SENSITIVITY-AWARE DIFFERENTIALLY PRIVATE DECENTRALIZED LEARNING WITH ADAPTIVE NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

Most existing decentralized learning methods with differential privacy (DP) employ fixed-level Gaussian noise during training, regardless of gradient convergence, which compromises model accuracy without providing additional privacy benefits. In this paper, we propose a novel Differentially Private Decentralized learning approach, termed AdaD^{2}P, which employs Adaptive noise leveraging the real-time estimation of sensitivity for local updates based on gradient norms and works for time-varying communication topologies. Compared with existing solutions, the integration of adaptive noise enables us to enhance model accuracy while preserving the (ϵ, δ)-DP privacy budget. We prove that AdaD^{2}P achieves a utility bound of $\mathcal{O}\left(\sqrt{d\log\left(\frac{1}{\delta}\right)}/(\sqrt{n}J\epsilon)\right)$, where J and n are the number of local samples and nodes, respectively, and d the dimension of decision variable; this bound matches that of the distributed counterparts with server-client structures, without relying on the stringent bounded gradient assumption commonly used in previous works. Theoretical analysis reveals the inherent advantages of AdaD^{2}P employing adaptive noise as opposed to constant noise. Extensive experiments on two benchmark datasets demonstrate the superiority of AdaD^{2}P over its counterparts, especially under a strong level of privacy guarantee.

1 INTRODUCTION

Distributed learning has recently attracted significant attention due to its great potential in enhancing computing efficiency and has thus been widely adopted in various application domains (Langer et al., 2020). In particular, it can be typically modeled as a non-convex finite-sum optimization problem solved by a group of n nodes, as depicted as follows:

$$\min_{x\in\mathbb{R}^d} f(x) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad \text{where } f_i(x) = \frac{1}{J}\sum_{j=1}^{J} f_i(x;j), \tag{1}$$

where J denotes the local dataset size of each node, $f_i(x;j)$ denotes the loss function of the j-th data sample at node i with respect to the model parameter $x \in \mathbb{R}^d$, and $f_i(x)$ and $f(x)$ denote the local objective function at node i and the global objective function. All nodes collaborate to seek the optimal model parameter to minimize $f(x)$, and each node i can only evaluate local stochastic gradient $\nabla f_i(x;\xi_i)$ where $\xi_i \in \{1, 2, ..., J\}$.

Bottlenecks such as high communication overhead and the vulnerability of central nodes in parameter server-based methods (Li et al., 2014; Zinkevich et al., 2010; McMahan et al., 2017a), motivate researchers to investigate fully decentralized methods (Lian et al., 2017; Tang et al., 2018; Lian et al., 2018) to solve Problem (1), where the central node is not required and each node only communicates with its neighbors. The existing decentralized learning algorithms usually employ undirected graphs for communication, which can not be easily implemented due to the existence of deadlocks (Assran et al., 2019). It is desirable to consider more practical scenarios where communication graphs may be directed and even time-varying. Stochastic gradient push (SGP) proposed in (Assran et al., 2019), which builds on push-sum protocol (Kempe et al., 2003), is proven to be very effective in solving (1) over directed and time-varying communication graphs.

In decentralized learning systems, all nodes frequently exchange information such as model parameters with their neighbors. This raises significant concerns about privacy, as the exposure of

intermediate parameters could potentially be exploited to compromise the privacy of original data samples (Wang et al., 2019b). To safeguard each node from potential data privacy attack, differential privacy (DP), as a theoretical tool to provide rigorous privacy guarantees and quantify privacy loss, can be integrated into each node within decentralized learning systems to enhance privacy protection. Most existing decentralized learning algorithms with differential privacy guarantee for non-convex problems tend to either assume stochastic gradients are bounded by some constant G (Yu et al., 2021; Xu et al., 2021) or employ gradient clipping strategy with a fixed clipping bound C (Li & Chi, 2023), and they use constant G or C to estimate the l_2 sensitivity S of gradient update across all iterations. As a result, each node injects fixed-level DP Gaussian noises with a variance proportional to the estimated sensitivity S before performing local SGD at each iteration. However, our empirical observations indicate that the norm of gradient typically decay as training progresses and ultimately converges to a small value (c.f., Figure 1). This observation suggests that the aforementioned methods estimating l_2 sensitivity using constant G or C for all iterations may be conservative as gradient norms are often smaller than the constant G or C, especially in the later stage of training. Therefore, their added fixed-level Gaussian noise deems unnecessary and will, instead, degrade the model accuracy without providing additional privacy gain. To this end, the following question arises naturally:

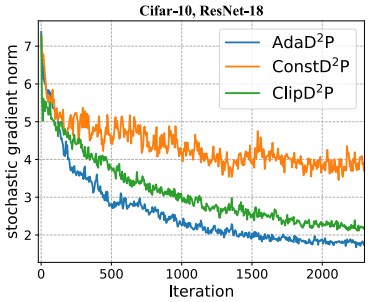


Figure 1: The evolution of gradient norm w.r.t the number of iterations for the proposed AdaD^{2}P and other two strategies (ConstD^{2}P and ClipD^{2}P) with constant noise.

"Can we design a decentralized learning method that adjusts the level of DP noise according to gradient norms during training while maintaining the privacy guarantee?"

To address this question, we develop a new differentially private learning method for non-convex problems in fully decentralized settings, which can adapt the noise level to the actual privacy requirements as the training progresses and thus enhance model accuracy given the same privacy budget. The key contributions are summarized as follows:

- **New efficient algorithm with adaptive DP noise.** We propose a differentially private decentralized learning method with adaptive DP noise (termed AdaD^{2}P) for non-convex problems, which works for time-varying directed communication topologies. In particular, each node adds noise with a variance calculated according to the noise scale and the sensitivity estimated based on real-time gradient norms. This adaptive mechanism allows adding smaller noise and thus enhancing model accuracy without compromising privacy budgets; importantly, it can be readily integrated into other existing decentralized algorithms.
- **Theoretical analysis and utility guarantees.** We prove that AdaD^{2}P achieves a utility bound of $\mathcal{O}\left(\sqrt{d\log\left(\frac{1}{\delta}\right)}/(\sqrt{n}J\epsilon)\right)$, which matches that of existing distributed methods with server-client structures (c.f., Table 1). Our proof involves constructing an intricate loop among the utility gap captured by the running average squared gradient norm, consensus error and error terms arising from injected DP noise; importantly, the proof *does not rely on the restrictive bounded gradient assumption* as commonly used by the previous works. Besides, we provide theoretical evidence that sheds light on the inherent advantages of AdaD^{2}P employing adaptive noise compared to that with fixed-level noise.
- **Extensive experimental evaluations.** Extensive experiments conducted on training ResNet-18 DNN (resp. 2-layer neural network) on the Cifar-10 (resp. Mnist) dataset in fully decentralized setting show that, when adhering to a same privacy budget constraint, our proposed AdaD^{2}P achieves superior model accuracy compared to its counterparts that employ fixed-level Gaussian noise, particularly in the strong privacy protection region.

2 Preliminary and Related Work

Differential privacy. Differential privacy (DP) was originally introduced in the seminal work by Dwork et al.(Dwork et al., 2006) as a foundational concept for quantifying the privacy-preserving

Algorithm	Privacy	Utility	Architecture	Without Assumption 6
DP-SGD (Abadi et al., 2016)	(ϵ, δ)-DP	$\frac{\sqrt{d\log\left(\frac{1}{\delta}\right)}}{J\epsilon}$	single node centralized	✗
Distributed DP-SRM[1] (Wang et al., 2019a)	(ϵ, δ)-DP global	$\frac{\sqrt{d\log\left(\frac{1}{\delta}\right)}}{nJ\epsilon}$	n nodes server-client	✗
LDP SVRG/SPIDER (Lowy et al., 2023)	(ϵ, δ)-DP for each node	$\frac{\sqrt{d\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}J\epsilon}$	n nodes server-client	✗
SoteriaFL-SAGA/SVRG (Li et al., 2022)	(ϵ, δ)-DP for each node	$\frac{\sqrt{(1+\omega)d\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}J\epsilon}$	n nodes server-client	✗
AdaD^{2}P (Algorithm 1)	(ϵ, δ)-DP for each node	$\frac{\sqrt{d\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}J\epsilon}$	n nodes decentralized	✓

[1] The global (ϵ, δ)-DP is considered therein, which only protects the privacy for the entire dataset while we consider (ϵ, δ)-DP for each node i, protecting the local dataset at the node's level.

Table 1: Comparison of existing differentially private stochastic algorithms for non-convex problems. Communication compression is employed in SoteriaFL-SAGA/SVRG with ω being the compression parameter. The Big $\mathcal{O}$ notation is omitted for simplicity.

capabilities of randomized algorithms. DP has now found widespread applications in a variety of domains that necessitate safeguarding against unintended information leakage, such as principle component analysis (Ge et al., 2018), meta learning (Li et al., 2019a), personalized recommendation (Shin et al., 2018), empirical risk minimization (Chaudhuri et al., 2011) and wireless network (Wei et al., 2021b). The standard definition of DP is provided as follows.

Definition 1 ((ϵ, δ)-DP (Dwork et al., 2014))**.** *A randomized mechanism $\mathcal{M}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies (ϵ, δ)-differential privacy (or (ϵ, δ)-DP), if for any two adjacent inputs $\mathrm{x}, \mathrm{x}' \in \mathcal{D}$ differing on a single entry and for any subset of outputs $S \subseteq \mathcal{R}$, it holds that*

$$Pr\left[\mathcal{M}\left(\mathrm{x}\right) \in S\right] \leqslant e^{\epsilon} Pr\left[\mathcal{M}\left(\mathrm{x}'\right) \in S\right] + \delta, \tag{2}$$

where the privacy budget ϵ denotes the privacy lower bound to measure a randomized query and δ is the probability of breaking this bound.

A commonly employed technique to ensure a (ϵ, δ)-differential privacy guarantee is through the use of the Gaussian mechanism as provided below.

Proposition 1 (Gaussian mechanism (Dwork et al., 2014))**.** *Let $f : \mathcal{D} \to \mathbb{R}$ be a real-valued function with S being f's l_2 sensitivity. Then, adding Gaussian noise $\mathcal{N}(0, \sigma^2 S^2)$ to f such that $\mathcal{M}(\mathrm{x}) = f(\mathrm{x}) + \mathcal{N}(0, \sigma^2 S^2)$ satisfies (ϵ, δ)-DP if the noise scale $\sigma \geqslant \frac{\sqrt{2\log\left(1.25/\delta\right)}}{\epsilon}$.*

The above proposition illustrates an inverse relationship between the noise scale σ and privacy budget ϵ for a fixed δ, and the fact that the noise variance $\sigma^2 S^2$ is dependent on both the noise scale σ and l_2 sensitivity S. For iterative training processes, the cumulative privacy spending can be calculated using the basic composition theorem (Dwork et al., 2006; Dwork & Lei, 2009) and advanced composition theorem (Dwork et al., 2010; Bun & Steinke, 2016). To achieve a more precise estimate of the overall privacy budget throughout the entire training process, Abadi et al. (2016) introduced the moments accountant method that tracks higher moments. In the rest of this section, we will review existing research works related to achieving differential privacy guarantees in machine learning and highlight their limitations inherent in decentralized scenarios.

Decentralized learning methods with privacy guarantee. DP guarantee is initially integrated to centralized (single-node) setting for designing differentially private stochastic learning algorithms (Abadi et al., 2016; Wang et al., 2017; Iyengar et al., 2019; Chen et al., 2020; Wang et al., 2020). Further, DP guarantee is considered in distributed learning with server-client structures and the representative works include (McMahan et al., 2017b; Li et al., 2019b; Wang et al., 2019a; Wu

et al., 2020; Zhang et al., 2020; Wei et al., 2020; Zeng et al., 2021; Wei et al., 2021a; Ding et al., 2021; Li et al., 2022; Liu et al., 2022; Lowy et al., 2023; Wang et al., 2023; Zhou et al., 2023; Wei et al., 2023). Recently, there have been few works aiming to achieve DP guarantees for fully decentralized learning algorithms. For example, Cheng et al. (2018; 2019) achieve DP in fully decentralized learning for only strongly convex problems. Wang & Nedic (2022) achieve DP in fully decentralized architectures by tailoring gradient methods for deterministic optimization problems. For non-convex stochastic optimization problems as we consider in this work, Yu et al. (2021) present a differentially private decentralized learning method (DP^2-SGD) based on D-PSGD (Lian et al., 2017), which relies on a fixed communication topology and uses the basic composition theorem to bound the overall privacy loss. To have a tight privacy guarantee, Xu et al. (2021) propose a differentially private asynchronous decentralized learning method (A(DP)^{2}SGD) based on AD-PSGD (Lian et al., 2018), which provides privacy guarantee in the sense of Rényi differential privacy (RDP) (Mironov, 2017). However, it should be noted that the aforementioned two algorithms (Yu et al., 2021; Xu et al., 2021) work only for undirected communication graphs which is often not satisfied in practical scenarios, and they rely on the bounded gradient assumption. Most recently, Li & Chi (2023) achieve DP guarantee in decentralized learning for non-convex problems without bounded gradient assumption by employing gradient clipping strategy with a fixed clipping bound C, while their method is only applicable to time-invariant communication topologies.

Learning with Adaptive DP Gaussian noise level. For the aforementioned differentially private decentralized methods designed for non-convex stochastic optimization problems (Yu et al., 2021; Xu et al., 2021; Li & Chi, 2023), the injected noise level may exceeds what is actually needed for privacy requirements as training progresses, especially during the later stages of training, since their estimated sensitivity based on fixed G (Yu et al., 2021; Xu et al., 2021) or C (Li & Chi, 2023) may not reflect the actual value of sensitivity. The overestimate of sensitivity may, indeed, lead to a waste of unnecessary privacy budget during training process (Wei et al., 2023). There has been few works dedicated to precisely estimate the sensitivity in a real-time manner. For instance, a scheme of decaying gradient clipping bound has been employed to estimate the sensitivity in differentially private centralized learning (Du et al., 2021; Wei & Liu, 2021), yielding decreasing amount of noise injection. In the realm of distributed learning, the similar strategy of adaptive clipping bounds are utilized in (Andrew et al., 2021) to estimate the sensitivity. Most recently, Wei et al. (2023) use the minimum value of properly decaying clipping bound and current gradient norm to more accurately estimate the l_2 sensitivity, leading to a less amount of noise injection. However, these distributed methods (Andrew et al., 2021; Fu et al., 2022; Wei et al., 2023) only focus on the server-client architecture and no theoretical guarantee on model utility is provided therein. In contrast, we aim to design a differentially private decentralized learning method which incorporates adaptive noise levels in fully distributed settings and provide a rigorous theoretical utility guarantee.

