A GENERAL DIFFERENTIALLY PRIVATE LEARNING FRAMEWORK FOR DECENTRALIZED DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Decentralized consensus learning has been hugely successful, which minimizes a finite sum of expected objective functions over a network of nodes. However, the local communication across neighbouring nodes in the network may lead to the leakage of private information. To address this challenge, we propose a general differentially private (DP) learning framework for decentralized data that applies to many non-smooth learning problems. We show that the proposed algorithm retains the performance guarantee in terms of stability, generalization, and finite sample performance. We investigate the impact of local privacy-preserving computation on the global DP guarantee. Further, we extend the discussion by adopting a new class of noise-adding DP mechanisms based on generalized Gaussian distributions to improve the utility-privacy trade-offs. Our numerical results demonstrate the effectiveness of our algorithm and its better performance over the state-of-the-art baseline methods in various decentralized settings.

1 INTRODUCTION

Decentralized consensus computation is a form of distributed learning to achieve a consensus target that relies on communication between nodes through a network diagram without a centralized server. On large-scale optimization tasks involving distributed training samples, decentralized algorithms consider a low-communication-overhead alternative network structure to learn a global optimal solution in a parallel and distributed fashion, while, allowing multiple agents to keep their own datasets unexposed. Existing approaches to decentralized learning problems mainly consist of subgradient-based algorithms (Nedic et al., 2008), Alternating Direction Method of Multipliers (ADMM) based algorithms (Boyd et al., 2011; Shi et al., 2014), and the composite of subgradient descent and ADMM (Bianchi et al., 2014), which have shown superior performance on many application domains, such as distributed sensing in wireless sensor networks, healthcare, and multi-agent robotic systems (Rabbat & Nowak, 2004; Bullo et al., 2009; Sayed et al., 2014; Yu et al., 2021a). However, the lack of adaptation of these existing algorithms has limited their application: most of the existing works require the objective functions to be smooth or strongly convex. To fill the gap, we provide a general framework of decentralized algorithms, stochastic Decentralized Krasnosel'skiĭ–Mann (D-KM) iteration, that comprises diverse decentralized algorithms with straightforward, high-performance implementation. Depending on various iterative operations, D-KM offers intuitive interfaces to implement a spectrum of decentralized algorithms, which is adaptive to objectives with different functional properties (e.g., non-smoothness).

Although the decentralized algorithm obtains the consensus solution through local computation, namely, each agent solves its own problems independently and locally with only intermediate parameters that need to be shared. Past experience has demonstrated the possibility of privacy leakage in such an iterative process: the attacker can recover the sensitive information from shared parameters as pointed in Shokri et al. (2017),Fredrikson et al. (2015). One defence procedure is to adopt a private variant of the learning algorithm using Differential Privacy (DP) to secure the iterative process. While the developed DP-based methods mainly focus on standalone systems or centralized learning (Song et al., 2013; Li et al., 2022), how to ensure data privacy in decentralized systems is still an open question due to their complex communication typologies and distinct learning characteristics. Moreover, most of the existing variants of learning algorithms consider simply adding Gaussian noise to iterates to ensure differential privacy where the noise could be extremely large (Farokhi, 2022). However, enforcing and optimizing DP protection with unbounded noise ad-

dition could severely affect the learning efficiency and degrade the performance of the trained model under differential privacy guarantee.

To mitigate privacy risks in decentralized settings, in this paper, we propose a general decentralized privacy-preserving algorithm, DP-KM. Specifically, the proposed algorithm enforces differential privacy protection of communication, and it can be applicable to much of the existing work on optimization and consensus computation with mild assumptions on objectives. We validate the privacy and effectiveness of the proposed approach through a rigorous theoretical analysis, which shows the performance guarantees of the proposed method. To enable the D-KM to be noise-resilient, we further propose a class of truncated generalized Gaussian noise-adding mechanisms. The key feature of the DP-KM with the mechanisms is to leverage the D-KM iteration with truncated generalized Gaussian noise in the iteration procedure to reduce the noise scale and achieve higher utility under differential privacy guarantee for general objective functions. Empirically, we conduct comprehensive experiments to demonstrate that our approach outperforms prior works in various decentralized settings. The main contributions of this paper are summarized as follows.

1. Leveraging the decentralized learning procedure to train models under privacy constraints, we propose a general differentially private decentralized algorithm, DP-KM.
2. We provide a rigorous sensitivity analysis of D-KM and establish the generalization and optimization error bound of DP-KM under mild conditions. Additionally, we investigate the impact of the local privacy-preserving computation on global differential privacy.
3. We introduce a new class of truncated generalized Gaussian mechanism that can achieve higher utility under differential privacy guarantee for general objective functions.
4. We conduct numerical analysis to validate the effectiveness of DP-KM with truncated Laplace and Gaussian noise in decentralized learning settings comparing with the state-of-the-art methods.

The rest of this paper is organized as follows. Section 2 introduces problem setting, formal definitions of decentralized systems, and differential privacy. The stability of D-KM is presented in Section 3. Section 4 establishes the generalization bound of the DP-KM in a decentralized setting, and the effect of local privacy-preserving computation on global differential privacy. Section 5 illustrates a new class of noise-adding mechanisms, especially, two truncated mechanisms: Truncated Laplacian Mechanism and Truncated Gaussian Mechanism, for synchronous training models. Section 6 presents our privacy analysis and the performance of our approach in various settings. We conclude this paper in Section 7.

2 Problem Statement

In this section, we first start with the problem setting. We then present preliminaries about decentralized learning schemes as well as differential privacy.

Problem setting: Consider a network with M agents, each of which holds a dataset $\Xi_m = \{\xi_{i(m)}\}_{i=1}^{N_i}$, for $m = 1, \cdots, M$, where N_i is the number of training samples in the dataset Ξ_m, $\xi_{i(m)} \in \mathbb{R}^p$ is the i-th sample stored in the m-th agent. For ease of presentation, we assume that data are evenly collected and each agent has a sample size of N.

In this paper, we consider solving a stochastic decentralized optimization approximated by its corresponding empirical risk minimization problem,

$$\widehat{L}(x) = \min_{x \in \mathbb{R}^p} \frac{1}{MN} \sum_{m=1}^{M} \sum_{i=1}^{N} \ell\left(x, \xi_{i(m)}\right), \quad \widehat{x} = \arg\min_{x \in \mathbb{R}^p} \frac{1}{MN} \sum_{m=1}^{M} \sum_{i=1}^{N} \ell\left(x, \xi_{i(m)}\right),$$

where $x \in \mathbb{R}^p$ is the target parameter, and $\ell(\cdot)$ is the loss function used to measure the quality of the trained model. Throughout, we assume the loss function is convex, closed, and proper but not necessarily differentiable. The goal of our problem is to learn a consensus parameter $\bar{x} = \frac{1}{M}\sum_{m=1}^{M} x(m)$ on M agents across a network diagram, where $x(m)$ is the solution of the local parameter on the m-th agent. We stress that, in practice, each agent operates independently and the average is only taken in the last iteration.

The decentralized optimization is associated with a given network topology that can be formulated mathematically by a mixing matrix (Alghunaim et al., 2019; Ying et al., 2021). The properties of the graph required to learn a global parameter can be summarized in Definition 1. The agent's communication with their neighbours is responsible for message passing and aggregation accordingly. To formalize the definition, we define the connected network by, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V} = \{1, \ldots, M\}$ and edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Edge $(m, l) \in \mathcal{E}$ represents the interconnection between agent m and its neighbors $l \in \mathcal{N}(m)$.

Definition 1 (Mixing Matrix) *For any given graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the mixing matrix $\mathbf{W} = [w_{m,l}] \in \mathbb{R}^{M \times M}$ is defined on the edge set $\mathcal{V}$ that satisfies: (1) if $m \neq l$ and $(m, l) \notin \mathcal{E}$, then $w_{m,l} = 0$; otherwise, $w_{m,l} > 0$; (2) $\mathbf{W} = \mathbf{W}^\top$; (3) $\mathrm{null}\{\mathbb{I} - \mathbf{W}\} = \mathrm{span}\{\mathbf{1}\}$; (4) $\mathbb{I} \succeq \mathbf{W} \succ -\mathbb{I}$.*

We remark that the mixing matrix $\mathbf{W}$ is a doubly stochastic gossip matrix that characterizes the underlying topology and the matrix is non-unique for a given graph (Ying et al., 2021; Sun et al., 2021). An essential constant that characterizes the connectivity of gossip communications is measured by the spectral gap as $1 - \lambda$ (Zhu et al., 2022), where $\lambda := \max\{|\lambda_2|, |\lambda_M|\}$, and λ_i denotes the ith largest eigenvalue of $\mathbf{W} \in \mathbb{R}^{M \times M}$. The definition of the mixing matrix implies that $0 \leq \lambda < 1$. We now give a general decentralized learning algorithm on the m-th agent.

Definition 2 (Stochastic Decentralized Krasnosel'skiĭ–Mann (D-KM)) *Suppose the training sample set $\Xi := \bigcup_{m=1}^{M} \Xi_m$ is distributed-stored in M agents with total sample size NM, where Ξ_m is a training dataset located in the m-th agent for $m = 1, \cdots, M$. We assume that $\xi_{i(m)} \sim \mathbb{P}$ with $\xi_{i(m)} \in \Xi_m$ for any m, i. For each agent, given a nonexpansive operator T, the iterative formula of the stochastic D-KM algorithm, $\mathcal{A}$, is defined as,*

$$x^{k+1}(m) = \mathcal{A}\left(x^k(m); \Xi\right) = \sum_{l \in \mathcal{N}(m)} w_{ml} x^k(l) + \alpha_k \left(T\left(x^k(m); \xi_{i_k(m)}\right) - x^k(m)\right), \quad (1)$$

where w_{ml} is the element of a given matrix $\mathbf{W}$ satisfying Definition 1, $\alpha_k \in (0, 1]$. i_k is i.i.d. variable drawn from the uniform distribution over $\{1, \cdots, N\}$ at the k-th iteration.

The form of T in Definition 2 depends on the specific algorithm we adopt. Let I be an identity operator, ∂f, ∇f be the sub-gradient and gradient of function f, respectively. Denote prox_f, relf_f as the proximal and reflection operator of f (Davis & Yin, 2016). The forms of T in Definition 2 for Stochastic Gradient Descent (SGD), Stochastic Proximal Gradient Descent (SPGD) and ADMM algorithms are summarized in Table 1. Further, let $\mathbf{X} = [x(1), \cdots, x(M)]^\top \in \mathbb{R}^{M \times p}$ that stores all local parameters across the network, $\mathbf{T}(\mathbf{X}; \Xi) = [T(x(1); \Xi_1), \cdots, T(x(M); \Xi_M)]^\top \in \mathbb{R}^{M \times p}$ stacking all local updating *w.r.t.* the first argument. In this way, the matrix form of Equation equation 1 is

$$\mathbf{X}^{k+1} = \mathbf{W}\mathbf{X}^k + \alpha_k \left(\mathbf{T}\left(\mathbf{X}^k; \Xi\right) - \mathbf{X}^k\right).$$

Table 1: Overview of several first-order algorithms

Algorithm	Operator identity (T)	Subgradient identity
SGD	$I - \gamma \nabla f$	$x^{k+1} = \mathbf{W}x^k - \gamma_k \nabla f(x^k)$
SPGD	$(I + \gamma \partial g)^{-1}(I - \gamma \nabla f)$	$x^{k+1} = \mathbf{W}x^k + (\mathrm{prox}_{\gamma_k g}(I - \gamma_k \nabla f(x^k)) - x^k)$
ADMM	$(I + \gamma \partial f)^{-1}\left[(I + \gamma \partial g)^{-1}(I - \gamma \partial f) + \gamma \partial f\right]$	$x^{k+1} = \mathbf{W}x^k + \frac{1}{2}\,\mathrm{refl}_{\gamma_k \partial f} \circ \mathrm{refl}_{\gamma_k \partial g}\left(x^k\right)$

Privacy Concern: In the stochastic D-KM algorithm, despite each agent communicating with its neighbours by sending parameters instead of directly exchanging raw data, the risk of leaking information still exists: the attacker can recover the sensitive information of data from shared parameters as pointed in Shokri et al. (2017), Fredrikson et al. (2015). This motivates us to consider privacy preserving iteration procedure with efficient communication while it retains a performance guarantee. Differential privacy, introduced by Dwork et al. (2006), is a widely adopted definition due to its important advantages over other privacy techniques. It quantifies to what extent individual privacy in a data set is preserved while releasing aggregated information.

