|^2\right]. \quad (27)$$

We want to next observe how the weight vector $\mathbf{w}^{(t)}$ converges to its optimum $\mathbf{w}^*$ by bounding the expected mean square error $\mathbb{E}\left[\left\|\mathbf{w}^{(t)} - \mathbf{w}^*\right\|^2\right]$. We first obtain a recurrence formula for $\left\|\mathbf{w}^{(t)} - \mathbf{w}^*\right\|^2$ by using the weight update equation (3) with $\bar{\mathbf{g}}_k^{(t)}$ replaced by $\tilde{\mathbf{g}}_k^{(t)}$ as follows:

$$\left\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\right\|^2 = \left\|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta_t \cdot \sum_{k=1}^K \frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|}\tilde{\mathbf{g}}_k^{(t)}\right\|^2 = \left\|\mathbf{w}^{(t)} - \mathbf{w}^*\right\|^2 + \eta_t^2\left\|\sum_{k=1}^K \frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|}\tilde{\mathbf{g}}_k^{(t)}\right\|^2 - 2\eta_t\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \sum_{k=1}^K \frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|}\tilde{\mathbf{g}}_k^{(t)}\rangle. \quad (28)$$

First, the second term can be decomposed into

$$\eta_t^2\left\|\sum_{k=1}^K \frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|}\tilde{\mathbf{g}}_k^{(t)}\right\|^2 = \eta_t^2\left\|\sum_{k=1}^K \frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|}(\bar{\mathbf{g}}_k^{(t)} + \mathbf{Z}_k)\right\|^2 = \eta_t^2\left\|\sum_{k=1}^K \frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|}\bar{\mathbf{g}}_k^{(t)} + \mathbf{Z}\right\|^2 \quad (29)$$

where $\mathbf{Z} = \sum_{k=1}^K \frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|}\mathbf{Z}_k$. The square of the ℓ_2-norm can be expanded as

$$\eta_t^2\left\|\sum_{k=1}^K \frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|}\bar{\mathbf{g}}_k^{(t)}\right\|^2 + 2\eta_t^2\sum_{k=1}^K \frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|}\langle \bar{\mathbf{g}}^{(t)}, \mathbf{Z}\rangle + \eta_t^2\left\|\mathbf{Z}\right\|^2.$$

It is straightforward to show that $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, C^2\sigma^2\mathbf{I}_d)$ where $\sigma^2 = \frac{\sum_{k=1}^K (|\mathcal{D}_k| q_k \sigma_k)^2}{(\sum_{k=1}^K |\mathcal{D}_k| q_k)^2}$. The expected value of the second term can be obtained as follows.

$$\mathbb{E}\left[\eta_t^2\left\|\sum_{k=1}^K \frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|}\bar{\mathbf{g}}_k^{(t)}\right\|^2\right] + \mathbb{E}\left[2\eta_t^2\sum_{k=1}^K \frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|}\langle \bar{\mathbf{g}}, \mathbf{Z}\rangle\right]^{0} + \mathbb{E}[\eta_t^2\left\|\mathbf{Z}\right\|^2] \leq \eta_t^2 C^2 + d\eta_t^2 C^2\sigma^2 = \eta_t^2 C^2(1 + d\sigma^2) \quad (30)$$

which follows because every element of $\mathbf{Z}$ is a zero-mean Gaussian, and taking the inner product is a linear transformation. Next, the third term of (28) can be decomposed into the inner products that can be locally calculated as

$$-2\eta_t\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \sum_{k=1}^K \frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|}\tilde{\mathbf{g}}_k^{(t)}\rangle = -2\eta_t\sum_{k=1}^K \frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|}\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \tilde{\mathbf{g}}_k^{(t)}\rangle. \quad (31)$$

We bound the expected value of each summand in (31) as

$$\mathbb{E}\left[\frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|}\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \tilde{\mathbf{g}}_k^{(t)}\rangle\right] = \mathbb{E}\left[\frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|}\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \bar{\mathbf{g}}_k^{(t)}\rangle\right] + \mathbb{E}\left[\frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|}\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{Z}_k\rangle\right]^{0} \geq \mathbb{E}\left[\frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|}\frac{C}{G}\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{g}_k^{(t)}\rangle\right]. \quad (32)$$