3 Proposed Algorithm

We consider solving Problem (1) over the following general network model.

Network Model. The communication topology is modeled as a sequence of time-varying directed graph $\mathcal{G}^k = \left(\mathcal{V}, \mathcal{E}^k\right)$, where $\mathcal{V} = \{1, 2, ..., n\}$ denotes the set of nodes and $\mathcal{E}^k \subset \mathcal{V} \times \mathcal{V}$ denotes the set of directed edges/links at iteration k. We associate each graph $\mathcal{G}^k$ with a non-negative mixing matrix $P^k \in \mathbb{R}^{n\times n}$ such that $(i, j) \in \mathcal{E}^k$ if $P^k_{i,j} > 0$, i.e., node i receiving a message from node j at iteration k. Without loss of generality, we assume that each node is an in-neighbor of itself.

The following assumptions are made on the mixing matrix and graph for the above network model to facilitate the subsequent utility analysis for the proposed algorithm.

Assumption 1 (Stochasicity of Mixing Matrix)**.** *The non-negative mixing matrix $P^k, \forall k$ is column-stochastic, i.e., $\mathbf{1}^\top P^k = \mathbf{1}^\top$, where $\mathbf{1}$ is a vector with all of its elements equal to* 1.

Assumption 2 (B-strongly Connected)**.** *There exists finite, positive integers B and $\triangle$, such that the graph with edge set $\bigcup_{k=lB}^{(l+1)B-1} \mathcal{E}^k$ is strongly connected and has diameter at most $\triangle$ for $\forall l \geqslant 0$.*

Algorithm Development. Now we present our proposed AdaD^{2}P, a novel differentially private decentralized stochastic learning algorithm for non-convex problems with adaptive DP Gaussian noise level, which can work over general time-varying directed communication topologies; the complete pseudocode is summarized in Algorithm 1. At a high level, AdaD^{2}P is comprised of local SGD

and the averaging of neighboring information, following a framework similar to SGP (Assran et al., 2019). This framework involves the use of the Push-Sum protocol (Kempe et al., 2003), which can tackle the unblanceness of directed topologies by asymptotically estimating the Perron–Frobenius eigenvector of transition matrices. However, the key distinction lies in the injection of adaptive DP Gaussian noise before preforming local SGD. In particular, each node i maintains three variables during the learning process: i) the model parameter x_i^k; ii) the scalar Push-Sum weight w_i^k and iii) the de-biased parameter $z_i^k = x_i^k / w_i^k$, with the initialization of $x_i^0 = z_i^0 \in \mathbb{R}^d$ and $w_i^0 = 1$ for all nodes $i \in \mathcal{V}$. At each iteration k, each node i updates as follows:

$$\underbrace{x_i^{k+\frac{1}{2}} = x_i^k - \gamma \left(\nabla f_i(z_i^k; \xi_i^k) + N_i^k\right)}_{\textbf{Differentially private local SGD}}, \quad \underbrace{x_i^{k+1} = \sum_{j=1}^n P_{i,j}^k x_j^{k+\frac{1}{2}}, w_i^{k+1} = \sum_{j=1}^n P_{i,j}^k w_j^k}_{\textbf{Neighboring information averaging}}, \underbrace{z_i^{k+1} = \frac{x_i^{k+1}}{w_i^{k+1}}}_{\textbf{De-bias}},$$

where $\gamma > 0$ is the step size and $\nabla f_i(z_i^k; \xi_i^k)$ is the gradient evaluated on the de-biased parameter z_i^k and training sample with index ξ_i^k at node i. The injected randomized noise N_i^k ensuring differential privacy guarantee for node i is drawn from the Gaussian distribution (c.f., (3)) with a variance calculated according to the noise scale σ and dynamic sensitivity estimated based on gradient norms It should be noted that gradient norm is a tighter estimation of actual sensitivity for noise injection than fixed G and C in most cases, especially at the later stage of training (Wei et al., 2023).

Algorithm 1 Differentially Private Decentralized Learning with Adaptive Noise (**AdaD^{2}P**)

1: **Initialization:** $x_i^0 = z_i^0 \in \mathbb{R}^d$, $w_i^0 = 1$, step size $\gamma > 0$, total number of iterations K and privacy budget (ϵ, δ).
2: **for** $k = 0, 1, 2, ..., K-1$, at node i, **do**
3: Randomly samples a local training data ξ_i^k with the sampling probability $\frac{1}{J}$;
4: Computes stochastic gradient at z_i^k: $\nabla f_i(z_i^k; \xi_i^k)$;
5: Draws randomized noise N_i^k from the Gaussian distribution

$$N_i^k \sim \mathcal{N}\left(0, \sigma^2 \left\|\nabla f_i\left(z_i^k; \xi_i^k\right)\right\|^2 \mathbb{I}_d\right), \tag{3}$$

where the noise scale σ is defined in Proposition 2;
6: Differentially private local SGD:

$$x_i^{k+\frac{1}{2}} = x_i^k - \gamma(\nabla f_i(z_i^k; \xi_i^k) + N_i^k); \tag{4}$$

7: Sends $\left(x_i^{k+\frac{1}{2}}, w_i^k\right)$ to all out-neighbors ;
8: Receives $\left(x_j^{k+\frac{1}{2}}, w_j^k\right)$ from all in-neighbors ;
9: Updates x_i^{k+1} by: $x_i^{k+1} = \sum_{j=1}^n P_{i,j}^k x_j^{k+\frac{1}{2}}$;
10: Updates w_i^{k+1} by: $w_i^{k+1} = \sum_{j=1}^n P_{i,j}^k w_j^k$;
11: Updates z_i^{k+1} by: $z_i^{k+1} = x_i^{k+1} / w_i^{k+1}$.
12: **end for**

Remark 1. *For comparison, we also present two counterparts ConstD^{2}P and ClipD^{2}P, which employ fixed-level noise with variance calculated according to fixed l_2 sensitivity, estimated using uniform gradient bound G and fixed gradient clipping bound C respectively. The complete pseudocodes of ConstD^{2}P and ClipD^{2}P can be found in Algorithm 2 and 3 in the appendix, respectively.*

4 Theoretical Guarantees

In this section, we provide the privacy and utility guarantee for our proposed AdaD^{2}P. In particular, we show that DP guarantee for each node can be achieved by setting the DP Gaussian noise scale σ properly according to the given certain privacy budget (ϵ, δ) and the total number of iterations K, which is summarized in the following proposition.

Proposition 2 (Privacy guarantee). *There exist constants c_1 and c_2 such that, for any $\epsilon < \frac{c_1 K}{J^2}$ and $\delta \in (0,1)$, (ϵ, δ)-DP can be guaranteed for each node i for AdaD^{2}P, ConstD^{2}P and ClipD^{2}P after K iterations if we set the noise scale*

$$\sigma = \frac{c_2 \sqrt{K \log\left(\frac{1}{\delta}\right)}}{J\epsilon}. \tag{5}$$

Proof. The proof of the above result can be easily adapted from Theorem 1 in (Abadi et al., 2016) by knowing the fact that the sampling probability is $\frac{1}{J}$ for each node i at each iteration. □

Remark 2. *The above theorem demonstrates that the variance of injected Gaussian noise for each node i at each iteration k for AdaD^{2}P is*

$$\mathbb{E}\left[\left\|N_i^k\right\|^2\right] \overset{(3)}{=} d\sigma^2 \left\|\nabla f_i\left(z_i^k; \xi_i^k\right)\right\|^2 \overset{(5)}{=} \underbrace{\frac{d c_2^2 \log\left(\frac{1}{\delta}\right)}{J^2 \epsilon^2}}_{\beta} \cdot K \left\|\nabla f_i\left(z_i^k; \xi_i^k\right)\right\|^2, \tag{6}$$

which is proportional to the real-time gradient norm $\left\|\nabla f_i\left(z_i^k; \xi_i^k\right)\right\|$.

Next, we make the following blanket assumptions for the utility analysis of AdaD^{2}P.

Assumption 3 (L-smooth). *For each function $f_i, i \in \mathcal{V}$, there exists a constant $L > 0$ such that $\|\nabla f_i(x) - \nabla f_i(y)\| \leqslant L \|x - y\|$.*

Assumption 4 (Unbiased gradient). *For $\forall x \in \mathbb{R}^d$, the expectation of stochastic gradients of node i is its aggregated gradient, i.e.,*

$$\mathbb{E}\left[\nabla f_i(x; \xi_i)\right] = \nabla f_i(x). \tag{7}$$

Assumption 5 (Bounded variance). *There exist finite positive constants ζ^2 and b^2 such that for any node i and $\forall x \in \mathbb{R}^d$,*

$$\mathbb{E}\left[\|\nabla f_i(x; \xi_i) - \nabla f_i(x)\|\right] \leqslant \zeta^2 \tag{8}$$

and

$$\|\nabla f_i(x) - \nabla f(x)\|^2 \leqslant b^2. \tag{9}$$

With the above assumptions, by properly choosing the total number of iterations K and the step size γ, we can obtain the utility guarantee of AdaD^{2}P (Algorithm 1) without relying on the bounded gradient assumption, which is presented in the following Theorem 1.

Theorem 1 (Utility guarantee). *Suppose Assumptions 1-5 hold and $J \geqslant n^{\frac{3}{2}} c_2 \sqrt{d \log\left(\frac{1}{\delta}\right)}/\epsilon$ for a given privacy budget (ϵ, δ). There exist constants C and $q \in [0,1)$, which depend on the diameter of the network $\triangle$ and the sequence of mixing matrices P^k, such that, if we set $\gamma = 1/(\frac{J\epsilon}{c_2\sqrt{nd\log\left(\frac{1}{\delta}\right)}} + \hat{\gamma}(C,q)^{-1})$, $K = \frac{J^2\epsilon^2}{dc_2^2 \log\left(\frac{1}{\delta}\right)}$ and the noise scale $\sigma = c_2 \sqrt{K \log\left(\frac{1}{\delta}\right)}/(J\epsilon)$, AdaD^{2}P can achieve (ϵ, δ)-DP guarantee for each node and has the following utility bound*

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right] \leqslant \mathcal{O}\left(\frac{\sqrt{d \log\left(\frac{1}{\delta}\right)}}{\sqrt{n} J \epsilon}\right), \tag{10}$$

where C and q can be found in Lemma 4 and the definition of constant $\hat{\gamma}(C,q)$ can be found at (46) in the appendix, respectively. The Big $\mathcal{O}$ notation hides all constants involved in our setting, e.g., $L, \zeta, b, C, q, \sum_{i=1}^n \left\|x_i^0\right\|^2$ and $f\left(\bar{x}^0\right) - f^*$, where $f^* = \min_{x \in \mathbb{R}^d} f(x)$.

Proof. The complete proof can be found in Section A.3 in the appendix. □

Remark 3. *Table 1 provides a detailed comparison of our AdaD^{2}P with existing centralized/server-client algorithms, where the bounded gradient assumption is all assumed in their utility analysis, except our AdaD^{2}P. AdaD^{2}P achieves a utility bound of $\mathcal{O}\left(\sqrt{d \log\left(\frac{1}{\delta}\right)}/(\sqrt{n} J \epsilon)\right)$, matching that*

of distributed methods with server-client structures, such as LDP SVRG/SPIDER, and SoteriaFL-SAGA/SVRG without communication compression ($\omega = 0$). Furthermore, AdaD^{2}P recovers the utility bound of centralized DP-SGD with $n = 1$. For completeness, we provide the derivation of the utility bound of the baseline centralized DP-SGD in Section A.6 in the appendix.

Now, we provide the theoretical rationale behind the superior model performance of AdaD^{2}P in comparison to ConstD^{2}P under the same level of privacy protection. To this end, we split the upper bound of the metric (utility gap) into two distinct components: the standard term associated with SGP (Assran et al., 2019) and the term related to privacy noise, without specifying the value of K. As a result, we derive the following result for the proposed AdaD^{2}P algorithm.

Proposition 3. *Suppose Assumptions 1-5 hold. If the step size $\gamma \leqslant \min\left\{\frac{1-q}{6LC}, \frac{1}{L}\right\}$ and the noise scale $\sigma = c_2\sqrt{K\log\left(\frac{1}{\delta}\right)}/\left(J\epsilon\right)$, AdaD^{2}P (Algorithm 1) can achieve (ϵ, δ)-DP guarantee for each node after K iterations and has the following error bound*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right] \leqslant \Xi + \underbrace{\left(\frac{12\gamma^2L^2C^2\beta}{\left(1-q\right)^2} + \frac{2\gamma L\beta}{n}\right)\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla f_i\left(z_i^k;\xi_i^k\right)\right\|^2\right]}_{\text{caused by adaptive privacy noise}}, \tag{11}$$

where $\Xi = \frac{4\left(f\left(\bar{x}^0\right)-f^*\right)}{\gamma K} + \frac{2\gamma L}{n}\zeta^2 + \frac{12\gamma^2L^2C^2}{(1-q)^2}\left(\zeta^2+3b^2\right) + \frac{12L^2C^2}{(1-q)^2nK}\sum_{i=1}^{n}\left\|x_i^0\right\|^2$ is the standard error term of SGP algorithm, and β is defined at (6).

Proof. The complete proof can be found in Section A.4 in the appendix. □

Next, we provide a similar result for ConstD^{2}P (Algorithm 2) which employs fixed-level noise relying on the following bounded gradient assumption.