Definition 3 (**(ε, δ)-Differential Privacy Dwork et al. (2006)**) *A stochastic algorithm $\mathcal{A}$ is called (ε, δ)-differential privacy if for any subset $\mathbb{R}_0 \subset \mathbb{R}^p$ and any neighbouring sample set pair Ξ and Ξ' which differs by only one sample, we have,*

$$\log\left[\frac{\mathbb{P}_{\mathcal{A}(\Xi)}\left(\mathcal{A}(\Xi) \in \mathbb{R}_0\right) - \delta}{\mathbb{P}_{\mathcal{A}(\Xi')}\left(\mathcal{A}\left(\Xi'\right) \in \mathbb{R}_0\right)}\right] \leq \varepsilon.$$

Differential privacy provides privacy guarantees by requiring the capability of indistinguishable in identifying whether an individual is in the data set or not based on the released information. The parameters ε and δ are privacy budgets indicating the strength of privacy protection from the algorithm. The classic differential privacy is called ε-differential privacy with $\delta = 0$, which imposes an upper bound e^{ε} on the multiplicative distance of probability distributions of randomized query outputs for any two neighbouring data sets (Dong et al., 2019). The common interpretation of (ε, δ)-differential privacy is that it is ε-differential privacy except with probability δ (Mironov, 2017).

3 SENSITIVITY OF THE STOCHASTIC D-KM ALGORITHM

In this section, we estimate the l_2 norm sensitivity of the stochastic D-KM, laying the foundation for noise addition in the generalized Gaussian mechanisms in Section 5. Note that the derivation of sensitivity of our proposed D-KM algorithm does not require the assumption of smoothness and strong convexity of objective functions. Before formalizing the result, we introduce the definition of sensitivity of algorithms in a decentralized learning setting, and assumptions.

Definition 4 (Sensitivity) *For a specific algorithm $\mathcal{A}$ acting on training samples, Ξ', Ξ'' which are two adjacent datasets that differ by one data point. Until iteration K, define the Δ_K-sensitivity of algorithm $\mathcal{A}$,*

$$\Delta_K := \sup_{\Xi', \Xi''} \left\|\mathcal{A}\left(\Xi'\right) - \mathcal{A}\left(\Xi''\right)\right\|.$$

Assumption 1 *The loss function is convex, closed, proper and sub-differentiable with respect to x, and the fixed-point iteration is bounded by the finite constant B, i.e., $\max_{x,\xi} \|T(x;\xi) - x\| \leq B$.*

$\|T(x;\xi) - x\|$ is defined as a fixed point residual in the literature which typically relates to the gradient of the objective (Davis & Yin, 2016). And it has shown that Assumption 1 is a weaker and common assumption in the optimization literature (Sun et al., 2021). We are now establishing the Δ_K-sensitivity of the D-KM algorithm. That is, we, through Theorem 1, provide the boundedness on Δ_K due to the only one different point for any two adjacent datasets.

Theorem 1 (**Δ_K-Sensitivity**) *Let $x^K = \frac{\sum_{m=1}^{M} x^K(m)}{M}$, $y^K = \frac{\sum_{m=1}^{M} y^K(m)}{M}$. Denote x^K and y^K as the corresponding outputs of the D-KM algorithm applied to two sets Ξ', Ξ'' of size NM which differ at only one point. Assume the initial value $\mathbf{X} = \mathbf{0}$. With Assumption 1 satisfied, given relaxed parameter, $\{\alpha_k\}_{k=0}^{K} \in (0, 1]$, the Δ_K-sensitivity of the D-KM algorithm has the upper bound,*

$$\mathbb{E}\Delta_K \leq \frac{2B\sum_{k=0}^{K-1}\alpha_k}{NM} + 4B\sum_{k=0}^{K-1}(1 + 2\alpha_k)\sum_{j=0}^{k-1}\alpha_j\lambda^{k-1-j}.$$

Theorem 1 quantifies the accumulated deviation bound between two trajectories of iterates based on two datasets which differ at only one point. We note that Sun et al. (2021) shows similar results for the stochastic gradient descent under a convex setting. With a fixed iteration number K, as the data size, M, N increases and λ decreases, Δ_K gets smaller for both diminishing and constant learning rates. However, it fails to have the sensitivity under control when K increases, which also suggests the risk of privacy that, with the higher iterative step, it will be easier to identify the specific sample. Moreover, different topologies affect the bound of Δ_K as the sensitivity decreases with decreasing λ, see Table 2 for details.

Note that compared to the centralized optimization setting (Hardt et al., 2016), there is an additional term of $4B\sum_{k=0}^{K-1}(1 + 2\alpha_k)\sum_{j=0}^{k-1}\alpha_j\lambda^{k-1-j}$ in Theorem 1, which quantifies the impact of the network topology on the learning procedure. We emphasize that, although decentralization has certain advantages compared with centralization, especially in the reduction of the expense of communication on the central server, it sacrifices data privacy and learning stability.

Table 2: Δ_K-Sensitivity under Different Graph Topology

Graph topology	Spectral gap (1-λ)	Sensitivity ($\alpha_k = \alpha$)	Sensitivity ($\alpha_k = (1/(k+1))$)
Ring	$\mathcal{O}(\frac{1}{M^2})$	$\mathcal{O}(\frac{\alpha K}{MN} + M^2\alpha K)$	$\mathcal{O}(\frac{\ln(K)}{MN} + M^2\ln(K))$
Grid	$\mathcal{O}(\frac{1}{M\log(M)})$	$\mathcal{O}(\frac{\alpha K}{MN} + M\log(M)\alpha K)$	$\mathcal{O}(\frac{\ln(K)}{MN} + M\log(M)\ln(K))$
Star	$\mathcal{O}(\frac{1}{M})$	$\mathcal{O}(\frac{\alpha K}{MN} + M\alpha K)$	$\mathcal{O}(\frac{\ln(K)}{MN} + M\ln(K))$
Exponential	$\mathcal{O}(\frac{1}{\log(M)})$	$\mathcal{O}(\frac{\alpha K}{MN} + \log(M)\alpha K)$	$\mathcal{O}(\frac{\ln(K)}{MN} + \log(M)\ln(K))$
Full connected	1	$\mathcal{O}(\frac{\alpha K}{MN} + \alpha K)$	$\mathcal{O}(\frac{\ln(K)}{MN} + \ln(K))$

4 PERFORMANCE OF DECENTRALIZED LEARNING ALGORITHMS WITH DIFFERENTIAL PRIVACY

Existing DP schemes in decentralized learning setting typically rely on the perturbation of objective functions, gradients, but are limited to iterates (Huang et al., 2019; Yu et al., 2021b; Asi et al., 2021). Such methods usually introduce extra noise that has privacy preservation but it is hard to examine how their algorithm and analysis can be adapted to the privacy multi-agent setting (McGregor et al., 2010), especially, the privacy and performance trade-off in the generalization of DP algorithms (He et al., 2021). In this paper, we consider iterate independent noise addition mechanisms (Definition 5) to preserve DP: a random noise is added to the iterate to reduce leakage information. Before introducing noise addition mechanisms used in the paper that satisfy differential privacy, in this section, we do a rigorous analysis by establishing a generalization error bound, as well as a finite sample guarantee of decentralized learning algorithms when these algorithms satisfy differential privacy. These results illustrate the effectiveness of using D-KM with the proposed noise addition mechanism in applications. We next proceed by quantifying the bound introduced by noise addition, and computing the end-to-end differential privacy guarantee across M agents over a network system.

Definition 5 (Noise-adding Mechanisms Geng et al. (2018)) *Given a data set Ξ, a query-output independent noise-adding mechanism $\widetilde{\mathcal{A}}$ will release the query output $\tilde{x}^k = \widetilde{\mathcal{A}}(x^k; \Xi)$ corrupted by an additive random noise d,*

$$\tilde{x}^k = x^k + d.$$

Let $L(x) = \mathbb{E}_{\xi\sim\mathbb{P}}[\ell(x,\xi)]$ and $x^\star$ be its optimal solution. Note that, for a specific stochastic algorithm $\mathcal{B}$ on Ξ with sample size NM with output $\mathcal{B}(\Xi)$, the excess generalization error of $\mathcal{B}$ defined as, $\mathbb{E}_{\Xi,\mathcal{B}}\left[L(\mathcal{B}(\Xi)) - L\left(x^\star\right)\right]$, can be decomposed into three terms

$$\underbrace{\mathbb{E}_{\Xi,\mathcal{B}}\left[L(\mathcal{B}(\Xi)) - \widehat{L}(\mathcal{B}(\Xi))\right]}_{\text{generalization error}} + \underbrace{\mathbb{E}_{\Xi,\mathcal{B}}\left[\widehat{L}(\mathcal{B}(\Xi)) - \widehat{L}(\widehat{x})\right]}_{\text{optimization error}} + \underbrace{\mathbb{E}_{\Xi,\mathcal{B}}\left[\widehat{L}(\widehat{x}) - \widehat{L}\left(x^\star\right)\right]}_{\text{test error}}. \tag{2}$$

Here, the expectation is taken over the algorithm and the data. We establish the boundedness of generalization error in Theorem 2 that reflects joint effects caused by the data Ξ and the algorithm $\mathcal{B}$.

Assumption 2 *The loss function $\ell(x,\xi)$ is nonnegative and Lipschitz continuous with respect to x, i.e., $\|\ell(x,\xi) - \ell(y,\xi)\| \leq Lip\|x-y\|$ for all $x, y \in \mathbb{R}^p$, $\xi \sim \mathbb{P}$.*

Theorem 2 (Generalization Bound) *Assume that the decentralized learning algorithm $\mathcal{B} : \Xi \mapsto \mathbb{R}^p \times \{1, \cdots, M\}$ is (ε, δ)-differentially private, and the loss function $\|\ell\|_\infty \leq R$ on the domain. Under Assumption 2, we have that,*

$$|\mathop{\mathbb{E}}_{\Xi\sim\mathbb{P}^{MN},\mathcal{B}(\Xi)}\left[L(\mathcal{B}(\Xi)) - \widehat{L}(\mathcal{B}(\Xi))\right]| \leq (1-e^{-\varepsilon})R + e^{-\varepsilon}M\delta.$$

Theorem 3 (Finite Sample Guarantee) *Under the Assumption of Theorem 2, for any $\epsilon > 0$, we have,*

$$\mathbb{P}(L(\mathcal{B}(\Xi)) \leq \widehat{L}(\mathcal{B}(\Xi)) + \epsilon) \geq \frac{\epsilon - (1-e^{-\varepsilon})R - e^{-\varepsilon}M\delta}{\epsilon + R}.$$

These two theorems represent the gap between the empirical loss based on finite samples and its expectation. It demonstrates the impact of differential privacy on in-sample and out-of-sample performance. Although ensuring data privacy sacrifices the generalization, these results show that a good privacy-preserving mechanism still retains a certain level of generalization as well as a finite sample guarantee. We further examine the optimization error bound in formula (2) that caused by adding noise to the query output.

Let $\widetilde{\mathbf{X}} = [\tilde{x}(1), \cdots, \tilde{x}(M)]^\top$ be the released iterates corrupted by an additive random noise for each agent, $\mathbf{X}^\star = [x^\star, \cdots, x^\star]^\top \in \mathbb{R}^{M\times M}$. With Assumption 2, the error bound can be controlled by the difference between the iterates and its true parameter,

$$\|\mathbf{W}\widetilde{\mathbf{X}}^k + \alpha_k \left(\mathbf{T}\left(\mathbf{X}^k;\Xi\right) - \mathbf{X}^k\right) - \mathbf{X}^\star\| \leq \|\mathbf{X}^{k+1} - \mathbf{X}^\star\| + \|\mathbf{W}\left(\mathbf{X}^k - \widetilde{\mathbf{X}}^k\right)\|,$$

where the first term is the same as in the non-privacy setting, which depends on the convergence properties of a given algorithm (Wotao, 2019). The second indicates the deviation by using the privacy mechanisms.