And then we bound the expected value of (31) as follows:

$$
\begin{aligned}
&-2\eta_t\mathbb{E}\left[\langle \mathbf{w}^{(t)}-\mathbf{w}^*, \sum_{k=1}^{K}\frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|}\tilde{\mathbf{g}}_k^{(t)}\rangle\right]\\
&\le -2\eta_t\sum_{k=1}^{K}\mathbb{E}\left[\frac{|\mathcal{J}_k^{(t)}|}{|\mathcal{J}^{(t)}|}\frac{C}{G}\langle \mathbf{w}^{(t)}-\mathbf{w}^*, \mathbf{g}_k^{(t)}\rangle\right]\\
&\overset{(a)}{\le} -2\eta_t\frac{C}{G}\mathbb{E}\left[\mathcal{L}(\mathbf{w}^{(t)})-\mathcal{L}(\mathbf{w}^*)+\frac{\lambda}{2}\left\|\mathbf{w}^{(t)}-\mathbf{w}^*\right\|^2\right]\\
&\overset{(b)}{\le} -2\eta_t\frac{C}{G}\mathbb{E}\left[\lambda\left\|\mathbf{w}^{(t)}-\mathbf{w}^*\right\|^2\right] \qquad (33)
\end{aligned}
$$

where (a) follows by (25) and (b) follows by (26). Using (30) and (33), we can bound the expected value of (28) as

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{w}^{(t+1)}-\mathbf{w}^*\right\|^2\right]\\
&\le \mathbb{E}\left[\left\|\mathbf{w}^{(t)}-\mathbf{w}^*\right\|^2\right]+\eta_t^2C^2(1+d\sigma^2)\\
&\qquad -2\eta_t\frac{C}{G}\lambda\mathbb{E}\left[\left\|\mathbf{w}^{(t)}-\mathbf{w}^*\right\|^2\right]\\
&=(1-2\eta_t\frac{C}{G}\lambda)\mathbb{E}\left[\left\|\mathbf{w}^{(t)}-\mathbf{w}^*\right\|^2\right]+\eta_t^2C^2(1+d\sigma^2). \qquad (34)
\end{aligned}
$$

Assume a learning rate of $\eta_t=\frac{G}{C\lambda t}$, so we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{w}^{(t+1)}-\mathbf{w}^*\right\|^2\right]\\
&\le (1-\frac{2}{t})\mathbb{E}\left[\left\|\mathbf{w}^{(t)}-\mathbf{w}^*\right\|^2\right]+\frac{G^2(1+d\sigma^2)}{\lambda^2t^2}. \qquad (35)
\end{aligned}
$$

One can infer the following explicit bound for each term for some $a_1>0$ from the above recurrence relation:

$$
\mathbb{E}\left[\left\|\mathbf{w}^{(t)}-\mathbf{w}^*\right\|^2\right]\le a_1\frac{G^2}{\lambda^2t}. \qquad (36)
$$

According to [21, Lemma 2], the first term $\mathbb{E}\left[\left\|\mathbf{w}^{(1)}-\mathbf{w}^*\right\|^2\right]$ can be bounded as $\mathbb{E}\left[\left\|\mathbf{w}^{(1)}-\mathbf{w}^*\right\|^2\right]\le\frac{4G^2}{\lambda^2}$. Thus, it is necessary to have $a_1\ge 4$. The next term with $t=2$ can be bounded by using the recurrence relation and the non-negativity of the mean square error (MSE) as

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{w}^{(2)}-\mathbf{w}^*\right\|^2\right]\le -\mathbb{E}\left[\left\|\mathbf{w}^{(1)}-\mathbf{w}^*\right\|^2\right]+\frac{G^2(1+d\sigma^2)}{\lambda^2}\\
&\le\frac{2G^2(1+d\sigma^2)}{2\lambda^2}. \qquad (37)
\end{aligned}
$$

To include this term, we should satisfy $a_1\ge 2(1+d\sigma^2)$. Similarly, the next term with $t=3$ can be bounded as

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{w}^{(3)}-\mathbf{w}^*\right\|^2\right]\le\frac{G^2(1+d\sigma^2)}{4\lambda^2}\\
&\le\frac{\frac{3}{4}G^2(1+d\sigma^2)}{3\lambda^2}. \qquad (38)
\end{aligned}
$$

Thus, we should satisfy also $a_1\ge\frac{3}{4}(1+d\sigma^2)$. For all terms with $t\ge 3$ the coefficient $(1-2/t)$ is always positive, so the bound can be shown by mathematical induction. Assume that (36) is true for $t=\tau\ge 3$. By the recurrence formula, we can show that (36) also holds for $t=\tau+1$ as follows.