Assumption 6 (Bounded gradient)**.** *For any $z \in \mathbb{R}^d$ and $\xi_i \in \{1, 2, ..., J\}$, there exists finite positive constant G such that*

$$\left\|\nabla f_i\left(z;\xi_i\right)\right\| \leqslant G. \tag{12}$$

Proposition 4. *Under the same condition of Proposition 3 and suppose Assumption 6 holds, ConstD^{2}P (Algorithm 2) can achieve (ϵ, δ)-DP guarantee for each node after K iterations and*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right] \leqslant \Xi + \underbrace{\left(\frac{12\gamma^2L^2C^2\beta}{\left(1-q\right)^2} + \frac{2\gamma L\beta}{n}\right)\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}G^2}_{\text{caused by fixed privacy noise}}. \tag{13}$$

Proof. The complete proof can be found in Section A.5 in the appendix. □

Remark 4 (Insights for the adavantage of AdaD^{2}P)**.** *The comparison between the upper bounds in (11) and (13) reveals a significant difference in the components related to privacy noise. In particular, as the stochastic gradient norm tends to decay throughout the training process, it becomes evident that the component related to privacy noise in AdaD^{2}P is much tighter compared to that of ConstD^{2}P. This insight sheds light on the reason behind AdaD^{2}P outperforming ConstD^{2}P under the same level of privacy protection, as demonstrated in the experiments in Section 5.*

5 Experiments

We conduct several experiments to verify the performance of AdaD^{2}P (Algorithm 1), with comparison to the counterparts algorithms ConstD^{2}P (Algorithm 2) and ClipD^{2}P (Algorithm 3) which both employ fixed-level noise. All experiments are deployed in a high performance server with Intel Xeon E5-2680 v4 CPU @ 2.40GHz and 8 Nvidia RTX 3090 GPUs, and are implemented with distributed communication package *torch.distributed* in PyTorch (Paszke et al., 2017), where a process serves as a node, and inter-process communication is used to mimic communication among nodes.

Experimental setup. We compare three algorithms in fully decentralized setting composed of 20 nodes, on two non-convex learning tasks (i.e., deep CNN ResNet-18 training and shallow 2-layer

neural network training). For all experiments, we split shuffled datasets evenly to 20 nodes and use time-varying directed exponential graph (refer to Section C in the appendix for its definition) as communication topology for three algorithms. The learning rate is set as 0.05 for ResNet-18 training and 0.03 for 2-layer neural network training. It is worth noting that bounded gradient (Assumption 6) is required for ConstD^{2}P (Algorithm 2). To obtain this upper bound G, we run non-private SGP algorithm (no privacy noise) 5 times in advance and use maximum norm of stochastic gradient of the training process to be the estimate of G. In addition, according to Proposition 2, we know that when fixing δ (usually set as 10^{-5}), the privacy budget ϵ depends on the noise scale σ and the total number of iterations K. That is to say, if we run three algorithms for the same total iterations with the same noise scale σ, the privacy protection levels for three algorithms are the same.

5.1 Deep CNN ResNet-18 training

The first task is to train CNN model ResNet-18 (He et al., 2016) on Cifar-10 dataset (Krizhevsky et al., 2009). In this setting, the value of G is estimated to be 8.5 using our aforementioned approach. For ClipD^{2}P (Algorithm 3), we test the fixed clipping bound C with three different values chosen from the set $\{2, 3, 5\}$. We run three algorithms for 3500 iterations.

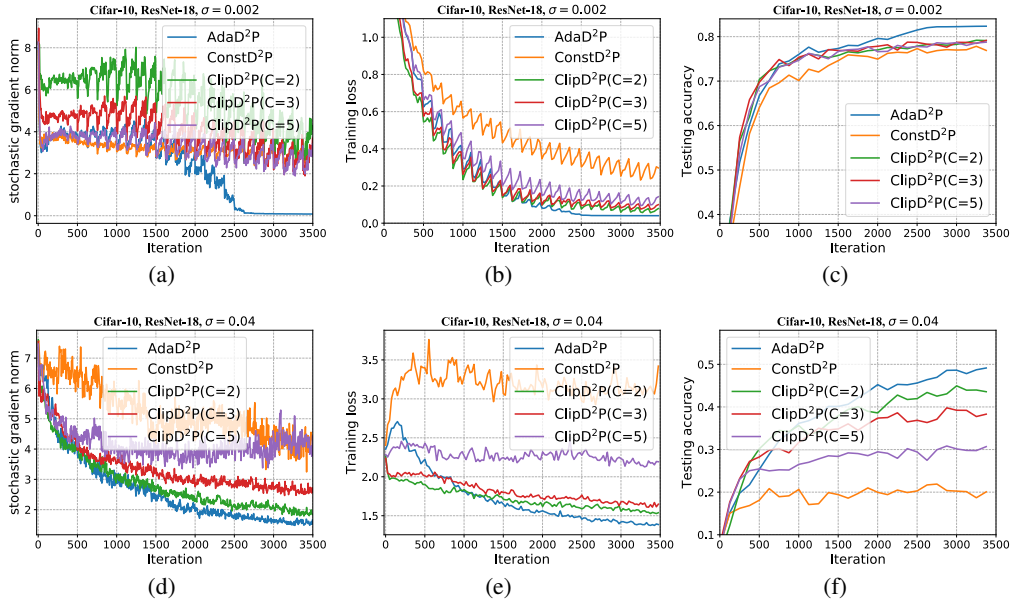
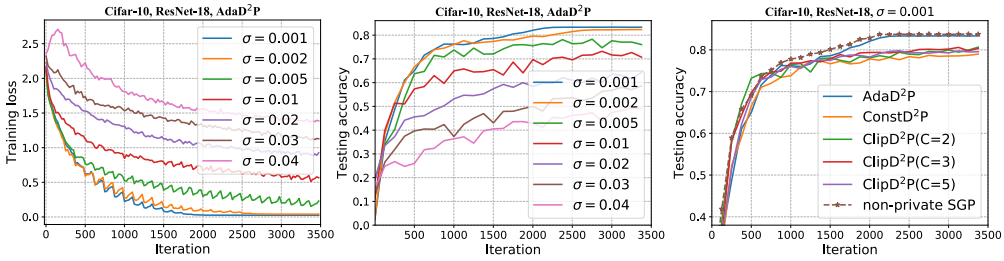


Figure 2: Performance comparison of training ResNet-18 for AdaD^{2}P with ConstD^{2}P and ClipD^{2}P under the same noise scale: $\sigma = 0.002$ for (a) (b) (c); $\sigma = 0.04$ for (d) (e) (f).

Performance comparison under the same level of privacy protection. We first set a relatively small $\sigma = 0.002$ for all three algorithms, which indicates a relatively modest level of privacy protection. The results depicted in Figures 2(a), 2(b) and 2(c) illustrate that, AdaD^{2}P outperforms the other two algorithms, in terms of the convergence of gradient norm, training loss and model accuracy. It is evident that when approaching the end of training, the gradient norm converges to a very small value near 0 for AdaD^{2}P, which results in a very minor amount of added noise, contributing positively to model accuracy further. In contrast, the other two algorithms inject fixed-level noise even during the later stages of the training process, leading to a degradation in model accuracy. -When setting a relatively larger $\sigma = 0.04$ which implies a relatively higher level of privacy protection, it follows from Figures 2(d), 2(e) and 2(f) that AdaD^{2}P still outperforms the other two algorithms, and shows more pronounced advantage in model accuracy (achieves a 30% higher model accuracy than ConstD^{2}P). In appendix D, we present additional experimental results of using other values of σ, and we have the same experimental observations.

Trade off between model utility and privacy protection level. We vary the value of noise scale σ from the set $\{0.001, 0.002, 0.005, 0.01, 0.02, 0.03, 0.04\}$, for AdaD^{2}P. The results presented in



(a) (b) (c)

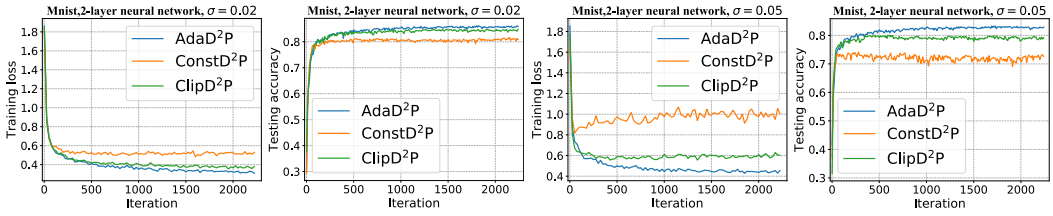
Figure 3: Performance comparison of training ResNet-18 for AdaD^{2}P under different noise scale σ in terms of (a) training loss and (b) testing accuracy; (c) Performance comparison of three algorithms with non-private (no noise) SGP algorithm under the noise scale $\sigma = 0.001$

.

Figures 3(a) and 3(b) show that, as noise scale σ increases which imply stronger privacy protection, the model utility (testing accuracy) deteriorates, illustrating the trade off between model utility and privacy protection level. Moreover, it follows from Figure 3(c) that when the noise scale is set as $\sigma = 0.001$, AdaD^{2}P is able to achieve model performance *almost without accuracy loss* compared to non-private SGP, while ConstD^{2}P and ClipD^{2}P still suffer from significant accuracy loss.

5.2 Shallow 2-layer neural network training

Next we consider a simple shallow 2-layer neural network training task on Mnist (Deng, 2012) dataset. For this task, the value of G is estimated to be 3.5 using the same approach. We set the clipping bound $C = 1$ for ClipD^{2}P and run three algorithms for the same 2200 iterations, and compare their performance under the same noise scale σ. It follows from the experimental results in Figure 4 that: under the same level of privacy protection (same σ), AdaD^{2}P outperforms the other two algorithms, and the advantage in model accuracy becomes more pronounced at a higher level of privacy protection (larger σ), verifying the superior performance of adaptive noise mechanism. Additional experimental tests for various σ values are provided in appendix D, and we can observe the same experimental phenomenon.



(a) (b) (c) (d)

Figure 4: Performance comparison of training 2-layer neural network for AdaD^{2}P with ConstD^{2}P and ClipD^{2}P under the same noise scale: $\sigma = 0.02$ for (a) (b); $\sigma = 0.05$ for (c) (d).

6 Conclusion

In this paper, we proposed a differentially private decentralized learning method for non-convex problems (termed AdaD^{2}P), which employs adaptive noise level and works for general time-varying communication topologies. Without relying on the bounded gradient assumption, we proved that AdaD^{2}P achieves a utility bound which matches that of distributed counterparts with server-client structures. Theoretical analysis revealed the inherent advantages of AdaD^{2}P employing adaptive noise as opposed to constant noise. We conducted extensive experiments to verify the superior performance of AdaD^{2}P compared to its counterparts which employ fixed-level noise.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

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

Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, and Mike Rabbat. Stochastic gradient push for distributed deep learning. In *International Conference on Machine Learning*, pp. 344–353. PMLR, 2019.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography: 14th International Conference, TCC 2016-B, Beijing, China, October 31-November 3, 2016, Proceedings, Part I*, pp. 635–658. Springer, 2016.

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

Xiangyi Chen, Steven Z Wu, and Mingyi Hong. Understanding gradient clipping in private sgd: A geometric perspective. *Advances in Neural Information Processing Systems*, 33:13773–13782, 2020.

Hsin-Pai Cheng, Patrick Yu, Haojing Hu, Feng Yan, Shiyu Li, Hai Li, and Yiran Chen. Leasgd: an efficient and privacy-preserving decentralized algorithm for distributed learning. *arXiv preprint arXiv:1811.11124*, 2018.

Hsin-Pai Cheng, Patrick Yu, Haojing Hu, Syed Zawad, Feng Yan, Shiyu Li, Hai Li, and Yiran Chen. Towards decentralized deep learning with differential privacy. In *International Conference on Cloud Computing*, pp. 130–145. Springer, 2019.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

Jiahao Ding, Guannan Liang, Jinbo Bi, and Miao Pan. Differentially private and communication efficient collaborative learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7219–7227, 2021.

Jian Du, Song Li, Xiangyi Chen, Siheng Chen, and Mingyi Hong. Dynamic differential-privacy preserving sgd. *arXiv preprint arXiv:2111.00173*, 2021.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pp. 371–380, 2009.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual international conference on the theory and applications of cryptographic techniques*, pp. 486–503. Springer, 2006.

Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 51–60. IEEE, 2010.

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.

Jie Fu, Zhili Chen, and Xiao Han. Adap dp-fl: Differentially private federated learning with adaptive noise. In *2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pp. 656–663. IEEE, 2022.

Jason Ge, Zhaoran Wang, Mengdi Wang, and Han Liu. Minimax-optimal privacy-preserving sparse pca in distributed systems. In *International Conference on Artificial Intelligence and Statistics*, pp. 1589–1598. PMLR, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Roger Iyengar, Joseph P Near, Dawn Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 299–316. IEEE, 2019.

David Kempe, Alin Dobra, and Johannes Gehrke. Gossip-based computation of aggregate information. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pp. 482–491. IEEE, 2003.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Matthias Langer, Zhen He, Wenny Rahayu, and Yanbo Xue. Distributed training of deep learning models: A taxonomic perspective. *IEEE Transactions on Parallel and Distributed Systems*, 31 (12):2802–2818, 2020.

Boyue Li and Yuejie Chi. Convergence and privacy of decentralized nonconvex optimization with gradient clipping and communication compression. *arXiv preprint arXiv:2305.09896*, 2023.

Jeffrey Li, Mikhail Khodak, Sebastian Caldas, and Ameet Talwalkar. Differentially private meta-learning. *arXiv preprint arXiv:1909.05830*, 2019a.

Mu Li, David G Andersen, Jun Woo Park, Alexander J Smola, Amr Ahmed, Vanja Josifovski, James Long, Eugene J Shekita, and Bor-Yiing Su. Scaling distributed machine learning with the parameter server. In *11th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 14)*, pp. 583–598, 2014.

Yanan Li, Shusen Yang, Xuebin Ren, and Cong Zhao. Asynchronous federated learning with differential privacy for edge intelligence. *arXiv preprint arXiv:1912.07902*, 2019b.

Zhize Li, Haoyu Zhao, Boyue Li, and Yuejie Chi. Soteriafl: A unified framework for private federated learning with communication compression. *Advances in Neural Information Processing Systems*, 35:4285–4300, 2022.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

Xiangru Lian, Wei Zhang, Ce Zhang, and Ji Liu. Asynchronous decentralized parallel stochastic gradient descent. In *International Conference on Machine Learning*, pp. 3043–3052. PMLR, 2018.

Tianyu Liu, Boya Di, Bin Wang, and Lingyang Song. Loss-privacy tradeoff in federated edge learning. *IEEE Journal of Selected Topics in Signal Processing*, 16(3):546–558, 2022.

Andrew Lowy, Ali Ghafelebashi, and Meisam Razaviyayn. Private non-convex federated learning without a trusted server, 2023.

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

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

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.

Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration. *PyTorch: Tensors and dynamic neural networks in Python with strong GPU acceleration*, 6(3):67, 2017.