Theorem 4 (Boundedness of local iterates based on Gaussian additive noise) *Given a data set Ξ, assume an iterative independently noise-adding mechanism $\tilde{\mathcal{A}}$ releases the output $\tilde{\mathcal{A}}(x^k;\Xi) := x^k + d$ corrupted by an additive random noise d, where d follows a Gaussian distribution with mean μ and variance σ^2. The error bound caused by the additive noise is*

$$\mathbb{E}\left[\left\|\mathbf{W}\left(\mathbf{X}^k - \widetilde{\mathbf{X}}^k\right)\right\|_F^2 \leq p\left(\sigma^2 + \mu^2\right)\left[(M-1)\lambda^2 + 1\right],\right.$$

where $\|\cdot\|_F$ is the Frobenius norm of $\cdot$.

Generalization of local computation In real applications, each agent acts independently, where there is less likely to reach an agreement on a consistent (ε, δ)-DP across all agents (Bellet et al., 2018). Therefore, we proceed by investigating how the local computation would affect global differential privacy as a composition theorem. A similar result is also established in Kairouz et al. (2015).

Theorem 5 (Composition Theorem) *Denote iterates generated by the specific stochastic algorithm with K steps as $\{x^k\}_{k=1}^K$. For the m-th agent, denote $\tilde{\mathcal{A}}_m : \Xi \mapsto \{\tilde{x}^k(m)\}_{k=1}^K$, where $\tilde{x}^k(m)$ is the iterates corrupted by noise. Let $\{\tilde{X}(m)\}_{m=1}^M = (\tilde{X}(1), \cdots, \tilde{X}(M))$ with $\tilde{X}(1) = (\tilde{x}^1(m), \cdots, \tilde{x}^K(m))$. For any fixed m, if $\tilde{\mathcal{A}}_m$ is $(\varepsilon_m, \delta_m)$-differential private, then $\left\{\tilde{X}(m)\right\}_{m=1}^M$ is (ε', δ')-differential private, where,*

$$\delta' = 1 - \left\{\prod_{m=1}^M (1 - e^{a_m}\frac{\delta_m}{1+e^{\varepsilon_m}})\right\} + \left\{1 - \prod_{m=1}^M (1 - \frac{\delta_m}{1+e^{\varepsilon_m}})\right\},$$

$\varepsilon' = \min\{\varepsilon_1, \varepsilon_2, \varepsilon_3\}$,

$$\varepsilon_1 = \sum_{m=1}^M \varepsilon_m, \quad \varepsilon_2 = \sum_{m=1}^M C_{KL}(m) + \sqrt{2log(\frac{1}{\delta'})(\sum_{m=1}^M \varepsilon_m^2)},$$

$$\varepsilon_3 = \sum_{m=1}^M \frac{(e^{\varepsilon_m} - 1)\varepsilon_m}{e^{\varepsilon_m} + 1} + \sqrt{\sum_{m=1}^M 2\varepsilon_m^2 \log\left(e + \frac{\sqrt{\sum_{m=1}^M \varepsilon_m^2}}{\tilde{\delta}}\right)},$$

for some $0 < a_m \leq \varepsilon_m$, $\sum_{m=1}^M a_m = \varepsilon'$, and a real constant $\tilde{\delta}$.

5 Differential Privacy Via Truncated Generalized Gaussian Mechanisms

While the commonly adopted Gaussian noise-adding mechanism for a single iterate can guarantee DP (Croft et al., 2022; Ghosh et al., 2012; Cormode et al., 2019), the extremely large noise

Algorithm 1 DP-KM iteration

Initialize: $\mathbf{X}^0$, weight matrix $\mathbf{W}$, $\alpha_k \in (0,1]$, number of iterations K, variance of noise σ, scale parameter b
while $k \leq K$ **do**
  for $m \in V$ ($m \in [1, M]$) **do**
    Let $\mathbf{X}^k = \tilde{\mathbf{X}}^{k-1}$
      1. Local computation
        $x^k(m) = \mathbf{W}\mathbf{X}^k(m) + \alpha_k(T(x^k(x(m))) - x^k(m))$
  end for
  for $m \in V$ **do**
      2. Add noise for privacy guarantee, $\varepsilon_m^k \sim GG(0, \sigma, b)$
        $\tilde{x}^k(m) = x^k(m) + \varepsilon_m^k$
        Broadcast $\tilde{x}^k(m)$ to all neighbours $j \in \mathcal{N}(m)$
  end for
end while
Output: $\mathbf{X}^K = (x^K(1), \cdots, x^K(M))$ and $\bar{x}^K = \frac{1}{M}\sum_{m=1}^{M} x^K(m)$

will severely affect a learning process and degrade the performance of the trained model under differential privacy guarantee. Beyond satisfying DP, once the iterate is posed, especially for constrained optimization that requests an output fall into a close-set $\mathcal{R}$, *i.e.*, it also require, $\forall k, \mathbb{P}_d\left(x^k + d \in \mathcal{R} \mid x^k\right) \geq 1 - \epsilon$, for some $\epsilon > 0$. To rectify this and therefore improve the utility, in this section, we consider a new noise-adding mechanism by adding controlled noise to iterates. The proposed mechanism unifies the Laplace and the Gaussian mechanism in a general family. [Liu (2018)] considers a similar mechanism for the generalized Gaussian mechanism by truncating it to the valid range.

Truncated generalized Gaussian mechanism Different from the common mechanisms, we truncate the probability density function used for the generation of noise with a careful determination of an appropriate bounding parameter. In the reminder, we consider truncated Generalized Gaussian (GG) distribution $\mathcal{P}_d := \mathrm{GG}(0, \sigma, b)$ with location parameter 0, scale parameter $\sigma > 0$, shape parameter $b > 0$. Its probability density function is,

$$p(z \mid 0, \sigma, b) = C_{gg} \exp\left\{-\left(\frac{|z|}{\sigma}\right)^b\right\}, \text{ where } z \in [-A, A],$$

where C_{gg} is a constant to guanrantee $\int_{-A}^{A} p(z \mid 0, \sigma, b)dz = 1$. The differentially private D-KM with truncated GG noise is shown in Algorithm 1. In applications, we consider the case of $b = 1, 2$, which represents the truncated Laplace distribution and truncated normal distribution, see Definition $6 - 7$. We address that it can be extended to the case $b \geq 3$ to preserve a differential privacy guarantee.

Definition 6 (Truncated Laplacian Distribution Geng et al. (2018)) *Given the privacy parameters* $0 < \delta < \frac{1}{2}, \varepsilon > 0$ *and iterates sensitivity* $\Delta > 0$*, the probability density function of the truncated Laplacian distribution is defined as,*

$$p_{Lap}(z) := \begin{cases} C_{Lap} e^{-\frac{|z|}{\lambda}}, & \textit{for } z \in [-A, A], \\ 0, & \textit{otherwise,} \end{cases}$$

where $\lambda := \frac{\Delta}{\varepsilon}$, $A := \frac{\Delta}{\varepsilon}\log\left(1 + \frac{e^{\varepsilon}-1}{2\delta}\right)$, $C_{Lap} := \frac{1}{2\lambda\left(1 - e^{-\frac{A}{\lambda}}\right)} = \frac{1}{2\frac{\Delta}{\varepsilon}\left(1 - \frac{1}{1+\frac{e^{\varepsilon}-1}{2\delta}}\right)}$.

Geng et al. (2018) shows the optimality of the truncated Laplacian mechanism for minimizing the noise amplitude and noise power under (ε, δ)-differential privacy guarantee. In the following, we establish that truncated Gaussian noise also preserves (ε, δ)-differential privacy.

Definition 7 (Truncated Gaussian Distribution) *The probability density function of the truncated Gaussian distribution is defined as,*

$$p_{nor}(z) = C_{nor} \exp\left\{-\left(\frac{|z|}{\sigma}\right)^2\right\}, \quad \textit{for } z \in [-A, A],$$

where C_{nor} is a constant to guarantee $\int_{-A}^{A} p_{nor}(z)dz = 1$, $\sigma^2 \geq \varepsilon^{-1}\Delta^2$. In practice, the C_{nor} and A are determined by the equation,

$$\begin{cases} C_{nor} \cdot \sum_{l=0}^{\infty}(-1)^l \cdot \frac{A^{2l+1}}{\sigma^{2l} l!(2l+1)} = \frac{1}{2}, \\ C_{nor} \cdot \sum_{l=0}^{\infty}(-1)^l \frac{A^{2l+1} - (A-\Delta)^{2l+1}}{\sigma^{2l} l!(2l+1)} = \delta. \end{cases}$$

Theorem 6 *The truncated Gaussian mechanism preserves (ε, δ)-differential privacy.*

An important property of the truncated GG mechanism is that the range of addition noises are bounded to $[-A, A]$ while the DP still holds. We emphasize that, however, it cannot be applied as a way to achieve adherence to the valid range. Given constraints on the range of output, to ensure the utility, designing appropriate mechanisms are needed to be adaptive to the various constraint configurations, which will be one of the core aspects of our future work.

6 Numerical Experiment

In this section, we evaluate the performance of DP-KM by considering ℓ_1 regularized least square regression,

$$\min_{x \in \mathbb{R}^p} \frac{1}{MN} \sum_{m=1}^{M} \sum_{i=1}^{N} (A_{mi}x - b_{mi})^2, \quad \min_{x \in \mathbb{R}^p} \frac{1}{MN} \sum_{m=1}^{M} \sum_{i=1}^{N} (A_{mi}x - b_{mi})^2 + \lambda \|x\|.$$

We consider using SGD, SPGD and ADMM algorithms and evaluate the performance of its privacy variants on several decentralized settings: ring, star and full connected graph. We compare our proposed DP-KM algorithm with three baseline algorithms: (a) non-private decentralized approach; (b) private decentralized approach with Laplace noise; (c) private decentralized approach with Gaussian noise. Bias and average root mean squared errors (RMSEs) are used to quantify estimation performance and prediction accuracy. The resluts are summaried in Figure 1-4.

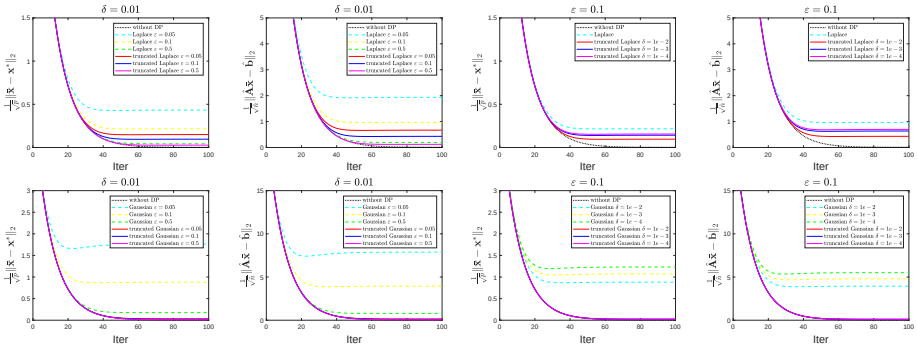

Figure 1: SGD with a fully connected graph.

Figure 1 shows that a truncated mechanism with larger ε and larger δ has better convergence. And the proposed mechanism has smaller estimation errors and prediction errors compared with Laplace and Gaussian mechanisms. In addition, our results demonstrate privacy-utility trade-offs of our approach. When privacy leakage increases, our truncated Laplace approach achieves better utility. We, then, fix the privacy budget and discuss the performance of the truncated approach under the settings that have different numbers of distributed data sources and different typologies, as shown in Figure 2, 3, 4. We demonstrate the effects of the structure of connected graphs. Compared with the ring graph, these results show a better convergence with the star and full-connected graphs.