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{w}^{(\tau+1)}-\mathbf{w}^*\right\|^2\right]\le(1-\frac{2}{\tau})a_1\frac{G^2}{\lambda^2\tau}+\frac{G^2(1+d\sigma^2)}{\lambda^2\tau^2}\\
&=a_1\frac{G^2}{\lambda^2(\tau+1)}\frac{1}{\tau^2}((\tau+1)(\tau-2)+\frac{\tau+1}{a_1}(1+d\sigma^2))\\
&=a_1\frac{G^2}{\lambda^2(\tau+1)}(1-\frac{1}{\tau}(1-\frac{1+d\sigma^2}{a_1})-\frac{1}{\tau^2}(2-\frac{1+d\sigma^2}{a_1}))\\
&\overset{(a)}{\le} a_1\frac{G^2}{\lambda^2(\tau+1)} \qquad (39)
\end{aligned}
$$

where (a) holds if $a_1>1+d\sigma^2$. Thus, (36) holds for every time instance if $a_1\ge\max\{4,2(1+d\sigma^2)\}$, i.e., we have

$$
\mathbb{E}\left[\left\|\mathbf{w}^{(t)}-\mathbf{w}^*\right\|^2\right]\le\max\{2,1+d\sigma^2\}\frac{2G^2}{\lambda^2t}. \qquad (40)
$$

Using this bound, a lower bound of the convergence rate can be obtained with the μ-smoothness condition (27) as follows:

$$
\begin{aligned}
&\mathbb{E}[\mathcal{L}(\mathbf{w}^{(t)})-\mathcal{L}(\mathbf{w}^*)]\le\frac{\mu}{2}\mathbb{E}\left[\left\|\mathbf{w}^{(t)}-\mathbf{w}^*\right\|^2\right] \qquad (41)\\
&\le\max\{2,1+d\sigma^2\}\frac{\mu G^2}{\lambda^2t}. \qquad (42)
\end{aligned}
$$

The lower bound of (11) is obtained by taking the inverse of the convergence rate bound in each case.

Consider next the transmission rate bound, for which we denote the noisy gradient as a sum of the clipped gradient and noise to exploit their boundedness and statistical properties, respectively, as follows:

$$
h(\tilde{\mathbf{g}}_k^{(t)})=h(\bar{\mathbf{g}}_k^{(t)}+\mathbf{Z}_k)\le h(\bar{\mathbf{g}}_k^{(t)})+h(\mathbf{Z}_k) \qquad (43)
$$

such that

$$
h(\bar{\mathbf{g}}_k^{(t)})\le\frac{1}{2}\log\det(2\pi e\mathbf{K})=\frac{1}{2}\log\left((2\pi e)^d\det(\mathbf{K})\right) \qquad (44)
$$

where $\mathbf{K}$ is the covariance matrix of $\bar{\mathbf{g}}_k^{(t)}$ and $\det(\cdot)$ represents the determinant of a matrix. Denote the element of $\mathbf{K}$ in i-th row and j-th column as $K_{i,j}$ for all $i,j=1,2,\ldots,d$. The determinant of the covariance matrix can be bounded as

$$
\det(\mathbf{K})\overset{(a)}{\le}\prod_{i=1}^{d}K_{i,i}=\prod_{i=1}^{d}\mathrm{Var}(\bar{g}_{k,i}^{(t)})\le\prod_{i=1}^{d}\mathbb{E}[(\bar{g}_{k,i}^{(t)})^2] \qquad (45)
$$

where (a) follows from the Hadamard's inequality since a covariance matrix is positive-semidefinite. We can further bound $\det(\mathbf{K})$ by using the boundedness of the ℓ_2-norm of the clipped gradient, i.e., $\sum_{i=1}^{d}\mathbb{E}[(\bar{g}_{k,i}^{(t)})^2]=\mathbb{E}\big[\sum_{i=1}^{d}(\bar{g}_{k,i}^{(t)})^2\big]\le$

$\mathbb{E}[C^2] = C^2$. Using the inequality of arithmetic and geometric means, the right hand side of (45) can be bounded as

$$\prod_{i=1}^{d} \mathbb{E}\big[(\bar{g}_{k,i}^{(t)})^2\big] \leq \left(\frac{1}{d}\sum_{i=1}^{d} \mathbb{E}\big[(\bar{g}_{k,i}^{(t)})^2\big]\right)^d \leq \left(\frac{C^2}{d}\right)^d. \quad (46)$$

Combining (44)-(46), we obtain

$$h(\bar{\mathbf{g}}_k^{(t)}) \leq \frac{d}{2}\log\Big(\frac{2\pi e C^2}{d}\Big). \quad (47)$$

Furthermore, the differential entropy of the noise $\mathbf{Z}_k \sim \mathcal{N}(\mathbf{0}, C^2\sigma_k^2\mathbf{I})$ is

$$h(\mathbf{Z}_k) = \frac{1}{2}\log\det(2\pi e C^2\sigma_k^2\mathbf{I}_d) = \frac{d}{2}\log(2\pi e C^2\sigma_k^2). \quad (48)$$

Combining (43), (47), and (48), the proof of (12) follows.