Hyejin Shin, Sungwook Kim, Junbum Shin, and Xiaokui Xiao. Privacy enhanced matrix factorization for recommendation with local differential privacy. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1770–1782, 2018.

Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu. D^2: Decentralized training over decentralized data. In *International Conference on Machine Learning*, pp. 4848–4856. PMLR, 2018.

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30, 2017.

Di Wang, Hanshen Xiao, Srinivas Devadas, and Jinhui Xu. On differentially private stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning*, pp. 10081–10091. PMLR, 2020.

Lingxiao Wang, Bargav Jayaraman, David Evans, and Quanquan Gu. Efficient privacy-preserving stochastic nonconvex optimization. *arXiv e-prints*, pp. arXiv–1910, 2019a.

Lingxiao Wang, Bargav Jayaraman, David Evans, and Quanquan Gu. Efficient privacy-preserving stochastic nonconvex optimization. In *Uncertainty in Artificial Intelligence*, pp. 2203–2213. PMLR, 2023.

Yongqiang Wang and Angelia Nedic. Tailoring gradient methods for differentially-private distributed optimization. *arXiv preprint arXiv:2202.01113*, 2022.

Zhibo Wang, Mengkai Song, Zhifei Zhang, Yang Song, Qian Wang, and Hairong Qi. Beyond inferring class representatives: User-level privacy leakage from federated learning. In *IEEE INFOCOM 2019-IEEE conference on computer communications*, pp. 2512–2520. IEEE, 2019b.

Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020.

Kang Wei, Jun Li, Ming Ding, Chuan Ma, Hang Su, Bo Zhang, and H Vincent Poor. User-level privacy-preserving federated learning: Analysis and performance optimization. *IEEE Transactions on Mobile Computing*, 21(9):3388–3401, 2021a.

Kang Wei, Jun Li, Chuan Ma, Ming Ding, Cailian Chen, Shi Jin, Zhu Han, and H Vincent Poor. Low-latency federated learning over wireless channels with differential privacy. *IEEE Journal on Selected Areas in Communications*, 40(1):290–307, 2021b.

Wenqi Wei and Ling Liu. Gradient leakage attack resilient deep learning. *IEEE Transactions on Information Forensics and Security*, 17:303–316, 2021.

Wenqi Wei, Ling Liu, Jingya Zhou, Ka-Ho Chow, and Yanzhao Wu. Securing distributed sgd against gradient leakage threats. *IEEE Transactions on Parallel and Distributed Systems*, 2023.

Nan Wu, Farhad Farokhi, David Smith, and Mohamed Ali Kaafar. The value of collaboration in convex machine learning with differential privacy. In *2020 IEEE Symposium on Security and Privacy (SP)*, pp. 304–317. IEEE, 2020.

Jie Xu, Wei Zhang, and Fei Wang. A (dp)ˆ 2sgd: Asynchronous decentralized parallel stochastic gradient descent with differential privacy. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

Dongxiao Yu, Zongrui Zou, Shuzhen Chen, Youming Tao, Bing Tian, Weifeng Lv, and Xiuzhen Cheng. Decentralized parallel sgd with privacy preservation in vehicular networks. *IEEE Transactions on Vehicular Technology*, 70(6):5211–5220, 2021.

Yiming Zeng, Yixuan Lin, Yuanyuan Yang, and Ji Liu. Differentially private federated temporal difference learning. *IEEE Transactions on Parallel & Distributed Systems*, (01):1–1, 2021.

Xin Zhang, Minghong Fang, Jia Liu, and Zhengyuan Zhu. Private and communication-efficient edge learning: a sparse differential gaussian-masking distributed sgd approach. In *Proceedings of the Twenty-First International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pp. 261–270, 2020.

Yipeng Zhou, Xuezheng Liu, Yao Fu, Di Wu, Jessie Hui Wang, and Shui Yu. Optimizing the numbers of queries and replies in convex federated learning with differential privacy. *IEEE Transactions on Dependable and Secure Computing*, 2023.

Martin Zinkevich, Markus Weimer, Lihong Li, and Alex Smola. Parallelized stochastic gradient descent. *Advances in neural information processing systems*, 23, 2010.

Appendix

CONTENTS

A PROOF OF MAIN RESULTS

To facilitate our analysis, we first rewrite the 9^{th} step of the proposed AdaD^{2}P (c.f., Algorithm 1) in a compact form:

$$X^{k+1} = \left(X^k - \gamma \left(\nabla F\left(Z^k; \xi^k\right) + N^k\right)\right) \left(P^k\right)^\top \tag{14}$$

where $\left(P^k\right)^\top \in \mathbb{R}^{n\times n}$ is the transpose of the mixing matrix P^k at iteration k, and

$X^k := \left[x_1^k, x_2^k, \cdots, x_n^k\right] \in \mathbb{R}^{d\times n}$: concatenation of all the nodes' parameters at iteration k;

$Z^k := \left[z_1^k, z_2^k, \cdots, z_n^k\right] \in \mathbb{R}^{d\times n}$: concatenation of all the nodes' de-biased parameters at iteration k;

$\nabla F(Z^k; \xi^k) := \left[\nabla f_1(z_1^k; \xi_1^k), \nabla f_2(z_2^k; \xi_2^k), ..., \nabla f_n(z_n^k; \xi_n^k)\right] \in \mathbb{R}^{d\times n}$: concatenation of all the nodes' stochastic gradients at iteration k;

$N^k := \left[N_1^k, N_2^k, ..., N_n^k\right] \in \mathbb{R}^{d\times n}$: concatenation of all the nodes' added Gaussian noise at iteration k.

Now, let $\bar{x}^k = \frac{1}{n} X^k \mathbf{1} = \frac{1}{n} \sum_{i=1}^n x_i^k \in \mathbb{R}^d$ denote the average of all nodes' parameters at iteration k. Then, the update of average system of (14) becomes

$$\bar{x}^{k+1} = \bar{x}^k - \gamma \cdot \left(\frac{1}{n} \sum_{i=1}^n \nabla f_i(z_i^k; \xi_i^k) + \frac{1}{n} \sum_{i=1}^n N_i^k) \right) \tag{15}$$

which can be easily obtained by right multiplying $\frac{1}{n}\mathbf{1}$ from both sides of (14) and using the column-stochastic property of P^k (c.f., Assumption 1). The above average system will be useful in subsequent analysis.

In addition, we denote $\mathcal{F}^k := \left\{ \bigcup_{i=1}^n \left(x_i^0, z_i^0, \xi_i^0, N_i^0, \cdots, x_i^{k-1}, z_i^{k-1}, \xi_i^{k-1}, N_i^{k-1}, x_i^k, z_i^k \right) \right\}$ as filtration of the history sequence upto k, and define $\mathbb{E}\left[\cdot \middle| \mathcal{F}^k \right]$ the conditional expectation given $\mathcal{F}^k$.

Outline of Proof. Our proof of utility guarantee for AdaD^{2}P (c.f., Theorem 1) consists of three parts: i) we first provide several technical lemmas to facilitate the subsequent analysis (c.f., Section A.1); ii) we then provide two supporting lemmas to establish two key inequalities (c.f., Section A.2) where the first inequality is obtained by applying descent lemma (c.f., (18)) and the second inequality is to upper bound the consensus error (c.f.,(26)); iii) we finally upper bound the error term arising from injected DP noise to obtain the third key inequality (c.f., (34)). As a result, the utility bound can be derived by constructing a loop for the above three key inequalities with proper choice of certain parameters such as the step size γ and the total number of iterations K (c.f., Section A.3).

A.1 Important Upper Bounds

In this section, we first provide several technical lemmas to facilitate the subsequent analysis.

Lemma 1. *Let $\left\{ v^k \right\}_{k=0}^{\infty}$ be a non-negative sequence and $\lambda \in (0,1)$. Then, we have*

$$\left(\sum_{l=0}^k \lambda^{k-l} v^l \right)^2 \leqslant \frac{1}{1-\lambda} \sum_{l=0}^k \lambda^{k-l} \left(v^l \right)^2. \tag{16}$$

Proof. Using Cauchy-Swarchz inequality, we have

$$\begin{aligned}
\left(\sum_{l=0}^k \lambda^{k-l} v^l \right)^2 &= \left(\sum_{l=0}^k \lambda^{\frac{k-l}{2}} \left(\lambda^{\frac{k-l}{2}} v^l \right) \right)^2 \\
&\leqslant \sum_{l=0}^k \left(\lambda^{\frac{k-l}{2}} \right)^2 \cdot \sum_{l=0}^k \left(\lambda^{\frac{k-l}{2}} v^l \right)^2 \\
&\leqslant \frac{1}{1-\lambda} \sum_{l=0}^k \lambda^{k-l} \left(v^l \right)^2,
\end{aligned}$$

which completes the proof. □

Lemma 2. *Suppose Assumptions 3 and 5 hold. Then, we have*

$$\left\| \nabla f_i \left(z_i^k \right) \right\|^2 \leqslant 3L^2 \left\| z_i^k - \bar{x}^k \right\|^2 + 3b^2 + 3 \left\| \nabla f \left(\bar{x}^k \right) \right\|^2. \tag{17}$$

Proof. Using Assumption 3 (L-smooth) and Assumption 5, we have

$$\begin{aligned}
\left\| \nabla f_i \left(z_i^k \right) \right\|^2 &= \left\| \nabla f_i \left(z_i^k \right) - \nabla f_i \left(\bar{x}^k \right) + \nabla f_i \left(\bar{x}^k \right) - \nabla f \left(\bar{x}^k \right) + \nabla f \left(\bar{x}^k \right) \right\|^2 \\
&\leqslant 3 \left\| \nabla f_i \left(z_i^k \right) - \nabla f_i \left(\bar{x}^k \right) \right\|^2 + 3 \left\| \nabla f_i \left(\bar{x}^k \right) - \nabla f \left(\bar{x}^k \right) \right\|^2 + 3 \left\| \nabla f \left(\bar{x}^k \right) \right\|^2 \\
&\leqslant 3L^2 \left\| z_i^k - \bar{x}^k \right\|^2 + 3b^2 + 3 \left\| \nabla f \left(\bar{x}^k \right) \right\|^2,
\end{aligned}$$

which completes the proof. □

A.2 SUPPORTING LEMMAS

The following lemma is crucial to the convergence analysis in the non-convex stochastic optimization, which is obtained by applying the descent lemma recursively from $k = 0$ to K.

Lemma 3. *Suppose Assumption 3, 4 and 5 hold. For a given constant step size γ, we have*

$$\begin{aligned}
&\frac{\gamma}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]+\frac{\gamma\left(1-\gamma L\right)}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i\left(z_i^k\right)\right\|^2\right]\\
\leqslant& f\left(\bar{x}^0\right)-f^*+\frac{\gamma L^2}{2}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|z_i^k-\bar{x}^k\right\|^2\right]+\frac{\gamma^2 LK}{2n}\zeta^2\\
&+\frac{\gamma^2 L}{2n}\cdot\underbrace{\beta K\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla f_i(z_i^k;\xi_i^k)\right\|^2\right]}_{\textit{caused by injected DP noise}}.
\end{aligned}\tag{18}$$

Proof. Applying the descent lemma to f at $\bar{x}^k$ and $\bar{x}^{k+1}$, we have

$$\begin{aligned}
f\left(\bar{x}^{k+1}\right)\leqslant& f\left(\bar{x}^k\right)+\left\langle\nabla f\left(\bar{x}^k\right),\bar{x}^{k+1}-\bar{x}^k\right\rangle+\frac{L}{2}\left\|\bar{x}^{k+1}-\bar{x}^k\right\|^2\\
\overset{(15)}{=}& f\left(\bar{x}^k\right)-\gamma\left\langle\nabla f\left(\bar{x}^k\right),\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(z_i^k;\xi_i^k)+\frac{1}{n}\sum_{i=1}^{n}N_i^k\right\rangle\\
&+\frac{\gamma^2 L}{2}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(z_i^k;\xi_i^k)+\frac{1}{n}\sum_{i=1}^{n}N_i^k\right\|^2\\
=& f\left(\bar{x}^k\right)-\gamma\left\langle\nabla f\left(\bar{x}^k\right),\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(z_i^k;\xi_i^k)+\frac{1}{n}\sum_{i=1}^{n}N_i^k)\right\rangle+\frac{\gamma^2 L}{2}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(z_i^k;\xi_i^k)\right\|^2\\
&+\frac{\gamma^2 L}{2}\left\|\frac{1}{n}\sum_{i=1}^{n}N_i^k\right\|^2+\gamma^2 L\left\langle\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(z_i^k;\xi_i^k),\frac{1}{n}\sum_{i=1}^{n}N_i^k\right\rangle.
\end{aligned}$$

Taking the expectation of both sides conditioned on $\mathcal{F}^k$ for the above inequality, we obtain

$$\begin{aligned}
&\mathbb{E}\left[f\left(\bar{x}^{k+1}\right)\left|\mathcal{F}^k\right.\right]\\
\leqslant& f\left(\bar{x}^k\right)-\gamma\mathbb{E}\left[\mathbb{E}\left[\left\langle\nabla f\left(\bar{x}^k\right),\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(z_i^k;\xi_i^k)+\frac{1}{n}\sum_{i=1}^{n}N_i^k)\right\rangle\left|\mathcal{F}^k,\ \xi^k\right.\right]\left|\mathcal{F}^k\right.\right]\\
&+\frac{\gamma^2 L}{2}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(z_i^k;\xi_i^k)\right\|^2\left|\mathcal{F}^k\right.\right]+\frac{\gamma^2 L}{2}\mathbb{E}\left[\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}N_i^k\right\|^2\left|\mathcal{F}^k,\ \xi^k\right.\right]\left|\mathcal{F}^k\right.\right]\\
&+\gamma^2 L\mathbb{E}\left[\mathbb{E}\left[\left\langle\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(z_i^k;\xi_i^k),\frac{1}{n}\sum_{i=1}^{n}N_i^k\right\rangle\left|\mathcal{F}^k,\ \xi^k\right.\right]\left|\mathcal{F}^k\right.\right]\\
\overset{(7)}{=}& f\left(\bar{x}^k\right)\underbrace{-\gamma\left\langle\nabla f\left(\bar{x}^k\right),\frac{1}{n}\sum_{i=1}^{n}\nabla f_i\left(z_i^k\right)\right\rangle}_{A_1}+\underbrace{\frac{\gamma^2 L}{2}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(z_i^k;\xi_i^k)\right\|^2\left|\mathcal{F}^k\right.\right]}_{A_2}\\
&+\underbrace{\frac{\gamma^2 L}{2}\mathbb{E}\left[\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}N_i^k\right\|^2\left|\mathcal{F}^k,\ \xi^k\right.\right]\left|\mathcal{F}^k\right.\right]}_{A_3}.
\end{aligned}\tag{19}$$

For A_1 in (19), we have

$$\begin{aligned}
A_1 =& -\frac{\gamma}{2}\left\|\nabla f\left(\bar{x}^k\right)\right\|^2 - \frac{\gamma}{2}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i\left(z_i^k\right)\right\|^2 + \frac{\gamma}{2}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i\left(z_i^k\right) - \nabla f\left(\bar{x}^k\right)\right\|^2 \\
=& -\frac{\gamma}{2}\left\|\nabla f\left(\bar{x}^k\right)\right\|^2 - \frac{\gamma}{2}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i\left(z_i^k\right)\right\|^2 + \frac{\gamma}{2}\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_i\left(z_i^k\right) - \nabla f_i\left(\bar{x}^k\right)\right)\right\|^2 \\
\leqslant& -\frac{\gamma}{2}\left\|\nabla f\left(\bar{x}^k\right)\right\|^2 - \frac{\gamma}{2}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i\left(z_i^k\right)\right\|^2 + \frac{\gamma}{2n}\sum_{i=1}^n \left\|\nabla f_i\left(z_i^k\right) - \nabla f_i\left(\bar{x}^k\right)\right\|^2 \\
\overset{(a)}{\leqslant}& -\frac{\gamma}{2}\left\|\nabla f\left(\bar{x}^k\right)\right\|^2 - \frac{\gamma}{2}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i\left(z_i^k\right)\right\|^2 + \frac{\gamma L^2}{2n}\sum_{i=1}^n \left\|z_i^k - \bar{x}^k\right\|^2,
\end{aligned} \tag{20}$$

where in (a) we used Assumption 3.