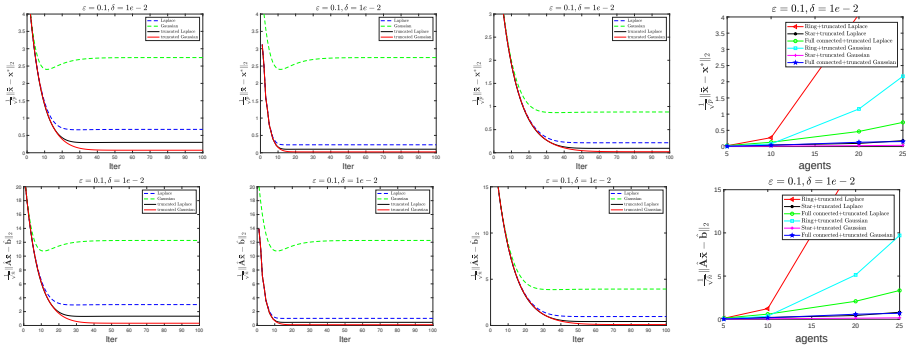

Figure 2: SGD with three graphs.

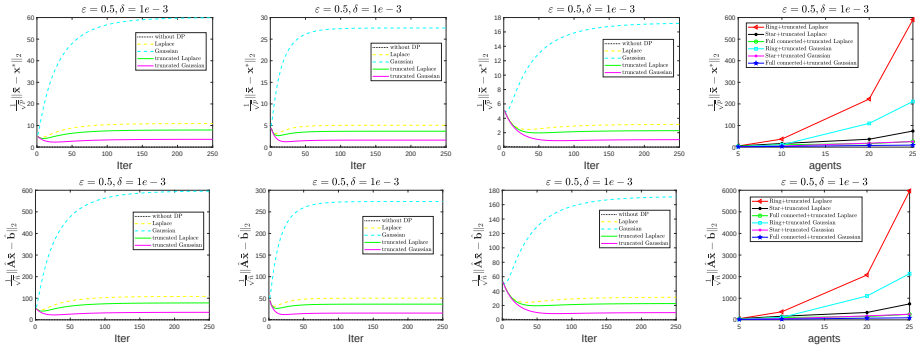

Figure 3: SPGD with three graphs.

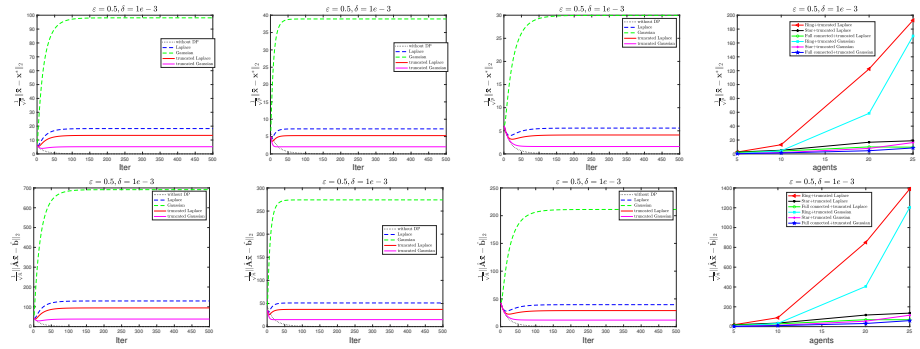

Figure 4: ADMM with three graphs.

7 Conclusion

Leveraging the decentralized learning procedure to train models under privacy constraints, in this paper, we have proposed a general framework of privacy-preserving algorithm, DP-KM, that is applicable to much of the existing work on optimization and consensus computation and, show that the proposed algorithm retains the performance guarantee regarding stability, generalization, and finite sample performance. The effect of local privacy-preserving computation on global differential privacy is also derived. To avoid extremely large addition noise added to the shared information that will severely affect and degrade the performance of learning process, we have introduced truncated generalized Gaussian mechanism, in which we demonstrate privacy and utility trade-offs under differential privacy guarantee. Experiments have demonstrated that our algorithm is highly effective in decentralized settings and that it performs better than state-of-the-art baseline algorithms.

REFERENCES

Sulaiman Alghunaim, Kun Yuan, and Ali H Sayed. A linearly convergent proximal gradient algorithm for decentralized optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Hilal Asi, John Duchi, Alireza Fallah, Omid Javidbakht, and Kunal Talwar. Private adaptive gradient methods for convex optimization. In *International Conference on Machine Learning*, pp. 383–392. PMLR, 2021.

Aurélien Bellet, Rachid Guerraoui, Mahsa Taziki, and Marc Tommasi. Personalized and private peer-to-peer machine learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 473–481. PMLR, 2018.

Pascal Bianchi, Walid Hachem, and Franck Iutzeler. A stochastic primal-dual algorithm for distributed asynchronous composite optimization. In *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 732–736. IEEE, 2014.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

Francesco Bullo, Jorge Cortés, and Sonia Martinez. *Distributed control of robotic networks: a mathematical approach to motion coordination algorithms*, volume 27. Princeton University Press, 2009.

Graham Cormode, Tejas Kulkarni, and Divesh Srivastava. Constrained private mechanisms for count data. *IEEE Transactions on Knowledge and Data Engineering*, 33(2):415–430, 2019.

William Croft, Jörg-Rüdiger Sack, and Wei Shi. Differential privacy via a truncated and normalized laplace mechanism. *Journal of Computer Science and Technology*, 37(2):369–388, 2022.

Damek Davis and Wotao Yin. Convergence rate analysis of several splitting schemes. In *Splitting Methods in Communication, Imaging, Science, and Engineering*, pp. 115–163. Springer, 2016.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*, 2019.

John Duchi and Ryan Rogers. Lower bounds for locally private estimation via communication complexity. In *Conference on Learning Theory*, pp. 1161–1191. PMLR, 2019.

Cynthia Dwork and Guy N Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pp. 265–284. Springer, 2006.

Farhad Farokhi. Distributionally-robust machine learning using locally differentially-private data. *Optimization Letters*, 16(4):1167–1179, 2022.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1322–1333, 2015.

Quan Geng, Wei Ding, Ruiqi Guo, and Sanjiv Kumar. Privacy and utility tradeoff in approximate differential privacy. *arXiv preprint arXiv:1810.00877*, 2018.

Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing*, 41(6):1673–1693, 2012.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 1225–1234. PMLR, 2016.

Fengxiang He, Bohan Wang, and Dacheng Tao. Tighter generalization bounds for iterative differentially private learning algorithms. In *Uncertainty in Artificial Intelligence*, pp. 802–812. PMLR, 2021.

Zonghao Huang, Rui Hu, Yuanxiong Guo, Eric Chan-Tin, and Yanmin Gong. Dp-admm: Admm-based distributed learning with differential privacy. *IEEE Transactions on Information Forensics and Security*, 15:1002–1012, 2019.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International Conference on Machine Learning*, pp. 1376–1385. PMLR, 2015.

Yiwei Li, Shuai Wang, Tsung-Hui Chang, and Chong-Yung Chi. Federated stochastic primal-dual learning with differential privacy. *arXiv preprint arXiv:2204.12284*, 2022.

Fang Liu. Generalized gaussian mechanism for differential privacy. *IEEE Transactions on Knowledge and Data Engineering*, 31(4):747–756, 2018.

Andrew McGregor, Ilya Mironov, Toniann Pitassi, Omer Reingold, Kunal Talwar, and Salil Vadhan. The limits of two-party differential privacy. In *IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 81–90. IEEE, 2010.

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

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Angelia Nedic, Alex Olshevsky, Asuman Ozdaglar, and John N Tsitsiklis. Distributed subgradient methods and quantization effects. In *The 47th IEEE Conference on Decision and Control*, pp. 4177–4184. IEEE, 2008.

Michael Rabbat and Robert Nowak. Distributed optimization in sensor networks. In *Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks*, pp. 20–27, 2004.

Ali H Sayed et al. Adaptation, learning, and optimization over networks. *Foundations and Trends® in Machine Learning*, 7(4-5):311–801, 2014.

Wei Shi, Qing Ling, Kun Yuan, Gang Wu, and Wotao Yin. On the linear convergence of the admm in decentralized consensus optimization. *IEEE Transactions on Signal Processing*, 62(7):1750–1761, 2014.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (SP)*, pp. 3–18. IEEE, 2017.

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pp. 245–248. IEEE, 2013.

Tao Sun, Dongsheng Li, and Bao Wang. Stability and generalization of decentralized stochastic gradient descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9756–9764, 2021.

Yin Wotao. Operator splitting methods for decentralized optimization. *Mathematica Numerica Sinica*, 41(3):225, 2019.

Bicheng Ying, Kun Yuan, Hanbin Hu, Yiming Chen, and Wotao Yin. Bluefog: Make decentralized algorithms practical for optimization and deep learning. *arXiv preprint arXiv:2111.04287*, 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, 2021a. doi: 10.1109/TVT.2021.3064877.

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, 2021b.

Tongtian Zhu, Fengxiang He, Lan Zhang, Zhengyang Niu, Mingli Song, and Dacheng Tao. Topology-aware generalization of decentralized sgd. In *International Conference on Machine Learning*, pp. 27479–27503. PMLR, 2022.

A APPENDIX

The appendix is divided into three parts. We first introduce some preliminaries before giving the detailed proof of the main results. We finally discuss more details for truncated mechanisms.

A.1 PRELIMINARIES

We know that the connection between local servers is described by a mixing matrix $\mathbf{W}$ for different topologies. The mixing matrix has a special property given in Lemma 1.

Lemma 1 (Convergence of Mixing Matrix Zhu et al. (2022)) *Let $\mathbf{P} \in \mathbb{R}^{M\times M}$ be a matrix whose elements are all $1/M$. Given any $t \in \mathbb{Z}^+$, the mixing matrix $\mathbf{W} \in \mathbb{R}^{M\times M}$ satisfies,*

$$\left\|\mathbf{W}^t - \mathbf{P}\right\|_{op} \leq \lambda^t,$$

where $\|\cdot\|_{op}$ denotes the spectral norm. Note that $\mathbf{W}$ corresponds to some Markov chain's transition matrix, and the parameter $0 \leq \lambda < 1$ characterizes the speed of convergence to the stationary state.

In this paper, we consider three topologies: ring, star, and full connected graph, which is presented in Figure 5.

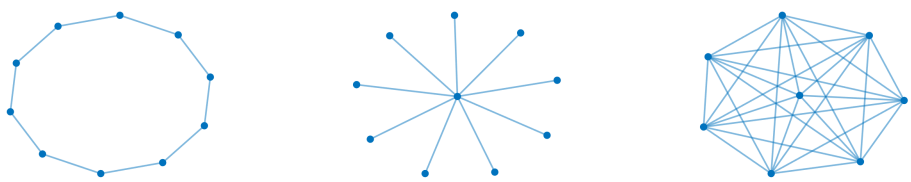

Figure 5: Structures of the connected graph used in numerical experiment, from left to right: ring, star, full connected.

Definition 8 (Nonexpansive Operator) *A mapping $T : \mathbb{X} \to \mathbb{X}$ is nonexpansive if $\|Tx - Ty\| \leq \|x - y\|$ holds for all $x, y \in \mathbb{X}$.*

We next introduce differential privacy based on KL divergence, which measures the distance between two distributions. The bound given in Definition 9 is useful for establishing the composition theorem.