For A_2 in (19), we have

$$\begin{aligned}
A_2 =& \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_i(z_i^k;\xi_i^k) - \nabla f_i(z_i^k)\right) + \frac{1}{n}\sum_{i=1}^n \nabla f_i(z_i^k)\right\|^2 \left|\mathcal{F}^k\right.\right] \\
\overset{(7)}{=}& \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(z_i^k)\right\|^2 \left|\mathcal{F}^k\right.\right] + \frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^n \left(\nabla f_i(z_i^k;\xi_i^k) - \nabla f_i(z_i^k)\right)\right\|^2 \left|\mathcal{F}^k\right.\right] \\
=& \left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(z_i^k)\right\|^2 + \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|\nabla f_i(z_i^k;\xi_i^k) - \nabla f_i\left(z_i^k\right)\right\|^2 \left|\mathcal{F}^k\right.\right] \\
\overset{(8)}{\leqslant}& \left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(z_i^k)\right\|^2 + \frac{\zeta^2}{n}.
\end{aligned} \tag{21}$$

For A_3 in (19), we have

$$\begin{aligned}
A_3 \overset{(6)}{=}& \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^n \beta K\left\|\nabla f_i(z_i^k;\xi_i^k)\right\|^2 \left|\mathcal{F}^k\right.\right] \\
=& \frac{\beta K}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|\nabla f_i(z_i^k;\xi_i^k)\right\|^2 \left|\mathcal{F}^k\right.\right].
\end{aligned} \tag{22}$$

Then, substituting (20), (21) and (22) into (19) yields

$$\begin{aligned}
&\mathbb{E}\left[f\left(\bar{x}^{k+1}\right)\left|\mathcal{F}^k\right.\right] \\
\leqslant& f\left(\bar{x}^k\right) - \frac{\gamma}{2}\left\|\nabla f\left(\bar{x}^k\right)\right\|^2 - \frac{\gamma\left(1-\gamma L\right)}{2}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i\left(z_i^k\right)\right\|^2 + \frac{\gamma L^2}{2n}\sum_{i=1}^n \left\|z_i^k - \bar{x}^k\right\|^2 \\
&+ \frac{\gamma^2 L\zeta^2}{2n} + \frac{\gamma^2 L}{2}\cdot\frac{\beta K}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|\nabla f_i(z_i^k;\xi_i^k)\right\|^2 \left|\mathcal{F}^k\right.\right].
\end{aligned} \tag{23}$$

Taking total expectation on both sides of (23), yields

$$\begin{aligned}
\mathbb{E}\left[f\left(\bar{x}^{k+1}\right)\right] \leqslant& \mathbb{E}\left[f\left(\bar{x}^k\right)\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right] - \frac{\gamma\left(1-\gamma L\right)}{2}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i\left(z_i^k\right)\right\|^2\right] \\
&+ \frac{\gamma L^2}{2n}\sum_{i=1}^n \mathbb{E}\left[\left\|z_i^k - \bar{x}^k\right\|^2\right] + \frac{\gamma^2 L\zeta^2}{2n} + \frac{\gamma^2 L}{2}\cdot\frac{\beta K}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|\nabla f_i(z_i^k;\xi_i^k)\right\|^2\right].
\end{aligned}$$

Summing the above inequality form $k=0$ to $K-1$, we obtain (18), which completes the proof. □

The following lemma bounds the distance between the de-biased parameters z_i^k at each node i and the node-wise average $\bar{x}^k$, which can be adapted from Lemma 3 in (Assran et al., 2019).

Lemma 4. *Suppose that Assumptions 1 and 2 hold. Let ε be the minimum of all non-zero mixing weights, $\lambda = 1 - n\varepsilon^{\triangle B}$ and $q = \lambda^{\frac{1}{\triangle B+1}}$. Then, there exists a constant*

$$C < \frac{2\sqrt{d}\varepsilon^{-\triangle B}}{\lambda^{\frac{\triangle B+2}{\triangle B+1}}}, \tag{24}$$

such that for any $i \in \mathcal{V}$ and $k \geqslant 0$, we have

$$\left\|z_i^k - \bar{x}^k\right\| \leqslant Cq^k \left\|x_i^0\right\| + \gamma C \sum_{s=0}^{k} q^{k-s} \left\|\nabla f_i\left(z_i^s; \xi_i^s\right) + N_i^s\right\|. \tag{25}$$

Now, we attempt to upper bound the accumulative consensus error $\sum_{k=0}^{K-1} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\left\|z_i^k - \bar{x}^k\right\|^2\right]$ using $\sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]$, which is summarized in the following lemma.

Lemma 5. *Suppose Assumptions 1-5 hold. Then, we have*

$$\begin{aligned}
&\left(1 - \frac{9\gamma^2 L^2 C^2}{(1-q)^2}\right) \sum_{k=0}^{K-1} \cdot \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\left\|z_i^k - \bar{x}^k\right\|^2\right] \\
\leqslant &\frac{9\gamma^2 C^2}{(1-q)^2} \sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right] + K \cdot \frac{3\gamma^2 C^2}{(1-q)^2}\left(\zeta^2 + 3b^2\right) \\
&+ \frac{3C^2}{(1-q)^2 n} \sum_{i=1}^{n} \left\|x_i^0\right\|^2 + \frac{3\gamma^2 C^2}{(1-q)^2} \cdot \underbrace{\beta K \sum_{k=0}^{K-1} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\left\|\nabla f_i(z_i^k; \xi_i^k)\right\|^2\right]}_{\text{caused by injected DP noise}}.
\end{aligned} \tag{26}$$

Proof. According to (25), we have

$$\left\|z_i^k - \bar{x}^k\right\| \leqslant Cq^k \left\|x_i^0\right\| + \gamma C \sum_{s=0}^{k} q^{k-s} \left\|\nabla f_i\left(z_i^s; \xi_i^s\right)\right\| + \gamma C \sum_{s=0}^{k} q^{k-s} \left\|N_i^s\right\|. \tag{27}$$

Squaring on both sides of (27), we have

$$\begin{aligned}
\left\|z_i^k - \bar{x}^k\right\|^2 \leqslant &\left(Cq^k \left\|x_i^0\right\| + \gamma C \sum_{s=0}^{k} q^{k-s} \left\|\nabla f_i\left(z_i^s; \xi_i^s\right)\right\| + \gamma C \sum_{s=0}^{k} q^{k-s} \left\|N_i^s\right\|\right)^2 \\
\leqslant &3C^2 q^{2k} \left\|x_i^0\right\|^2 + 3\gamma^2 C^2 \left(\sum_{s=0}^{k} q^{k-s} \left\|\nabla f_i\left(z_i^s; \xi_i^s\right)\right\|\right)^2 + 3\gamma^2 C^2 \left(\sum_{s=0}^{k} q^{k-s} \left\|N_i^s\right\|\right)^2 \\
\leqslant &3C^2 q^{2k} \left\|x_i^0\right\|^2 + \frac{3\gamma^2 C^2}{1-q} \sum_{s=0}^{k} q^{k-s} \left\|\nabla f_i\left(z_i^s; \xi_i^s\right)\right\|^2 + \frac{3\gamma^2 C^2}{1-q} \sum_{s=0}^{k} q^{k-s} \left\|N_i^s\right\|^2,
\end{aligned} \tag{28}$$

where we used $(a+b+c)^2 \leqslant 3a^2 + 3b^2 + 3c^2$ in the second inequality and Lemma 1 in the last inequality, respectively.

Taking total expectation on both sides of (28) yields

$$
\begin{aligned}
&\mathbb{E}\left[\left\|z_i^k-\bar{x}^k\right\|^2\right] \\
\leqslant & 3C^2q^{2k}\left\|x_i^0\right\|^2+\frac{3\gamma^2C^2}{1-q}\sum_{s=0}^{k}q^{k-s}\mathbb{E}\left[\left\|\nabla f_i\left(z_i^s;\xi_i^s\right)\right\|^2\right]+\frac{3\gamma^2C^2}{1-q}\sum_{s=0}^{k}q^{k-s}\mathbb{E}\left[\left\|N_i^s\right\|^2\right] \\
\overset{(a)}{=} & 3C^2q^{2k}\left\|x_i^0\right\|^2+\frac{3\gamma^2C^2}{1-q}\sum_{s=0}^{k}q^{k-s}\mathbb{E}\left[\left\|\nabla f_i\left(z_i^s;\xi_i^s\right)-\nabla f_i\left(z_i^s\right)+\nabla f_i\left(z_i^s\right)\right\|^2\right] \\
&+\frac{3\gamma^2C^2\beta K}{1-q}\sum_{s=0}^{k}q^{k-s}\mathbb{E}\left[\left\|\nabla f_i\left(z_i^s;\xi_i^s\right)\right\|^2\right] \\
\overset{(b)}{=} & 3C^2q^{2k}\left\|x_i^0\right\|^2+\frac{3\gamma^2C^2}{1-q}\sum_{s=0}^{k}q^{k-s}\mathbb{E}\left[\left\|\nabla f_i\left(z_i^s;\xi_i^s\right)-\nabla f_i\left(z_i^s\right)\right\|^2\right] \\
&+\frac{3\gamma^2C^2}{1-q}\sum_{s=0}^{k}q^{k-s}\mathbb{E}\left[\left\|\nabla f_i\left(z_i^s\right)\right\|^2\right]+\frac{3\gamma^2C^2\beta K}{1-q}\sum_{s=0}^{k}q^{k-s}\mathbb{E}\left[\left\|\nabla f_i\left(z_i^s;\xi_i^s\right)\right\|^2\right],
\end{aligned}
\tag{29}
$$

where we used (6) in (a), and (7) in (b).

Further, using (8) and Lemma 2, the above inequality becomes

$$
\begin{aligned}
&\mathbb{E}\left[\left\|z_i^k-\bar{x}^k\right\|^2\right] \\
\leqslant & 3C^2q^{2k}\left\|x_i^0\right\|^2+\frac{3\gamma^2C^2}{1-q}\sum_{s=0}^{k}q^{k-s}\cdot\zeta^2+\frac{3\gamma^2C^2\beta K}{1-q}\sum_{s=0}^{k}q^{k-s}\mathbb{E}\left[\left\|\nabla f_i\left(z_i^s;\xi_i^s\right)\right\|^2\right] \\
&+\frac{3\gamma^2C^2}{1-q}\sum_{s=0}^{k}q^{k-s}\mathbb{E}\left[3L^2\left\|z_i^s-\bar{x}^s\right\|^2+3b^2+3\left\|\nabla f\left(\bar{x}^s\right)\right\|^2\right] \\
\leqslant & 3C^2q^{2k}\left\|x_i^0\right\|^2+\frac{3\gamma^2C^2}{(1-q)^2}\left(\zeta^2+3b^2\right)+\frac{9\gamma^2L^2C^2}{1-q}\sum_{s=0}^{k}q^{k-s}\mathbb{E}\left[\left\|z_i^s-\bar{x}^s\right\|^2\right] \\
&+\frac{9\gamma^2C^2}{1-q}\sum_{s=0}^{k}q^{k-s}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^s\right)\right\|^2\right]+\frac{3\gamma^2C^2\beta K}{1-q}\sum_{s=0}^{k}q^{k-s}\mathbb{E}\left[\left\|\nabla f_i\left(z_i^s;\xi_i^s\right)\right\|^2\right].
\end{aligned}
\tag{30}
$$

Summing (30) from $i=1$ to n and dividing by n, we have

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|z_i^k-\bar{x}^k\right\|^2\right] \\
\leqslant & \frac{3C^2q^{2k}}{n}\sum_{i=1}^{n}\left\|x_i^0\right\|^2+\frac{3\gamma^2C^2}{(1-q)^2}\left(\zeta^2+3b^2\right)+\frac{9\gamma^2L^2C^2}{1-q}\sum_{s=0}^{k}q^{k-s}\cdot\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|z_i^s-\bar{x}^s\right\|^2\right] \\
&+\frac{9\gamma^2C^2}{1-q}\sum_{s=0}^{k}q^{k-s}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^s\right)\right\|^2\right]+\frac{3\gamma^2C^2\beta K}{1-q}\sum_{s=0}^{k}q^{k-s}\cdot\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla f_i\left(z_i^s;\xi_i^s\right)\right\|^2\right].
\end{aligned}
\tag{31}
$$

Summing (31) from $k=0$ to $K-1$, we obtain

$$\begin{aligned}
&\sum_{k=0}^{K-1}\cdot\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|z_i^k-\bar{x}^k\right\|^2\right]\\
\leqslant&\frac{3C^2}{(1-q)^2 n}\sum_{i=1}^{n}\left\|x_i^0\right\|^2+K\cdot\frac{3\gamma^2C^2}{(1-q)^2}\left(\zeta^2+3b^2\right)+\frac{9\gamma^2L^2C^2}{1-q}\sum_{k=0}^{K-1}\sum_{s=0}^{k}q^{k-s}\cdot\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|z_i^s-\bar{x}^s\right\|^2\right]\\
&+\frac{9\gamma^2C^2}{1-q}\sum_{k=0}^{K-1}\sum_{s=0}^{k}q^{k-s}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^s\right)\right\|^2\right]+\frac{3\gamma^2C^2\beta K}{1-q}\sum_{k=0}^{K-1}\sum_{s=0}^{k}q^{k-s}\cdot\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla f_i\left(z_i^s;\xi_i^s\right)\right\|^2\right]\\
\overset{(c)}{\leqslant}&\frac{3C^2}{(1-q)^2 n}\sum_{i=1}^{n}\left\|x_i^0\right\|^2+K\cdot\frac{3\gamma^2C^2}{(1-q)^2}\left(\zeta^2+3b^2\right)+\frac{9\gamma^2L^2C^2}{(1-q)^2}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|z_i^k-\bar{x}^k\right\|^2\right]\\
&+\frac{9\gamma^2C^2}{(1-q)^2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]+\frac{3\gamma^2C^2\beta K}{(1-q)^2}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla f_i\left(z_i^k;\xi_i^k\right)\right\|^2\right].
\end{aligned}\tag{32}$$

where in (c) we used the fact that $\sum_{k=0}^{K-1}\sum_{s=0}^{k}q^{k-s}a_s\leqslant\frac{1}{1-q}\sum_{k=0}^{K-1}a_k$.

Rearranging the term in (32), we obtain

$$\begin{aligned}
&\left(1-\frac{9\gamma^2L^2C^2}{(1-q)^2}\right)\sum_{k=0}^{K-1}\cdot\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|z_i^k-\bar{x}^k\right\|^2\right]\\
\leqslant&\frac{9\gamma^2C^2}{(1-q)^2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]+K\cdot\frac{3\gamma^2C^2}{(1-q)^2}\left(\zeta^2+3b^2\right)\\
&+\frac{3C^2}{(1-q)^2 n}\sum_{i=1}^{n}\left\|x_i^0\right\|^2+\frac{3\gamma^2C^2\beta K}{(1-q)^2}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla f_i\left(z_i^k;\xi_i^k\right)\right\|^2\right],
\end{aligned}\tag{33}$$

which completes the proof. □

A.3 Proof of Theorem 1

With two supporting lemmas (Lemma 3 and Lemma 5) in the previous section, we are now ready to prove Theorem 1. We first upper bound the error term arising from injected DP noise as follows:

$$\begin{aligned}
&\beta K\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla f_i(z_i^k;\xi_i^k)\right\|^2\right]\\
=&\beta K\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla f_i\left(z_i^k;\xi_i^k\right)-\nabla f_i\left(z_i^k\right)+\nabla f_i\left(z_i^k\right)\right\|^2\right]\\
=&\beta K\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla f_i\left(z_i^k;\xi_i^k\right)-\nabla f_i\left(z_i^k\right)\right\|^2\right]+\beta K\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla f_i\left(z_i^k\right)\right\|^2\right]\\
\overset{(a)}{\leqslant}&\beta K\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\zeta^2+\beta K\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[3L^2\left\|z_i^k-\bar{x}^k\right\|^2+3b^2+3\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]\\
=&\beta K^2\left(\zeta^2+3b^2\right)+3\beta K\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]+3L^2\beta K\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|z_i^k-\bar{x}^k\right\|^2\right],
\end{aligned}\tag{34}$$

where in (a) we used (8) and Lemma 2.

Substituting (34) into (18), we have

$$\begin{aligned}
&\frac{\gamma}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]+\frac{\gamma\left(1-\gamma L\right)}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i\left(z_i^k\right)\right\|^2\right]\\
\leqslant& f\left(\bar{x}^0\right)-f^*+\frac{\gamma L^2}{2}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|z_i^k-\bar{x}^k\right\|^2\right]+\frac{\gamma^2 LK}{2n}\zeta^2+\frac{\gamma^2 L\beta K^2}{2n}\left(\zeta^2+3b^2\right)\\
&+\frac{3\gamma^2 L\beta K}{2n}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]+\frac{3\gamma^2 L^3\beta K}{2n}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|z_i^k-\bar{x}^k\right\|^2\right].
\end{aligned}\tag{35}$$

Rearranging terms in the above inequality, we can obtain

$$\begin{aligned}
&\left(\frac{\gamma}{2}-\frac{3\gamma^2 L\beta K}{2n}\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]+\frac{\gamma\left(1-\gamma L\right)}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i\left(z_i^k\right)\right\|^2\right]\\
\leqslant& f\left(\bar{x}^0\right)-f^*+\left(\frac{\gamma L^2}{2}+\frac{3\gamma^2 L^3\beta K}{2n}\right)\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|z_i^k-\bar{x}^k\right\|^2\right]+\frac{\gamma^2 LK}{2n}\zeta^2\\
&+\frac{\gamma^2 L\beta K^2}{2n}\left(\zeta^2+3b^2\right).
\end{aligned}\tag{36}$$

Similarly, substituting (34) into (26), we have

$$\begin{aligned}
&\left(1-\frac{9\gamma^2 L^2C^2}{\left(1-q\right)^2}\right)\sum_{k=0}^{K-1}\cdot\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|z_i^k-\bar{x}^k\right\|^2\right]\\
\leqslant&\frac{9\gamma^2C^2}{\left(1-q\right)^2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]+\left(1+\beta K\right)K\cdot\frac{3\gamma^2C^2}{\left(1-q\right)^2}\left(\zeta^2+3b^2\right)+\frac{3C^2}{\left(1-q\right)^2 n}\sum_{i=1}^{n}\left\|x_i^0\right\|^2\\
&+\frac{9\gamma^2C^2\beta K}{\left(1-q\right)^2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]+\frac{9\gamma^2L^2C^2\beta K}{\left(1-q\right)^2}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|z_i^k-\bar{x}^k\right\|^2\right].
\end{aligned}$$

Rearranging terms in the above inequality, we can get

$$\begin{aligned}
&\left(1-\left(1+\beta K\right)\cdot\frac{9\gamma^2L^2C^2}{\left(1-q\right)^2}\right)\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|z_i^k-\bar{x}^k\right\|^2\right]\\
\leqslant&\left(1+\beta K\right)\cdot\frac{9\gamma^2C^2}{\left(1-q\right)^2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]+\left(1+\beta K\right)K\cdot\frac{3\gamma^2C^2}{\left(1-q\right)^2}\left(\zeta^2+3b^2\right)\\
&+\frac{3C^2}{\left(1-q\right)^2 n}\sum_{i=1}^{n}\left\|x_i^0\right\|^2.
\end{aligned}\tag{37}$$

Substituting $K=\frac{1}{\beta}\overset{(6)}{=}\frac{J^2\epsilon^2}{dc_2^2\log\left(\frac{1}{\delta}\right)}$ into (36) and (37), those two inequalities become now

$$\begin{aligned}
&\left(\frac{1}{2}-\frac{3\gamma L}{2n}\right)\cdot\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]+\frac{1-\gamma L}{2K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i\left(z_i^k\right)\right\|^2\right]\\
\leqslant&\frac{f\left(\bar{x}^0\right)-f^*}{\gamma K}+\left(\frac{L^2}{2}+\frac{3\gamma L^3}{2n}\right)\cdot\frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|z_i^k-\bar{x}^k\right\|^2\right]+\frac{\gamma L}{2n}\left(2\zeta^2+3b^2\right)
\end{aligned}\tag{38}$$

and

$$\begin{aligned}&\left(1-\frac{18\gamma^2L^2C^2}{(1-q)^2}\right)\cdot\frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|z_i^k-\bar{x}^k\right\|^2\right]\\\leqslant&\frac{18\gamma^2C^2}{(1-q)^2}\cdot\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]+\frac{6\gamma^2C^2\left(\zeta^2+3b^2\right)}{(1-q)^2}+\frac{3C^2}{(1-q)^2nK}\sum_{i=1}^{n}\left\|x_i^0\right\|^2.\end{aligned}\tag{39}$$

If γ satisfies

$$\gamma\leqslant\frac{1-q}{6LC},\tag{40}$$

(39) becomes

$$\begin{aligned}\frac{1}{2K}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|z_i^k-\bar{x}^k\right\|^2\right]\leqslant&\frac{18\gamma^2C^2}{(1-q)^2}\cdot\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]\\&+\frac{6\gamma^2C^2\left(\zeta^2+3b^2\right)}{(1-q)^2}+\frac{3C^2}{(1-q)^2nK}\sum_{i=1}^{n}\left\|x_i^0\right\|^2.\end{aligned}\tag{41}$$

Substituting (41) into (38), we have

$$\begin{aligned}&\left(\frac{1}{2}-\frac{3\gamma L}{2n}-\frac{18\gamma^2L^2C^2}{(1-q)^2}-\frac{54\gamma^3L^3C^2}{(1-q)^2n}\right)\cdot\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]\\&+\frac{1-\gamma L}{2}\cdot\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i\left(z_i^k\right)\right\|^2\right]\\\leqslant&\frac{f\left(\bar{x}^0\right)-f^*}{\gamma K}+\left(L^2+\frac{3\gamma L^3}{n}\right)\left(\frac{6\gamma^2C^2\left(\zeta^2+3b^2\right)}{(1-q)^2}+\frac{3C^2}{(1-q)^2nK}\sum_{i=1}^{n}\left\|x_i^0\right\|^2\right)\\&+\frac{\gamma L}{2n}\left(2\zeta^2+3b^2\right).\end{aligned}\tag{42}$$

If γ further satisfies

$$\gamma\leqslant\min\left\{\frac{(1-q)^2}{24nLC^2},\frac{n}{12L},\frac{1}{L}\right\},\tag{43}$$

for (42), we have

$$\begin{aligned}&\frac{1}{4}\cdot\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]\\\leqslant&\left(\frac{1}{2}-\frac{3\gamma L}{2n}-\frac{18\gamma^2L^2C^2}{(1-q)^2}-\frac{54\gamma^3L^3C^2}{(1-q)^2n}\right)\cdot\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]\\&+\frac{1-\gamma L}{2}\cdot\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i\left(z_i^k\right)\right\|^2\right]\\\leqslant&\frac{f\left(\bar{x}^0\right)-f^*}{\gamma K}+\frac{1}{K}\cdot\frac{6L^2C^2}{(1-q)^2n}\sum_{i=1}^{n}\left\|x_i^0\right\|^2+\frac{12\gamma^2L^2C^2\left(\zeta^2+3b^2\right)}{(1-q)^2}\\&+\frac{\gamma L}{2n}\left(2\zeta^2+3b^2\right),\end{aligned}\tag{44}$$

i.e.,

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]\leqslant&\frac{4\left(f\left(\bar{x}^0\right)-f^*\right)}{\gamma K}+\frac{1}{K}\cdot\frac{24L^2C^2}{(1-q)^2n}\sum_{i=1}^{n}\left\|x_i^0\right\|^2\\&+\frac{48\gamma^2L^2C^2\left(\zeta^2+3b^2\right)}{(1-q)^2}+\frac{2\gamma L}{n}\left(2\zeta^2+3b^2\right).\end{aligned}\tag{45}$$

By now, the step size γ need to satisfy (40) and (43), i.e.,

$$\gamma \leqslant \underbrace{\min\left\{\frac{(1-q)^2}{24nLC^2}, \frac{1-q}{6LC}, \frac{n}{12L}, \frac{1}{L}\right\}}_{\triangleq\hat{\gamma}(C,q)}. \tag{46}$$

Now we set the step size γ as

$$\gamma = \frac{1}{\sqrt{K/n}+\hat{\gamma}(C,q)^{-1}} = \frac{1}{\sqrt{\frac{1}{n\beta}}+\hat{\gamma}(C,q)^{-1}} \overset{(6)}{=} \frac{1}{\frac{J\epsilon}{c_2\sqrt{nd\log\left(\frac{1}{\delta}\right)}}+\hat{\gamma}(C,q)^{-1}}, \tag{47}$$

then (45) can be further bounded as

$$\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right] \\
\leqslant&\frac{4\left(f\left(\bar{x}^0\right)-f^*\right)}{\gamma K}+\frac{1}{K}\cdot\frac{24L^2C^2}{(1-q)^2 n}\sum_{i=1}^{n}\left\|x_i^0\right\|^2 \\
&+\frac{48\gamma^2L^2C^2\left(\zeta^2+3b^2\right)}{(1-q)^2}+\frac{2\gamma L}{n}\left(2\zeta^2+3b^2\right) \\
\overset{(47)}{\leqslant}&\frac{4\left(f\left(\bar{x}^0\right)-f^*\right)}{\sqrt{nK}}+\frac{2L\left(2\zeta^2+3b^2\right)}{\sqrt{nK}}+\frac{4\left(f\left(\bar{x}^0\right)-f^*\right)}{\hat{\gamma}(C,q)K} \\
&+\frac{1}{K}\cdot\frac{24L^2C^2}{(1-q)^2 n}\sum_{i=1}^{n}\left\|x_i^0\right\|^2+\frac{1}{K}\cdot\frac{48nL^2C^2\left(\zeta^2+3b^2\right)}{(1-q)^2} \\
=&\frac{4\left(f\left(\bar{x}^0\right)-f^*\right)+2L\left(2\zeta^2+3b^2\right)}{\sqrt{nK}} \\
&+\frac{1}{K}\left(\frac{24L^2C^2}{(1-q)^2 n}\sum_{i=1}^{n}\left\|x_i^0\right\|^2+\frac{48nL^2C^2\left(\zeta^2+3b^2\right)}{(1-q)^2}+\frac{4\left(f\left(\bar{x}^0\right)-f^*\right)}{\hat{\gamma}(C,q)}\right).
\end{aligned} \tag{48}$$

Knowing that K is chosen as

$$K = \frac{1}{\beta} = \frac{J^2\epsilon^2}{dc_2^2\log\left(\frac{1}{\delta}\right)}, \tag{49}$$