Definition 9 ((ε, α)-Rényi Differential Private Duchi & Rogers (2019)) *For distributions P and Q, $D_\alpha(P\|Q) := \frac{1}{\alpha-1}\log\int(dP/dQ)^\alpha dQ$ is the Rényi α-divergence. For $\alpha \geq 1$, $\mathcal{A}$ is (ε, α)-Rényi differential private if for all Ξ, Ξ', $D_\alpha\left(\mathcal{A}(\Xi)\|\mathcal{A}\left(\Xi'\right)\right) \leq \varepsilon$. By taking $\alpha = 1$, we obtain ε-KL-privacy. If $\mathcal{A}$ is ε-differential private, then for any $\alpha \geq 1$, it also satisfies,*

$$D_\alpha\left(\mathcal{A}(\Xi)\|\mathcal{A}\left(\Xi'\right)\right) \leq \min\left\{2(\alpha-1)\varepsilon^2 + \min\left\{2, (e^\varepsilon - 1)\right\}\varepsilon, \varepsilon\right\}.$$

Lemma 2 (Mohri et al. (2018)) *Suppose $\{Y_i\}_{i=1}^N$ is a sequence of random variables, $Y_i \in [-c_i, c_i]$. Let $\{X_i\}_{i=1}^N$ be a sequence of random variables such that $\mathbb{E}[Y_i \mid X_{i-1}, \cdots, X_i] \leq C_i$, where $\{C_i\}_{i=1}^N$ is sequence of constant real numbers. Then,*

$$\mathbb{P}\left(\sum_{i=1}^N Y_i \geq \sum_{i=1}^N C_i + t\sqrt{\sum_{i=1}^N c_i^2}\right) \leq e^{-\frac{t^2}{2}}.$$

Lemma 3 (Dwork & Rothblum (2016)) *For any random variables Y, Z, we have that $D_\infty^\delta(Y\|Z) \leq \varepsilon$, $D_\infty^\delta(Z\|Y) \leq \varepsilon$ if and only if there exist r.v.s, Y', Z' such that $\Delta\left(Y\|Y'\right) \leq \frac{\delta}{e^{\varepsilon+1}}$, $\Delta\left(Z\|Z'\right) \leq \frac{\delta}{e^{\varepsilon+1}}$, $D_\infty\left(Y'\|Z'\right) \leq \varepsilon, D_\infty\left(Z'\|Y'\right) \leq \varepsilon$, where $D_\infty(X\|Y) = \max_{\nu\in\text{supp}(X)}\left[\log\frac{P(X\in U)}{P(Y\in U)}\right]$, $D_\infty^\delta(X\|Y) = \max_{U\in\text{supp}(i):P(Y\in U)\geq\delta}\left[\log\frac{P(X\in U)-\delta}{P(Y\in U)}\right]$, and $\Delta(X\|Y) = \max_U |P(X\in U) - P(Y\in U)|$.*

A.2 PROOF OF MAIN RESULTS

A.2.1 PROOF OF LEMMA 4

Lemma 4 *$x^k(m)$ is generated by (1) for a given relaxed parameter $\alpha_k \in (0,1]$. Define $x^k = \frac{1}{M}\sum_{m=1}^{M} x^k(m)$ is an average iterate of all local agent, $m \in \{1,\cdots,M\}$. Under Assumption 1, we have*

$$\left[\sum_{m=1}^{M}\|x^k(m)-x^k\|^2\right]^{\frac{1}{2}} \leq 2\sqrt{M}B\sum_{j=0}^{k-1}\alpha_j\lambda^{k-1-j}.$$

Proof. Let

$$\begin{aligned}
&\mathbf{X}^k := [x^k(1),\cdots,x^k(M)]^\top \in \mathbb{R}^{M\times p},\\
&\Delta^k := \alpha_k[T(x^k(1);\xi_{i_k(1)})-x^k(1),\cdots,T(x^k(M);\xi_{i_k(M)})-x^k(M)]^\top.
\end{aligned}$$

The matrix form of (1) is as follows,

$$\mathbf{X}^{k+1} = \mathcal{A}(\mathbf{X}^k) = \mathbf{W}\mathbf{X}^k + \Delta^k,$$

where $\mathbf{W}$ satisfies Definition 1 and Lemma 1. We then have,

$$(\mathbb{I}-\mathbf{P})\mathbf{X}^{k+1} = (\mathbb{I}-\mathbf{P})\left[\mathbf{W}\mathbf{X}^k+\Delta^k\right].$$

Therefore,

$$\begin{aligned}
\|(\mathbb{I}-\mathbf{P})\mathbf{X}^{k+1}\| &\leq \|\left(\mathbf{W}\mathbf{X}^k+\Delta^k\right)-\left(\mathbf{P}\mathbf{W}\mathbf{X}^k\right)\| + \|\Delta^k\|\\
&\leq \|\mathbf{W}\mathbf{X}^k+\Delta^k-\mathbf{P}\mathbf{W}\mathbf{X}^k\| + \alpha_k\sqrt{M}B\\
&\leq \|\mathbf{W}\mathbf{X}^k-\mathbf{P}\mathbf{W}\mathbf{X}^k\| + 2\alpha_k\sqrt{M}B\\
&= \left\|(\mathbf{W}-\mathbf{P})(\mathbb{I}-\mathbf{P})\mathbf{X}^k\right\| + 2\alpha_k\sqrt{M}B\\
&\leq \lambda\left\|(\mathbb{I}-\mathbf{P})\mathbf{X}^k\right\| + 2\alpha_k\sqrt{M}B\\
&\leq \lambda^{k+1}\|(\mathbb{I}-\mathbf{P})\mathbf{X}^0\| + 2\sqrt{M}B\sum_{l=0}^{k}\alpha_l\lambda^{k-l}.
\end{aligned}$$

The result of Lemma 4 follows.

A.2.2 PROOF OF THEOREM 1

Without loss of generality, we assume that two random sets Ξ' and Ξ'' differ at one point in the first N samples of agent 1, and the iteration starts with samples selected from this agent. Then with probability, $1-\frac{1}{N}$, the sample selected by D-KM is the same in both Ξ' and Ξ''. Note that,

$$\begin{aligned}
&x^{K+1}-y^{K+1}\\
&= \frac{1}{M}\sum_{m=1}^{M}\left\{\left[\sum_{l=1}^{M}\omega_{ml}x^K(l)-\alpha_K\left(T(x^K(m);\xi_{i_K(m)})-x^K(m)\right)\right]\right.\\
&\quad\left.-\left[\sum_{l=1}^{M}\omega_{ml}y^K(l)-\alpha_K\left(T(y^K(m);\xi_{i_K(m)})-y^K(m)\right)\right]\right\}\\
&= \frac{1}{M}\sum_{m=1}^{M}\left\{\left[x^K-\alpha_K\left(T(x^K;\xi_{i_K(m)})-x^K\right)\right]-\left[y^K-\alpha_K\left(T(y^K;\xi_{i_K(m)})-y^K\right)\right]\right\}+\varepsilon_K,
\end{aligned}$$

where,

$$
\begin{aligned}
\|\varepsilon_K\| \leq & \frac{1}{M}\sum_{m=1}^{M}\{\sum_{l=1}^{M}\omega_{ml}\left\|x^K(l)-x^K\right\| \\
& + \alpha_K\|\left(T(x^K(m);\xi_{i_K(m)})-x^K(m)\right)-\left(T(x^K;\xi_{i_K(m)})-x^K\right)\|\} \\
& + \frac{1}{M}\sum_{m=1}^{M}\{\sum_{l=1}^{M}\omega_{ml}\left\|y^K(l)-y^K\right\| \\
& + \alpha_K\|\left(T(y^K(m);\xi_{i_K(m)})-y^K(m)\right)-\left(T(y^K;\xi_{i_K(m)})-y^K\right)\|\} \\
\leq & \frac{1}{M}\sum_{m=1}^{M}\left\{\sum_{l=1}^{M}\omega_{ml}\left\|x^K(l)-x^K\right\|+2\alpha_K\|x^K(m)-x^K\|\right\} \\
& + \frac{1}{M}\sum_{m=1}^{M}\left\{\sum_{l=1}^{M}\omega_{ml}\left\|y^K(l)-y^K\right\|+2\alpha_K\|y^K(m)-y^K\|\right\} \\
\leq & \frac{1}{M}\left\{\sum_{l=1}^{M}\left\|x^K(l)-x^K\right\|+2\alpha_K\sum_{m=1}^{M}\|x^K(m)-x^K\|\right\} \\
& + \frac{1}{M}\left\{\sum_{l=1}^{M}\left\|y^K(l)-y^K\right\|+2\alpha_K\sum_{m=1}^{M}\|y^K(m)-y^K\|\right\} \\
\leq & \frac{1}{M}\left\{(1+2\alpha_K)\sqrt{M}\left[\sum_{m=1}^{M}\|x^K(m)-x^K\|^2\right]^{\frac{1}{2}}\right\} \\
& + \frac{1}{M}\left\{(1+2\alpha_K)\sqrt{M}\left[\sum_{m=1}^{M}\|y^K(m)-y^K\|^2\right]^{\frac{1}{2}}\right\} \\
\leq & 4(1+2\alpha_K)B\sum_{j=0}^{K-1}\alpha_j\lambda^{K-1-j}+2\lambda^{K+1}(1+2\alpha_K)\|(\mathbb{I}-\mathbf{P})\mathbf{X}^0\|/\sqrt{M} \\
= & 4(1+2\alpha_K)B\sum_{j=0}^{K-1}\alpha_j\lambda^{K-1-j}.
\end{aligned}
\tag{A1}
$$

Due to the nonexpansiveness of the operator T, we have that,

$$
\begin{aligned}
& \left\|\frac{1}{M}\sum_{m=1}^{M}\left\{\left[x^K-\alpha_K\left(T(x^K;\xi_{i_K(m)})-x^K\right)\right]-\left[y^K-\alpha_t\left(T(y^K;\xi_{i_K})-y^K\right)\right]\right\}\right\| \\
& \leq \|x^K-y^K\| = \delta_K.
\end{aligned}
$$

That is,

$$
\delta_{K+1} \leq \delta_K + 4(1+2\alpha_K)B\sum_{j=0}^{K-1}\alpha_j\lambda^{K-1-j}. \tag{A2}
$$

On the other hand, with probability $\frac{1}{N}$, D-KM selects the one sample to update in which Ξ' and Ξ'' differ.

$$
\begin{aligned}
& x^{K+1} - y^{K+1} \\
=& \frac{1}{M} \sum_{m=2}^{M} \left\{ \left[\sum_{l=1}^{M} \omega_{ml} x^K(l) - \alpha_K \left(T(x^K; \xi_{i_K(m)}) - x^K \right) \right] \right. \\
& - \left[\sum_{l=1}^{M} \omega_{ml} y^K(l) - \alpha_K \left(T(y^K; \xi_{i_K(m)} - y^K \right) \right] \\
& + \frac{1}{M} \left\{ \left[\sum_{l=1}^{M} \omega_{1l} x^K(l) - \alpha_K \left(T(x^K; \xi_{i_K(1)}) - x^K \right) \right] \right. \\
& - \left[\sum_{l=1}^{M} \omega_{1l} y^K(l) - \alpha_K \left(T(y^K; \xi'_{i_K(1)}) - y^K \right) \right] + \Im_K \\
=& \frac{1}{M} \sum_{m=2}^{M} \left\{ \left[x^K - \alpha_K \left(T(x^K; \xi_{i_K(m)}) - x^K \right) \right] - \left[y^K - \alpha_K \left(T(y^K; \xi_{i_K(m)}) - y^K \right) \right] \right\} \\
& + \frac{1}{M} \left\{ \left[x^K - \alpha_K \left(T(x^K; \xi_{i_K(1)}) - x^K \right) \right] - \left[y^K - \alpha_K \left(T(y^K; \xi'_{i_K(1)}) - y^K \right) \right] \right\} + \Im_K.
\end{aligned}
$$

From the inequality A1, we have $\|\Im_K\| \leq 4\left(1+2\alpha_K\right) B \sum_{j=0}^{K-1} \alpha_j \lambda^{K-1-j}$. We conclude that

$$
\|x^{K+1} - y^{K+1}\| \leq \frac{M-1}{M} \delta_K + \frac{\delta_K + 2\alpha_K B}{M} + 4\left(1+2\alpha_K\right) B \sum_{j=0}^{K-1} \alpha_j \lambda^{K-1-j}. \tag{A3}
$$

Inequalities A2, A3 imply that

$$
\mathbb{E}\delta_{K+1} \leq \mathbb{E}\delta_K + \frac{2\alpha_t B}{MN} + 4\left(1+2\alpha_K\right) B \sum_{j=0}^{K-1} \alpha_j \lambda^{K-1-j},
$$

which indicates that,

$$
\mathbb{E}\delta_{K+1} \leq \frac{2B \sum_{k=0}^{K} \alpha_k}{MN} + 4B \sum_{k=0}^{K} (1+2\alpha_k) \sum_{j=0}^{k-1} \alpha_j \lambda^{k-1-j}.
$$