(48) becomes

$$\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right] \\
\overset{(46)}{\leqslant}&\frac{c_2\sqrt{d\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}J\epsilon}\cdot\left[4\left(f\left(\bar{x}^0\right)-f^*\right)+2L\left(2\zeta^2+3b^2\right)\right] \\
&+\frac{c_2^2 d\log\left(\frac{1}{\delta}\right)}{J^2\epsilon^2}\cdot\left[\frac{24L^2C^2}{(1-q)^2 n}\sum_{i=1}^{n}\left\|x_i^0\right\|^2+\frac{48nL^2C^2\left(\zeta^2+3b^2\right)}{(1-q)^2}\right] \\
&+\frac{c_2^2 d\log\left(\frac{1}{\delta}\right)}{J^2\epsilon^2}\cdot 4\left(f\left(\bar{x}^0\right)-f^*\right)\cdot\max\left\{\frac{24nLC^2}{(1-q)^2}, \frac{6LC}{1-q}, \frac{12L}{n}, L\right\}.
\end{aligned} \tag{50}$$

According to the mild assumption

$$J \geqslant \frac{n^{\frac{3}{2}}c_2\sqrt{d\log\left(\frac{1}{\delta}\right)}}{\epsilon}, \tag{51}$$

(50) can be further bounded as

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right] \\
\leqslant&\frac{c_2\sqrt{d\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}J\epsilon}\cdot\left[4\left(f\left(\bar{x}^0\right)-f^*\right)+2L\left(2\zeta^2+3b^2\right)\right] \\
&+\frac{c_2\sqrt{d\log\left(\frac{1}{\delta}\right)}}{J\epsilon}\cdot\frac{1}{n^{3/2}}\cdot\left[\frac{24L^2C^2}{\left(1-q\right)^2n}\sum_{i=1}^{n}\left\|x_i^0\right\|^2+\frac{48nL^2C^2\left(\zeta^2+3b^2\right)}{\left(1-q\right)^2}\right] \\
&+\frac{c_2\sqrt{d\log\left(\frac{1}{\delta}\right)}}{J\epsilon}\cdot\frac{1}{n^{3/2}}\cdot 4\left(f\left(\bar{x}^0\right)-f^*\right)\cdot\max\left\{\frac{24nLC^2}{\left(1-q\right)^2},\frac{6LC}{1-q},\frac{12L}{n},L\right\} \\
\leqslant&\frac{c_2\sqrt{d\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}J\epsilon}\cdot\left[4\left(f\left(\bar{x}^0\right)-f^*\right)+2L\left(2\zeta^2+3b^2\right)\right] \\
&+\frac{c_2\sqrt{d\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}J\epsilon}\cdot\left[\frac{24L^2C^2}{\left(1-q\right)^2n}\sum_{i=1}^{n}\left\|x_i^0\right\|^2+\frac{48L^2C^2\left(\zeta^2+3b^2\right)}{\left(1-q\right)^2}\right] \\
&+\frac{c_2\sqrt{d\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}J\epsilon}\cdot 4\left(f\left(\bar{x}^0\right)-f^*\right)\cdot\max\left\{\frac{24LC^2}{\left(1-q\right)^2},\frac{6LC}{1-q},13L\right\} \\
=&\mathcal{O}\left(\frac{\sqrt{d\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}J\epsilon}\right),
\end{aligned}
$$

which completes the proof of Theorem 1.

A.4 Proof of Proposition 3

Now we provide the complete proof of Proposition 3.

If the step size γ satisfies

$$
\gamma\leqslant\frac{1-q}{\sqrt{18}LC}, \tag{52}
$$

(26) becomes

$$
\begin{aligned}
&\frac{1}{2}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|z_i^k-\bar{x}^k\right\|^2\right] \\
\leqslant&\frac{9\gamma^2C^2}{\left(1-q\right)^2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]+K\cdot\frac{3\gamma^2C^2}{\left(1-q\right)^2}\left(\zeta^2+3b^2\right) \\
&+\frac{3C^2}{\left(1-q\right)^2n}\sum_{i=1}^{n}\left\|x_i^0\right\|^2+\frac{3\gamma^2C^2\beta K}{\left(1-q\right)^2}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla f_i\left(z_i^k;\xi_i^k\right)\right\|^2\right].
\end{aligned} \tag{53}
$$

Substituting (53) into (18), we obtain

$$
\begin{aligned}
&\left(\frac{\gamma}{2}-\frac{9\gamma^3L^2C^2}{\left(1-q\right)^2}\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]+\frac{\gamma\left(1-\gamma L\right)}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i\left(z_i^k\right)\right\|^2\right] \\
\leqslant&f\left(\bar{x}^0\right)-f^*+\frac{\gamma^2LK}{2n}\zeta^2+K\cdot\frac{3\gamma^3L^2C^2}{\left(1-q\right)^2}\left(\zeta^2+3b^2\right)+\frac{3\gamma L^2C^2}{\left(1-q\right)^2n}\sum_{i=1}^{n}\left\|x_i^0\right\|^2 \\
&+\left(\frac{3\gamma^3L^2C^2\beta K}{\left(1-q\right)^2}+\frac{\gamma^2L\beta K}{2n}\right)\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla f_i\left(z_i^k;\xi_i^k\right)\right\|^2\right].
\end{aligned} \tag{54}
$$

Dividing by γK on both sides of the above inequality yields

$$\begin{aligned}
&\left(\frac{1}{2}-\frac{9\gamma^2L^2C^2}{(1-q)^2}\right)\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]+\frac{1-\gamma L}{2K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i\left(z_i^k\right)\right\|^2\right]\\
\leqslant&\frac{f\left(\bar{x}^0\right)-f^*}{\gamma K}+\frac{\gamma L}{2n}\zeta^2+\frac{3\gamma^2L^2C^2}{(1-q)^2}\left(\zeta^2+3b^2\right)+\frac{3L^2C^2}{(1-q)^2nK}\sum_{i=1}^{n}\left\|x_i^0\right\|^2\\
&+\left(\frac{3\gamma^2L^2C^2\beta}{(1-q)^2}+\frac{\gamma L\beta}{2n}\right)\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla f_i\left(z_i^k;\xi_i^k\right)\right\|^2\right].
\end{aligned}\tag{55}$$

If the step size γ satisfies

$$\gamma\leqslant\min\left\{\frac{1-q}{6LC},\frac{1}{L}\right\},\tag{56}$$

(55) becomes now

$$\begin{aligned}
&\frac{1}{4}\cdot\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]\\
\leqslant&\frac{f\left(\bar{x}^0\right)-f^*}{\gamma K}+\frac{\gamma L}{2n}\zeta^2+\frac{3\gamma^2L^2C^2}{(1-q)^2}\left(\zeta^2+3b^2\right)+\frac{3L^2C^2}{(1-q)^2nK}\sum_{i=1}^{n}\left\|x_i^0\right\|^2\\
&+\left(\frac{3\gamma^2L^2C^2\beta}{(1-q)^2}+\frac{\gamma L\beta}{2n}\right)\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla f_i\left(z_i^k;\xi_i^k\right)\right\|^2\right].
\end{aligned}\tag{57}$$

By now, the step size γ needs to satisfy (52) and (56), i.e.,

$$\gamma\leqslant\min\left\{\frac{1-q}{6LC},\frac{1}{L}\right\}.\tag{58}$$

We thus complete the proof of Proposition 3.

A.5 Proof of Proposition 4

The proof of Proposition 4 shares the similarity with that of Proposition 3, except for the processing of privacy noise-related terms.

According to (83) in Algorithm 2 and (5) in Theorem 2, we have the variance of injected Gaussian noise for each node i at each iteration k as follows

$$\mathbb{E}\left[\left\|N_i^k\right\|^2\right]=d\sigma^2G^2=\underbrace{\frac{dc_2^2\log\left(\frac{1}{\delta}\right)}{J^2\epsilon^2}}_{\beta}\cdot KG^2.\tag{59}$$

Therefore, A_3 in (22) becomes

$$A_3=\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}N_i^k\right\|^2\right]=\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\left\|N_i^k\right\|^2\right]\overset{(59)}{=}\frac{\beta K}{n^2}\sum_{i=1}^{n}G^2.\tag{60}$$

Following the proof of Lemma 3 with the above new A_3, we have

$$\begin{aligned}
&\frac{\gamma}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]+\frac{\gamma\left(1-\gamma L\right)}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i\left(z_i^k\right)\right\|^2\right]\\
\leqslant&f\left(\bar{x}^0\right)-f^*+\frac{\gamma L^2}{2}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|z_i^k-\bar{x}^k\right\|^2\right]+\frac{\gamma^2LK}{2n}\zeta^2+\frac{\gamma^2L\beta K}{2n}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}G^2.
\end{aligned}\tag{61}$$

Following the proof of Lemma 5 with the new noise variance $\mathbb{E}\left[\left\|N_i^k\right\|^2\right]$ (c.f., (59)), we have

$$\begin{aligned}
&\left(1-\frac{9\gamma^2L^2C^2}{(1-q)^2}\right)\sum_{k=0}^{K-1}\cdot\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|z_i^k-\bar{x}^k\right\|^2\right]\\
\leqslant&\frac{9\gamma^2C^2}{(1-q)^2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]+K\cdot\frac{3\gamma^2C^2}{(1-q)^2}\left(\zeta^2+3b^2\right)\\
&+\frac{3C^2}{(1-q)^2n}\sum_{i=1}^{n}\left\|x_i^0\right\|^2+\frac{3\gamma^2C^2\beta K}{(1-q)^2}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}G^2.
\end{aligned}\tag{62}$$

With (61) and (62) being the alternatives of (18) and (26) respectively, following the proof of Proposition 3 (c.f., (52)-(58)), we can easily obtain

$$\begin{aligned}
&\frac{1}{4}\cdot\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\bar{x}^k\right)\right\|^2\right]\\
\leqslant&\frac{f\left(\bar{x}^0\right)-f^*}{\gamma K}+\frac{\gamma L}{2n}\zeta^2+\frac{3\gamma^2L^2C^2}{(1-q)^2}\left(\zeta^2+3b^2\right)+\frac{3L^2C^2}{(1-q)^2nK}\sum_{i=1}^{n}\left\|x_i^0\right\|^2\\
&+\left(\frac{3\gamma^2L^2C^2\beta}{(1-q)^2}+\frac{\gamma L\beta}{2n}\right)\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}G^2,
\end{aligned}\tag{63}$$

and the step size γ need to satisfy

$$\gamma\leqslant\min\left\{\frac{1-q}{6LC},\frac{1}{L}\right\}.\tag{64}$$

We thus complete the proof of Proposition 4.

A.6 Derivation of utility bound for baseline centralized DP-SGD

We first make the following blanket assumptions for our theoretical analysis of centralized DP-SGD.

Assumption 7 (L-smooth)**.** *For any model parameter x and y, we have*

$$\|\nabla f(x)-\nabla f(y)\|\leqslant L\|x-y\|.\tag{65}$$

Assumption 8 (Unbiased gradient)**.** *For any model parameter x^k, we have*

$$\mathbb{E}\left[\nabla f\left(x^k;\xi^k\right)\right]=\nabla f\left(x^k\right).\tag{66}$$

Assumption 9 (Bounded variance)**.** *For any model parameter x^k, we have*

$$\mathbb{E}\left[\left\|\nabla f\left(x^k;\xi^k\right)-\nabla f\left(x^k\right)\right\|^2\right]\leqslant\zeta^2.\tag{67}$$

Assumption 10 (Bounded gradient)**.** *For any $x\in\mathbb{R}^d$ and $\xi\in\{1,2,...,J\}$, there exists finite positive constant G such that*

$$\|\nabla f(x;\xi)\|\leqslant G.\tag{68}$$

The update of centralized differentially private SGD is:

$$x^{k+1}=x^k-\gamma\left(\nabla f\left(x^k;\xi^k\right)+n^k\right),\tag{69}$$

where the randomized noise n^k is drawn from the Gaussian distribution

$$n^k\sim\mathcal{N}\left(0,\sigma^2G^2\mathbb{I}_d\right),\tag{70}$$

and σ is defined in Proposition 2.

With the above, we have

$$\mathbb{E}\left[\left\|n^k\right\|^2\right] = d \cdot \sigma^2 G^2 = \underbrace{\frac{dc_2^2 \log\left(\frac{1}{\delta}\right)}{J^2\epsilon^2}}_{\beta} \cdot G^2 K. \tag{71}$$

Applying the descent lemma to f at x^k and x^{k+1}, we have

$$\begin{aligned} f\left(x^{k+1}\right) &\overset{(65)}{\leqslant} f\left(x^k\right) + \left\langle \nabla f\left(x^k\right), x^{k+1} - x^k \right\rangle + \frac{L}{2}\left\|x^{k+1} - x^k\right\|^2 \\ &\overset{(69)}{=} f\left(x^k\right) - \gamma\left\langle \nabla f\left(x^k\right), \nabla f\left(x^k;\xi^k\right) + n^k \right\rangle + \frac{\gamma^2 L}{2}\left\|\nabla f\left(x^k;\xi^k\right) + n^k\right\|^2 \\ &= f\left(x^k\right) - \gamma\left\langle \nabla f\left(x^k\right), \nabla f\left(x^k;\xi^k\right) + n^k \right\rangle + \frac{\gamma^2 L}{2}\left\|\nabla f\left(x^k;\xi^k\right)\right\|^2 \\ &\quad + \frac{\gamma^2 L}{2}\left\|n^k\right\|^2 + \gamma^2 L\left\langle \nabla f\left(x^k;\xi^k\right), n^k \right\rangle \end{aligned} \tag{72}$$

Taking expectation of both sides of the above inequality conditioned on x^k, we have

$$\begin{aligned} &\mathbb{E}\left[f\left(x^{k+1}\right)\middle| x^k\right] \\ \leqslant& f\left(x^k\right) - \gamma\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\gamma^2 L}{2}\mathbb{E}\left[\left\|\nabla f\left(x^k;\xi^k\right)\right\|^2\middle| x^k\right] + \frac{\gamma^2 L}{2}\mathbb{E}\left[\left\|n^k\right\|^2\middle| x^k\right] \\ \overset{(a)}{=}& f\left(x^k\right) - \gamma\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\gamma^2 L}{2}\mathbb{E}\left[\left\|\nabla f\left(x^k;\xi^k\right) - \nabla f\left(x^k\right) + \nabla f\left(x^k\right)\right\|^2\middle| x^k\right] \\ &+ \frac{\gamma^2 L}{2}\cdot\beta G^2 K \\ =& f\left(x^k\right) - \gamma\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\gamma^2 L}{2}\mathbb{E}\left[\left\|\nabla f\left(x^k;\xi^k\right) - \nabla f\left(x^k\right)\right\|^2\middle| x^k\right] \\ &+ \frac{\gamma^2 L}{2}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\gamma^2 L}{2}\cdot\beta G^2 K \\ \overset{(b)}{\leqslant}& f\left(x^k\right) - \gamma\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\gamma^2 L}{2}\left\|\nabla f\left(x^k\right)\right\|^2 + \frac{\gamma^2 L}{2}\zeta^2 + \frac{\gamma^2 L}{2}\cdot\beta G^2 K. \end{aligned} \tag{73}$$

where we used (71) in (a) and (67) in (b).