A.2.3 Proof of Theorem 2

Let Ξ^{ξ_0} be the sample set with size MN which differs Ξ at only point ξ_0. Note that

$$
\begin{aligned}
& \mathop{\mathbb{E}}_{\Xi \sim \mathbb{P}^{MN}} \left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)} \left[\hat{L}(\mathcal{B}(\Xi)) \right] \right] = \mathop{\mathbb{E}}_{\Xi \sim \mathbb{P}^{MN}} \left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)} \left[\mathop{\mathbb{E}}_{\xi \sim \Xi_m} \left[\ell\left(x_{\mathcal{B}(S)}, \xi \right) \right] \right] \right] \\
&= \mathop{\mathbb{E}}_{\Xi \sim \mathbb{P}^{MN}} \left[\mathop{\mathbb{E}}_{\xi \sim \Xi} \left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)} \left[\ell\left(x_{\mathcal{B}(\Xi)}, \xi_{m_{\mathcal{B}(\Xi)}} \right) \right] \right] \right] \\
&= \mathop{\mathbb{E}}_{\Xi \sim \mathbb{P}^{MN}} \left[\mathop{\mathbb{E}}_{\xi \sim \Xi} \left[\int_0^R \mathbb{P}\left(\ell\left(x_{\mathcal{B}(\Xi)}, \xi_{m_{\mathcal{B}(\Xi)}} \right) > t \right) dt \right] \right] \\
&= \mathop{\mathbb{E}}_{\Xi \sim \mathbb{P}^{MN}} \left[\mathop{\mathbb{E}}_{\xi \sim \Xi} \left[\sum_{m=1}^{M} \int_0^R \mathbb{P}\left(\ell\left(x_{\mathcal{B}(\Xi)}, \xi_m \right) > t, m_{\mathcal{B}(\Xi)} = m \right) dt \right] \right] \\
&\leq \mathop{\mathbb{E}}_{\Xi \sim \mathbb{P}^{MN}} \left[\mathop{\mathbb{E}}_{\xi \sim \Xi, \xi^0 \sim \mathbb{P}} \left[\sum_{m=1}^{M} \int_0^R \left(e^{\varepsilon} \mathbb{P}\left(\ell(x_{\mathcal{B}(\Xi^{\xi_0})}, \xi_m) > t, m_{\mathcal{B}(\Xi)} = m \right) + \delta \right) \mathrm{d}t \right] \right] \\
&\leq \sum_{m=1}^{M} e^{\varepsilon} \mathop{\mathbb{E}}_{\Xi \sim \mathbb{P}^{MN}} \left[\mathop{\mathbb{E}}_{\xi \sim \Xi, \xi^0 \sim \mathbb{P}} \left[\int_0^R \left(\mathbb{P}\left(\ell(x_{\mathcal{B}(\Xi^{\xi_0})}, \xi_m) > t, m_{\mathcal{B}(\Xi)} = m \right) \right) \mathrm{d}t \right] \right] + M\delta \\
&\leq \sum_{m=1}^{M} e^{\varepsilon} \mathop{\mathbb{E}}_{\Xi' \sim \mathbb{P}^{MN-1}} \left[\mathop{\mathbb{E}}_{\xi_m \sim \mathbb{P}, \xi^0 \sim \mathbb{P}} \left[\int_0^R \left(\mathbb{P}\left(\ell\left(x_{\mathcal{B}(\Xi' \cup \{\xi_0\})}, \xi_m \right) > t, m_{\mathcal{B}(\Xi)} = m \right) \right) \mathrm{d}t \right] \right] + M\delta,
\end{aligned}
$$

where the first inequality holds because the algorithm $\mathcal{B}(\Xi)$ is (ε, δ)-differentially private. Let $\Xi = \Xi' \cup \{\xi^0\}$, and ξ_m, ξ_0 are i.i.d following the distribution $\mathbb{P}$. We have

$$
\begin{aligned}
&\sum_{m=1}^{M} e^{\varepsilon} \mathop{\mathbb{E}}_{\Xi' \sim \mathbb{P}^{MN-1}} \left[\mathop{\mathbb{E}}_{\xi_m \sim \mathbb{P}, \xi^0 \sim \mathbb{P}} \left[\int_0^R \left(\mathbb{P}\left(\ell\left(x_{\mathcal{B}(\Xi' \cup \{\xi_0\})}, \xi_m \right) > t, m_{\mathcal{B}(\Xi)} = m \right) \right) \mathrm{d}t \right] \right] + M\delta \\
&= \sum_{m=1}^{M} e^{\varepsilon} \mathop{\mathbb{E}}_{\Xi \sim \mathbb{P}^{MN}} \left[\mathop{\mathbb{E}}_{\xi \sim \mathbb{P}} \left[\int_0^R \left(\mathbb{P}\left(\ell\left(x_{\mathcal{B}(\Xi)}, \xi \right) > t, m_{\mathcal{B}(\Xi)} = m \right) \right) \mathrm{d}t \right] \right] + M\delta \\
&= e^{\varepsilon} \mathop{\mathbb{E}}_{\Xi \sim \mathbb{P}^{MN}} \left[\mathop{\mathbb{E}}_{\xi \sim \mathbb{P}} \left[\int_0^R \left(\mathbb{P}\left(\ell\left(x_{\mathcal{B}(\Xi)}, \xi \right) > t \right) \right) \mathrm{d}t \right] \right] + M\delta \\
&= e^{\varepsilon} \mathop{\mathbb{E}}_{\Xi \sim \mathbb{P}^{MN}} \left[\mathop{\mathbb{E}}_{\xi \sim \mathbb{P}} \left[\mathbb{E}^{\mathcal{B}(\Xi)} \left[\ell\left(x_{\mathcal{B}(S)}, \xi \right) \right] \right] \right] + M\delta.
\end{aligned}
$$

Thus we obtain,

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\Xi \sim \mathbb{P}^{MN}} \left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)} \left[\hat{L}(\mathcal{B}(\Xi)) \right] \right] \leq e^{\varepsilon} \mathop{\mathbb{E}}_{\Xi \sim \mathbb{P}^{MN}} \left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)} \left[L(\mathcal{B}(\Xi)) \right] \right] + M\delta, \\
&- \mathop{\mathbb{E}}_{\Xi \sim \mathbb{P}^{MN}} \left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)} \left[L(\mathcal{B}(\Xi)) \right] \right] \leq -e^{-\varepsilon} \mathop{\mathbb{E}}_{\Xi \sim \mathbb{P}^{MN}} \left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)} \left[\hat{L}(\mathcal{B}(\Xi)) \right] \right] + e^{-\varepsilon} M\delta.
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\Xi \sim \mathbb{P}^{MN}} \left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)} \left[\hat{L}(\mathcal{B}(\Xi)) \right] \right] - \mathop{\mathbb{E}}_{\Xi \sim \mathbb{P}^{MN}} \left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)} \left[L(\mathcal{B}(\Xi)) \right] \right] \\
&\leq e^{-\varepsilon} M\delta - e^{-\varepsilon} \mathop{\mathbb{E}}_{\Xi \sim \mathbb{P}^{MN}} \left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)} \left[\hat{L}(\mathcal{B}(\Xi)) \right] \right] + \mathop{\mathbb{E}}_{\Xi \sim \mathbb{P}^{MN}} \left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)} \left[\hat{L}(\mathcal{B}(\Xi)) \right] \right] \\
&= (1 - e^{-\varepsilon}) \mathop{\mathbb{E}}_{\Xi \sim \mathbb{P}^{MN}} \left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)} \left[\hat{L}(\mathcal{B}(\Xi)) \right] \right] + e^{-\varepsilon} M\delta \\
&\leq (1 - e^{-\varepsilon}) R + e^{-\varepsilon} M\delta.
\end{aligned}
$$

The other side can be similarly obtained. Now we complete the proof.

A.2.4 Proof of Theorem 3

Due to $\hat{L}(\mathcal{B}(\Xi)) \geq 0$, for any $\epsilon > 0$, we have,

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\Xi \sim \mathbb{P}^{MN}} \left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)} \left[L(\mathcal{B}(\Xi)) \right] \right] &\geq \mathop{\mathbb{E}}_{\Xi \sim \mathbb{P}^{MN}} \left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)} \left[L(\mathcal{B}(\Xi)) \mathbb{I}_{\{L(\mathcal{B}(\Xi)) \geq \hat{L}(\mathcal{B}(\Xi)) + \epsilon\}} \right] \right] \\
&\geq \mathop{\mathbb{E}}_{\Xi \sim \mathbb{P}^{MN}} \left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)} \left[(\hat{L}(\mathcal{B}(\Xi)) + \epsilon) \mathbb{I}_{\{L(\mathcal{B}(\Xi)) \geq \hat{L}(\mathcal{B}(\Xi)) + \epsilon\}} \right] \right].
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\Xi\sim\mathbb{P}^{MN}}\left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)}\left[L(\mathcal{B}(\Xi))\right]\right]-\mathop{\mathbb{E}}_{\Xi\sim\mathbb{P}^{MN}}\left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)}\left[\hat{L}(\mathcal{B}(\Xi))\right]\right]\\
&\geq\mathop{\mathbb{E}}_{\Xi\sim\mathbb{P}^{MN}}\left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)}\left[(\hat{L}(\mathcal{B}(\Xi))+\epsilon)\mathbb{I}_{\{L(\mathcal{B}(\Xi))\geq\hat{L}(\mathcal{B}(\Xi))+\epsilon\}}\right]\right]\\
&\quad-\mathop{\mathbb{E}}_{\Xi\sim\mathbb{P}^{MN}}\left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)}\left[\hat{L}(\mathcal{B}(\Xi))\mathbb{I}_{\{L(\mathcal{B}(\Xi))\geq\hat{L}(\mathcal{B}(\Xi))+\epsilon\}}\right]\right]\\
&\quad-\mathop{\mathbb{E}}_{\Xi\sim\mathbb{P}^{MN}}\left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)}\left[\hat{L}(\mathcal{B}(\Xi))\mathbb{I}_{\{L(\mathcal{B}(\Xi))<\hat{L}(\mathcal{B}(\Xi))+\epsilon\}}\right]\right]\\
&=\mathop{\mathbb{E}}_{\Xi\sim\mathbb{P}^{MN}}\left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)}\left[\epsilon\mathbb{I}_{\{L(\mathcal{B}(\Xi))\geq\hat{L}(\mathcal{B}(\Xi))+\epsilon\}}\right]\right]\\
&\quad-\mathop{\mathbb{E}}_{\Xi\sim\mathbb{P}^{MN}}\left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)}\left[\hat{L}(\mathcal{B}(\Xi))\mathbb{I}_{\{L(\mathcal{B}(\Xi))<\hat{L}(\mathcal{B}(\Xi))+\epsilon\}}\right]\right]\\
&\geq\epsilon\cdot\mathbb{P}(L(\mathcal{B}(\Xi))\geq\hat{L}(\mathcal{B}(\Xi))+\epsilon)-R\cdot\mathbb{P}(L(\mathcal{B}(\Xi))\leq\hat{L}(\mathcal{B}(\Xi))+\epsilon)\\
&=\epsilon-(\epsilon+R)\cdot\mathbb{P}(L(\mathcal{B}(\Xi))\leq\hat{L}(\mathcal{B}(\Xi))+\epsilon).
\end{aligned}
\tag{A4}
$$

Note that Theorem 2 indicates

$$
\mathop{\mathbb{E}}_{\Xi\sim\mathbb{P}^{MN}}\left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)}\left[L(\mathcal{B}(\Xi))\right]\right]-\mathop{\mathbb{E}}_{\Xi\sim\mathbb{P}^{MN}}\left[\mathop{\mathbb{E}}_{\mathcal{B}(\Xi)}\left[\hat{L}(\mathcal{B}(\Xi))\right]\right]\leq(1-e^{-\varepsilon})R+e^{-\varepsilon}M\delta.
$$

This result combined with equation A4 implies that

$$
\mathbb{P}(L(\mathcal{B}(\Xi))\leq\hat{L}(\mathcal{B}(\Xi))+\epsilon)\geq\frac{\epsilon-(1-e^{-\varepsilon})R-e^{-\varepsilon}M\delta}{\epsilon+R}.
$$