Taking total expectation on both sides of the above inequality, we have

$$\mathbb{E}\left[f\left(x^{k+1}\right)\right] \leqslant \mathbb{E}\left[f\left(x^k\right)\right] - \gamma\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] + \frac{\gamma^2 L}{2}\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] + \frac{\gamma^2 L}{2}\zeta^2 + \frac{\gamma^2 L}{2}\cdot\beta G^2 K. \tag{74}$$

Summing (74) from $k = 0$ to $K - 1$, we have

$$\left(\gamma - \frac{\gamma^2 L}{2}\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] \leqslant f\left(x^0\right) - f^* + \frac{\gamma^2 L}{2}\zeta^2 \cdot K + \frac{\gamma^2 L}{2}\cdot\beta G^2 K^2. \tag{75}$$

Dividing γK on both sides, we have

$$\left(1 - \frac{\gamma L}{2}\right)\cdot\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] \leqslant \frac{f\left(x^0\right) - f^*}{\gamma K} + \frac{\gamma L}{2}\zeta^2 + \frac{LG^2\beta}{2}\cdot\gamma K. \tag{76}$$

If γ satisfies

$$\gamma \leqslant \frac{1}{L}, \tag{77}$$

(76) becomes

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] \leqslant \frac{2\left(f\left(x^0\right) - f^*\right)}{\gamma K} + \gamma L\zeta^2 + LG^2\beta\cdot\gamma K. \tag{78}$$

Now we set γ as

$$\gamma = \frac{1}{\sqrt{K} + L}, \tag{79}$$

then (78) becomes

$$\begin{aligned} \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] \leqslant & \frac{2\left(f\left(x^0\right)-f^*\right)}{\sqrt{K}} + \frac{2L\left(f\left(x^0\right)-f^*\right)}{K} + \frac{L\zeta^2}{\sqrt{K}} + LG^2\beta\sqrt{K} \\ \leqslant & \frac{2\left(1+L\right)\left(f\left(x^0\right)-f^*\right)+L\zeta^2}{\sqrt{K}} + LG^2\beta\sqrt{K}. \end{aligned} \tag{80}$$

Regarding the right hand side of (80) as a function of K, we can obtain the optimal value of K by minimizing this function, and the optimal value of K is

$$K = \frac{2\left(1+L\right)\left(f\left(x^0\right)-f^*\right)+L\zeta^2}{LG^2\beta} \overset{(71)}{=} \frac{2\left(1+L\right)\left(f\left(x^0\right)-f^*\right)+L\zeta^2}{LG^2} \cdot \frac{J^2\epsilon^2}{dc_2^2\log\left(\frac{1}{\delta}\right)}. \tag{81}$$

and the utility (80) becomes

$$\begin{aligned} \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] &\leqslant 2G\sqrt{2\left(1+L\right)L\left(f\left(x^0\right)-f^*\right)+L^2\zeta^2}\cdot\sqrt{\beta} \\ &= 2c_2G\sqrt{\left(2\left(1+L\right)L\left(f\left(x^0\right)-f^*\right)+L^2\zeta^2\right)}\cdot\frac{\sqrt{d\log\left(\frac{1}{\delta}\right)}}{J\epsilon} \\ &= \mathcal{O}\left(\frac{\sqrt{d\log\left(\frac{1}{\delta}\right)}}{J\epsilon}\right). \end{aligned} \tag{82}$$

B MISSING PSEUDOCODES OF ALGORITHMS

Algorithm 2 Differentially Private Decentralized Learning with Constant Gaussian Noise (**ConstD^{2}P**)

1: **Initialization:** $x_i^0 = z_i^0 \in \mathbb{R}^d$, $w_i^0 = 1$, learning rate $\gamma > 0$, total number of iterations K and privacy budget (ϵ, δ).
2: **for** $k = 0, 1, 2, ..., K-1$, at node i, **do**
3: Randomly samples a local training data ξ_i^k with the sampling probability $\frac{1}{J}$;
4: Computes stochastic gradient at z_i^k: $\nabla f_i(z_i^k; \xi_i^k)$;
5: Draws randomized noise N_i^k from the Gaussian distribution:

$$N_i^k \sim \mathcal{N}\left(0, \sigma^2 G^2 \mathbb{I}_d\right), \tag{83}$$

where σ is defined in Proposition 2, and G is defined at (12);
6: Differentially private local SGD:

$$x_i^{k+\frac{1}{2}} = x_i^k - \gamma(\nabla f_i(z_i^k; \xi_i^k) + N_i^k). \tag{84}$$

7: Follows the 7^{th}-11^{th} steps of Algorithm 1.
8: **end for**

Algorithm 3 Differentially Private Decentralized Learning with Fixed Gradient Clipping Bound (**ClipD^{2}P**)

1: **Initialization:** $x_i^0 = z_i^0 \in \mathbb{R}^d$, $w_i^0 = 1$, learning rate $\gamma > 0$, total number of iterations K, privacy budget (ϵ, δ) and fixed clipping bound C.
2: **for** $k = 0, 1, 2, ..., K-1$, at node i, **do**
3: Randomly samples a local training data ξ_i^k with the sampling probability $\frac{1}{J}$;
4: Computes stochastic gradient at z_i^k: $\nabla f_i(z_i^k; \xi_i^k)$;
5: Clips stochastic gradient by:

$$g_i^k = \frac{\nabla f_i(z_i^k; \xi_i^k)}{\max\left\{1, \frac{\|\nabla f_i(z_i^k; \xi_i^k)\|}{C}\right\}}; \tag{85}$$

6: Draws randomized noise N_i^k from the Gaussian distribution:

$$N_i^k \sim \mathcal{N}\left(0, \sigma^2 C^2 \mathbb{I}_d\right), \tag{86}$$

where σ is defined in Proposition 2;
7: Differentially private local SGD:

$$x_i^{k+\frac{1}{2}} = x_i^k - \gamma(g_i^k + N_i^k); \tag{87}$$

8: Follows the 7^{th}-11^{th} steps of Algorithm 1.
9: **end for**

C Missing Definition of Time-varying Directed Exponential Graph

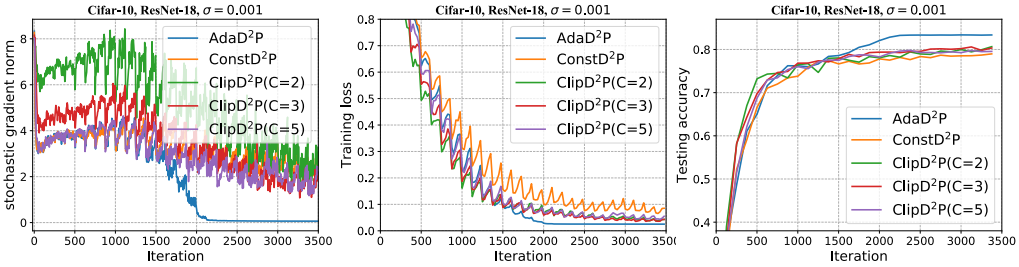
We supplement the definition of time-varying directed exponential graph (Assran et al., 2019) we missed in the main text. Specifically, n nodes are ordered sequentially with their rank 0,1...,$n-1$, and each node has out-neighbours that are 2^0,2^1,...,$2^{\lfloor \log_2(n-1) \rfloor}$ hops away. Each node cycles through these out-neighbours, and only transmits messages to one of its out-neighbours at each iteration. For example, each node sends message to its 2^0-hop out-neighbour at iteration k, and to its 2^1-hop out-neighbour at iteration $k+1$, and so on. The above procedure will be repeated within the list of out-neighbours. Note that each node only sends and receives a single message at each iteration.

D Additional Experiments Results

In this section, we present additional experimental results.

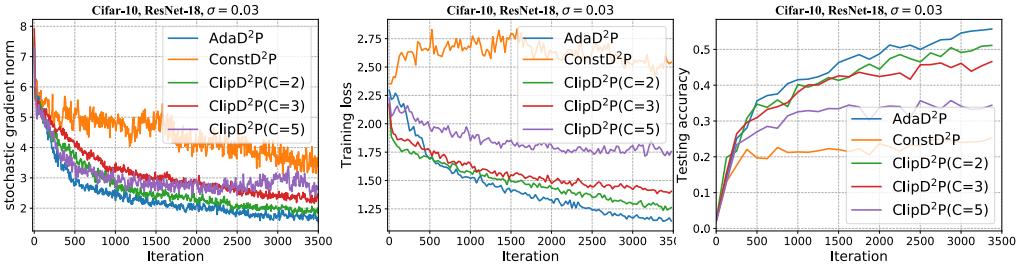
Deep CNN ResNet-18 training. For the task of training deep CNN model ResNet-18 on Cifar-10 dataset, we run AdaD^{2}P, ConstD^{2}P and ClipD^{2}P for 3500 iterations and compare their performance under the same noise scale σ selected from the set $\{0.001, 0.03\}$. It can be observed form Figures 5 and 6 that AdaD^{2}P always outperforms other algorithms ConstD^{2}P and ClipD^{2}P which both employ fixed-level DP noise under the same level of privacy protection, in terms of the convergence of gradient norm, training loss and model accuracy. These experiments further confirm the superiority of our proposed AdaD^{2}P employing adaptive noise level, compared to its counterparts employing fixed-level noise.

Shallow 2-layer neural network training. For the task of training shallow 2-layer neural network on Mnist dataset, we run AdaD^{2}P, ConstD^{2}P and ClipD^{2}P for 2200 iterations and compare their convergence performance under the same noise scale σ selected from the set $\{0.03, 0.04, 0.06\}$. The experiments presented in Figures 7, 8 and 9 demonstrate that: under the same level of privacy protection (same σ), AdaD^{2}P achieves superior model accuracy compared to its counterparts that employ fixed-level noise, which verifies the effectiveness of our adaptive noise mechanism.



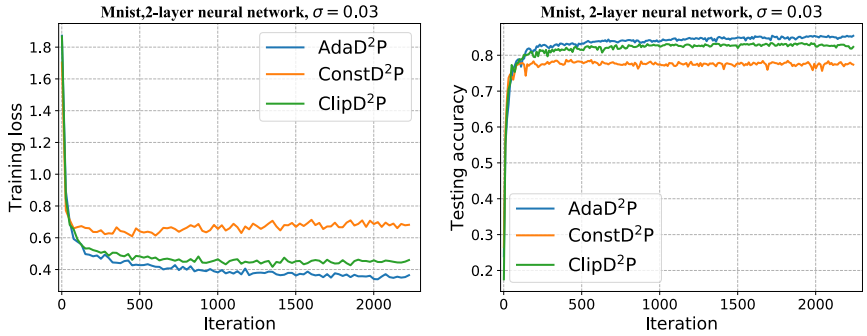
(a) (b) (c)

Figure 5: Performance comparison of training ResNet-18 for AdaD^{2}P with ConstD^{2}P and ClipD^{2}P under the same noise scale $\sigma = 0.001$.



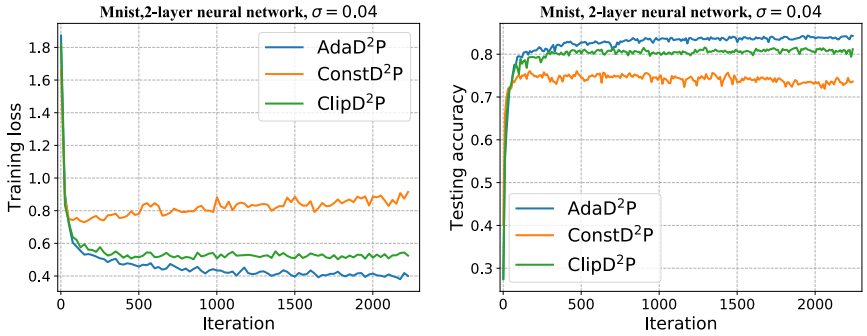
(a) (b) (c)

Figure 6: Performance comparison of training ResNet-18 for AdaD^{2}P with ConstD^{2}P and ClipD^{2}P under the same noise scale $\sigma = 0.03$.



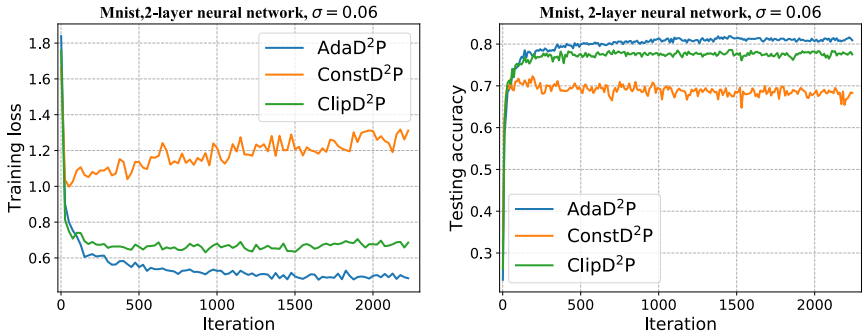
(a) (b)

Figure 7: Performance comparison of training 2-layer neural network for AdaD^{2}P with ConstD^{2}P and ClipD^{2}P under the same noise scale $\sigma = 0.03$.



(a) (b)

Figure 8: Performance comparison of training 2-layer neural network for AdaD^{2}P with ConstD^{2}P and ClipD^{2}P under the same noise scale $\sigma = 0.04$.



(a) (b)

Figure 9: Performance comparison of training 2-layer neural network for AdaD^{2}P with ConstD^{2}P and ClipD^{2}P under the same noise scale $\sigma = 0.06$.