A.2.5 Proof of Theorem 4

Denote $x_m^k := x^k(m)$, $x^{v,k}$ as the v-th element of $x^k\in\mathbb{R}^p$. We know that

$$
\begin{aligned}
&\left\|\mathbf{W}\left(\mathbf{X}^k-\widetilde{\mathbf{X}}^k\right)\right\|_F^2=\sum_{m=1}^M\left\|\sum_{l=1}^M w_{ml}\left(x_l^k-\widetilde{x}_l^k\right)\right\|^2=\sum_{v=1}^p\sum_{m=1}^M\left\|\sum_{l=1}^M w_{ml}\left(x_l^{v,k}-\widetilde{x}_l^{v,k}\right)\right\|^2\\
&=\sum_{v=1}^p\sum_{m=1}^M\sigma_{k,m}^2\left(\sum_{l=1}^M w_{ml}^2\right)\left\{\frac{\sum_{l=1}^M w_{ml}\left[\left(x_l^{v,k}-\tilde{x}_l^{v,k}\right)-\mu_{k,m}^v\right]+\sum_{l=1}^M w_{ml}\mu_{k,l}^v}{\sigma_{k,m}\sqrt{\sum_{l=1}^M w_{ml}^2}}\right\}^2\\
&\leq\sum_{v=1}^p\sum_{m=1}^M\sigma_{k,m}^2\left(\sum_{l=1}^M w_{ml}^2\right)\left\{\frac{\sum_{l=1}^M w_{ml}\left[\left(x_l^{v,k}-\widetilde{x}_l^{v,k}\right)-\mu_{km}^v\right]}{\sigma_{k,m}\sqrt{\sum_{l=1}^M w_{ml}^2}}\right\}^2+\mu^2pM\cdot\sum_{m=1}^M\sum_{l=1}^M w_{ml}^2\\
&+\sum_{v=1}^p\sum_{m=1}^M\left(\sum_{l=1}^M w_{ml}^2\right)\frac{\sum_{l_1=1}^M\sum_{l_2=1}^M w_{ml_1}w_{ml_2}\mu_{k,l_2}^v\left[\left(x_{l_1}^{v,k}-\tilde{x}_{l_1}^{v,k}\right)-\mu_{k,l_1}^v\right]}{\sqrt{\sum_{l_1=1}^M w_{ml_1}^2}\sqrt{\sum_{l_2=1}^M w_{ml_2}^2}}.
\end{aligned}
$$

The last inequality holds since $\sum_{v=1}^p\left(\mu_{k,l}^v\right)^2\leq p\mu^2$. Considered that the weight difference is normally distributed,

$$
\left(x_m^k-\tilde{x}_m^k\right)\sim\mathcal{N}\left(\mu_{k,m},\sigma_{k,m}^2I_p\right),\quad m=1,\ldots,M,
$$

with $\mu_{k,m}$ satisfying $\|\mu_{k,m}\|_2^2\leq p\mu^2$ and $\sigma_{k,m}^2\in\mathbb{R}$ being bounded by σ^2, we obtain,

$$
\frac{\sum_{l=1}^M w_{ml}\left[\left(x_l^{v,k}-\tilde{x}_l^{v,k}\right)-\mu_{k,m}^v\right]}{\sigma_{k,m}\sqrt{\sum_{l=1}^M w_{ml}^2}}\overset{i.i.d.}{\sim}\mathcal{N}(0,1).
$$

It follows that

$$\left\{\frac{\sum_{l=1}^{M} w_{m,l}\left[\left(x_l^{v,k}-\widetilde{x}_l^{v,k}\right)-\mu_{k,m}^{v}\right]}{\sigma_{k,m}\sqrt{\sum_{l=1}^{M} w_{ml}^2}}\right\}^2 \sim \mathcal{X}^2(1), \quad l=1,\ldots,M.$$

Further, $\forall l, \mathbb{E}\left[\left(x_l^{v,k}-\widetilde{x}_l^{v,k}\right)-\mu_{k,l}^{v}\right]=0$, therefore, we have,

$$\mathbb{E}\left[\sum_{v=1}^{p}\sum_{m=1}^{M}\left(\sum_{l=1}^{M} w_{ml}^2\right)\frac{\sum_{l_1=1}^{M}\sum_{l_2=1}^{M} w_{ml_1}w_{ml_2}\mu_{k,l_2}^{v}\left[\left(x_{l_1}^{v,k}-\widetilde{x}_{l_1}^{v,k}\right)-\mu_{k,l_1}^{v}\right]}{\sqrt{\sum_{l_1=1}^{M} w_{ml_1}^2}\sqrt{\sum_{l_2=1}^{M} w_{ml_2}^2}}\right]=0,$$

As a consequence,

$$\mathbb{E}\left[\left\|\mathbf{W}\left(\mathbf{X}^k-\widetilde{\mathbf{X}}^k\right)\right\|_F^2\right. \leq p\sigma^2\sum_{m=1}^{M}\lambda_m^2+p\mu^2\sum_{m=1}^{M}\lambda_m^2 \leq p\left(\sigma^2+\mu^2\right)\left[1+(M-1)\lambda^2\right].$$

A.2.6 Proof of Theorem 5

For any subsets $H, H_1, \cdots, H_M$, with Definition 3, we have that,

$$\log\frac{\mathbb{P}_{\Xi}\left(\{\tilde{X}(m)\}_{m=1}^{M}\in H\right)}{\mathbb{P}_{\Xi'}\left(\{\tilde{X}(m)\}_{m=1}^{M}\in H\right)}=\log\prod_{m=1}^{M}\frac{\mathbb{P}_{\Xi}\left(\tilde{X}(m)\in H_m\right)}{\mathbb{P}_{\Xi'}\left(\tilde{X}(m)\in H_m\right)}$$
$$=\sum_{m=1}^{M}\log\frac{\mathbb{P}_{\Xi}\left(\tilde{X}(m)\in H_m\right)}{\mathbb{P}_{\Xi'}\left(\tilde{X}(m\in H_m)\right)}=\sum_{m=1}^{M}\log\frac{\mathbb{P}_{\Xi}\left(\tilde{\mathcal{A}}_m(\Xi_m)\in H_m\mid\tilde{X}^0(m)\right)}{\mathbb{P}_{\Xi'}\left(\tilde{\mathcal{A}}_m(\Xi'_m)\in H_m\mid\tilde{X}^0(m)\right)}$$

And, we have,

$$\mathbb{E}_{\Xi}\left\{\log\left(\frac{\mathbb{P}(\tilde{\mathcal{A}}_m(\Xi)\in H_m)}{\mathbb{P}(\tilde{\mathcal{A}}_m(\Xi')\in H_m)}\right)\mid\tilde{X}^0(m)\right\}\leq D_{KL}\left(\tilde{\mathcal{A}}(\Xi))\|\tilde{\mathcal{A}}(\Xi'))\right)$$
$$\leq\min\left\{\min\left\{2,e^{\varepsilon_m}-1\right\}\varepsilon_m,\varepsilon_m\right\}:=C_{KL}(m).$$

Based on Lemma 2, we have that,

$$\mathbb{P}\left(\sum_{m=1}^{M}\log\left(\frac{\mathbb{P}\left(\tilde{\mathcal{A}}_m(\Xi)\in H_m\right)}{\mathbb{P}\left(\tilde{\mathcal{A}}_m(\Xi')\in H_m\right)}\right)\geq\sum_{m=1}^{M}C_{kL}(m)+t\cdot\sqrt{\sum_{m=1}^{M}\varepsilon_m^2}\right)\leq e^{-\frac{t^2}{2}}.$$

Let $\tilde{\delta}=e^{-\frac{t^2}{2}}$, that is, $t=\sqrt{2\log\frac{1}{\delta'}}, \forall\tilde{\delta}>0$. It follows

$$\mathbb{P}\left(\sum_{m=1}^{M}\log\left(\frac{\mathbb{P}\left(\tilde{\mathcal{A}}_m(\Xi)\in H_m\right)}{\mathbb{P}\left(\tilde{\mathcal{A}}_m(\Xi')\in H_m\right)}\right)\geq\sum_{m=1}^{M}C_{kL}(m)+\sqrt{2\log\frac{1}{\tilde{\delta}}}\cdot\sqrt{\sum_{m=1}^{M}\varepsilon_m^2}\right)\leq\tilde{\delta}.$$

Therefore, $\{\tilde{X}(m)\}_{m=1}^{M}$ is $(\varepsilon',\tilde{\delta})$-pDP with $\varepsilon'=\sum_{m=1}^{M}C_{kL}(m)+\sqrt{\left(2\log\frac{1}{\tilde{\delta}}\right)\left(\sum_{m=1}^{M}\varepsilon_m^2\right)}$.

Since $\left\{\tilde{x}^k(m)\right\}_{k=0}^{K}$ is (ε_m,δ_m)-DP, for each $m=1,\cdots,M$. Based on Lemma 3, we have that there exists r.v.s. $\tilde{Y}_{\Xi}(m)$, $\tilde{Z}_{\Xi'}(m)$, such that,

$$\Delta\left(\tilde{X}_{\Xi}(m)\|\tilde{Y}_{\Xi}(m)\right)\leq\frac{\delta_m}{e^{\varepsilon_m}+1},\quad\Delta\left(\tilde{X}_{\Xi'}(m)\|\tilde{Z}_{\Xi'}(m)\right)\leq\frac{\delta_m}{e^{\varepsilon_m}+1}$$

$$D_{\infty}\left(\tilde{Y}_{\Xi}(m)\|\tilde{Z}_{\Xi'}(m)\right)\leq\varepsilon_m,\quad D_{\infty}\left(\tilde{Z}_{\Xi'}(m)\|\tilde{Y}_{\Xi}(m)\right)\leq\varepsilon_m.$$

Since $\{\tilde{X}(m)\}_{m=1}^{M}$ is ε'-pDP, we need to prove that,

$$\prod_{m=1}^{M} \mathbb{P}\left(\tilde{X}_{\Xi}(m) \in H_m\right) \leq e^{\varepsilon'} \prod_{m=1}^{M} \mathbb{P}\left(\tilde{X}_{\Xi'}(m) \in H_m\right) + \delta'.$$

Note that,

$$\mathbb{P}\left(\tilde{Y}_{\Xi}(m) \in H_i\right) \leq \min\left\{e^{\varepsilon_m}, \frac{1}{\mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right)}\right\} \mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right)$$

which indicates that,

$$\prod_{m=1}^{M} \mathbb{P}\left(\tilde{Y}_{\Xi}(m) \in H_i\right) \leq \prod_{m=1}^{M} \min\left\{e^{\varepsilon_m}, \frac{1}{\mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right)}\right\} \mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right).$$

Case 1: $\prod_{m=1}^{M}\left\{e^{\varepsilon_m}, \frac{1}{\mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right)}\right\} \leq e^{\varepsilon'}$

$$\begin{aligned}
&\prod_{m=1}^{M} \mathbb{P}\left(\tilde{Y}_{\Xi}(m) \in H_m\right) \\
&\leq \prod_{m=1}^{M} \min\left\{e^{\varepsilon_m}, \frac{1}{\mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right)}\right\} \mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right) \\
&+ \prod_{m=1}^{M} \min\left\{e^{\varepsilon_m}, \frac{1}{\mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right)}\right\} \mathbb{P}\left(\tilde{X}_{\Xi'}(m) \in H_m\right) \\
&- \prod_{m=1}^{M} \min\left\{e^{\varepsilon_m}, \frac{1}{\mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right)}\right\} \left\{\mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right) - \frac{\delta_m}{1+e^{\varepsilon_m}}\right\}.
\end{aligned}$$

Because

$$\left\{\mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right) - \frac{\delta_m}{1+e^{\varepsilon_m}}\right\} \leq \mathbb{P}\left(\tilde{X}_{\Xi'}(m) \in H_m\right)$$

$$\begin{aligned}
&\prod_{m=1}^{M} \min\left\{e^{\varepsilon_m}, \frac{1}{\mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right)}\right\} \mathbb{P}\left(\tilde{X}_{\Xi'}(m) \in H_m\right) \\
&- \prod_{m=1}^{M} \min\left\{e^{\varepsilon_m}, \frac{1}{\mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right)}\right\} \left\{\mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right) - \frac{\delta_m}{1+e^{\varepsilon_m}}\right\} \\
&\leq 1 - \prod_{m=1}^{M} \left\{1 - \min\left\{e^{\varepsilon_m}, \frac{1}{\mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right)}\right\} \frac{\delta_m}{1+e^{\varepsilon_m}}\right\} := \Delta_1.
\end{aligned}$$

We have that,

$$\prod_{m=1}^{M} \mathbb{P}\left(\tilde{Y}_{\Xi}(m) \in H_m\right) \leq \prod_{m=1}^{M} \min\left\{e^{\varepsilon_m}, \frac{1}{\mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right)}\right\} \mathbb{P}\left(\tilde{X}_{\Xi'}(m) \in H_m\right) + \Delta_1.$$

Combing the inequality $\mathbb{P}\left(\tilde{Y}_{\Xi}(m) \in H_m\right) \geq \mathbb{P}\left(\tilde{X}_{\Xi}(m) \in H_m\right) - \frac{\delta_m}{1+e^{\varepsilon_m}}$. We have,

$$
\begin{aligned}
&\prod_{m=1}^{M} \mathbb{P}\left(\tilde{X}_{\Xi}(m) \in H_m\right) \\
\leq &\prod_{m=1}^{M} \min\left\{e^{\varepsilon_m}, \frac{1}{\mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right)}\right\} \mathbb{P}\left(\tilde{X}_{\Xi'}(m) \in H_m\right) + \Delta_1 + \prod_{m=1}^{M} \frac{\delta_m}{1+e^{\varepsilon_m}} \\
\leq &e^{\varepsilon'} \prod_{m=1}^{M} \mathbb{P}\left(\tilde{X}_{\Xi'}(m) \in H_m\right) + \Delta_1 \\
&+ \prod_{m=1}^{M} \mathbb{P}\left(\tilde{X}_{\Xi}(m) \in H_m\right) - \prod_{m=1}^{M} \left\{\mathbb{P}\left(\tilde{X}_{\Xi}(m) \in H_m\right) - \frac{\delta_m}{1+e^{\varepsilon_m}}\right\} \\
\leq &e^{\varepsilon'} \prod_{m=1}^{M} \mathbb{P}\left(\tilde{X}_{\Xi'}(m) \in H_m\right) + \Delta_1 + 1 - \prod_{m=1}^{M} \left(1 - \frac{\delta_m}{1+e^{\varepsilon_m}}\right).
\end{aligned}
$$

That is, $\delta' = 1 - \prod_{m=1}^{M} \left(1 - \frac{\delta_m}{1+e^{\varepsilon_m}}\right) + \Delta_1$.

Case 2: $\prod_{m=1}^{M} \left\{e^{\varepsilon_m}, \frac{1}{\mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right)}\right\} > e^{\varepsilon'}$

There exists a sequence of real numbers $\{a_m\}_{m=1}^{M}$, such that,

$$
e^{a_m} \leq \min\left\{e^{\varepsilon_m}, \frac{1}{\mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right)}\right\}, \quad \sum_{m=1}^{M} a_m = \varepsilon'.
$$

Note that

$$
\begin{aligned}
&\prod_{m=1}^{M} \mathbb{P}\left(\tilde{Y}_{\Xi}(m) \in H_m\right) \\
\leq &\prod_{m=1}^{M} \min\left\{e^{\varepsilon_m}, \frac{1}{\mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right)}\right\} \mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right) \\
&+ \prod_{m=1}^{M} e^{a_m} \mathbb{P}\left(\tilde{X}_{\Xi'}(m) \in H_m\right) - \prod_{m=1}^{M} e^{a_m} \left\{\mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right) - \frac{\delta_m}{1+e^{\varepsilon_m}}\right\}.
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
&\prod_{m=1}^{M} e^{a_m} \mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_M\right) - \prod_{m=1}^{M} e^{a_m} \left\{\mathbb{P}\left(\tilde{Z}_{\Xi'}(m) \in H_m\right) - \frac{\delta_m}{1+e^{\varepsilon_m}}\right\} \\
\leq &1 - \prod_{m=1}^{M} \left\{1 - e^{a_m} \frac{\delta_m}{1+e^{\varepsilon_m}}\right\} := \Delta_2.
\end{aligned}
$$

which indicates,

$$
\prod_{m=1}^{M} \mathbb{P}\left(\tilde{Y}_{\Xi}(m) \in H_m\right) \leq \prod_{m=1}^{M} e^{a_m} \mathbb{P}\left(\tilde{X}_{\Xi'}(m) \in H_m\right) + \Delta_2.
$$

In further,

$$
\prod_{m=1}^{M} \mathbb{P}\left(\tilde{Y}_{\Xi}(m) \in H_m\right) \geq \prod_{m=1}^{M} \left\{\mathbb{P}\left(\tilde{X}_{\Xi}(m) \in H_m\right) - \frac{\delta_m}{1+e^{\varepsilon_m}}\right\}
$$

$$\prod_{m=1}^{M}\left\{\mathbb{P}\left(\tilde{X}_{\Xi}(m)\in H_m\right)-\frac{\delta_m}{1+e^{\varepsilon_m}}\right\}\leq\prod_{m=1}^{M}e^{a_m}\mathbb{P}\left(\tilde{X}_{\Xi'}(m)\in H_m\right)+\Delta_2.$$

In general,

$$\begin{aligned}
&\prod_{m=1}^{M}\mathbb{P}\left(\tilde{X}_{\Xi}(m)\in H_i\right)\\
\leq&\prod_{m=1}^{M}e^{a_m}\mathbb{P}\left(\tilde{X}_{\Xi'}(m)\in H_m\right)+\Delta_2\\
&-\left\{\prod_{m=1}^{M}\left[\mathbb{P}\left(\tilde{X}_{\Xi}(m)\in H_m\right)-\frac{\delta_m}{1+e^{\varepsilon_m}}\right]\right\}+\prod_{m=1}^{M}\mathbb{P}\left(\tilde{X}_{\Xi}(m)\in H_m\right)\\
\leq&e^{\varepsilon'}\prod_{m=1}^{M}\mathbb{P}\left(\tilde{X}_{\Xi'}(m)\in H_m\right)+\Delta_2+\left\{1-\prod_{m=1}^{M}(1-\frac{\delta_m}{1+e^{\varepsilon_i}})\right\}
\end{aligned}$$

Let $\delta'=1-\left\{\prod_{m=1}^{M}(1-e^{a_m}\frac{\delta_m}{1+e^{\varepsilon_m}})\right\}+\left\{1-\prod_{m=1}^{M}(1-\frac{\delta_m}{1+e^{\varepsilon_m}})\right\}$.
Thus, we have $\varepsilon'=\min\{\varepsilon_1,\varepsilon_2,\varepsilon_3\}$, where,

$$\varepsilon_1=\sum_{m=1}^{M}\varepsilon_m,\quad \varepsilon_2=\sum_{m=1}^{M}C_{KL}(m)+\sqrt{2\log(\frac{1}{\delta'})(\sum_{m=1}^{M}\varepsilon_m^2)},$$

$$\varepsilon_3=\sum_{m=1}^{M}\frac{(e^{\varepsilon_m}-1)\varepsilon_m}{e^{\varepsilon_m}+1}+\sqrt{\sum_{m=1}^{M}2\varepsilon_m^2\log\left(e+\frac{\sqrt{\sum_{m=1}^{M}\varepsilon_m^2}}{\tilde{\delta}}\right)}.$$

A.2.7 Proof of Theorem 6

We discuss the cases by the value of z. Note that, for any $S\in\mathbb{R}$, $z\in\left(-\infty,-\frac{\Delta}{2}\right),\mathbb{P}_z(z)\leq\mathbb{P}_z(z+v)$. Therefore, $\mathbb{P}_z(S)-e^{\varepsilon}\mathbb{P}_z(S+v)\leq 0\leq\delta$.

If $z\in\left[-\frac{\Delta}{2},0\right),z+v\in\left[-\frac{\Delta}{2},0\right)$, then $0\leq v<\frac{\Delta}{2}$,

$$\frac{p_z(z)}{p_z(z+v)}=e^{\frac{v^2+2vz}{\sigma^2}}\leq 1\leq e^{\varepsilon}.$$

Therefore, $\mathbb{P}_z(S)-e^{\varepsilon}\mathbb{P}_z(S+v)=\int_S\left(p_z(z)-e^{\varepsilon}p_z(z+v)\right)dz\leq 0\leq\delta$.

If $z\in\left[-\frac{\Delta}{2},0\right),z+v\in[0,\Delta)$, $e^{\frac{v^2+2vz}{\sigma^2}}\leq e^{\frac{\Delta^2}{\sigma^2}}\leq e^{\varepsilon}$. Therefore, $\mathbb{P}_z(S)-e^{\varepsilon}\mathbb{P}_z(S+v)\leq\delta$.

For any $z\in[0,A-\Delta]$, we still have $\frac{p_z(z)}{p_z(z+v)}\leq\frac{p_z(z)}{p_z(z+\Delta)}\leq e^{\varepsilon}$.

For any $z\in[A-\Delta,+\infty)$,

$$\begin{aligned}
\mathbb{P}_z(S)-e^{\varepsilon}\mathbb{P}_z(S+v)&\leq\mathbb{P}_z(A-\Delta)-e^{\varepsilon}\mathbb{P}_z(A-\Delta+v)\\
&=\int_{A-\Delta}^{A-\Delta+v}p_z(z)dz+\int_{A-\Delta+v}^{+\infty}p_z(z)dz-e^{\varepsilon}\int_{A-\Delta+v}^{+\infty}p_z(z)dz\\
&=\int_{A-\Delta}^{A-\Delta+v}p_z(z)dz+\int_{A-\Delta+v}^{+\infty}\left(1-e^{\varepsilon}\right)p_z(z)dz\\
&\leq\int_{A-\Delta}^{A-\Delta+v}p_z(z)dz\leq\int_{A-\Delta}^{A}p_z(z)dz=\delta.
\end{aligned}$$

Based on the above inequality, for any $S\in\mathbb{R},|d|\leq\Delta,\mathbb{P}_z(S)-e^{\varepsilon}\mathbb{P}_z(S+d)\leq\delta$, we have the conclusion of the theorem.

A.3 TRUNCATED GENERALIZED GAUSSIAN MECHANISMS

In this section, we give more details to determine an appropriate bounding parameter in noise addition. Given the privacy budget ε and δ, we are interested in deriving the minimum amount of noise added to shared parameters to achieve the highest utility while preserving differential privacy. Motivated by the observation that under (ε, δ)-differential privacy, the decay rate defined as $p(z)/p(z+\Delta)$ shall be as high as possible without exceeding e^{ε}, except for a set of points with a probability mass δ [Mironov (2017)]. We then truncate a probability density function (pdf) with an appropriate normalization factor to restore the integral over the truncated span to be 1 and calculate an appropriate scaling parameter. The required scaling parameter depends upon the privacy parameter (ε, δ). Specifically,

- The probability mass δ is equal to the area under the pdf in the last interval with length Δ over the support of pdf, i.e., the interval $[A-\Delta, A]$.
- The decay rate $\frac{p(z)}{p(z+\Delta)}$ is exactly e^{ε} for $z \in [0, A-\Delta)$.

The parameter A is then derived by solving the equations $\int_{-A}^{A} p(z)dz = 1$ and $\int_{A-\Delta}^{A} p(z)dz = \delta$.

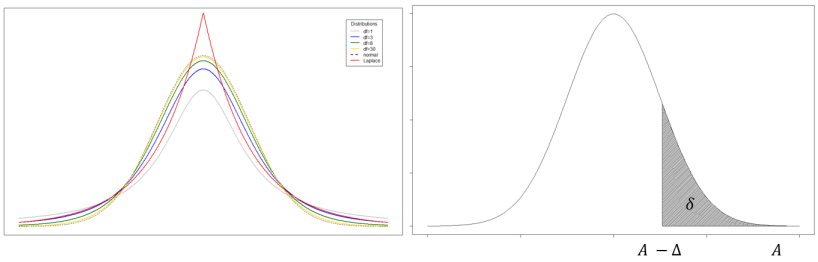

Figure 6: Noise probability density function.